ACOUSTIC MODELLING OF COCHLEAR IMPLANTS

by

Mariëtte Conning

Submitted in partial fulfillment 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

October 2005

ACOUSTIC MODELLING OF COCHLEAR IMPLANTS by

Mariëtte Conning

Supervisor: Prof JJ Hanekom

Co-supervisor: Dr T Hanekom

Department of Electrical, Electronic and Computer Engineering

Master of Engineering (Bio-Engineering)

Summary

High levels of speech recognition have been obtained with cochlear implant users in quiet conditions. In noisy environments, speech recognition deteriorates considerably, especially in speech-like noise. The aim of this study was to determine what underlies measured speech recognition in cochlear implantees, and furthermore, what underlies perception of speech in noise. Vowel and consonant recognition was determined in ten normal-hearing listeners using acoustic simulations. An acoustic model was developed in order to process vowels and consonants in quiet and noisy conditions; multi-talker babble and speech-like noise were added to the speech segments for the noisy conditions. A total of seven conditions were simulated acoustically; namely for recognition in quiet and as a function of signal-to-noise ratio (0 dB, 20 dB and 40 dB speech-like noise and 0 dB, 20 dB and 40 dB multi-talker babble). An eight- channel SPEAK processor was modelled and used to process the speech segments. A number of biophysical interactions between simulated nerve fibres and the cochlear implant were simulated by including models of these interactions in the acoustic model. Biophysical characteristics that were modelled included dynamic range compression and current spread in the cochlea. Recognition scores deteriorated with increasing noise levels, as expected. Vowel recognition was better than consonant recognition in general. In quiet conditions, the features transmitted most efficiently for recognition of speech segments were duration and F_2 for vowels and burst and affrication for consonants. In noisy conditions, listeners mainly depended on the duration of vowels for recognition and the burst of consonants. As the SNR decreased, the number of features used to recognise speech segments also became fewer. This suggests that the addition of noise reduces the number of acoustic features available for recognition.

Efforts to improve the transmission of important speech features in cochlear implants should improve recognition of speech in noisy conditions.

Keywords: acoustic model, simulation, speech-like noise, confusion matrix, biophysics, acoustic analysis.

AKOESTIESE MODELLERING VAN KOGLÊERE INPLANTINGS deur

Mariëtte Conning

Leier: Prof JJ Hanekom

Medeleier: Dr T Hanekom

Departement Elektriese, Elektroniese en Rekenaar-Ingenieurswese

Meester van Ingenieurswese (Bio-Ingenieurswese)

Opsomming

In 'n ruislose omgewing word hoë vlakke van spraakherkenning verkry met koglêere implantings. Die herkenning van spraak verminder egter drasties in omstandighede waar ruis teenwoordig is, spesifiek spraakagtige ruis. Die doel van hierdie studie is om te bepaal wat onderliggend is aan spraakherkenning in persone met koglêere inplantings, en verder ook wat onderliggend is aan spraakherkenning in ruis. Vokaal- en konsonantherkenning is bepaal vir tien normaalhorende luisteraars deur gebruik te maak van akoestiese simulasies. 'n Akoestiese model is ontwikkel sodat vokale en konsonante geprosesseer kon word in stil en ruiserige omgewings. Vir die ruiserige omgewing is multispreker-babbelklanke en spraakagtige ruis by die spraaksegmente gevoeg. 'n Totaal van sewe kondisies is akoesties gesimuleer; naamlik herkenning in ruislose omstandighede en herkenning as 'n funksie van sein-tot-ruis verhouding (0 dB, 20 dB en 40 dB spraakagtige ruis en 0 dB, 20 dB en 40 dB multispreker babbelklanke). 'n Agt-kanaal SPEAK prosesseerder is gemodelleer en gebruik om die spraaksegmente te prosesseer. Biofisiese interaksies tussen die gestimuleerde senuweeselle en die koglêere inplanting is ook gesimuleer deur modelle van hierdie interaksies in die akoestiese model in te sluit. Biofisiese eienskappe wat ingesluit is, is onder andere dinamiesebereik-samedrukking en stroomverspreidings in die koglea. Herkenning van spraak het afgeneem met 'n toename in ruisvlakke, soos verwag kon word. Vokaalherkenning was oor die algemeen hoër as konsonantherkenning. In ruislose omstandighede is die akoestiese eienskappe van spraaksegmente wat die effektiefste oorgedra word die tydsduur en F_2-formantfrekwensie van vokale en die ploffing en affrikasie van konsonante. In ruiserige omstandighede het luisteraars hoofsaaklik staatgemaak op die tydsduur van vokale en die ploffing van konsonante. Soos die sein-tot-

ruis vlakke afgeneem het, het die aantal eienskappe wat gebruik word vir spraakherkenning ook afgeneem. Dit dui daarop dat die aantal akoestiese eienskappe beskikbaar vir spraakherkenning afneem met die byvoeging van ruis. Herkenning van spraak in ruis met koglêere prosteses kan verbeter word deur die oordrag van belangrike eienskappe in spraak te verbeter.

Sleutelwoorde: akoestiese model, simulasie, spraakagtige ruis, biofisika, akoestiese analise, verwarringsmatriks.

List of abbreviations

CI	Cochlear implant
CIS	Continuous Interleaved Sampling (Cochlear implant speech processing algorithm)
dB	Decibels
FEM	Finite Element Modelling
FFT	Fast Fourier Transform
F_1	First formant
F_2	Second formant
FITA	Feature Information Transmission Analysis
HINT	Hearing in noise test
IIR	Infinite impulse response
LPC	Linear Predictive Coding
NMT	Nucleus Matlab Toolbox
pdf	Probability density function
pps	pulses per second
RMS	Root-Mean-Square
SPEAK	Spectral Peak (Cochlear implant speech processing algorithm)
SPL	Sound Pressure Level
SNR	Signal to Noise Ratio
VAF	Variance Accounted For (Multidimensional scaling parameter)

Table of contents

CHAPTER 1 INTRODUCTION

1.1 BACKGROUND AND SCOPE OF WORK

This dissertation investigates the encoding of speech with cochlear implants (CI). However, before discussing the issues concerning cochlear implants, it is necessary to familiarise the reader with the context.

In normal hearing, the outer ear picks up acoustic pressure waves. The middle ear then converts these waves to mechanical vibrations. A number of small bones are responsible for this conversion. The mechanical vibrations are transformed to vibrations in fluid in the inner ear, the cochlea. When the fluids of the cochlea undergo pressure variations, the basilar membrane undergoes displacements that contain information about the frequency and temporal information of the acoustic signal. The displacement of the basilar membrane causes hair cells in the cochlea to deform. Neurotransmitter is released when the hair cells are deformed, which causes neurons to fire, indicating that there is excitation in the inner ear at a specific cochlear place. Electric signals, generated from the firing neurons, are carried on the auditory nerve and convey information about the acoustic signal to the brain.

The auditory system cannot transform acoustic pressure waves (sound) to neural impulses when the hair cells are damaged, causing hearing impairment. The hair cells can be damaged in various ways, including by diseases such as meningitis and Meniere's disease, congenital disorders and some drug treatments (Bhatia, Gibbin, Nikolopoulos and O'Donoghue, 2004; Matsui and Cotanche, 2004). The auditory neurons can degenerate as a result of the damaged hair cells (Miller, Chi, O'Keeffe, Kruszka, Raphael and Altschuler, 1997; Ohlemiller and Gagnon, 2004; Whitlon, 2004). Loizou (1998), describes a person with a large number of damaged auditory nerves or hair cells in the cochlea as profoundly deaf.

It has been shown (Hinojosa and Marion, 1983) that the loss of hair cells, rather than the loss of auditory neurons, is the most common cause of deafness. This is a promising find, as the remaining neurons can be excited directly through electrical stimulation. The normal hearing mechanism (outer, middle and part of the inner ear) can be bypassed by a device called a cochlear implant that stimulates the auditory neurons directly. The challenge faced by researchers is to find a way in which to stimulate the auditory neurons so that useful information about speech is conveyed to the brain, i.e. amplitude, frequency and temporal information.

A cochlear implant is a device that picks up sound with a microphone and sends the speech signal to a speech processor. This processor converts sound to electric signals that are transmitted to an implanted electrode array (consisting of multiple electrodes), implanted in the cochlea by a surgeon. Different auditory nerve fibres are stimulated at different places in the cochlea, depending on the information in the speech signal. The signal processor divides the incoming signal into a number of frequency bands or channels and calculates the energy in the band; the energy determines the amplitude of the electric pulses used to activate the electrodes. The electrodes at the base of the cochlea are used to stimulate with information about high frequency signals and those at the apex are used to stimulate with information about low frequency signals. The cochlear implant is used to mimic the function of a healthy cochlea (Clark, 2003; Loizou, 1999a; Loizou, 1998; Waltzman and Cohen, 2000; Zeng, Popper and Fay, 2004).

To test the ability of a cochlear implant user to recognise speech, the recognition of sentences, monosyllabic words, vowels and consonants are normally determined (Fu, Shannon and Wang, 1998; Loizou, 1998; Tyler, Preece and Lowder, 1987). Recognition scores for sentence tests are usually higher than for other tests, owing to knowledge about grammar and context. Initially, single-channel implants were used, but performance was poor, scores for word identification ranged from 2 % to 4 %, as reported in Danhauer, Ghadialy, Eskwitt and Mendel (1990). Speech recognition was greatly improved with the introduction of multichannel implants. Recognition scores for vowels and consonants are reported as approximately 70 % (Loizou, 1998), a significantly higher figure than for single-channel implants. Today, only multichannel implants are used.

The success of cochlear implants differs among individuals, and can be affected by various factors, such as insertion depth (Dorman, Loizou and Rainey, 1997a; Faulkner, Rosen and Stanton, 2003), number of independent spectral channels (Faulkner, Rosen and Wilkinson, 2001; Friesen, Shannon, Baskent and Wang, 2001) and speech-processing strategy (Skinner, Fourakis, Holden, Holden and Demorest, 1996; Whitford, Seligman, Everingham, Antognelli, Skok, Hollow, Plant, Gerin, Staller, McDermott, Gibson and Clark, 1995). It is difficult to determine to what extent a certain factor influences speech perception because of the interaction between factors. For example, meningitis is associated with bone growth in the cochlea, which will cause an obstruction when the electrode is inserted into the cochlea. A particular question that may arise in this instance, is whether a user has poor speech recognition as a result of hair cell loss, or as a result of shallow electrode insertion.

Cochlear implants still have numerous limitations, including decreased performance in noise (Faulkner et al., 2001; Müller, Schön and Helms, 2002; ter Keurs, Festen and Plomp, 1993b). For speech recognition in noise, it becomes difficult to separate different sources of sound. It has been reported that cochlear implantees have little appreciation for music signals, indicating that information needed to perceive music is lost during processing (Koelsch, Wittfoth, Wolf, Müller and Hahne, 2004; Kong, Cruz, Jones and Zeng, 2004; McDermott, 2004; McDermott and McKay, 1997).

The problem addressed in this dissertation, is to determine what underlies measured speech recognition in cochlear implantees, and furthermore, what underlies perception of speech in noise.

1.2 APPROACH

There are numerous ways in which the effect of a specific characteristic of a cochlear implant can be examined, including speech recognition experiments with implantees and psychoacoustic experiments. One way is to change characteristics of an existing implant, such as insertion depth of electrodes or the number of spectral channels, and observe the effect. Another is to locate a cochlear implant user who has an implant processor or

electrode array with a specific characteristic and analyse the speech recognition of this particular implant user.

A more systematic way to examine the effect of implant characteristics is to use acoustic simulations. For this dissertation, the approach followed to investigate the effect of cochlear implant characteristics, is using acoustic simulations (Dorman et al., 1997a; Dorman, Loizou and Rainey, 1997b; Shannon, Zeng, Kamath, Wygonski and Ekelid, 1995; Shannon, Zeng and Wygonski, 1998). An acoustic simulation is an algorithm that processes speech exactly like a cochlear implant processor but presents sounds to normal-hearing persons. The amplitudes that result from the processing are used to modulate noisebands instead of electric current pulses for the simulation. The modulated noisebands are played back acoustically to a normal-hearing listener for recognition.

With an acoustic simulation[1], different scenarios can be set up and speech can be processed through the model. These sounds can then be played back to normal-hearing persons to determine the effect that each factor has on speech recognition. As mentioned previously, sentences, monosyllabic words, vowels and consonants are used as tests for speech recognition. Acoustic models may be used to assist engineers and other professionals in developing custom maps[2] and/or custom algorithms for cochlear implant users, or to determine to what extent a cochlear implant could be successful. By being able to hear what cochlear implant users hear, a clearer understanding can be acquired as to what underlies speech recognition. Conducting experiments becomes easier when normal-hearing persons are used instead of cochlear implant users. More normal-hearing persons are available for experiments than cochlear implant users, so that more experiments can be done in a shorter time.

[1]An acoustic model is developed in order to perform acoustic simulations on speech.

[2]A cochlear implant map refers to the specific set of parameters that the audiologist can manipulate in a cochlear implant. This includes setting the highest and lowest levels of stimulation to be used on each electrode.

There are many factors that may be considered in an acoustic simulation, as will be discussed later. A trade-off must be made between the simplicity of the acoustic simulation and its accuracy. To obtain a good approximation for a cochlear implant, it may be necessary to include complex processing, originating from the complexity of the biophysics of the cochlea. However, some of the more complex detail of the biophysics of a cochlear implant may be ignored in the simulation; still the overall operation of the simulation must be the same as that of an implant. When more factors are included, the interaction between these factors must also be determined and integrated into the simulation, making the acoustic model more complex.

The approach that will be followed to develop the acoustic model, is to perform signal processing on normal speech in exactly the same way that cochlear implant processors do. At the point where electric signals are generated for stimulation in the cochlear implant, an appropriate substitute must be used to stimulate the healthy cochlea of a normal-hearing person acoustically. A sum of noise bands with specific centre frequencies are used to simulate the electric stimulation of nerve cells in the cochlea (figure 1.1). A speech signal is reconstructed from the noise bands and played back to normal-hearing persons. Details will be discussed in chapter 3.

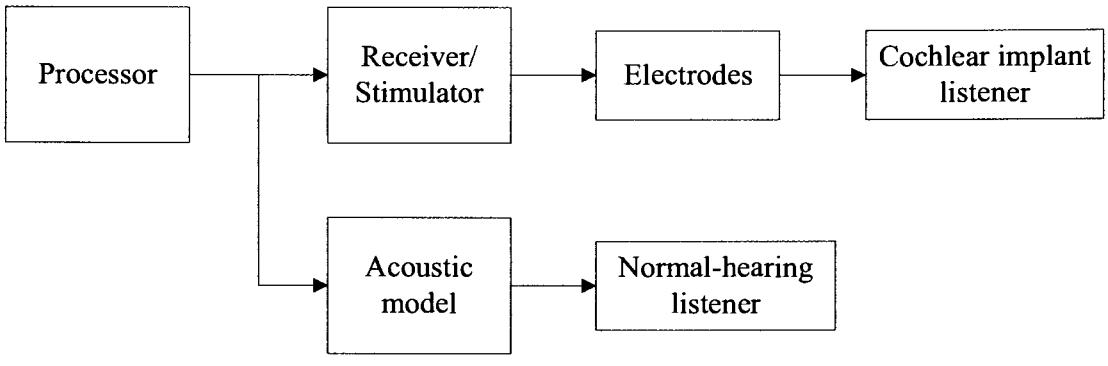

Figure 1.1. Different processing paths for electric hearing and acoustic hearing

1.3 OBJECTIVES

This study had three primary objectives. It was necessary to develop the needed tools, i.e. the acoustic model, before specific research questions could be answered. The first

objective was to develop the model, which was then used in the remainder of the study to investigate particular characteristics of the biophysics of cochlear implants. A second objective was to perform experiments with normal-hearing persons instead of cochlear implantees and thirdly, experiments in noise were performed to determine speech recognition in noisy conditions. These objectives are discussed in more detail below.

First of all, an acoustic model was developed to simulate cochlear implants as closely as possible. There are two definite components of the acoustic model – the implant-processing model and the biophysical model. The processing model can be controlled to some degree, whereas the biophysical model is user-specific and may vary in different cochlear implant users. To determine whether the model was successfully implemented, the results achieved with the model were compared with existing results. Results of vowel and consonant recognition with cochlear implant users (Fu et al., 1998; Pretorius, Hanekom, Van Wieringen and Wouters, 2005) were compared with the results achieved when normal-hearing persons listened to speech processed through the acoustic model. As results compared satisfactorily, it was assumed that the model is a good approximation of a cochlear implant. This is discussed in chapter 4.

Next, acoustic simulations were used to determine the influence on speech recognition when dynamic range compression of the stimulus current was included. These results were compared with results achieved when the compression was excluded. By performing these experiments, a clearer understanding can be achieved of the effect of using electric signals to stimulate the cochlea to produce the sensation of sound. These experiments were all conducted with normal-hearing persons after the vowels and consonants were processed through the model. Analyses of the processed vowels and consonant signals were conducted to determine the effect that dynamic range compression, spread of electric current and quantisation of current had on the signal properties of the speech sounds.

Lastly, the model was used to determine the effect that noise has on the recognition of vowels and consonants and to investigate what underlies this performance. It is evident that speech-recognition performance is poorer in the presence of noise, especially speech-like noise (Fetterman and Domico, 2002; Hochberg, Boothroyd, Weiss and Hellman, 1992;

Kiefer, Müller, Pfennigdorff, Schön, Helms, Von Ilberg, Baumgartner, Gstöttner, Ehrenberger, Arnold, Stephan, Thumfart and Baur, 1996; Müller-Deile, Schmidt and Rudert, 1995). The waveform of the speech signals was analysed to observe any major differences between the speech signals in quiet and in noise. Experiments with normal-hearing persons were conducted to determine the type of confusions present in noisy environments. From the confusion matrices, conclusions were made as to the cause of the confusions in noise.

A secondary objective of this study is to develop a tool that may be used to assist with improved designs of cochlear implants. By using the model, specific characteristics of the processor or hardware can be changed and the model can be analysed to explain why a cochlear implant user hears what he or she hears. This is beneficial in the sense that acoustic simulations are the bridge between understanding speech recognition of cochlear implant users and acoustic properties of processed speech. With an appropriate acoustic model, it becomes possible to analyse the acoustic properties of processed speech in a cochlear implant as well as speech after taking into account the biophysics of the implant. The analyses can then be used to explain speech recognition directly from signal characteristics. Figure 1.2 shows that analyses of speech after the biophysics can be done only with the use of the model. This is a valuable contribution of the model.

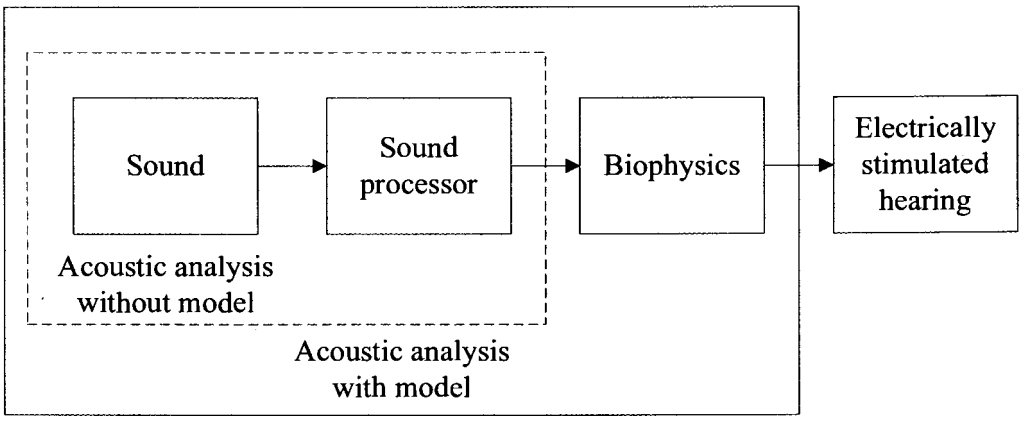

Figure 1.2. Flow diagram showing the value of the acoustic model – acoustic analyses can be performed at any stage of processing, including after the modelled biophysics, which cannot be done without the model.

1.4 HYPOTHESIS AND RESEARCH QUESTIONS

The hypothesis in the current study is that an acoustic model can provide insight into the performance of cochlear implants in quiet and noisy conditions. By developing an acoustic model that includes the processing and biophysics of the cochlear implant (figure 1.3), specific characteristics can be isolated and examined. Biophysics are a very important aspect of cochlear implants and an accurate model can give better insight into the effect that biophysics have on speech recognition. The term 'biophysics' includes several aspects relating to the electrode-nerve interface, including the spread of stimulation current and insertion depth of the electrodes.

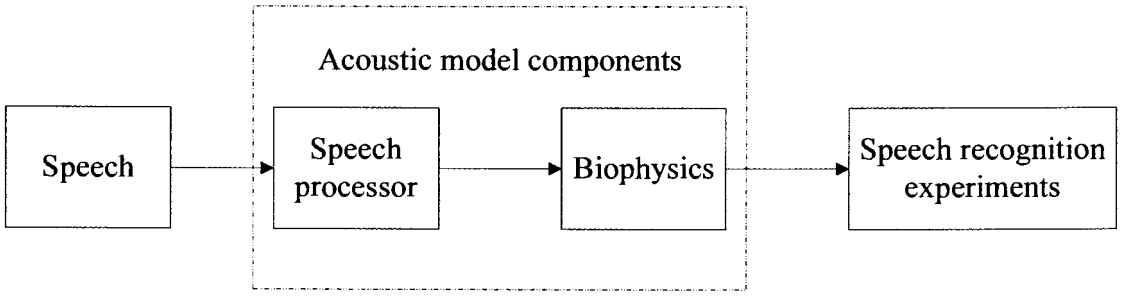

Figure 1.3. Graphical representation of the components of the acoustic model. The input to the model is speech; the processed speech is used in experiments.

It is important that the component of the acoustic model that simulates the processor of the cochlear implant, closely emulates that of the cochlear implant. To simulate the biophysics, results from the literature are used. Some of these are experimental results and others are from models of specific aspects of the cochlear implant. It is assumed that these results are accurate enough to be used in this study. The combination of the model of the cochlear implant processor and biophysics yielded an accurate model. Acoustic simulations were generated to determine speech recognition using normal-hearing persons instead of cochlear implant users. The same feature information transmission analyses were performed on the results achieved with the acoustic simulations as those performed on results achieved with cochlear implantees.

With respect to the results of the experiments before and after dynamic range compression of the stimulus current, the more complete model (inclusion of dynamic range

compression) is expected to compare better with results from experiments with cochlear implantees. The hypothesis is that the percentage correct scores for recognition of vowels and consonants will be closer to the results found with cochlear implant users with the inclusion of dynamic range compression than for the exclusion of dynamic range compression.

By assuming that the acoustic model with dynamic range compression simulates cochlear implants accurately, the study can be taken further to determine the effect of noise on speech recognition. An advantage of using the acoustic simulations in this case, is that acoustic analyses can be performed on the output signals of the acoustic model so that the underlying effect of noise on speech recognition can be determined in terms of acoustic properties.

Specifically, the research questions investigated in this dissertation are:

- What are the components needed for an accurate acoustic simulation and how should these be implemented?
- Using the developed model, what underlies speech recognition in cochlear implants?
- What are the effects on speech recognition of electric stimulation with signals that have undergone dynamic range compression?
- What are the effects of speech-like noise and white noise on the recognition of vowels and consonants?
- What are the acoustic cues for speech recognition in the presence of dynamic range compression and speech-like noise and white noise, using acoustic analyses?

1.5 OUTLINE

In the following chapters, the process of the development of a model is given. Before any development on the model could be done, a thorough background study was conducted. The technology used in cochlear implants is described, as well as the models currently used for experiments with normal-hearing persons. From the literature study, the

opportunities are identified so that the models may be implemented more accurately. The implementation of existing models is used as a starting point for the development of a more comprehensive acoustic model to be used for the acoustic simulations.

The development of the acoustic model is described in chapter 3. All the individual functions included in the model are described in detail and the reasoning behind the choice of the specific implementation is also given. The development of acoustic models for both the CIS and SPEAK strategies are described. After careful reflection it was decided that only the SPEAK strategy would be considered for the experimental study. From results found for the CIS strategy, it appeared that the high stimulation rate of the CIS strategy could not be implemented in a meaningful way. This has a direct impact on the quality of the acoustic simulation, as the high stimulation rate is one of the most important aspects of the CIS strategy. Also in this chapter, the methodology followed to do the acoustic experiments for normal-hearing persons is given.

The results from the model and experiments described in chapter 3 are reported and discussed in chapter 4. Acoustic properties obtained from the results achieved with the acoustic simulations are used to explain confusions found in the experiments. Acoustic analyses are performed on both the vowels and consonants in chapter 4 for quiet and noisy conditions. Various analyses are performed on the confusion matrices to reach a clearer understanding of what acoustic properties are important for speech recognition.

The relation of this study to the literature is discussed in chapter 5. Insights gained into various aspects are reported and the implications of what is learnt from this study are evaluated. Finally, in chapter 6, a conclusion is drawn from all the results achieved through the completion of this study. The objectives that were met with the conclusion of the study are summarised. This study has contributed to the current state of literature and this is also presented in the final chapter. Possible improvements that can be made to the model developed in this study, are stated in this chapter and any studies that might flow from this study are suggested in this chapter.

CHAPTER 2 LITERATURE STUDY

2.1 CHAPTER OBJECTIVES

The previous chapter outlined the study. The problem, to determine what underlies measured speech recognition in cochlear implantees in quiet and noise, was introduced. To solve this problem, it is necessary to gain insight into previous work. A thorough discussion of the relevant literature is given in this chapter. Gaps in the current knowledge will become apparent from the material discussed here. This chapter includes background on the modelling of cochlear implant speech processors (section 2.3) and biophysics (section 2.4 - see figure 1.3) as well as studies performed previously with the use of acoustic simulations (section 2.5).

2.2 INTRODUCTION

To simulate the effect of electric stimulation in the cochlea acoustically, two aspects of cochlear implants need to be modelled. First of all, there are the processing steps that need to be exactly like those performed in the cochlear implant processor and secondly, the biophysics must be taken into account. In the aforementioned categories, there are many parameters, such as insertion depth of electrodes and number of independent spectral channels, that are different for each user. The effect that these variables have on speech recognition in quiet and noisy conditions are investigated.

To determine what effect each of these variables have on the recognition of speech, a good understanding of each variable is vital. In the following sections, these variables will be discussed and from this background, an acoustic model is developed in chapter 3. Both the cochlear implant processing steps and the biophysics will be discussed.

As discussed in chapter 1, the implementation of the signal-processing strategy and model of biophysics are two separate components of the acoustic model. The signal-processing strategies and biophysics will therefore be discussed separately in the following section.

2.3 SPEECH-PROCESSING STRATEGIES

In earlier years, cochlear implants were implemented using single-channel implants (Danhauer et al., 1990; Hochmair-Desoyer, Hochmair and Stiglbrunner, 1985). Electrical stimulation was presented at a single location in the cochlea using one electrode. Implants have developed to such a degree that only multichannel implants are used today; general consensus has been reached that multichannel implants result in better speech recognition (Loizou, 1999b; Loizou, 1998; McDermott, McKay and Vandali, 1991).

Multichannel implants use an array of electrodes so that electrical stimulation can be provided at several locations in the cochlea (Waltzman and Cohen, 2000; Zeng et al., 2004). The tonotopic coding of frequencies in the cochlea is therefore used to be able to transfer spectral information (Greenwood, 1990). Different electrodes are used to stimulate different nerve cells. The nerve cells near the apex are stimulated with low frequency signals and those near the base are stimulated with high frequency signals.

With the introduction of multichannel implants, many questions arose (Loizou, 1999b; Loizou, 1998). These include questions such as how many channels are sufficient for high levels of speech understanding (Dorman et al., 1997b; Friesen et al., 2001) and what type of information should be extracted from speech for transmission to the electrodes (McKay, Vandali, McDermott and Clark, 1994; Whitford, Seligman, Blamey, McDermott and Patrick, 1993). To answer these questions, researchers developed different devices with a varying number of spectral channels. Typically, there is a fixed number of implanted electrodes. A subset of these electrodes is activated depending on the spectral resolution of the implemented processor.

The exact number of channels needed for good speech recognition is still under investigation; an extended study is reported in the next section (Dorman et al., 1997b; Faulkner et al., 2001; Friesen et al., 2001). Many different signal-processing strategies have been developed to extract different features from the speech signal (Kiefer et al., 1996; Loizou, 1999b; McKay et al., 1994; Skinner, Arndt and Staller, 2002; Skinner et al., 1996; Whitford et al., 1993; Whitford, Seligman, Everingham, Antognelli, Skok, Hollow,

Plant, Gerin, Staller, McDermott, Gibson and Clark, 1995a). They can be divided into three main groups, namely waveform, feature-extraction and hybrid.

Waveform strategies present a waveform, either pulsatile or analog, obtained by filtering a speech signal into separate frequency bands. A strategy is feature-extracting when spectral features or temporal features of a speech signal are presented to the electrodes (Dowell, Seligman, Blamey and Clark, 1987). To obtain these spectral or temporal features, specific algorithms are used. Hybrid strategies use a mixture of feature-extracting and waveform strategies (Whitford et al., 1995b).

Examples of the different strategies are Compressed-Analog (CA) and Continuous Interleaved Sampling (CIS) (waveform strategies), F_0/F_2, $F_0/F_1/F_2$ and MPEAK (feature-extraction strategies), Interleaved Processor, Spectral Maxima Sound Processor, SPEAK and ACE (hybrid strategies).

As mentioned in Throckmorton and Collins (2002), results for speech understanding did not differ significantly for the CIS and SPEAK strategies, suggesting that the performance of these two strategies compares well. Holden, Skinner, Holden and Binzer (1995) report a significant difference between SPEAK and MPEAK strategies, with SPEAK producing considerably better results than MPEAK. For the purpose of this study, a description of only the CIS approach and SPEAK processor will be presented. These are the strategies most widely used at present.

2.3.1 Continuous Interleaved Sampling Strategy (CIS)

The CIS strategy was originally designed to address the problem of channel interactions when stimulating all the electrodes simultaneously (White, Merzenich and Gardi, 1984). The CIS approach uses pulses that are nonsimultaneous and interleaved to stimulate nerve cells. The nerve cells are stimulated by sending biphasic pulse trains to the electrodes, with only one electrode stimulated at any given time (Loizou, 1999b; Loizou, 1999a).

The first processing step in the CIS strategy is the pre-emphasis of the higher frequencies. A pre-emphasis filter is used to attenuate the low frequencies so that the frequency band has equal loudness across the whole speech spectrum (Hartman, 1998). The signal is now passed through a bank of bandpass filters. The number of filters depends on the number of spectral channels used. To extract the envelopes of the filtered waveforms, full-wave rectification and low-pass filtering are used. The typical cut-off frequency for the low-pass filter is 200 or 400 Hz. The outputs of the filters are compressed and used to modulate biphasic pulses (Loizou, 1999b). The logarithmic compression of the signal depends on the particular user's dynamic range of electrically evoked hearing. The amplitudes of the trains of balanced biphasic pulses are proportional to the envelopes of the processed waveforms. The pulses are then delivered to the electrodes at a constant rate and in a sequential manner.

The rate of stimulation appears to have a significant influence on speech recognition – the number of pulses per second (pps) that is used varies from user to user. Some users achieve optimum performance with 833 pps while others achieve optimum performance with 1 365 pps (Loizou, 1998; Wilson, Lawson and Zerbi, 1995). Pulse rates vary from as low as 100 pps to as high as 2 500 pps. Skinner et al. (2002) report that users obtain the best speech recognition scores for stimulation rates that vary between 900 and 2 400 pps.

The order of stimulation of the electrodes does not have a large impact on speech recognition (Dorman and Loizou, 1997; Wilson et al., 1995). This parameter is user-dependent and can be changed as desired. There are various orders in which to stimulate the electrodes, including apex-to-base, base-to-apex (natural order) and staggered order. Different orders have been implemented in previous studies. In the model developed, the order of stimulation can be changed as desired (Loizou, 1998; Skinner et al., 2002).

2.3.2 SPEAK Strategy

The SPEAK strategy is an "*n-of-m*" strategy, where the speech signal is filtered into m frequency bands and the processor selects the n ($n < m$) outputs with the largest energy in the envelope in any one stimulation cycle (Loizou, 1999b; Whitford et al., 1995b). Only

the *n* electrodes corresponding to these selected outputs are then activated.

A pre-emphasis filter is used, similar to that of the CIS strategy, to produce speech that is equal in loudness across the frequency spectrum. The speech signal is then filtered into 20 frequency bands with centre frequencies ranging from 250 Hz to 10 kHz. The outputs of the filters are rectified and low-passed filtered (cut-off frequency of 200 Hz). The SPEAK processor now selects a number of maxima at 4 ms intervals to modulate the amplitude of the stimulating pulse train. The number of maxima varies between five and 10, with an average of six. "Maxima" does not necessarily refer to the amplitude of spectral peaks within the signal, but to the energy content within a frequency band. Figure 2.1 shows a block diagram of the processing steps involved in the SPEAK strategy.

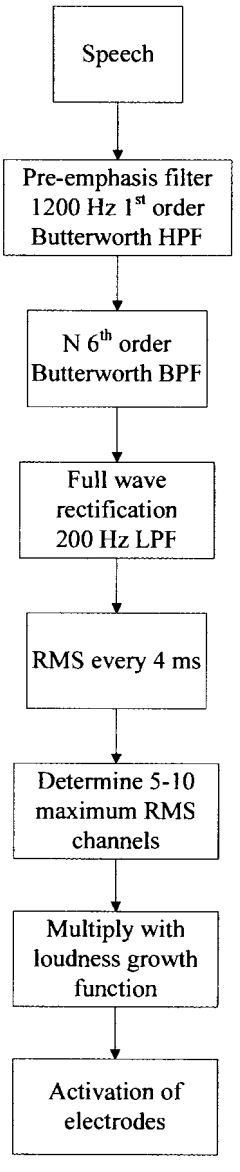

Figure 2.1. Block diagram of processing steps for SPEAK strategy

The electrodes are organised according to the tonotopic order within the cochlea; each output of the bandpass filters is allocated to a specific electrode. For example, the most apical electrode corresponds to the output of the filter with the lowest centre frequency. The stimulation rate of the electrodes varies between 180 pps and 300 pps (Loizou, 1999b) a rate of 250 pps is reported in Skinner et al. (2002).

The rate of stimulation depends on the number of selected maxima as well as the particular user's parameters. When more maxima are selected for a broader spectrum, the stimulation rate is reduced (Whitford et al., 1995b). Temporal information is increased when the spectral content is reduced and the stimulation rate is increased. Note that there is a trade-off between spectral content and temporal information.

The specific characteristics of the CIS and SPEAK strategies are important for the implementation of the acoustic simulations. The acoustic models should process speech as closely as possible to the cochlear implant processors. To summarise, a list of the important characteristics of the signal-processing model that should be included in an acoustic simulation is given below:

- Pre-emphasis filter cut-off frequency
- Bandpass filter bandwidths and centre frequencies
- Stimulation rate, which affects the interval length in which the maxima are determined
- Cut-off frequency of the low-pass filter
- Number of spectral channels for analysis
- Dynamic range compression of stimulus current
- Quantisation of stimulus current

Chapter 3 will explain how these signal-processing steps were included in the acoustic model.

2.3.3 Modelling of signal-processing strategies in existing acoustic models

Many factors affecting speech recognition in cochlear implant users have been investigated, including the number of channels needed for speech understanding (Dorman et al., 1997b; Faulkner et al., 2001; Friesen et al., 2001), insertion depth of electrodes (Baskent and Shannon, 2005; Dorman et al., 1997a; Faulkner et al., 2003), dynamic range compression of stimulus current (Fu and Shannon, 1998) and interaction between channels (Throckmorton and Collins, 2002; White et al., 1984).

Modelling the signal processing of the cochlear implant is the first step in developing an acoustic model. Before any of the existing models are discussed that deal with biophysical modelling, the modelling of signal processing strategies is described.

For the CIS speech processing strategy, the speech signals are first processed through a pre-emphasis filter for equal intensity across the frequency range (Hartman, 1998) – a second order 1 200 Hz or 2 000 Hz high-pass filter. The pre-emphasised data are filtered into L ($2 \leq L \leq 20$) logarithmic frequency bands using 6^{th} order Butterworth filters. The signal's envelope is extracted with full-wave rectification and low-pass filtering (400 Hz cut-off). The amplitudes of the sinusoids are computed from the root-mean-square of the envelopes every 4 ms. Lastly, the sinusoids are summed and played back to normal-hearing persons (Dorman, Loizou, Fitzke and Tu Z, 1998; Dorman et al., 1997b; Dorman, Loizou, Spahr and Maloff, 2002). Another way to present the processed speech signals (Dorman et al., 1997b) is to present the processed speech signal as a sum of noise bands the width of the channels. The amplitudes of the noise bands are computed in the same manner as discussed earlier.

For the SPEAK strategy, only the peaks in the short-term spectrum are used to modulate the stimulating pulses. The processing is similar to that of the CIS strategy. The number of filter banks are 16-20 (McKay and Henshall, 2002), in contrast to the 2-20 of the CIS strategy. Of these 16-20 filter banks, only the 6-8 channels with the maximum energy content are used in any one stimulation cycle. The RMS (root-mean-square) of the filtered waveforms is calculated for every 4 ms as for the CIS strategy.

2.4 MODELLING OF BIOPHYSICAL CHARACTERISTICS

Speech recognition performance of cochlear implants are influenced by the biophysical characteristics of the implant, such as the number of independent spectral channels, the insertion depth of electrodes and current spread within the cochlea. Even though standard signal-processing strategies are used for all cochlear implants, the interaction between the stimulating electrodes and the nerve cells in the cochlea vary among cochlear implant users. There are models and physical measurements that can predict general behaviour

under specific circumstances. These measurements and models are used to develop a biophysical model that simulates these interactions. Specific biophysics involved with cochlear implants and how they can be modelled are discussed in the following section.

The most important factor that should be included in the biophysical model, and also the most basic, is the number of channels. Any other factor that may be important enough to be included in the model can be added by altering this basic model. As can be seen from the literature, the insertion depth of the electrode can be simulated by altering the analysis and carrier frequencies of the basic model. Different aspects of insertion depth can be investigated using this model.

When speaking of the number of channels in a cochlear implant, one refers to the number of areas in the cochlea that are stimulated using a specific centre frequency. The number of available, independent channels plays an important role in the level of speech recognition.

For high levels of speech understanding, a minimum number of independent channels are needed for stimulation. The optimum number needed must be determined, as speech recognition against number of channels reaches a plateau at a specific number of channels. The effect of the number of channels on speech recognition has been investigated by Dorman et al. (1997b) and Shannon et al. (1995). The number of channels is difficult to determine using implant users, because of other factors that may play a role in speech understanding, e.g. the number of surviving ganglion cells. The use of acoustic simulations is ideal to examine the effect of this specific factor.

For the simulations, speech is processed with algorithms similar to those implemented in the implant processor. The processed signals are then presented to normal-hearing persons as a sum of sinusoids or noise bands. It is important to note that the only factor being varied is the number of channels; all other factors are held constant. Loizou (1998) determined that between five and eight independent channels are needed for good speech recognition. For vowel recognition, a minimum of eight channels are needed and for sentence recognition, five channels are needed. High levels of recognition will therefore

be obtained with five to eight independent channels. The number of channels that should be implemented in the acoustic model should be between five and eight.

The insertion depth of electrode arrays for cochlear implants has a distinct influence on the performance of speech understanding. Dorman et al. (1997a) explained this phenomenon with the following example: when an electrode array is inserted approximately 27 mm into the cochlea, the most apical electrode will lie near the area for 350 Hz signals. The centre frequency of the first filter in an eight-channel prosthesis will be at 350 Hz. In this case there will be little, if any, frequency mismatching. On the other hand, when the electrode array is inserted only 22 mm into the cochlea, the first filter's centre frequency (350 Hz) will be used to stimulate near the 800 Hz area of the cochlea. It is clear then that frequency up-shifting will take place, and this will have an effect on speech perception. The following figure shows the frequency bands associated with the insertion of the electrode array in the cochlea:

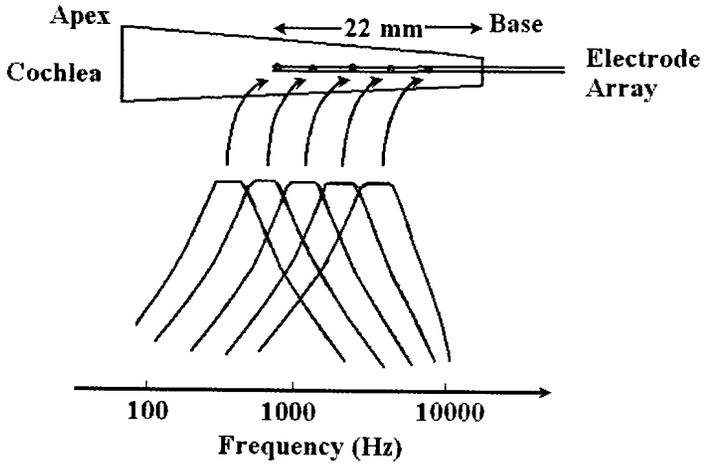

Figure 2.2. Insertion depth of electrodes

Loizou (1998) described experiments conducted to determine the effect of insertion depth of the electrode array on speech understanding, using simulations played back to normal-hearing persons. Insertion depths of 22 mm to 25 mm for the CIS strategy were simulated. Again, the only factor varied for the simulations was the insertion depth; a fixed number of electrodes was used. The signal processing is similar to that of the experiment with the

number of channels. In this case only five channels were used for the CIS strategy, as described in Dorman et al. (1997a) and Loizou (1999a). Five amplitudes were calculated at fixed frequencies that would be used in the simulation. The set of amplitudes was different for the four simulation scenarios. The sine wave output frequencies that simulated the different electrode depths were calculated using the frequency-to-place equations found in Greenwood (1990).

The results show that insertion depth has a significant effect on speech recognition. For the 25 mm and 24 mm insertion depth, reasonable speech understanding was obtained, but for 23 mm and 22 mm, recognition was very poor. This shows that when the cochlea is obstructed in such a way that the electrode array cannot be inserted deeper than 23 mm, it is possible that poor speech recognition will be achieved by the cochlear implant user.

Shannon et al. (1998) found that analysis and carrier bands must be matched in frequency. The specific cut-off frequencies of the bands are not as critical, and carrier bands that are shifted in frequency decrease performance considerably. From the work of Baskent, Shannon and Baskent (2003) and Faulkner et al. (2003) it is clear that the best speech recognition results, pertaining to insertion depth of electrodes, will be obtained when acoustic frequency information is mapped onto the appropriate cochlear place. For this study, frequency-place mapping was implemented, as described in chapter 3.

Channel interactions are a very real problem with cochlear implants. In acoustical hearing the nerve cells are stimulated tonotopically, according to the energy present at different frequencies. For good performance in speech-recognition tasks, the optimal number of channels needs to be stimulated independently. This is achieved when only the neural fibres in the immediate area of the stimulating electrode are excited. However, this does not happen. The current spreads through the whole cochlea and excites fibres that are far away from the electrode as well (Hawkins, McMullen, Popper and Fay, 1996). This causes channel interactions (Vanpoucke, Zarowski and Peeters, 2004). Channel interactions cause the number of perceptually independent frequency channels to be less than the number of electrodes available.

In Hanekom (2001), a three-dimensional finite element model (FEM) was developed to determine the potential distribution within the human cochlea. These models are complex and an approximation may be used instead for the modelling of current spread, as described in Vanpoucke et al. (2004) and Jolly, Spelman and Clopton (1996). In Vanpoucke et al. (2004) the potential distribution in the cochlea is determined for every electrode stimulated separately. The potentials at the neighbouring electrodes are measured to determine the spread of current through the cochlea as a function of distance.

Asynchronous stimulation is used in cochlear implants in an attempt to minimise channel interactions. A smaller group of nerve cells are stimulated with the biphasic pulse. This has an influence on the quality of the perceived speech signal. In the model developed by Fu and Galvin III (2001), the channels are desynchronised by introducing a delay for each channel; the delay of each channel is a fraction of the total pulse rate. The order of delay of the different channels depends on the order of stimulation of electrodes in cochlear implants. For example, for the "apex-to-basal" order, the highest frequency sinusoidal signal is delayed with the longest time delay and the lowest sinusoidal signal does not have a delay (Loizou, 1999b; Skinner et al., 2002).

Cochlear implants work on the basis of stimulating the auditory neurons directly without going through the normal mechanism for a healthy cochlea (outer, middle and part of the inner ear). Information about speech is transmitted directly to the auditory nerve connected to the brain. It is therefore very important that the auditory neurons are still intact, even though the hair cells in the inner ear may be damaged (Hinojosa and Marion, 1983; Loizou, 1999a).

The effectiveness of the cochlear implant depends on the distribution of healthy nerve cells in the spiral ganglion. If there are no living ganglion cells present at a specific place in the cochlea, it is impossible to convey any information about speech to the brain for the corresponding frequency. Hearing loss as a function of frequency is reported in Starr, Isaacson, Michalewski, Zeng, Kong, Beale, Paulson, Keats and Lesperance (2004), where typical distributions of dead ganglion cells can be observed. Every cochlear implant user has a different neural survival spread, making it difficult to assume a general survival

spread. It has been proven though that auditory cells in the basal area of the cochlea are damaged before auditory cells in the apical area, causing hearing loss for high-frequency speech components (Starr et al., 2004). This auditory nerve spread must be included in the acoustical simulation.

The biophysics of cochlear implants, as described in this section, were implemented in the acoustic model as a separate processing block (see figure 1.3). The model can also be used in further studies to investigate the effect on speech recognition of biophysical characteristics not mentioned in this chapter.

2.5 PREVIOUS RESEARCH CONDUCTED WITH ACOUSTIC MODELS

As reported in the previous section, various studies have been performed with acoustic models to determine the effects of specific characteristics of cochlear implants on speech recognition. The characteristics are either part of the speech processor or the biophysics of the cochlear implant. Specific experimental studies performed using normal-hearing listeners and/or cochlear implant users, are discussed in this section in order to demonstrate typical studies that can be performed using an acoustic model.

Dorman et al. (2002) conducted a study to determine the number of channels needed for good speech understanding in quiet and noise. A comparison was made between the CIS and SPEAK strategies. For quiet conditions, six channels for SPEAK and eight fixed channels for CIS were determined to be adequate for high levels of speech understanding. In noise, nine channels for SPEAK and 10 channels for CIS were the minimum number of channels for good understanding. Similar to the results obtained by Dorman et al. (2002) are those obtained by Friesen et al. (2001) and Fu et al. (1998). They investigated the effect of noise and spectral resolution on speech recognition. Their findings were that it is necessary to improve the effective number of spectral channels for better performance in noise. Cochlear implant users are not able to use spectral information received from the electrodes of their implant fully. These results must also be taken into account when developing an acoustic model to mimic the cochlear implant accurately.

Another factor related to the number of channels needed for good speech understanding is the minimum spectral contrast required. Loizou and Poroy (2001) conducted a study in which they used a model that manipulated the channel amplitudes of the basic model. The new model implemented a 1 – 10 dB (decibel) spectral contrast for the channel amplitudes. The results showed that cochlear implant listeners needed 4 – 6 dB more spectral contrast than normal-hearing listeners for high recognition levels.

The differences in spectral shape resolution abilities among cochlear implant listeners, and between cochlear implant and normal-hearing listeners were investigated by Henry and Turner (2003). The effect of varying the number of channels on spectral shape resolution was examined. To determine spectral shape resolution, the spacing where an interchange in peak and valley position could be detected was measured. The participants in the experiment were 21 cochlear implant users and eight normal-hearing persons. The specific experiments conducted were spectral ripple resolution and vowel recognition, using vowel stimuli recorded by Hillenbrand, Getty, Clark and Wheeler (1995). Their findings were that there is a wide variation in the ability of different cochlear implant listeners to determine spectral shapes in the acoustic signal; spectral shape resolution was poorer in cochlear implant users than normal-hearing listeners with the same number of channels; normal-hearing listeners were able to make use of more channels to determine spectral peaks than cochlear implant listeners and there was a significant correlation between vowel recognition and spectral shape resolution for cochlear implant users.

Recognition of spectrally asynchronous sentences by normal-hearing persons and cochlear implant listeners with varying spectral resolution and fine spectral structure was examined by Fu and Galvin III (2001). A CIS processor with four or 16 channels was used for the cochlear implant listeners and either a full-spectrum or a noise-band processor, also four or 16 channels, for normal-hearing persons. The output of each channel was time-shifted with respect to the other channels, with the delay ranging from 0 to 240 ms. Six normal-hearing persons and five cochlear implant users participated in the study. The Hearing in Noise Test (HINT) (Nilsson, Soli and Sullivan, 1994) sentence set was used. They found that a detailed auditory analysis of the short-term spectrum is not necessary for understanding speech. The loss of fine spectral information has nevertheless a distinct

negative effect on speech intelligibility when cross-channel spectral asynchrony is present.

Loizou and Poroy (2001), performed a study to determine the minimum spectral contrast needed for vowel identification by normal-hearing persons and cochlear implant users. The vowels used for the experiments were produced by a male speaker and selected from the vowel database used by Hillenbrand et al. (1995). For normal-hearing persons, the speech data were processed in a similar way to the CIS strategy and presented to the participants. Throughout the experiments, the peak-to-trough ratio and number of channels were varied correspondingly. It was found that minimum spectral contrast was dependent on the spectral resolution of the processed signal. Six cochlear implant users and nine normal-hearing persons participated in the study.

The effect of reduced spectral resolution and distorted spectral distribution of temporal envelope cues on consonant, vowel and sentence recognition was measured in Shannon et al. (1998). In all the experiments, an acoustical model was used to process speech similar to the CIS strategy. Eight normal-hearing persons participated in the study. The vowel and consonant tokens that were used were taken from the sound track of the Iowa audiovisual speech perception laser video disc (Tyler et al., 1987). The words used for recognition in sentences were taken from the sound track from the City University of New York laser videodisc everyday sentences (Boothroyd, Hnath-Chisolm and Hanin, 1985). The experiments included location of band divisions, frequency-shifting envelope cues, warping the spectral distribution of envelope cues and spectral smearing. The experiments showed that the exact cutoff frequencies which define the bands are not critical for speech recognition. The warping of the spectral distribution of envelope cues causes speech to be completely unintelligible. Poor intelligibility results from a tonotopic shift of the envelope pattern. Another finding from the study was that the selectivity of the envelope carrier bands was not critical for speech recognition.

The effect of channel interactions on speech recognition can be investigated with acoustical models. Different kinds of channel interactions can also be investigated and compared. In Throckmorton and Collins (2002) a thorough study of five different kinds of interactions are conducted and the results are compared. Initially experiments with

normal-hearing persons were conducted to determine the effect of pitch reversals, electrode discrimination and forward masking effects. Two additional models were later added based on the earlier results, namely pitch gap models and a modified set of forward masking models. Eleven normal-hearing persons participated in the study. The experiments conducted were vowel recognition, consonant recognition, sentence recognition (using the CID Everyday Sentence Lists) and word and phoneme recognition (using the NU #6 Monosyllabic Words Lists). The results from this study indicate that various channel interactions affect speech recognition to different degrees. The effects of channel interactions are frequency dependent, spectral interactions that affect lower-frequency information have a more significant effect on speech recognition than interactions affecting higher-frequency information.

Acoustic simulations were used to determine the effect of compression of frequency-to-place mapping on speech recognition in Baskent et al. (2003). Expansion of the frequency-to-place mapping was also measured. A vocoder was used to process consonants, vowels and sentences similar to the CIS strategy, using four, eight and 16 channels. Six normal-hearing persons participated in the study. The vowel tokens were taken from the materials recorded by Hillenbrand et al. (1995) and the sentences were taken from the TIMIT sentence materials. Results from this study suggests that speech recognition is dependent on the mapping of acoustic frequency information onto the appropriate place in the cochlea.

The recognition of frequency-shifted and spectrally degraded vowels were investigated by Fu and Shannon (1999a) for acoustic and electric hearing. Experiments with five normal-hearing subjects showed that vowel recognition is sensitive to both spectral resolution and frequency shifting. However, the effect of a frequency shift does not appear to interact with spectral resolution. The results from five cochlear implant users were similar to those from the normal-hearing subjects. The speech signals used for the experiments were taken from Hillenbrand et al. (1995).

In Friesen et al. (2001), speech recognition in noise with a varying number of spectral channels was compared for acoustic and electric hearing. Recognition in four different

signal-to-noise ratios (15, 10, 5 and 0 dB) was measured, while varying the number of spectral channels. Five normal-hearing listeners and 19 cochlear implant listeners participated in the study. Vowel and consonant recognition using the speech signals from Hillenbrand et al. (1995) was tested, as well as monosyllable word and sentence recognition using the CNC words test and HINT sentence test. The results demonstrate that most cochlear implant users are unable to fully utilise the spectral information provided by the number of electrodes used in their implant.

An interesting observation was made by Becken, Donaldson, Kimberley and Nelson (2005), who performed a study to determine the relationship between psychophysical measures of electric hearing and neural survival in cochlear implant users. Their conclusion was that there is no consistent relationship between spiral ganglion cell survival and threshold, maximum acceptable loudness level and dynamic range. The conclusion is that there is a complex relationship among electrode location, neural survival and behavioural measures of hearing. The proximity and number of surviving spiral ganglion cells have no influence on threshold, maximum acceptable loudness level and dynamic range. These measures are affected by other factors.

2.6 GAPS IN THE CURRENT LITERATURE

From knowledge obtained from literature, a number of gaps was identified. Research was not done to fill all the gaps; this study contributes to the existing acoustic models by addressing some of these issues. The model to be developed must be able to determine the effect on speech recognition of many combined cochlear implant characteristics, which does not appear to be available at present.

No literature appears to be available on models that simulate current distribution in the cochlea. This also has an effect on speech recognition and should be investigated in order to explain the influence of current spread on acoustic features. When the cochlea is stimulated with current, not only the nerve cells in the immediate area are stimulated, but also nerve cells in the area of the current spread. This causes the stimulated frequency band to increase and it is expected that a broadband signal will be perceived by the

cochlear implantee. This current spread only exists for electric stimulation; it is important to model it for acoustic hearing.

It is also apparent from the literature that the relationship between spectral resolution and electrode insertion depth is a topic that still needs a great deal of research. At the moment, the number of independent channels do not achieve their full potential, because of limiting factors such as insertion depth.

One aspect of the current study is to determine the effect that dynamic range compression has on speech recognition with specific reference to the SPEAK strategy. Dynamic range compression refers to the stimulating current levels of the implant. Every implant user has specific comfortable (C) and threshold (T) current levels. All the amplitudes of the stimulating current pulses must be in the range between T and C, resulting in dynamic range compression (Fu and Shannon, 1998). To determine the effect, experiments were conducted with speech signals processed through the developed model before and after dynamic range compression. The experiments were conducted with normal-hearing persons using acoustic simulations.

Although many studies have investigated the effect of noise on speech recognition, it appears that no one has done research on speech recognition in noise with the inclusion of dynamic range compression. Noise will contribute to the energy in a specific frequency band, increasing the stimulus current. As the energy content in a frequency band increases, the stimulus currents will approach the comfortable level and will reach a plateau. This will have a direct impact on speech recognition, as the envelope of a signal might be lost in the presence of noise.

These gaps led to the primary research question, "What underlies measured speech recognition in cochlear implantees, and furthermore, what underlies perception of speech in noise?"

2.7 DEVELOPMENT OF AN ACOUSTIC MODEL

The models proposed in the literature studied are parsimonious; a comprehensive acoustic model will be developed and evaluated to determine whether the results compare better with results found with cochlear implant users. To evaluate the model, experiments were set up involving cochlear implant users and normal-hearing persons, and the results can be compared. Results from Friesen et al. (2001) suggest that a relative comparison must be made between results from cochlear implant users and normal-hearing persons. Using absolute values is not recommended.

A good starting point for the development of an acoustic model would be to use existing models that simulate the effect of the number of channels on speech recognition. One must remember though that the number of channels is only one aspect of the functioning of the cochlear implant and acoustic simulation. A possible solution would be to give the user the option to change different factors, including number of channels, type of signal processing, dynamic range compression of stimulus current and insertion depth of electrodes.

The fundamental signal processing that will be done on speech signals to simulate any influencing factor will always be done in a similar way to that described in section 2.4. One can easily see that any model will have a specific number of channels. The other factors that may have an influence will be simulated by changing characteristics associated with a particular factor. Thus, even though a user may have the use of all available electrodes, it is still very important to understand and implement the signal processing associated with the number of channels.

Most of the models already in use implement very similar signal-processing techniques. It is therefore safe to use these fundamental models and combine or expand them to build a model that incorporates more than one factor. The challenge would be to build a cogent argument that will demonstrate the influences of all the different factors and their combination and interaction. Although two separate factors have a distinct influence on speech recognition, it is not safe to assume that one can just add the effects for the overall

evaluation of speech recognition. The studies on channel interactions have shown that there is more than one possible consequence of interaction between factors.

2.8 SUMMARY

In Chapter 2, the literature on which the study is based was summarised. The background needed for the development of this specific model is condensed into one chapter. Based on the existing models described in this chapter, a new and more comprehensive model was developed (discussed in the following chapters). Experiments conducted in previous studies were also discussed. The experiments that will be conducted in this study will be based on the procedures recorded in the studies mentioned. The method followed to develop the acoustic simulation and to conduct the needed experiments will be described in detail in the following section, chapter 3.

CHAPTER 3 METHODS

3.1 CHAPTER OBJECTIVES

Following the background given in the previous chapter, the development of a model for cochlear implants is given in this section. The section describing the methods implemented is divided into two parts, the building of the model, and the experiments done using the simulations from the model in order to meet the specific objectives. Analysis of vowels and consonants was done using the programs PRAAT and Matlab. The spectral information of the speech segments was analysed before and after being processed with the model.

3.2 INTRODUCTION

The approach followed to develop the acoustic model was to separate the model into biophysics and signal processing parts. For some aspects of the model, characteristics of previous models were incorporated. The intention was to develop a more detailed model that incorporated aspects not previously found in acoustic models. The speech processing part of the model should obviously emulate the cochlear implant speech processing as close as possible. To this end, the Nucleus Matlab Toolbox (NMT) from Cochlear Pty Ltd was used.

Cochlear Pty Ltd[1] developed this toolbox for Matlab in order to process speech similar to a cochlear implant. Current signals that can be used to activate a cochlear implant electrode array directly are generated by the toolbox in order to perform controlled experiments with cochlear implant users. The processing steps used in this toolbox are exactly the same as those in the cochlear implant processor. These steps were analysed to determine which steps are important for the acoustic simulation. As will be shown, some of the steps are particular to cochlear implant users and not appropriate in an acoustic simulation that

[1]www.cochlear.com

normal-hearing listeners will listen to. The real challenge lies in the simulation of the biophysics so that a model can be developed that can simulate what happens in the cochlea after stimulation with current pulses. To do this, several aspects of the nerve-electrode interface were considered, including current spread within the cochlea and current-loudness mapping within the cochlea.

In the sections that follow, the development of the model is described, based on the processing steps from the NMT and previously developed models. A detailed description of the processing steps of the model is given below.

3.3 DEVELOPMENT OF AN ACOUSTIC SIMULATION

3.3.1 Processing steps in Nucleus speech processor as performed in the NMT

The NMT implements two types of speech-processing strategies as they appear in the Nucleus cochlear implant processor. The one focuses on temporal information in the speech signal (CIS) while the other focuses on spectral information (SPEAK). The processing steps of the SPEAK strategy were analysed and included in the model, with modifications introduced. The steps for CIS and SPEAK are summarised below as implemented in the NMT. Typically, six channels are used for the CIS processing strategy; this is less than for SPEAK in order to have a higher stimulation rate (Loizou, 1999b). For SPEAK, the incoming speech is divided into 20 frequency bands and the eight channels with the highest energy content in any one stimulation cycle are used for stimulation.

The speech signal is divided into time windows with an overlap of 75 %. These are windowed with a Hanning window to prevent sharp transitions in the speech signal in the time domain. When blocks of speech are extracted from a signal in the time domain, the speech signal is in effect modulated with a square wave, introducing broadband frequency components to the speech signal's spectrum. This is due to the broadband frequency response of a square wave. By using Hanning windows, the speech signal is modulated with a smooth function, minimising the spectral spread. The length of the windows is

fixed at 8 ms. The signal is then divided into frequency bands using a Fast Fourier Transform (FFT); the frequency bins are predetermined for both CIS and SPEAK. There are six CIS frequency bands and as mentioned before, 20 frequency bands for SPEAK. Note that in the developed model, a number of bandpass filters were used instead of frequency bins, based on existing models (Loizou, 1998).

The energy content in each band for a specific time window is determined next. Each frequency band has an amplitude gain according to the characteristics of the implant for a specific user; these are applied with the determination of energy content. The length of one time window is 128 samples (8 ms at a sampling frequency of 44.1 kHz). Because of the 75 % overlap, new samples are available every 2 ms. The overall maximum stimulation rate is 14 400 pps in the Nucleus speech processor. The stimulation rate of a single electrode depends on the number of channels[2] in use. For example, for a six-channel implant, the stimulation rate of a single electrode will be 14 400 pps divided by six, which is 2 400 pps. For the CIS strategy all the channels are used for further processing. Only the eight channels with the highest energy content per time window is used for the SPEAK strategy for further processing. The corresponding channel position is retained for the SPEAK strategy for when the electrodes are activated at a later stage.

The calculated energy levels, described above, are mapped to current levels that will be used to stimulate the nerve cells in the cochlea. The maximum current level is the comfort (C) level and the minimum level is the threshold (T) level. Every cochlear implantee has a specific comfort and threshold current level for each electrode pair. The threshold level of each electrode pair is the minimum current value for a just-audible stimulus. The comfort level is the maximum current value that can be used for stimulation before it becomes uncomfortably loud. These values are user-specific and can be changed within the NMT (also in the acoustic simulation). Usually these are set by the audiologist and are commonly known as a "map". In the NMT the current levels are clipped so that all values fall within the C and T levels.

[2]'Channel' refers to two electrodes in an electrode array, one electrode is the stimulating electrode and the other is the return electrode for the stimulating current

A logarithmic loudness growth function is next applied to the computed current values in the NMT. (This is in fact not a loudness growth function even though the NMT calls this function as such.) This step is not included in the acoustic simulation, as it is shown in the next section that the loudness growth function linearises the current-loudness function only when stimulating nerve cells with current signals in cochlear implants.

The stimulation order for a cochlear implant can be changed as needed; by default the channels are activated from the most basal position to the most apical position. In the NMT (in the acoustic simulation as well), the order of stimulation is set to the default order by sorting the channels according to their centre frequency. The final step in the processing is to map the current values, as determined in the previous steps, to particular electrodes that will stimulate specific places in the cochlea.

3.3.2 Processing steps in the acoustic model

The processing part of the acoustic model must be as close as possible to the actual signal processing done by the processor of a cochlear implant. All the filtering, amplitude calculations, frequency mapping and other steps must be performed exactly as the processor does. The acoustic model was programmed in Matlab. The speech signals used for the experiments were processed by Matlab code and the output saved as a .wav file for easy manipulation afterwards.

The following block diagram, figure 3.1, gives the building blocks of the acoustic model, followed by a detailed description of each processing step.

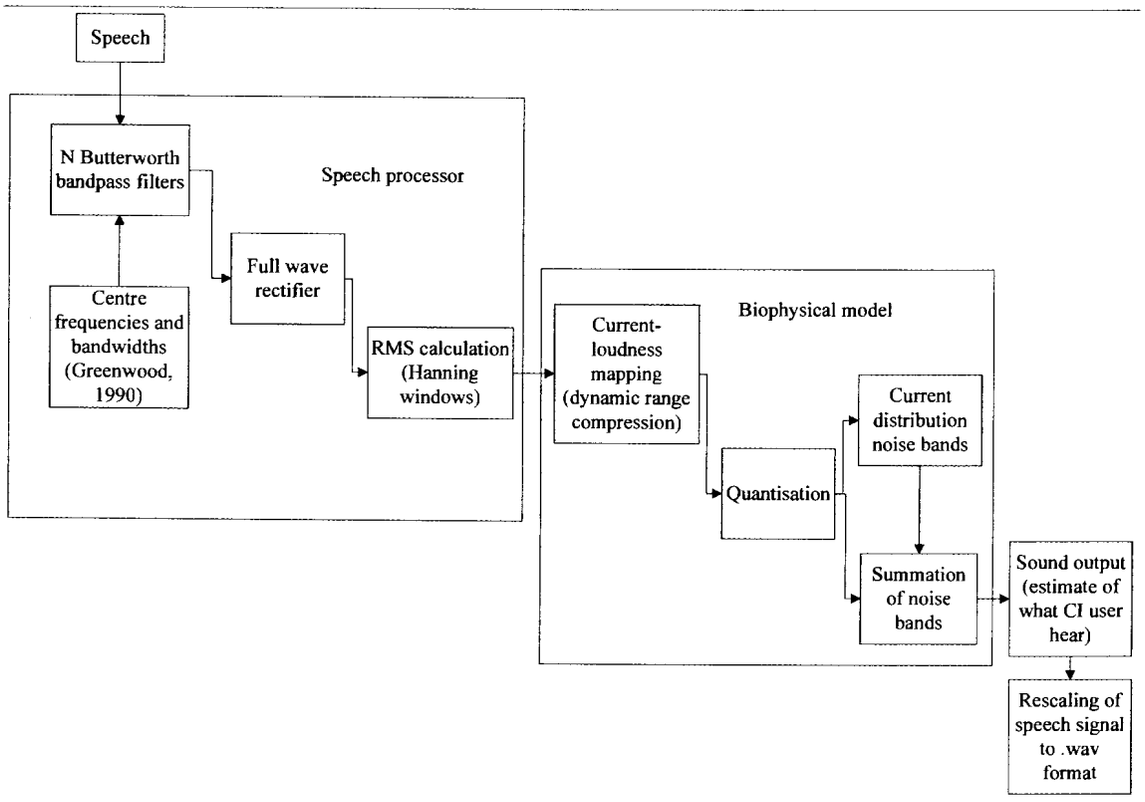

Figure 3.1. Block diagram of basic acoustic model

Examples of the output of the processing steps used in the model are shown where possible. The processing was done on an English sentence, "The fire is very hot". The original signal is shown in figures 3.2 and 3.3 for the time and frequency domain.

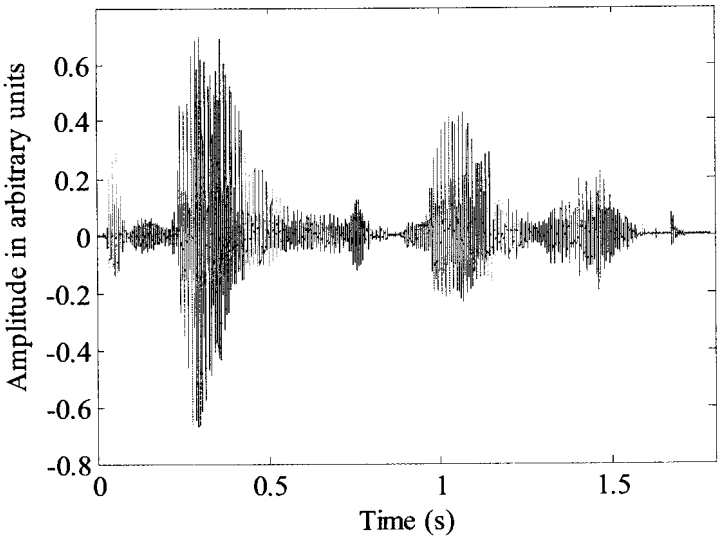

Figure 3.2. Time domain representation of the original sentence, "The fire is very hot"

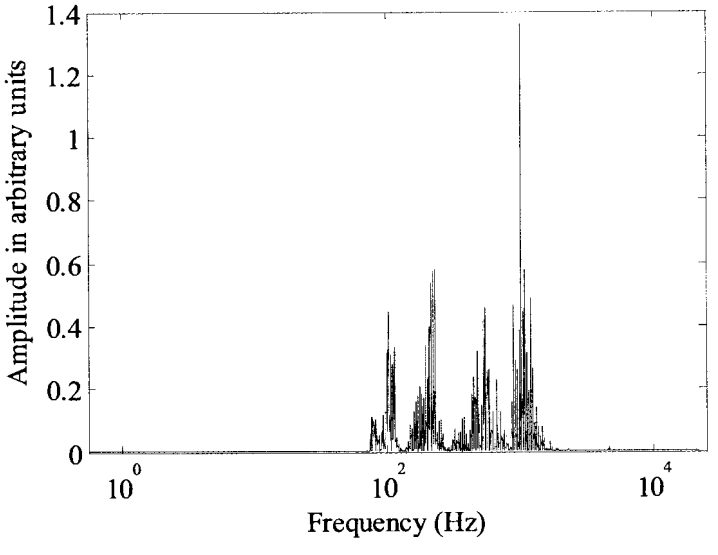

Figure 3.3. Frequency domain representation of the original sentence, "The fire is very hot"

3.3.2.1 Bandpass filters

The original data are read from a .wav file and then filtered into 20 different frequency bands. The bandwidths and centre frequencies are determined by Greenwood's frequency-to-place equation (Greenwood, 1990) for an insertion depth of 25 mm and a distance of 0.75 mm in between electrodes. Greenwood's frequency-to-place equation, as calculated

for human cadavers, is

$$f = 165.4 \times \left(10^{0.06 \times (-d+35)} - 1\right) \text{ Hz}, \qquad (3.1)$$

where f is the frequency corresponding to a specific distance (d) from the base of the cochlea. The 20 bandpass filters are sixth-order Butterworth filters, which were used because of their flat bandpass response. The number of filters corresponds to the number of places in the cochlea that will be stimulated with the electrode array. In table 3.1, the upper and lower -3 dB cut-off frequencies of each frequency band are summarised. The filter was implemented as an IIR filter. A typical transfer function is shown in figure 3.4. An example of the output of the bandpass filters is shown in figures 3.5 and 3.6.

Table 3.1. -3 dB cut-off frequencies for bandpass filters as determined by Greenwood's frequency-to-place equations; frequency bands are defined for bipolar stimulation

Channel	Lower cut-off frequency (Hz)	Upper cut-off frequency (Hz)
1	460	528
2	528	604
3	604	688
4	688	781
5	781	884
6	884	999
7	999	1 126
8	1 126	1 267
9	1 267	1 423
10	1 423	1 597
11	1 597	1 789
12	1 789	2 003
13	2 003	2 239
14	2 239	2 502
15	2 502	2 793
16	2 793	3 116
17	3 116	3 474
18	3 474	3 871
19	3 871	4 312
20	4 312	4 801

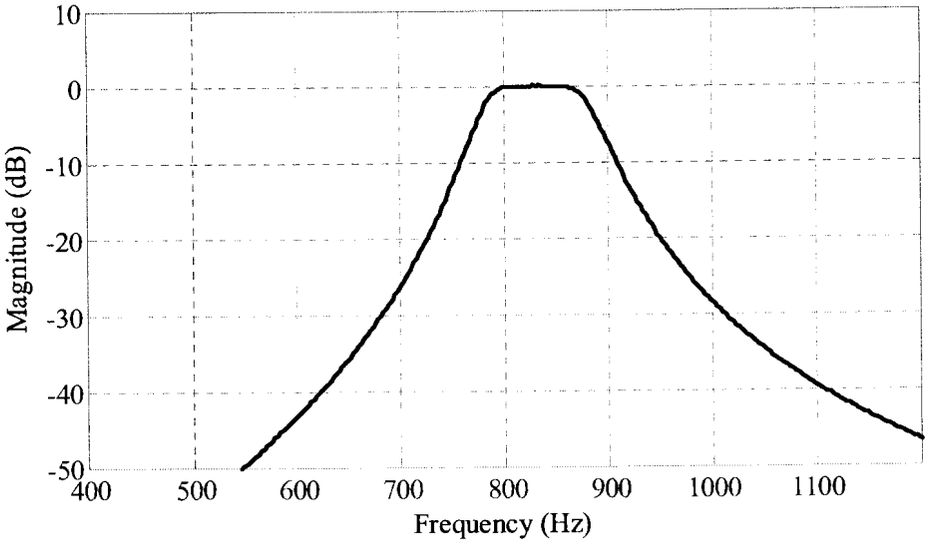

Figure 3.4. Transfer function of bandpass filter for channel 5

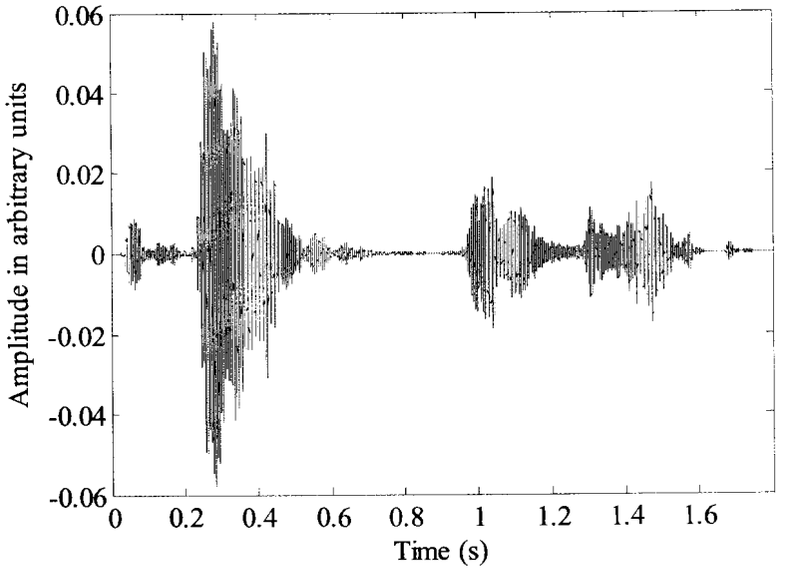

Figure 3.5. Example of time domain representation of bandpass filtered speech for channel 5

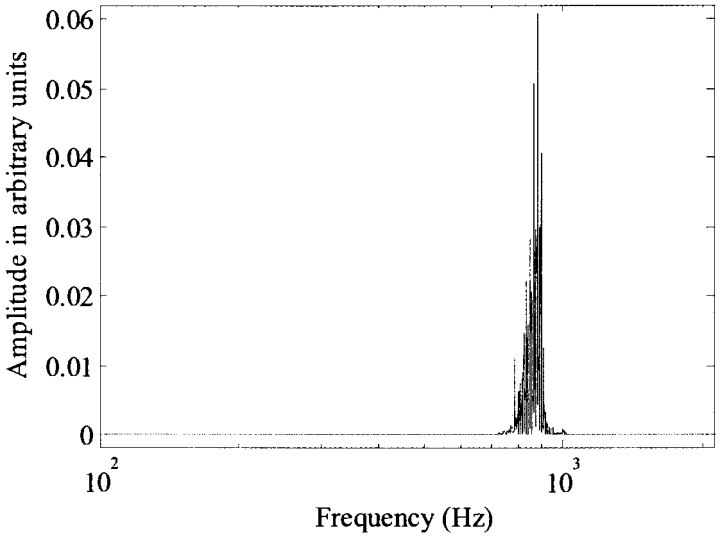

Figure 3.6. Frequency domain representation of bandpass filtered speech for channel 5

3.3.2.2 Calculation of energy in each band

An indication of the energy content in each analysis band is determined by full wave rectification and root-mean-square calculation. The envelope of each band is extracted by full wave rectification using the equation

$$A_{FWR} = |A_{BPF}|, \tag{3.2}$$

where A_{BPF} is the bandpass filtered signal amplitude and A_{FWR} is the full wave rectified amplitude. When full wave rectification is performed on a signal centred at f_c, the effect is similar to multiplying the signal with itself (squaring function). This in effect is modulation of the signal with itself, causing frequency components to appear at $(f_c - f_c)$ Hz and $(f_c + f_c)$ Hz. Full wave rectification of the original speech signal will result in components at the double frequencies and at 0 Hz. This is exactly the reason why full wave rectification is used. In lowpass filtering of the full wave rectified data, the only components that will remain are the components mixed down to lower frequencies. From the low-frequency components, an indication of the energy content in the speech signal is obtained by determining the RMS of the signal envelope.

A second order Butterworth lowpass filter with -3 dB cut-off frequency at 400 Hz is used to filter the rectified data. This will remove the double frequency components originating from the full wave rectification. The IIR filter is implemented in Matlab and a typical transfer function is shown in figure 3.7. In figures 3.8 and 3.9, the time and frequency domain data are shown for a speech signal.

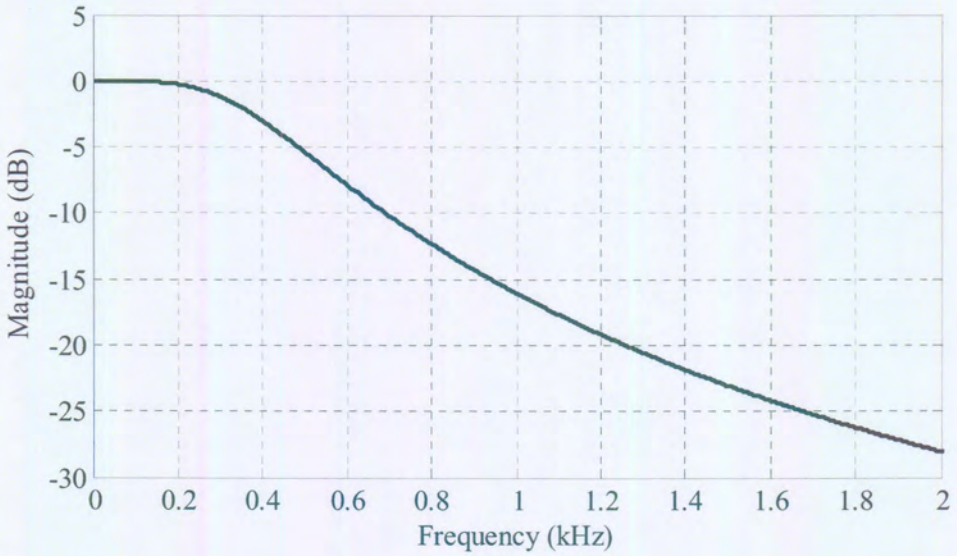

Figure 3.7. Transfer function of 400 Hz lowpass filter

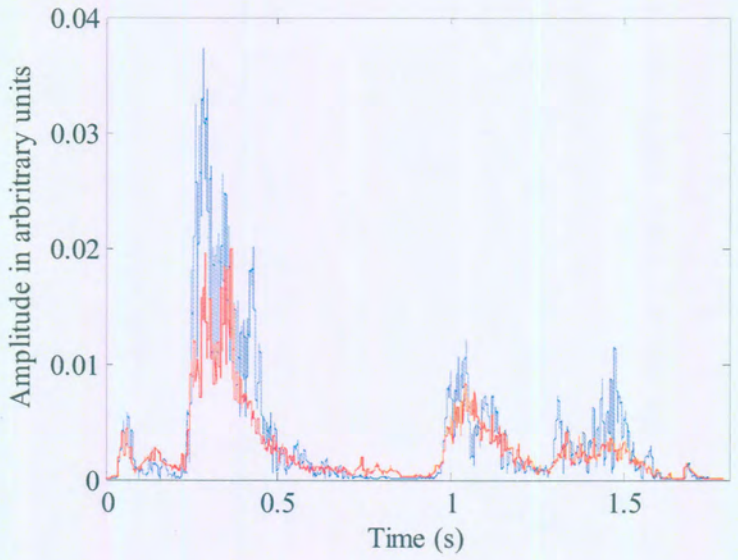

Figure 3.8. Example of time domain representation of full wave rectified speech for channel 5. Values in red are the calculated RMS values in a 2 ms window

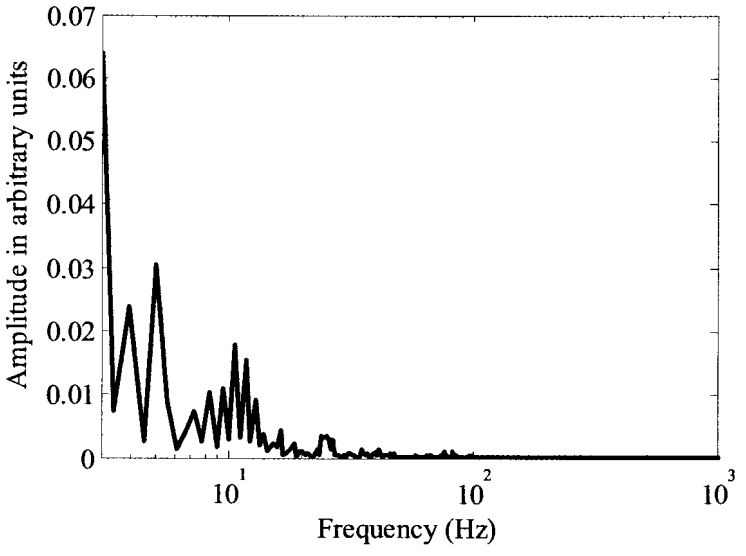

Figure 3.9. Example of frequency domain representation of full wave rectified speech for channel 5

3.3.2.3 Root-mean-square calculation

An indication of the energy content in the signal is calculated by determining the root-mean-square of the full wave rectified data for each channel. Each channel is divided into a number of time windows with 75 % overlap and the RMS is calculated for every window. This value represents the energy content in a specific frequency band for a given window of time. The equation used to calculate the RMS is

$$A_{RMS} = \sqrt{\frac{1}{N}\sum A_{FWRi}^{2}}, \qquad (3.3)$$

where N is the number of samples in the full wave rectified data vector and A_{FWRi} is the i^{th} data point in the vector of full wave rectified samples.

By extracting blocks of data from the speech signal, the signal is in effect modulated with a square wave, changing the spectral shape of the speech signal over a broad frequency band. The spectrum of a Hanning window has a narrow frequency band, reducing the spread of frequencies when extracting windows. The Hanning windows ensure that there are no high

frequency components present in the speech signal that will cause a click when listening to the processed signal. In the time domain, the signal is smoothed by the Hanning windows so that there are no abrupt transitions from one window to another. The length of the windows is fixed to be 8 ms long, irrespective of the number of channels used and the stimulation rate. The number of samples in a window depends on the sampling frequency of the original signal and is calculated using the equation

$$N_{samples} = 8ms \times \frac{1}{(1/f_s)ms/sample}, \tag{3.4}$$

where $N_{samples}$ is the number of samples in the Hanning window and f_s is the sampling frequency measured in kHz. For the speech signals processed in this study, there were 353 samples in an analysis window. The weights of the Hanning window are shown in figure 3.10. The full wave rectified vector is divided into blocks of 353 points and multiplied with the values shown in this graph. Calculated RMS values are shown in figures 3.11 and 3.12.

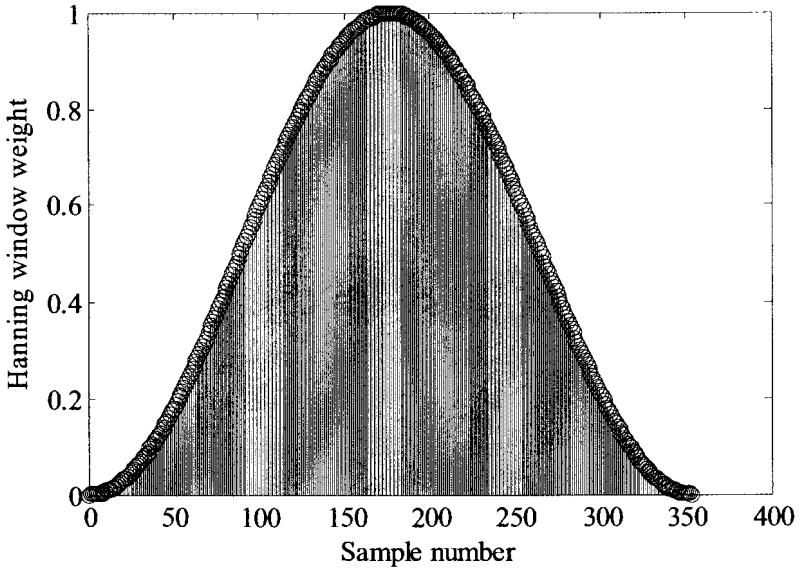

Figure 3.10. 8 ms Hanning window with 353 samples for a sampling frequency of 44.1 kHz

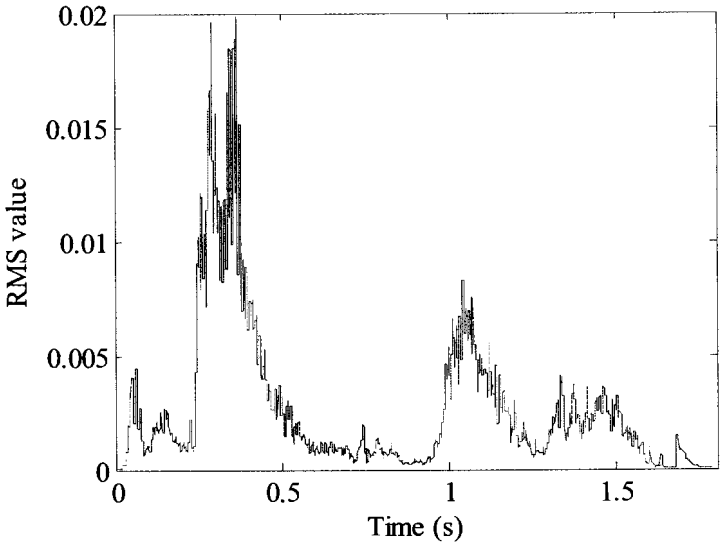

Figure 3.11. Calculated RMS values for speech signal as in figure 3.8

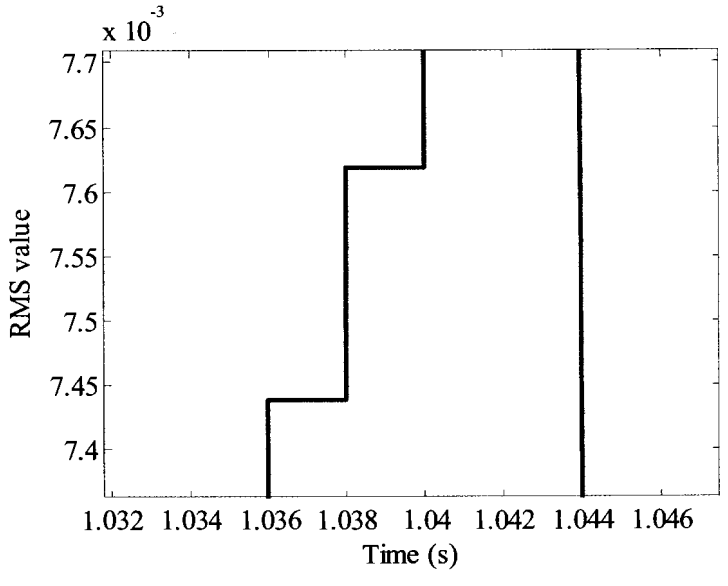

Figure 3.12. A closer look at three RMS values; it can be seen that the RMS values remain constant for 2 ms (effective length of Hanning windows)

3.3.2.4 Current to loudness mapping

The RMS values calculated in the previous step must be mapped to stimulation current magnitudes. This is done so that processing can be performed with the current values further on in the model, including the determination of the bandwidths of the noise bands

(to be explained later) used to reconstruct the speech signal. The signal energy, reflected in the RMS values, is converted to signal intensity (dB SPL) using the equation

$$10 \times \log(A_{RMS}) . \tag{3.5}$$

These values are then mapped to current values ranging from a comfort level (C) to threshold level (T), over a range of 30 dB, reflected in figure 3.13 (Fu and Shannon, 1998; Zeng, Grant, Niparko, Galvin III, Shannon, Opie and Segel, 2002). This 30 dB is used owing to the reduction in dynamic range for cochlear implants. The dynamic range can comprise a range of current levels, according to the comfortable and threshold current levels. Typical values are 1 mA for C and 100 μA for T (Bruce, White, Irlicht, O'Leary, Dynes, Javel and Clark, 1999). In the acoustic simulation, the maximum amplitude of all the channels in dB SPL is used to normalise the amplitude vector of each channel separately so that the maximum normalised amplitude across all channels is 0 dB, as indicated in figure 3.14. For the mapping of signal intensity to current levels, C corresponds to 0 dB SPL and T to -30 dB SPL. The values of C and T can be changed by the user or audiologist.

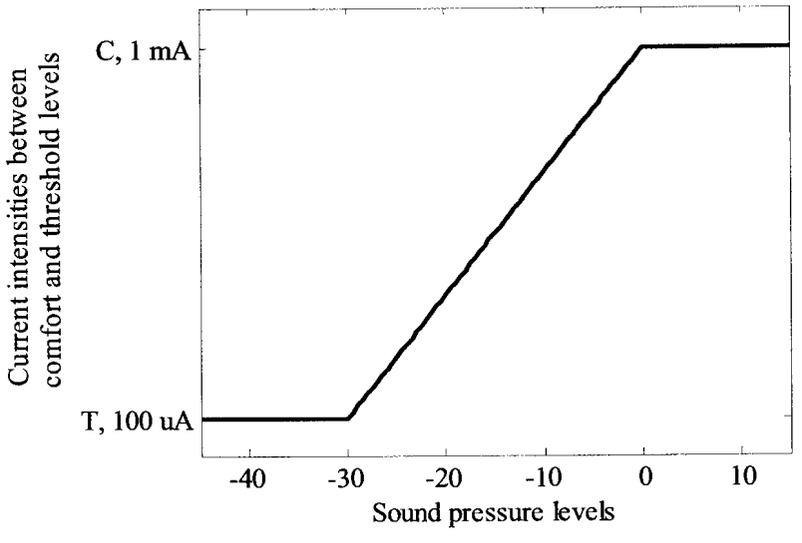

Figure 3.13. Sound pressure levels mapped to current intensities on a log scale (example values are used for T and C)

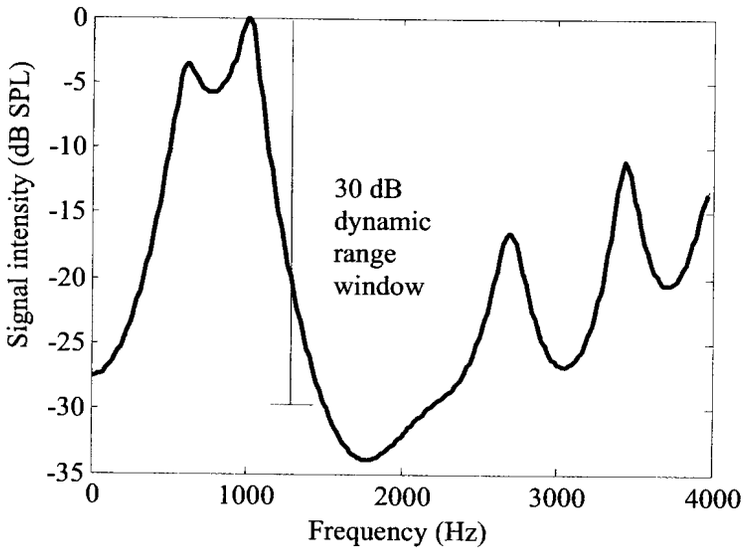

Figure 3.14. Example of frequency response of speech signal that shows the normalisation of the processed speech and the 30 dB dynamic range window

Current levels calculated in this processing step are used to stimulate the cochlea in a cochlear implant to produce a hearing sensation. It is important for these current values to be used to 'stimulate' the normal-hearing person with the dynamic range compression included. For the cochlear implant, the dynamic range is compressed because of the limits on the currents that can be used for stimulation, for example the individual's comfort and threshold levels as well as the quantisation of current. Linear mapping is used for this step. The loudness growth function applied in the NMT is not included here, an explanation for the exclusion of this step is given in the next paragraphs.

The current-loudness function for cochlear implants was obtained from Chatterjee (1999). The effect of applying the loudness growth function is shown in figures 3.15 to 3.17. A current-loudness graph for cochlear implant users is shown in figure 3.15. When the stimulation pulses are processed through the loudness growth function (figure 3.16), the result is a linear current-loudness relationship (figure 3.17) for cochlear implants. In Chatterjee (1999), a number of exponents for the current-loudness function were reported. For this demonstration an exponent of 0.02 was used.

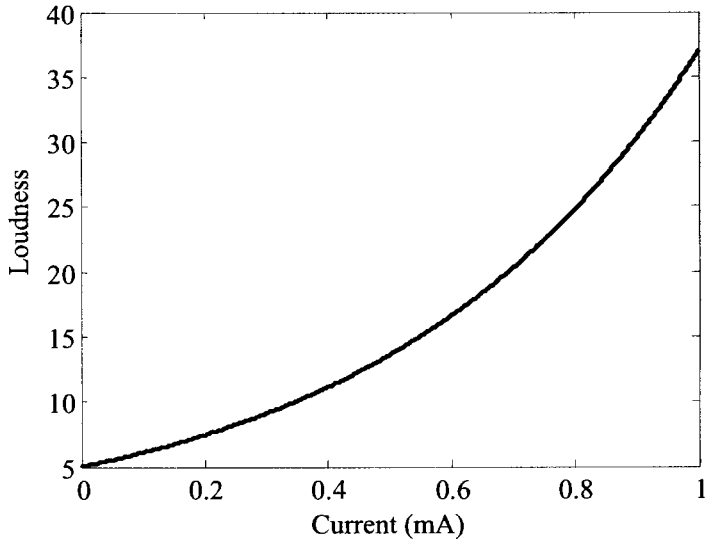

Figure 3.15. (Log) Loudness perception of cochlear implant users as a function of stimulus current

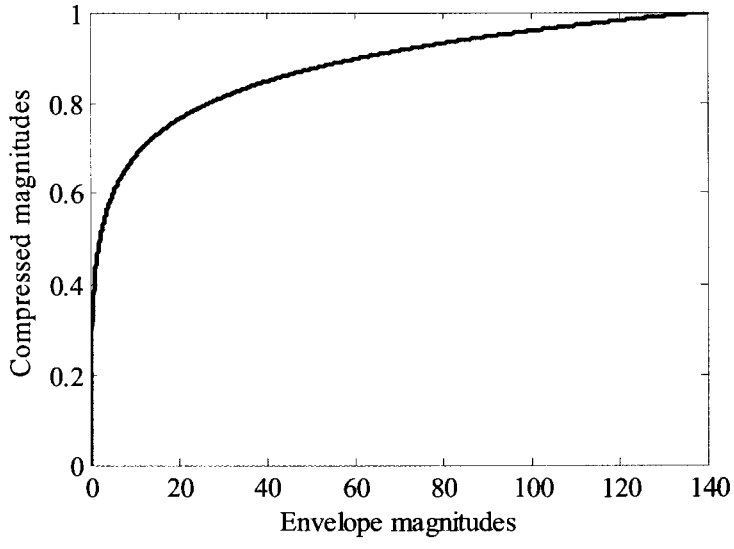

Figure 3.16. Loudness growth function (obtained from the NMT) applied to envelope magnitudes to linearise the relationship between stimulus current (proportion of dynamic range) and perceived loudness for cochlear implant users

The equation used for the loudness growth function, as obtained from the NMT, is

$$CM = \frac{\log(1 + LGF\alpha \times EM)}{\log(1 + LGF\alpha)}, \qquad (3.6)$$

where CM is the compressed magnitudes (fraction of dynamic range), EM is the envelope

magnitudes and LGFα is a loudness growth parameter obtained from the NMT.

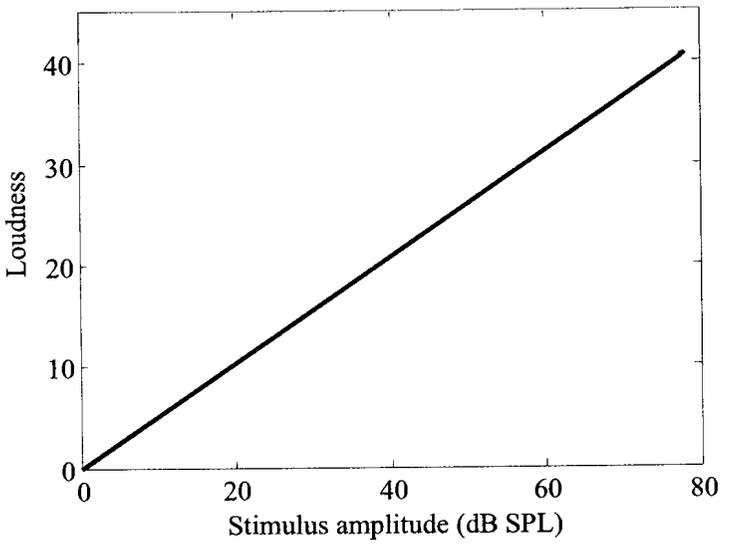

Figure 3.17. Resulting linear relationship of current-loudness (log) distribution

The equation used to obtain the linear relationship is

$$L = 20\log(e^{\beta \times CM}), \tag{3.7}$$

where L is the loudness, β is a current-loudness exponent and CM is the compressed magnitudes from equation 3.6.

For normal-hearing listeners, the exponential current-loudness relationship (figure 3.15) does not exist, therefore the processing done by the loudness growth function is not necessary for the acoustic simulation. There is already a linear relationship between stimulus amplitude (in dB SPL) and loudness (Hartman, 1998). The output of this step in the acoustic simulation is a vector with current amplitudes that are clipped at specific values to simulate the limitations of cochlear implant stimulation (see figure 3.18).

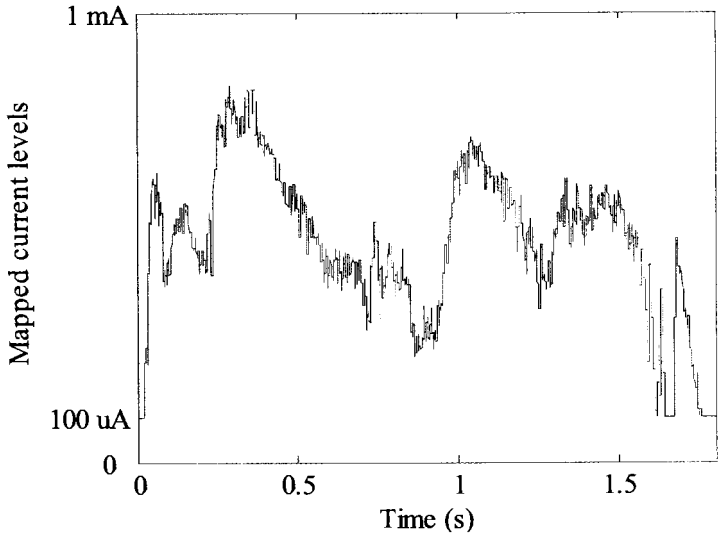

Figure 3.18. Sound intensity mapped to current levels for the same speech signal as in figure 3.2 (channel 5)

3.3.2.5 Current distribution - Noise bands

To reconstruct the speech signal, bandlimited noise bands are used. When an electrode is activated in the cochlea, not only the nerve fibres in the immediate area will be stimulated by the current pulses, but also nerve fibres some distance away. The use of bandlimited noise bands is a good approximation to the spread of current in the cochlea. The noise bands have specific centre frequencies (to simulate a specific place in the cochlea) and bandwidths (to simulate the current spread in the cochlea). The centre frequencies of the noise bands are the same as for the analysis bands, the bandwidths for the noise bands are wider than for the analysis bands. The bandwidth of the noise bands depends on the intensity of the current used to stimulate with an electrode, the spread in the cochlea for normal listeners is not taken into account. For the noise bands, the upper -3 dB cut off frequency of one band is the lower -3 dB cut off frequency of the next frequency band. The current threshold where a nerve will be activated is determined by the distance of the electrode from the nerve cell and the current distribution in the cochlea. The current distribution depends on the stimulation mode. In this simulation bipolar configuration is assumed, so that the current decay is assumed to be 4 dB/mm (Bruce et al., 1999).

For an electrode 1 mm away from the nerve cells and a threshold current of T = 100 μA,

the current level 4 dB below 100 μA will be the minimum current that will still yield an audible sound. The current level is 63.1 μA for the assumed distance of 1 mm; the distance can be changed in the acoustic simulation. For mapped current values above 63.1 μA, the bandwidth for the specific channel will be non-zero, for values below 63.1 μA, the bandwidth will be zero and there will be no stimulation for this specific channel. The magnitude of the stimulation current can be used dynamically to determine the bandwidth of the noise band in mm with the equations

$$I_T(dB) = T(dB) - 4(dB/mm) \times D_E \tag{3.8}$$

$$BW = \left| \frac{(I - 4(dB/mm)) \times D_E - I_T(dB)}{4} \right| \qquad I \geq I_T, \tag{3.9}$$

where I_T is the minimum current at the nerve cell, T is the threshold current (i.e. 100 μA), D_E is the distance of the nerve cell from the activated electrode, I is the stimulating current at the nerve cell and BW is the bandwidth. The bandwidths and centre frequencies of the noise bands are calculated in mm and translated into Hz using Greenwood's frequency-to-place equations. For all $I < I_T$, the bandwidth is zero. The current decay is demonstrated in figures 3.19 and 3.20 for the case of threshold stimulation and above-threshold stimulation.

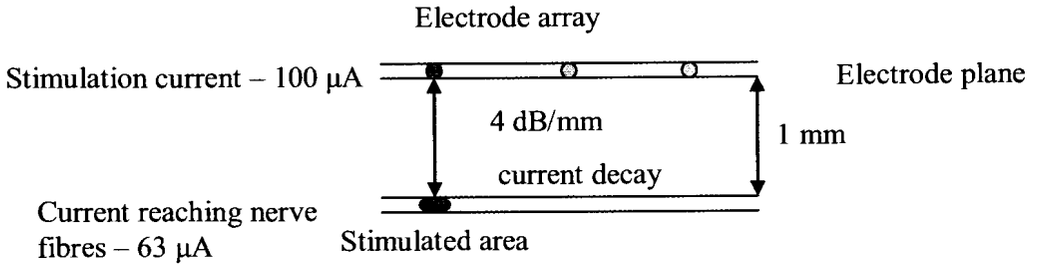

Figure 3.19. Figure to demonstrate the current decay in the cochlea for threshold stimulation (I_S = 100 μA)

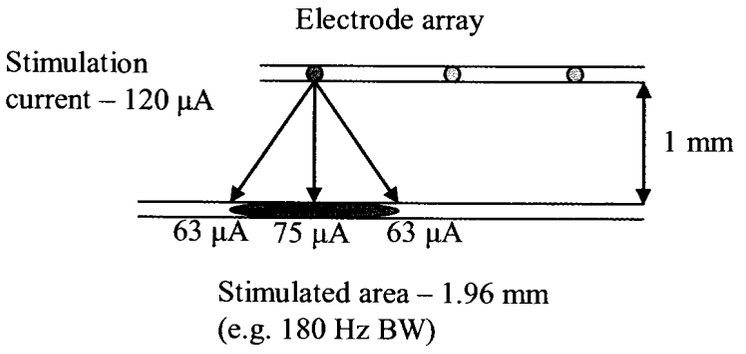

Figure 3.20. Figure to demonstrate the current spread in the cochlea for above-threshold stimulation (I_S = 120 μA)

In the actual implementation, the bandwidth was not varied. The average and standard deviations of the bandwidths of the channels were calculated for a number of sentences. It was found that the standard deviation of the bandwidth of a channel is less than 5 % of the centre frequency when the bandwidths are calculated dynamically using the stimulation current. This is a very small percentage, given that implementing a different order filter for the bandpass filters can change the bandwidth of a channel by 5 %. It was therefore decided to calculate and use the average bandwidths across a number of speech signals and remove the dynamic allocation of bandwidths because of the almost negligible effect and computational overhead introduced. The bandwidths used for the channels at an insertion depth of 25 mm are given in table 3.2. Figure 3.21 shows a band-limited white Gaussian noise signal used in the summation of all the channels. A broadband noise signal was generated in Matlab and filtered with a sixth-order Butterworth bandpass filter.

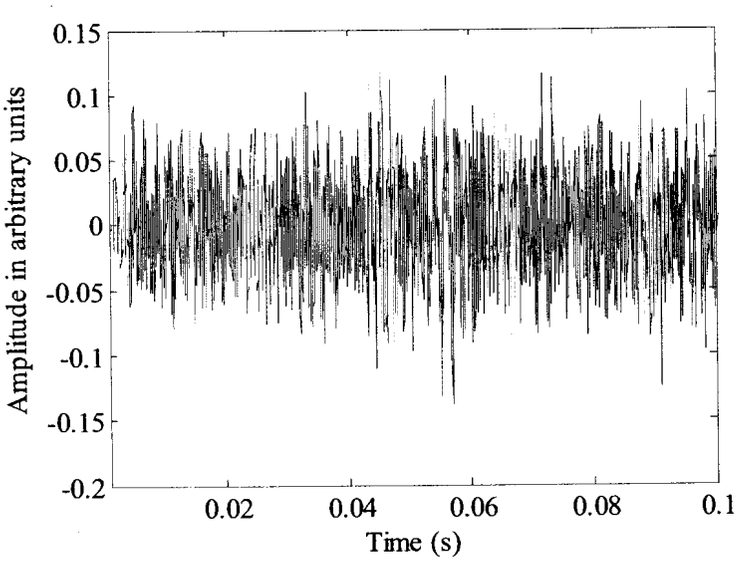

Figure 3.21. Example of the time domain representation of bandlimited noise for channel 5

Table 3.2. Bandwidths used for filters to generate noise bands

Channel	Bandwidth (Hz)		Channel	Bandwidth (Hz)
1	90		11	275
2	105		12	305
3	120		13	340
4	133		14	375
5	150		15	420
6	165		16	460
7	180		17	515
8	200		18	570
9	225		19	630
10	250		20	700

As a comment, another way of simulating the current distribution in the cochlea is by frequency modulating a sinusoidal carrier signal. The centre frequency for the sinusoidal carriers are chosen according to the frequency-to-place mapping of the processed signal. The frequency of the modulating signal depends on the magnitude of the stimulating current. The bandwidth is therefore effectively equal to the frequency of the modulating signal. The modulated signal will have energy at the carrier frequency and at the carrier

frequency plus and minus the modulating frequency (Proakis and Salehi, 2002). This is shown in figure 3.22. This approach was not followed, however, as the use of noise bands was assumed to be a better approximation to the actual sound sensation.

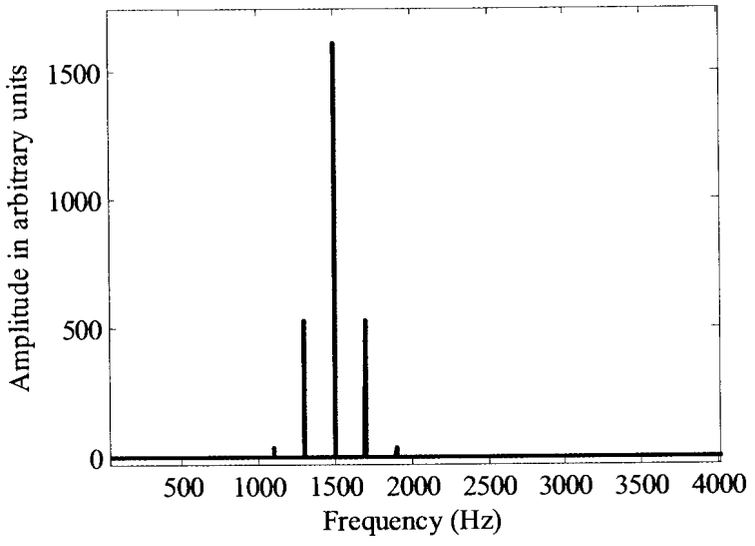

Figure 3.22. Frequency modulated carrier that shows an increase in bandwidth that may be used for the summation of sinusoids

3.3.2.6 Quantisation

The final step before summation of all the channels is the quantisation of the stimulation current values. In the Nucleus cochlear implant there are only 238 available current levels by which the nerve cells can be stimulated, ranging from T to C, a section of the range is shown in figure 3.23. As mentioned, T may typically be around 100 μA and C around 1 mA. The stimulation intensity is also quantised into 238 linear steps in the acoustic simulation. The 20 dB intensity range is divided into 238 steps. This function causes a reduction of the intensity resolution available for a more realistic simulation of what happens in the cochlear implant. The function 'quant' in Matlab is used to quantise the continuous speech signal so that it is represented in 238 steps.

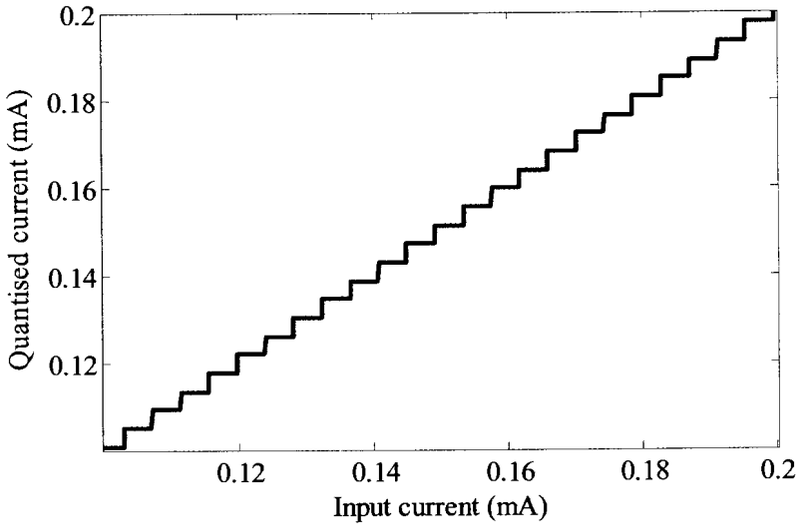

Figure 3.23. Input current quantised into 238 discrete levels; this figure shows only 24 current levels between 0.1 mA and 0.2 mA

3.3.2.7 Summation of all the channels

To reconstruct the speech signal, all the individual channels must be added together. This is done by amplitude modulating noise bands (representing the 20 channels) with the magnitudes obtained from the quantisation step. The bandwidths of the noise bands are also calculated from these magnitudes, as described earlier. When the channels are summed, the stimulation rate and asynchronous stimulation are taken into account, as will be explained in the following sections.

The channels of a cochlear implant are stimulated in a specific order at a rate of 14 400 pps (the maximum rate is assumed, a slower rate might be a more accurate simulation of the SPEAK strategy). They are not stimulated simultaneously; only one channel can be active during a 1/(14 400 pps) period. The stimulation rate of a single channel therefore depends on the number of channels that are used in a specific cochlear implant. For example, for an eight-channel implant, each channel will be stimulated every 8/14 400 seconds (555.56 μs). This corresponds to a frequency of 1 800 Hz. Each channel is effectively modulated with a pulse train with a frequency of 1 800 pps and a duty cycle of 0.125 (1/8).

The values used to modulate the noisebands are calculated from the quantised intensities. The effective window length that is used to calculate the RMS values is 2 ms, while the period of stimulation in the cochlea is 69.44 µs per single channel. This means that the analysis windows are much longer than the stimulation windows[3]. The quantised values are constant for a 2 ms window, the amplitudes of the modulated noise bands will also remain constant over a period of 2 ms, irrespective of the stimulation period, similar to a sample and hold function. This is demonstrated in figure 3.24.

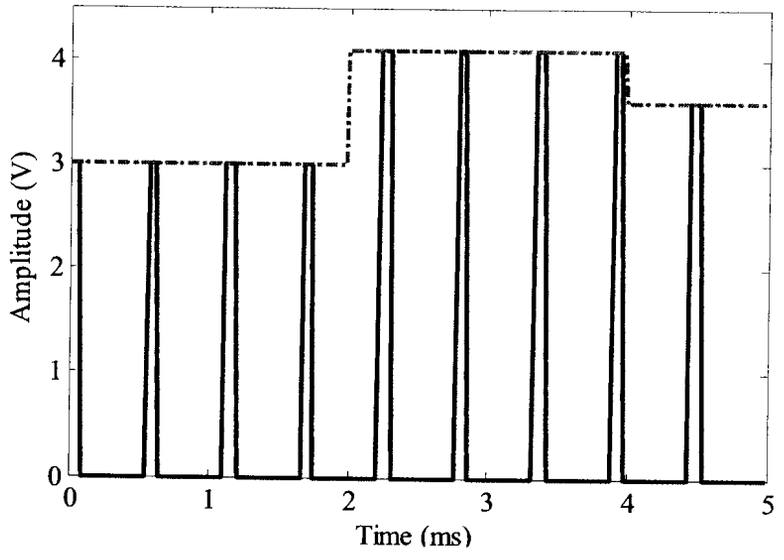

Figure 3.24. Figure that shows that the modulated stimulating pulses' amplitude (solid line) follows the amplitudes from the 2 ms analysis windows (dashed line) for one channel. The length of a stimulation pulse is 69.44 µs

One of the important challenges encountered during the development of the simulation was finding a way to simulate the periodicity of stimulation while providing enough information for a normal-hearing person to perceive a specific frequency. When a stimulating frequency of 14 400 pps is used, one stimulation period equals 69.44 µs. When a signal with a frequency of less than 14 400 Hz must be used for the acoustic stimulation, a whole period of the signal will not be completed within the stimulation window. It is important that at least half a period must be completed for the acoustic stimulation, since a healthy cochlea uses frequency information for the sensation of

[3]Stimulation pulses used for electrical stimulation become stimulation windows in the acoustic simulation

hearing. If a fraction of a period is used as stimulus, the frequency specificity is lost and the normal-hearing listener will not be able to perceive the specific frequency, as shown in figure 3.25.

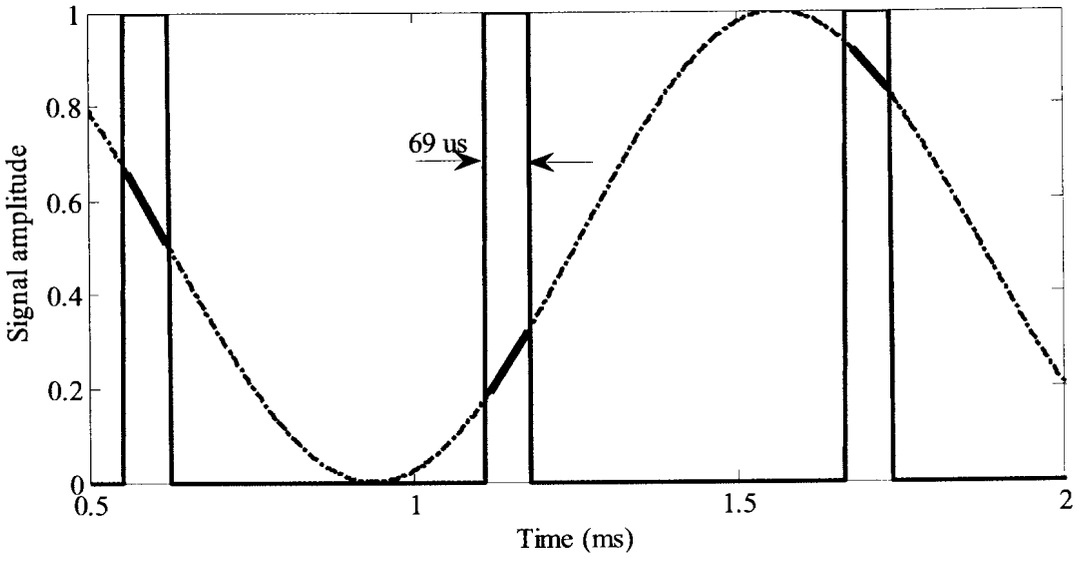

Figure 3.25. Graph that shows the stimulation pulses (solid line) and an acoustic 800 Hz pure tone inside the frequency range of speech (dashed line). For example, the 800 Hz pure tone is used to acoustically simulate stimulation at the 800 Hz position in the cochlea. It is clear from the graph that the stimulation window is too short to transmit frequency information about the signal – the highlighted sections of the pure tone do not convey a 800 Hz sensation

The short periods of the stimulation pulses introduce a problem, as most speech signals are typically in the frequency range of 80 Hz to 6 kHz. For electric stimulation this problem does not exist, as the electrodes stimulate nerve fibres at the exact place where a specific frequency sensation is generated. To simulate the processing correctly using acoustic signals, a method must be developed so that the effect of the stimulation rate is still present, but the frequency information is retained for the acoustic stimulation of normal-hearing listeners.

One way of solving the problem is to simulate all the channels continuously, but to amplify the channel that represents the activated electrode in a cochlear implant for one stimulation period. This will still give the effect of the periodicity of stimulating the electrodes in a basal-to-apex order.

A pilot experiment was done to assist in deciding how to simulate the periodicity of stimulation effectively. The simulation generated a pure tone which was then modulated with a square wave. The square wave's frequency was determined by the equation

$$f_{stim} = \frac{14400}{N_{channels}} \quad \text{Hz}, \tag{3.10}$$

where f_{stim} is the stimulation frequency of 1 channel, 14 400 is the maximum stimulation frequency (pps) and $N_{channels}$ are the number of channels used in the simulation. The duty cycle of the square wave was determined by the number of channels used in the simulation. For an eight channel simulation, the duty cycle was 0.125. The modulation depth of the square wave was 0.5, which meant that the active channel was amplified with 1 while the rest of the channels were amplified with 0.5 for each stimulation period. The outcome of the experiment was that the pitch of the modulated signal was the same as the pure tone, with an audible pitch representing the modulation signal. This modulation scheme was used in combination with the method explained in the following section to simulate the stimulation rate while still preserving the pitch of the acoustic signal.

When the harmonics of a fundamental tone are added together and used as an acoustic stimulus, the pitch of the signal will be the same as the pitch of the fundamental tone (Terhardt, 1979; Terhardt, Stoll and Seewann, 1982). This can be used effectively to simulate the pitch of a low-frequency signal when this signal can only be turned on for a short period of time. The harmonics will complete at least half a period in the stimulation window as shown in figure 3.26, while the pitch will be the same as for the fundamental tone and frequency specificity will not be lost with the acoustic simulation.

The same principle for pitch applies to noise bands with a specific centre frequency and bandwidth. When noise bands are generated centred around the three harmonic frequencies of the fundamental centre frequency, the pitch of the noise will be the same as the noise band centred at the fundamental frequency. The choice of the three harmonics is done in such a way that the smallest harmonic will be able to complete a half period within the stimulation period of 69.44 µs. For the maximum stimulation rate of 14 400 pps, the minimum harmonic frequency needed to complete a half period is 7.2 kHz, which is half

the maximum stimulation rate. When a signal is halfwave rectified, as is done by modulating the smallest harmonic with a square wave, the frequency of that signal will be present in the spectrum.

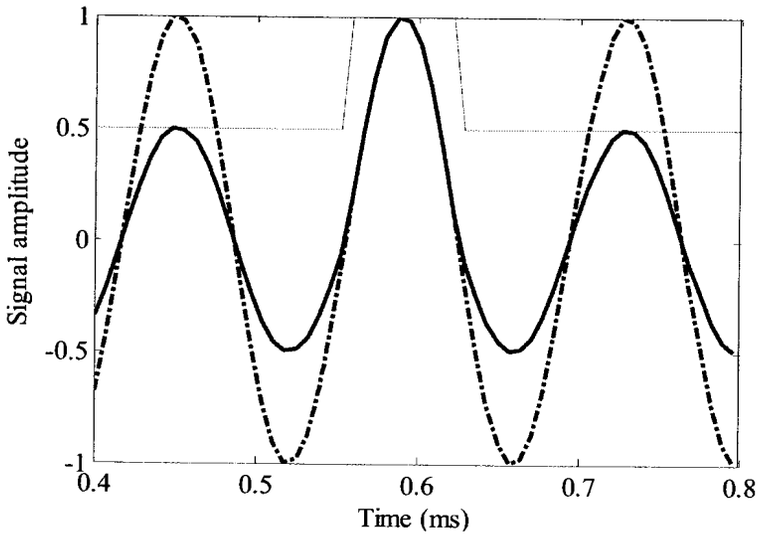

Figure 3.26. Demonstration that half a period of the second harmonic of 3.2 kHz (dashed line) is completed in a time period of 69.44 μs (thin solid line). The signal modulated with the stimulation window is also shown (thick solid line)

Table 3.3. Number of harmonics included in the summation of the noise bands

Channel	Fundamental tone (Hz)	Number of harmonics used
1	493	17
2	565	15
3	645	14
4	733	12
5	831	11
6	940	10
7	1 061	9
8	1 195	9
9	1 343	8
10	1 508	7
11	1 690	7
12	1 893	6
13	2 118	6
14	2 367	6
15	2 644	5
16	2 950	5
17	3 290	5
18	3 668	4
19	4 086	4
20	4 550	4

The frequencies in table 3.3 represent the harmonics used for an insertion depth of 25 mm. When all the harmonics presented in table 3.3 and the fundamental frequency are summed to reconstruct the speech signal, the speech signal will be less shrill. The higher harmonics were included so that the period of the signal would be short enough to fit into the simulation window. However, by doing this, more bands are introduced to the simulation and it is not possible to simulate the effect that the number of spectral channels has on speech recognition. There will always be more spectral channels present in the reconstructed signal than the number of frequency bands used to analyse the speech signal. For this study, the increased number of spectral channels does not have a significant influence. All the simulations were done at a fixed number of channels, namely eight. It was shown in Dorman et al. (2002) that six channels are adequate for good speech understanding. With more than six channels, speech understanding stabilises even though

the number of channels increases. With eight channels the plateau of speech understanding has already been reached and it will not increase with more spectral channels. Figure 3.27 shows the noise bands with centre frequencies at the harmonics of the fundamental centre frequency. The amplitudes of the noise bands are attenuated with increasing centre frequency of the harmonics.

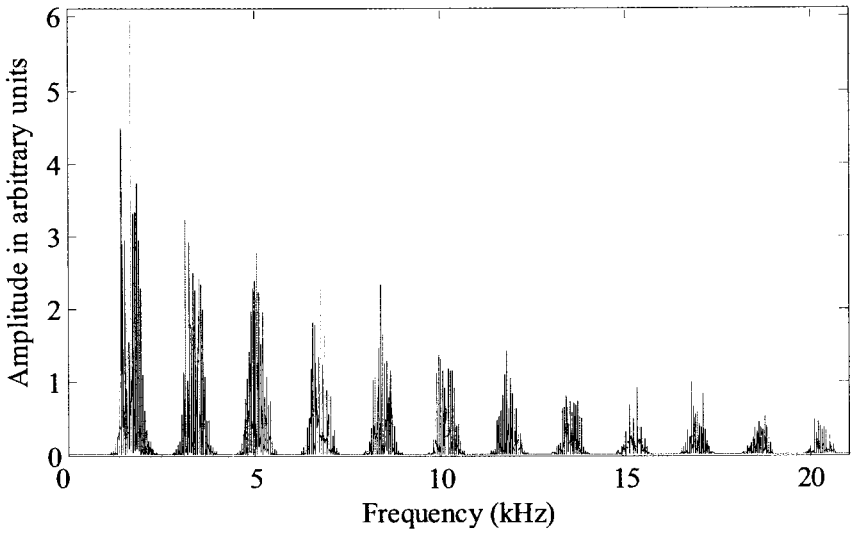

Figure 3.27. Example of noise bands centred at the harmonics of a fundamental tone centred at 1 700 Hz

To demonstrate the effective simulation of the stimulation rate, a 4 160 Hz pure tone was modulated with the abovementioned scheme and analysed to determine the perceived pitch. When listening to the modulated pure tone, there was a pitch sensation at two frequencies – one for the modulating frequency (1 800 Hz) and one for the frequency of the pure tone. In figure 3.28 it is shown that the modulated pure tone has energy content at 4 160 Hz, but also at (4 160 – 1 800) Hz and (4 160 + 1 800) Hz due to the 1 800 Hz stimulation rate. The depth of the modulation determines which pitch will be most prominent. The modulation depth refers to the difference between the maximum and minimum level of the modulation signal. It has been found that a modulation depth of 0.5 produces a balanced pitch – both pitches can be heard distinctly.

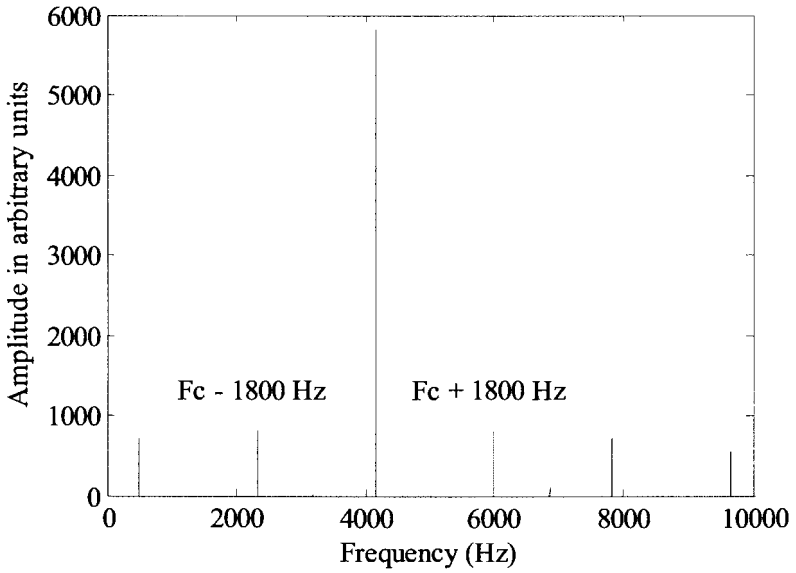

Figure 3.28. Modulation of a 4 160 Hz pure tone with the developed modulation scheme to demonstrate the effect of the stimulation rate on the frequency components of the signal

Because a square wave is used for modulation, there will be frequency components (that are 1 800 Hz apart) present in the whole frequency band, making the 1 800 Hz pitch very prominent in the processed signal. This frequency spread is because of the sinc(f) FFT of the square wave. This modulation signal can be adapted so that there will not be a wide frequency spread, but a better choice for the modulation signal was not investigated in this study.

The desired effect of hearing a pitch dependent on the stimulation rate is therefore achieved by modulating the speech signals with a square wave and including the harmonics of the fundamental noise band. Both the stimulation rate pitch and place pitch have been simulated.

In figures 3.29 and 3.30 an example of the final output of the acoustic model is shown for the time and frequency domain.

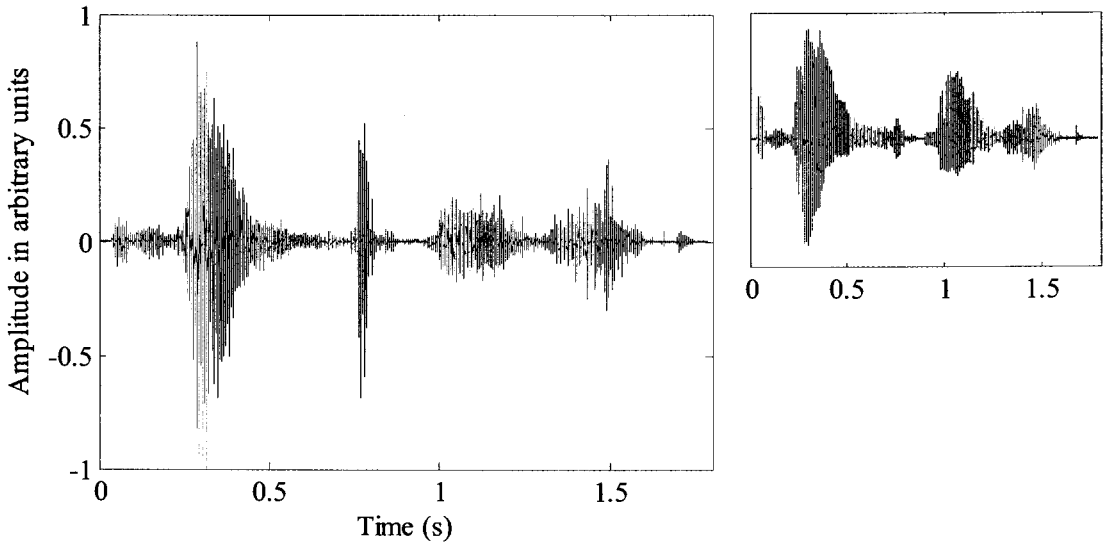

Figure 3.29. Example of the final output of the acoustic model for the same speech signal as in figure 3.5 (shown in the top right corner) in the time domain.

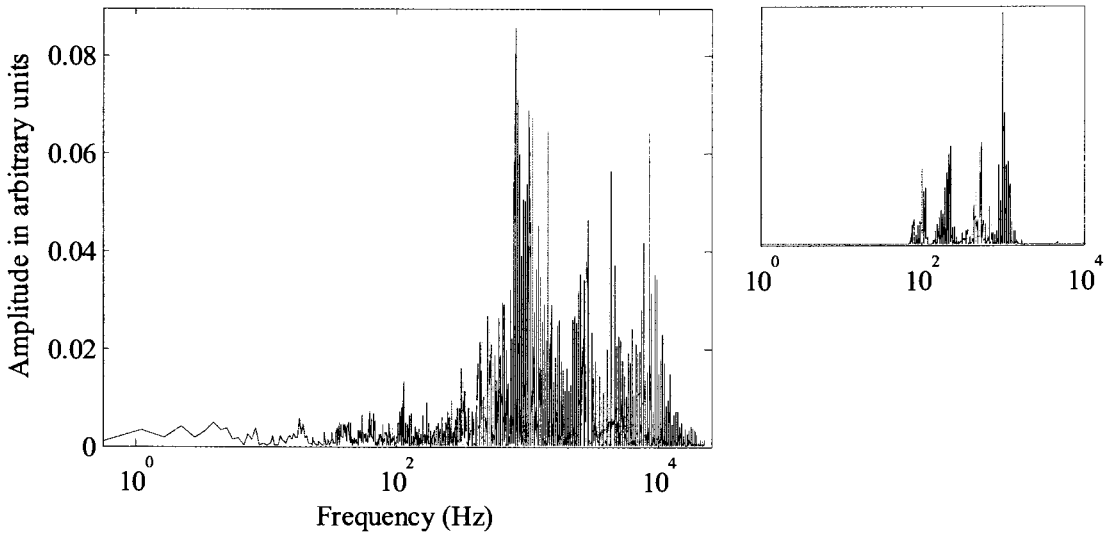

Figure 3.30. Frequency representation for the same segment of speech as for figure 3.29; the original spectrum is shown in the top right corner. The high-frequency components present in the processed signal that are absent in the original signal is due to the use of harmonics for the reconstructed signal

Finally, all the channels are simulated asynchronously to obtain the effect of activating each channel separately in the cochlear implant.

3.4 EXPERIMENTAL STUDY

3.4.1 Listeners

The processed speech segments were presented to seven female listeners and three male listeners. All the listeners were normal-hearing persons between the ages of 19 and 26. Experiments were done with native Afrikaans-speaking persons.

3.4.2 Stimuli

The acoustic model was used to process 12 vowels (in the context of /p/-VOWEL-/t/) and 15 consonants (in the context /a/-CONSONANT-/a/). The original utterances, spoken by an Afrikaans male speaker, were recorded at 44.1 kHz (16 bit resolution) at the University of Pretoria. The outputs from the acoustic model were scaled to conform to .wav file specifications, the amplitudes were normalised between -1 and 1, and these were then used in the experiments.

The processed speech segments were also used in acoustic analyses to determine which features of the segments are used for recognition of phonemes. From the results of the analyses, predictions can be made on which speech segments will be confused in the experiments.

The stimuli presented for the vowels are /æ/ (pat), /ɑ/ (pad), /u/ (poet), /œ/ (put), /y/ (puut), /e/ (peet), /ɑː/ (paat), /i/ (piet), /ə/ (pit), /ɔ/ (pot), /ɛː/ (pêt) and /ɛ/ (pet) and for the consonants /k/ (aka), /b/ (aba), /p/ (apa), /n/ (ana), /m/ (ama), /l/ (ala), /r/ (ara), /s/ (asa), /z/ (aza), /f/ (afa), /v/ (awa), /t/ (ata), /d/ (ada), /j/ (aja), /x/ (aga).

To present the processed utterances, a software application called Baby Apex was used (Pretorius et al., 2005). This application generates a matrix from specified .wav files and presents the speech segments in random order. Every phoneme is played to the listener through a loudspeaker and he or she must then choose the stimulus that he or she heard. An example of the screen view is shown in figure 3.31. Each stimulus is repeated 10 times

for each experiment. It is possible to give the listener feedback and the listener can also do a few practice runs before the actual experiment. However, this was not permitted in these experiments.

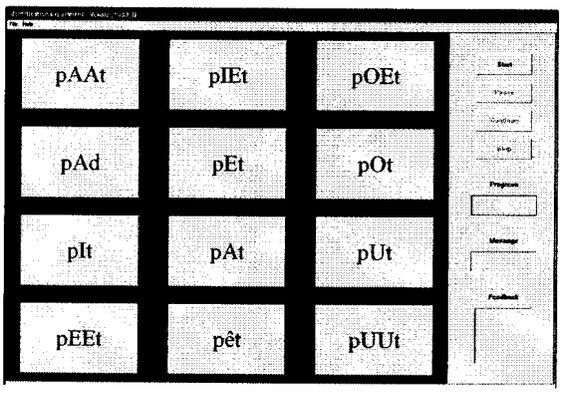

Figure 3.31. Screen view of choices for a vowel recognition experiment

The listener was introduced to the nature of the processed sound from the acoustic simulation by playing eight English sentences[4] to the listener. Both the processed and original sentences were played back. No practice sessions for the phoneme presentation were scheduled. The speech tokens were presented inside a soundproof booth at an average sound pressure level (dB SPL) of 70 dB. The utterances were presented through a Yamaha MS101 II speaker approximately 1 m away from the listener. One session was completed in approximately 30 minutes and included, on average, four experiments.

During the experiment, the speech segments are played back in random order and the listener must choose the speech segment that he or she heard by clicking on the corresponding block. Confusion matrices are generated according to the response to a specific stimulus.

Normal-hearing persons do not use cochlear implants on a daily basis and will need to become accustomed to the sound of the acoustic simulation. The subjects conducted more experiments than the number used for the final results. A definite learning curve could be observed from the first experiment to approximately the fourth experiment. The data for the fourth and fifth experiments were used for the final results.

[4]Sentences were obtained from http://www.utdallas.edu/~loizou/cimplants/tutorial/

3.4.3 Experimental conditions investigated

Experiments were set up for conditions before and after dynamic range compression in a quiet environment. For conditions after dynamic range compression, the processing step of mapping the current levels between C and T was included as well as quantisation (refer to sections 3.3.2.4 and 3.3.2.6). This was done to determine the effect of the dynamic range compression on speech recognition. From the results it would also be possible to determine what underlies speech recognition with cochlear implants. For the simulations, eight channels were used at an insertion depth of 25 mm.

Experiments in noise were conducted to determine what underlies speech recognition under noisy conditions. Initially, experiments were done with white Gaussian noise, but the recognition proved to deteriorate very little, if at all. Other types of noise were therefore implemented to simulate everyday conversational scenarios.

The speech tokens were mixed with speech-like noise (CCITT Recommendation 227) (Fastl, 1987; Müller et al., 2002; Zwicker and Fastl, 1999) and multi-talker babble (Ferguson and Kewley-Port, 2002; Killion, Niquette, Gudmundsen, Revit and Banerjee, 2004; Nie, Stickney and Zeng, 2005) at signal-to-noise ratios (SNRs) of 40 dB, 20 dB and 0 dB before processing through the acoustic simulator. The SNR is defined as the 10log of the ratio between the power in the speech segment as a whole and the power present in the noise, measured in decibels. The speech tokens were normalised at a level of 70 dB and the noise power adapted accordingly. For example, for 20 dB SNR, the noise signal's power level was set to 50 dB SPL, giving a 20 dB difference of signal power to noise power. The levels of noise were based on the results obtained in pilot experiments. For the range of 0 – 40 dB SNR, the results varied from near chance to average recognition.

Vowel and consonant recognition was measured in the 10 normal-hearing persons at the three different signal-to-noise ratios: 0 dB, 20 dB and 40 dB. The recognition of vowels and consonants was determined in the presence of CCITT noise and multi-talker babble. A total of 12 different experiments were conducted for recognition in noise, three sets of experiments for vowels in CCITT noise (0 dB SNR, 20 dB SNR and 40 dB SNR), three

sets for vowels in multi-talker babble (0 dB SNR, 20 dB SNR and 40 dB SNR), three sets for consonants in CCITT noise (0 dB SNR, 20 dB SNR and 40 dB SNR) and three sets for consonants in multi-talker babble (0 dB SNR, 20 dB SNR and 40 dB SNR).

Pilot experiments were done using vowels and consonants mixed with Gaussian white noise (Pollack and Pickett, 1957), but no significant deterioration in speech recognition was detected. At a SNR of 0 dB for vowels, the percentage of vowels recognised correctly was 72 %, which is almost the same as for quiet conditions. This is ascribed to the fact that white noise has energy over a very broad frequency band. Speech only has energy up to approximately 5 kHz. The acoustic simulation filters all the high-frequency noise components so that there is little noise energy in the final speech token. The final SNR is therefore not as high as initially intended.

Speech-like noise and multi-talker babble were used instead of Gaussian white noise (Dubno, Horwitz and Ahlstrom, 2005; Ferguson and Kewley-Port, 2002; Friesen et al., 2001; Fu et al., 1998; Killion et al., 2004; Müller et al., 2002; Nie et al., 2005; ter Keurs et al., 1993b; Yang and Fu, 2005). Existing .wav files[5] of noise signals were used with permission from E Hennix. The CCITT noise and multi-talker babble have frequency components in the same frequency bands as speech, thus simulating a more realistic hearing environment for cochlear implant users.

The speech segments were presented at an average sound pressure level of 70 dB inside a soundproof room, as for the experiments without noise. As for the earlier experiments, confusion matrices were determined using the application Baby Apex. These confusion matrices were analysed through multidimensional scaling to determine the effect of noise on speech recognition, as will be discussed in the following chapter.

[5] http://www.e.kth.se/ and http://www.mmk.e-technik.tu-muenchen.de/

3.5 SUMMARY

In this chapter, the method followed to develop the acoustic model was reported. Considerations encountered during the development are recorded here as well as the proposed solutions. The solutions implemented are explained in detail. This chapter also presented the method followed to perform experiments with acoustic simulations. The results from the experiments performed in quiet and the results from the experiments performed in noisy conditions will be presented separately in the next chapter. An analysis of the processed speech from the acoustic simulations will be done in order to predict confusions between speech segments in terms of acoustic features.

CHAPTER 4 RESULTS

4.1 CHAPTER OBJECTIVES

The acoustic simulation is evaluated in this chapter. A comparison is made between the electrodogram obtained from the NMT and the acoustic simulation. From the two electrodograms it can be seen that the signal processing of the acoustic model is a good approximation of the processing done in cochlear implants. Results from acoustic analyses are used to explain confusion encountered in experiments with vowels and consonants. The results from experiments before and after dynamic range compression will be given, as well as the results of experiments done in noisy conditions.

4.2 RESULTS OF ACOUSTIC SIMULATION

The processing steps in the acoustic model were shown in chapter 3. The final outcome of the model is a processed speech signal. It is evaluated here by comparing the spectrogram of the acoustic simulation of a section of speech with the electrodogram of the same speech processed with the SPEAK strategy. The NMT has the functionality of displaying the current pulses used to stimulate a cochlear implant as a function of time and electrode. This is called an electrodogram, which is similar to a spectrogram[1]. By plotting these spectrograms for both the SPEAK strategy implemented in the NMT and the acoustic simulation, a meaningful comparison can be done. From the comparisons, it can be determined whether the processor part of the developed acoustic model produces the same result as the SPEAK processor implemented in the NMT. This gives a measure of how accurately the processor part has been implemented in the acoustic model.

The electrodogram from the NMT and spectrogram from the acoustic simulation are shown in figures 4.2 and 4.3. The spectrogram for the original speech signal is shown in figure 4.1.

[1]A spectrogram gives the space-time stimulation pattern of an electrode array

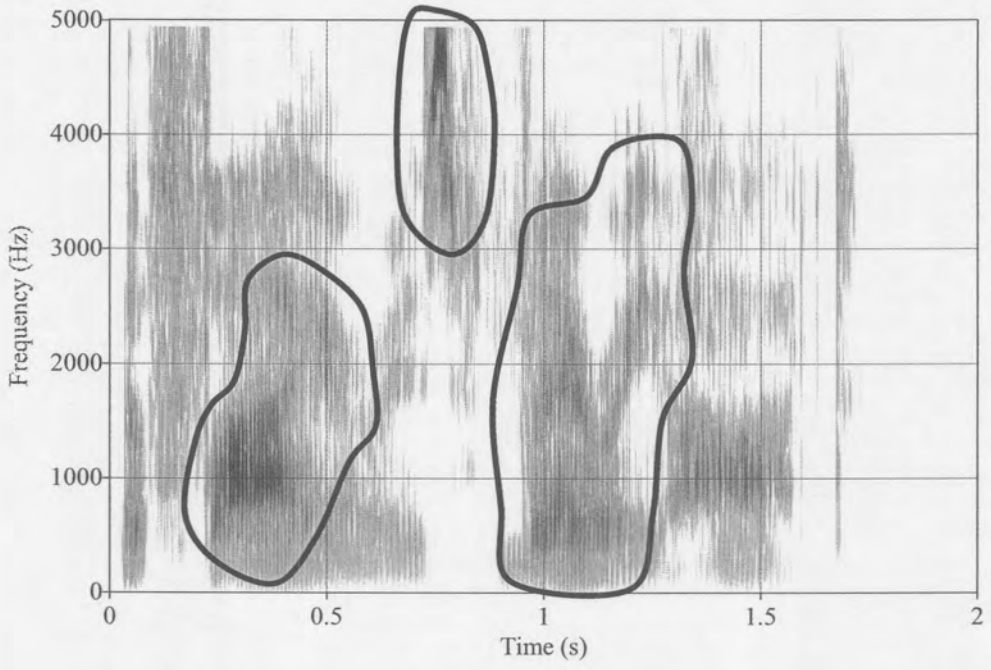

Figure 4.1. Spectrogram of original speech signal, 'The fire is very hot'

In figure 4.2 each vertical line represents a stimulation pulse. Line length represents stimulation pulse amplitude. Thus, the electrodogram shows the actual stimulation pulse trains that the SPEAK algorithm would generate if the input should be the speech signal with spectrogram shown in figure 4.1. ICE stands for internal cochlear electrode and ECE for external cochlear electrode.

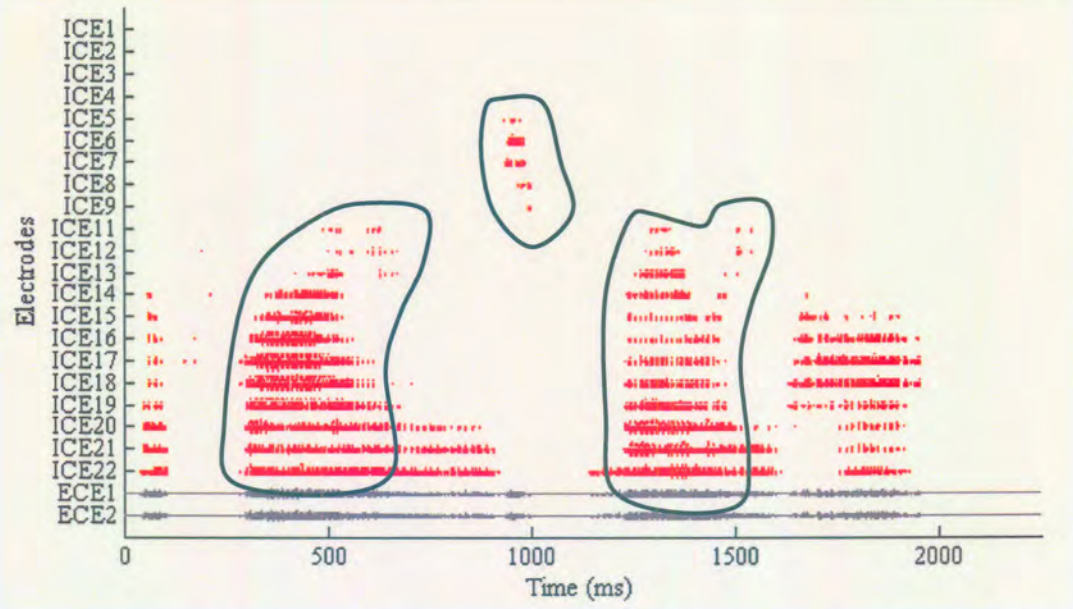

Figure 4.2. Electrodogram of the sentence 'The fire is very hot' using the NMT for the SPEAK strategy

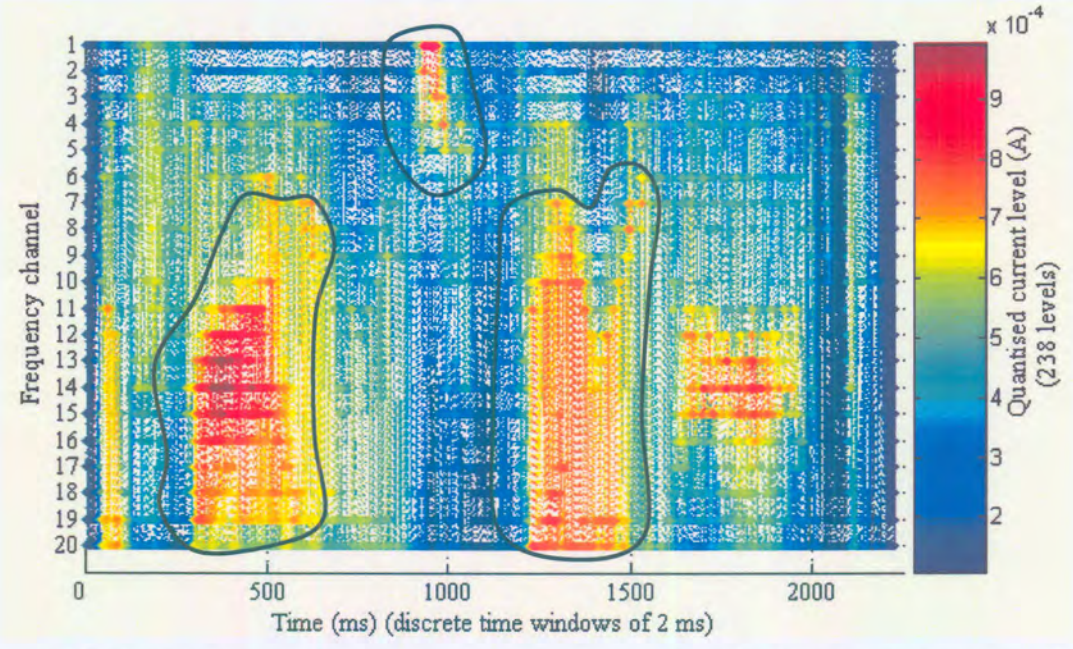

Figure 4.3. Spectrogram of the sentence 'The fire is very hot' obtained from the output of the acoustic simulation for the SPEAK strategy

From the electrodogram and spectrogram, the spread of energy is very similar across the electrodes. One difference though is that for the SPEAK algorithm implementation of the NMT, the electrodes are activated only from the fifth electrode as opposed to the first

electrode already active in the acoustic simulation. The reason for the acoustic simulation to be activated from the first electrode is the frequency-place mapping used in the acoustic simulation, as described in section 2.4.

The electrodograms for the acoustic simulation and the SPEAK algorithm implemented in the NMT appear very similar, indicating that the acoustic model may contain speech cues similar to those found in the electrode space-time stimulation pattern. Take note of the similar patterns marked on figures 4.1 to 4.3. The same stimulation patterns in the two electrodograms can be observed, for example the high energy content around 500 ms, and at 1 000 ms they have the same high frequency components. At 100 ms, there is more energy content in the spectrogram of the acoustic simulation than in the electrodogram of the SPEAK processing strategy implemented in the NMT. The same energy distribution across electrodes is indeed expected, as the speech signal is divided into similar frequency bands for both the acoustic simulation and the SPEAK strategy.

4.3 USING THE ACOUSTIC MODEL TO PREDICT CONFUSIONS AND RESULTS FROM EXPERIMENTAL STUDY

By using the acoustic model, the characteristics of the processed speech segments can be analysed in order to explain confusions of vowels and consonants. By identifying specific characteristics of consonants and vowels, predictions can be made of possible speech recognition trends. By using a Feature Information Transmission Analysis (FITA) (Miller and Nicely, 1955; Van Tassel, Soli, Kirby and Widin, 1987; Wang and Bilger, 1973), confusion matrices are analysed to determine the information transmitted and conclusions are made as to which characteristics are transmitted most effectively with the acoustic simulation. Some of the characteristics of vowels that are important for recognition are the formant frequencies and duration of vowels. For consonants it is the duration, peak and median energy levels, minimum to peak ratio of energy levels and the envelope variation of individual speech segments (Pretorius et al., 2005; Van Wieringen and Wouters, 1999). Consonants can also be classified according to manner, voicing, nasality, liquidity, place and affrication (Miller and Nicely, 1955). These classifications give an indication of how consonants are produced acoustically.

4.3.1 Vowel confusions

4.3.1.1 Acoustic analysis of vowels at output of acoustic model

For the analysis of the vowels, specific signal characteristics were calculated. These characteristics are typically used as cues to recognise speech (Borden and Harris, 1994). For the vowels; the duration, F_1 and F_2 were determined using the program PRAAT, a phonetic software package (Boersma and Weenink, 2004). In order to analyse the formant frequencies of the vowels, linear predictive coding was performed on the speech segments (Rabiner and Schafer, 1978). The first and second formant frequencies (F_1 and F_2) of the vowels were estimated by means of linear predictive coding (16^{th} order, 25 ms time windows) and by visual inspection. For example, the formants of /a/ are shown in figure 4.4.

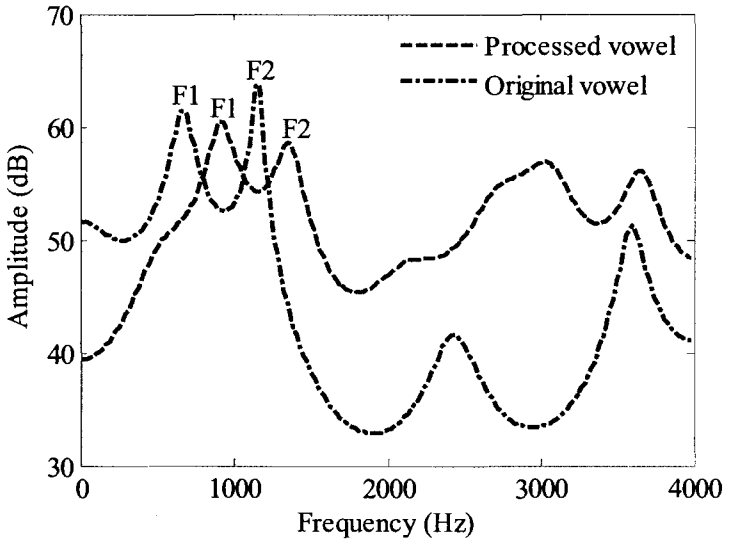

Figure 4.4. Formants for the vowel /a/ (as uttered by a male speaker). "Processed" refers to the speech signal at the output of the acoustic model

The procedure for determining the formant frequencies of the vowels is as follows: the vowel is extracted from the speech segment by visual inspection, then processed using a linear predictive coding algorithm (Rabiner and Schafer, 1978) in PRAAT. Linear predictive coding is performed on the vowel to obtain the envelope of the spectrum of the vowel. Although PRAAT has a function that can determine mean formant frequencies

over a period of time, this function was not used to determine the formant frequencies, as it proved to be unreliable and gave erroneous values in some tests. Rather, the first and second formant frequencies were determined from both the LPC spectrum and visual inspection of the spectrogram of the vowel. Table 4.1 summarises the results from these analyses for the conditions before and after dynamic range compression. The analysis of speech segments with processing before and after dynamic range compression is done in order to determine the effect that dynamic range compression has on the acoustic signal characteristics and recognition of speech segments. It is suspected that the inclusion of dynamic range compression has a dramatic effect on speech recognition, as a significant amount of information for recognition of speech is removed through this step.

Table 4.1. Values for duration (ms), F1 (Hz) and F2 (Hz) used to plot the vowel spaces

		Original vowels			**Without dynamic range compression**			**With dynamic range compression**		
		Duration (ms)	F_1 (Hz)	F_2 (Hz)	Duration (ms)	F_1 (Hz)	F_2 (Hz)	Duration (ms)	F_1 (Hz)	F_2 (Hz)
pAAt	ɑ:	218	765	1 074	205	590	1 016	226	747	1 266
pIEt	i	87	258	2 031	67	480	1 920	70	540	1 955
pOEt	u	84	319	1 057	67	450	1 128	67	440	1 120
pAd	a	100	783	1 143	88	720	1 216	87	740	1 174
pEt	ɛ	87	508	1 966	85	583	1 757	88	541	1 941
pOt	ɔ	102	525	954	101	520	1 066	118	540	1 100
pIt	ə	73	479	1 588	61	450	1 473	69	598	1 644
pAt	æ	135	664	1 506	111	648	1 433	104	690	1 430
pUt	œ	92	508	1 524	70	480	1 448	84	526	1 502
pEEt	e:	198	337	2 104	137	460	1 720	132	441	1 955
pêt	ɛ:	274	416	1 904	225	509	1 890	228	526	1 756
pUUt	y	91	285	2 069	73	440	1 919	71	484	2 034

From figure 4.4, it can be seen that the first and second formants are displaced slightly after processing. This is due to the fixed frequency bands used for the reconstruction of the speech signal. It is notable though that the formant frequencies still have the same pattern, even though it moved slightly in frequency. The formant patterns are retained despite the low frequency resolution available. The two important formants are the first

and second – the higher formants are less important in vowel recognition. From the figure it can be seen that the third formant splits into two peaks during processing. It is assumed that this will not have a great impact on vowel recognition. Another important result seen in figure 4.4 is the reduced spectral contrast of the processed vowel. The spectral contrast between the valleys and peaks is reduced significantly, which may contribute to a reduction in recognition of vowels (Sidwell and Summerfield, 1985; ter Keurs et al., 1993b).

The fact that the formant pattern maintains primary characteristics after processing through the acoustic model is noteworthy. The modulated noise bands do indeed appear to convey the important information of the speech signal, so that recognition of speech should be possible.

4.3.1.2 Predictions of vowel confusion from acoustic analyses

When looking at the vowel space[2] of the original, unprocessed speech, each vowel has either its own formant space or belongs to a definite group of vowels having approximately the same formant frequencies. Vowels are mostly recognised by the information that is transmitted by the formant frequencies and the duration of the vowel (Borden and Harris, 1994; Van Wieringen and Wouters, 1999). Therefore, when different vowels' formant frequencies (both F_1 and F_2) are close together, there is the possibility that these vowels can be confused with each other after processing.

In the following few paragraphs the formant frequency space of the processed speech will be discussed so that predictions can be made as to which vowels will be confused. Firstly, predictions will be made by visually inspecting the vowel spaces. Thereafter a physical measure is introduced by which the predictions can be quantified. In this study, a three-dimensional Euclidean distance is used for the predictions. The Euclidean distance is measured between all the vowels using normalised F_1, F_2 and the duration.

[2]A vowel space is defined as a multidimensional space where a number of vowels are plotted as a function of their signal characteristics, specifically their formant frequencies F_1 and F_2

For the original speech, one can see a number of distinct groups of vowels in figure 4.5. The vowels /œ/ (put) and /ə/ (pit) are grouped very close together and when they are processed through the acoustic model, they are expected to be confused regularly. Their duration is also very similar, 92 ms and 73 ms respectively.

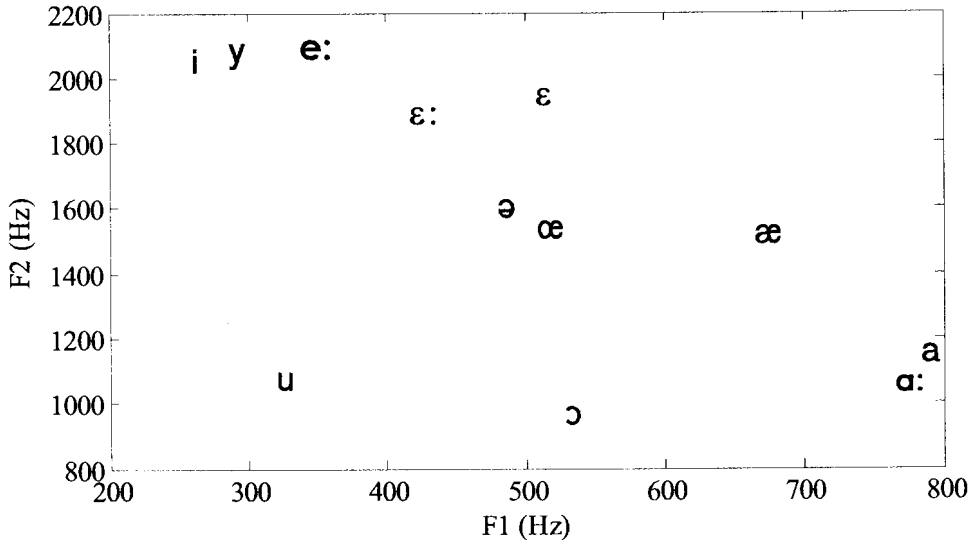

Figure 4.5. First and second formant frequencies of original speech for a male speaker

The duration of /a/ (pad) and /ɑ:/ (paat) differ notably, 100 ms and 218 ms respectively. The time cue should result in reasonable differentiation between the two vowels, even though they are very close together in the formant frequency space.

The group of vowels /i/ (piet), /y/ (puut), /ɛ/ (pet) and /e:/ (peet) are close together in the vowel space and can potentially be confused. When looking at the time duration of the vowels, it is 87 ms, 91 ms, 87 ms and 198 ms respectively. From these values and the vowel space, it is clear that /i/ (piet), /ɛ/ (pet) and /y/ (puut) should be confused regularly. Their formant frequencies are very close together and their duration is almost the same. The longer duration of /e:/ (peet) causes this vowel to be distinguished from the others in the group more often.

Each of the cues of the vowels /ɛ:/ (pêt), /u/ (poet), /ɔ/ (pot) and /æ/ (pat) are distinct so that it is not expected that they will be confused with one another. The duration of /u/

(poet) and /ɔ/ (pot) are very similar, 84 ms and 102 ms respectively, making them the only two vowels in this group that might possibly be confused.

The first and second formant frequencies of the processed and original vowels were normalised using Lobanov's z-score transformation (Adank, Smits and van Hout, 2004). This is necessary to be able to compare formant spaces across various conditions; with the processor there might be an offset added to the formant frequencies of a specific vowel space. These offsets will be removed by normalisation. It is possible to perform comparisons between vowel spaces because the transformation normalises the formant frequencies of a single vowel with respect to the average and standard deviation of the formant frequencies of all the vowels. These normalised formant frequencies were plotted to compare the vowel space of the original vowels and the processed vowels. The vowel spaces without normalisation are shown in figures 4.6 and 4.8; the normalised vowel spaces are shown in figures 4.7 and 4.9.

The use of the Lobanov z-score transformation was chosen because it is a vowel-extrinsic procedure, taking into account the average and standard deviation of the formant frequencies of all the vowels. The normalisation of a formant frequency, using a vowel-extrinsic procedure, depends not only on the formant frequency of the relevant vowel, but also on the formant frequencies of all the vowels in the set of vowels. Procedures that include information across vowels and information within formants perform better at normalising a set of formant frequencies than procedures that do not include these (Adank et al., 2004). The Lobanov z-score transformation is an example of such a procedure. Other vowel-extrinsic procedures that might also have been used and given the same results are the Nearey1 and Gerstman procedures (in contrast to the Neary2 procedure) (Adank et al., 2004).

In Borden and Harris (1994), it is suggested that patterns are used for vowel identification rather than absolute formant frequency values. This explains why a person can still recognise vowels even when speakers' vowel spaces are different, as is the case when the speaker is a man, a woman or a child. A vowel space is defined by the point vowels; these vowels are used as reference points to normalise formant frequency values. The point

vowels are /i/, /a/ and /u/. These three are on the edges of the processed vowel space.

When looking at the normalised vowel space of the processed and original vowels (figures 4.7 and 4.9) it can be seen that the vowel space is transformed to some degree. The processed vowels have shifted around in the formant space. This is because of the fixed centre frequencies of the bandpass filters used in the processor. The frequency bands are fixed for the noise bands to reconstruct the output of the acoustic model. The processed vowels still represent a vowel space similar to that of the original vowel space. This transformation of the formant space might be the source of many confusions in the experiments (discussed later). There are a few characteristics that are worth mentioning:

- the vowels /i/, /y/ and /e:/ are still grouped together and are at the edge of the vowel space, even though in the group itself they moved,
- the vowels /æ/, /œ/, /ɛ:/ and /ə/ remain approximately at the centre of the vowel space, and
- the edges of the vowel space are still formed by the vowels /a/, /u/, /i/, /ɔ/, /ɑ:/ and /y/.

Possible confusions might be among the point vowels in the initial experiments. The vowel space may be foreign to the listener with the vowels moving significantly relative to each other. For normal speech the vowel space is clearly defined and used by the listeners on a daily basis. A new vowel space needs to be defined before the listener can recognise other vowels relative to the point vowels. The listener therefore first has to become accustomed to the new vowel space. Before this, the specific vowels that define this new space might easily be confused with one another, for example /a/, /ɔ/, /u/ and /i/ (Borden and Harris, 1994). As the listener becomes familiar with this new vowel space, confusions should decrease.

The vowels located in the centre of the vowel space would possibly be confused from the start of the experiments, as they do not move as much relative to the point vowels. It can be seen from figures 4.7 and 4.9, the normalised vowel spaces, that the /œ/ and /ə/ vowels are clustered in the middle of the vowel space before and after processing.

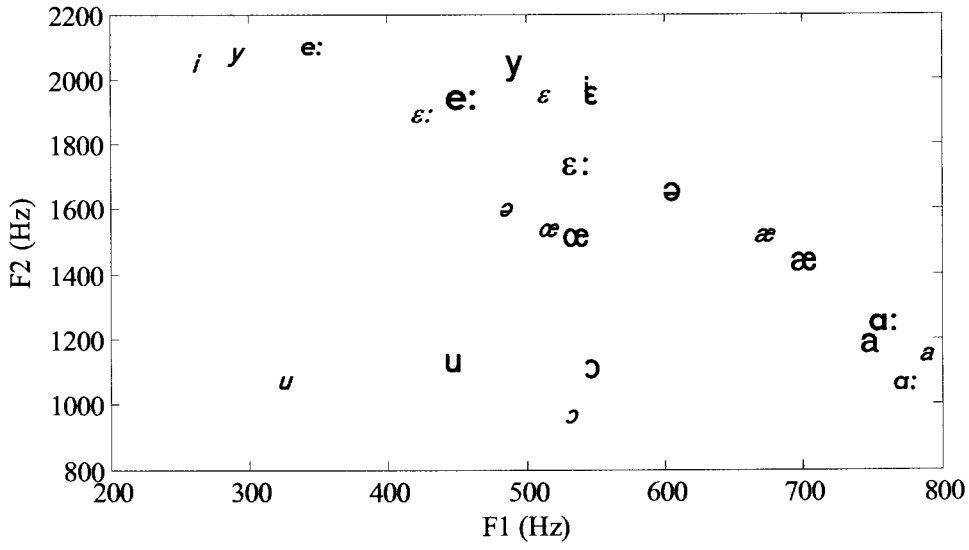

Figure 4.6. Vowel space of original (small italic font) and processed (large font) vowels with dynamic range compression

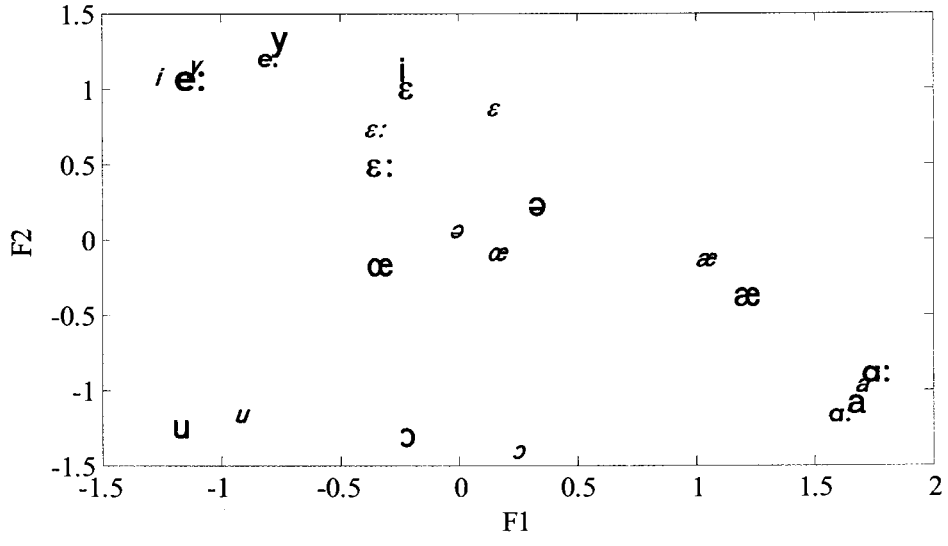

Figure 4.7. Vowel space for formant frequencies normalised using the Lobanov algorithm, processed with dynamic range compression. The processed vowels are larger than the original vowels, which are also printed in italics

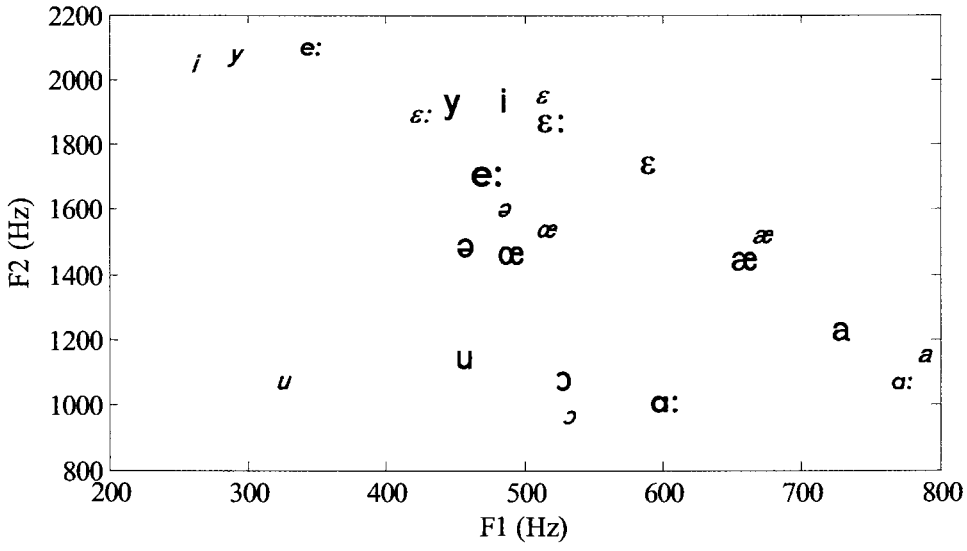

Figure 4.8. Vowel space of original (small italic font) and processed (large font) vowels without dynamic range compression

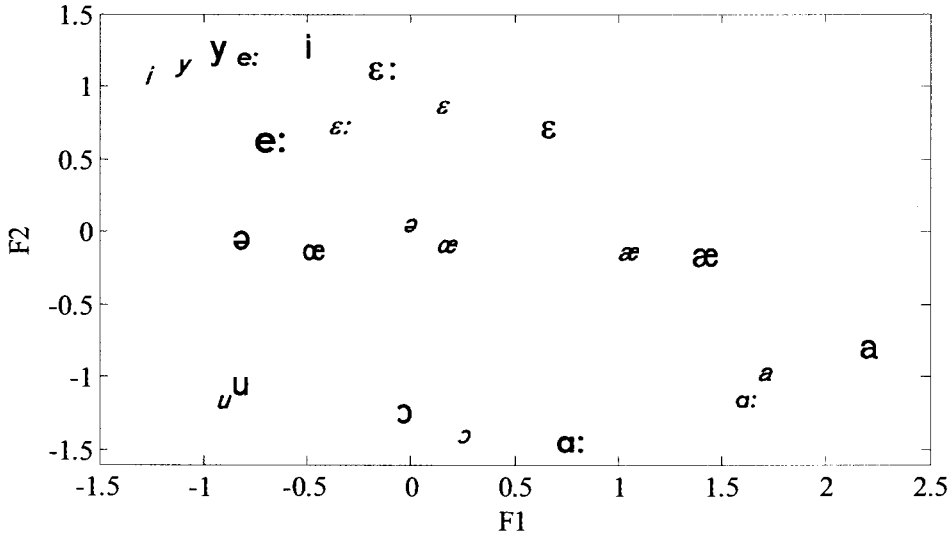

Figure 4.9. Vowel space for formant frequencies normalised using the Lobanov algorithm, processed without dynamic range compression. The processed vowels are larger than the original vowels, which are also printed in italics

Up to now, F_1 and F_2 as cues for vowel recognition have been discussed. Another important cue for vowel recognition, which is part of the three-dimensional vowel space, is the duration of a vowel. A distinction can be made between longer and shorter vowels

even though their formant frequencies are close together, for example /ɑː/ and /a/, which have a very similar formant frequency space, are recognised correctly based on the duration cue of the vowel.

The duration of the processed vowels is determined using PRAAT. The vowel must be isolated from the consonants /p/ and /t/ to record the duration of the vowel alone. The transition from the first consonant to the vowel is not very distinct. The transition is determined subjectively by listening to the speech token and splitting the speech token into separate parts. The duration is therefore an approximation, yet it still gives an accurate enough measure for analysis of the vowel characteristics. For the FITA analysis (discussed later), the vowels are separated into groups and classified as having either a longer or shorter duration. With the FITA analysis, the information of a specific group that is transmitted is determined according to these classifications. The exact duration is not extremely important for the FITA analysis, as the information transmitted to the listener is determined for groups of speech segments. Therefore, the vowels with a shorter duration are put in a group and the longer vowels are also grouped together. When the analysis is performed, the information transmitted for the groups with different durations is determined, not for the exact duration of a vowel.

When the vowel space is viewed in terms of the duration and F_1 (figures 4.10 and 4.11), one can predict a number of possible confusions. There is a specific group of vowels, /u, œ, y, i, ə/, with approximately the same duration and with a first formant frequency in the region of 440 Hz – 600 Hz. Based on the duration and F_1, it is predicted that these vowels will be confused often.

The duration of /æ/ and /a/ is very similar. Both the first and second formants are of the same order. Based on this and the comparable duration, these two vowels may easily be confused.

The vowel /eː/ has approximately the same duration as /ɛː/ and /ɑː/ in the original vowel space. In the processed vowel space the duration was considerably reduced, decreasing the chance of confusion with the other two vowels with a longer duration. In the processed

vowel space (figure 4.10), /e:/ and /ɔ/ moved closer to each other, generating the possibility of confusion between these two vowels.

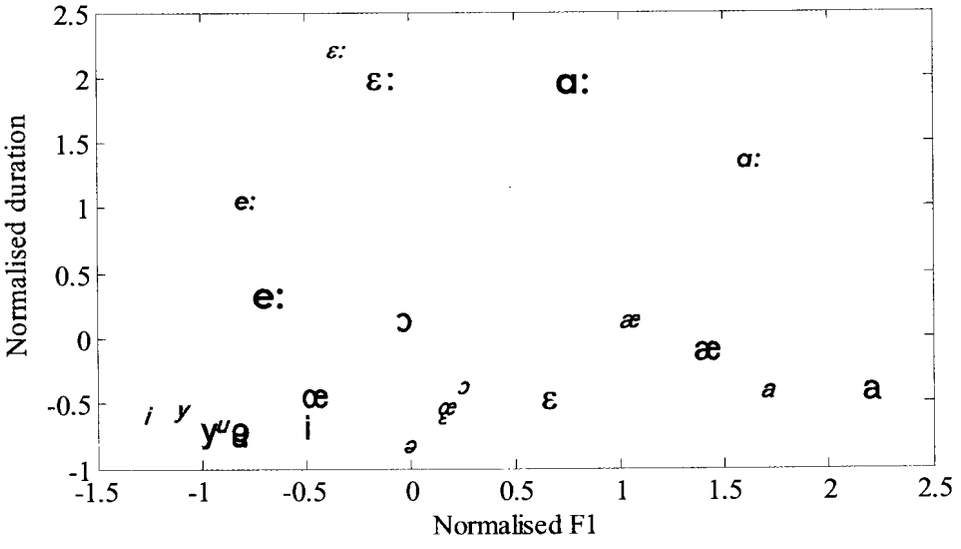

Figure 4.10. Normalised duration vs normalised F_1 for original (small italic font) and processed (large font) vowels with dynamic range compression

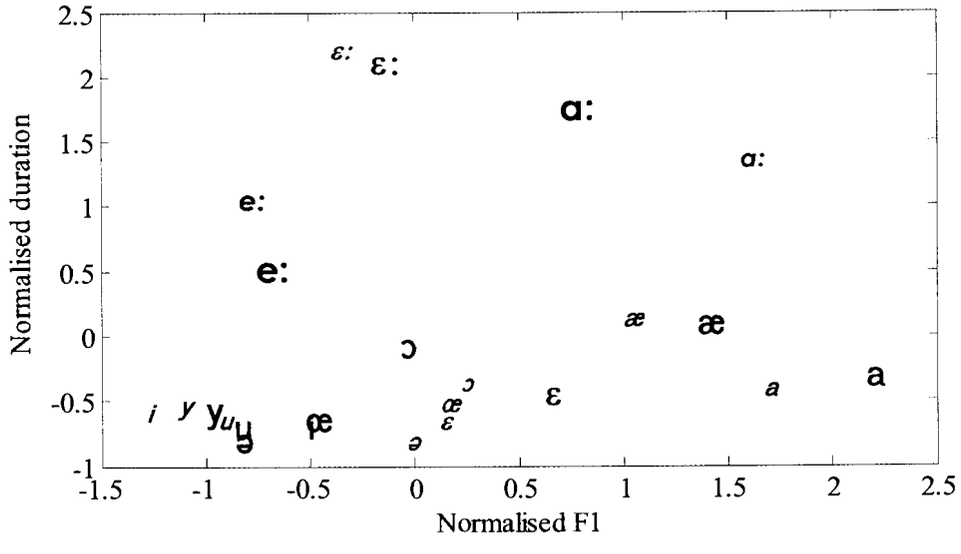

Figure 4.11. Normalised duration vs normalised F_1 for original (small italic font) and processed (large font) vowels without dynamic range compression

In figures 4.12 to 4.17, the Euclidean distance matrices are shown. The distance measures were obtained by determining an Euclidean distance between each vowel's normalised F_1,

F_2 and duration measures with the compared vowels' normalised F_1, F_2 and duration measures. The signal characteristics of the processed vowels, as summarised in table 4.1, are normalised and used to determine the Euclidean distances. By examining the distance measures, the vowels that are close to each other in the formant and duration spaces can be identified. The shorter the distance between two vowels, the more likely they are to be confused. These matrices serves as a prediction confusion matrix for vowels.

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	0.0	3.9	4.0	2.4	3.6	2.8	3.3	2.2	3.3	3.9	2.5	4.3
pIEt	i		0.0	2.5	2.9	0.3	2.5	1.0	2.1	1.3	1.4	2.8	0.6
pOEt	u			0.0	2.9	2.5	1.3	2.1	2.6	1.4	2.6	3.4	2.6
pAd	a				0.0	2.9	2.0	1.9	0.9	2.2	3.7	3.6	3.4
pEt	ɛ					0.0	2.4	1.0	2.0	1.2	1.2	2.5	0.7
pOt	ɔ						0.0	1.8	1.7	1.3	2.6	2.7	2.8
pIt	ə							0.0	1.2	0.8	2.0	2.9	1.5
pAt	æ								0.0	1.6	2.8	2.8	2.7
pUt	œ									0.0	1.7	2.6	1.6
pEEt	e:										0.0	1.9	1.2
pêt	ɛ:											0.0	2.9
pUUt	y												0.0

Figure 4.12. Three-dimensional Euclidean distance for duration, F1 and F2 of vowels with dynamic range compression

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	0.0	3.9	3.0	2.6	3.1	2.1	3.3	2.2	3.0	2.8	2.8	4.0
pIEt	i		0.0	2.4	3.4	1.3	2.6	1.4	2.5	1.4	1.4	2.9	0.5
pOEt	u			0.0	3.1	2.4	1.0	1.0	2.5	1.0	2.2	3.7	2.3
pAd	a				0.0	2.2	2.3	3.2	1.1	2.8	3.4	4.0	3.8
pEt	ɛ					0.0	2.2	1.8	1.3	1.5	1.7	2.7	1.7
pOt	ɔ						0.0	1.6	1.8	1.3	2.1	3.3	2.7
pIt	ə							0.0	2.4	0.4	1.6	3.3	1.3
pAt	æ								0.0	2.0	2.3	2.9	2.8
pUt	œ									0.0	1.5	3.1	1.5
pEEt	e:										0.0	1.8	1.3
pêt	ɛ:											0.0	2.9
pUUt	y												0.0

Figure 4.13. Three-dimensional Euclidean distance for duration, F1 and F2 of vowels without dynamic range compression

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	0.0	2.8	2.9	0.3	2.7	2.0	1.8	0.7	2.2	3.5	2.5	3.3
pIEt	i		0.0	2.5	2.9	0.0	2.4	1.0	2.1	1.3	0.9	0.6	0.6
pOEt	u			0.0	2.8	2.5	0.9	2.1	2.5	1.3	2.4	2.0	2.6
pAd	a				0.0	2.9	1.9	1.9	0.9	2.2	3.6	2.6	3.4
pEt	ɛ					0.0	2.4	1.0	2.0	1.2	0.9	0.5	0.6
pOt	ɔ						0.0	1.6	1.7	1.1	2.6	1.9	2.7
pIt	ə							0.0	1.1	0.8	1.7	0.8	1.5
pAt	æ								0.0	1.6	2.8	1.8	2.6
pUt	œ									0.0	1.5	0.7	1.6
pEEt	e:										0.0	1.0	0.5
pêt	ɛ:											0.0	0.9
pUUt	y												0.0

Figure 4.14. Two-dimensional Euclidean distance of F1 and F2 for vowels with dynamic range compression

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	0.0	2.9	1.6	1.6	2.2	0.8	2.1	1.4	1.8	2.5	2.7	3.1
pIEt	i		0.0	2.4	3.4	1.2	2.6	1.4	2.4	1.4	0.6	0.3	0.4
pOEt	u			0.0	3.0	2.4	0.8	1.0	2.4	1.0	1.7	2.3	2.3
pAd	a				0.0	2.2	2.3	3.1	1.0	2.8	3.3	3.1	3.8
pEt	ɛ					0.0	2.2	1.7	1.2	1.5	1.4	0.9	1.7
pOt	ɔ						0.0	1.4	1.8	1.2	2.0	2.4	2.7
pIt	ə							0.0	2.2	0.3	0.7	1.4	1.3
pAt	æ								0.0	1.9	2.3	2.1	2.7
pUt	œ									0.0	0.8	1.3	1.5
pEEt	e:										0.0	0.7	0.6
pêt	ɛ:											0.0	0.8
pUUt	y												0.0

Figure 4.15. Two-dimensional Euclidean distance of F1 and F2 for vowels without dynamic range compression

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	0.0	3.3	4.0	2.4	3.1	2.7	3.1	2.2	3.2	3.3	2.1	3.7
pIEt	i		0.0	0.9	1.9	0.3	0.8	0.5	1.5	0.3	1.4	2.8	0.5
pOEt	u			0.0	2.9	1.0	1.3	1.5	2.5	0.9	1.1	2.9	0.4
pAd	a				0.0	1.9	2.0	1.4	0.6	2.0	2.9	3.2	2.4
pEt	ɛ					0.0	0.5	0.6	1.4	0.2	1.2	2.4	0.6
pOt	ɔ						0.0	1.0	1.4	0.6	1.0	1.9	1.0
pIt	ə							0.0	1.1	0.7	1.8	2.9	1.1
pAt	æ								0.0	1.6	2.4	2.7	2.0
pUt	œ									0.0	1.2	2.5	0.5
pEEt	e:										0.0	1.9	1.1
pêt	ɛ:											0.0	2.8
pUUt	y												0.0

Figure 4.16. Duration-F1 distance measure between vowels with dynamic range compression

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	0.0	2.8	3.0	2.6	2.2	2.0	3.1	1.8	2.8	1.9	1.0	2.9
pIEt	i		0.0	0.3	2.7	1.2	0.8	0.4	2.0	0.1	1.3	2.9	0.5
pOEt	u			0.0	3.0	1.5	1.0	0.1	2.4	0.3	1.3	3.0	0.2
pAd	a				0.0	1.5	2.3	3.1	0.9	2.7	3.0	3.4	3.1
pEt	ɛ					0.0	0.8	1.6	0.9	1.2	1.7	2.7	1.6
pOt	ɔ						0.0	1.1	1.4	0.7	0.9	2.3	1.0
pIt	ə							0.0	2.4	0.4	1.4	3.1	0.2
pAt	æ								0.0	2.0	2.2	2.6	2.4
pUt	œ									0.0	1.2	2.8	0.5
pEEt	e:										0.0	1.7	1.2
pêt	ɛ:											0.0	2.9
pUUt	y												0.0

Figure 4.17. Duration-F1 distance measure between vowels without dynamic range compression

When the matrices are normalised row by row, the distances between vowels can be rounded to 0, 0.25, 0.5 and 0.75 (figures 4.18 and 4.19). For a distance of 0.75, the vowels are expected to be confused more often than not. For a distance of 0.5, the vowels might be confused often; for a distance of 0.25, the vowels are expected to be confused sometimes and for a distance of 0, no confusions are expected. For the diagonal, the distance measure gives an indication of how many times the vowel will not be confused with other vowels. For a distance measure of 1, it is expected that the vowel will almost never be confused with any other vowel.

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	1.00	0.00	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.00	0.25	0.00
pIEt	i		0.75	0.25	0.00	0.75	0.25	0.50	0.25	0.50	0.50	0.25	0.75
pOEt	u			0.75	0.25	0.25	0.50	0.25	0.25	0.50	0.25	0.00	0.25
pAd	a				0.75	0.00	0.25	0.25	0.75	0.25	0.00	0.00	0.00
pEt	ɛ					0.50	0.25	0.50	0.25	0.50	0.50	0.25	0.75
pOt	ɔ						1.00	0.25	0.25	0.50	0.00	0.00	0.00
pIt	ə							0.50	0.50	0.50	0.25	0.00	0.50
pAt	æ								0.75	0.25	0.00	0.00	0.00
pUt	œ									0.50	0.25	0.00	0.50
pEEt	e:										0.75	0.25	0.50
pêt	ɛ:											1.00	0.00
pUUt	y												0.50

Figure 4.18. Summary of predictions of confusions for vowels processed with dynamic range compression

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	1.00	0.00	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.25	0.00
pIEt	i		0.75	0.25	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.50
pOEt	u			0.75	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.00	0.25
pAd	a				1.00	0.25	0.25	0.00	0.25	0.25	0.00	0.00	0.00
pEt	ɛ					0.75	0.25	0.25	0.25	0.25	0.25	0.00	0.25
pOt	ɔ						0.75	0.25	0.25	0.25	0.25	0.00	0.00
pIt	ə							0.75	0.25	0.50	0.25	0.00	0.25
pAt	æ								0.75	0.25	0.00	0.00	0.00
pUt	œ									0.75	0.50	0.00	0.50
pEEt	e:										0.75	0.25	0.50
pêt	ɛ:											1.00	0.25
pUUt	y												0.75

Figure 4.19. Summary of predictions of confusions for vowels processed without dynamic range compression

4.3.1.3 Results from experimental study on vowel confusions

Results from the experiments conducted with normal-hearing persons in the form of confusion matrices are presented in this section. Section 3.4 explained the procedure followed and the experimental parameters. A confusion matrix is compiled by recording the response of a listener to a specific stimulus. The diagonal of the matrix represents the stimuli recognised correctly, while incorrect responses are scattered across the matrix. By examining these matrices, typical confusions between vowels are determined. The results obtained with normal-hearing listeners are compared to results obtained with cochlear implantees by comparing results for FITA analyses.

The recognition of vowels improved with the number of experiments completed, which is expected because the normal-hearing listeners become accustomed to the sound of the processed vowels. It can be seen from figures 4.20 and 4.21 that after about three or four experiments, the percentage of correctly recognised vowels start to stabilise at a specific value – the slope between experiments three and five is less than the slope between experiments one and three. The percentage of correct scores is determined by the summation of the diagonal of a confusion matrix and the division of this value by the total number of stimuli.

The percentage recognised correctly is generally higher in cases where dynamic range compression is not included – the recognition starts at a higher percentage and tends to stabilise at a higher percentage.

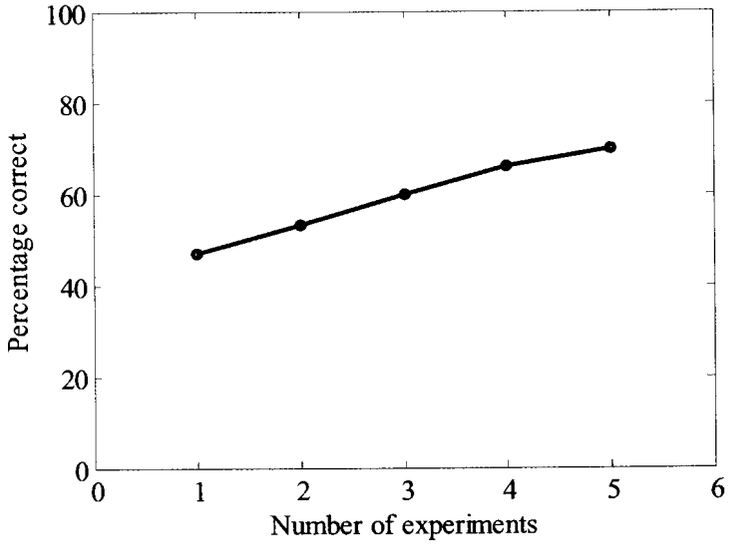

Figure 4.20. Learning curve over time for vowels with dynamic range compression

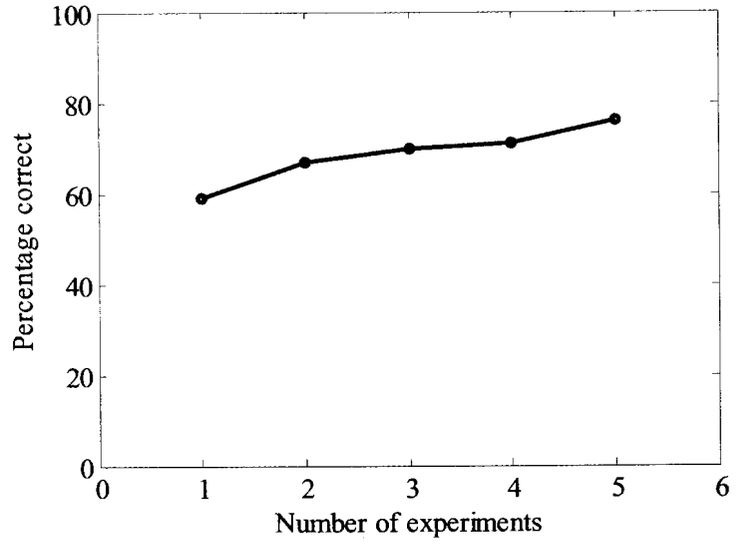

Figure 4.21. Learning curve over time for vowels without dynamic range compression

When considering the recognition of individual vowels, in the earlier experiments the vowels on the edges of the vowel space (figure 4.6) are often confused. As the listener becomes accustomed to the sound of the speech tokens, these confusions decrease and more confusions appear between vowels clustered together in the vowel space.

Because vowels are recognised according to the formant frequencies relative to one

another, the processed vowels that lie close to the original vowels in the normalised vowel space may possibly be confused. From the results of the earlier experiments (not shown here), this does not appear to be the case; it is rather the vowels that lie on the edges of the processed vowel space that are confused in the earlier experiments. In the later experiments, it is the vowels that are clustered together in the F_1-F_2 and F_1-duration spaces that are confused, even though the clusters have moved relative to each other.

When comparing the trends of confusions for the vowels processed with dynamic range compression and those without the compression, there are no notable differences. The confusions have the same type of distribution, but with different magnitudes. The overall percentage recognised correctly for the processing without dynamic range compression is higher than for the vowels processed without dynamic range compression, indicating that the individual vowels will also be recognised more accurately for the vowels without compression. This is evident in the results. This indicates that the inclusion of dynamic range compression does not change the type of confusions, it only reduces the overall percentage of accurate recognition.

In the following paragraphs, all the confusions for the vowels with dynamic range compression will be analysed and explained with respect to their F_1, F_2 and duration characteristics. The discussion of vowels without dynamic range compression follows the discussion of vowels with dynamic range compression. The confusion matrices for the results of the experiments with the vowels before and after dynamic range compression are presented in figures 4.22 to 4.23. The confusion matrices were obtained by pooling all the participants' results for experiments four and five.

Stimulus		Response: pAAt ɑ:	pIEt i	pOEt u	pAd a	pEt ɛ	pOt ɔ	pIt ə	pAt æ	pUt œ	pEEt e:	pêt ɛ:	pUUt y
pAAt	ɑ:	199	0	0	3	0	0	0	0	0	0	0	0
pIEt	i	0	141	2	0	2	0	8	0	5	0	0	42
pOEt	u	0	0	93	32	1	39	13	12	10	0	0	0
pAd	a	0	0	13	173	0	12	0	2	0	0	0	0
pEt	ɛ	0	42	3	0	85	1	36	2	18	0	1	12
pOt	ɔ	0	1	33	7	1	137	0	1	1	19	0	0
pIt	ə	0	10	1	0	23	0	116	1	49	0	0	0
pAt	æ	0	0	0	8	0	11	0	180	0	0	1	0
pUt	œ	0	6	2	7	32	36	41	4	61	0	0	11
pEEt	e:	0	0	0	0	0	0	1	0	0	199	0	0
pêt	ɛ:	5	0	0	0	3	0	0	1	0	0	171	20
pUUt	y	0	109	1	0	6	0	11	0	7	1	1	64

Figure 4.22. Confusion matrix for vowels with dynamic range compression

Stimulus		Response: pAAt ɑ:	pIEt i	pOEt u	pAd a	pEt ɛ	pOt ɔ	pIt ə	pAt æ	pUt œ	pEEt e:	pêt ɛ:	pUUt y
pAAt	ɑ:	199	0	0	0	0	0	0	0	0	0	0	1
pIEt	i	0	165	7	0	0	0	0	0	2	0	0	26
pOEt	u	0	0	126	47	0	6	0	18	3	0	0	0
pAd	a	0	0	0	196	0	2	0	2	0	0	0	0
pEt	ɛ	0	3	1	0	73	5	69	14	34	0	0	1
pOt	ɔ	0	0	0	0	0	199	0	0	1	0	0	0
pIt	ə	0	0	2	1	2	34	54	1	105	0	0	1
pAt	æ	0	0	0	4	0	0	0	196	0	0	0	0
pUt	œ	0	0	2	0	12	52	26	2	106	0	0	0
pEEt	e:	1	0	0	0	0	0	0	0	0	199	0	0
pêt	ɛ:	1	0	0	0	0	0	0	0	0	0	179	20
pUUt	y	0	85	16	0	0	0	0	0	2	0	0	97

Figure 4.23. Confusion matrix for vowels without dynamic range compression

In figures 4.24 and 4.25, the probability of a correct response is plotted against the calculated Euclidean distance. As the Euclidean distance between two vowels increases,

the probability of a correct response will increase. The procedure to determine the probability of a correct response is as follows: firstly the confusion matrices from the experiments are normalised so that the sum of each row is 1. The values in the normalised matrices are used as the probabilities of an incorrect response for a specific stimulus. All the probabilities at a particular Euclidean distance are added together, this sum is multiplied with the number of times the particular Euclidean distance occurred and divided by the total number of Euclidean distances in the matrix. This is then subtracted from 1 to reach the probability of a correct response. These probabilities are then plotted against the corresponding Euclidean distance.

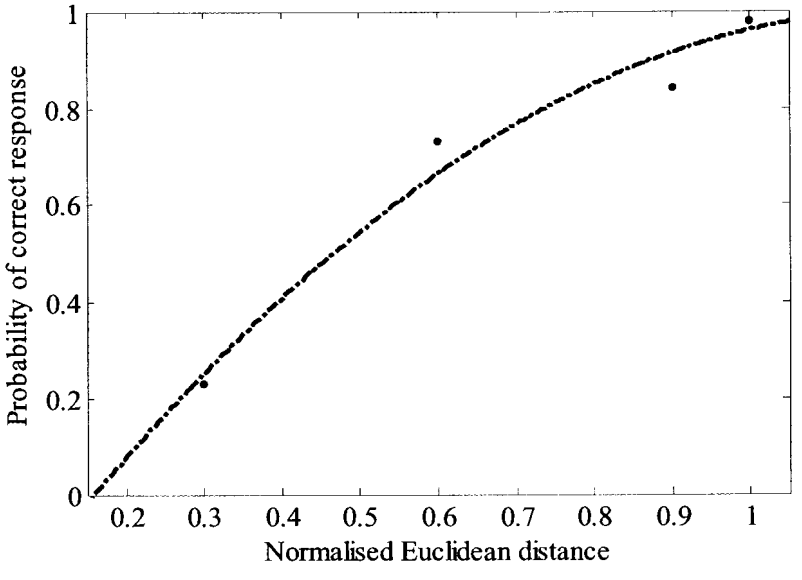

Figure 4.24. Probability of correct response to stimuli as a function of normalised Euclidean distance for vowels after dynamic range compression

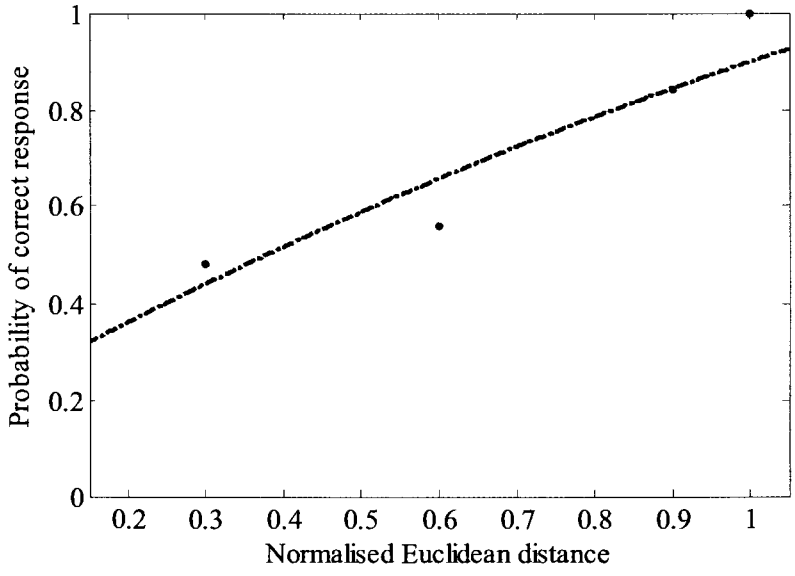

Figure 4.25. Probability of correct response to stimuli as a function of normalised Euclidean distance for vowels before dynamic range compression

Vowels with dynamic range compression

The overall percentage that is correctly recognised for vowels with dynamic range compression is 67 %, pooled over all listeners. For the processed vowels with dynamic range compression, confusions were apparent, as seen from an inspection of figures 4.18 and 4.19. For every confusion a possible explanation is given. In section 4.3.2.1, possible confusions were predicted by visually inspecting the formant spaces. In this section, the actual confusions are explained in terms of a physical measure, the Euclidean distances between vowels.

When /ɑː/ is used as stimulus, it is confused with /a/. This can be ascribed to the very short distance measure in the two-dimensional space F_1-F_2. The vowels that are closest to /ɑː/ in the three-dimensional space of figure 4.12, are /a/ and /æ/. There are a few confusions with /a/, which is natural, but there is none with /æ/. When looking at the two-dimensional space, figure 4.14, it can be seen that the distance between /ɑː/ and /a/ is significantly less than for /æ/, explaining the more regular confusion with /a/.

The vowels /i, u, y, ə, œ, ɛ/ all have approximately the same duration and also have first formant frequencies that are grouped together, all with an Euclidean distance of less than 1.5 between each other. In the F_1-duration vowel space they form a cluster. For the stimulus /i/, all these vowels are recorded as responses, with /y/ the main confusing vowel. The two vowels /i/ and /y/ are not only close to each other in the F_1-duration vowel space, but also in the F_1-F_2 vowel space.

For the stimulus /u/, the response is often /ɔ/. These two vowels have a small three-dimensional Euclidean distance; it is the smallest distance between /u/ and any of the other vowels, thus explaining the regular confusion. Other vowels that are also confused for /u/ are /a, œ, ə/. For these vowels it is the duration that causes the confusions. These vowels all have a similar duration, as can be seen in the duration-F_1 distance measure (figure 4.16).

The vowels /u, ɔ, a/ have similar second formant frequencies and are confused to a lesser degree. The duration of /æ/ is close to that of /a/, which explains the confusion between these two vowels. However, /a/ is not confused very often with /æ/, suggesting the possibility that duration is not the most significant characteristic affecting vowel recognition.

The most significant confusion for vowels with dynamic range compression is between /ɛ, ə, ɔ, œ/, with only 31 % recognised correctly when /œ/ is used as stimulus. When /y/ is used as stimulus, the response is /i/ almost 55 % of the time. When /i/ is used as stimulus, 17 % of the time the response is /y/. These are significant figures. The other vowels that are confused very often are /œ, ɔ, ə/ and for the stimulus /ɛ/, the response is often /i, y, ə, œ/.

Vowels without dynamic range compression

The confusions for the vowels without dynamic range compression are very similar to those for the vowels with dynamic range compression. The overall percentage of vowels recognised correctly are higher: 75 % as opposed to 67 % pooled over all the listeners.

The vowels /i, y/ have a very short three-dimensional distance between them, 0.5, explaining the confusions between these two vowels. The confusions are present whether /i/ or /y/ is used as the stimulus.

For the stimulus /u/, the response is often /a/. From the vowel spaces described here, it is not clear why they are confused so regularly. The confusions with /œ/ and /ɔ/ can be explained by the small three-dimensional Euclidean distance between them. Other vowels that are also confused for /u/ are /ɛ/ and /ə/. For these vowels it is also the short three-dimensional distance that explains the confusions. These vowels all have a similar duration, as can be seen in the duration vowel space, figure 4.17.

The significant confusions between /œ/ and /ə/, for the stimulus /ə/, can easily be explained when looking at the three-dimensional distance matrix, figure 4.13. Together with the confusion between /ɛ/ and /ə/, with /ɛ/ as stimulus, they are the most common. The minimum distance between any two vowels is, not surprisingly, between /œ/ and /ə/. It is therefore not unexpected that these two vowels are confused regularly.

The confusions between /ɛː/ and /y/ can be explained by the short two-dimensional duration-F_1 distance between them. The difference in duration for /ɛː/ and /ɑː/ is almost zero, contributing to the confusion between these two vowels. They are far removed in the F_1-F_2 vowel space. This will explain why they are not confused regularly.

The most significant confusion, as for the vowels with dynamic range compression, is between /ɛ, ə, œ/, the responses being almost evenly distributed when the stimulus is /ɛ/. When /y/ is used as stimulus, the response is /i/ more than 40 % of the time, similar to the recognition of vowels with dynamic range compression.

4.3.1.4 FITA analysis

A FITA analysis was done for both the vowel and consonant recognition confusion matrices in order to determine which cues are transmitted most effectively and to compare data recorded for cochlear implant users. The output of the FITA analysis is a measure of

covariance between input and output. This measure is calculated through the procedure described next. If the input variable is x with probability p_i, i = 1, 2, ..., k, the mean logarithmic probability (MLP) is defined as

$$MLP(x) = E(-\log p_i) = -\sum_i p_i \log p_i \,. \tag{4.1}$$

A similar expression is defined for the output y with probability p_j, j = 1, 2, ..., m. A measure of covariance of input with output is given as

$$T(x;y) = MLP(x) + MLP(y) - MLP(xy) = \sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}}, \tag{4.2}$$

where p_{ij} is the probability of the joint occurrence of input i and output j. T(x;y) is the transmission from x to y. When a response is closely correlated with a specific stimulus, the transmission of a specific feature is good and T(x;y) will be near unity (Miller and Nicely, 1955). In this study, the stimulus for vowels is classified as in table 4.2 and the response is a confusion matrix from the experimental study. T(x;y) is determined for each of the features in table 4.2.

The characteristics of the vowels that were used were duration, F_1 and F_2. The classifications were different for the processed and original vowels (Pretorius et al., 2005). They are classified as shown in tables 4.2 and 4.3. The classifications of the vowels are determined using the guideline summarised in table 4.4. The vowels are grouped together according to their classifications to determine the percentage information transmitted for a specific characteristic. For example, /ɑ:/ has a long duration (2), a high F_1 value (2) and a medium F_2 value (2). The confusion matrices are analysed using these classifications to determine whether long vowels are distinguished from short vowels, vowels with low F_1 frequencies are distinguished from vowels with high F_1 frequencies and so forth.

Note that the classification of the duration of /æ/ and /ɔ/ changed after processing. This is due to the noise that causes the start and end of the vowels to be less apparent as before processing.

Table 4.2. Classification of processed vowels for FITA analysis

	pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
Duration	2	1	1	1	1	2	1	2	1	2	2	1
F_1	2	1	1	2	2	1	1	2	1	1	1	1
F_2	2	3	2	2	3	2	2	2	2	3	3	3

Table 4.3. Classification of original vowels for FITA analysis (Pretorius et al., 2005)

	pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
Duration	2	1	1	1	1	1	1	1	1	2	2	1
F_1	2	1	1	2	2	2	2	2	2	1	1	1
F_2	2	3	1	2	3	1	2	2	2	3	3	3

Table 4.4. Ranges of duration, F1 and F2 used for classification of processed vowels

	Duration	F_1	F_2
1	**0 - 100**	**0 - 540**	**0 - 960**
2	**> 100**	**540 - 900**	**960 - 1 700**
3		**> 900**	**> 1 700**

The results of the FITA analyses for the vowels are summarised in tables 4.5 to 4.7. The confusion matrices used for the FITA analyses were the pooled confusion matrices from all the participants. The results for the cochlear implant users were obtained from Pretorius et al. (2005). Nine post lingual deaf adults completed the vowel tests and eleven post lingual deaf adults completed the consonant tests. All the participants used the Nucleus implant, using either the SPEAK or ACE speech processor.

Table 4.5 . Results of FITA analysis for cochlear implant users listening to the original vowels

	Percentage information transmitted
Duration	63
F_1	43
F_2	50

Table 4.6. Results of FITA analysis for processed vowels with dynamic range compression

	Percentage information transmitted
Duration	60
F_1	43
F_2	57

Table 4.7. Results of FITA analysis for processed vowels without dynamic range compression

	Percentage information transmitted
Duration	67
F_1	55
F_2	65

In figures 4.26 and 4.27 the percentage information transmitted for all the normal-hearing listeners is shown. Figure 4.26 reflects the transmitted information for the vowels processed with dynamic range compression and figure 4.27 reflects the vowels processed without dynamic range compression.

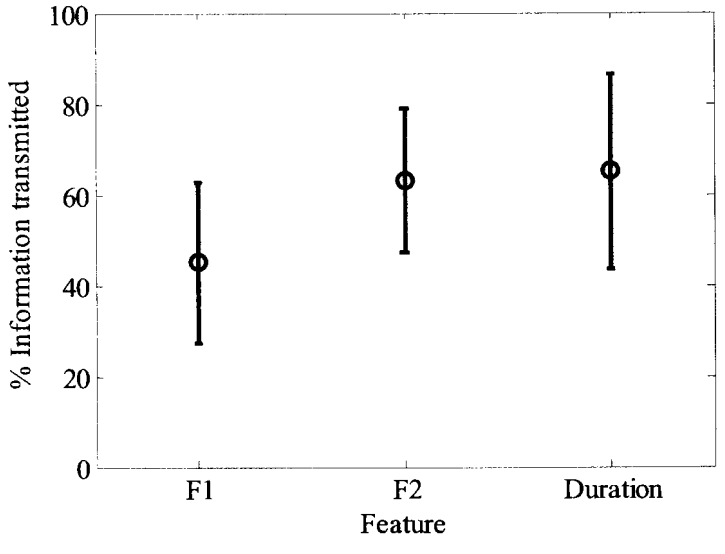

Figure 4.26. Indication of information transmitted to normal-hearing listeners for vowels processed with dynamic range compression. The average and standard deviation percentage information transmitted is shown for duration, F_1 and F_2

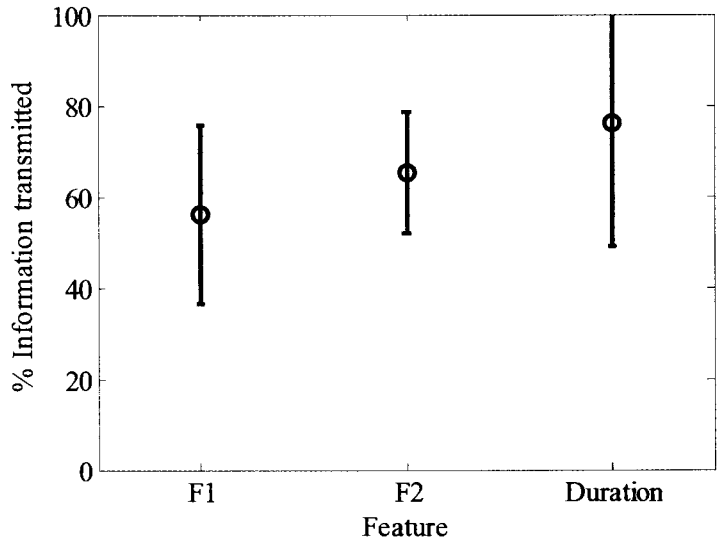

Figure 4.27. Indication of information transmitted to normal-hearing listeners for vowels processed without dynamic range compression. The average and standard deviation percentage information transmitted is shown for duration, F_1 and F_2

From the FITA analysis, it can be seen that for the acoustic simulation with dynamic range compression, the results correspond very well with those found for the original vowels presented to cochlear implant users. The information transmitted for F_1 is in both cases only 43 %. Information transmitted for duration is 3 % less for the processed vowels with dynamic compression than for the original vowels. The difference of 7 % for the information transmitted for F_2 is less than 10 %. The feature that is transmitted most effectively is the duration of a vowel. Referring to the low percentage of information transmitted for F_1, it appears that a lot of the information of F_1 is lost during the processing of the vowels. This result is consistent for the experiments before and after dynamic range compression and for experiments with cochlear implant listeners. This gives a strong indication that duration of vowels is transmitted most effectively and F_1 information is transmitted poorly for the processing of Afrikaans vowels when processed through the acoustic model.

The vowels without dynamic range compression yield better results than either the original vowels recognised by cochlear implant users or vowels recognised with the dynamic range compression. The difference in FITA results between the original vowels and the vowels

without the dynamic range compression is in the order of 10 % on average. This can be seen in figure 4.28. This shows that more information is transmitted on average with the acoustic simulation before dynamic range compression.

A linear fit was performed on the results from the FITA analysis for cochlear implantees and for FITA analysis performed on the normal-hearing listeners' results (the confusion matrices were not pooled). The curve fitting is shown in figure 4.28 for the FITA results before and after dynamic range compression and for results found with cochlear implant users. A t-test was performed between the results obtained before dynamic range compression and results from cochlear implant users, and between the results obtained after dynamic range compression and results from cochlear implant users. Three separate t-tests were performed for F_1, F_2 and duration respectively. Each listener's results were analysed separately through FITA and used in the t-test. From the results of the t-test, it appears that only the results obtained after dynamic range compression belong to the same probability density function as the results from the experiments done with the cochlear implant users with a significance level of 5 %. The FITA results from the acoustic simulation before dynamic range compression do not belong to the same pdf as those from the experiments done with cochlear implant users.

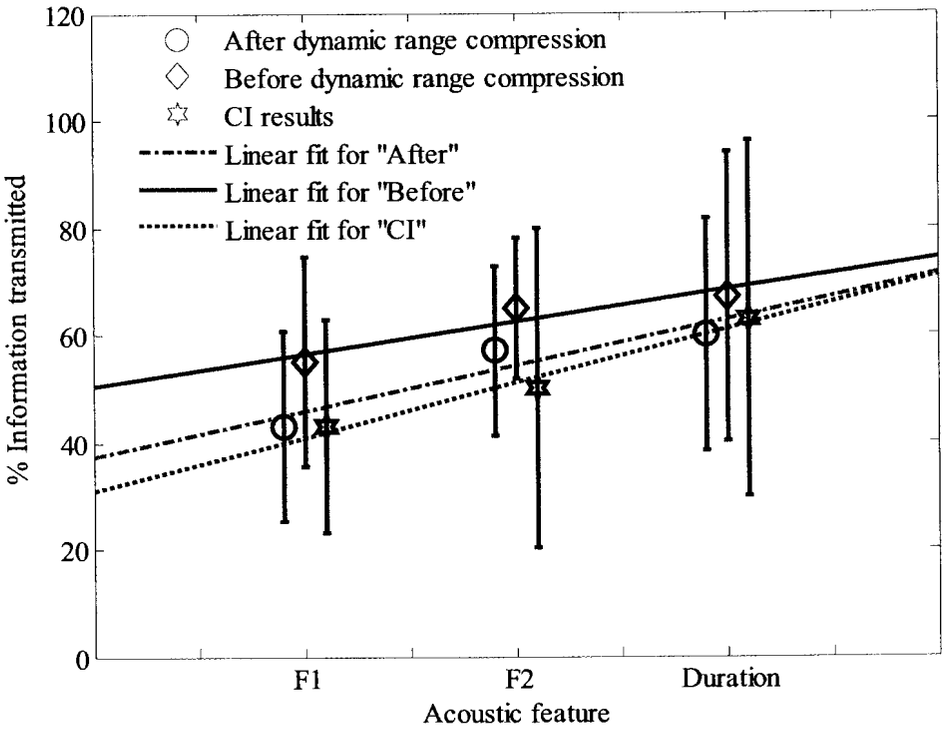

Figure 4.28. Linear fit for FITA results of vowels using the acoustic simulation with dynamic range compression (r = 0.863), acoustic simulation without dynamic range compression (r = 0.858) and cochlear implant users (r = 0.985)

The linear fit and t-test suggests that the acoustic simulation with dynamic range compression models cochlear implants slightly better than the simulation without the dynamic range compression. This might be expected because of the presence of dynamic range compression in cochlear implants; the compression should also be modelled in the acoustic simulation.

4.3.2 Consonant confusions

4.3.2.1 Acoustic analysis of consonants at the output of the acoustic model

The results for the experiments with consonants were analysed using multidimensional scaling. This is a procedure where coordinates according to confusions between the consonants are determined. These coordinates give an indication of which consonants are

confused most often. Coordinates for a number of dimensions are determined. These coordinates, grouped according to a specific dimension, can then be correlated with specific acoustic features and signal characteristics to determine which features or characteristics are transmitted to the listener in order to recognise the consonants. In this study, an Individual Differences Scaling (INDSCAL) (Carroll and Chang, 1970a) analysis was performed on confusion matrices, using the software package PRAAT.

Consonants are classified according to specific features generated during the articulation process, including duration and manner of articulation. By grouping the consonants according to these features, predictions can be made as to which consonants might be confused in actual experiments. Listeners recognise consonants based on these features (Borden and Harris, 1994; Miller and Nicely, 1955). Consonants that are grouped according to their acoustic features also have distinct signal characteristics. When a consonant is presented acoustically to a listener, the consonant is recognised according to the transmitted signal characteristics (acoustic cues), for example the envelope variation, energy content and duration of consonants. Possible confusions include confusions between the plosives, between the fricatives and between the nasals, glides and liquids.

An analysis of the confusion matrices can also be performed by dividing the consonants into specific groups of stimuli and their respective responses (Rosen, 1992; Wang and Bilger, 1973). This will give an indication of confusions within groups with similar acoustic cues rather than confusions between groups with different acoustic cues. By doing this, a smaller confusion matrix is formed and confusions can be analysed according to the sets of acoustic cues. When a new score is calculated for the percentage correctly recognised, it will be greater for the smaller matrix than for the original matrix. This is because a high percentage of confusions occur within a specific group and all the confusions within a group that were wrong originally, will now be considered 'correct'.

In the following section, the chosen set of acoustic features for the classification of consonants are described in more detail. Thereafter the signal characteristics that are linked to the acoustic features are discussed. Both the acoustic features and signal characteristics can be correlated with the coordinates obtained from the INDSCAL analysis

to determine what signal characteristics and acoustic features are responsible for the recognition of consonants.

Articulatory features

Consonants are divided into groups with similar acoustic features. It is predicted that the consonants within a group will be confused with one another because of the similar production of the sounds. Different groups of acoustic features that can be used to classify the consonants are described in this section.

There are different ways of producing a consonant; this is described as the manner of articulation. Manners of articulation include nasal, plosive, fricative and liquid, as will be discussed in more detail next (Borden and Harris, 1994; Miller and Nicely, 1955).

For consonants produced with a burst (plosive consonants), the flow of air is totally blocked somewhere in the vocal tract. These consonants include /p/ and /b/. The consonants /l/ and /r/ are both liquid consonants. They are voiced consonants produced by constricting the vocal tract.

Voiced consonants are differentiated from voiceless consonants in articulatory terms according to the manner in which the vocal cords are used. For the voiced consonants, /b, d, m, n, r, l, j, z, v, x/, the vocal cords vibrate and for the voiceless consonants, /p, t, k, s, f/, the vocal cords do not vibrate (Miller and Nicely, 1955). In acoustic terms, the voiceless consonants are aperiodic or noisy in character and the voiced consonants have a definite periodic or line-spectrum component superimposed on the noise. Other characteristics that are also worth mentioning is that voiceless consonants seem to have a higher signal intensity. In speech segments presented as /a/-CONSONANT-/a/, the voiceless stops have a kind of breathy noise between the release of pressure and the beginning of the following vowel, they may also be a bit shorter than the voiced stops. Voiced stops include /b, d/, the air is totally blocked before the consonant is produced using the vocal cords.

When referring to a nasal consonant, the articulation can be described as the lips being closed and the pressure released through the nose. This nasal articulation provides a distinct acoustic feature by which a consonant can be recognised. The two nasal consonants are /m/ and /n/. It also appears that the two nasals are longer in duration and more intense than the stop or fricative consonants. These two consonants are periodic and do not have the aperiodic component of noisiness.

When the articulators are brought close together and air is forced between them, a turbulence or friction noise is produced. This friction noise distinguishes /s, z, f, v, x/ from /p, t, k, b, d, m, n, r, l, j/; for the latter the articulators are closed completely producing a stop or nasal consonant. The turbulence is characteristic of the fricative consonants. The stops are characterised by a silence followed by a pop and the nasals are characterised by a periodic, almost vowel-like resonance.

Although the consonants /s, f, v, z/ are distinguished as long, intense, high-frequency noises, the more important characteristic is their duration. They are somewhat longer than the other consonants and it is believed that this feature sets them apart from the other consonants.

The place in the mouth where the major constriction of the vocal passage occurs can be divided into three positions: front, middle and back. The consonants can be grouped as /p, b, f, v, m/ for the front, /t, d, s, z, n, l, r/ in the middle and /k, g, j/ at the back. It is difficult for a listener to determine the place of articulation of a consonant by listening to speech, the information transmitted through this feature is the most difficult to identify. For a detailed explanation on how place of articulation is used as an acoustic feature, refer to Miller and Nicely (1955).

The specific features of the consonants are summarised in table 4.8. The classifications used to group the consonants are explained as follows: a consonant is either plosive (1) or not (2), voiced (2) or voiceless (1), is produced in the front (1), middle (2) or back (3) of the mouth, is either a nasal (2) or not (1), has liquidity characteristics (2) or not (1) and is either a fricative (1) or not (2). The manner of articulation is classified as plosive (1),

liquid, nasal or glide (2), or fricative (3). These classifications group the consonants produced together in a similar way to analyse the consonants according to information transmitted.

Table 4.8 is used to analyse the confusion matrices obtained from the experiments. When the classifications from table 4.8 are applied to the confusion matrices, the percentage information transmitted for each acoustic feature can be obtained from a FITA analysis. In this way it can be determined whether information about a specific feature is transmitted sufficiently or not.

Table 4.8. Classification of consonants for FITA and INDSCAL analysis

	Burst	Voicing	Manner	Place	Nasality	Liquidity	Affrication
aPa	1	1	1	1	1	1	2
aTa	1	1	1	2	1	1	2
aKa	1	1	1	3	1	1	2
aBa	1	2	1	1	1	1	2
aDa	1	2	1	2	1	1	2
aMa	2	2	3	1	2	1	2
aNa	2	2	3	2	2	1	2
aRa	2	2	3	2	1	2	2
aLa	2	2	3	2	1	2	2
aJa	2	2	3	3	1	1	2
aSa	2	1	2	2	1	1	1
aZa	2	2	2	2	1	1	1
aFa	2	1	2	3	1	1	1
aWa	2	2	3	3	1	1	2
aGa	2	2	2	3	1	1	1

Acoustic properties

In order to obtain non-arbitrary perceptual dimensions based on the confusions of consonants that can be associated with signal characteristics, the multidimensional scaling approach is followed. This is discussed later. To confirm the predictions made using the acoustic features (refer to section 1), the results from the multidimensional scaling were used in addition to the FITA analysis. Through this approach, it can be proven that

articulatory features do represent valid perceptual dimensions (Wang and Bilger, 1973). The specific acoustic properties that were analysed for the speech tokens used in this study are summarised in table 4.9. These acoustic properties were also correlated with the acoustic features recorded in table 4.8 to determine the relationship between them. In order to calculate the relevant signal characteristics that are used as acoustic cues, the output of the acoustic model was used, as will be explained in the following paragraphs.

Both the consonants before and after dynamic range compression were analysed to determine their acoustic properties (tables 4.9 and 4.10). These acoustic properties were chosen based on those described in Van Wieringen and Wouters, (1999). The duration was measured, using spectrograms in PRAAT, from the onset of the consonant, including the silence of the voiceless plosives. The root-mean-square (RMS) of a sliding window was calculated to determine the energy content of the consonant. The sliding window was used for a smooth transition from one window to another. The energy in the consonants was calculated using 512-sample time windows with 75 % overlap (implemented with a Hanning window similar to the analysis explained in section 3.3.2.3) – this vector of values was used to determine the peak, median and the ratio of minimum to peak energy of the signal as in equation 4.3,

$$\begin{aligned} P &= \max(20\log(V_{RMS})) \\ M &= median(20\log(V_{RMS})) \\ M/P &= \frac{\min(V_{RMS})}{\max(V_{RMS})} \end{aligned} \quad . \qquad (4.3)$$

Table 4.9. Acoustic properties of consonants with dynamic range compression – duration (ms), peak and median level energy (dB), minimum to peak energy ratio, peak and median level energy after low-pass filtering (dB) and envelope variation (dB) between 20 and 200 Hz

	D (ms)	P (dB)	M (dB)	Minimum/ peak	P-LPF (dB)	M-LPF (dB)	EV (dB)
aPa	266	-15.98	-34.23	0.0889	-53.338	-71.799	7.290
aTa	217	-16.05	-34.42	0.0938	-50.943	-71.513	6.416
aKa	297	-15.86	-21.96	0.0998	-51.134	-60.125	7.565
aBa	210	-14.85	-27.77	0.1321	-49.921	-63.337	6.948
aDa	244	-14.82	-23.67	0.0938	-49.852	-62.638	6.741
aRa	228	-17.04	-20.18	0.3259	-51.713	-56.507	3.458
aLa	255	-16.05	-18.22	0.6161	-49.322	-56.560	3.157
aJa	187	-17.38	-20.09	0.6088	-51.580	-57.457	3.579
aMa	283	-15.85	-18.47	0.5486	-50.690	-56.482	3.732
aNa	237	-16.99	-19.63	0.5542	-48.387	-57.616	3.821
aSa	223	-17.11	-19.85	0.5716	-50.919	-59.592	3.951
aFa	326	-16.20	-21.53	0.4099	-51.953	-60.418	3.671
aZa	253	-17.86	-20.56	0.5877	-50.704	-58.415	3.810
aWa	236	-15.97	-22.27	0.2293	-51.507	-60.909	5.090
aGa	279	-15.31	-18.22	0.5401	-48.269	-56.506	3.820

Table 4.10. Acoustic properties of consonants without dynamic range compression – duration (ms), peak and median level energy (dB), minimum to peak energy ratio, peak and median level energy after low-pass filtering (dB) and envelope variation (dB) between 20 and 200 Hz

	D (ms)	P (dB)	M (dB)	Minimum/ peak	P-LPF (dB)	M-LPF (dB)	EV (dB)
aPa	245	-19.82	-77.72	0.0007	-55.553	-111.570	20.863
aTa	238	-23.10	-78.92	0.0007	-54.498	-112.250	22.464
aKa	226	-23.61	-68.14	0.0007	-57.595	-103.610	22.532
aBa	207	-15.87	-55.58	0.0042	-48.813	-92.041	15.747
aDa	206	-21.08	-58.99	0.0008	-58.780	-94.958	19.139
aRa	219	-19.78	-30.79	0.0296	-54.544	-68.318	6.576
aLa	210	-18.23	-27.15	0.2674	-53.066	-65.600	4.776
aJa	235	-18.72	-32.30	0.0654	-53.424	-74.664	7.932
aMa	222	-19.74	-36.52	0.0794	-55.157	-75.530	6.806
aNa	225	-18.91	-34.72	0.1000	-55.707	-73.375	6.300
aSa	257	-20.92	-28.88	0.0628	-57.340	-70.085	4.920
aFa	315	-20.35	-42.99	0.0265	-54.039	-82.091	8.641
aZa	234	-23.08	-36.45	0.0758	-57.999	-76.002	6.920
aWa	169	-21.16	-46.72	0.0150	-54.963	-84.319	11.566
aGa	292	-18.38	-32.59	0.0838	-56.191	-70.786	5.143

Two analyses were developed (using Matlab) to determine possible candidates for a physical measure of amplitude envelope. For the first analysis, each consonant was low-pass filtered with a first-order Butterworth filter with a cutoff frequency of 20 Hz. By doing this, most of the temporal information was preserved while the spectral information was lost. From the envelope, the peak and median energy levels were determined by using an equation similar to 4.3, where the V_{RMS} is now calculated from the low-passed signal. In the second analysis, the variation of the envelope was determined. The isolated consonant was bandpass filtered between 20 and 200 Hz and full-wave rectified. From this output the RMS was calculated using 512-sample time windows with 75 % overlap (implemented with Hanning windows). The standard deviation from the RMS values (in dB) gives an indication of the variation in the envelope, as calculated by

$$EV = Stdev(\log(V_{RMS})), \tag{4.4}$$

where V_{RMS} is the RMS voltage of the filtered signal, calculated similar to equation 3.4.

4.3.2.2 Predictions of consonant confusions from acoustic analyses

The acoustic properties of consonants calculated in the previous section can be used to predict possible confusions. The acoustic properties were normalised using the Lobanov z-score transformation (Adank et al., 2004), similar to the analysis of the vowels. Three of the most important acoustic properties were chosen to calculate the Euclidean distances between consonants; the three acoustic properties were chosen according to the results found in Van Wieringen and Wouters (1999). The acoustic properties used for predicting confusions were envelope variation, ratio of minimum to peak energy and the duration of the consonant. The Euclidean distance measures between the consonants are shown in figures 4.29 and 4.30.

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	0.0	1.5	0.9	1.6	0.7	2.8	3.5	3.9	3.1	3.1	3.2	3.1	3.1	1.7	3.0
aTa	t		0.0	2.3	0.4	0.8	2.1	3.3	3.0	3.2	2.7	2.6	3.7	2.9	1.1	3.1
aKa	k			0.0	2.4	1.5	3.3	3.8	4.5	3.1	3.5	3.7	2.9	3.4	2.3	3.1
aBa	b				0.0	1.0	2.4	3.4	3.1	3.4	2.8	2.7	4.0	3.1	1.4	3.3
aDa	d					0.0	2.3	3.2	3.4	3.0	2.8	2.8	3.3	2.9	1.2	2.9
aMa	m						0.0	1.5	1.7	1.8	1.1	1.2	2.7	1.4	1.1	1.7
aNa	n							0.0	1.9	0.9	0.7	1.0	2.2	0.4	2.2	0.8
aRa	r								0.0	2.7	1.4	1.0	3.9	1.8	2.4	2.6
aLa	l									0.0	1.3	1.7	1.3	0.8	2.1	0.1
aJa	j										0.0	0.4	2.5	0.5	1.7	1.2
aSa	s											0.0	2.9	0.8	1.7	1.6
aZa	z												0.0	2.2	2.8	1.4
aFa	f													0.0	1.9	0.7
aWa	v														0.0	2.0
aGa	x															0.0

Figure 4.29. Three-dimensional Euclidean distance measures for envelope variation, minimum to peak energy ratio and duration of consonants with dynamic range compression

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	0.0	0.3	0.6	1.3	1.1	2.3	4.6	2.1	2.4	2.6	2.5	2.7	2.3	2.6	2.9
aTa	t		0.0	0.3	1.3	1.0	2.4	4.7	2.3	2.6	2.8	2.8	3.0	2.5	2.5	3.2
aKa	k			0.0	1.1	0.8	2.4	4.7	2.4	2.6	2.8	2.9	3.3	2.5	2.3	3.4
aBa	b				0.0	0.5	1.4	4.1	1.7	1.8	2.0	2.3	3.3	1.8	1.2	3.1
aDa	d					0.0	1.9	4.4	2.1	2.2	2.4	2.7	3.5	2.2	1.5	3.4
aMa	m						0.0	3.5	0.7	0.7	1.0	1.2	2.7	0.8	1.6	2.2
aNa	n							0.0	3.0	2.8	2.5	3.3	4.6	2.9	4.0	3.5
aRa	r								0.0	0.5	0.6	0.8	2.3	0.2	2.1	1.7
aLa	l									0.0	0.3	1.1	2.8	0.3	1.9	2.0
aJa	j										0.0	1.1	2.8	0.4	2.2	1.9
aSa	s											0.0	1.8	0.7	2.8	1.0
aZa	z												0.0	2.4	4.2	1.2
aFa	f													0.0	2.2	1.7
aWa	v														0.0	3.7
aGa	x															0.0

Figure 4.30. Three-dimensional Euclidean distance measures for envelope variation, minimum to peak energy ratio and duration of consonants without dynamic range compression

Following a similar procedure as for the vowels, these distance measures are translated into a prediction confusion matrix. The Euclidean distance matrices are normalised row by row and rounded to 0, 0.25, 0.5 and 0.75 (figures 4.31 and 4.32). Consonants are expected to be confused almost always at a distance of 0.75; for a distance of 0.5 confusions might be expected often; for a distance of 0.25, only a few confusions are expected and at a distance of 0, no confusions are expected. The values for the diagonal are calculated slightly differently. This distance measure gives an indication of how often the consonant will not be confused with other consonants. A distance measure of 1 indicates that no confusions are expected.

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	0.75	0.50	0.75	0.50	0.75	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
aTa	t		0.75	0.25	0.75	0.75	0.25	0.00	0.00	0.00	0.25	0.25	0.00	0.00	0.50	0.00
aKa	k			0.75	0.25	0.50	0.25	0.00	0.00	0.25	0.00	0.00	0.25	0.00	0.25	0.25
aBa	b				0.50	0.75	0.25	0.00	0.00	0.00	0.25	0.25	0.00	0.00	0.50	0.00
aDa	d					0.75	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
aMa	m						0.75	0.50	0.25	0.25	0.50	0.50	0.00	0.50	0.50	0.25
aNa	n							0.75	0.25	0.75	0.75	0.50	0.25	0.75	0.25	0.75
aRa	r								0.75	0.25	0.50	0.75	0.00	0.50	0.25	0.25
aLa	l									0.75	0.50	0.50	0.50	0.75	0.25	0.75
aJa	j										0.75	0.75	0.25	0.75	0.50	0.50
aSa	s											0.75	0.00	0.75	0.50	0.50
aZa	z												0.75	0.25	0.25	0.50
aFa	f													0.75	0.25	0.75
aWa	v														0.75	0.25
aGa	x															0.75

Figure 4.31. Predictions of confusions for consonants processed with dynamic range compression

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	0.50	0.75	0.75	0.50	0.75	0.50	0.00	0.50	0.25	0.25	0.25	0.25	0.25	0.25	0.25
aTa	t		0.50	0.75	0.50	0.75	0.25	0.00	0.50	0.25	0.25	0.25	0.25	0.25	0.25	0.25
aKa	k			0.50	0.75	0.75	0.25	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.50	0.25
aBa	b				0.50	0.75	0.50	0.00	0.50	0.50	0.50	0.25	0.00	0.50	0.50	0.25
aDa	d					0.50	0.50	0.00	0.50	0.50	0.25	0.25	0.00	0.25	0.50	0.00
aMa	m						0.75	0.00	0.75	0.75	0.50	0.50	0.00	0.75	0.50	0.25
aNa	n							1.00	0.25	0.25	0.25	0.25	0.00	0.25	0.00	0.25
aRa	r								0.75	0.75	0.75	0.50	0.00	0.75	0.25	0.25
aLa	l									0.75	0.75	0.50	0.00	0.75	0.25	0.25
aJa	j										0.75	0.50	0.00	0.75	0.00	0.25
aSa	s											0.75	0.25	0.75	0.00	0.50
aZa	z												0.75	0.25	0.00	0.50
aFa	f													0.75	0.25	0.25
aWa	v														0.75	0.00
aGa	x															0.75

Figure 4.32. Predictions of confusions for consonants processed without dynamic range compression

4.3.2.3 Results from experimental study on consonant confusions

The recognition of consonants improved with the number of experiments completed, similar to the results found for the recognition of vowels. It can be seen from figures 4.33 and 4.34 that after approximately three to four experiments, the percentage of consonants recognised correctly start to stabilise at specific values – the slope between experiments three and five is less than the slope between experiments one and three. The mean percentage scores for the pooled consonants recognised are 63 % and 52 % for processing before and after dynamic range compression respectively. The percentage recognised correctly is higher in cases where the dynamic range compression is not included, similar to the recognition of vowels. The confusion matrices for the consonants after dynamic range compression and before dynamic range compression are given in figures 4.35 and 4.36 respectively.

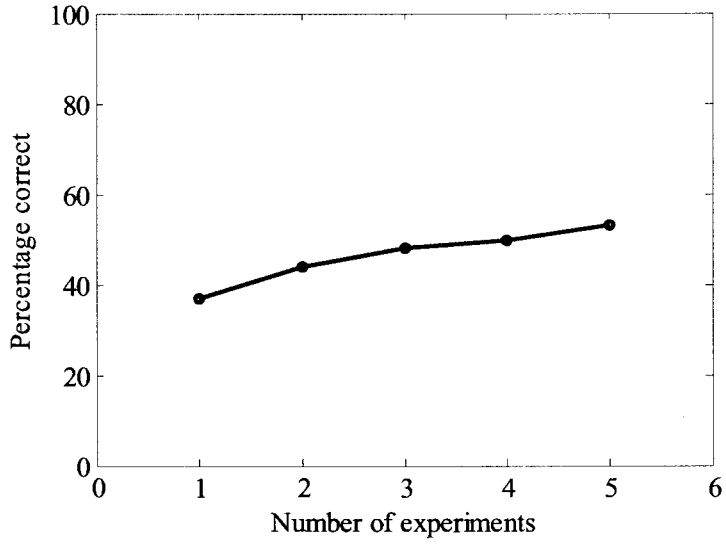

Figure 4.33. Learning curve for recognition of consonants with dynamic range compression

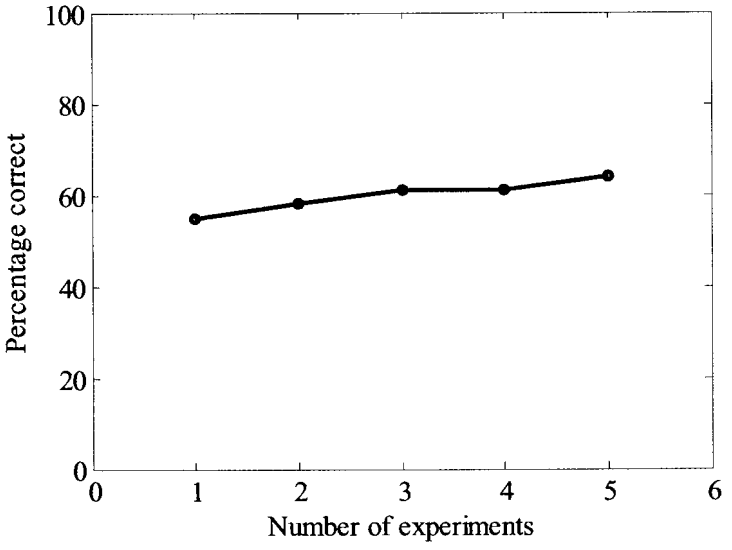

Figure 4.34. Learning curve for recognition of consonants without dynamic range compression

Stimulus		Response: aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	80	27	19	25	9	0	0	0	0	0	0	0	23	16	1
aTa	t	15	84	47	5	40	0	0	0	0	0	0	0	6	2	1
aKa	k	19	26	102	12	24	0	0	1	0	0	0	0	5	11	0
aBa	b	31	0	0	104	1	44	1	0	0	0	0	0	7	10	2
aDa	d	0	30	31	0	137	0	0	1	0	1	0	0	0	0	0
aMa	m	0	0	0	0	0	116	20	0	26	2	0	0	0	36	0
aNa	n	0	0	0	0	0	62	74	0	31	13	1	12	1	6	0
aRa	r	0	0	0	0	0	0	0	200	0	0	0	0	0	0	0
aLa	l	0	0	0	0	0	0	5	1	129	43	1	9	0	5	12
aJa	j	0	0	0	0	0	0	43	0	82	36	3	35	0	1	0
aSa	s	0	0	0	0	0	15	2	0	4	0	129	49	1	0	0
aZa	z	0	0	0	1	0	5	3	0	15	18	19	121	3	3	12
aFa	f	0	0	0	0	0	17	2	0	3	1	98	44	34	1	0
aWa	v	0	0	0	4	0	11	5	0	5	24	0	0	21	130	0
aGa	x	0	0	0	0	0	11	10	5	8	25	9	17	27	8	80

Figure 4.35. Confusion matrix obtained from experiments for the condition after dynamic range compression

		Response														
		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
Stimulus	aPa p	49	72	0	18	32	0	0	0	0	0	0	0	6	13	0
	aTa t	0	109	9	0	71	0	1	0	0	0	0	0	0	0	0
	aKa k	19	1	170	0	0	0	0	0	0	0	0	0	0	0	0
	aBa b	41	0	0	115	0	0	0	1	0	0	0	0	0	33	0
	aDa d	0	53	1	0	136	0	0	0	0	0	0	0	0	0	0
	aMa m	0	0	0	0	0	92	1	0	20	0	0	0	52	25	0
	aNa n	1	8	0	0	7	1	112	2	14	22	1	17	1	4	0
	aRa r	0	0	0	0	0	0	0	190	0	0	0	0	0	0	0
	aLa l	0	0	0	0	0	1	28	1	122	17	0	0	0	10	11
	aJa j	0	1	0	0	19	0	3	0	44	80	1	42	0	0	0
	aSa s	0	0	0	0	3	3	0	0	0	2	169	13	0	0	0
	aZa z	0	0	0	0	6	0	1	3	3	6	34	136	0	0	1
	aFa f	2	0	0	18	0	25	1	0	0	0	8	0	124	9	0
	aWa v	9	0	1	1	0	14	7	1	22	17	0	0	20	116	2
	aGa x	0	0	0	0	0	2	2	0	1	1	2	2	82	14	94

Figure 4.36. Confusion matrix obtained from experiments for the condition before dynamic range compression

The predictions made from the three-dimensional Euclidean distances give an indication of what confusions to expect. When the predictions and actual results are compared, most of the confusions are predicted reasonably well by the prediction confusion matrix. In figures 4.37 and 4.38, the probability of a correct response is plotted against the calculated Euclidean distance. As the Euclidean distance between two consonants increases, the probability of a correct response will also increase. The procedure to determine the probability of a correct response is similar to that followed for the vowels: firstly the confusion matrices from the experiments are normalised so that the sum of each row is 1. The values in the normalised matrices are used as the probabilities of an incorrect response for a specific stimulus. All the probabilities at a particular Euclidean distance are added together, this sum is multiplied with the number of times the particular Euclidean distance occurred and divided by the total number of Euclidean distances in the matrix. This is then subtracted from 1 to reach the probability of a correct response. These probabilities are then plotted against the corresponding Euclidean distance.

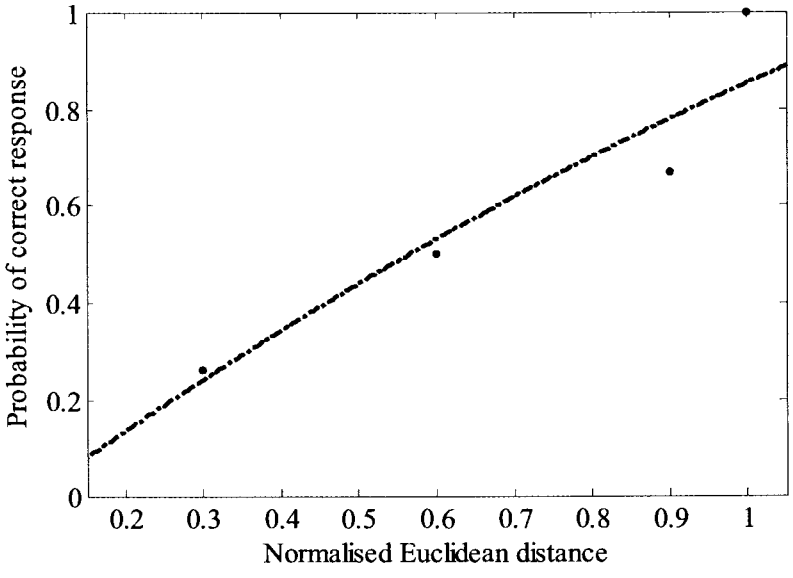

Figure 4.37. Probability of correct response to stimuli as a function of normalised Euclidean distance for consonants after dynamic range compression

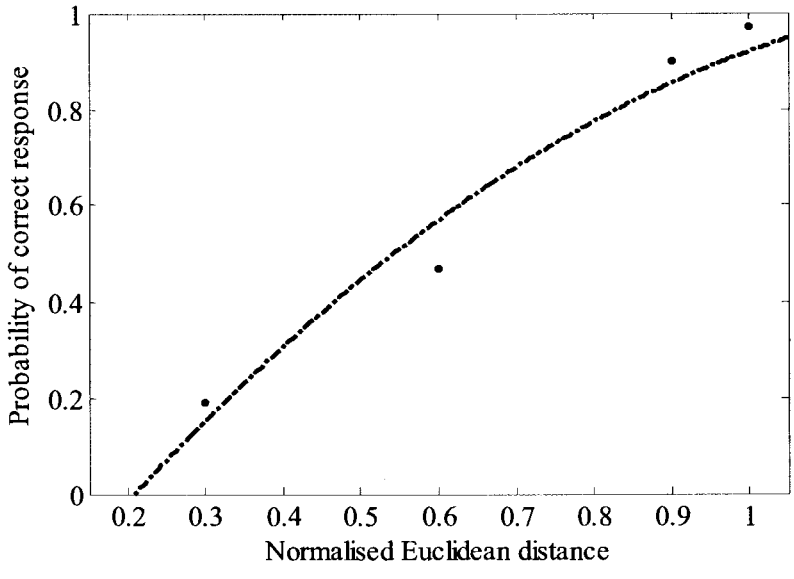

Figure 4.38. Probability of correct response to stimuli as a function of normalised Euclidean distance for consonants before dynamic range compression

Multidimensional scaling

An INDSCAL analysis (Carroll and Chang, 1970b) was performed on the consonant confusion matrices and the acoustic features obtained from the acoustic analysis were correlated with the results from the INDSCAL analysis. The INDSCAL analysis is a multidimensional scaling technique that examines the relationship between specific stimuli and their respective responses. By performing this analysis, it can be determined whether the auditory and articulatory features discussed earlier are reflected by the important perceptual dimensions as obtained by INDSCAL from the measured confusion matrices. The analysis was performed on the pooled matrices from all the subjects using the software package PRAAT (Boersma and Weenink, 2004). This was done for the consonants before and after dynamic range compression.

The output of an INDSCAL analysis is coordinates for each of the consonants for a fixed number of dimensions and the variance accounted for (VAF) for each dimension. The VAF is an indication of how much of the variance in one variable is accounted for by the variance in another. The coordinates within a specific dimension group the consonants according to the information transmitted. By plotting two dimensions against each other, the confusions of consonants can be interpreted by looking at the graphic representation. For example, all the plosives are expected to be grouped together and the nasals are expected to be grouped together.

For the consonant data to be subjected to INDSCAL analysis, there are a number of pre-processing steps that need to be performed. All the pre-processing was done in PRAAT; the program has the capability of performing INDSCAL analysis and all the pre-processing needed for the analysis. The input confusion matrices are normalised to fractions by dividing the responses by the total number of stimuli. These were then symmetrised using the algorithm suggested by Houtgast (Klein, Plomp and Pols, 1970; Van Wieringen and Wouters, 1999) and converted to dissimilarity matrices (symmetric tables representing dissimilarities between responses). For the actual INDSCAL analysis, the number of dimensions was increased until only small increases in VAF were observed. The number of dimensions that still contribute more than 0.04 % (Van Tassel et al., 1987) was three for

the consonants before and after dynamic range compression.

The coordinates of the dimensions for each consonant were correlated with the consonant classifications using a Pearson correlation. By examining the correlation coefficients of the dimension coordinates with the consonant classifications and speech production features (or acoustic features), a specific dimension can be assigned to a corresponding feature. For example, for the results shown in table 4.11, dimension 1 has the highest correlation with the burst of the consonants. From this correlation, it appears that dimension 1 from the INDSCAL analysis is represented by the presence or absence of burst in the consonants.

Consonants with dynamic range compression

For the first dimension, the presence or absence of burst yielded the highest correlation ($r = 0.92$). The second dimension correlated well with the feature affrication ($r = 0.75$). Lastly, the liquidity of consonants is the acoustic feature with the highest correlation with the third dimension ($r = 0.51$). The burst dimension separates the stops, /p, t, k, b, d/, from the other consonants. The fricative consonants, /s, z, f, x/, are separated from the other consonants by the second dimension, affrication. Lastly, /r, l/, are separated from the consonants without liquidity by the third dimension. These results assign the dimensions from the INDSCAL analysis to specific acoustic features, as summarised in table 4.11. The normalised dimension weights are determined during the INDSCAL analysis. This gives an indication of the percentage contribution that each dimension makes to the recognition of consonants. The results from the INDSCAL analysis for the acoustic features of consonants are given in table 4.12.

Table 4.11. Total VAF, normalised dimension weights, highest correlating speech production feature and correlation coefficient for consonants after dynamic range compression

	VAF R^2	Normalised dimension weights	Highest correlating speech production feature	r
1st Dimension	0.575	0.766	Burst	0.92
2nd Dimension	0.663	0.380	Affrication	0.75
3rd Dimension	0.707	0.322	Liquidity	0.51

Table 4.12. Total VAF, normalised dimension weights, highest correlating acoustic feature and correlation coefficient for consonants after dynamic range compression

	VAF R^2	Normalised dimension weights	Highest correlating acoustic feature	r
1st Dimension	0.575	0.766	Envelope variation	0.897
2nd Dimension	0.663	0.380	Peak energy level	0.361
3rd Dimension	0.707	0.322	Median energy level	0.237

For each of the acoustic cues, the correlation with the consonant classifications was determined to establish which signal property contributes primarily to a specific classification. From these correlations, it can also be determined which acoustic cues contribute most to the recognition of consonants. The acoustic cues that contribute most to the recognition of consonants were found to be the envelope variation and the ratio of minimum to peak energy in this study, similar to the results found in Faulkner and Rosen (1999). For the classification of burst, a correlation coefficient of $r = 0.957$ was obtained with envelope variation. For both affrication and liquidity the highest correlation was with minimum to peak energy ratio, $r = 0.450$ and $r = 0.389$ respectively. These were the acoustic cues with the highest correlations with the consonant classifications.

In figures 4.39 and 4.40, the consonants are plotted using the coordinates obtained from the INDSCAL analysis. For the condition of processing after dynamic range compression, two graphs are shown – dimension 1 (burst) vs. dimension 2 (affrication) and dimension 1 (burst) vs. dimension 3 (liquidity). The consonants that are confused regularly are clustered together in the graphs. In the graph for dimension 1 vs. dimension 2, the groups of consonants that are confused regularly are more clearly defined than in any of the other graphs. This flows from the fact that dimension 1 represents the feature with the highest VAF value, with dimension 2 and 3 contributing progressively less to the final VAF value. The INDSCAL analysis gives a valuable graphical representation of the confusions of the consonants. These confusions can be linked to the predicted confusions as discussed earlier.

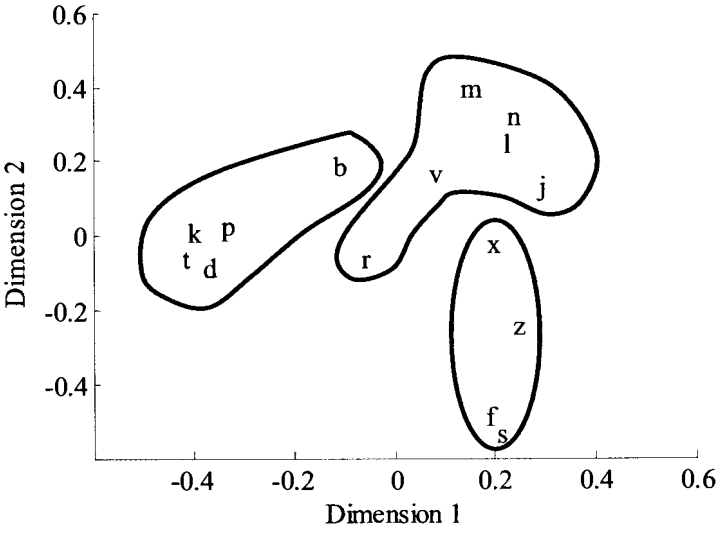

Figure 4.39. Dimension 1 (burst) vs. dimension 2 (affrication) for consonants after dynamic range compression. From the groupings shown, the plosives are separated from the fricatives and the nasals, glides and liquids

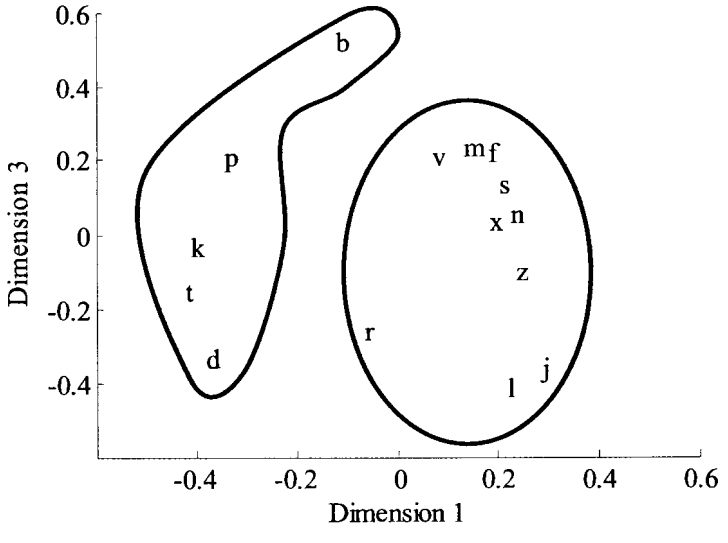

Figure 4.40. Dimension 1 (burst) vs. dimension 3 (liquidity) for consonants after dynamic range compression. Dimension 1 clearly divides the consonants according to the plosive feature

The results shown in figures 4.39 and 4.40 show the groupings of the consonants according to the coordinates calculated from the confusion matrices. The consonants that are confused regularly are grouped together.

Consonants without dynamic range compression

The highest correlation for the first dimension was the classification of burst, (r = 0.735), the same classification as for the consonants processed with dynamic range compression. The second dimension correlated well with the feature of manner (r = 0.522), which separates the nasals, fricatives and affricates. Lastly, the affrication of consonants correlated highly with the third dimension (r = 0.484). As mentioned above, the burst dimension separates the stops, /p, t, k, b, d/, from the other consonants. The fricative consonants, /s, z, f, x/, are separated from the other consonants by the third dimension for the consonants processed before dynamic range compression, as opposed to the second dimension separating these consonants after dynamic range compression. It is interesting to note that there are very similar features that separate the consonants for both processing before and after dynamic range compression. This should be expected as the processing is similar except for the inclusion of the dynamic range compression. The results for the consonants before dynamic range compression are summarised in table 4.13 (speech production features) and table 4.14 (acoustic features).

Table 4.13. Total VAF, normalised dimension weights, highest correlating speech production feature and correlation coefficient for consonants before dynamic range compression

	VAF R^2	Normalised dimension weights	Highest correlating speech production feature	r
1st Dimension	0.303	0.600	Burst	0.735
2nd Dimension	0.538	0.500	Manner	0.522
3rd Dimension	0.607	0.356	Affrication	0.484

Table 4.14. Total VAF, normalised dimension weights, highest correlating acoustic feature and correlation coefficient for consonants before dynamic range compression

	VAF R^2	Normalised dimension weights	Highest correlating acoustic feature	r
1st Dimension	0.303	0.600	Envelope variation	0.729
2nd Dimension	0.538	0.500	Minimum/peak ratio	0.504
3rd Dimension	0.607	0.356	Duration	0.465

As for the consonants after dynamic range compression, each of the acoustic cues was correlated with the consonant classifications to determine which signal property contributes primarily to a specific classification. From these correlations, it can also be determined which acoustic cues contribute mainly to the recognition of consonants. The acoustic cues that contribute most to the recognition of consonants before dynamic range compression are the envelope variation, the median energy level and the duration of the consonants. For the classification burst, a correlation coefficient of $r = 0.945$ was obtained with envelope variation, for manner the highest correlation was found with the median energy level after low-pass filtering, $r = 0.809$ and for affrication, the highest correlation was with the duration of the consonant, $r = 0.729$. These acoustic cues contribute primarily to the recognition of consonants. This compares well with the results found in Van Wieringen and Wouters (1999).

In the figures that follow, the consonants are plotted using the coordinates obtained from the INDSCAL analysis. For the condition of processing before dynamic range compression, two graphs are shown – dimension 1 (burst) vs. dimension 2 (manner) and dimension 1 (burst) vs. dimension 3 (affrication). The consonants that are confused regularly are clustered together in the graphs. As for the consonants after dynamic range compression, the graph for dimension 1 vs. dimension 2 groups the consonants that are confused regularly more clearly than any of the other graphs. This flows from the fact that dimension 1 represents the feature with the highest VAF value, with dimension 2 and 3 contributing progressively less to the final VAF value. In figures 4.41 and 4.42, the confusions of the consonants before dynamic range compression are displayed graphically using the coordinates obtained from the INDSCAL analysis.

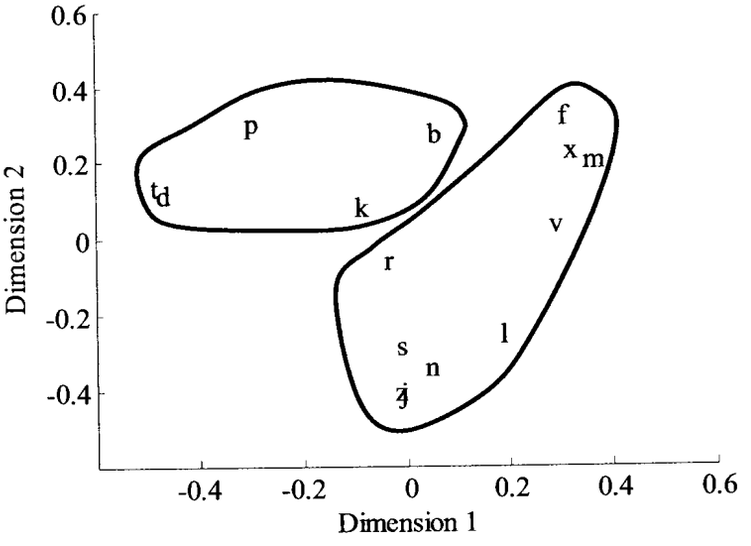

Figure 4.41. Dimension 1 (burst) vs. dimension 2 (manner) for consonants before dynamic range compression. Similar to results for consonants after dynamic range compression, dimension 1 separates the consonants according to the plosiveness in the speech segment

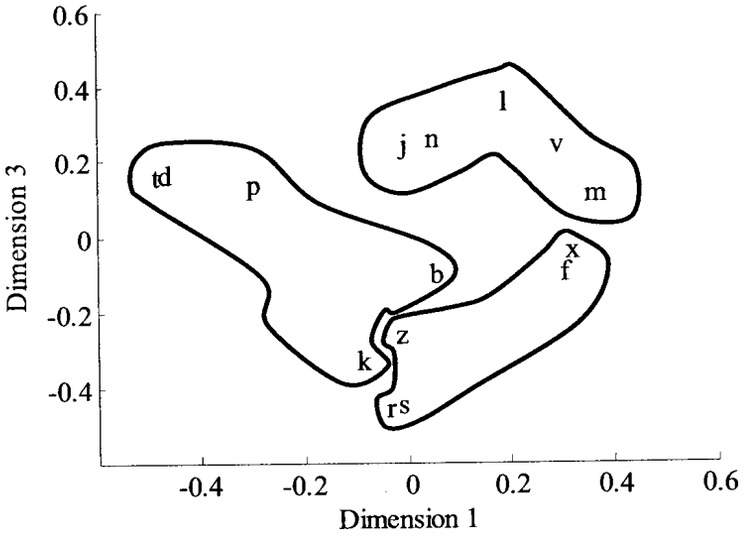

Figure 4.42. Dimension 1 (burst) vs. dimension 3 (affrication) for consonants before dynamic range compression. The plosives are separate from the other consonants; the fricatives are also separated from the other consonants, except for /r/

The point of the INDSCAL analysis was to show that the chosen cues for the recognition of consonants are indeed meaningful choices. The confusions between consonants can be explained by dividing the consonants according to these cues.

4.3.2.4 FITA analysis

For the consonants, the classifications that have been used for the FITA analysis were presented in table 4.8. These classifications were used together with the confusion matrices obtained from the acoustic experiments to perform a FITA analysis. The results for the FITA analysis are given in tables 4.15 to 4.17.

Table 4.15. Results of FITA analysis for original consonants – cochlear implantees

	Percentage information transmitted
Burst	70
Voicing	54
Manner	76
Place	50
Nasality	54
Liquid	51
Affrication	82

Table 4.16. Results of FITA analysis for processed consonants after dynamic range compression

	Percentage information transmitted
Burst	72
Voicing	30
Manner	60
Place	22
Nasality	31
Liquid	44
Affrication	45

Table 4.17. Results of FITA analysis for processed consonants before dynamic range compression

	Percentage information transmitted
Burst	74
Voicing	29
Manner	62
Place	40
Nasality	25
Liquid	51
Affrication	50

FITA analyses were also performed to determine the percentage information transmitted for the acoustic properties, as listed in tables 4.9 and 4.10. According to the results in tables 4.18 and 4.19, the acoustic property that is transmitted most effectively is the envelope variation (72 % for both before and after dynamic range compression). This confirms the result found with the INDSCAL analysis; the acoustic property that corresponds to dimension 1 is also the envelope variation.

Table 4.18. Results of FITA analysis for processed consonants with dynamic range compression; classifications are done according to acoustic properties, refer to table 4.9

	Percentage information transmitted
D (ms)	33
P (dB)	36
M (dB)	39
Minimum/peak	34
P-LPF (dB)	18
M-LPF (dB)	31
EV (dB)	72

Table 4.19. Results of FITA analysis for processed consonants without dynamic range compression, classifications are done according to acoustic properties, refer to table 4.10

	Percentage information transmitted
D (ms)	49
P (dB)	28
M (dB)	45
Minimum/peak	41
P-LPF (dB)	20
M-LPF (dB)	34
EV (dB)	72

The results from the FITA analysis confirm the results from the multidimensional scaling. The four classifications of consonants with the highest information transmitted were burst, manner, affrication and liquidity. These were also the four dimensions that were associated with the dimensions for the multidimensional scaling. The data suggest that these classifications are transmitted most effectively in the recognition of consonants processed through the acoustic model.

In figures 4.43 and 4.44 the percentage information transmitted for consonants is shown. The average and standard deviation of FITA scores are determined for all the listeners. Error bars[3] represent one standard deviation. Figure 4.43 reflects the transmitted information for the consonants processed with dynamic range compression and figure 4.44 reflects the consonants processed without dynamic range compression.

[3]An error bar on the graphs has a length equal to two standard deviations and is centred at the average of the relevant value

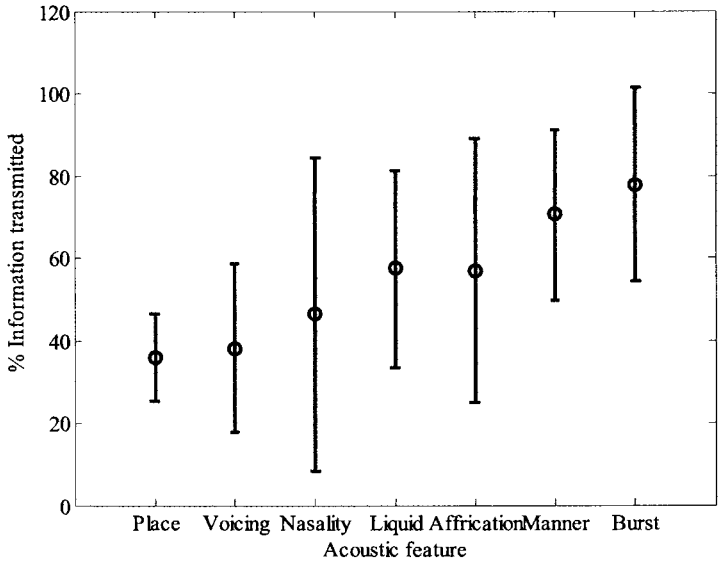

Figure 4.43. Indication of information transmitted to normal-hearing listeners for consonants processed with dynamic range compression. The percentage information transmitted is shown for seven acoustic features; one error bar represents one standard deviation

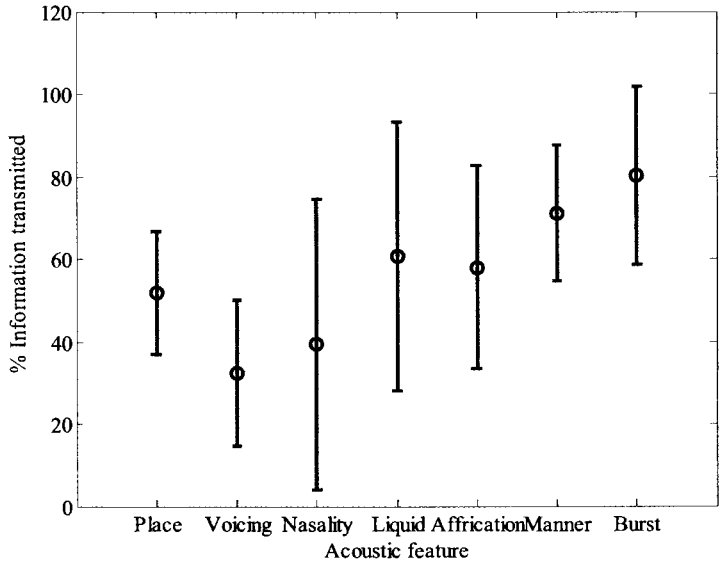

Figure 4.44. Indication of information transmitted to normal-hearing listeners for consonants processed without dynamic range compression. The percentage information transmitted is shown for seven acoustic features

When comparing the results from the FITA analysis for the confusions obtained when cochlear implant users listened to the original consonants with the processed consonants that were recognised by normal-hearing listeners, the three classifications transmitting most information was affrication, manner and burst. Even though the order of significance differs for the processed consonants, the information transmitted most effectively for the recognition of consonants is the same for the processed consonants and original consonants.

As for the vowels, a linear fit was performed on the results from the FITA analysis for cochlear implantees and for FITA analysis performed on the normal-hearing listeners' results. The curve-fitting results are shown in figure 4.45. The error bars represent seven acoustic features, similar to figures 4.43 and 4.44. A t-test was performed between the results obtained before and after dynamic range compression; each normal-hearing listener's results were analysed through FITA and used in the t-test. From the results of the t-test, it appears that five of the seven features from the FITA results obtained before and after dynamic range compression belong to the same probability density function with a significance level of 5 %. They also belong to the same probability density function as the

FITA results for cochlear implant users. The two features that do not belong to the same probability density function are voicing and nasality. For the acoustic simulations, the information transmitted through the nasality and voicing of the consonant is poor compared to the cochlear implant.

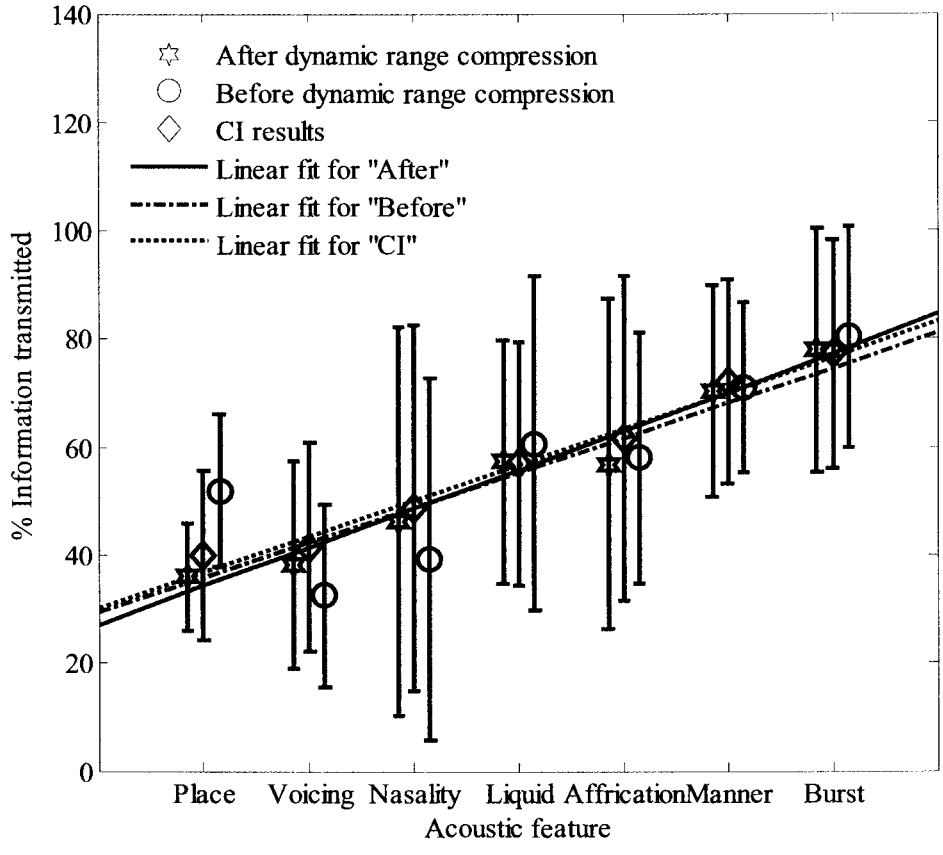

Figure 4.45. Linear fit for FITA results of consonants using the acoustic simulation with dynamic range compression (r = 0.680), acoustic simulation without dynamic range compression (r = 0.614) and cochlear implant users (r = 0.944)

With analysis of the output of the acoustic model, specific predictions were made as to which speech segments would be confused (for either vowels or consonants). These predictions appear to be good enough to gain an understanding of what underlies speech recognition of vowels and consonants in electric hearing.

The results found with the acoustic simulations were compared with the results found for electric hearing. The results from the acoustic model with dynamic range compression compared well with those found for electric hearing. From these results, we know that the developed model is a good acoustic simulation of electric hearing.

The developed model can now be applied to various scenarios in order to investigate and predict speech recognition under specific conditions. In this study, speech recognition in noise was investigated further and the results are discussed next.

4.3.3 Experiments in noise

The material up to now suggests that the researchers understand which cues are needed for vowel and consonant recognition and that the acoustic model is effective in predicting measurements with cochlear implants. The model can now be used to investigate the effect of noise on electric hearing using normal-hearing persons listening to the output of the model. The relationship between the recognition of vowels and consonants and the masking of the cues used for recognition can also be determined.

The approach followed for the experiments in noise is similar to that for the experiments on the consonants and vowels, while the procedure followed for the analysis is the same as for the consonants. Confusion matrices were obtained from the experiments with normal-hearing persons and analysed with multidimensional scaling to determine which acoustic features are transmitted amid noise. This was done for both the vowels and consonants, for the conditions of speech-like noise (as standardised by CCITT Recommendation 227) and multi-talker babble. The experimental set-up for the experiments is described in detail in section 3.4.3.

Figures 4.46 to 4.57 show the confusion matrices obtained for the experiments in noise, in the following order:

- vowels for 0 dB SNR multi-talker[4] babble and speech-like noise,
- vowels for 20 dB SNR multi-talker babble and speech-like noise,
- vowels for 40 dB SNR multi-talker babble and speech-like noise,
- consonants for 0 dB SNR multi-talker babble and speech-like noise,
- consonants for 20 dB SNR multi-talker babble and speech-like noise, and
- consonants for 40 dB SNR multi-talker babble and speech-like noise.

The multi-talker has the same spectral characteristics as normal speech; the speech-like noise is white Gaussian noise filtered to have the same spectral content as normal speech.

For all the confusion matrices, the stimulus is on the vertical axis and the response on the horizontal axis.

[4]Multi-talker babble refers to babble originating from various speakers talking simultaneously

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	35	4	7	5	2	7	1	4	6	14	10	5
pIEt	i	9	14	4	11	5	5	6	8	10	14	7	7
pOEt	u	14	6	2	10	8	12	5	9	9	9	8	8
pAd	a	26	5	1	17	4	8	3	6	4	11	12	3
pEt	ɛ	11	10	4	14	3	9	6	13	7	9	8	6
pOt	ɔ	17	1	6	7	5	12	3	11	6	10	16	6
pIt	ə	7	10	5	15	10	9	9	16	13	2	3	1
pAt	æ	23	11	4	6	4	8	1	6	2	11	16	8
pUt	œ	23	6	6	8	5	12	5	4	6	9	11	5
pEEt	e:	12	13	5	1	4	4	2	3	5	19	8	24
pêt	ɛ:	32	4	5	8	3	7	0	2	6	9	15	9
pUUt	y	10	5	4	5	9	9	6	8	4	15	11	14

Figure 4.46. Confusion matrix for vowels at 0 dB SNR – multi-talker babble

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	27	4	0	7	7	18	1	8	6	6	13	3
pIEt	i	14	10	3	9	12	9	5	13	5	7	9	4
pOEt	u	11	8	5	7	11	12	5	12	10	8	6	5
pAd	a	22	5	6	8	9	13	3	6	5	8	13	2
pEt	ɛ	9	5	3	11	10	15	14	13	4	4	9	3
pOt	ɔ	19	6	3	6	17	16	2	9	7	6	6	3
pIt	ə	7	5	4	23	7	16	11	14	4	3	4	2
pAt	æ	26	3	2	10	5	18	4	11	2	6	10	3
pUt	œ	18	8	2	9	10	13	9	10	4	9	5	3
pEEt	e:	13	21	2	7	10	14	2	7	2	14	5	3
pêt	ɛ:	24	5	5	4	6	11	2	5	6	13	12	7
pUUt	y	10	22	4	7	9	12	10	6	4	5	8	3

Figure 4.47. Confusion matrix for vowels at 0 dB SNR – speech-like noise

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	99	0	0	1	0	0	0	0	0	0	0	0
pIEt	i	0	11	0	8	2	4	51	3	19	0	0	2
pOEt	u	0	0	11	29	1	5	26	17	11	0	0	0
pAd	a	0	0	10	59	1	17	6	1	5	1	0	0
pEt	ɛ	0	5	1	13	7	9	44	7	11	0	0	3
pOt	ɔ	0	4	19	4	1	56	13	0	3	0	0	0
pIt	ə	0	0	1	4	2	9	59	0	25	0	0	0
pAt	æ	0	4	0	13	4	1	8	63	7	0	0	0
pUt	œ	0	12	3	1	4	8	44	0	25	0	2	1
pEEt	e:	0	3	0	0	1	1	0	1	1	83	0	10
pêt	ɛ:	14	0	0	0	1	0	0	0	0	3	64	18
pUUt	y	0	65	0	8	0	2	13	1	11	0	0	0

Figure 4.48. Confusion matrix for vowels at 20 dB SNR – multi-talker babble

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	71	2	1	8	1	3	0	1	3	4	4	2
pIEt	i	0	16	5	14	8	5	25	1	6	3	0	17
pOEt	u	0	5	7	16	4	32	9	5	20	0	1	1
pAd	a	2	5	1	71	3	3	0	5	3	4	2	1
pEt	ɛ	1	10	3	12	15	4	26	5	15	0	2	7
pOt	ɔ	5	6	10	5	12	29	7	6	10	2	0	8
pIt	ə	0	4	7	0	13	8	46	0	22	0	0	0
pAt	æ	0	2	2	43	7	2	1	28	4	6	2	3
pUt	œ	5	4	1	8	4	16	10	5	43	2	0	2
pEEt	e:	2	1	1	1	0	1	0	0	1	81	12	0
pêt	ɛ:	6	1	0	2	2	0	0	0	0	5	83	1
pUUt	y	0	24	13	11	0	0	2	10	4	4	0	32

Figure 4.49. Confusion matrix for vowels at 20 dB SNR – speech-like noise

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	94	0	2	1	0	3	0	0	0	0	0	0
pIEt	i	0	77	1	0	0	0	10	0	7	0	0	5
pOEt	u	1	0	1	20	5	12	39	9	13	0	0	0
pAd	a	0	0	0	100	0	0	0	0	0	0	0	0
pEt	ɛ	0	30	1	2	18	5	26	0	13	0	1	4
pOt	ɔ	0	0	9	25	6	57	2	0	0	1	0	0
pIt	ə	0	0	0	0	1	12	74	1	12	0	0	0
pAt	æ	0	0	0	22	0	0	1	77	0	0	0	0
pUt	œ	0	0	3	2	1	34	32	1	27	0	0	0
pEEt	e:	1	0	0	0	0	0	0	0	0	99	0	0
pêt	ɛ:	3	0	0	0	0	0	0	0	0	0	80	17
pUUt	y	0	76	0	5	0	0	9	0	7	0	0	3

Figure 4.50. Confusion matrix for vowels at 40 dB SNR – multi-talker babble

		pAAt	pIEt	pOEt	pAd	pEt	pOt	pIt	pAt	pUt	pEEt	pêt	pUUt
		ɑ:	i	u	a	ɛ	ɔ	ə	æ	œ	e:	ɛ:	y
pAAt	ɑ:	90	0	0	0	1	0	0	0	0	6	3	0
pIEt	i	0	30	8	2	5	1	4	0	10	0	0	40
pOEt	u	0	4	38	26	1	25	1	4	1	0	0	0
pAd	a	0	0	2	84	1	6	5	0	2	0	0	0
pEt	ɛ	0	10	6	3	10	7	10	3	46	0	0	5
pOt	ɔ	0	4	9	7	7	58	2	2	6	0	0	5
pIt	ə	0	3	3	3	6	17	46	3	19	0	0	0
pAt	æ	1	1	0	33	0	7	0	57	0	0	0	1
pUt	œ	1	4	0	6	10	16	23	2	35	1	0	2
pEEt	e:	4	0	0	0	0	1	0	0	0	67	28	0
pêt	ɛ:	18	0	0	0	0	0	0	0	0	0	82	0
pUUt	y	0	41	2	1	1	2	0	1	3	0	0	49

Figure 4.51. Confusion matrix for vowels at 40 dB SNR – speech-like noise

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	2	3	0	5	4	7	10	10	15	10	11	3	3	8	9
aTa	t	1	2	5	4	10	12	16	11	9	11	5	1	2	7	4
aKa	k	1	4	3	4	6	10	14	8	13	11	1	6	9	6	4
aBa	b	0	3	1	7	6	5	22	5	19	14	4	5	2	5	2
aDa	d	0	4	1	4	16	3	20	12	8	8	7	5	2	5	5
aMa	m	0	0	1	4	5	7	14	12	20	7	7	3	7	8	5
aNa	n	2	0	2	6	8	3	14	7	18	10	7	7	3	5	8
aRa	r	4	1	1	7	5	6	10	15	20	11	5	2	2	5	6
aLa	l	0	2	1	7	8	15	12	7	26	4	5	5	1	3	4
aJa	j	0	0	3	4	9	5	17	9	10	8	3	4	4	15	9
aSa	s	2	1	3	8	0	8	11	7	15	16	4	2	6	7	10
aZa	z	2	3	1	7	8	12	11	8	11	13	9	2	1	5	7
aFa	f	1	1	1	3	8	5	21	20	11	7	5	4	4	5	4
aWa	v	3	7	1	5	5	10	11	7	9	15	7	1	4	5	10
aGa	x	1	3	1	12	8	7	12	7	10	8	5	6	6	8	6

Figure 4.52. Confusion matrix for consonants at 0 dB SNR – multi-talker babble

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	2	1	4	0	2	10	10	8	8	10	7	11	6	14	7
aTa	t	1	3	4	3	3	11	12	11	3	14	7	6	3	15	4
aKa	k	3	1	3	2	7	5	21	5	8	8	6	5	4	18	4
aBa	b	1	3	1	3	6	7	24	8	5	3	4	16	4	14	1
aDa	d	2	2	4	5	5	10	13	3	4	5	11	14	2	14	6
aMa	m	1	5	4	2	1	5	14	8	11	13	7	5	3	16	5
aNa	n	3	3	0	3	3	4	16	3	9	9	5	9	3	15	15
aRa	r	4	3	3	2	4	10	13	9	5	5	11	8	1	18	4
aLa	l	1	2	5	1	6	11	21	9	3	4	10	8	3	12	4
aJa	j	3	5	5	1	5	5	9	10	4	5	11	13	3	17	4
aSa	s	2	5	0	2	2	7	10	10	5	12	8	9	3	19	6
aZa	z	2	3	4	5	4	5	6	7	5	15	5	8	7	17	7
aFa	f	0	3	2	3	7	6	7	13	5	10	2	6	3	29	4
aWa	v	2	4	3	7	4	5	11	10	8	11	6	8	4	17	0
aGa	x	1	0	1	1	7	7	10	4	7	6	11	8	5	16	16

Figure 4.53. Confusion matrix for consonants at 0 dB SNR – speech-like noise

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	20	17	23	10	11	3	2	0	4	1	0	1	5	3	0
aTa	t	21	38	12	6	19	2	1	0	0	1	0	0	0	0	0
aKa	k	1	27	50	0	21	0	0	0	1	0	0	0	0	0	0
aBa	b	21	2	0	24	9	14	11	0	9	4	0	2	0	4	0
aDa	d	0	25	0	2	73	0	0	0	0	0	0	0	0	0	0
aMa	m	1	0	0	1	0	48	10	2	12	7	0	0	0	19	0
aNa	n	0	0	0	0	1	11	47	2	9	7	7	10	1	5	0
aRa	r	0	0	0	0	1	0	0	99	0	0	0	0	0	0	0
aLa	l	0	0	0	0	0	2	10	4	29	14	3	13	0	6	19
aJa	j	1	1	0	1	14	1	18	0	3	3	26	31	0	0	1
aSa	s	0	0	0	0	0	6	2	0	0	0	38	22	7	1	24
aZa	z	0	0	0	0	0	5	4	1	10	8	23	29	2	13	5
aFa	f	0	0	0	0	1	3	2	0	1	1	17	11	32	13	19
aWa	v	0	3	1	1	6	1	7	1	2	12	0	1	4	55	6
aGa	x	0	1	0	1	0	8	3	0	4	4	4	14	18	0	43

Figure 4.54. Confusion matrix for consonants at 20 dB SNR – multi-talker babble

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	4	1	0	50	2	17	0	0	0	0	1	0	19	6	0
aTa	t	3	32	11	24	19	5	1	0	0	0	0	0	0	5	0
aKa	k	0	35	40	0	21	0	0	1	0	0	0	0	0	3	0
aBa	b	2	0	0	54	5	20	3	0	3	1	0	1	0	11	0
aDa	d	0	23	3	2	67	0	4	0	0	0	0	0	0	1	0
aMa	m	1	0	1	0	1	30	37	5	8	6	0	1	3	5	2
aNa	n	0	0	0	1	1	26	40	1	25	2	0	1	0	3	0
aRa	r	0	0	0	0	1	0	0	99	0	0	0	0	0	0	0
aLa	l	0	0	1	0	2	1	18	0	66	9	0	1	0	2	0
aJa	j	0	0	1	0	5	0	17	0	10	9	24	30	2	2	0
aSa	s	0	0	0	1	2	16	3	0	2	2	35	18	13	2	6
aZa	z	0	0	0	0	0	0	1	0	2	0	25	34	4	26	8
aFa	f	0	0	0	0	0	0	1	0	0	1	17	13	42	3	23
aWa	v	1	0	2	1	13	0	3	0	3	11	1	5	3	43	14
aGa	x	0	0	0	1	2	11	1	1	5	9	9	3	27	2	29

Figure 4.55. Confusion matrix for consonants at 20 dB SNR – speech-like noise

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	66	0	4	18	0	1	0	0	0	0	0	0	8	3	0
aTa	t	34	16	30	13	6	1	0	0	0	0	0	0	0	0	0
aKa	k	5	2	82	4	7	0	0	0	0	0	0	0	0	0	0
aBa	b	37	1	1	47	2	9	2	0	0	0	0	0	0	1	0
aDa	d	2	16	28	0	54	0	0	0	0	0	0	0	0	0	0
aMa	m	0	0	0	0	0	54	12	0	8	4	2	1	2	17	0
aNa	n	2	9	0	0	9	8	45	0	5	5	5	7	0	5	0
aRa	r	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
aLa	l	0	0	0	0	0	3	15	0	34	30	0	3	0	7	8
aJa	j	0	0	0	0	14	0	16	0	17	6	14	32	0	1	0
aSa	s	0	0	0	0	0	1	9	0	2	0	60	25	0	2	1
aZa	z	0	0	0	0	0	0	15	0	21	6	17	35	0	2	4
aFa	f	0	0	0	0	0	4	1	0	0	1	34	16	37	6	1
aWa	v	0	0	0	0	1	13	12	0	3	14	1	1	3	51	1
aGa	x	0	0	0	0	0	7	7	0	1	6	9	10	6	11	43

Figure 4.56. Confusion matrix for consonants at 40 dB SNR – multi-talker babble

		aPa	aTa	aKa	aBa	aDa	aMa	aNa	aRa	aLa	aJa	aSa	aZa	aFa	aWa	aGa
		p	t	k	b	d	m	n	r	l	j	s	z	f	v	x
aPa	p	13	11	2	49	12	0	0	0	0	0	0	0	2	10	1
aTa	t	0	41	32	0	23	0	0	0	0	0	0	0	3	1	0
aKa	k	0	37	43	1	10	0	0	0	0	0	0	0	4	5	0
aBa	b	4	0	1	75	3	2	0	0	2	0	0	0	0	13	0
aDa	d	0	11	12	0	77	0	0	0	0	0	0	0	0	0	0
aMa	m	0	0	0	0	0	71	5	1	15	7	0	0	0	1	0
aNa	n	0	0	6	1	2	8	62	0	14	6	0	0	0	1	0
aRa	r	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
aLa	l	0	0	0	0	0	2	11	2	70	14	0	0	0	0	1
aJa	j	0	0	1	0	0	1	52	0	19	25	0	1	0	0	1
aSa	s	0	0	0	0	0	4	11	0	0	0	52	30	2	1	0
aZa	z	0	0	0	0	1	0	6	0	12	3	14	63	0	0	1
aFa	f	0	0	0	0	0	0	0	0	1	0	53	20	20	4	2
aWa	v	0	0	0	1	1	0	0	0	11	6	0	0	1	70	10
aGa	x	0	0	0	0	0	0	0	0	1	7	33	9	11	11	28

Figure 4.57. Confusion matrix for consonants at 40 dB SNR – speech-like noise

The results from the confusion matrices are summarised in figures 4.58 to 4.61. In these figures, the FITA scores for a number of features are plotted as a function of SNR. It is clear that as the noise level increases, the information transmitted becomes less. This is expected because the noise masks the information available to the listener; with increased levels of noise, the available features become less prominent. At 0 dB, almost no information is transmitted to the listener.

A repeated-measures analysis of variance demonstrated a significant effect of noise (at levels of 0 dB, 20 dB, 40 dB and quiet) on the information transmitted by the features. The analysis was performed for vowels in multi-talker babble ($F = 37.23$, $p < 0.0001$), vowels in speech-like noise ($F = 31.91$, $p < 0.0001$), consonants in multi-talker babble ($F = 58.05$, $p < 0.0001$) and consonants in speech-like noise ($F = 50.21$, $p < 0.0001$).

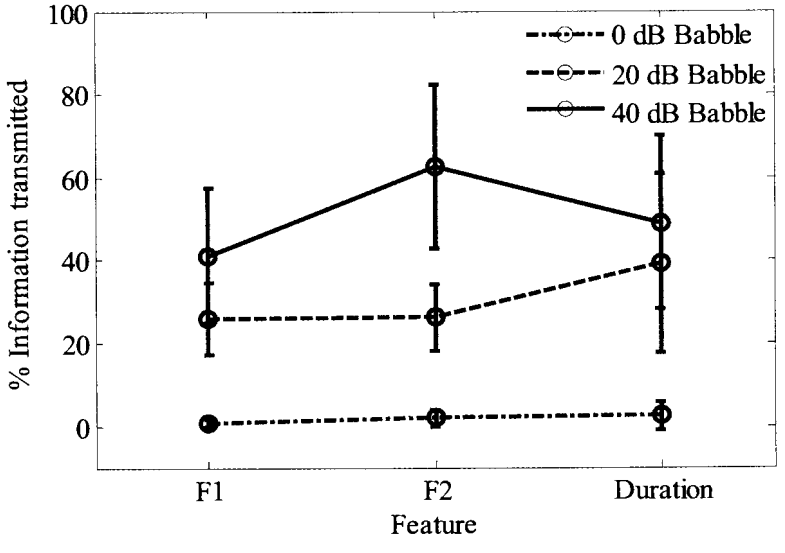

Figure 4.58. Average FITA scores of vowels for experiments in multi-talker babble. Scores were calculated for each listener separately. One error bar represents one standard deviation in all listeners

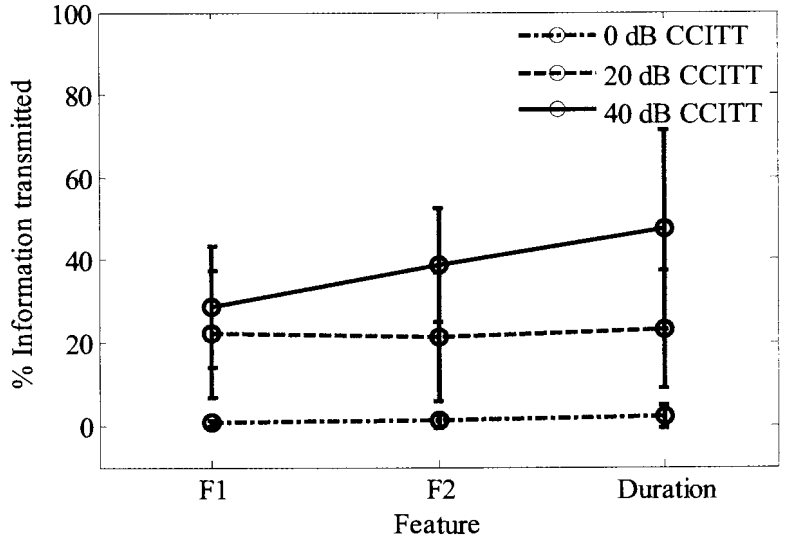

Figure 4.59. Average FITA scores of vowels for experiments in speech-like noise (CCITT Recommendation 227). Scores were calculated for each listener separately. One error bar represents one standard deviation in all listeners

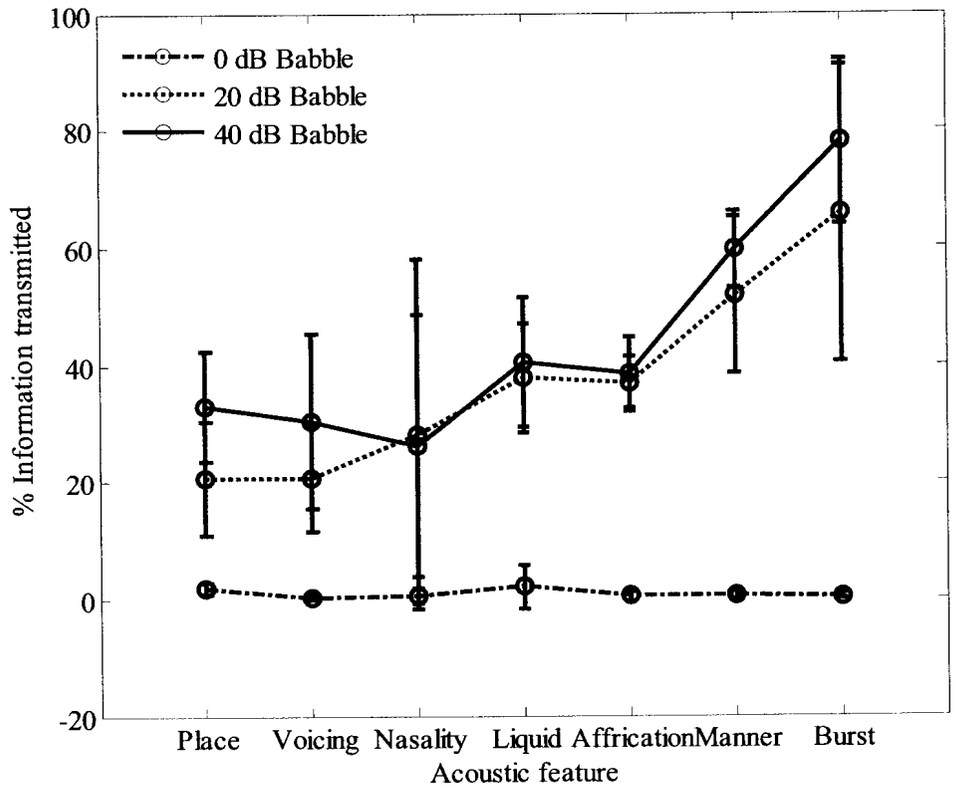

Figure 4.60. Average FITA scores of consonants for experiments in multi-talker babble. One error bar represents one standard deviation in all listeners

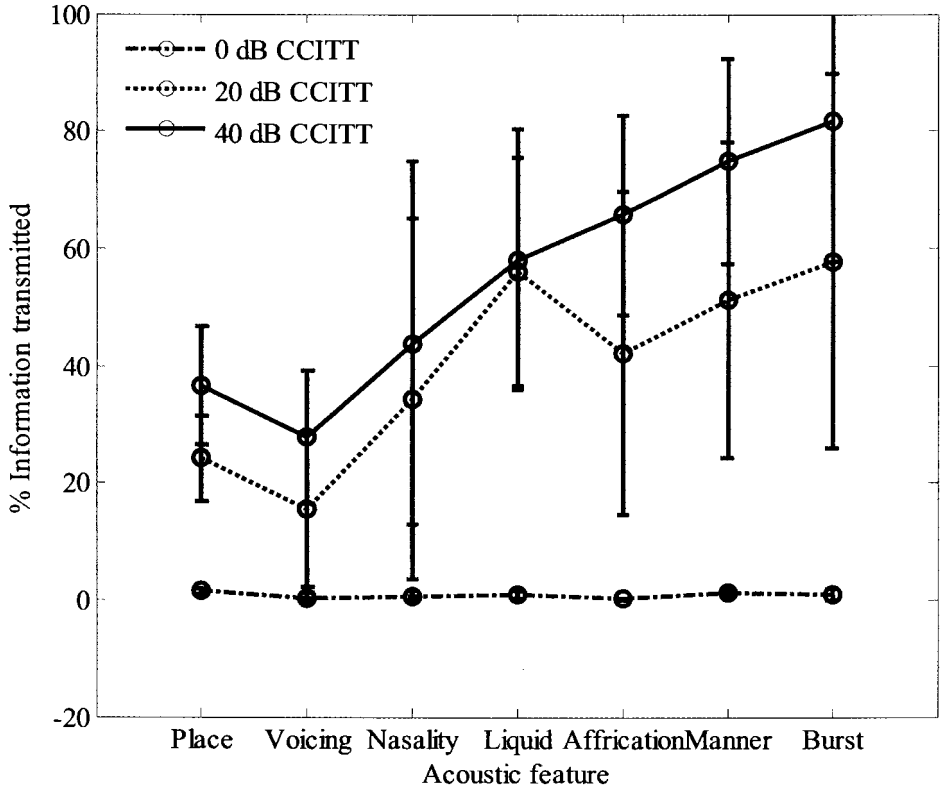

Figure 4.61. Average FITA scores of consonants for experiments in speech-like noise (CCITT Recommendation 227); one error bar represents one standard deviation in all listeners

The mean percentage scores of vowels and consonants recognised compare well with those reported in literature (Dubno et al., 2005; Fu et al., 1998; Killion et al., 2004; Nie et al., 2005), ranging from near chance to average levels. Results for this study are shown in figure 4.62 for all noise conditions. The SNRs were determined as described in section 3.4.3. It is not an easy task to determine the acoustic features that contribute to the recognition of vowels and consonants in noise by inspecting the confusion matrices. The formant frequencies are masked by the noise, making it difficult to determine these frequencies. Because of the difficulty in determining the formant frequencies, a vowel space cannot be determined for noisy conditions in the way it was for quiet conditions. Multidimensional scaling was performed on the confusion matrices and correlated with results obtained from acoustic analyses of the degraded speech segments to determine which acoustic cues are transmitted most effectively.

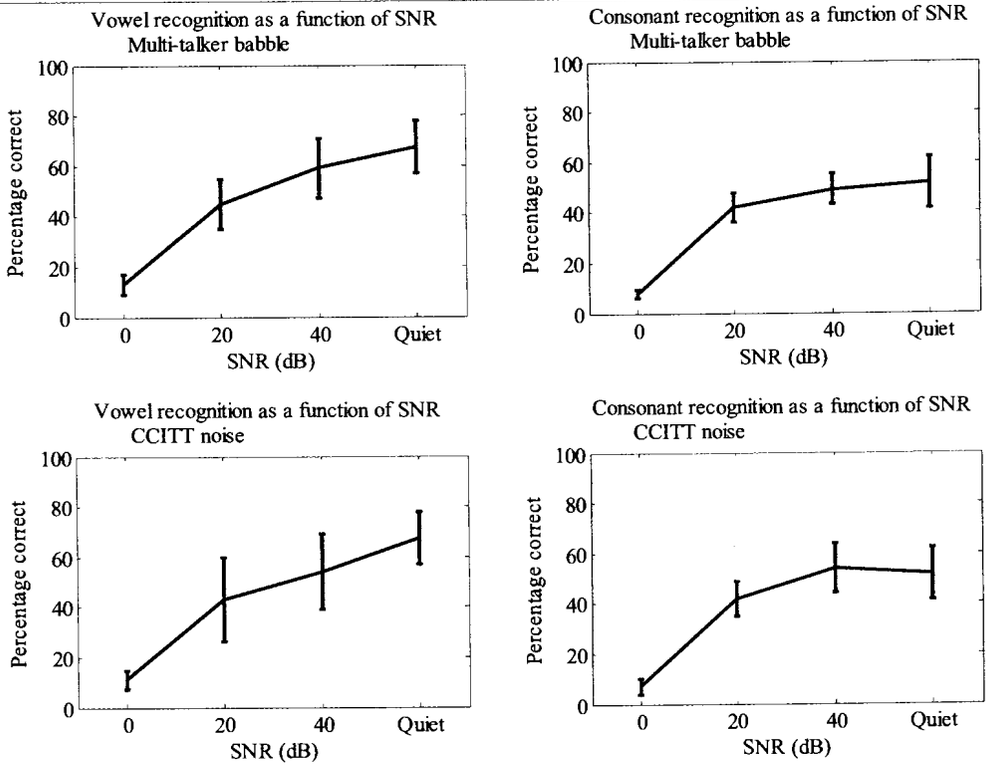

Figure 4.62. Results for vowel and consonant recognition in noise – multi-talker babble and speech-like noise (CCITT Recommendation 227). One error bar represents one standard deviation

The acoustic features that are used for the analysis of the vowels in noise are slightly different from those used for the vowels in quiet. For quiet, the vowel space obtained from the values of F_1, F_2 and duration were used to determine the acoustic features contributing to the recognition of vowels. Identification of F_1 and F_2 becomes more difficult with the addition of noise (Leek and Summers, 1996), owing to a reduction in spectral contrast, which will decrease the probability of identifying F_1 or F_2. The speech-like and babble noise has the same formant structure as the vowels, causing the identification of the formant frequencies to become more difficult.

It is reported in the literature that spectral contrast of processed vowels decreases significantly, even more so in the presence of noise. LPC analysis (figure 4.63) confirmed the decrease in spectral contrast of the processed vowels with decreasing SNR (Alcantara and Moore, 1995; Loizou and Poroy, 2001; Summers and Leek, 1994; ter Keurs, Festen and Plomp, 1993a; ter Keurs, Festen and Plomp, 1992). Limited spectral contrast will

reduce the probability of a person identifying the formants used in vowel recognition.

The spectral contrast[5] of F_2 for the example of figure 4.63 was calculated using the peak of F_2 and the following valley. The spectral contrast for the original signal is 31 dB, for the processed signal in quiet it is 14 dB, for the processed signal in 40 dB noise it is 13.5 dB and for the processed signal in 20 dB noise, it is 4 dB. A steady reduction in spectral contrast is obvious.

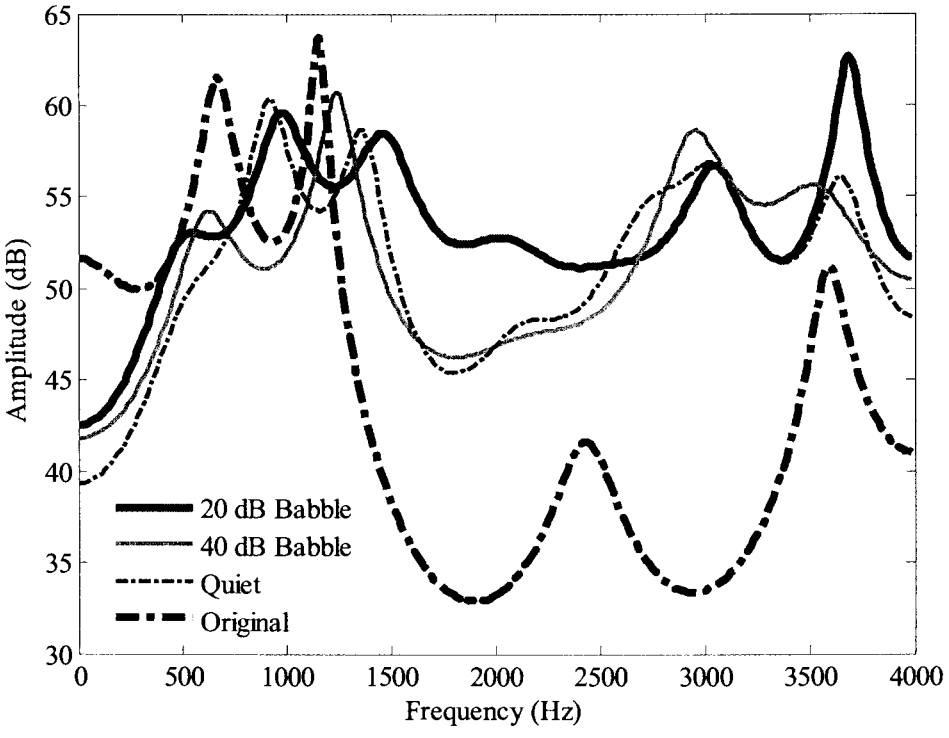

Figure 4.63. Formants of the vowel /a/ to show reduction in spectral contrast with increasing level of noise. The four conditions displayed are for 20 dB SNR (babble noise), 40 dB SNR (babble noise), processed in quiet and the original unprocessed vowel

Results from studies performed previously indicate that the minimum spectral contrast needed for the recognition of vowels is 1 dB for normal-hearing persons (Leek and Summers, 1996) and 4-6 dB for hearing-impaired persons (Loizou and Poroy, 2001). Reduction in spectral contrast is therefore an important aspect that needs to be investigated for acoustic simulations in noise.

[5]Spectral contrast is defined in Leek and Summers (1994) as the contrast between spectral peaks and valleys of harmonic complexes

Vowel features used in the multidimensional scaling analysis are the duration, spectral contrast of F_1 (which corresponds to the probability of identifying F_1), spectral contrast of F_2 (which corresponds to the probability of identifying F_2), magnitude of the intensity of F_1 and F_2 (obtained from the formant analysis of original vowels), height of peaks above valleys for F_1 and F_2 and the -3 dB bandwidth of F_1 and F_2 LPC formant peaks. These features were correlated with the coordinates obtained from the multidimensional scaling to determine which of these are important for vowel recognition in noise.

The -3 dB bandwidth was determined to give an indication of the overlap of the formant spaces of the individual vowels. Figure 4.64 shows, for example, the overlapping bandwidths for the condition of multi-talker babble. Looking only at the graphic representation does not give a good indication of the effect caused by the reduction in spectral contrast. The formant spaces overlap completely for some of the vowels, yet they are still differentiated from one another. In the processed vowel space without noise, some of the vowels also overlap, but not as much as the processed vowels in noise. The -3 dB bandwidth measure does not appear to be a meaningful characteristic to use for the determination of the cause of increased vowel confusions with the increase in noise level.

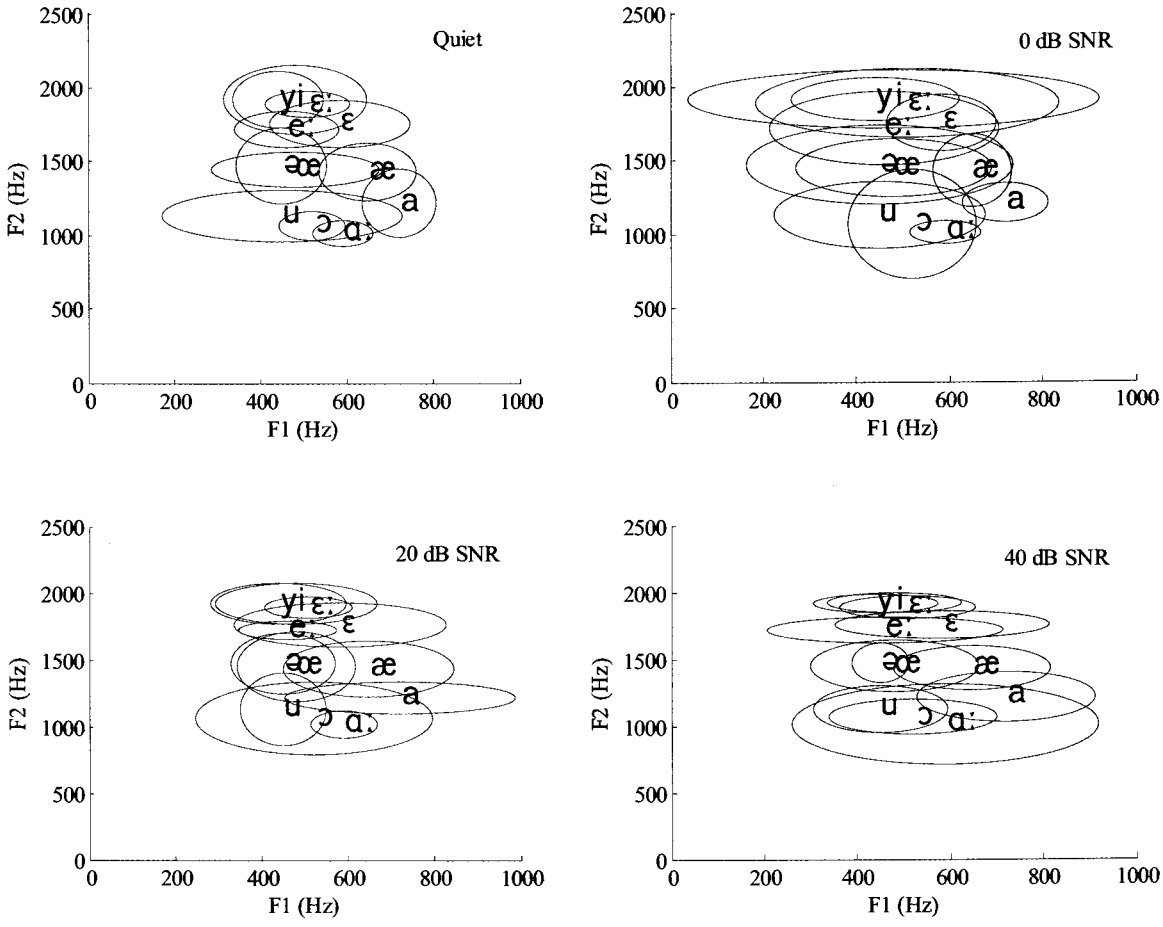

Figure 4.64. Vowel space of processed vowels in multi-talker babble with -3 dB bandwidths indicated. The specific SNRs are shown in the top right corner

VAF obtained from the multidimensional scaling indicates that the number of significant dimensions used for recognition decreases with increasing noise levels. Typically only one feature transmits information for recognition of vowels/consonants in low SNRs. This is a direct consequence of the presence of significant noise levels; SNRs of 0 dB have the least number of dimensions. The results from the correlations between the acoustic features and multidimensional scaling coordinates are shown in table 4.20.

Table 4.20. VAF, normalised weights and correlations with important features obtained from multidimensional scaling analysis for vowels

	VAF R^2	Normalised dimension weights	Highest correlating acoustic feature	r
Vowels - 0 dB multi-talker babble				
1st Dimension	0.726	0.855	Duration	0.717
2nd Dimension	0.914	0.482	F1 height above valley	0.305
3rd Dimension	0.878	0.135	-3 dB Bandwidth of F2	0.545
Vowels - 0 dB speech-like noise				
1st Dimension	0.66	0.817	Duration	0.742
2nd Dimension	0.876	0.508	F2 height above valley	0.633
3rd Dimension	0.851	0.190	-3 dB Bandwidth of F1	0.388
Vowels - 20 dB multi-talker babble				
1st Dimension	0.602	0.782	Duration	0.935
2nd Dimension	0.673	0.359	-3 dB Bandwidth of F1	0.506
3rd Dimension	0.723	0.330	Probability of identifying F2	0.705
Vowels - 20 dB speech-like noise				
1st Dimension	0.588	0.774	Duration	0.865
2nd Dimension	0.696	0.427	Magnitude of F2 peak	0.473
3rd Dimension	0.733	0.274	Probability of identifying F1	0.552
Vowels - 40 dB multi-talker babble				
1st Dimension	0.341	0.629	Probability of identifying F1	0.726
2nd Dimension	0.594	0.516	Duration	0.522
3rd Dimension	0.67	0.365	F1 height above valley	0.491
Vowels - 40 dB speech-like noise				
1st Dimension	0.524	0.734	Duration	0.915
2nd Dimension	0.709	0.480	Magnitude of F1 peak	0.636
3rd Dimension	0.771	0.346	F1 height above valley	0.685

The indication of spectral contrast for F_1 and F_2 was obtained using an algorithm written in Matlab, using signal information extracted from the vowel. The LPC of the relevant vowel was calculated and used to determine the height of F_1 and F_2 above the noise present in the

signal. The standard deviations of these heights were determined to give an indication of the power present in the specific formant peak (Proakis and Salehi, 2002). The mean of the noise was determined as 0, with the mean of the peak calculated with reference to the valley between the F_1 and F_2 peaks. To determine the probability of identifying F_1 or F_2, signal detection theory was used (Gelfand, 1990; Proakis and Salehi, 2002).

For the example of figure 4.65, the probability density function (pdf) on the left represents the noise at 0 dB SNR and the pdf on the right represents the mean height of the F_1 peak for /i/ at 0 dB multi-talker babble. If the intersection of the two graphs is at position r, the probability of correctly detecting the presence of F_1 amid noise is Q(r) (assuming that the listener places the detection criterion at r and that the pdfs have the same variance). Q(r) is the error function defined in equation 4.5 for the upper bound and 4.6 for the lower bound (Proakis and Salehi, 2002),

$$Q(r) < \frac{e^{-0.5r^2}}{r\sqrt{2\pi}} \quad , \tag{4.5}$$

and

$$Q(r) > \frac{e^{-0.5r^2}}{r\sqrt{2\pi}}\left(1-\frac{1}{r^2}\right) \quad . \tag{4.6}$$

By determining the probabilities, correct detection of F_1 and F_2, an indication of successful identification of formant frequencies with reduced spectral contrast, can be determined. Lower values for Q(r) indicate poorer detection of the formant peaks.

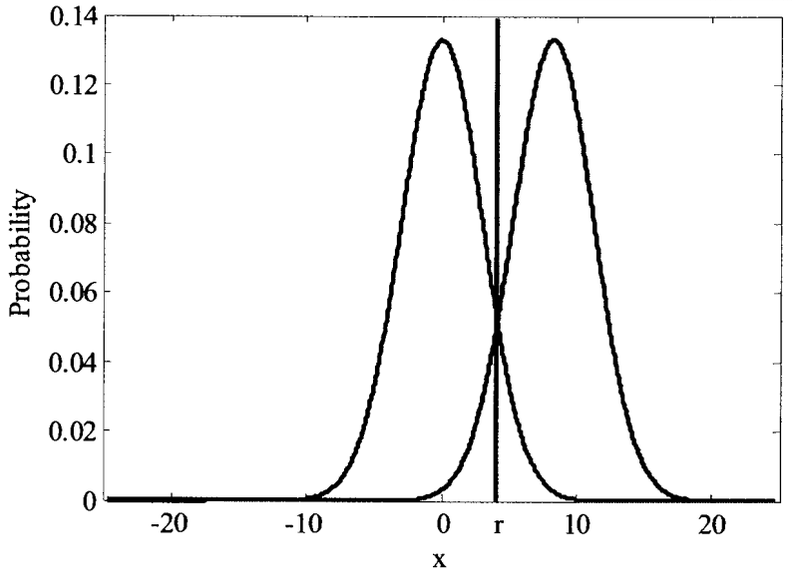

Figure 4.65. Probability density functions for noise (lefthand side) and F1 (righthand side) for the vowel /i/

From the multidimensional scaling, the acoustic feature that is transmitted most effectively is the duration. Five out of six noise conditions have high correlation between the first dimension's coordinates and the duration of the vowel, with normalised dimension weights approximately double those of the second most important feature, which is usually spectral contrast. This means that, although spectral contrast is important for the recognition of formants, the presence of noise masks spectral information and forces listeners to concentrate on duration (the temporal information is more robustly transmitted). The spectral contrast is used gradually to a lesser extent as noise levels increase. At low SNRs there are only a few dimensions and features that are used to identify vowels. Duration is therefore one of the acoustic features transmitted most effectively for the recognition of vowels in noise, similar to that found for recognition in quiet.

The acoustic features extracted from the consonants in noise are the same as those determined for the consonants in quiet, i.e. duration of consonant, peak level energy, median level energy, minimum to peak energy ratio, peak and median level energy after low-pass filtering (20 Hz cutoff) and envelope variation between 20 and 200 Hz. Correlations between the multidimensional scaling coordinates and the acoustic features and classifications were obtained to determine which feature or classification is responsible

for recognition of consonants.

The correlations between the speech production features and multidimensional scaling coordinates are shown in table 4.21, results from the acoustic analysis and multidimensional scaling are summarised in table 4.22.

In contrast with the results of the vowels in noise, the number of significant dimensions in the multidimensional scaling does not decrease with increasing noise levels. It seems that there are still approximately three features that contribute to the recognition of consonants in noise. For five out of the six noise conditions, burst (a temporal feature) plays the most important role in recognition. This corresponds with the results found for quiet conditions, showing that this feature is still present in noisy conditions.

For the lower SNRs, the other features that are important are nasality and liquidity. The high percentage recognition of /r/ contributes to the fact that liquidity is transmitted effectively. As the SNR increases, the speech segments' signal characteristics approach those of consonants in quiet and at high SNRs, the same information should be transmitted as in quiet. This is shown in the results; affrication, place and manner are transmitted most effectively at higher SNR results, similar to the results in quiet conditions.

Table 4.21. VAF, weights and correlations with important features obtained from multidimensional scaling analysis for consonants (speech production feature)

	VAF R^2	Normalised dimension weights	Highest correlating speech production feature	r
Consonants - 0 dB multi-talker babble				
1st Dimension	0.483	0.708	Burst	0.267
2nd Dimension	0.658	0.470	Liquidity	0.694
3rd Dimension	0.697	0.316	Nasality	0.468
Consonants - 0 dB speech-like noise				
1st Dimension	0.441	0.679	Affrication	0.482
2nd Dimension	0.624	0.478	Nasality	0.496
3rd Dimension	0.711	0.379	Place	0.166
Consonants - 20 dB multi-talker babble				
1st Dimension	0.677	0.827	Burst	0.889
2nd Dimension	0.746	0.356	Nasality	0.594
3rd Dimension	0.792	0.325	Liquidity	0.714
Consonants - 20 dB speech-like noise				
1st Dimension	0.425	0.669	Burst	0.812
2nd Dimension	0.610	0.480	Nasality	0.693
3rd Dimension	0.716	0.400	Manner	0.576
Consonants - 40 dB multi-talker babble				
1st Dimension	0.635	0.802	Burst	0.936
2nd Dimension	0.700	0.351	Place	0.312
3rd Dimension	0.733	0.307	Affrication	0.597
Consonants - 40 dB speech-like noise				
1st Dimension	0.43	0.672	Burst	0.937
2nd Dimension	0.660	0.520	Affrication	0.834
3rd Dimension	0.719	0.344	Voice	0.233

Table 4.22. VAF, weights and correlations with important features obtained from multidimensional scaling analysis for consonants (acoustic feature)

	VAF R^2	Normalised dimension weights	Highest correlating acoustic feature	r
Consonants - 0 dB multi-talker babble				
1st Dimension	0.483	0.708	Median LPF energy level	0.287
2nd Dimension	0.658	0.470	Median energy level	0.381
3rd Dimension	0.697	0.316	Envelope variation	0.381
Consonants - 0 dB speech-like noise				
1st Dimension	0.441	0.679	Median energy level	0.334
2nd Dimension	0.624	0.478	Median LPF energy level	0.537
3rd Dimension	0.711	0.379	Envelope variation	0.346
Consonants - 20 dB multi-talker babble				
1st Dimension	0.677	0.827	Minimum/Peak ratio	0.818
2nd Dimension	0.746	0.356	Duration	0.449
3rd Dimension	0.792	0.325	Median LPF energy level	0.341
Consonants - 20 dB speech-like noise				
1st Dimension	0.425	0.669	Median energy level	0.664
2nd Dimension	0.610	0.480	Median LPF energy level	0.369
3rd Dimension	0.716	0.400	Peak energy level	0.250
Consonants - 40 dB multi-talker babble				
1st Dimension	0.635	0.802	Envelope variation	0.870
2nd Dimension	0.700	0.351	Peak energy level	0.257
3rd Dimension	0.733	0.307	Peak LPF energy level	0.497
Consonants - 40 dB speech-like noise				
1st Dimension	0.43	0.672	Median energy level	0.924
2nd Dimension	0.660	0.520	Duration	0.535
3rd Dimension	0.719	0.344	Envelope variation	0.466

Correlation coefficients for the lower SNRs are low and show that there are few features that contribute significantly to recognition of consonants. With the increase in SNR, the correlations approach 1, indicating stronger contributions to consonant recognition. It

appears that the recognition of consonants at lower SNRs is at chance level (<10 %). The multidimensional scaling also shows that there is no notable difference in speech recognition in speech-like noise or multi-talker babble.

4.4 SUMMARY

In this chapter, the results of the experimental studies were given. Analysis of the output from the acoustic simulation was used to predict possible confusions of vowels and consonants. Important results recorded in this chapter that were used in the predictions and explanations of confusions are the first and second formant frequencies (F_1 and F_2) of the vowels, the duration of the vowels and the signal properties of the consonants. All these values were obtained from the processed vowels and consonants using PRAAT or Matlab.

Results from the acoustic simulation experiments were presented for both quiet and noisy conditions. From these results, it can be seen that the model is a good simulation of cochlear implants. Results reported for the experimental study are: confusion matrices for vowels and consonants before and after dynamic range compression, Euclidean distance matrices for vowels, FITA analyses, growth functions of percentage correctly recognised and recognition of vowels and consonants in speech-like noise and multi-talker babble. The results presented in this chapter were also discussed where necessary. The final discussion of all the results will be done in the following chapter. Results obtained in this study will be compared to those found in the literature.

CHAPTER 5 DISCUSSION

5.1 CHAPTER OBJECTIVES

In this chapter the results presented previously will be discussed and a general discussion of all the results will be given. The results from the development of the model as well as the results from the experimental study will be linked with the results from other studies reported in the literature. The implication of the work and suggestions for future work will also be presented.

5.2 CONTRIBUTIONS

A comprehensive acoustic model was developed, consisting of two distinct models, the signal processing model and the biophysics model. It is especially the development of the biophysics model that contributes to the current state of the literature.

In this study, the effect of dynamic range compression on speech recognition was investigated and the confusions of speech segments were explained through the analysis of the acoustic model. There appears to be a limited number of studies that examine the effect of performing dynamic range compression on the amplitudes used for stimulation on speech recognition (Fu and Shannon, 1998; Loizou, Dorman and Fitzke, 2000).

Another important contribution is the modelling of the effect of current spread on the area of stimulated nerve fibres in the cochlea. This was simulated by changing the bandwidth of the noise bands used for the reconstruction of a sound signal.

The modelling of the pitch and stimulation rate (which is also asynchronous) of electrodes was effectively implemented in this study (using the harmonics of a speech signal). It is possible to determine what the effect of the stimulation rate is on speech recognition by analysing the acoustic simulation and results found through experimental studies.

By performing acoustic analyses (multidimensional scaling and Euclidean distance measures) on the processed speech segments, a connection can be established between the acoustic properties of speech and the confusions among speech segments. The inclusion of these acoustic analyses is an important part of this study and presents an effective method of determining what underlies speech recognition.

Multidimensional scaling for both vowel and consonant recognition in noise gives an indication of the acoustic features that contribute to the recognition of these speech segments. Duration (temporal information) of vowels is transmitted well even in the presence of noise. The multidimensional scaling indicated that for lower SNRs, this feature is mostly used to recognise vowels. The spectral contrast of vowels is reduced considerably in the presence of noise, contributing to a reduction in speech recognition in noisy conditions. As the SNR decreases, so does the total number of features transmitted for speech recognition. For consonants, at higher signal-to-noise ratios, the important features are similar to those in quiet conditions, namely burst (again temporal information) and affrication. At 0 dB multi-talker babble and speech-like noise, the major contributing factor to recognition of consonants is burst, with nasality and liquidity also playing an important role. It appears that temporal information is retained after processing and transmitted effectively to the listener while spectral information is lost and not available to the listener in order to recognise speech. The temporal characteristics of speech are robust in the presence of noise.

As mentioned previously, a more comprehensive model has been developed that includes the simulation of various aspects of a cochlear implant. This model can now be used in different experiments to determine what the effect on speech recognition will be when some of the parameters of the acoustic model are changed.

5.3 DISCUSSION OF RESEARCH QUESTIONS

With respect to the research questions posed in chapter 1, the following conclusions were made:

- For the successful implementation of an acoustic model, there are a number of specific processing steps that need to be incorporated, including dynamic range compression, current spread in the cochlea, stimulation rate and asynchronous stimulation.
- A clearer understanding has been reached as to what underlies speech recognition in cochlear implants; temporal information is transmitted effectively while spectral information is lost through the processing of speech.
- The inclusion of dynamic range compression has a significant effect on vowel recognition; there is a statistically significant difference between the results found before dynamic range compression and after dynamic range compression. There is no statistically significant difference between the results found with electric hearing and simulations after dynamic range compression. For consonants, there is no significant difference between any of the conditions, it appears that dynamic range compression does not have a significant effect on consonant recognition.
- In the presence of speech-like noise, spectral information is lost while temporal information is still present for the recognition of speech, which is also consistent with the results found in quiet.
- The important acoustic cues used to recognise speech in the presence of dynamic range compression and speech-like noise, are spectral contrast and duration for vowels and burst for consonants.

5.4 COMPARISON WITH OTHER ACOUSTIC MODELS

The performance of normal-hearing persons listening to the acoustic simulations presented in this study compare well with the results found for previously implemented acoustic models. It has been found in previous studies that normal-hearing listeners adapt to the sound of degraded speech (Throckmorton and Collins, 2002), as was shown in this study. There are a number of experiments that serve as training, after which the recognition of speech segments starts to stabilise.

The percentage correct scores for the acoustic simulations in this study compare reasonably well with the results found using previously implemented acoustic models. The percentage

scores for phonemes are plotted in Throckmorton and Collins (2002) for a number of acoustic simulations. On average, the percentage scores are around 65 %, which compares well with the scores of 67 % for vowels in this study. The percentage score for consonants, 52 %, is lower than the percentage scores in Throckmorton and Collins (2002).

For a four-channel CIS strategy acoustic simulation (Fu and Shannon, 1999), the percentage score for vowels compares well with that found in the present study. At an insertion depth of 25 mm (used in this study as well), the percentage correct score is around 62 %. The four-channel CIS strategy at 25 mm compares well with the SPEAK strategy implemented in the current acoustic model, also at an insertion depth of 25 mm.

The percentage scores for acoustic simulations presented in Dorman et al. (2002) are higher in general than the percentage scores obtained in this study. For an eight-channel SPEAK processor (as implemented in this study), the percentage score for vowels is approximately 90 %, which is high compared to the results found with electric hearing. Also, in noise, the percentage scores do not degrade as much as was found in this study. At -2 dB SNR, the percentage correct score is around 80 %, in contrast to an average of 10 % found in this study. It appears that the acoustic model developed in this study is more realistic than the model implemented in Dorman et al., (2002).

For the results found in chapter 4, there is no literature that presents the results after each processing step to enable a comparison with the results found in this study. Some of the results can, however, be compared to those found with the NMT. One of the results that is specifically important is the comparison of the spectrogram for the acoustic simulation and the NMT.

The preprocessing of the speech signals (pre-emphasis filter, bandpass filters and full-wave rectification) yielded the desired results. The pre-emphasis filter is a general processing step found in the majority of existing models. The bandpass filters were implemented similar to Dorman et al. (1997b), Loizou (1999b), Loizou (1998) and Shannon et al. (1995) using Greenwood's frequency-place equation (Greenwood, 1990). The original and processed signals have the same shape in the time domain. The processed signals have a

broader frequency spectrum owing to the use of noise bands and their harmonics. The next processing step of calculating the RMS of the envelope of the signal was performed in a similar manner to the procedure followed in other studies (Dorman et al., 1998; Dorman et al., 1997b; Loizou, 1999a; Throckmorton and Collins, 2002), calculating the RMS every 8 ms with overlapping Hanning windows.

Conversion of the RMS values to current levels does not appear to be reported in the literature. This is a necessary step to simulate the biophysical interaction between the stimulating current and sound perception. Including this step in the acoustic model expands models in the present literature. The quantised current levels are a good approximation to the dynamic range compression present in cochlear implants. From the results it can be seen that by including this step, the results compare better with those obtained with cochlear implant users than when the dynamic range compression is not included.

Most of the models implemented previously use either sinusoidal signals or noise bands for the final summation of channels for the acoustic simulation (Dorman et al., 1998; Dorman et al., 1997b; Dorman et al., 2002; Friesen et al., 2001; Fu et al., 1998; Loizou, 1999b). None of the models implementing noise bands allocated the bandwidth of the noise bands dynamically according to the magnitude of the stimulation current. In this study, the bandwidths used for the noise bands are calculated according to the intensity of the stimulating current. Initially, the bandwidths were calculated dynamically, but because of processing overhead, this was removed for the acoustic simulations that were performed. It was determined that the bandwidths do not change significantly for a fixed insertion depth. This step is important, however, in cases where the insertion depth is varied. The bandwidths will change owing to the change in the position of a specific electrode. The dynamic allocation of bandwidths should be included when the insertion depth of an electrode array is varied, which was not done in this study.

For the summation of all the channels using noise bands, the main contribution of this model is the simulation of asynchronous stimulation and stimulation rate. In previous studies, the asynchronous stimulation was modelled by simply introducing a time delay

(Fu and Galvin III, 2001). This is not sufficient; the effect that stimulation rate has on speech recognition is not simulated accurately. In this study, a more realistic implementation of asynchronous stimulation was included in the model. One of the strong points of this model is the simulation of the stimulation rate in such a way that low frequency signals are presented at a high stimulation rate (Terhardt, 1979; Terhardt et al., 1982). One weakness of this implementation is that the number of spectral channels increases by adding the harmonics of the low frequency components, i.e. a more robust, spread-spectrum-like representation is created. This is not desirable when the effect of the number of channels is investigated. Because of the increased number of spectral channels, this model cannot be used to simulate the effect of number of channels on speech recognition. For simulating the effect of number of channels, the simulation of the stimulation rate may need to be improved.

5.5 COMPARISON WITH COCHLEAR IMPLANT DATA

Before comparisons can be made between results found with the acoustic simulations and electric hearing, it is important to understand which acoustic cues are used in the recognition of phonemes. This was determined through the comparison of the predictions from the acoustic model and results found with normal-hearing listeners. To make sure the acoustic model is appropriate for modelling cochlear implants, data from experiments for electric hearing were compared with the results found with the acoustic simulations in quiet conditions. The results for electric hearing and the acoustic simulations compared well, indicating that the implemented acoustic model is an appropriate model of cochlear implants. Because the model is trustworthy in quiet conditions, it can now be applied to various conditions. In this study speech recognition in noise was investigated. The output of the model provides a means to gain access to measurements deeper inside the auditory system than is possible without the simulation. With this in mind, the results from the acoustic model are presented in this section and compared to results found for electric hearing.

The acoustic model developed in this study was used to process vowels and consonants for recognition under conditions of quiet (before and after dynamic range compression) and

noise (only after dynamic range compression). The noise conditions were at 0 dB, 20 dB and 40 dB SNR for multi-talker babble and speech-like noise (CCITT Recommendation 227) (Dubno et al., 2005).

The percentage recognised correctly for quiet conditions, 67 % for vowels and 52 % for consonants (pooled across listeners), compare relatively well with the results found in Skinner, Holden, Holden, Demorest and Fourakis (1997) (70 % for vowels and 66 % for consonants). The percentage score for consonants is lower than most of the results found previously. The acoustic model simulates the processing of speech in cochlear implants better for vowels than for consonants. The scores are in the same range as found in Pretorius et al. (2005), where the percentage recognised is 63 % and 72 % for vowels and consonants respectively. In Van Wieringen and Wouters (1999), scores of 42 % and 33 % were reported for Laura cochlear implantees. The results from this study compare well overall with those found in the literature. Personal differences have been noted between cochlear implant users previously. The variation in results between all the studies can be ascribed to these differences between participating subjects.

Confusions of vowels are consistent with the results found in Pretorius et al. (2005), Skinner et al. (1997) and Van Wieringen and Wouters (1999). FITA analyses show that the same information is transmitted in this study using a model, as for the vowels recognised in Pretorius et al. (2005) by cochlear implantees. Distance measures for F_1, F_2 and duration have been shown to be important for analysis of the recognition of vowels using the acoustic model. From the FITA analysis, the feature transmitting most information for the vowels is the duration of the vowel followed by the value of F_2.

The recognition of consonants is poorer than for vowels when using the acoustic model, similar to the results found in Van Wieringen and Wouters (1999) and Skinner et al. (1997) for cochlear implant users. This is in contrast to the results found in Pretorius et al. (2005), where the consonants were recognised at a significantly higher percentage in four subjects. The consonant confusions are grouped, similar to those reported in Pretorius et al. (2005), indicating that the same information is transmitted for both cases. This can also be seen when comparing the results from the FITA analysis. From the multidimensional scaling

analysis (Rosen, 1992; Wang and Bilger, 1973), the important signal characteristics found in the present study are consistent with those found in Van Wieringen and Wouters, (1999) for cochlear implant users.

From the multidimensional scaling, the features transmitted most effectively for the consonants are burst and affrication. The SPEAK strategy does not convey temporal information as effectively as the CIS strategy. It is surprising that the burst feature is transmitted so effectively through the acoustic simulation (this is true for cochlear implants as well). The signal characteristic of envelope variation is very important for the recognition of consonants, being the most prominent acoustic feature for both the conditions before and after dynamic range compression. In Van Wieringen and Wouters (1999), envelope variation was the second most important feature after turbulence, indicating that the same information is transmitted in the acoustic simulations as for electric hearing.

Results for recognition in noise follow the same trend as the results reported in Killion et al. (2004) and Nie et al. (2005). There is a definite improvement in recognition with an increase in SNR. The percentage recognised correctly ranges from near chance to average recognition; for the vowels the mean scores were 13 %, 45 % and 59 % for 0 dB, 20 dB and 40 dB multi-talker babble noise respectively and 8 %, 42 % and 49 % for the consonants in multi-talker babble respectively. In speech-like noise, the scores were 11 %, 44 % and 54 % for the vowels and 7 %, 42 % and 54 % for the consonants, for 0 dB, 20 dB and 40 dB respectively. These results compare well with results obtained for Nucleus-22 cochlear implant listeners for various SNRs, showing that the acoustic model still simulates electric hearing with increasing levels of noise (Dubno et al., 2005; Friesen et al., 2001; Fu et al., 1998; Yang and Fu, 2005).

In general, it appears that speech recognition is similar in conditions of speech-like noise and in multi-talker babble. This indicates that both types of noise degrade speech in a similar way, which is expected, as both have spectral components in the same frequency bands. Pilot experiments performed with white noise (Pollack and Pickett, 1957) showed that speech recognition in white noise did not deteriorate as it did in the presence of

speech-like noise. This may be ascribed to the fact that in the acoustic simulation the spectral components are spread over a broad frequency band owing to the use of harmonics (as explained in section 3.3.2.8).

5.6 CURRENT-LOUDNESS MAPPING

The difference between the percentage recognised correctly for the experiments before and after dynamic range compression is on average 8 %, before compression resulting in higher recognition scores. In general, the results for the experiments after compression compare better with those reported in the literature. This indicates that the inclusion of the mapping and quantisation of stimulation currents is necessary. There have been limited studies that included the mapping of current in an acoustic model. Fu and Shannon (1998) and Loizou, Dorman and Fitzke (2000) included the mapping of current in their model but did not include the quantisation of the current levels.

5.7 GENERAL DISCUSSION

Results from the developed acoustic model and the experimental study compare well with those found in literature and the NMT. Confusions for vowels and consonants are consistent with those found in Pretorius et al. (2005) for quiet conditions and those found in Dubno et al. (2005), Friesen et al. (2001), Fu et al. (1998) and Yang and Fu (2005) for noisy conditions. The results suggest that the developed acoustic model is an appropriate simulation of cochlear implants.

The model has a number of strong and weak points that merit discussion. The use of harmonics in the summation of the final output to simulate the stimulation rate in the cochlea is effective. This approach, however, will only be sufficient under conditions making provision for a fixed number of channels. When the effect of number of independent channels is investigated, other possible implementations need to be considered.

Furthermore, the acoustic simulation of the high stimulation rate would have to be implemented differently for the CIS strategy than for the SPEAK strategy. Sufficient information about the speech signal must be transmitted within a short simulation window. To do this, an acoustic model for the CIS strategy should be considered that conveys low frequency information in a small simulation window. It is specifically this higher stimulation rate for the CIS speech processing strategy that makes it valuable (Wilson, Lawson, Finley and Wolford, 1991). It was found that a meaningful implementation for the CIS strategy needs to be investigated in future.

Models from other studies (Bruce et al., 1999; Hanekom, 2001) were used to determine the spread of stimulation current inside the cochlea. The model that was developed contributed to the present state of the literature by using the magnitude and spread of the stimulation current to determine the bandwidth of simulation frequency bands. This is an important aspect of cochlear implants incorporated into the current model. As mentioned previously, the dynamic allocation of bandwidths was removed from this model owing to processing overhead and the constancy of the bandwidths over a period of time. When the insertion depth is changed for specific simulations, this processing step must be included to calculate realistic bandwidths of the noise bands at specific places in the cochlea. By improving the dynamic allocation in terms of processing time, the model can be improved.

In the processing of the acoustic models studied (Dorman et al., 1997a; Friesen et al., 2001; Fu et al., 1998; Loizou, 1999a; Loizou, 1998), the magnitudes of the noise bands or sinusoidal waves were determined directly from the RMS values. This is not exactly what happens in the cochlear implant processor – the RMS values are transformed to quantised current levels used for stimulation. In the model developed in this study, this step was included for a more accurate acoustic simulation.

In the analysis of confusion matrices for the vowels, a multidimensional scaling approach may have been appropriate. Through this approach, the recognition of vowels can be linked to specific signal characteristics and an even better understanding of what underlies speech recognition in cochlear implants can be reached. Multidimensional scaling was not performed because it was possible to explain the confusions between vowels adequately

with the distance measures determined. Various acoustic analyses were performed to determine the underlying features responsible for successful vowel (FITA with Euclidean distances) and consonant recognition (FITA and multidimensional scaling).

For the experiments in noise, only three different SNRs were used to process the speech tokens. From the results it can be seen that there is no significant difference between multi-talker babble and speech-like noise. The number of experiments can be reduced by using only one of these noise sources.

In general, the results show that the implementation of this acoustic model was successful. This model can now be used in further studies to determine the effect that various cochlear implant parameters have on speech recognition.

CHAPTER 6 CONCLUSION

This final chapter will describe what has been done and what has been achieved and making suggestions for future work. The literature study, methods, results and discussion will be observed as a whole and the relevant conclusions will be made.

A literature study was performed on cochlear implants and existing acoustic models. From this study, a number of research gaps were identified and some of these issues were addressed in this dissertation. A thorough study was carried out to determine the exact processing steps of a cochlear implant processor and the biophysics associated with cochlear implants.

From the information gathered through the literature study, an acoustic model was developed for the SPEAK strategy. The acoustic model consists of two parts: the speech processing model (analogous to that of the cochlear implant processor) and the biophysical model (simulation of the biophysical interaction between the cochlea and cochlear implant).

The first objective of this study was completed by the implementation of the acoustic model, incorporating the exact processing steps of the SPEAK speech processing strategy and the biophysics of a cochlear implant. The similarity of the electrodograms for the NMT and the acoustic simulation confirms the similar processing of speech with the cochlear implant processor and the model that was developed.

With reference to the second objective, the effect of dynamic range compression was determined using acoustic simulations from the acoustic model. Experiments were conducted with normal-hearing persons for vowels and consonants before and after dynamic range compression. Acoustic analyses were performed on the processed speech segments in order to determine which features are transmitted most effectively for the recognition of vowels and consonants. These analyses were completed together with analyses of the confusion matrices, using either multidimensional scaling or FITA analysis. By doing this, the underlying source of the recognition of vowels and consonants was

determined in quiet and noisy conditions.

Finally, research was done on the recognition of vowels and consonants in the presence of speech-like or multi-talker babble noise. The specific acoustic features of vowels and consonants that are transmitted effectively for the recognition of speech segments in noise were determined. Through analysis of the confusion matrices, it was possible to establish which acoustic features are removed through the processing of speech and are not available for the recognition of speech.

In summary, the following conclusions can be drawn from the results.

- The acoustic model with the dynamic range compression mimics cochlear implants more closely than the acoustic model without dynamic range compression. This shows that dynamic range compression has an effect on speech recognition.
- The information transmitted most effectively in quiet and noise is the duration of vowels and burst of consonants.
- The formants are not clearly defined for processed vowels (in quiet and noisy conditions), reducing recognition of vowels, especially after dynamic range compression.
- Spectral contrast is important for the recognition of vowels; when spectral contrast is reduced, the recognition of vowels becomes poor.
- Temporal information, including duration and burst, is transmitted effectively by the acoustic model and by the speech processor for cochlear implantees for both vowels and consonants.
- The information about manner of articulation is lost after processing, reducing the number of features available for recognition of consonants.
- Envelope variation is an important signal characteristic for consonant recognition in quiet; in noise median energy levels is important.
- The total number of features used for recognising vowels and consonants becomes lower after processing, especially with dynamic range compression and in noisy conditions.

6.1 FUTURE WORK

The development of the acoustic model in this study answered many questions and raised a number of questions as well. During the development of the model, a number of stumbling blocks were encountered. Some of the problems were solved, while others were left unanswered. One of the main areas where future research is necessary is in the development of an acoustic model for the CIS speech processing strategy. It was found in this study that the approach must be different for the development of a model for the CIS speech processing strategy. An important characteristic of the CIS strategy is the implementation of a high stimulation rate in contrast with the SPEAK strategy that focuses on an increased number of spectral channels.

For the summation of the noise bands or sinusoidal waves, more research can be done to find an efficient way to simulate the effect of the stimulation rate. The problem lies in the fact that the period of stimulation is significantly shorter than the period of typical speech signals, making it difficult to stimulate the cochlea acoustically using the output of the model. With electric stimulation, there is no problem with stimulating high frequency information with a short stimulation pulse, considering the fact that electric current is used at a specific place in the cochlea to produce a sensation of a particular frequency. With the implementation in this study, the increased number of spectral channels owing to the use of harmonics does not have a significant effect on speech recognition. When specific studies are performed on the effect of the number of available spectral channels on speech recognition, this implementation will not be sufficient.

Acoustic simulations should be performed to determine the effect of insertion depth on speech recognition. The simulations can be used to determine the difference in speech recognition under two conditions: frequency-place mapping as opposed to the case where a frequency band is fixed for a specific electrode. From these results, it is possible to optimise the speech processor for an individual, if the insertion depth of the person's cochlear implant is known. For these studies to be conducted, the dynamic allocation of noise band bandwidths must be included.

Although a number of important cues for speech recognition were identified, it is still not clear if all the signal characteristics that are important for the recognition of vowels and consonants have been identified. There may still be acoustic properties that are important for recognition of speech that were not discussed in this study. The developed model can be used to identify such acoustic properties by performing predictions of speech recognition and performing analyses on the confusions and acoustic properties from the output of the acoustic model.

The developed model can now be used to develop better maps for cochlear implants. Various scenarios can be set up and the effect of changing a specific parameter can be determined by analysing the output of the acoustic model.

A suggestion to improve current cochlear implant processors is to find a method to convey frequency information more effectively. This should improve the transmission of important acoustic cues such as F_1 and F_2 in the cochlear implant, which will in turn improve recognition of vowels. By analysing the output of the acoustic model, other possible improvements may also be identified. The developed model will therefore make a contribution to the improvement of current cochlear implant processors.

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.

Alcantara, J. I. and Moore, B. C. J. (1995). The identification of vowel-like harmonic complexes: Effects of component phase, level, and fundamental frequency, *Journal of the Acoustical Society of America*, **97**(6): 3813-3824.

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

Baskent, D., Shannon, R. V. and Baskent, D. (2003). Speech recognition under conditions of frequency-place compression and expansion, *Journal of the Acoustical Society of America*, **113**(4 I): 2064-2076.

Becken, E. T., Donaldson, G. S., Kimberley, B. P. and Nelson, D. A. (2005). Neural survival and psychophysical measures of electric hearing in cochlear implant users, *Otolaryngology - Head and neck surgery*, **131**(2): 157-157.

Bhatia, K., Gibbin, K. P., Nikolopoulos, T. P. and O'Donoghue, G. M. (2004). Surgical complications and their management in a series of 300 consecutive pediatric cochlear implantations, *Otology and Neurotology*, **25**: 730-739.

Boersma, P. and Weenink, D. (2004). Praat, a system for doing phonetics by computer, version 3.4., *Report of the Institute of Phonetic Sciences Amsterdam nr 132*, **182**.

Boothroyd, A., HnathChisolm, T. and Hanin, L. (1985). A sentence test of speech perception: Reliability, set-equivalence, and short-term learning, *(Internal report RCI 10). New York: City University of New York.*

Borden, G. J. and Harris, K. S. (1994). *Speech Science Primer* Williams and Wilkins, c1994., Baltimore, Md.

Bruce, I. C., White, M. W., Irlicht, L. S., O'Leary, S. J., Dynes, S., Javel, E. and Clark, G. M. (1999). A stochastic model of the electrically stimulated auditory nerve: single-pulse response, *IEEE Transactions on Biomedical Engineering*, **46**(6): 617-628.

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

Chatterjee, M. (1999). Effects of stimulation mode on threshold and loudness growth in multielectrode cochlear implants, *Journal of the Acoustical Society of America*, **105**(2): 850-860.

Clark, G. M. (2003). *Cochlear implants: fundamentals and applications* AIP Press, New York.

Danhauer, J., Ghadialy, F., Eskwitt, D. and Mendel, L. (1990). Performance of 3M/House cochlear implant users on tests of speech perception, *Journal of the American Academy of Audiology*, **1**: 236-239.

Dorman, M. F. and Loizou, P. C. (1997). Changes in speech intelligibility as a function of time and signal processing strategy for an Ineraid patient fitted with Continuous Interleaved Sampling (CIS) processors, *Ear and Hearing*, **18**: 147-155.

Dorman, M. F., Loizou, P. C., Fitzke, J. and Tu Z (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels, *Journal of the Acoustical Society of America*, **104**(6): 3583-3585.

Dorman, M. F., Loizou, P. C. 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. F., Loizou, P. C. 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.

Dorman, M. F., Loizou, P. C., Spahr, A. J. and Maloff, E. (2002). A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants, *Journal of Speech, Language and Hearing Research*, **45**: 783-788.

Dowell, R., Seligman, P. M., Blamey, P. J. and Clark, G. M. (1987). Evaluation of a two-formant speech processing strategy for a multichannel cochlear prosthesis, *Annals of Otology, Rhinology and Laryngology*, **96**: 132-134.

Dubno, J. R., Horwitz, A. R. and Ahlstrom, J. B. (2005). Recognition of filtered words in noise at higher-than-normal levels: Decreases in scores with and without increases in masking, *Journal of the Acoustical Society of America*, **118**(2): 923-933.

Fastl, H. (1987). A background noise for speech audiometry, *Audiology Acoustics*, **26**: 2-13.

Faulkner, A. and Rosen, S. (1999). Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception, *Journal of the Acoustical Society of America*, **106**(4): 2063-2073.

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.

Faulkner, A., Rosen, S. and Wilkinson, L. (2001). Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking through a simulated cochlear implant speech processor, *Ear and Hearing*, **22**(5): 431-438.

Ferguson, S. H. and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners, *Journal of the Acoustical Society of America*, **112**(1): 259-271.

Fetterman, B. L. and Domico, E. H. (2002). Speech recognition in background noise for cochlear implant patients, *Otolaryngology - Head and neck surgery*, **126**: 257-263.

Friesen, L. M., Shannon, R. V., 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.

Fu, Q. J. and Galvin III, J. 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. V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing, *Journal of the Acoustical Society of America*, **105**(3): 1889-1900.

Fu, Q. J. and Shannon, R. V. (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., Shannon, R. V. and Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing, *Journal of the Acoustical Society of America*, **104**(6): 3586-3596.

Gelfand, S. A. 1990, "Theory of signal detection," in *Hearing. An introduction to psychological and physiological acoustics*, Marcel Dekker Inc., New York, pp. 313-324.

Greenwood, D. (1990). A cochlear frequency-position function for several species - 29 years later, *Journal of the Acoustical Society of America*, **87**: 2592-2605.

Hanekom, T. (2001). Three-Dimensional spiraling finite element model of the electrically stimulated cochlea, *Ear and Hearing*, **22**(4): 300-315.

Hartman, W. M. (1998). *Signals, sound, and sensation* Springer Science, Woodbury, N.Y.

Hawkins, H. L., McMullen, T. A., Popper, A. N. and Fay, R. R. (1996). *Auditory computation* Springer, New York.

Henry, B. A. and Turner, C. W. (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.

Hillenbrand, J., Getty, L. A., Clark, M. J. and Wheeler, K. (1995). Acoustic characteristics of American English vowels, *Journal of the Acoustical Society of America*, **97**: 3099-3111.

Hinojosa, R. and Marion, M. (1983). Histopathology of profound sensorineural deafness, *Annals of the New York Academy of Sciences*, **Vol. 405**: 459-484.

Hochberg, I., Boothroyd, A., Weiss, M. and Hellman, S. (1992). Effects of noise and noise suppression on speech perception by cochlear implant users, *Ear and Hearing*, **13**: 263-271.

Hochmair-Desoyer, I., Hochmair, E., & Stiglbrunner, H. 1985, "Psychoacoustic temporal processing and speech understanding in cochlear implant patients," in *Cochlear Implants*, R. Schindler & M. Merzenich, eds., Raven Press, New York, pp. 291-304.

Holden, L. K., Skinner, M. W., Holden, T. A. and Binzer, S. M. (1995). Comparison of the multipeak and spectral peak speech coding strategies of the Nucleus 22-channel cochlear implant system, *American Journal of Audiology*, **4**: 49-54.

Jolly, C. N., Spelman, F. A. and Clopton, B. M. (1996). Quadrupolar stimulation for cochlear prostheses: modeling and experimental data, *IEEE Transactions on Biomedical Engineering*, **43**(8): 857-865.

Kiefer, J., Müller, J., Pfennigdorff, T., Schön, F., Helms, J., Von Ilberg, C., Baumgartner, W. D., Gstöttner, W., Ehrenberger, K., Arnold, W., Stephan, K., Thumfart, W. and Baur, S. (1996). Speech understanding in quiet and in noise with the CIS speech coding strategy (MED-EL Combi-40) compared to the multipeak and spectral peak strategies (Nucleus), *ORL*, **58**: 127-135.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J. and Banerjee, S. (2004). Development of quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners, *Journal of the Acoustical Society of America*, **116**(4): 2395-2405.

Klein, W., Plomp, R. and Pols, L. C. (1970). Vowel spectra, vowel spaces, and vowel identification, *Journal of the Acoustical Society of America*, **48**(4): 999-1009.

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**(4): 966-972.

Kong, Y. Y., Cruz, R., Jones, J. A. and Zeng, F. G. (2004). Music Perception with Temporal Cues in Acoustic and Electric Hearing, *Ear and Hearing*, **25**(2): 173-185.

Leek, M. R. and Summers, V. (1996). Reduced frequency selectivity and the preservation of spectral contrast in noise, *Journal of the Acoustical Society of America*, **100**(3): 1796-1806.

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

Loizou, P. C. (1999b). Introduction to cochlear implants, *IEEE Engineering in Medicine and Biology*, **18**(1): 32-42.

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

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

Loizou, P. C. 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.

Matsui, J. I. and Cotanche, D. A. (2004). Sensory hair cell death and regeneration: two halves of the same equation, *Current Opinion in Otolaryngology & Head and Neck Surgery*, **12**: 418-425.

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

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

McDermott, H. J., McKay, C. M. and Vandali, A. E. (1991). An improved sound processor for a multiple-channel cochlear implant,13, pp. 1903-1904.

McKay, C. M. and Henshall, K. R. (2002). Frequency-to-electrode allocation and speech perception with cochlear implants, *Journal of the Acoustical Society of America*, **111**(2): 1036-1044.

McKay, C. M., Remine, M. D. and McDermott, H. J. (2001). Loudness summation for pulsatile electrical stimulation of the cochlea: Effects of rate, electrode separation, level, and mode of stimulation, *Journal of the Acoustical Society of America*, **110**(3 I): 1514-1524.

McKay, C. M., Vandali, A. E., McDermott, H. J. and Clark, G. M. (1994). Speech processing for multichannel cochlear implants: Variations of the spectral maxima sound processor strategy, *Acta Oto-Laryngologica*, **114**(1): 52-58.

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

Miller, J. M., Chi, D. H., O'Keeffe, L. J., Kruszka, P., Raphael, Y. and Altschuler, R. A. (1997). Neurotrophins can enhance spiral ganglion cell survival after inner hair cell loss, *International Journal for the Development of Neuroscience*, **15**(4/5): 631-643.

Müller, J., Schön, F. and Helms, J. (2002). Speech understanding in quiet and noise in bilateral users of the MED-EL COMBI 40/40+ cochlear implant system, *Ear and Hearing*, **23**(3): 198-206.

Müller-Deile, J., Schmidt, B. J. and Rudert, H. (1995). Effects of noise on speech discrimination in cochlear implant patients, *Annals of Otology, Rhinology and Laryngology*, **166**: 303-306.

Nie, K., Stickney, G. and Zeng, F. G. (2005). Encoding frequency modulation to improve cochlear implant performance in noise, *IEEE Transactions on Biomedical Engineering*, **52**(1): 64-73.

Nilsson, M., Soli, S. D. and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise, *Journal of the Acoustical Society of America*, **95**(2): 1085-1099.

Ohlemiller, K. K. and Gagnon, P. M. (2004). Apical-to-basal gradients in age-related cochlear degeneration and their relationship to "primary" loss of cochlear neurons, *The Journal of Comparative Neurology*, **479**(1): 103-116.

Pollack, I. and Pickett, J. M. (1957). Masking of speech by noise at high sound levels, *Journal of the Acoustical Society of America*, **30**(2): 127-130.

Pretorius, L. L., Hanekom, J. J., Van Wieringen, A. and Wouters, J. (2005). 'n Analitiese tegniek om die foneemherkenningsvermoë van Suid-Afrikaanse kogleêre inplantingsgebruikers te bepaal, *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie*, **Submitted for publication**.

Proakis, J. G. and Salehi, M. (2002). *Communication systems engineering*, 2 ed, Prentice-Hall, New Jersey.

Rabiner, L. R. & Schafer, R. W. 1978, "Linear Predictive Coding of Speech," in *Digital processing of speech signals*, R. W. Schafer & L. Rabiner, eds., Prentice Hall New York, pp. 396-405.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects, *Philosophical transactions of the Royal Society of London.Series B: Biological sciences*, **336**(1278): 367-373.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science*, **270**: 303-304.

Shannon, R. V., Zeng, F. G. and Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues, *Journal of the Acoustical Society of America*, **104**(4): 2467-2476.

Sidwell, A. and Summerfield, Q. (1985). The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise, *Journal of the Acoustical Society of America*, **78**(2): 495-506.

Skinner, M. W., Arndt, P. L. and Staller, J. S. (2002). Nucleus 24 advanced encoder conversion study: performance versus preference, *Ear and Hearing*, **23**(1): 2S-17S.

Skinner, M. W., Fourakis, M. S., Holden, T. A., Holden, L. K. and Demorest, M. E. (1996). Identification of speech by cochlear implant recipients with the Multipeak (MPEAK) and Spectral Peak (SPEAK) speech coding strategies I. Vowels, *Ear and Hearing*, **17**(3): 182-197.

Skinner, M. W., Holden, L. K., Holden, T. A., Demorest, M. E. and Fourakis, M. S. (1997). Speech recognition at simulated soft, conversational, and raised-to-loud vocal efforts by adults with cochlear implants, *Journal of the Acoustical Society of America*, **101**(6): 3766-3782.

Starr, A., Isaacson, B., Michalewski, H. J., Zeng, F. G., Kong, Y. Y., Beale, P., Paulson, G. W., Keats, B. J. B. and Lesperance, M. M. (2004). A dominantly inherited progressive deafness affecting distal auditory nerve and hair cells, *Journal of the Association for Research in Otolaryngology*, **5**: 411-426.

Summers, V. and Leek, M. R. (1994). The internal representation of spectral contrast in hearing-impaired listeners, *Journal of the Acoustical Society of America*, **95**(6): 3518-3528.

ter Keurs, M., Festen, J. M. and Plomp, R. (1992). Effect of spectral envelope smearing on speech reception. I, *Journal of the Acoustical Society of America*, **91**(5): 2872-2880.

ter Keurs, M., Festen, J. M. and Plomp, R. (1993a). Effect of spectral envelope smearing on speech reception. II, *Journal of the Acoustical Society of America*, **93**(3): 1547-1552.

ter Keurs, M., Festen, J. M. and Plomp, R. (1993b). Limited resolution of spectral contrast and hearing loss for speech in noise, *Journal of the Acoustical Society of America*, **94**(2): 1307-1314.

Terhardt, E. (1979). Calculating virtual pitch, *Hearing Research*, **1**(2): 155-182.

Terhardt, E., Stoll, G. and Seewann, M. (1982). Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions, *Journal of the Acoustical Society of America*, **71**(3): 671-678.

Throckmorton, C. S. and Collins, L. M. (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.

Tyler, R. S., Preece, J. P. and Lowder, M. W. (1987). The Iowa audiovisual speech perception laser videodisc, Laser Videodisc and Laboratory Report.

Van Tassel, D. J., Soli, S., Kirby, V. M. and Widin, G. P. (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.

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

Waltzman, S. B. and Cohen, N. L. (2000). *Cochlear implants* Thieme, New York.

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

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

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

Whitford, L. A., Seligman, P. M., Everingham, C. E., Antognelli, T., Skok, M. C., Hollow, R. D., Plant, K. L., Gerin, E. S., Staller, S. J., McDermott, H. J., Gibson, W. R. and Clark, G. M. (1995b). Evaluation of the Nucleus Spectra 22 processor and new speech processing strategy (SPEAK) in postlinguistically deafened adults, *Acta Oto-Laryngologica*, **115**(5): 629-637.

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

Whitlon, D. S. (2004). Cochlear development: hair cells don their wigs and get wired, *Current Opinion in Otolaryngology & Head and Neck Surgery*, **12**: 449-454.

Wilson, B. S., Lawson, D. T., Finley, C. C. and Wolford, R. D. (1991). Coding strategies for multichannel cochlear prosthesis, *American Journal of Otology Supplement*, **12**: 56-61.

Wilson, B. S., Lawson, D. T. and Zerbi, M. (1995). Advances in coding strategies for cochlear implants, *Advances in Otolaryngology - Head and Neck Surgery*, **9**: 105-129.

Yang, L. P. and Fu, Q. J. (2005). Spectral subtraction-based speech enhancement for cochlear implant patients in background noise (L), *Journal of the Acoustical Society of America*, **117**(3): 1001-1004.

Zeng, F. G., Grant, G., Niparko, J., Galvin III, J. J., Shannon, R. V., 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.

Zeng, F. G., Popper, A. N. and Fay, R. R. (2004). *Cochlear implants - Auditory prostheses and electric hearing* Springer, New York.

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics - facts and models*, 2nd ed, Springer Berlin.