UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

# OBJECTIVE DETERMINATION OF VOWEL INTELLIGIBILITY OF A COCHLEAR IMPLANT MODEL

by

Joe van Zyl

Submitted in partial fulfilment of the requirements for the degree

Master of Engineering (Bio-Engineering)

in the

Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

November 2008

**ACKNOWLEDGEMENTS**

To my mom, dad, brother and sisters for their unyielding support during a tough and prolonged journey. I have learned so much from your unending love and belief in me and I am a stronger person because of you.

**OBJECTIVE DETERMINATION OF VOWEL INTELLIGIBILITY OF A COCHLEAR IMPLANT MODEL by Joe van Zyl**

**Supervisor: Prof JJ Hanekom**

**Department of Electrical, Electronic and Computer Engineering**

**Master of Engineering (Bio-Engineering)**

## SUMMARY

The goal of this study was to investigate the methodology in designing a vowel intelligibility model that can objectively predict the outcome of a vowel confusion test performed with normal hearing individuals listening to a cochlear implant acoustic model. The model attempts to mimic vowel perception of a cochlear implantee mathematically. The output of the model is the calculated probability of correct identification of vowel tokens and the probability of specific vowel confusions in a subjective vowel confusion test. In such a manner, the model can be used to aid cochlear implant research by complementing subjective listening tests. The model may also be used to test hypotheses concerning the use and relationship of acoustic cues in vowel identification.

The objective vowel intelligibility model consists of two parts: the speech processing component (used to extract the acoustic cues which allow vowels to be identified) and the decision component (simulation of the decision making that takes place in the brain). Acoustic cues were extracted from the vowel sounds and used to calculate probabilities of identifying or confusing specific vowels. The confusion matrices produced by the objective vowel perception model were compared with results from subjective tests performed with normal hearing listeners listening to an acoustic cochlear implant model. The most frequent confusions could be predicted using the first two formant frequencies and the vowel duration as acoustic cues. The model could predict the deterioration of vowel recognition when noise was added to the speech being evaluated. The model provided a first approximation of vowel intelligibility and requires further development to completely predict speech perception of cochlear implantees.

**Keywords:** cochlear implant, speech intelligibility, acoustic model, confusion matrix, acoustic analysis, objective speech perception model, objective speech intelligibility model

**OBJEKTIEWE BEPALING VAN VOKAAL VERSTAANBAARHEID VAN 'N KOGLÊERE INPLANTINGMODEL deur Joe van Zyl**

**Studieleier: Prof JJ Hanekom**

**Departement Elektriese, Elektroniese en Rekenaar-Ingenieurswese**

**Meester van Ingenieurswese (Bio-Ingenieurswese)**

## OPSOMMING

Hierdie studie het ten doel gehad die ondersoek na die metodologie in die ontwerp van 'n vokaal-verstaanbaarheidsmodel wat die uitkoms van 'n vokaal-verwarringstoets kan voorspel wat toegepas is op normalhorende persone wat na 'n koglêere inplantingsmodel luister. Die model poog om die vokaalpersepsie van 'n persoon met 'n koglêere implanting wiskundig na te boots. Die uitsette van die model verteenwoordig die wiskundige waarskynlikhede vir die identifikasie van vokaalklanke deur 'n persoon met 'n koglêere implanting, en die waarskynlikheid van spesifieke vokaalverwarrings. Die model kan gevolglik gebruik word om koglêere prostese navorsing te baat deur subjektiewe gehoortoetse te vervang of aan te vul. Die model kan ook gebruik word om hipoteses te toets rakende die gebruik en verwantskap van akoestiese leidrade in vokaalherkenning.

Die objektiewe vokaal-verstaanbaarheidsmodel bestaan uit twee dele: die spraak prosseseringskomponent (word gebruik om die akoestiese leidrade te ontleed wat toelaat dat vokale geidentifiseer kan word); en die besluitskomponent (simulasie van die besluitneming wat in die brein plaasvind). Akoestiese leidrade is uit die vokaalklanke geneem en is gebruik om die waarskynlikhede vir die identifisering of verwarring van spesifieke vokale te bereken. Die verwarringsmatrikse wat voortgebring is deur die objektiewe vokaalwaarnemingsmodel is vergelyk met die resultate van die subjektiewe toetse toegepas op normaalhorende persone wat geluister het na die akoestise koglêere inplantingmodel. Die mees dikwelse verwarrings kon voorspel word deur die eerste twee formant frekwensies en die vokaaltydsduur te gebruik as akoestiese leidrade. Die model kon die verslegting in vokaalherkenning voorspel, wanner ruis toegevoeg is tot die spraak wat beoordeel moes word. Die model was 'n eerste stap na redelike vokaalverstaanbaarheid en benodig verdere ontwikkeling om die spraakpersepsie van persone met koglêere inplantings te voorspel.

**Sleutelwoorde:** koglêere inplanting, spraakverstaanbaarheid, akoestiese model, verwarringsmatriks, akoestiese analise, objektiewe spraak-verstaanbaarheidsmodel.

## List of abbreviations

| | |
|---|---|
| ASR | Automatic Speech Recognition |
| CI | Cochlear implant |
| CIs | Cochlear implants |
| CIS | Continuous Interleaved Sampling (Cochlear implant speech processing algorithm) |
| dB | Decibels |
| DTW | Dynamic Time Warping |
| F1 | Formant 1 Frequency |
| F2 | Formant 2 Frequency |
| FITA | Feature Information Transmission Analysis |
| HMM | Hidden Markov Model |
| Hz | Hertz |
| jnd | just noticeable difference |
| LPC | Linear Predictive Coding |
| MFCC | Mel Frequency Cepstral Coefficients |
| MPEAK | Multi Peak (Cochlear implant speech processing algorithm) |
| ms | millisecond(s) |
| pdf | probability density function |
| RMS | Root-Mean-Square |
| SNR | Signal to Noise Ratio |
| SPEAK | Spectral Peak (Cochlear implant speech processing algorithm) |
| TEC | Token Envelope Correlation |

# Table of Contents

**CHAPTER 1 INTRODUCTION 1**

1.1 BACKGROUND AND SCOPE OF WORK 1
*1.1.1 Physiology of Hearing 1*
*1.1.2 Cochlear Implants 2*
*1.1.3 Speech Perception of Cochlear Implantees 4*
*1.1.4 Evaluation of Speech Intelligibility of Cochlear Implantees 6*
1.2 APPROACH 8
1.3 HYPOTHESIS AND RESEARCH QUESTIONS 11
1.4 OBJECTIVES 12
1.5 OUTLINE 13

**CHAPTER 2 LITERATURE STUDY 15**

2.1 CHAPTER OBJECTIVES 15
2.2 INTRODUCTION 15
2.3 CUES USED FOR VOWEL IDENTIFICATION 16
*2.3.1 Formant Frequencies 16*
*2.3.2 Duration 18*
*2.3.3 Formant Movement 19*
*2.3.4 Consonant-Formant Transitions 19*
*2.3.5 Spectral Shape Features 20*
*2.3.6 Acoustic Cues Used in the Presence of Noise 21*
2.4 OBJECTIVE SPEECH EVALUATION METHODS 22
*2.4.1 Difference Measure Based Models 22*
Dynamic Time Warping (DTW) 23
Token Envelope Correlation (TEC) 25
Hidden Markov Model Based Model 26
*2.4.2 Feature Identification through the Neighbourhood Activation Model 28*
*2.4.3 Multidimensional Phoneme Identification (MPI) model 30*
*2.4.4 Neural Network Model 34*
2.5 GAPS IN THE CURRENT LITERATURE 35
2.6 SUMMARY 37

**CHAPTER 3 METHODS 38**

3.1 CHAPTER OBJECTIVES 38
3.2 INTRODUCTION 38
3.3 MATHEMATICAL MODELING OF VOWEL PERCEPTION BY COCHLEAR IMPLANTEES 40
3.4 DEVELOPMENT OF MODEL 42
*3.4.1 Processing Component 44*

3.4.1.1 Inputs 46
3.4.1.2 Down Sampling 46
3.4.1.3 Segmentation (Hanning Window) 47
3.4.1.4 Removing the Vowel From the Word 48
3.4.1.5 LPC Spectrum 50
3.4.1.6 Formant Tracking 55
3.4.1.7 Uncertainty Factors 60
Frequency Variance 62
Spectral Contrast 63
*3.4.2 Decision Component* *66*
3.4.2.1 Creation of 4D Gaussian Distribution Functions 71
3.4.2.2 Finding the Decision Axis 75
3.4.2.3 Find Gradient of Decision Plane 77
3.5 EXPERIMENTAL STUDY 81
*3.5.1 Listeners* *81*
*3.5.2 Stimuli* *81*
*3.5.3 Experimental Conditions Investigated* *82*
3.5.3.1 Acoustic CI Model 82
3.5.3.2 Background Noise Conditions 83
3.5.3.3 Evaluation of Results 84
*3.5.4 Summary* *84*

**CHAPTER 4 RESULTS** **85**

4.1 CHAPTER OBJECTIVES 85
*4.1.1 Formant Frequencies and Duration* *85*
*4.1.2 Accuracy of Acoustic Cue Tracking* *88*
*4.1.3 FITA Analysis* *90*
4.2 FREQUENCY VARIATION MODEL 92
*4.2.1 Speech Without Additional Background Noise* *92*
*4.2.2 Speech at 40dB SNR (Multi-Talker Babble Noise)* *99*
*4.2.3 Speech at 20 dB SNR (Multi-Talker Babble Noise)* *105*
*4.2.4 Speech at 0 dB SNR (Multi-Talker Babble Noise)* *112*
4.3 SPECTRAL CONTRAST MODEL 116
*4.3.1 Speech Without Additional Background Noise* *116*
*4.3.2 Speech at 40dB SNR (Multi-Talker Babble Noise)* *121*
*4.3.3 Speech at 20 dB SNR (Multi-Talker Babble Noise)* *128*
*4.3.4 Speech at 0 dB SNR (Multi-Talker Babble Noise)* *133*
4.4 SUMMARY 136

**CHAPTER 5 DISCUSSION** **138**

5.1 CHAPTER OBJECTIVES 138
5.2 EVALUATION OF THE OBJECTIVE MODELS 138

5.3 DIRECT COMPARISON BETWEEN THE TWO MODELS 141
5.4 RESEARCH INSIGHTS 149
5.5 COMPARISON WITH OTHER COCHLEAR IMPLANT RESEARCH 151
5.6 COMPARISON WITH OTHER MODELS 154
5.7 RESEARCH QUESTION FINDINGS 159

**CHAPTER 6 CONCLUSION 161**

6.1 FUTURE WORK 163

**REFERENCES 166**

# CHAPTER 1 INTRODUCTION

## 1.1 BACKGROUND AND SCOPE OF WORK

This dissertation investigates the methodology in developing an objective vowel intelligibility model which can predict vowel recognition and confusion of normal hearing individuals listening to a cochlear implant (CI) acoustic model. The model can be used to evaluate the effect of acoustic cue information on vowel identification and to predict confusions which might occur under specific testing conditions. Such a model can benefit cochlear implant development by substituting or complementing time-consuming subjective testing done with cochlear implantees.

Before discussing the issues concerning the evaluation of speech intelligibility for cochlear implantees however, it is necessary to provide the reader with background information on speech perception, cochlear implants, cochlear implant evaluation, and other related issues that are relevant for this study.

### 1.1.1 Physiology of Hearing

In a normal functioning ear, sound waves from the outer ear are transformed into mechanical displacement of the ossicular chain (malleus, incus and stapes) in the middle ear. (See Figure 1.1). The ossicular chain is connected to the oval window of the fluid-filled inner ear or cochlea and the displacement of the ossicles generates movement in the fluid. Resulting pressure variations cause displacement of the basilar membrane and subsequent deformation of the cochlear hair cells. The nature and magnitude of such deformation is determined by the spectral and temporal characteristics of the acoustic signal. Deformation of cochlear hair cells in turn initiate action potentials in associated nerve fibres, thereby encoding information regarding the original signal. Auditory perception is achieved as a result of the interpretation of these neural signals in the auditory cortex of the brain (Marieb, 2004).

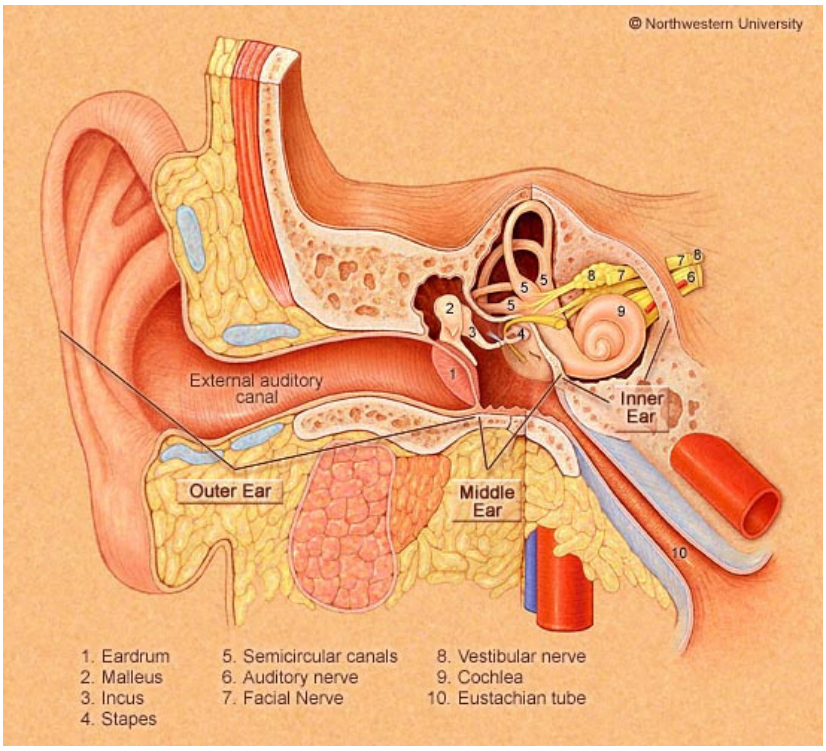


**Figure 1.1 Physiology of the human ear. Used with permission from http://www.tchain.com/otoneurology/testing/hearing_test.htm.**

### 1.1.2 Cochlear Implants

Although hearing impairments can result from damage to the auditory nerve, Hinojosa and Marion (1983) showed that the loss of cochlear hair cells, rather than the loss of auditory nerve fibres, is the most common cause of deafness. Loss of or damage to cochlear hair cells results in the failure of the auditory system to transform acoustic pressure waves to neural action potentials accurately, thereby causing hearing impairment (Ohlemiller and Gagnon, 2004). This is good news for the hearing-impaired: by inserting a prosthetic hearing device into the inner ear the auditory neurons can be stimulated directly (Loizou, 1999b; Whitlon, 2004). Today, cochlear implants allow profoundly deaf individuals to partially regain their sense of hearing (Loizou, 1998).

The objective of the cochlear implant is to mimic the function of a healthy cochlea (Clark, 2003; Loizou, 1998; Loizou, 1999a; Waltzman and Cohen, 2000). A cochlear implant

consists of an external and an internal part. The external unit of the device comprises a microphone, speech processor and a transmitting module. Figure 1.2 shows all the main components of the cochlear implant.


**Figure 1.2 ESPrit 3G and Nucleus 22 cochlear implant system. The components are: 1. The electrode array (which is placed in the inner ear). 2. The receiver for the electrode array. 3. The speech processor 4. Transmitting module and 5. Microphone (worn behind the ear). (Used with permission from http://www.cochlear.com)**

The microphone receives the acoustic signal from the environment and conveys it to the speech processor. The speech processor performs a complex signal processing strategy which includes signal analysis, compression and filtering. The reader is encouraged to consult Loizou (1998) for more information.

In simple terms the speech processor's main function is the separation of the input signal into its frequency components (Loizou, 1998), similar to the way a healthy cochlea would respond to acoustic input (Yost, 2006). Once the input has been separated into frequency bands, the speech processor calculates an indication of the energy within each band. The output signals are sent to the transcutaneous transmitting module, which allows signals to be transferred wirelessly to the internal components. The internal receiver-stimulator uses the energy of each filter band to modulate pulsatile signals that are applied to specific electrodes of the electrode array inserted into the scala tympani in the cochlea.

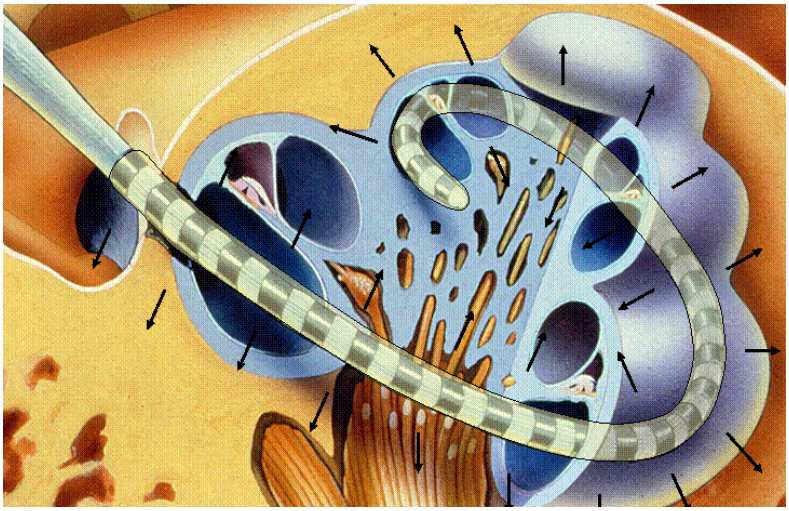

**Figure 1.3. Representation of how the electrodes of a cochlear implant fits into the cochlea. (With permission from http://www.sciencenow.org.au/bionic_ear/bionic_ear_institute.htm)**

Figure 1.3 gives a representation of how a typical electrode array fits into the cochlea. The electrode array consists of multiple electrodes, which, upon activation, stimulate specific regions of the auditory nerve fibres according to the spectral properties of the input signal. Successful functioning of the cochlear implant, therefore, requires sufficient auditory neuron survival in the vicinity of the inserted electrode array after damage to the cochlea or loss of neurons has occurred (Shepherd, Hatsushika and Clark, 1993; Shepherd and McCreery, 2006). Once stimulated, the nerve fibres fire and propagate neural impulses to the brain. The brain in turn interprets these pulses as sounds (Loizou, 1998).

### 1.1.3 Speech Perception of Cochlear Implantees

Although a cochlear implant enables a profoundly deaf individual to perceive speech to an extent that allows for meaningful participation in conversation, the quality of speech perception is not comparable to that of normal hearing listeners (Fu and Shannon, 1998; Fu, Shannon and Wang, 1998a). Cochlear implants have proven to be particularly successful in quiet environments, although there is still a lot of variation in performance between CI users. This variability may be due to patient-related factors (such as the type of hearing loss, the duration of deafness, the health and the location of the remaining auditory neurons, the insertion depth of the electrode array, amongst others) and/or to

processor-related factors (such as the number of electrodes/channels, the stimulation rate, the frequency-to-electrode allocation, etc.) (Liu and Fu, 2007). Cochlear implantees still have difficulty in recognizing speech in the presence of noise. Normal hearing listeners have the ability to mask background noise in order to still maintain a high level of speech perception. This can, in part, be attributed to parameters that affect spectral resolution, which is dependant on the number of electrodes of the cochlear implant, the restrictions of the bio-physical interface and the processing strategy (Fu and Shannon, 2000; Holden, Skinner, Holden and Demorest, 2002; Zeng, Grant, Niparko, Galvin III, Shannon, Opie and Segel, 2002).

Continuous studies are being conducted to determine if the acoustic cues in speech known to aid speech perception for normal hearing listeners are actually available to cochlear implantees. Although many studies have been completed and a vast array of knowledge has been compiled on speech perception, we do not know the relative relationship of acoustic cues and cochlear implant parameter settings that can optimally be employed by cochlear implantees to understand speech. Understanding the use of acoustic cues by cochlear implantees can aid in the improvement of current cochlear implant speech strategies. This can also be beneficial in other areas such as automatic speech recognition algorithms and speech synthesis.

In general, speech performance is strongly influenced by parameters that affect the spectral resolution (for example, the number of electrodes/channels). For both CI users and normal hearing subjects listening to an acoustic model of a CI, speech recognition improves with increasing numbers of spectral channels in the absence of any additional noise added to the signal (Liu and Fu, 2007).

### 1.1.4 Evaluation of Speech Intelligibility of Cochlear Implantees

Subjective listening tests are traditionally used to measure the effect on speech perception when changing processor parameters of a cochlear implant or manipulating the speech input signal. Subjective testing is based upon statistical information gathered from having human subjects evaluating sets of speech tokens (for example, groups of vowels, consonants or sentences). The output of these tests are often presented as confusion matrices or recognition trends against some parameter setting (Van Wieringen and Wouters, 1999). The results are valuable to researchers because they highlight information on speech perception that can be used for furthering cochlear implant development.

Speech recognition is measured through listening tests in a controlled environment such as a sound-proof room. These tests, however, require considerable effort and time in order to form a conclusive result. A number of cochlear implanted subjects are necessary with each individual spending up to a few hours listening to each condition of an experiment, which may include hundreds of repetitions of vowels, consonants and/or short sentences. In order to measure speech intelligibility on a regular basis, it is advantageous to avoid the time-consuming and expensive procedures of subjective determination of speech intelligibility.

In recent years, using acoustic models to test normal-hearing subjects for cochlear implant research is a widely used and well-accepted method for determining the effect of chosen parameters on speech intelligibility in cochlear implants (Dorman, Loizou, Spahr and Maloff, 2002; Svirsky, 2000; Van Wieringen and Wouters, 1999). An acoustic model of a cochlear implant is an algorithm that processes speech exactly like a cochlear implant processor, but, unlike the processor, additionally includes a model of the biophysical interface. It is used to present cochlear implant-like sound to normal-hearing persons. There are numerous advantages in using normal-hearing subjects listening to acoustic models in CI research. Normal-hearing subjects are more numerous and easier to recruit, the experimental setups tend to be less involved, and there are fewer subject variables (such as the experience of the user with cochlear implant devices, the type of implanted device, the cause of deafness, and the quality of implantation) that affect an individual

user's performance. These types of tests have been used in numerous studies and have been verified to provide good predictions of parameters measured with cochlear implantees in a number of instances (Fu and Shannon, 2000; Fu *et al.*, 1998a; Van Wieringen and Wouters, 1999).

Although acoustic models make research into speech perception and parameter settings easier, objective speech intelligibility prediction models can take this evolution one step further, by replacing the normal-hearing listener with an algorithm that listens to the output of the acoustic model. An objective speech prediction model attempts to mimic the transducing of acoustic information in the ear and the cognitive processing in decision-making of a listener identifying a vowel sound. Such a model predicts the psychophysical performance of a listener in vowel or consonant identification, and may give further insight into the acoustic cues hypothesized to be relevant. An objective speech quality or intelligibility measurement system cannot replace the human brain in the perception of speech, but it can give a good estimation of the possible reaction of a listener in a specific experiment as a first approximation.

Most speech intelligibility prediction models are usually only developed for listeners with normal hearing (Beerends, Hekstra, Rix and Hollier, 2002; Vainio, Suni, Jarvelainen, Jarvikivi and Mattila, 2005; Voran, 1999a). Only a small body of research is available that attempts to predict speech intelligibility for cochlear implantees (Remus and Collins, 2004; Svirsky, 2000). An objective speech evaluation method for CI users may save production development time in devising new speech processing algorithms by minimizing the time-consuming subjective testing that needs to be done in intermediate development steps. For every change made to a cochlear implant processor, subjective testing needs to be performed with cochlear implantees or normal hearing listeners using an acoustic model. An objective speech evaluation method could speed up the development process by cutting down on the time-consuming subjective evaluations.

The problem addressed in this thesis is the development of a methodology that may be used to predict vowel intelligibility of severely degraded sound (as would be received at the output of a CI acoustic model). Often speech recognition in cochlear implantees is

assessed through vowel and consonant confusion and sentence recognition tests (Loizou, Dorman, Poroy and Spahr, 2000; Loizou, Dorman and Powell, 1998; Loizou and Poroy, 2001b; Pretorius, Hanekom, Van Wieringen and Wouters, 2006; Van Wieringen and Wouters, 1999). The present study considers only vowel recognition. Although vowel and consonant intelligibility cannot be seen to be independent predictors of speech intelligibility, this is a first step towards developing algorithmic means of predicting speech intelligibility of cochlear implant speech. The eventual goal is not to replace subjective testing with testing using cochlear implantees completely, but, rather, to use an objective model for a first approximation of the possible results when developing new cochlear implant speech processing algorithms or testing a specific hypothesis. This method is developed by using an existing acoustic model (developed in our research group, but unpublished) for a cochlear implant together with signal processing and statistical theory. The rest of the chapter will be devoted to the elaboration of the problem, the approach followed in solving it, and the specific research questions considered in this study.

## 1.2 APPROACH

The model proposed in this study attempts to approximate the outcomes of vowel confusion tests that are used traditionally in experiments with profoundly deaf individuals (Blamey, Dowell and Brown, 1987; Ferguson and Kewley-Port, 2002; Fu *et al.*, 1998a; Pretorius *et al.*, 2006; Skinner, Fourakis, Holden, Holden and Demorest, 1996; Tyler, Tye-Murray and Otto, 1989; Van Wieringen and Wouters, 1999). The results of these tests provide information on the confusions that an individual might experience between different vowel or consonant sounds. This study will focus on developing an objective method that predicts the outcome of vowel confusion tests specifically. Figure 1.4 shows an outline of the approach followed in the development of the objective vowel prediction model.

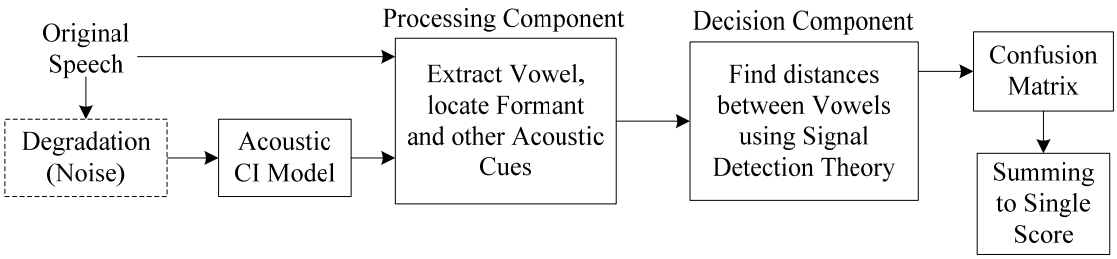


**Figure 1.4. Outline of the approach to the development of a method appropriate for quantifying vowel intelligibility of cochlear implantees.**

The input (pre-recorded vowel sounds) to the vowel perception model will first be processed through an acoustic model of a cochlear implant. The acoustic cochlear implant model transforms a sound signal through the same processing steps as in a cochlear implant. Biophysical interactions between simulated nerve fibres and the cochlear implant are also simulated by including models of these interactions in the acoustic model. This produces an output simulating the speech as heard by a cochlear implant user. The acoustic model (unpublished) used for this study was developed in our research group. Given the highly degraded nature of the speech at the output of the acoustic model, an unprocessed version of the vowel token is also used as a input for reference purposes. (This is explained in more detail later in this document.)

Various methodologies are implemented to predict speech perception in current objective perceptual models. Most of these models are comprised of two components: The processing component and the decision component. The processing component either extracts some information from the signal for evaluation or uses various signal processing transformations to prepare the speech token for evaluation. Perceptual processes that are most important to vowel or consonant discrimination, such as masking and loudness mapping, are usually implemented (Beerends *et al.*, 2002; Voran, 1998). The present study focuses on acoustic cues that listeners use to identify vowel sounds and that, therefore, are assumed to be important in vowel confusion experiments. These cues are extracted from each of the vowel sounds and used to generate a three dimensional vowel space. The features in the vowel sound that mask these acoustic cues are also extracted and used as uncertainty factors. The larger the uncertainty factor is, the larger the probability of confusion of vowels is going to be.

The decision component normally performs calculations in the form of spectral analysis or a difference measure between a reference token and a degraded token (Beerends *et al.*, 2002; Remus and Collins, 2004; Svirsky, 2000; Voran, 1999c). Various methods such as neural networks or Hidden Markov Models (HMMs) have been used to improve the accuracy of results obtained in speech perception models (Chang, Anderson and Loizou, 2001; Remus and Collins, 2005). In the present study, signal detection theory is utilized to determine the perceptual distance between tokens. The measured information of the acoustic cues and the uncertainty factors are used to calculate multidimensional probability density functions in order to predict the probability of confusing the different vowel sounds with each other. The probability of confusing each vowel with all other vowels is calculated and documented in a confusion matrix.

Most other speech evaluation models provide a single score (percentage correct) as an output (Beerends *et al.*, 2002; Rix, Beerends, Hollier and Hekstra, 2001; Thorpe and Wonho, 1999a; Voran, 1999a; Werner, Kamps, Tuisel, Beerends and Vary, 2003); the model in this study, however, attempted to produce an entire confusion matrix. Only two published models appeared to be available at the time of writing that produced results in a similar way (Remus and Collins, 2004; Svirsky, 2000). The predicted confusion matrix is used to approximate specific vowel confusions and the percentage correct answers that a cochlear implantee might produce for each vowel token.

The confusion matrices obtained from the new objective method were evaluated and compared to subjective test results through information transmission analysis using the same speech tokens as input. Information transmission analysis is used to determine how well acoustic cues were transmitted to the listener (Miller and Nicely, 1955). The tests were first completed with noiseless speech tokens and then with added multi-speaker babble background noise at different Signal to Noise Ratios (SNRs). A comparison is also made between each of the different tests using information transmission analysis. The variation of the percentage of correct answers as the SNR changes is also compared between the prediction model and the subjective tests. This is done to determine whether the model followed the trends of actual subjective tests at different SNRs which simulate

real world scenarios.

## 1.3 HYPOTHESIS AND RESEARCH QUESTIONS

The hypothesis in the present study is that an objective vowel intelligibility prediction model can be used to give an approximation of how well implantees will perform in subjective listening tests and of how well listeners with normal hearing will perform when listening to vowels degraded by an acoustic model of a cochlear implant. This study tested this hypothesis specifically for a objective vowel intelligibility model that is based on some of the acoustic cues generally accepted to be used by the listeners to recognize vowel sounds.

The results of this new model are compared with results of subjective vowel confusion tests conducted on normal hearing listeners listening to an acoustic cochlear implant model. Such a test gives developers insight into not only the performance of a cochlear implant speech strategy but also into which types of signal information are transmitted to the electrically-stimulated auditory system. Once the model is refined, these confusion results can aid in setting specific parameters of a cochlear implant for a specific individual and provide speech perception information for researchers.

The second hypothesis tested in this study is that the general trend of a vowel confusion matrix can be predicted by the vowel prediction model. A reliable prediction of this trend is more desirable than merely predicting a percentage of correct answers.

Furthermore, this study investigates the trend of vowel prediction by the model when background noise is added to the input vowel tokens. The hypothesis is that the vowel tokens are recognized more poorly than in high fidelity speech because the acoustic cues are masked by the noise. This deterioration can be predicted by selecting the appropriate cues and using signal detection theory to calculate a probability. The model will measure the disintegration of the spectral qualities of the vowel token as background noise is added, especially in terms of the lack of spectral contrast which has been shown to cause

ambiguity in speech recognition (Leek, Dorman and Summerfield, 1987; Leek and Summers, 1996b; Loizou and Poroy, 2001a). The new objective method should be able to predict this decline at different signal to noise ratios.

Specifically, the research questions investigated in this dissertation are:

- Will an objective vowel intelligibility model, implementing acoustic cue analysis, provide an approximation to the percentage of correct answers by normal hearing listeners listening to an acoustic CI model in subjective speech evaluation tests?
- Can such a model also predict the most frequent confusions made by these listeners?
- Will such a model adequately predict the deterioration of these results in the subjective evaluation test when noise is added to the vowel tokens?
- If uncertainty factors are used to calculate standard deviation in the vowel confusion probabilities, which uncertainty factor will perform better in predicting subjective testing results?

## 1.4 OBJECTIVES

The overall objective of this study is to build upon existing speech prediction models to aid engineers and scientists in cochlear implant speech processor design. Techniques were developed to assess the speech intelligibility of the vowel sounds received by listeners listening to an acoustic CI model. To achieve these goals, speech processing and statistical tools were utilized along with an existing acoustic CI model to form a new objective vowel prediction model.

More specifically this study had two primary objectives. The first objective was to determine which acoustic cues are used by listeners to identify degraded vowel sounds. Previous studies performed with cochlear implant users were researched to establish which acoustic cues in speech aid vowel identification. The research study also explores the methodologies and techniques used by current objective speech prediction and recognition

models. In addition to this, an investigation was done to identify factors that may contribute to vowel confusions for cochlear implant users.

The second objective concerns the development and evaluation of a new objective vowel prediction model for cochlear implant research. In particular the objective is to develop an algorithm that (i) extracts important acoustic cues from degraded vowel sounds, (ii) calculates the uncertainty level of the acoustic cue being used for identification, (iii) uses statistical analysis to form a confusion matrix which approximates the results of a subjective speech evaluation test, and (iv) compares these results with those obtained in subjective tests for high-fidelity and noise-degraded speech.

## 1.5 OUTLINE

In the following chapters, the process for the development of a new objective vowel prediction model for cochlear implantees is given. Before any development on the model could be done, a thorough background study was conducted. This investigation is summarized in the literature study in chapter 2. Speech perception is discussed, especially the cues used by cochlear implantees to identify a vowel sound. Current speech recognition and prediction models are also studied, including the methodologies that have been followed in their implementation, and the results produced by these methods. From the literature study, opportunities are identified for the development of a new vowel prediction model.

The methodology in developing an objective vowel intelligibility prediction model is described in chapter 3. All the individual steps implemented in the method are described in detail and the reasoning behind the choice of the specific implementation is also given. The new method is based on vowel confusion tests which are commonly performed with cochlear implant users. The output of this method is no longer a single score but, rather, a confusion matrix which shows the uncertainty an individual would have in identifying presented vowels.

The results obtained from the model developed in chapter 3 are reported and evaluated in chapter 4. These results are compared to results gained from subjective confusion listening tests. Various analyses are performed on the confusion matrices to determine how much information transmitted by the acoustic cues are used by the new model. The results are then compared to subjective testing with simulated cochlear implantees, i.e. normal hearing participants listening to an acoustic CI model. The experiments are first performed with clean speech (no noise added) followed by speech with different levels of background noise added.

The relation of this study to the literature available at present is discussed in chapter 5. Insights gained into various aspects of speech perception are reported and the implications of what was learnt from this study are discussed. Conclusions are drawn from the results achieved and the objectives that were met by the completion of this study are summarized. The contribution of this study to the current state of literature is also presented in this chapter.

Finally, in chapter 6, the study and all important findings are summarized. Possible improvements that can be made to the model, as well as any suggested studies that might follow from this dissertation brings the study to a conclusion.

# CHAPTER 2 LITERATURE STUDY

## 2.1 CHAPTER OBJECTIVES

The previous chapter outlined the study. The problem introduced was how to predict objectively speech intelligibility of individuals fitted with cochlear implants. 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 information on the mechanisms that listeners use to identify vowel sounds and how these and signal detection techniques have been applied to current objective intelligibility prediction models.

## 2.2 INTRODUCTION

A vowel intelligibility prediction model aims to predict how a listener, or in this study's case a cochlear implantee, would identify a vowel sound. The goal of this study is to learn more about how acoustic cues are used together to determine the identity of a vowel sound and to do an investigation into how current vowel intelligibility models are implemented.

The first part of the literature study focuses on what the acoustic cues in vowel sounds are that are thought to be used by normal hearing and cochlear implanted listeners. It will also focus on how these cues work together to form a perceptual vowel space in order for a listener to make a decision in interpreting the sound.

Since cochlear implantees have difficulty in discriminating speech in the presence of noise the effect of noise on discrimination of sound cues will be investigated. This study is extended to include other factors that influence confusion of vowel sounds and to investigate why vowel discrimination deteriorates in the presence of noise. Specific questions are: Is there spectral or temporal information in a vowel sound that assists a normal-hearing individual in speech perception under noisy conditions that is not available

to cochlear implantees? How can the deterioration of vowel intelligibility with added background noise be captured into an objective evaluation method? Answers to these questions can be incorporated into a vowel intelligibility model in order to improve predictions of the model in the presence of additional noise.

In general, objective speech evaluation tests have been developed to refine subjective testing or to replace subjective tests completely, to standardize results and speed up the testing procedures. The literature study considers techniques used in current speech intelligibility prediction methods. This study examines speech evaluation methods designed to predict speech intelligibility for cochlear implantees. It also considers what types of results these methods provide.

It will become apparent why research into a vowel intelligibility model for cochlear implantees is necessary. The study will also show how such a method can aid the improvement of cochlear implants.

## 2.3 CUES USED FOR VOWEL IDENTIFICATION

Many speech quality or intelligibility methods use signal processing on the acoustic signal as a means to measure speech quality or intelligibility. The present study, however, attempts to create a vowel intelligibility model that emulates the psychoacoustic processes that a listener would use to interpret a vowel sound. First, the mechanisms underlying human vowel recognition are considered.

### 2.3.1 Formant Frequencies

It has long been recognized that a primary acoustic cue aiding a listener in identifying a vowel sound are peaks in the spectrum called formants (Hillenbrand, Getty, Clark and Wheeler, 1995; Kewley-Port and Watson, 1994; Kewley-Port and Zheng, 1999; Liu and Kewley-Port, 2004b). The original study conducted by Peterson and Barney (1952) is still one of the most frequently quoted articles on the perception of vowels. Peterson and Barney measured the formant frequencies of vowel sounds and presented these vowels to

listeners in order to determine how vowel identification is correlated to the formant frequencies. The noise conditions of the sound recordings were not documented. The frequency measurements of the first two formants, Formant 1 (F1) and Formant 2 (F2), were taken at a single time slice of a vowel sound as spoken between two consonants 'h' and 'd.' An average of these formant frequencies was used to create a F1-F2 vowel space which showed the separation of vowels based on their formant frequencies. The results of the measurement study showed a strong relationship between the results of the listening test and the separation by formant frequencies. The overall error rate when listening to the vowel sounds was 5.6% and nearly all confusions involved confusions between adjacent vowels in the vowel space (Peterson and Barney, 1952).

This study was repeated and expanded by Hillenbrand *et al.* (1995). The researchers came to the same conclusion that formant frequencies are the most important factors in the recognition of vowel sounds. Hillenbrand and colleagues used a large group of speakers to take measurements of vowel duration, F0 contours, and formant frequencies. It was however found that the F1-F2 perceptual space was more crowded (vowels lay closer together) than the initial study by Peterson and Barney (1952). They concluded that formant frequencies are not the only acoustic cues used since a few of the vowels were well recognized in spite of lying very close together in the F1-F2 space. It could also mean that the vowel space is not orthogonal as assumed in the study.

The study performed by Remez, Rubin, Pisoni and Carrell (1981) supported the formant frequency theory and another approach of testing the hypothesis. Their study showed that speech synthesized from formant frequencies only (in terms of three-tone sinusoidal replicas) had sufficient information to convey a vowel's identity, despite removal of the rest of the spectral information.

There can be as many as five formants in a vowel sound, although it is still widely accepted among researchers that the frequencies of the lower formants (F1, F2, and sometimes F3) are the most important acoustic cues in identifying a vowel correctly, also by cochlear implantees (Delgutte, 1984; Greenberg, Ainsworth, Popper and Fay, 2004; Klatt, 1982; Summerfield and Assmann, 1989; Van Wieringen and Wouters, 1999). Van

Wieringen and Wouters (1999) specifically showed that the primary acoustic cues used by Laura cochlear implantees to identify vowels are the F1 frequency and vowel duration. This was done by presenting vowel sounds to twenty-five Laura cochlear implantees and analyzing the stimulus-response confusion matrices in terms of information transmission scores.

### 2.3.2 Duration

Although the spectral properties in the form of formant frequencies are considered to be the primary acoustic property, other properties have also been found to aid vowel discrimination. Results by Hillenbrand *et al.* (1995) showed that inclusion of duration in the parameter set resulted in consistent improvements in performance, especially when used only with the lower formant frequencies.

Similar findings were reported by Hillenbrand, Clark and Houde (2000). The results showed a modest but consistent improvement in classification accuracy with the addition of duration measures. In the study it was found that vowels were recognized well even when their original duration was altered (decrease of 5% accuracy), although there were a number of vowel sounds that were severely affected by changes in their duration. A study of Australian English vowels by Watson and Harrington (1999) showed an improvement in classification accuracy when duration measures were used to augment formant testing.

Recent experiments have shown that normal-hearing individuals and cochlear implant users make similar use of duration as cues for vowel discrimination (Iverson, Smith and Evans, 2006). The study by Van Wieringen and Wouters (1999) concluded that duration along with the F1 frequency are the most important acoustic cues used by Laura cochlear implantees for vowel identification.

### 2.3.3 Formant Movement

Formants are not always steady throughout the duration of a spoken vowel. In some vowel sounds there exists some formant movement in frequency over the duration of the sound. The effect of formant contour was tested by Hillenbrand and Nearey (1999). They asked listeners to identify naturally produced nonsense 'h'-vowel-'d' words using two synthetically generated versions. One set of synthesized signals was generated using the original measured formant contours and a second set of signals was synthesized with constant flat formants. When no additional noise was incorporated into the stimuli presented to the listeners, it was found that vowel recognition accuracy declines by about 15 to 23 percentage points when vowel formant movement is flattened in synthesized or signal processed speech (Assmann and Katz, 2005; Hillenbrand and Nearey, 1999). It has also been shown that vowels can be recognized even when the relatively steady-state portions, where the formant frequencies meet their targets, have been removed (Strange, 1989).

Iverson and colleagues (Iverson *et al.*, 2006) performed experiments with post-lingually deafened cochlear implantees to test their use of formant movement and duration as acoustic cues. The study suggested that patients with cochlear implants use formant movement and duration cues to the same extent as do normal-hearing listeners. Experiments showed that removing both formant movement and duration reduced vowel recognition accuracy for cochlear implant users by an average of 29.4%. In a second experiment, cochlear implantees were asked to rate words that had modified combinations of formant frequencies, formant movement and durations. It was found that the cochlear implantees' preferences for those secondary acoustic cues were less consistent than for the F1 and F2 frequencies.

### 2.3.4 Consonant-Formant Transitions

Studies have shown that normal-hearing listeners can correctly identify the vowel sound between two consonants even when the central part of the vowel is silenced (Jenkins, Strange and Edman, 1983; Strange, 1989). This is so because there exist transitions to and

from formants to consonants which can give indication to listeners as to the identity of the vowel sound. Along with the duration cue it was found that listeners could identify vowels at almost 100% accuracy. Even when the duration cue was removed (that is, the consonant transitions linked to each other), vowel identification was still 70% accurate.

Since these formant transitions happen very rapidly, it seems uncertain whether a cochlear implantee would be able to use these transitions as acoustic cues for vowel identification. It has, however, been established that cochlear implant users are able to recognize vowels above the chance level (averages being between 40 and 50%) based only on consonantal formant transitions (Kirk *et al.*, 1992). It must be mentioned that removing these formant transitions had no effect on vowel recognition.

### 2.3.5 Spectral Shape Features

Although formant frequencies are seen as the essential cue for vowel perception, there exist research results that suggest that the whole spectral shape (not just the formant peaks) is used for vowel identification (Zahorian and Jagharghi, 1993). Ito and co-workers (Ito, Tsuchida and Yano, 2001a) performed experiments to test this hypothesis. In the first experiment, they suppressed the first and second formant frequencies to determine whether this suppression would change vowel identification. The results proved to be very close to the results obtained using the normal spectrum. This implies that formant frequencies are not exclusive cues for vowel perception.

Experimentation by Beddor and Hawkins (1990) showed that spectral shape was an important cue if the lowest spectral prominence is weak. This was confirmed in the second experiment by Ito *et al.* (2001a) in which the frequencies below 1250 Hz were varied to change the amplitude ratio between high and low components of the spectrum.

Hillenbrand, Houde and Gayvert (2006) compared the effects on speech intelligibility between reconstructed speech based only on the distribution of spectral peaks and speech in which they had preserved the fine details of the spectral shape. In general the

information conveyed by the peaks-only test was similar to the detail-preserving test. There was however a 5 to 6 percentage point advantage to the detail-preserving test. They concluded that these results provided some support for the cochlear implant strategies such as MPEAK (Multipeak) and SPEAK (Spectral Peak) which rely on transmitting primarily the high energy components of the spectrum.

### 2.3.6 Acoustic Cues Used in the Presence of Noise

Many of the studies that have been used to determine which acoustic cues listeners use have been performed in sound-proof laboratory environments. This practice, however, does not give a true reflection of speech recognition in everyday life. Noise still poses a problem for cochlear implantees (Fetterman and Domico, 2002; Kiefer, Müller, Pfennigdorff, Schön, Helms, Von Ilberg, Baumgartner, Gstöttner, Ehrenberger, Arnold, Stephan, Thumfart and Baur, 1996; Skinner, Clark, Whitford, Seligman, Staller, Shipp, Shallop, Everingham, Menapace, Arndt, Antogenelli, Brimacombe, Pijl, Daniels, George, McDermott and Beiter, 1994), so it is important to look at speech recognition studies that have been conducted in the presence of additional noise. These studies give important information on how acoustic cues are used in noisy conditions.

In the study by Parikh and Loizou (2005), the effects of multi-talker background noise on the spectral shape of speech was investigated. The research showed that noise mostly affects the mid-frequency ranges (defined in their study as 1–2.7 kHz) and that listeners rely heavily on the F1 frequency to identify a vowel. Liu and Kewley-Port (2004a) tested vowel recognition in the presence of noise by manipulating the formant frequency, signal-to-noise ratio, and noise type. Results suggested that formant discrimination was significantly influenced by all three factors. The masking caused by noise showed significant decrease in formant discrimination even for normal hearing listeners.

The various studies discussed above showed that there are a number of acoustic cues that aid listeners in identifying a vowel sound. The lower formant frequencies, especially F1 and F2, are generally regarded as the primary acoustic cues in vowel identification. This holds true for normal hearing listeners as well as cochlear implantees. Studies have also

shown that duration is also an important acoustic cue used by cochlear implantees to identify vowels. The primary acoustic cues have been shown to be important for vowel identification even in the presence of noise (Parikh and Loizou, 2005). For this reason these cues have been selected as the primary input for decision making in the objective model. Secondary cues have also been found to aid vowel identification: these include formant glides (Hillenbrand and Nearey, 1999), formant-consonant transitions (Kirk, Tye-Murray and Hurtig, 1992) and overall spectral shape (Hillenbrand *et al.*, 2006). However, these secondary acoustic cues were not included in the model developed in this study. The objective vowel prediction model therefore extracted the F1, F2 and duration acoustic cues from vowel tokens being evaluated to approximate cochlear implantee vowel identification.

## 2.4 OBJECTIVE SPEECH EVALUATION METHODS

In addition to the above discussions on vowel identification cues, objective vowel intelligibility models have been developed that predict speech intelligibility in normal hearing and hearing impaired individuals. The rest of this chapter will look at the techniques used in these models and their performance in predicting results from subjective tests.

Objective vowel prediction models typically measure and process some property of the input signal. Thereafter a decision component produces an outcome through measurement or calculation. This study focused especially on the decision components used in these models. The remaining part of the literature study will concentrate on the methodology of published objective models and the mechanisms used to produce an outcome. The outputs and performance of these methods will also be considered.

### 2.4.1 Difference Measure Based Models

Three methods for predicting patterns of consonant and vowel confusion were developed by Remus and Collins (2004). The methods were based on signal processing techniques that calculate a probability per speech token; these are then used to find a quantitative

difference between speech tokens, and are tested using listening test results. The study by Remus and Collins is one of only two prediction methods (found at the time of writing this thesis) that attempted to provide the same type of results (in the form of confusion matrices) as the present study. The other study by Svirsky (2000) will be discussed later in this chapter.

The study by Remus and Collins (2004) evaluated the following three techniques in speech prediction:

1. Dynamic Time Warping (DTW) - using cepstrum representation.
2. Token Envelope Correction (TEC) - using the discrete envelope of the signal.
3. Hidden Markov Model (HMM) - using cepstrum representation and HMM

A brief description of the methodology used in each of these methods will be described. This is followed by a discussion of the results produced by these methods when compared to vowel confusion tests with normal hearing listeners through a CI acoustic model. The three prediction methods used metrics calculated as some measure of similarity or distance between two speech tokens, the stimulus and possible response. These were used to generate a complete confusion matrix as opposed to a single score produced by other models.

**Dynamic Time Warping (DTW)**

The dynamic time warping method creates a confusion matrix by finding the Euclidean distances between the cepstrum coefficients of the stimulus and the response. The Euclidean distance is calculated for all possible responses to complete the confusion matrix. Cepstrum coefficients can be seen as information about the rate of change in different spectrum bands. Mel-frequency cepstral coefficients (MFCC) has been proved to be very successful and is a popular front-end feature extraction method for the automatic speech recognition (ASR) field (Han, Chan, Choy and Pun, 2006; Skowronski and Harris, 2002; Zheng, Zhang and Song, 2001). Mel cepstrum coefficients are generally calculated generically as follows (Zheng *et al.*, 2001):

1. Calculate the Fourier transform for time windows of the speech signal.

2. Map the log amplitudes of the spectrum to the Mel-scale, using triangular overlapping windows.

3. Calculate the Discrete Cosine Transform of the list of Mel log-amplitudes.

4. The MFCC's are the amplitudes of the resulting spectrum.

In the DTW method, the cepstrum coefficients of the two tokens (for the stimulus and response) were used to create a prediction confusion matrix. The (ith, jth) entry in the prediction metric matrix is the value of the minimum cost mapping through a cost matrix of Euclidean distances between the cepstrum coefficients of the ith given token and the jth response token. To calculate the (ith, jth) entry in the prediction metric matrix, the cepstrum coefficients are computed from energy-normalized speech tokens.

The DTW method focuses on the cepstral properties of the speech signal – an aspect that seems important for the recognition of speech signals. It was assumed that using the cepstral properties could potentially be a good predictor of speech recognition. By using cepstrum coefficients the speech tokens are each assigned a location in a vowel space after which Euclidian distances are calculated in order to determine confusions between the tokens. A possible drawback in this method could result from the fact that no other known cues responsible for vowel and consonant recognition (like duration, formant movement, etc.) could be integrated in the model, since the location of each token in the vowel space is determined only by its cepstrum coefficients.

Cepstrum coefficients were also used in a later study by Liu and Fu (2005) to compare perceptual space successfully to a prediction acoustic vowel space for cochlear implantees. A vowel space is Cartesian (or geometric) coordinate system based representation of the relationship between vowel sounds. In Liu and Fu's model the perceptual space is defined as the representation of the relationship between vowels that a person uses to identify a vowel sound. The acoustic vowel space is the vowel space as measured from specific acoustic properties in the vowel sound. Liu and Fu's results suggested that acoustic distance between phonemes may well predict recognition performance of spectrally

degraded speech.

The aim of Liu and Fu's study was to create a model that could predict, automatically, the level of speech recognition, without the need of cochlear implantees or normal-hearing listeners using an acoustic CI model. An acoustic model was used to test the vowel recognition performance of normal hearing subjects for five varied speech processor parameters (spectral channels, amount of spectral shifting, amplitude mapping and the degree of spectral smearing and warping). The acoustic Euclidian distances between these vowel signals for each simulated processor condition were then calculated by using MFCCs to determine the location of each vowel in the vowel space. The acoustic vowel space was compared with the normal hearing subject's vowel recognition performance by using linear regression analysis. The outcome revealed that the predicted and actual obtained results were highly correlated when the number of spectral channels and amount of spectral smearing was varied. The results indicate that measuring acoustic space using dynamic time warped Mel-cepstrum coefficients could predict perception data.

**Token Envelope Correlation (TEC)**

For the Token Envelope Correlation, each entry in the prediction confusion matrix is the normalised inner product of the discrete envelopes of the stimulus and response. The stimulus is the token being presented and the current response is being calculated. The discrete envelope is obtained by first passing the signal through a high-pass equalization filter with cut-off at 1000 Hz prior to an anti-aliasing low-pass filter with cut-off frequency at 11 kHz. The signal is separated into either eight or twenty frequency bands or channels, after which the envelope is extracted by full-wave rectifying of each channel and passing it through a low pass filter at 110 Hz. The discrete envelopes of the two tokens are then aligned using dynamic time warping (or minimum cost path, as in the DTW method). After alignment the final value of the prediction confusion matrix can be calculated as:

$$M_{i,j} = \frac{x_i^T s_j}{\sqrt{x^T x_i}\sqrt{s^T s_j}}, \tag{2.1}$$

where $M_{i,j}$ is the prediction metric between the stimulus $i$ and response $j$, $x_i$ is the discrete envelope of the stimulus $i$ and $s_j$ is the discrete envelope of the response $j$.

The TEC method consistently underperforms to predict for three tests performed by Remus and Collins. The TEC method seemed to lack some basic information regarding cues that are important for vowel and consonant recognition. By analysing temporal cues only, the TEC method did not provide accurate results. This is to be expected since most of the cues mentioned previously in the literature study are based on the spectral properties of the speech signals.

**Hidden Markov Model Based Model**

In the third method, the prediction confusion matrix is calculated with Hidden Markov Models by using the Mel-cepstrum coefficients to represent each stimulus in a statistical fashion. Hidden Markov Models serve as a theoretical basis in a variety of applications and are especially useful in speech recognition (Rabiner, 1989). The Markov model uses a sequence of states to describe an observation. Using HMMs, each entry in the prediction metric matrix is the log-likelihood that the cepstrum of the given token is the observation produced by the HMM for the cepstrum of the response token. Training sets were used to obtain all the probabilities in the HMM. A training set of 100 speech samples was used to train each HMM, while the HMMs that were used consisted of three states each. No information is given in the article on the parameters used to maximize the probability of the observation sequence. The final entries in the prediction confusion matrix were described as the log likelihood of the model for each token and observation.

The effectiveness of each of the three prediction methods described above was evaluated by comparing the results obtained by these methods to results obtained from listening experiments done with twelve normal hearing subjects. These tests were done with noisy vowel and consonant speech tokens processed through two different cochlear implant acoustic models, each imitating the CIS and SPEAK processing techniques, respectively.

Three tests were done by Remus and Collins (2004) to determine which method, if any, would give an accurate prediction of vowel and consonant recognition. In the first test, successful near predictions were determined by measuring the most and least frequent incorrect responses. A near prediction for the vowel sounds was defined as the case where one token in a set of two MFIRs matches one token in the predicted set of two MFIRs. Where one speech token in the most or least frequently incorrect responses corresponded to a token in the predicted most or least frequently incorrect responses, it was classified as a successful prediction. This satisfied the study's objective of predicting certain patterns in the confusions. In the second test, each method was evaluated by its ability to predict the percentage of correct responses, as represented by the main diagonal of the confusion matrices. This was done by ranking the responses from least to most recognised. The third test evaluated the differences between the predicted and subjectively tested confusion matrices in terms of correctly identified speech tokens.

For the first test the DTW method performed the best (with accuracy of a 78% compared to the subjective vowel tests described earlier) with the HMM method performing at a similarly high level. The TEC method consistently underperformed. Linear regression showed that the HMM method performed very well for vowel recognition ranking (96%). The TEC and DTW methods scored poorly for the test. For the third test, DTW was the only method that appeared to have any success predicting the correct identification trends for different token sets (that is, different SNR levels). The predicted trends for the TEC and HMM methods did not accurately indicate the trends in the listening tests at different SNRs (Remus and Collins, 2004).

Failure of the TEC at the first task supported the conclusion that strictly temporal representations lack sufficient distinguishing characteristics. Since the HMM method performed very well on the first and second tasks, its failure in the last test was unexpected. Overall, it appeared that speech recognition predictions were more accurately made by implementing cepstral representations of the signals compared to temporal

envelopes. This confirms the information from the literature that the formant frequencies in vowel signals (and, therefore, the spectral information) are important for vowel recognition.

The three methods by Remus and Collins do provide some form of prediction of vowel confusions and correct identification of vowel sounds. The techniques used in their methods do not correspond with the objectives of the present study. The two methods which provided relatively successful results were based on cepstrum coefficients calculated from the speech. Cepstrum coefficients however do not provide information on how the acoustic cues are used by listeners to identify the vowel sounds.

Similarly the training of models using HMMs can improve the results produced by the model, but provides a statistical approach to predicting outcomes. This approach does not aid researchers in learning more about the use of cues in the identification of speech. The present study attempts to predict the vowel sound using the same acoustic cues shown in research to aid vowel identification and by doing so is aimed to provide information on speech perception.

### 2.4.2 Feature Identification through the Neighbourhood Activation Model

Luce and Pisoni (1998) researched human spoken word recognition and the relationship between sound patterns of words in memory and the effects of these relations on spoken word recognition. Their assumption was that the internal recognition system of the listener was a noisy system. This 'noisy' internal representation in memory implies that there are a number of words that are phonetically similar to the given stimulus word. These similar phonemes fall into a group called the similarity neighbourhood. The intelligibility of words is affected, therefore, by both the number of possible confusions, as well as the frequency of these words. The authors believed that the models of spoken word recognition at the time did not include an important technique which may improve lacked the importance of structural organisation of acoustic-phonetic patterns in the mental lexicon. For instance, when a person hears a word that might be any one of three, the person will choose the

word that is more predictable (used more often) in the language. The researchers found it unlikely that word recognition is accomplished by direct access to the acoustic phonetic representation in memory.

Based on this theory, Luce and Pisoni (1998) developed a spoken word recognition system called the Neighbourhood Activation Model (NAM) which could predict how listeners interpreted spoken words. The model focused primarily on structural issues concerning the process of lexical discrimination (taking acoustic phonetic properties and frequency of words into account). Equations expressed in terms of various probabilities were developed which took into account stimulus word intelligibility, stimulus word frequency, neighbourhood confusability, and neighbourhood frequency. The confusability of individual speech sounds was determined from confusion matrices for all initial consonants, vowels, and final consonants. Further information of these equations and the processing of the confusion matrices can be found in Luce and Pisoni, (1998) and Pisoni, Nusbaum, Luce and Slowiaczek (1985).

In a study by Meyer, Frisch, Pisoni, Miyamoto and Svirsky (2003) the NAM model was used to predict word recognition of postlingually deafened adults after cochlear implantation. The goal of the study was to use the model to gain insight into the psychoacoustic processes used by cochlear implant users in recognizing spoken words. Confusion tests from individual cochlear implantees were used to train the model. The probability of correctly identifying the stimulus word was based on the phoneme confusion probabilities and the relative frequency of occurrence of the target word compared with similar sounding neighbours.

The NAM model predicted word recognition at similar levels as actual cochlear implant users. The study concluded that these listeners use word frequency of occurrence and word similarity information to identify spoken words in a manner that is fundamentally similar to the way listeners with normal hearing recognize spoken words. The NAM model was also shown to be successful in predicting word recognition performance for

paediatric cochlear implant users (Frisch and Pisoni, 2000).

The principle of the NAM model could be relevant in the subjective confusion tests which were used to compare the data gained from the model developed in this study. It needs to be considered that, in a situation where a listener is uncertain as to which speech token was presented to him/her, the listener would rather respond with the speech token that he/she hears more frequently. The principles described in the NAM model were not used in this study but is important to note since it may be used at a later stage to improve on the model.

### 2.4.3 Multidimensional Phoneme Identification (MPI) model

The Multidimensional Phoneme Identification (MPI) model was first proposed and used in a study by Svirsky (2000). It was used to predict vowel perception of subjects implanted with Ineraid CI which uses a compressed-analogy (CA) simulation strategy. Svirsky's model appears to be the only model in the literature that uses perceptual acoustic cues in a mathematical model to predict vowel confusions in a fashion similar to what this present study attempted to do. The MPI model attempted to imitate how listeners encode and combine acoustic cues and how a decision is made on what sound they heard. This then allowed for testing specific hypotheses about phoneme identification in a more insightful manner instead of the "black box" approaches of other models (Svirsky, 2000).

The Svirsky study had a different goal than that of the present study. The present study attempted to develop techniques to predict results from subjective confusion tests and determine which of two acoustic cue models provide better predictors of the data. Svirsky studied the effect of vowels with "conflicting cues", specifically in terms of temporal and amplitude information as transmitted by the cochlear implant. This was done by presenting the temporal (waveform information) of one vowel while presenting cochlear channel amplitudes of another vowel to the cochlear implantee. Svirsky's mathematical model captures acoustic cues in separate dimensions in a vowel space and uses signal detection theory to generate a prediction confusion matrix. This is precisely the methodology which the present study attempted to follow.

The MPI model uses three components to generate a prediction confusion matrix. The model incorporates an internal noise model to account for sensitivity, a decision model that allows for response bias, and a multidimensional perceptual space. From the perceptual space a prediction confusion matrix is generated using signal detection theory.

The first step in implementing the MPI model was to determine which acoustic cues are to be used for prediction, and then to measure these cues from the stimulus tokens. Svirsky, for the evaluation of his model, chose those acoustic cues that were assumed to be used most frequently by users of the Compressed Analogue (CA) CI strategy (Hillenbrand *et al.*, 1995; Rosen, 1992). The first cue was the first formant (F1), encoded by a temporal cue in channel 1 of the implant. Another cue, believed to be used frequently, was the relationship between the amplitudes of stimulation delivered to different electrodes. In other words, the F1/F2 frequencies (encoded as amplitudes by the four channels in the CI) were used as acoustic cues.

Signal detection theory was used to determine a listener's response to these acoustic cues. Signal detection theory is commonly applied to psychophysics and is widely used to investigate human perception (Wickens, 2002). Physical characteristics of the stimulus and standard deviations were used to create a multidimensional perceptual space. Each stimulus was represented in the multidimensional space as a Gaussian distribution. The mean was determined by the measurements of each acoustic cue of the stimulus. The standard deviation along each dimension is equal to the listener's 'just-noticeable difference' (jnd) along the relevant perceptual dimension. The just-noticeable difference was measured by psychoacoustic testing of the listeners (see Figure 2.1.).

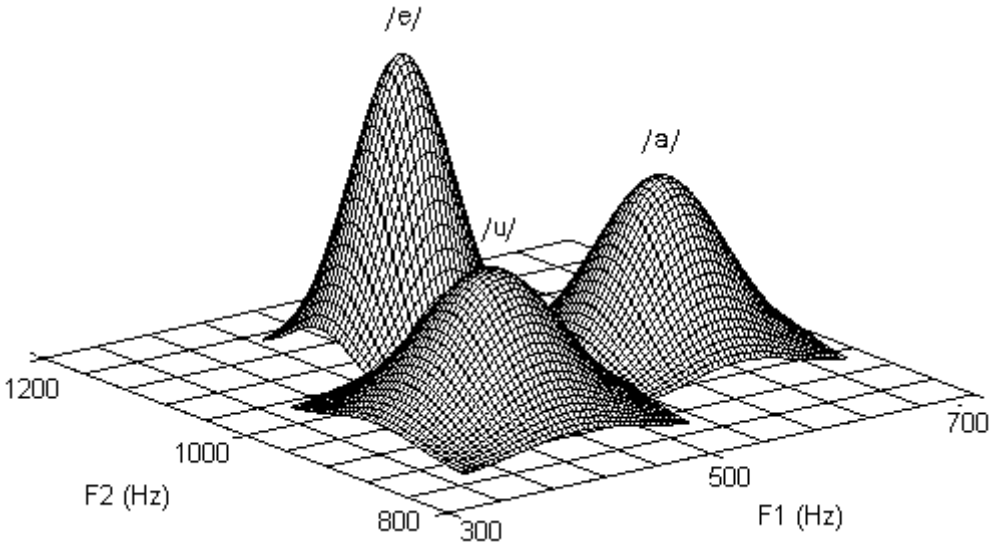


**Figure 2.1. Example of Gaussian probability functions for 3 vowel sounds in a two-dimensional space. The mean of each distribution is the value of the F1 and F2 frequencies and the variance is the *jnd* for each vowel sound.**

Each stimulus can be described mathematically for a perceptual space of M dimensions as a Gaussian probability function S associated with stimulus $E_i$ as:

$$\begin{aligned} S(E_i) &= S_i(x_1, x_2, ..., x_m) \\ &= \frac{1}{JND_1 JND_2 ... JND_m (2\sqrt{2\pi})^m} e^{-(x_1 - T_{i1})^2 / 2JND_1^2} e^{-(x_2 - T_{i2})^2 / 2JND_{21}^2} ... e^{-(x_m - T_{im})^2 / 2JND_m^2}, \end{aligned} \quad (2.2)$$

where $x_j$ is the value of stimulus $E_i$ along dimension $j$, $T_{ij}$ is the average value of stimulus $i$ over dimension $j$, and $JND$ is the subject's just-noticeable difference along dimension $j$.

Once the value of $S(E_i, n)$ has been determined, the decision model was applied to determine the subject's response to the stimulus. The decision model associated a response centre $R_k$ with each possible response, thus creating a partition of the space into response regions. The response region $r_k$ consisted of the points that are closer to $R_k$ than to any other response centre. If the stimulus falls in the response region $r_k$, the subject's response is equal to $k$. Each cell in the response matrix was then determined by the multiple integral

of the distribution $S_i$ over the region $r_k$.

The prediction response matrix for the set of stimuli $E_i$ is obtained by integrating all the S ($E_i$) distributions over each multi dimensional response region using equation 2.3.

$$P(cell_{ik}) = \int_{R_k} S_i(x_1, x_2, ..., x_m) dx_1 dx_2 ... dx_m \tag{2.3}$$

To evaluate the MPI model, various choices of dimensions were used. At first temporal-only (F1) and channel-amplitude-only (that is, A2/A1, A3/A1, A4/A1) dimensions were used, which proved insufficient to predict listeners confusions. The MPI model was then run using all four dimensions which proved to be the best fit when jnd-values of 120 Hz were used for F1 and 2.6 dB was used for the channel-amplitude ratios. The predicted matrix consisted of three vowel sounds and proved to be a good estimation of the observed data. There were no errors greater than 20% and the mean square error was less than 8%. This showed that the MPI model can be used for predicting/evaluating group and individual performance of cochlear implant users (Svirsky, 2000).

The methodology followed by Svirsky lines up closely with the objectives of the present study. Svirksy extracted acoustic cues from the output of the individual channels of a cochlear implant model. This can be seen as negative since calculation is not performed on the vowel sound as a whole, but rather on the individual CI channels. It may be a better approach to perform the calculation on a reconstructed signal using the acoustic cues found in literature to aid vowel identification. The model developed in the present study will also extract the acoustic cues from an acoustic CI model; however the acoustic cues will be extracted from the reconstructed vowel sound. This allows the model to base its predictions on the signal as it is believed that the brain receives it.

The use of probability theory to create probability density functions from the information from the vowel space was used in the present model. This seems to be a good means of predicting confusions based on the acoustic cues. Svirksy did not assign a bias to the

probability calculations because no reason was found to support this. The present study followed the same course.

The approach followed by Svirsky to estimate the 'just noticeable difference' seems to have a few flaws which can be highlighted. Firstly it requires manual measurement of a specific listener's performance in order to be of use. This defeats the objective of replacing subjective testing with a model. In Svirsky's study these values were estimated. This, however, means that the predictions of the model is not made from the information in the signal alone but is also based on assumptions.

The present model attempted to improve this part of the Svirsky model by rather measuring uncertainty factors (which was believed to produce confusion) directly from the signal. This allows for automated use of the model without the need for measurement of listener's performance. In measuring these values, the present model provides a more scientific approach to predicting the speech intelligibility. This approach will also allow the model to automatically measure the deterioration in vowel identification as background noise is added to the original vowel sounds.

### 2.4.4 Neural Network Model

Chang *et al.*, (2001) developed a method that automatically optimized the speech processing parameters for a given implant patient. A neural network was trained to mimic the CI user's performance on the vowel identification task. The neural network model was trained from confusion matrices of the CI user, and was used subsequently to adjust the cochlear implant channel amplitudes to optimize the subject's performance. Results showed that weighting the channel amplitudes from using the neural network method yielded a small, yet significant, improvement in vowel recognition performance.

The benefits of the approach followed by Chang *et al.* (2001) is that the implementation is relatively simple since no psychoacoustic measurements or signal detection models need to

be applied. The disadvantage of the system is that it does not provide further insight into how cochlear implants interpret vowel sounds, since the system assumes a basically black-box approach. Such a model can be used to predict speech intelligibility in specific circumstances and replace actual listeners for determining performance under environmental conditions. A neural network model however fails to explain the reason why intelligibility in a specific circumstance is either good or bad. Experimentation with such a model does not provide information on how to improve the speech perception for a listener. For this reason neural networks were not used to improve the results of the present study.

## 2.5 GAPS IN THE CURRENT LITERATURE

This literature review shows that there have been numerous studies and investigations into the acoustic cues used by normal hearing listeners and cochlear implantees to identify vowel sounds. There still remains a lot to be learnt about how these acoustic cues are used in relation to each other, especially in the presence of noise. The following limitations exist with current objective vowel prediction models, which the present model will attempt to address

- The final output of most models is a single number and does not provide further information which may not aid psychoacoustic research. The current model will extend upon these methods by predicting vowel confusions under various conditions.

- The two models which do provide a confusion matrix as output either need to be trained (HMM) or research needs to be done for each individual person to find a listener-specific ‘just noticeable difference’. The present model, in contrast, will extract acoustic cues and uncertainty factors from the signal to calculate its output.

- The right combination of acoustic cues has not been determined to build an accurate model. The present model will be based on the three acoustic cues which literature shows are the primary cues for cochlear implantees.

- Vowel prediction models are not flexible enough so that other acoustic cues can be inserted or so that the relationship between cues can be modified. The present

model will be developed in a modular fashion so that various environmental conditions and cue inputs can be tested.

These gaps in the existing models have led to the two primary research questions: "Can an objective vowel intelligibility model be developed which extracts these acoustic cues from a speech signal," and "Can processing of these cues be used to predict vowel discrimination and confusions for cochlear implant users?"

The acoustic cues used in the objective vowel intelligibility model was those documented in the literature to aid cochlear implantees most in the recognition of speech, namely the F1 frequency, the F2 frequency and the duration of the vowel sound. The F1 and duration cues were also found to play the largest role in vowel identification in the presence of background noise. These will be extracted automatically from the vowel tokens and used as inputs for the model. In addition, the model was modular so that any other acoustic cues can be used to generate the perceptual dimensions in the vowel space. This allows further testing to be done to evaluate if other acoustic cue combinations will provide better correlation with results from subjective testing.

The MPI model of Svirsky (2000) was the only model found in the literature to use acoustic cues and probability theory to predict vowel perception of cochlear implantees. The approach followed in this model aligns closely to the objectives of this present study. The Svirsky study required that psychoacoustic testing still had to be done to determine the 'just noticeable difference' for every group of listeners that was to be predicted; this study attempts to automate this step.

A new model was developed that attempts to measure the perceived quality of vowel identification by listeners fitted with cochlear implants. This model uses psychoacoustic modelling to predict speech intelligibility and not difference measures (as used in most prediction or evaluation models to date). Difference measures simply subtract certain spectral properties of the signal being evaluated from a reference signal. The current model rather facilitates the use of the acoustic cues in the signal (as used by human

listeners). This allows for better insight into vowel perception of CI users, since different acoustic cues can be evaluated and different weightings on acoustic cues can be assessed.

The focus of the present study falls on the methodology of classification rather than on the selection of acoustic cues. The present study also contributes to current literature by using uncertainty factors so that the model will be able to predict vowel confusion under various background noise conditions.

## 2.6 SUMMARY

In Chapter 2, the literature covering the field of speech perception and speech prediction was summarized. The acoustic cues that listeners use for identifying vowel sounds have been identified. The main acoustic cues will be used in the development of the objective prediction model. A study was also made of the speech prediction models that have been proposed in the literature to date. The methods used in these models have been evaluated and will be employed in the current objective model. The methodology followed in developing this new model will be described in the next chapter.

# CHAPTER 3 METHODS

## 3.1 CHAPTER OBJECTIVES

The literature study in the previous chapter investigated the elements in speech that are used to identify a vowel sound. It concluded that formant frequency and vowel duration were the most important cues in vowel identification. The literature study also looked at various objective speech intelligibility prediction models. A number of models were found which predicted speech intelligibility for cochlear implantees. Only one model was found to use acoustic cues to produce confusion matrices.

The development of a vowel prediction model for cochlear implantees will be described in this chapter. The focus of this study is the methodology of developing such a model, which will be described systematically. The purpose of the method is twofold: to predict the probability of the listener hearing a vowel correctly and to predict the trend of the confusions caused by other vowels. In the next chapter, the ability of the model to predict vowel intelligibility will be compared to subjective testing done with normal hearing listeners and a CI model.

## 3.2 INTRODUCTION

An objective speech evaluation model essentially makes a prediction of the results that would be obtained in a specific subjective speech assessment method. The output of the objective method could then be used to provide researchers with a first approximation of results without needing actual listeners present (as would be required using a subjective test). The objective speech evaluation model described in this chapter attempts to predict the results of a vowel confusion test when performed with a normal hearing listener through an acoustic model of a cochlear implant.

Afrikaans vowels were used in the development and evaluation of the model. In the tests,

the vowel sounds were recorded between the consonants 'p' and 't'. The pronunciation of the 12 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). This is stated here since the vowel sounds will be cited throughout the chapter (the complete test setup will be described at the end of this chapter.)

The predictions obtained are summarized in a confusion matrix. A confusion matrix shows which phonemes were presented to the listener and how the listener responded to each. It shows the percentage of correctly-identified phonemes and those phonemes with which a particular phoneme was confused. Figure 3.1 shows an example of a confusion matrix.

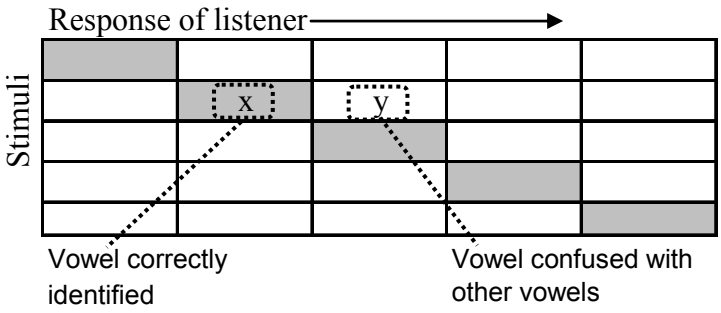


**Figure 3.1. Schema of a confusion matrix.**

Constructing a confusion matrix from an objective vowel prediction model means that acoustic analysis of each presented degraded vowel is necessary. The algorithm will have to extract the acoustic cues thought to be used by listeners to identify each phoneme. It will also have to calculate the probability of a listener either hearing the right vowel or confusing it with other specific vowels.

Confusion is created when the information carried by acoustic cues is not transmitted properly to the listener. Therefore, it must be established which factors of a degraded vowel cause uncertainty in the human auditory system to identify these cues.

Lastly, the confusion matrix is collapsed into a single number to give the predicted

percentage of correct answers given by a listener.

## 3.3 MATHEMATICAL MODELING OF VOWEL PERCEPTION BY COCHLEAR IMPLANTEES

As shown in the literature study, it is accepted among many researchers that the frequencies of the lower formants and vowel duration are the most important acoustic cues used in identifying a vowel correctly. The method defined in this chapter extracts these acoustic cues from the vowel sounds, along with other spectral features in order to produce a measure of uncertainty. The acoustic cues are used to create a 3-dimensional vowel space. A vowel space can be defined as a multidimensional domain where each vowel occupies a single point as a function of chosen signal characteristics. This allows various metrics to be used to measure the distances between vowels to account for how a person distinguishes vowel sounds from each other using specific acoustic cues.

In the present model, the measured frequencies of the formants, F1 and F2, and the duration of the vowel are used for each of the axes that determine where the vowels lie in the vowel space. It is assumed that vowels that lie closer together in this vowel space will have a greater probability of being confused with each other. Figure 3.2 gives a representation of a three dimensional vowel space for the Afrikaans vowels that were used in subjective vowel confusion tests. This vowel space was created by measuring the mentioned three acoustic cues for each of the vowels and then using the model to construct the space. The values of the acoustic cues were extracted by the model described in this chapter and the values are given in Table 4.1.

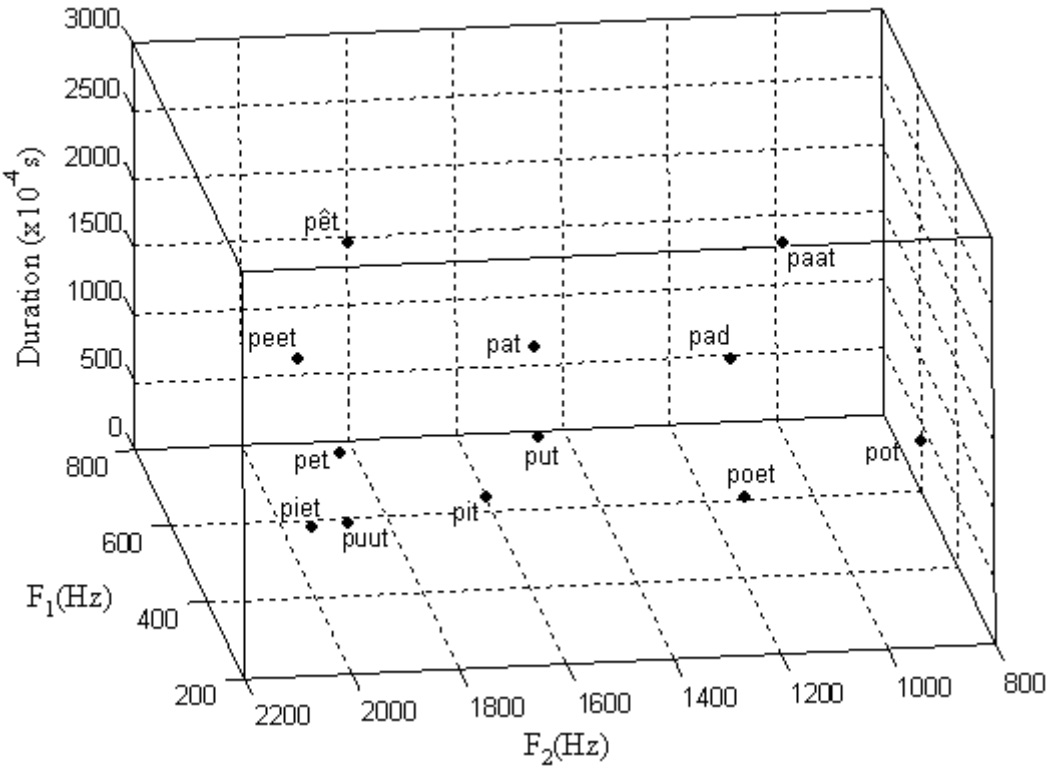


**Figure 3.2. The various Afrikaans vowels used as cues depicted in a 3-dimensional vowel space.**

The proposed model also measures degradation of the acoustic cues that might cause uncertainty to the listener as to the identity of a particular vowel. The degradation of cues is indicated as a variance that represents this uncertainty, so that each vowel is characterized by a pdf (probability density function) in the vowel space. Signal detection theory is then used to calculate the probability of a vowel being confused with any other vowel.

The distances between the means of the pdfs were calculated from the vowel space. The variance of each distribution was calculated from factors that cause uncertainty in the position of the formants and the duration. The literature study implicated a few aspects of a formant that may act to confuse the auditory system when it's task is to recognize where formants lie in a vowel space. One of these aspects is reduced spectral contrast caused by noise (Loizou and Poroy, 2001a). Another possible factor is the instability of formants; this is caused when spurious formants are apparently inserted into the spectrum due to noise. These factors will be discussed later in the chapter.

Finally, signal detection theory is used to reach the objective of producing a confusion matrix which predicts confusions made by cochlear implantees. This will approximate the process whereby a person, listening to sounds in the presence of additional noise, identifies a vowel sound and isolates the factors that contribute in creating confusions between the vowel presented to the listener and other vowels.

## 3.4 DEVELOPMENT OF MODEL

The model consists of two parts. Firstly, the *processing component* is described and the important features that a human would use to discriminate between vowels are identified. The second part is the *decision component*, which uses signal detection theory to predict the probability of discriminating between specific vowels . Finally the confusion matrix (as the output) is explained.

The individual steps that form the proposed model are described in more detail in the rest of the chapter. The system is based on comparison methods which are the most common in the perceptual speech evaluation field (Beerends *et al.*, 2002; Rix, Hollier, Hekstra and Beerends, 2002; Thorpe and Wonho, 1999b; Voran, 1998).

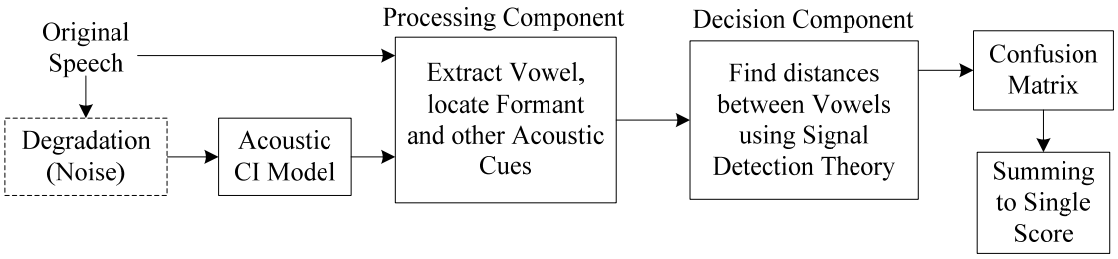


**Figure 3.3. Outline of the objective speech evaluation model. (Repeat of Figure 1.4.)**

The model accepts original and degraded speech as inputs (see Figure 3.3.). The original speech was recorded under quiet conditions in a double-walled sound booth (Pretorius *et al.*, 2006) and used as input. This serves as a reference to aid the model in extracting the acoustic cues from the degraded input.

For the second input the original speech was processed through an acoustic CI model before it entered the processing component as degraded speech. It is this degraded speech that will be processed by the model. The acoustic model was used to convert the input speech signal into an acoustic representation of what a person fitted with a CI is believed to hear. The acoustic model used for the experimentation was developed by Conning and Hanekom (2005, unpublished). It was designed to approximate the speech processing used in a Nucleus cochlear implant. The model also implements aspects of the biophysical interface that affect the signal in the cochlea. The acoustic CI model approximates the speech heard by a person with a cochlear implant (more information on the acoustic model is given in section 3.5.1 about the experimental setup). The output of the acoustic CI model is used as the degraded input that is to be processed by the prediction model.

The aim of the model was to predict the intelligibility of vowels as heard by cochlear implantees under specific conditions. Therefore, the CI acoustic model was an integral part in the development of this model and the output of the model was used to make decisions on how to implement certain functions. The output of the CI acoustic model produces severely degraded speech tokens; therefore, spectra of the vowels processed by the acoustic model were analysed to determine how uncertainty in the acoustic cues could be expressed.

In Figure 3.3, prior to the acoustic CI model, there exists an optional step that adds multi-talker babble noise to the original speech to simulate real-world environments. Initial testing was done with no additional degradation through noise added (see block diagram in Figure 3.3), because of the severity of degradation caused by the acoustic CI model. The ability of the model to track the results of the subjective test in the presence of different SNRs is used to evaluate the model in the results chapter.

Speech is presented to the model in consonant-vowel-consonant context, exactly as it is when presented to listeners in subjective tests. The same material was used for both. The processing component extracts the vowel part from the presented speech in the fashion

described in paragraph 3.4.1. It also performs acoustic analysis on the vowels to extract the selected features from the vowels for analysis. It does this for both the original speech and the degraded speech.

The decision component uses the information extracted from the vowels in the processing component to predict the probability of vowel intelligibility and confusion as described above. These probabilities were calculated for each vowel and used to construct a confusion matrix and to produce a single intelligibility score. The rest of the chapter will give a in-depth description of the entire model.

### 3.4.1 Processing Component

The first component of the proposed vowel perception model is the processing component. The processing component has the following three goals.

- Extract the vowel from the input speech signals. (Phonemes in a /CVC/ (consonant-vowel-consonant) form; in the present instance, in a 'p'-vowel-'t' context.)
- Extract the acoustic cues from the vowel sound.
- Estimate the factors that will lead to uncertainty in a listener.

An outline of the processing component is shown in Figure 3.4. Subsequently, each functional block will be discussed in detail.

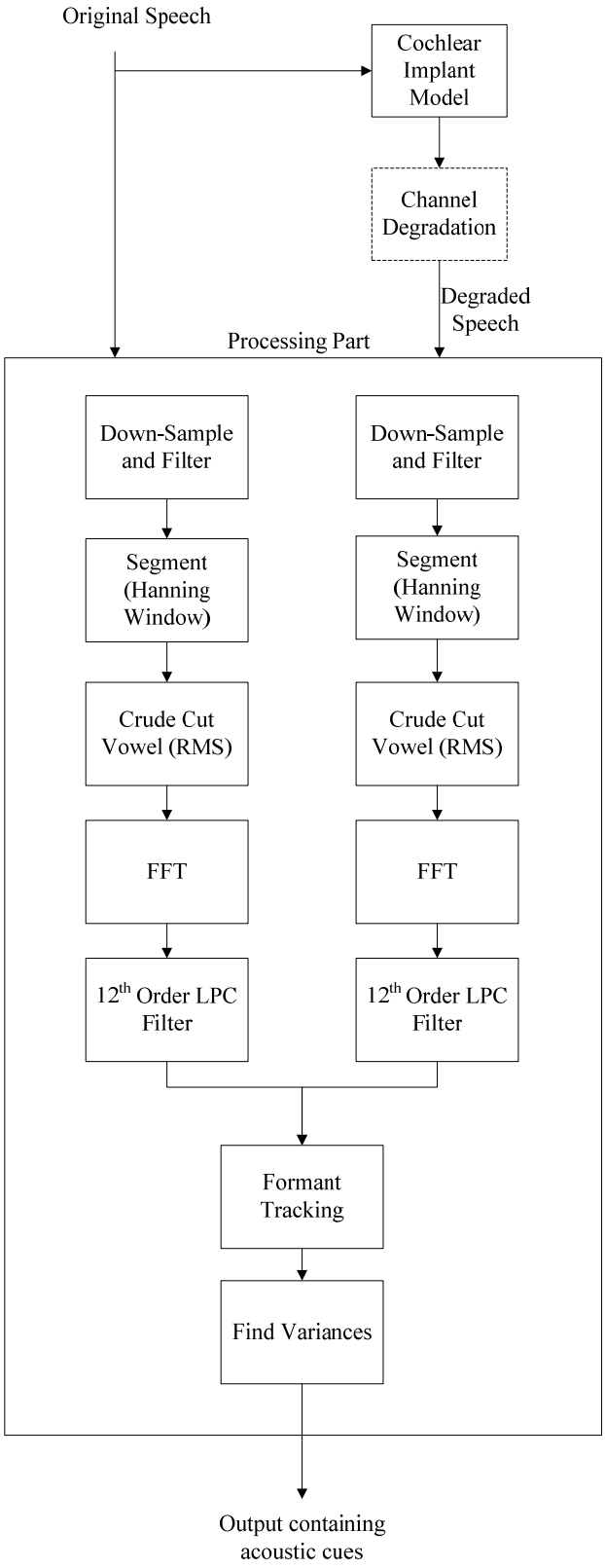


**Figure 3.4. Block diagram of processing component**

#### 3.4.1.1 Inputs

The processing component receives two inputs. Firstly, it receives the original speech signal, which is the vowel in /CVC/ form and in a quiet state (therefore, as a normal hearing person would perceive it when the sound is presented under ideal circumstances). Moreover, it receives the degraded signal which has gone through an CI acoustic model and an optional degradation channel. The degraded signal is the input that the algorithm evaluates.

Since there is significant degradation caused by the CI model, the original speech is used as an input to serve as a reference to where the actual formants lie and to what the actual duration of the vowel is. To enable comparison across SNR with actual listener performance, the degradation channel added multi-talker babble noise to the original speech to simulate real-world scenarios. However, the degradation channel can also be, for example, a telecommunication system or codec that needs to be evaluated for use by cochlear implantees. The degradation channel can also be omitted in the situation that certain CI parameters need to be evaluated for noiseless speech. Therefore, this step is indicated with a dashed-line box in Figure 3.4.

#### 3.4.1.2 Down Sampling

The original vowels that served as input to the system were sampled at 44kHz. Speech lies primarily in the frequency band from 100 Hz to 3400 Hz (Zwicker and Fastl, 1999); therefore, the original speech samples may be down sampled to 8000Hz. This procedure would ensure a Nyquist frequency of 4000 Hz and is a common practice in perceptual speech evaluation models (ETSI Standard EG 201 377-1, 2002; ITU-T Recommendation P.862, 2000). Down sampling is important to ensure that only the required speech information lying in this frequency band is used in the Fast Fourier Transform (FFT) and the Linear Predictive Coding (LPC) calculations in the later steps. This ensures that maximal precision is obtained in the spectrum that is produced since all the points of the FFT is focused on the band in question and not on irrelevant spectrum information.

Down sampling is usually followed by a low-pass filtering process to prevent the occurrence of aliasing in the base band. In the present model, therefore, a linear-phase anti-aliasing low pass FIR filter function was used with a cut-off frequency of 4000Hz. The standard built-in MATLAB low-pass filter function was used to accomplish this task.

#### 3.4.1.3 Segmentation (Hanning Window)

The analysis of vowel signals in the developed system required that the signals be segmented into overlapping blocks of evenly-spaced samples in time. Segmentation allows usable blocks of samples to be used to produce a spectrogram. The typical length of windowed slices that speech is segmented into is between 15 ms to 40 ms in length (ETSI Standard EG 201 377-1, 2002). In the present model the speech signal was segmented into 32 ms frames with an overlap of 50%. Overlapping successive blocks is a smoothing operation that avoids abrupt changes from segment to segment. The choice of segment length and overlapping percentage conform to those implemented in current speech evaluation algorithms (ITU-T Recommendation P.862, 2000; Rix *et al.*, 2001; Voran, 1999a).

Segmentation can be seen as the multiplication of each segment by a rectangular window. Because the frequency response of a rectangular window contains high side lobes (in the frequency domain), it is normally not recommended to make use of such a window. For this reason each segmental block was multiplied by a Hanning window. A Hanning window reduces the endpoints of each segmental block to zero, avoiding spectral leakage in the process. Spectral leakage is an effect in the frequency analysis of signals where small amounts of signal energy are observed in frequency components that do not exist in the original waveform. The window is computed from the following equation

$$h(k) = \frac{1}{2}(1 - \cos(2\pi k / N), \tag{3.1}$$

where N is the number of samples in the window. The number of samples is calculated in

terms of the sampling frequency using the following equation.

$$N = 32ms \times \frac{1}{(1000/f_s)ms/sample}, \quad (3.2)$$

where N is the number of samples in the window and $f_s$ is the sampling frequency.

#### 3.4.1.4 Removing the Vowel From the Word

Speech signals presented in vowel recognition tests are presented in a /CVC/ form: the vowel is presented between two chosen consonants. This is how the vowel is presented to a listener in a traditional subjective vowel confusion test. (Refer to the literature study for more information.) Given that the exact speech signal needs to be used in the objective evaluation model as it is used in the subjective test that it attempts to replace, the sound files that are used for the input to this system must be in the same form. Since the vowel portion of the file needs to be analyzed, it is necessary to determine where the vowel lies and, subsequently, the speech needs to be cropped so that only the vowel remains.

Acoustically, vowels differ from consonants in at least two ways. Firstly, a vowel contains more energy than a consonant. Secondly, a vowel is characterised by steady state or slowly gliding formants in the frequency spectrum. These two observations are harnessed to crop the vowel sound of the /CVC/ word that is presented to the listener. By calculating the Root Mean Square (RMS) for each of the 32ms time windows, the vowel can be distinguished from the rest of the word. The RMS value for each window through time of the Afrikaans word “peet” is plotted in Figure 3.5. It is visible in the figure that the high energy in the vowel sound distinguishes it from the rest of the word.

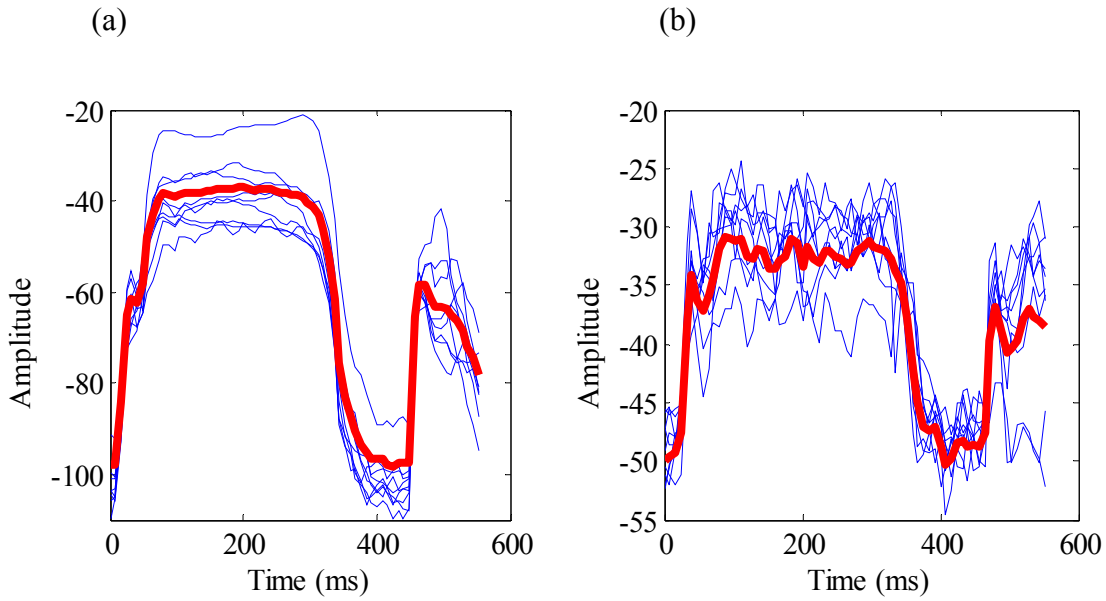


**Figure 3.5. Root Mean Square of the Original Speech (a) and Degraded Speech (b) of the /CVC/ word "peet". The thin blue lines represents the RMS of the frequency bands and the thick red line represents the average RMS of all frequency bands.**

These graphs were created by first separating the spectrum of the word into 10 separate frequency bands (shown as thin blue lines in the Figure 3.5.) This was done to create broad bands which show the energy of the different formants and valleys in the spectrum. The RMS of each time window was calculated to form the separate curves on each graph. Since all bands were relatively stable throughout the duration of the vowel all of the frequency bands were used to calculate the RMS average energy of the vowel sound. The calculated average is shown as a heavy red line in Figure 3.5. The maximum value is found throughout the graph and, by experimentation, it was determined that a threshold of 70% of that value would provide reliable estimates of the position of the vowel in the input token. The vowel is delimited by the first rise through the threshold and the first drop past it. This section is then identified as the vowel and cut from the rest of the word.

The vowel part of the degraded signal lies in the same windowed segments in time as the original, since the acoustic CI model has no effect on the time scale of the sound token (that is, there is no temporal distortion). Therefore, this calculation is only done on the original speech to improve the accuracy of the vowel cropping. The start and stop markers

are used as a reference in conducting the processing of the degraded speech tokens. The formant frequencies and uncertainty factors are still calculated from the degraded speech, since the degraded speech is being evaluated.

#### 3.4.1.5 LPC Spectrum

Linear prediction analysis of speech is historically one of the most important speech analysis and synthesis techniques (Makhoul, 1975b). In digital signal processing linear prediction is often called linear predictive coding (LPC).

Speech analysis with LPC exploits the predictable nature of speech signals. Cross-correlation, autocorrelation, and auto-covariance provide the mathematical tools to determine this predictability (Berouti, Schwartz and Makhoul, 1979; Hermansky, 1990). Speech analysis by linear prediction is based on the assumption that the short-time spectral envelope of the speech waveform can be represented by a number of poles (Makhoul, 1975a). The signal is modeled as a linear combination of its past values. This amounts to performing a linear prediction of the next sample as a weighted sum of past samples (Berouti *et al.*, 1979). The all-pole model of the speech spectrum is accurate for approximating vowel and vowel-like sounds (Atal and Schroeder, 1978). The transfer function *H(z)* of the filter is given by the equation

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{3.3}$$

where $G$ represents the gain, $p$ the order and $a_k$ the different coefficients of the transfer function in the z-domain.

The two most frequently implemented methods used to compute the coefficients for the all-pole filter (Equation 3.3) is the covariance method and the auto-correlation formulation. When using the auto-correlation formulation, the roots of the polynomial in the denominator of Equation 3.3 will always be inside the unit circle in the z domain. This will

guarantee stability for the filter *H(z)*. It is for this reason that the auto-correlation formulation was used to compute the coefficients in equation 3.3. The auto-correlation method requires the calculation of the auto-correlation equation

$$R_{xx}(k) = \sum_{n=n_0+1+k}^{n_0+N} x_n x_{n-k} \quad , \tag{3.4}$$

where 1<k<p with p being the order, N the number of samples and $n_0$ the first sample in the window. The values of $R_{xx}$ can be written in the matrix $R$ and matrix $B$ while the matrix of $A$ contains the filter coefficients as in Equations 3.5 and 3.6.

$$R = \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & R_{xx}(2) & \cdots\cdots & R_{xx}(p-2) & R_{xx}(p-1) \\ R_{xx}(1) & R_{xx}(0) & R_{xx}(1) & \cdots\cdots & R_{xx}(p-3) & R_{xx}(p-2) \\ R_{xx}(2) & R_{xx}(1) & R_{xx}(0) & \cdots\cdots & R_{xx}(p-4) & R_{xx}(p-3) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ R_{xx}(p-1) & R_{xx}(p-2) & R_{xx}(p-3) & \cdots\cdots & R_{xx}(1) & R_{xx}(0) \end{bmatrix} \tag{3.5}$$

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ \vdots \\ a_p \end{bmatrix} \qquad B = \begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ R_{xx}(3) \\ \vdots \\ \vdots \\ R_{xx}(p) \end{bmatrix} \tag{3.6}$$

In order to obtain the coefficients in the matrix $A$, the equation

$$RA = P \, , \tag{3.7}$$

needs to be evaluated. This can be done by rearranging Equation 3.7 in the following manner

$$A = R^{-1}P. \tag{3.8}$$

To solve for $A$ it is required that $R^{-1}$ be computed. It is important to notice that the matrix of $R$ is symmetric and that all the elements that are on a line parallel to the diagonal elements are equal. This type of matrix is called a Toeplitz matrix and there exist efficient

recursive algorithms to find its inverse. The Levinson-Durbin algorithm is one such algorithm and takes advantage of the properties of the Toeplitz matrix of *R*. The Levinson-Durbin algorithm is depicted in the flow diagram in Figure 3.6.

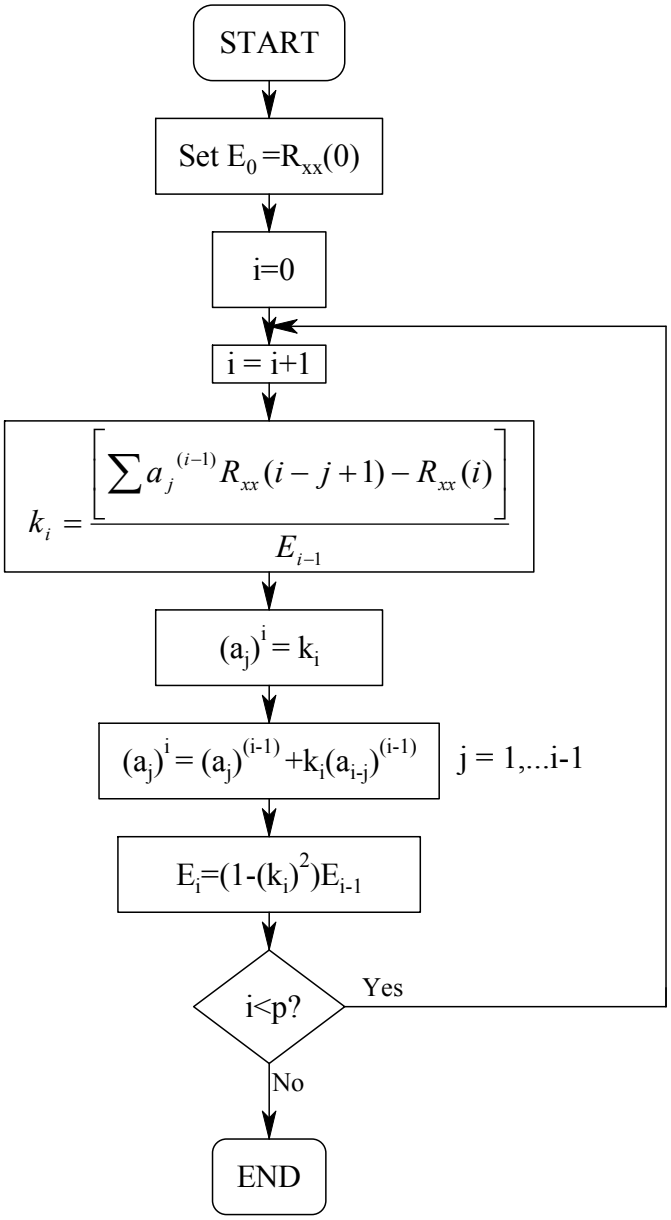


**Figure 3.6. Flow diagram of the Levinson-Durbin algorithm**

In the block diagram (Figure 3.6) the filter order is denoted with a superscript $a_i^{(j)}$ for a j'th order filter. The average mean squared error of a j'th order filter is denoted by $E_j$. For a

*p*'th order filter, the Levinson-Durbin algorithm calculates all filters that have an order of less than *p*, and hence it determines all order *N* filters where N=1,...,p-1.

The final LPC transfer function was formulated by using the coefficients obtained from the Levinson-Durbin algorithm, as well as the gain of the filter, which is defined by

$$G = R_{xx} \times A\,. \tag{3.8}$$

All the windows in the crude-cut vowel are filtered with the LPC filter to gain a smooth approximation of the FFT spectrum (that is, the envelope). A 12th order LPC filter was implemented to approximate the spectrum of the signal. This order will give a maximum of 6 peaks in the 4000Hz bandwidth of the signal. An order of more than 10 is recommended in the literature so that the spectrum forms enough peaks to accentuate the 4 -5 formants of the vowel which lie in the first 4000 Hz of a vowel sound (Ferguson and Kewley-Port, 2002; Snell and Milinazzo, 1993). Using a LPC filter higher than 10th order allows all of these peaks to be shown as can be seen in Figure 3.7.

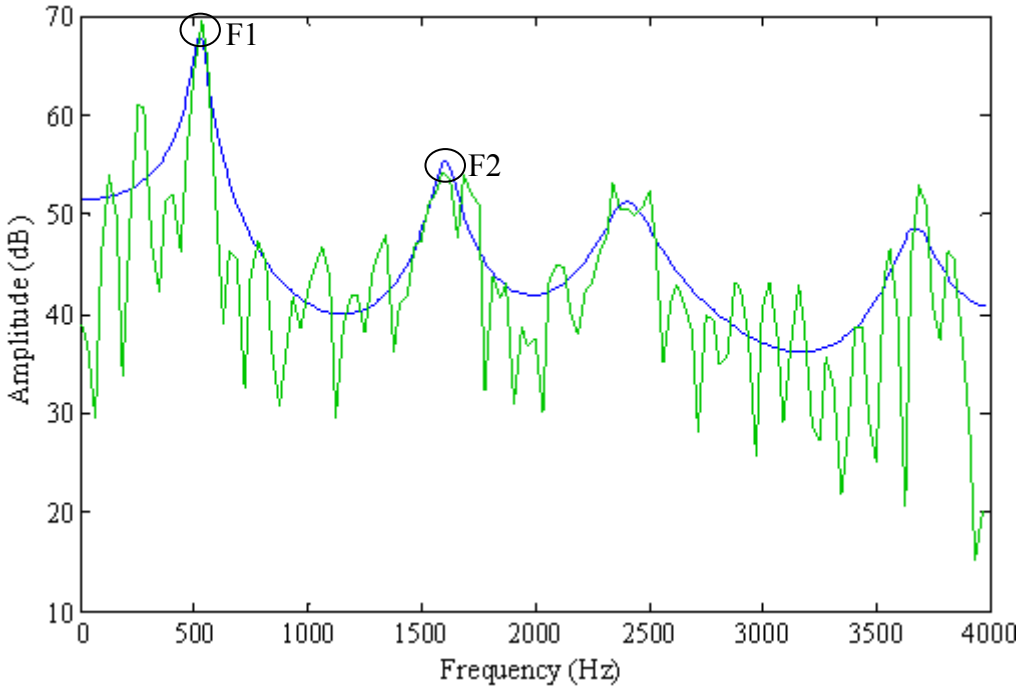


**Figure 3.7. FFT and LPC Spectrum of a single window of the word "pit". The smooth blue line represents the LPC spectrum.**

The Fast Fourier Transform (FFT) was implemented to obtain the discrete Fourier transform (DFT). The Fourier transform was calculated by using a 256 point FFT. The LPC was implemented by using a $12^{th}$ order all-pole LPC filter. The peaks in the LPC spectrum make it easy to identify the formants that a listener may use to identify the vowel. The most important of these are the lower formants, F1 and F2, represented by the first two peaks in the figure.

The LPC was calculated for each segmented window in the word to form a LPC spectrogram for the entire word. Figure 3.8 shows the LPC spectrogram for the entire /CVC/ word "pit". The vowel part of the word can be seen clearly as the high energy part of the word. After the vowel there is a break before the burst of sound that produces the last consonant "t".

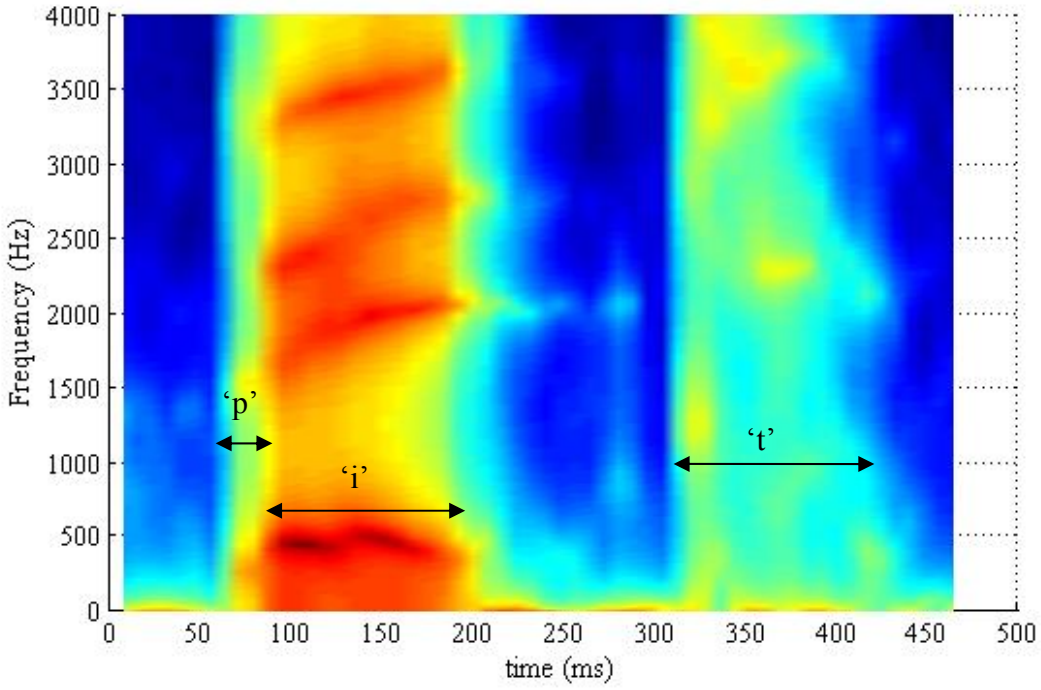


**Figure 3.8. LPC Spectrogram of the word "pit".**

This LPC spectrogram was then used in determining the frequencies of the F1 and F2 formants of the vowel.

### 3.4.1.6 Formant Tracking

The LPC spectrogram of the vowel 'i', cut from the original word "pit", is shown in Figure 3.9. By inspecting the spectrogram it is easy to determine where the formant frequencies lie. The first formant lies around 500 Hz, the second between 1550 and 1800 Hz, and the third between 2450 and 2700 Hz. There is very little uncertainty to human hearing as to where these formants lie. However, this is more difficult to see after the clean vowel has been processed through the acoustic CI model.

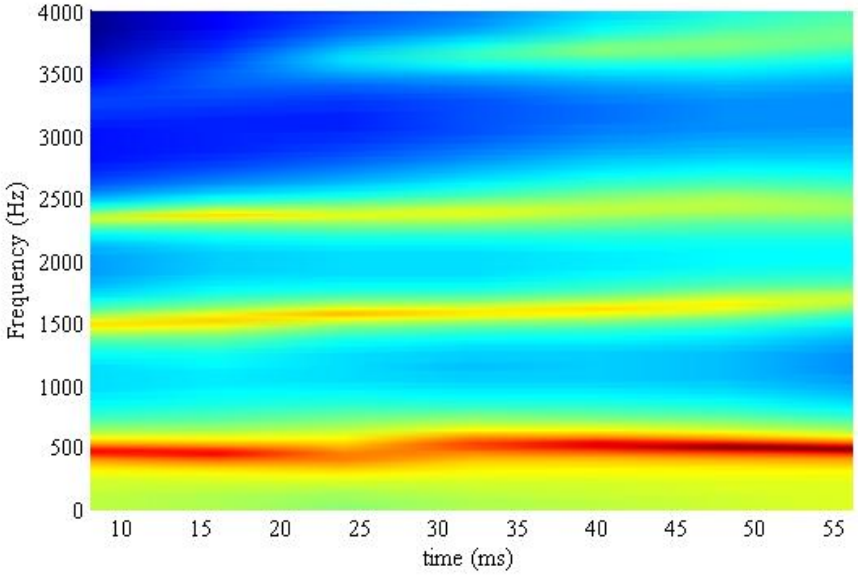


**Figure 3.9. LPC spectrum of only the vowel portion sliced out of the /CVC/ phrase "pit". The formants are clearly visible as red/yellow bands as consistent throughout the noiseless vowel.**

When the same word has been processed by the CI model, the frequency of the formants loses its exactness. This can be seen in Figure 3.10. The vowel is still visible although the formants are not as apparent as in the spectrum of the clean vowel. Since these are the cues that a person uses to identify a vowel, an uncertainty factor is introduced.

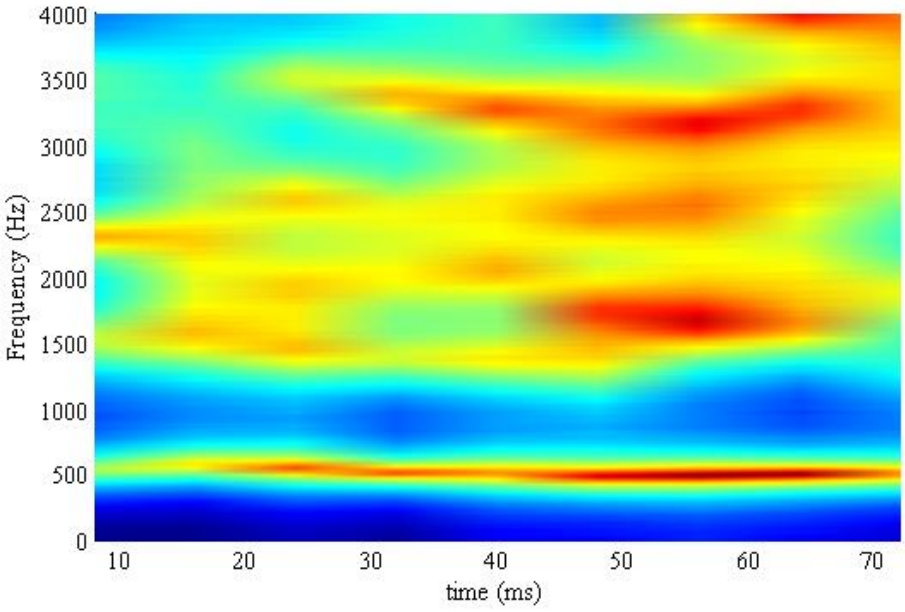


**Figure 3.10. LPC spectrum of only the vowel sliced out of the /CVC/ phrase "pit" (after processing in an acoustic model). The formants are clearly visible as stable throughout the noiseless vowel.**

Since the algorithm has to determine the perceived frequencies of the first two formants of the degraded vowels, simply picking the peaks in the signal as formants is not sufficient in finding the formants. At least two approaches to the problem are available – analysis by synthesis and peak picking from smoothed spectra (McCandless, 1974; Snell and Milinazzo, 1993).

In analysis by synthesis, an educated guess is made of the formant frequencies and bandwidths, and a spectrum is generated based on the educated guess. The formant frequencies for the synthesized spectrum are varied systematically until the differences between this and the actual spectrum are minimized (McCandless, 1974). A method for varying all three formant frequencies, using a Newton-Raphson technique to find a least-squares fit is described in .

In peak-picking, certain rules are applied to select the appropriate peaks from a smoothed LPC spectrum at each frame to identify the first two or three formants. The challenge is in recognizing which peaks are spurious and/or whether two formants have merged into one

peak (McCandless, 1974; Snell and Milinazzo, 1993). The peak picking method implemented will be explained in the following paragraphs.

The peak-picking method was implemented for this algorithm. An iterative approach was followed to select the correct peaks as formants. As a first approximation, the first two peaks were chosen as the first two formants. Therefore, the algorithm applied determined the peaks for each window and labelled the first peak F1 and the second peak F2. This simple implementation worked well for the clean vowels, but, once the vowel had been processed by the CI model, errors were made in determining the formants. This can be seen in Figure 3.11.

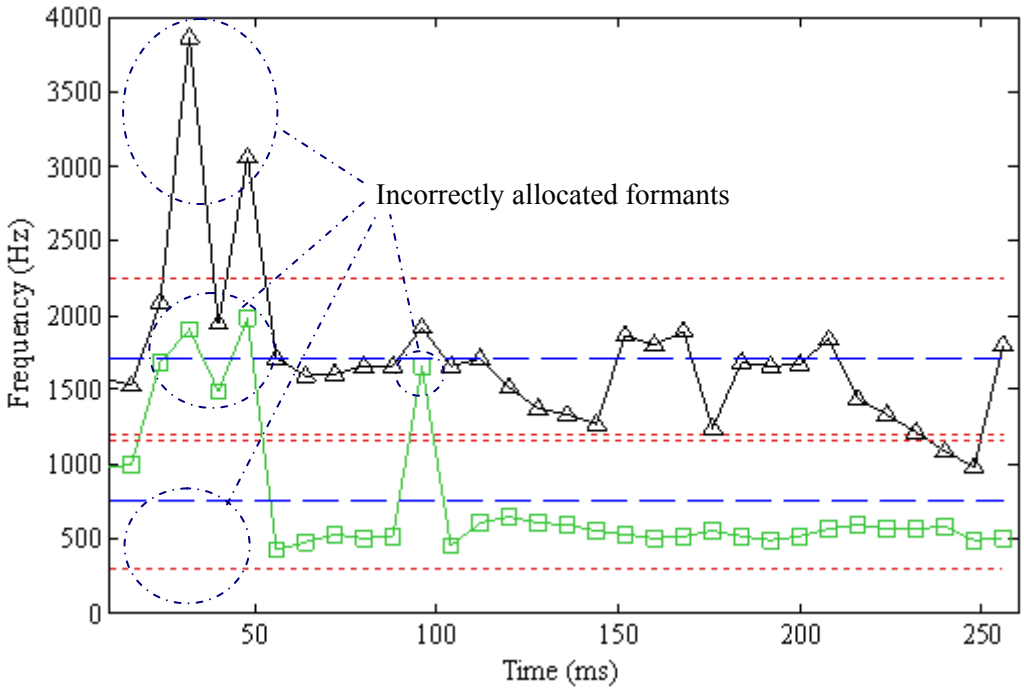


**Figure 3.11. Example of formant detection error in a degraded vowel in the degraded word "peet". The triangles depict the F2 frequencies and the squares represent the F1 frequencies. The blue dashed line depicts the standard deviation of each formant, and the red dotted lines show their standard deviation.**

The frequencies of the formants picked from individual LPC windowed segments in time from the vowel in the degraded word "peet" is shown in Figure 3.11. The line depicted by squares represents the frequencies picked by the algorithm as F1 frequencies. Similarly,

the line depicted by triangles represents the frequencies picked as F2 frequencies. The lines of long dashes represents the average frequency for each of the two formants. The lines of short dashes represent the standard deviation of each of the two formants.

Looking at Figure 3.11, there are obvious mistakes in identification of the formants. This demonstrates how an error can be made when picking the formants directly from the peaks in the LPC spectrum. The reason this happens in the degraded signal is that a peak might disappear for a small number of windows when predicting the vowels with a $12^{th}$ order LPC filter because of the low spectral contrast of the signal. It may also disappear when two formants merge or they are sufficiently close to one another so that the valley between them disappear in a LPC spectrum.

The first problem was solved by calculating the mean and standard deviations for the formants (shown in Figure 3.11 by dashed lines and dotted lines, respectively). Normally a mismatch is made when F1 is not detected and the peak for F2 is mistakenly recognized as F1. Therefore, if a frequency recognized for a formant lies outside the bounds given by the standard deviation, the frequency is reallocated. In other words, if the frequency for F1 lies inside the standard deviation bounds of F2, it is reassigned to F2. For that specific window in time there is then no peak allocated for F1 since there is no peak representing F1. It is assumed that a listener will not be able to have F1 available for that instance in time since there is no peak available. The formant frequency is then left out for that specific window, and the spectral contrast value (which will be explained later) is set to 0.

Cross-checking is also performed in time. Consequently (in Figure 3.11), if a selected peak creates a large jump from the selected peak in the previous window in time, then it is accepted that there is no relevant peak depicting that formant for that instant in time. No formant will be picked for that window, and the spectral contrast will once again be set to 0.

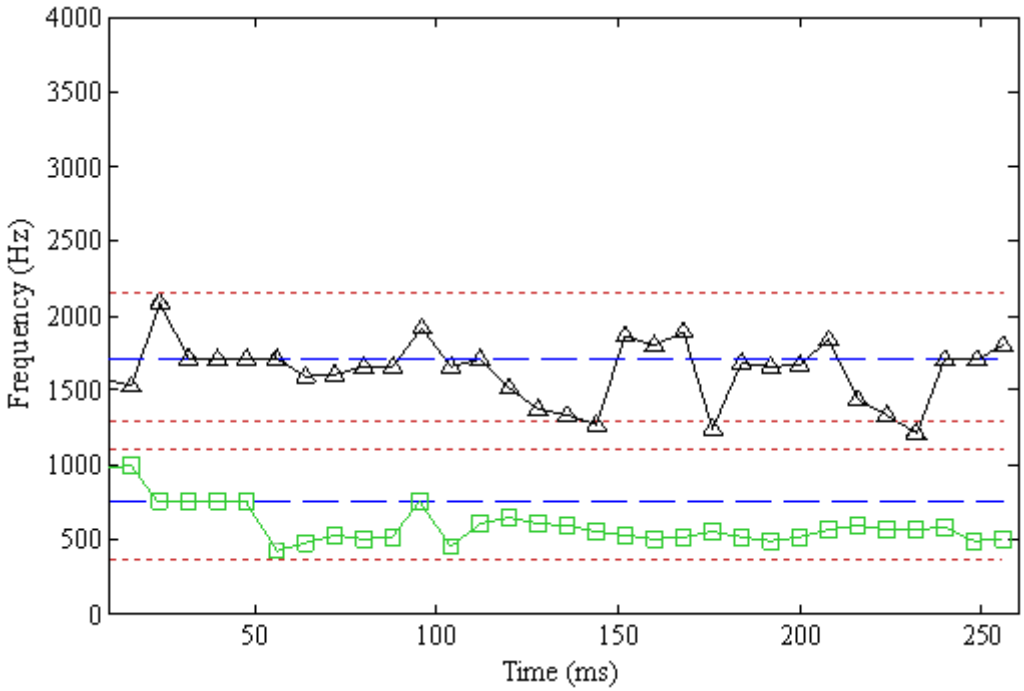


**Figure 3.12. Example of formant tracking after error correction has been implemented for the word "peet".**

Once these correcting features have been implemented, the errors in Figure 3.11 are corrected. The formant tracking shown in Figure 3.12 above shows that F2 peaks are no longer selected incorrectly as F1 frequencies. The average frequency for each formant across all windows in time is taken as the formant frequency for that specific vowel. To evaluate the formant tracking in the figure above, the formant frequencies were highlighted in the spectrogram of the same word in Figure 3.14. The thick solid lines show the tracking of formants which were done manually by inspection of the spectrogram. The dotted line gives the approximate average of each of the two formant frequencies. The comparison of the automated tracking in Figure 3.12 and the manual inspection of the formant frequencies in Figure 3.14 correspond well. Even by inspection, however, it is clear that the formants in a vowel processed by a cochlear implant is not well defined and hard to approximate.

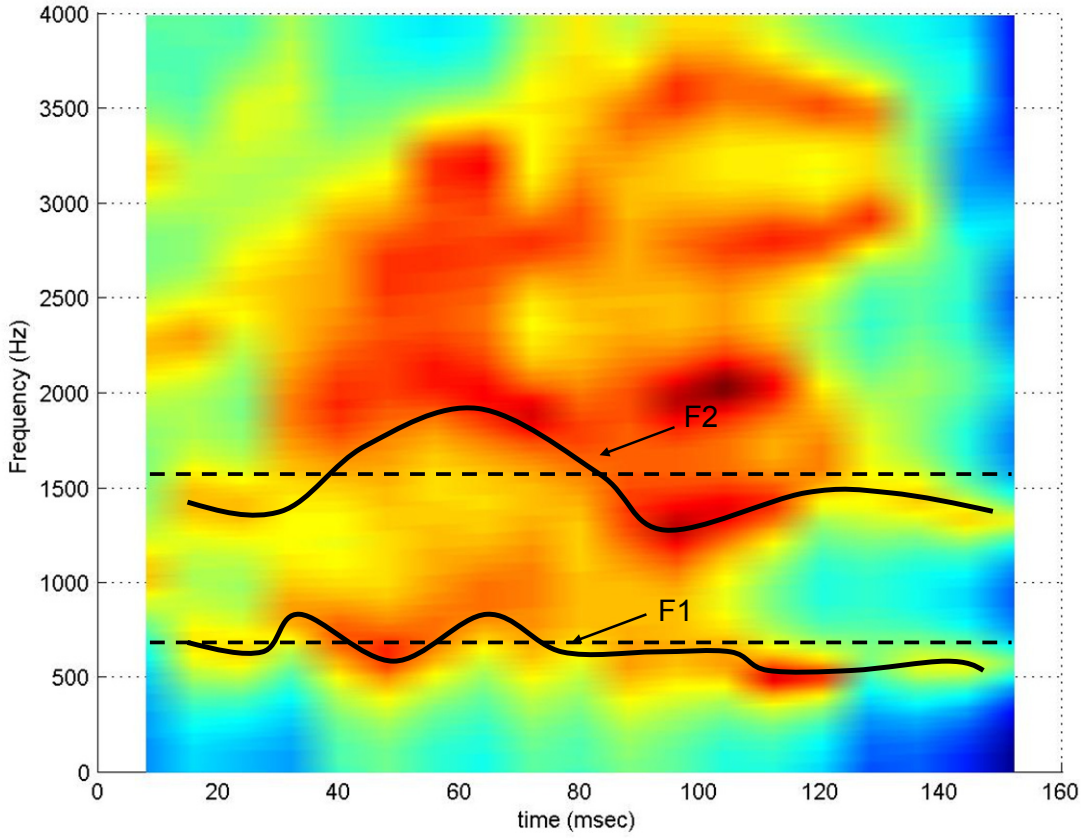


**Figure 3.13. Manual formant tracking by observation of spectrogram of the processed word "poet".**

### 3.4.1.7 Uncertainty Factors

Figure 3.14 shows the LPC spectra for three different windows evenly spaced in time for the degraded vowel in the word "pad". For a clean vowel the spectral peaks are normally constant over the duration of the vowel or show a slow increase or decrease throughout the time of the vowel (which would mean that all the graphs in the figure would look very similar). It is seen easily in Figure 3.13 that for a degraded vowel the formant peaks are not constant from window to window. The formants continuously shift in amplitude and frequency, and the spectral contrast changes continuously as formants appear and fade.

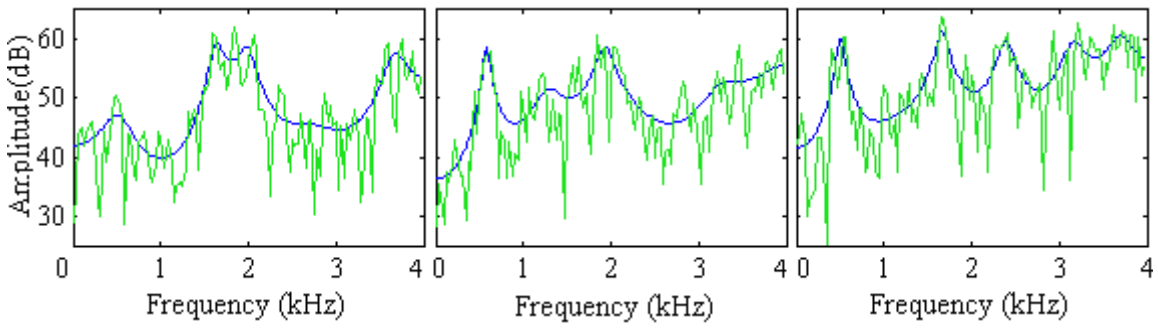


**Figure 3.14. Example of the instability of the LPC Spectrum between three time instances taken from the vowel portion of the degraded word "pad". The peaks in these spectra should look almost identical in a clean (non-degraded) vowel.**

It has been established earlier that formant peaks are important acoustic cues in recognizing a vowel sound. Therefore, the observed wavering of the spectral peaks between these windows in time suggests that uncertainty may be produced in a person's perception of where the formants lie. It is assumed that the inability of a person to recognize the formant positions may be one of the causes of vowel confusion for a cochlear implantee. The model has to be able to measure these confounding effects. From inspection of the degraded vowel, two main factors that have been identified as causing ambiguity in vowel identification (because of the masking of the formant frequency). These are:

- Frequency Variance
- Spectral Contrast

One objective of the present work was to identify which of these two masking effects is the larger cause of deterioration in vowel identification. Each of these factors will be implemented into the model separately. In the next chapter these methods will be evaluated against each other so as to determine which method predicts most accurately the results obtained from the subjective listening tests with the acoustic CI model. The rest of this section will cover how these two methods are used in quantifying these uncertainties.

**Frequency Variance**

From observation it was apparent that the formants in clean vowels have very stable frequencies where the formant frequency never fluctuates more than a 100 Hz in consecutive windows. Listeners can assumedly easily recognize these vowels because the formants are clearly defined. This can be seen in Figure 3.15 (a) which shows the LPC spectrogram of the vowel from the Afrikaans word "poet". Figure 3.15 (b) shows the frequencies of F1 and F2 tracked by the algorithm for each window of duration through time. The dashed lines show the mean of the two formants and the dotted lines show the standard deviation of the formants. The standard deviation in this case is clearly not very large due to the stable formant frequencies over the duration of the vowel.

(a) (b)

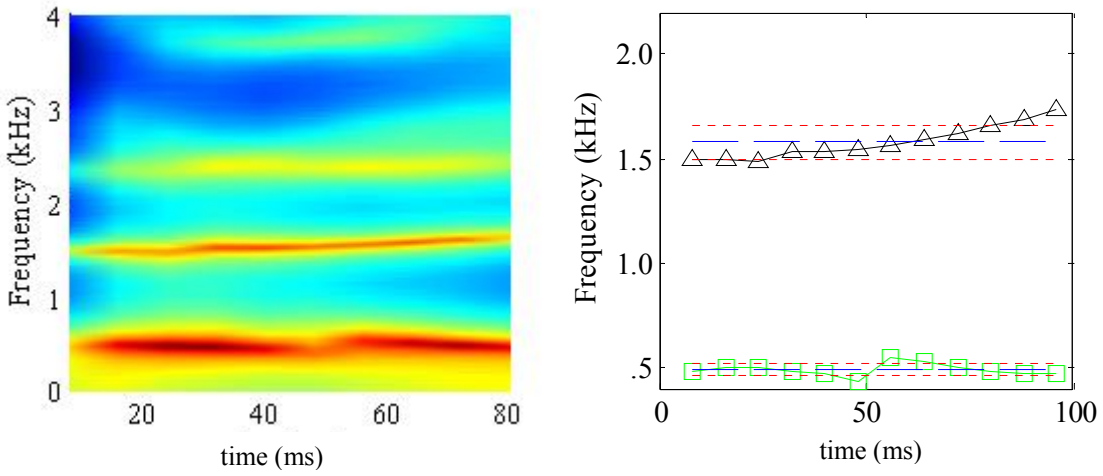


**Figure 3.15. (a) LPC spectrogram and (b) and formant tracks of the vowel sound /u/ extracted by the algorithm from the clean Afrikaans word "poet".**

In contrast to Figure 3.15 which represents a clean vowel, Figure 3.16 (a) shows the amount of noise visible in the spectrogram for the same vowel after it has been degraded by the acoustic CI model. It is difficult through inspection to see where the peaks that represent the formants in the spectrum are situated. When the algorithm has tracked these formants, Figure 3.16 (b) is produced. Although the first formant is still quite stable, much variation is produced in the frequency of F2. The mean of the frequency of F2 (depicted

by the dashed line) still lies close to that of the original. However, the standard deviation (shown by the red dotted line) has increased dramatically.

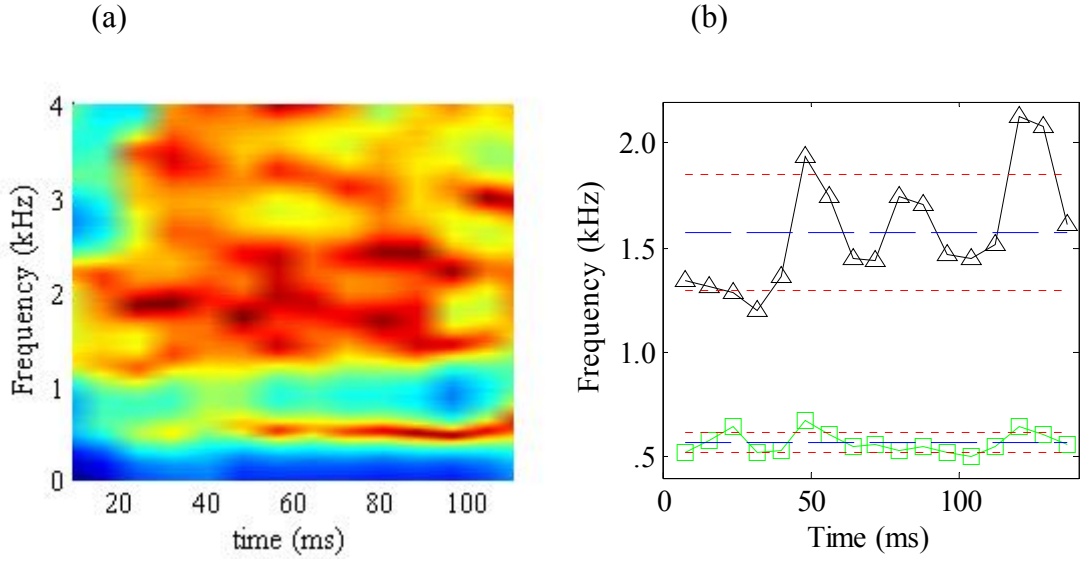


**Figure 3.16. (a) LPC Spectrogram and (b) Formant tracks of the vowel sound /ʊ/ extracted from the Afrikaans word "poet" after it has been processed through the CI model.**

It is assumed that the variability in the formant frequencies is one of the causes of vowel misinterpretation. This assumption is based on the fact that research defines a formant as being a stable spectral peak (Peterson and Barney, 1952; Hillenbrand *et al.*, 1995). Inspection of the vowel sound in Figure 3.16 shows that the F2 frequency undergoes erratic movement in time. Therefore, the standard deviation of the formant frequencies was used as the first measure of uncertainty. The standard deviation was calculated for the detected frequencies of the peaks of each window through time. An upper bound for the standard deviation of 1000Hz was applied. Since all formants lie within 1000Hz of each other, it is assumed that if a formant jumps more than 1000Hz between windows it was picked incorrectly. The calculated standard deviation was used for the confusion calculations in the decision component.

### Spectral Contrast

As the frequencies of the formants (spectral peaks) signal the identity of a vowel, the

differences in amplitude of peaks and valleys must be maintained to some degree for the human auditory system to interpret it correctly (Leek *et al.*, 1987). Even in the absence of significant noise, poor frequency resolution results in reduced definition of spectral peaks in a speech signal as the frequency regions become smeared together (Leek *et al.*, 1987).

Speech processed through normal auditory filters generally shows significant preservation of spectral peaks and valleys that serve to differentiate speech sounds. However, when speech is processed through a cochlear implant with broader auditory filters, a reduction in internal spectral contrast is produced. A smearing of the spectral information specifying the frequency locations of the formants then results (Summers and Leek, 1994; ter Keurs, Festen and Plomp, 1993a; ter Keurs, Festen and Plomp, 1993b).

Spectral contrast was used, therefore, as the second measure of the uncertainty of a person differentiating a vowel sound. Spectral contrast is defined as the height of the peaks of a formant in relation to the valleys around it (Leek *et al.*, 1987; Loizou and Poroy, 2001a). The higher this value is, the larger the spectral contrast becomes. It has been shown that high spectral contrast contributes to the accuracy of a person recognizing a vowel and that a reduction in spectral contrast creates confusion in the interpretation of vowel sounds (Sidwell and Summerfield, 1985; ter Keurs *et al.*, 1993b).

Figure 3.17 displays the LPC spectra of all windows through time superimposed on each other, in order to depict the spectral contrast for the entire clean vowel in the word “paat”. The blue lines in the graph represent the LPC spectra of individual windows through time. The thick red line shows the average LPC spectrum of the vowel. The first two peaks in the spectrum depict the first two formants. The valleys around these formants are used to measure the spectral contrast of each formant. For the first formant the spectral contrast is measured as the distance between the peak and the valley to the right of the peak. The valley to the left is not included because it assumedly does not contain any information that is necessary for a human to hear the first formant. The spectral contrast of the second and consecutive formants is measured as the distance between the peak of the formant and the average distance of the valleys both to the left and right of the formants.

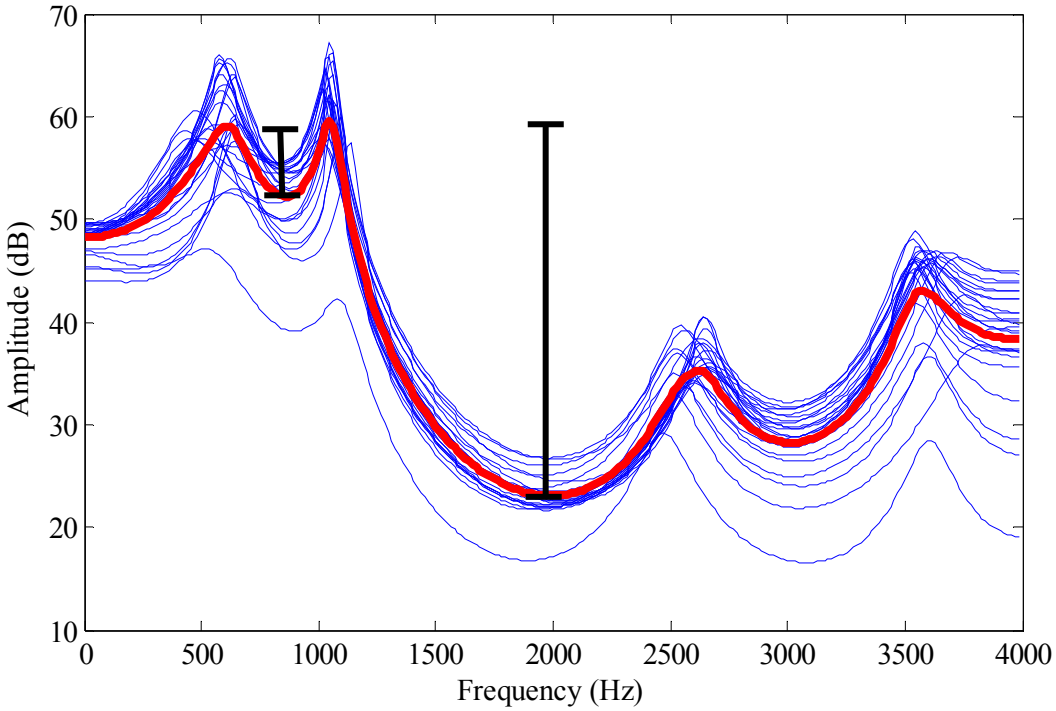


**Figure 3.17. LPC spectra showing the spectral contrast of the clean vowel (which has not been processed by the acoustic model) in the Afrikaans word "paat". The thick line represents the average of all windows (thin lines) through time of the vowel sound. The vertical bars represent the spectral contrast for the F1 and F2 frequency.**

Figure 3.18 shows that the spectral contrast between the valleys and peaks is reduced significantly, on average, after it has been processed through the CI model. The same measurements were used as described before. Experimentation showed that measuring the spectral contrast for each window individually and averaging this value does not give satisfactory results. This happens for the following reason: individually there may be good spectral contrast per window, but this contrast is very unsteady throughout. To overcome this, the LPC spectra of all the windows in the vowel was first averaged (shown as a thick red line in Figure 3.18.) The spectral contrast was then measured on the averaged spectrum instead of measured in the individual windows in time. Using this method was more pragmatic for the measurement of the spectral contrast.

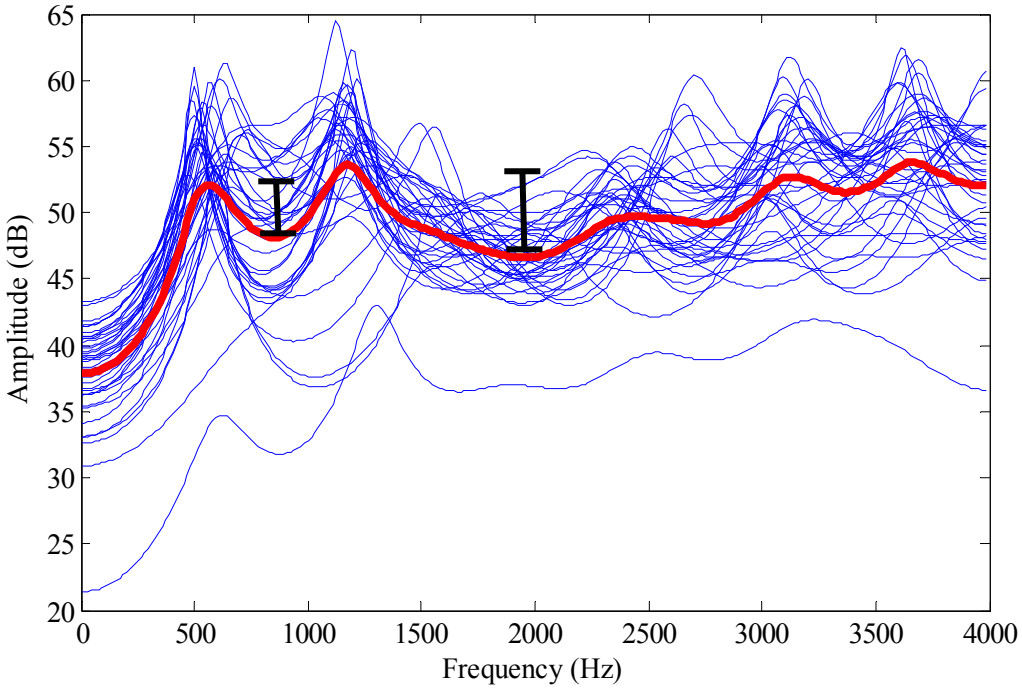


**Figure 3.18. LPC windows showing the spectral contrast of the processed vowel sound /ɑː/ in the Afrikaans word "paat". Vertical bars depict the average spectral contrast for the F1 and the F2 frequency.**

The spectral contrast was measured for each of the formants as the second type of uncertainty factor. This is used in the decision component to aid in the confusion measurements, described next.

### 3.4.2 Decision Component

Once all the necessary information for all the vowels had been gathered by the processing component, the decision component was used to predict the correct answers and possible confusions that could arise between the vowels being evaluated.

The decision component is based on a signal detection theory described in Gelfand (1990) and Green and Swets (1966). Signal detection theory provides a general framework to describe and study decisions that are made in uncertain or ambiguous situations. It is

commonly applied in psychophysics and in the most successful of the quantitative investigations into the processes of human decision-making and perception (Wickens, 2002).

The block diagram shown in Figure 3.18 summarizes the main functional steps that make up the decision component. The acoustic cues (those that were selected, as described previously) were used to determine perceptual distance measures in a vowel space of all the vowels presented. The position for each vowel was determined by the frequency of the first two formants and the duration of the vowel. These values were extracted from the degraded vowel sound being evaluated. The Euclidean distances between the vowels in terms of F1, F2 and duration were used to indicate the position of the vowels in the vowel space. The variances for each pdf were calculated from the uncertainty factors provided by the processing component. In calculating the probability of confusion, the greater the overlap between the pdfs the greater the probability of the vowels being confused. This is similar to the model developed in Svirsky (2000), however the variances in the present study were calculated and not estimated. The end product of the decision component is a confusion matrix which predicts the confusions a listener would experience in a subjective vowel confusion test.

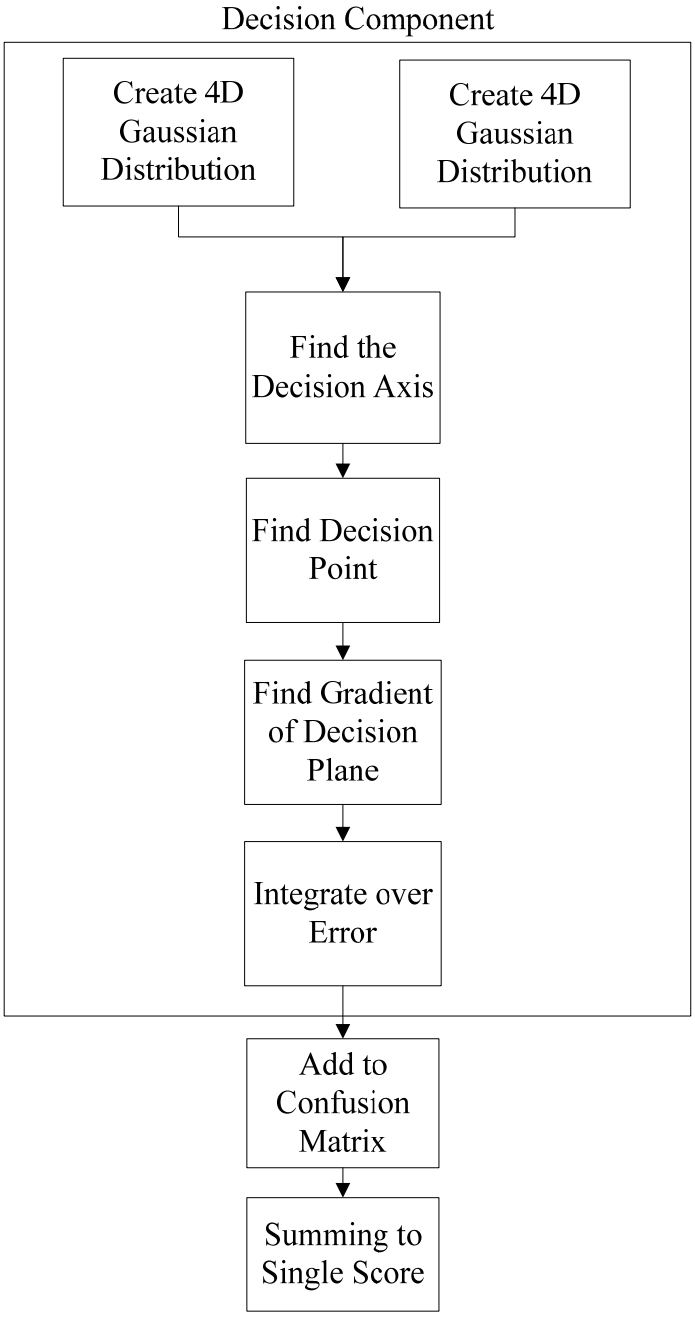


**Figure 3.19. Block diagram of the decision component.**

The operation of the decision component was based on statistical signal decision theory (Wickens, 2002). In signal detection theory, pdfs are used to represent stimuli and responses to the stimuli. To determine the probability of a correct answer, pdfs were generated for both the stimulus (the vowel being presented) and the listener's possible

response (the vowel perceived by the listener). The mean of the pdf was gained from the acoustic cues extracted in the processing component. The variance of the pdf was determined by the amount of noise in the given stimulus which masks the acoustic cues. The noise variance is directly proportional to the uncertainty that a person has in identifying a vowel. All pdfs were generated to have a Gaussian distribution. Although this was done for three variables in the model (creating a four dimensional pdf), for the sake of simplicity the explanation provided here will describe a scenario for a one dimensional variable (creating a two dimensional pdf). This will then be expanded later on in the chapter. The Gaussian distribution f(x) in one dimension is shown in Equation 3.9,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right), \tag{3.9}$$

where μ is the mean of the distribution and σ the variance.

A pdf was calculated for all the vowels being evaluated. The probability of the listener giving one specific response from all possible responses was calculated individually. These probabilities were calculated by integrating the tail of the pdf of the stimulus from a certain decision point (one dimensional example shown in shown in Figure 3.20). If no bias is assumed in the decision making process then this point is chosen as the crossing point of the stimulus pdf and the possible response that is being calculated. (See Section 3.4.2.2. of this dissertation for further information on how the decision point was determined.)

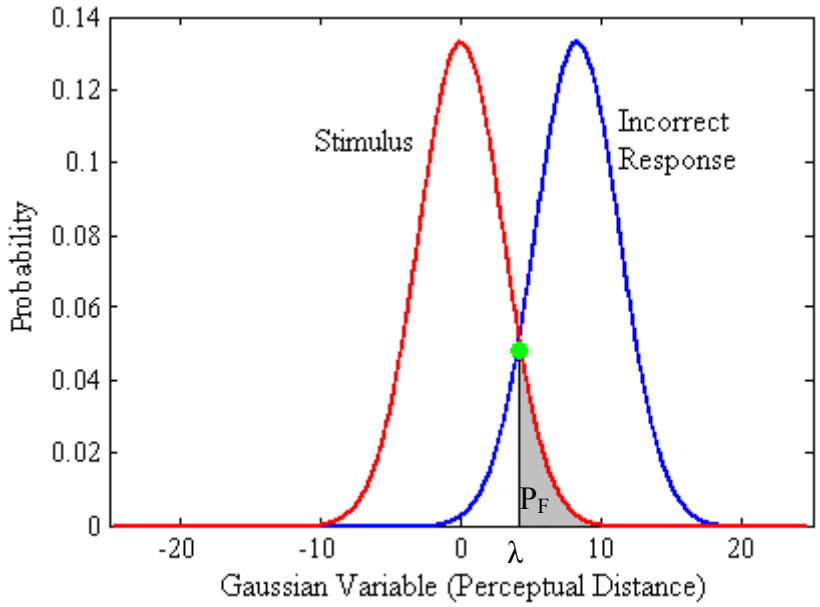


**Figure 3.20. Probability distribution functions representing the stimulus and response. The probability of giving the incorrect response, PF, is calculated from λ to the end of the stimulus pdf.**

Figure 3.20 is presented for illustration purposes only, and shows the case when only one random variable is used. The model will use three variables, one for each acoustic cue. The mean of the pdfs are the acoustic cue values (for example, the F1 frequency). The first pdf represents the stimulus (the vowel sound played) and the second pdf represents a possible response by a listener (one of the possible answers given by a listener). The variance of each pdf is determined from the uncertainty factor from the processing component (either frequency variation or spectral contrast). The cleaner the sound is, the better is the chance of the listener identifying the vowel correctly and the less is the variance in the pdf representing the vowel. The more uncertainty a person has in identifying a vowel, the larger the variance of the pdf becomes. Therefore, the more overlap there is between the two pdfs, the greater the probability of the listener responding in error. The distance between the centre points of the pdfs is determined from the Euclidean distance in terms of the formant frequencies F1 and F2, as well as the vowel duration.

The false response rate $P_F$ is the probability that an observation from the stimulus is incorrectly perceived. The shaded area in the distribution in Figure 3.20 corresponds to

this probability. It is assumed that no bias exists toward any vowel; therefore, the decision criterion λ was chosen as the point where the two distributions cross. The equation for $P_F$ can be written as follows (where $f_s$ is the pdf of the stimulus) .

$$\begin{aligned} P_F &= P(\text{Incorrect Response} \mid \text{Stimulus}) \\ &= \int_{\lambda}^{\infty} f_s(x)dx \end{aligned} \tag{3.10}$$

The probability of making all the possible responses is calculated in turn for each presented vowel. The probability of all responses being given for each presented vowel is then grouped into a confusion matrix. The probabilities are converted to a form that can be compared to confusion matrices resulting from subjective vowel confusion tests. A more in-depth explanation of each functional block in the decision component will be described next.

#### 3.4.2.1 Creation of 4D Gaussian Distribution Functions

In detection and identification models, the effect of a stimulus presentation is expressed by a univariate random value $X$ (as described above). This representation works well with a single stimulus; however, most signals incorporate more than one component carrying some information regarding the nature of a sound. The listener must integrate these various components into a single decision. For the present, objective model trivariate random variables were used for the selected acoustic cues, namely the F1 and F2 frequencies and the duration of the vowel.

The signal detection representation of a multidimensional stimulus is a generalization of the univariate representation. The effect of a single component stimulus corresponds to a random variable $X$. The effect of a three-component stimulus is described by a trivariate random variable $X = (X_1,X_2,X_3)$ and its instances by $x = (x_1,x_2,x_3)$. Instead of being represented as a point on a line (as in the description above) an observation is a point in three-dimensional space. The distribution of $X$ is expressed by a density function $f(\boldsymbol{x}) = (x_1,x_2,x_3)$ which associates a density with each point in the space. The Gaussian

distributions are expanded to trivariate form using the following equation (this will only be shown in matrix form for simplicity):

$$f(x) = \mathbf{C}_{\Sigma} \exp\left(-\frac{1}{2}\mathbf{Q}_{\Sigma}(x-\mu)\right), \tag{3.11}$$

where the constant $C_{\Sigma}$ and quadratic function $Q_{\Sigma}(\mathbf{x}\text{-}\mathbf{u})$ are given by the following matrices.

$$C_{\Sigma} = \left[(2\pi)^{d/2}|\Sigma|\right]^{-1} \quad \text{and} \quad Q_{\Sigma}(x) = x'\Sigma^{-1}x \tag{3.12}$$

The variable $\mathbf{\Sigma}$ represents the following covariance matrix.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 \end{bmatrix} \tag{3.13}$$

The random variables (F1,F2, vowel duration) are considered to be independent of each other. Hence, the correlation coefficient $\rho$ (representing the correlation between the random variables) have been set to zero. This simplifies the covariance matrix to the following:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \tag{3.14}$$

The density functions representing each vowel are characterized by their trivariate mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$. The means of the processed vowels for each distribution are based on the acoustic cues obtained in the processing component. If these three means were plotted, each vowel's trivariate mean could be plotted in a three-dimensional perceptual space. If all vowels are given the same variance in all three directions, it follows that the closer the vowels are to each other, the higher the probability of a person confusing them.

Figure 3.20 represents the modelled four-dimensional perceptual vowel space of a listener.

The centre points of the ellipsoids are defined by the duration of the vowel and its average F1 and F2 frequencies across the duration of the vowel sound. The sizes of the ellipsoids give an indication of the standard deviation of both formants. The figure shows the distances between the vowels and also gives an indication of which vowel is likely to be confused with another vowel. A representation such as Figure 3.20 does not provide the actual pdfs, but provides the reader with a clear picture of the distances between the vowels and the possible confusions between vowels. To show a pictorial representation of the pdfs would require a four-dimensional graphic.

The variances of the vowel, that is, the size of these ellipsoids in all three dimensions, are determined by the uncertainty that a listener has in hearing one of the three listening cues (namely, the vowel duration, the F1 frequency and the F2 frequency). The larger a particular uncertainty factor is, the larger the ellipsoid will be in the specific direction indicating that particular factor. According to the proposed model, the closer the ellipsoids representing the vowels are to each other, the larger is the probability that they will be confused with each other. In this model, if two ellipsoids do not intersect, there is a smaller probability that the vowels will be confused.

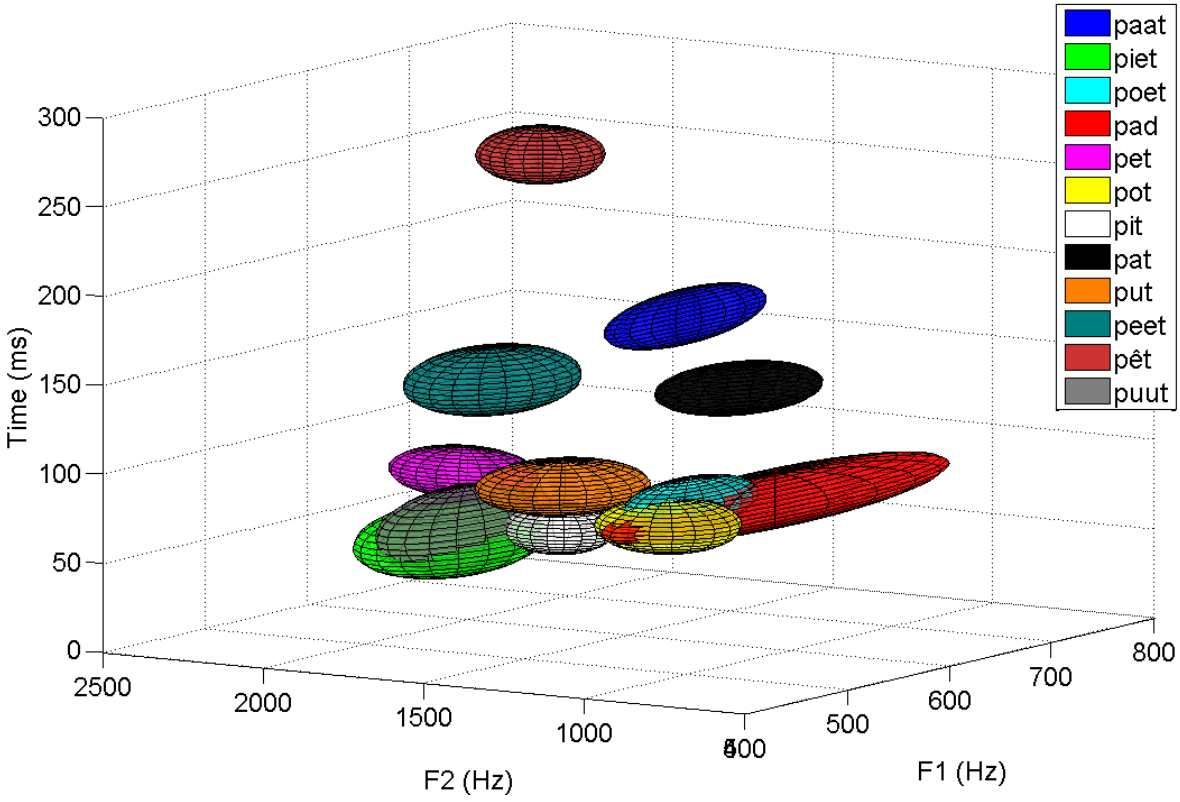


**Figure 3.21. A three-dimensional vowel space generated by the objective vowel prediction model**

As was mentioned earlier, the actual Gaussian distributed probability density functions used for determining the probability of vowels being confused exist in four dimensions (just as a one variable pdf requires a two-axis figure as representation, a trivariate pdf requires a four-axis figure). The three-dimensional vowel spaces only provide the reader with a representation of the trivariate random variables and their respective variances. From the vowel space in Figure 3.21, the model predicts which vowels will have a probability of being confused, for example, "puut" with "piet", and "pit" with "put."

The values from the vowel spaces are used to generate trivariate pdfs. The next steps are to find the decision axis and integration point between the stimulus vowel and the response so that the probability of providing the response can be calculated.

### 3.4.2.2 Finding the Decision Axis

The problem in calculating probabilities with multivariate information is to find a rule to reduce the multiple dimensions to a single decision. A common approach to solving this problem in signal detection theory is through geometry (Wickens, 2002). It is clear that the shortest distance between the centres of the two distributions lies along the line that connects their means. This line creates a natural decision axis. When the two multivariate distributions are projected onto this axis, they form two univariate distributions. These distributions can then be used to find the decision point. The decision point is the place at which the two pdfs intersect and will be used to find the plane from which the integration to the tail of the distribution of the stimulus will take place. Since the trivariate Gaussian distributions cannot be shown (they are four-dimensional), bivariate distributions will be used in figures to illustrate the concept. The equations given will be trivariate however, as used in the algorithm.

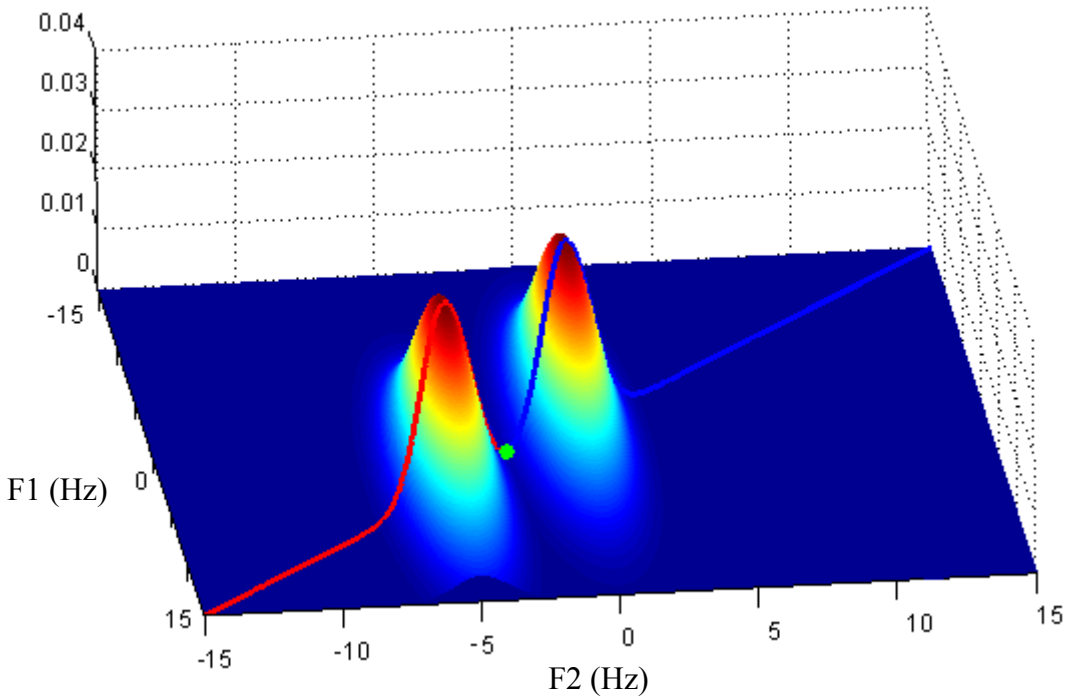


**Figure 3.22. Bivariate representation of using the line between the means of the distributions to find the decision criterion.**

Figure 3.22 shows how the decision plane is found. The red and blue lines show the univariate distributions along the line that connects the two means. This line represents the plane along which the bivariate pdfs (three-dimensional) collapse into univariate pdfs (two-dimensional). This procedure is done by finding the projection of both the pdfs onto the line between the two means, thus forming the function for the red and blue line in Figure 3.22. To do this, the vector between the two means is first calculated by

$$\mathbf{a} = < \mu_{S\mathbf{F1}} - \mu_{R\mathbf{F1}}, \mu_{S\mathbf{F}2} - \mu_{R\mathbf{F}2}, \mu_{SD} - \mu_{RD} >, \tag{3.15}$$

where $\mu_{S..}$ is the point of the mean of the Gaussian distribution of the stimulus in the F1, the F2 and the duration axes and $\mu_{R..}$ is the point of the mean of the Gaussian distribution of the response in the F1, the F2 and the duration axes.

Both distributions are then projected onto this vector in order to calculate the point of intersection. This is completed by finding the dot product of the distribution and the vector **a** as calculated by

$$P(\mathbf{a}) = P(F_1, F_2, D) \cdot \mathbf{a}, \tag{3.16}$$

where $P(\mathbf{a})$ is the univariate pdf along the vector **a**, $P(F_1,F_2,D)$ is the original trivariate pdf and **a** is the vector between the mean of the stimulus pdf and the response pdf.

Figure 3.23 is formed when these distributions are collapsed onto the vector between the two means. This allows for the decision criterion to be found as it would be found in univariate signal detection theory. It is assumed that there is no bias to either distribution: therefore, the decision criterion λ was chosen as the point where the two distributions cross in the graph. The crossing point between the two distributions, λ, minimizes the error even when the standard deviation between the two pdfs are not the same.

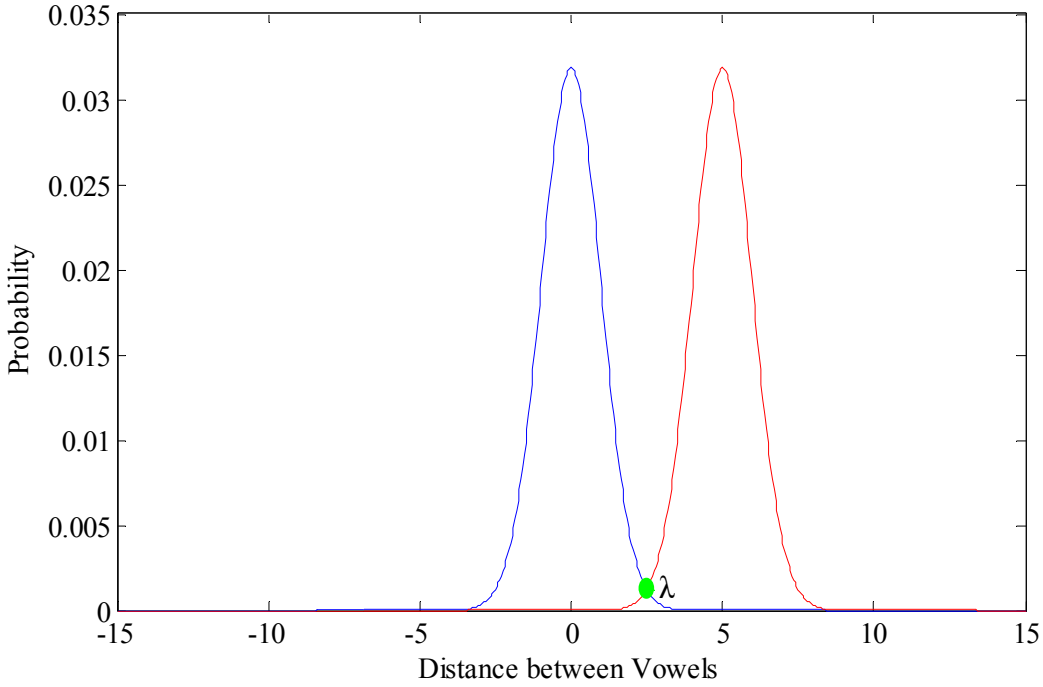


**Figure 3.23. Univariate distributions found along the line between the means of the bivariate distributions in Figure 3.22.**

### 3.4.2.3 Find Gradient of Decision Plane

In univariate form the decision criterion is a point from which integration can start to determine the probability of error. In the case of bivariate signal detection this point is extended to a line, whereas in the case of trivariate signal detection (as in this present model) the point needs to be extended to a plane.

The difficulty in finding the area of integration for a trivariate pdf, is that the orientation of the plane is not known and needs to be determined. The simplest approach is to choose the plane perpendicular to the line along the two means. This approximation performs accurately in an equal-variance model: it produces inaccurate results, however, when the three variances differ from each other. Figure 3.24 shows how an error is produced when setting the orientation of the plane orthogonal to the line between the means.

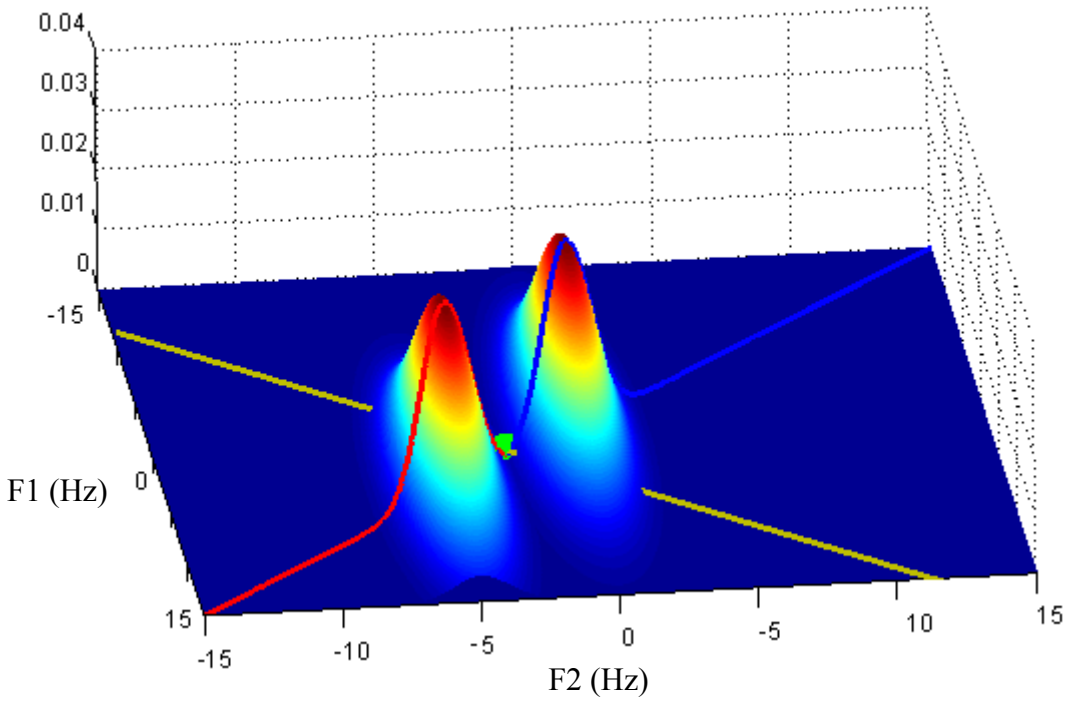


**Figure 3.24. 3-Dimensional representation of the error produced when choosing the decision line orthogonal to the line between the means of the distributions.**

The decision criterion calculated as the intersection between the univariate distributions along the decision line can be seen in Figure 3.23 as a green dot. The yellow line is calculated by finding the line orthogonal to the vector between the means of the two pdfs. This line is the place from which the integration to the tail of the distribution of the stimulus will take place in order to calculate the probability of error. By inspecting the figure, it becomes clear that there is almost no interaction between the two pdfs; therefore, the probability of error should also be very close to zero. Although there is very little interaction between the two Gaussian distributions, it is clear that the yellow line cuts into both distributions and, if the integration starts from this line, the predicted error will be larger than it should be.

This error in calculation is corrected by not choosing the plane perpendicular to the line between the means but, rather, finding the directional gradient of the distribution function in all three directions. Therefore, the tangent plane to the pdf is found at the point of decision (green dot in Figure 3.24). This is done by calculating the partial differential in

each direction using the following equations:

$$f_{F1} = \frac{\partial f}{\partial F1} P(F_1, F_2, D), \tag{3.17}$$

$$f_{F2} = \frac{\partial f}{\partial F2} P(F_1, F_2, D), \tag{3.18}$$

$$f_D = \frac{\partial f}{\partial D} P(F_1, F_2, D), \tag{3.19}$$

where $P(F_1,F_2,D)$ is the pdf for the stimulus distribution and $f_{F1}$, $f_{F2}$ and $f_D$ are the partial derivatives representing the slope in each direction. From these equations a tangent plane can be calculated by calculating the dot product of the partial derivatives and the tangent plane at the decision point λ. This is shown in equation 3.20.

$$TP = < \mathbf{p} - \boldsymbol{\lambda} > \cdot < f_{F1}(\lambda), f_{F2}(\lambda), f_D(\lambda) >, \tag{3.20}$$

where $TP$ is the new tangent plane function, $\mathbf{p}$ is the tangent plane vector, $f_{F1}$, $f_{F2}$ and $f_D$ are the partial derivatives in each direction and λ is the decision point.

By using the tangent plane the error can be minimized so that the least possible error is calculated since the integration to the tail happens at the point where the predicted error would be minimal. This method of finding the gradient will be referred to as the tangent method henceforth. Graphical portrayal in three dimensions (which means the plane is a line) is shown in Figure 3.25. (Remember that in the actual model the calculation is done for an integration plane for trivariate pdfs; therefore, in four dimensions the yellow line is actually a plane.)

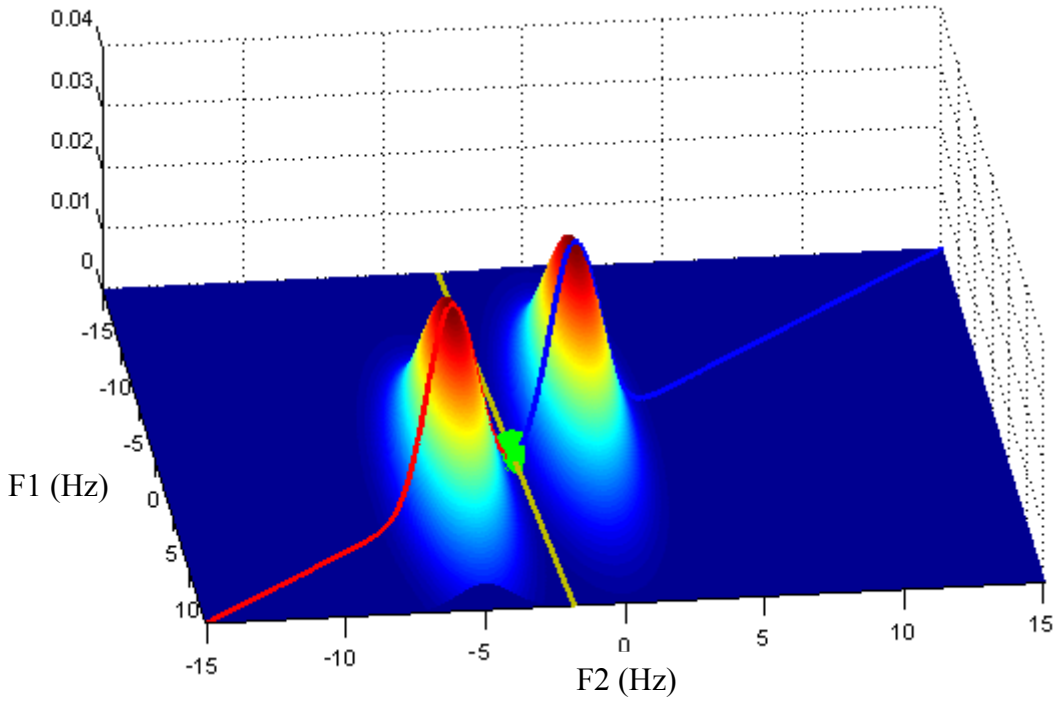


**Figure 3.25. 3-Dimensional representation of the error minimized by using the tangent method to find the decision line.**

Figure 3.25 gives a three-dimensional representation of how the tangent method improves on the perpendicular method. The yellow integration line shows how the error is fixed to give a more accurate answer to the interaction of the two distributions. The probability of confusing the stimulus with the response (error probability) is calculated by integrating to the tail of the pdf from the tangent plane using the equation

$$P(E) = \iiint_R P_S(F_1, F_2, D) dF_1 dF_2 dD \,, \tag{3.21}$$

where $P_S$ is the Gaussian probability density function of the stimulus in terms of the three cues (F1, F2 and duration). R, the region of integration, extends from the tangent plane (calculated above in Equation 3.20) to the tail of the pdf. For the examples shown in Figure 3.24 and Figure 3.25, it is clear by inspection that the probability of error is reduced by using the tangent method.

The probability of making an error is calculated using the method described in this chapter

for each vowel presented and each possible wrong answer given. Each of these results is then placed in a prediction confusion matrix. The confusion matrix shows the probability of a listener confusing the various vowel sounds. These confusion matrices are evaluated in the next chapter by comparing the results with confusion matrices obtained from subjective vowel confusion tests performed with normal hearing listeners listening to an acoustic CI model.

## 3.5 EXPERIMENTAL STUDY

To assess the newly developed vowel prediction model it was compared with results from a subjective vowel confusion test conducted with human participants. Care was taken in using the same speech stimulus material in testing the model that was used in the original subjective test. In order to achieve this, results from the experiments done by Conning and Hanekom (2005 unpublished) were used; thus the original speech material was available together with the results. The objective in the study by Conning and Hanekom was to evaluate the CI acoustic model, the same model used in this study. The details from the subjective experiments are given below.

### 3.5.1 Listeners

The processed speech segments were presented to seven female listeners and three male listeners. All the listeners were normal hearing people between the ages of 19 and 26, with Afrikaans as their home language.

### 3.5.2 Stimuli

The acoustic model was used to process 12 /CVC/ vowels (in the context of ‘p’-VOWEL-‘t’). The original speech, spoken by an Afrikaans male speaker, were recorded at 44.1 kHz (16 bit resolution) in a double walled sound booth at the University of Pretoria. A high quality Sennheiser microphone was used for the recordings. The stimuli presented were the Afrikaans vowels /æ/ (pat), /a/ (pad), /u/ (poet), /œ/ (put), /y/ (puut), /e/ (peet), /ɑː/ (paat), /i/ (piet), /ə/ (pit), /ɔ/ (pot), /ɛː/ (pêt) and /ɛ/ (pet). Each phoneme was played, in

random order, to the listener through a loudspeaker at an average sound pressure level of 70 dB SPL. The listener had to select (from a computer screen) which vowel sound he/she recognized each stimulus to represent. Each stimulus was repeated ten times for each experimental condition. The latter are described below. Since normal hearing listeners had to become accustomed to the sound of the CI model, the experiments were repeated five times and only the results from the fifth experiment were used.

### 3.5.3 Experimental Conditions Investigated

#### 3.5.3.1 Acoustic CI Model

The acoustic CI model used for the experiments was designed to closely follow the speech processor used in a cochlear implant and it also implemented the biophysical features that affect the signal in the cochlea. The flow diagram for this model is shown in Figure 3.26**.** In essence the model breaks the signal into 20 frequency bands. Each of these bands corresponds to an electrode in the electrode array situated into the cochlea. The biophysical part simulates the interaction between the electrodes and the nerve fibres in the cochlea.

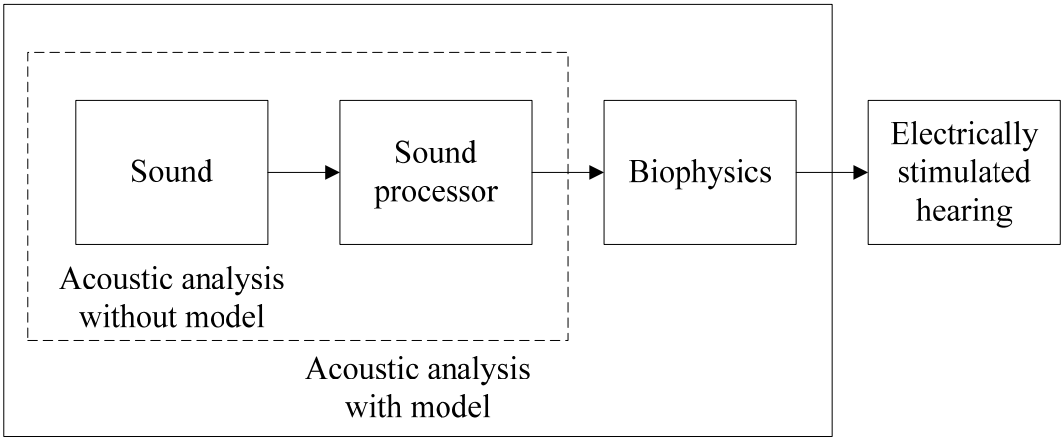


**Figure 3.26. Flow diagram showing the value of the acoustic model.**

The sound processing section of the acoustic model transforms the signals in the same manner that the processor does in a cochlear implant that uses the SPEAK strategy. For this strategy the incoming speech signal is divided into 20 frequency bands and the 8 channels with the highest energy content are used. To reconstruct the speech signal, band-limited noise bands are used as an approximation to the spread of current in the cochlea. All the channels are quantised to the stimulation current values and then all the active

channels are summed to provide output in the form of an audio file.

The acoustic CI model allows the user to set the number of channels, type of signal processing, dynamic range compression of the stimulus current and the insertion depth of the electrodes. For the simulations, the acoustic CI model was set to eight channels at an insertion depth of 25 mm with the SPEAK cochlear implant processing strategy.

#### 3.5.3.2 Background Noise Conditions

In the first condition, the experiment was conducted with the cochlear implant acoustic model with no background noise. In a second condition, Noise was added to the input speech signal to determine how speech recognition would deteriorate in the presence of noise.

Multi-talker babble noise was used to test different listening conditions because of its superior masking effects on speech (Dubno, Horwitz and Ahlstrom, 2005; Ferguson and Kewley-Port, 2002; Friesen, Shannon, Baskent and Wang, 2001; Fu *et al.*, 1998a; Killion, Niquette, Gudmundsen, Revit and Banerjee, 2004; Müller, Schön and Helms, 2002; Nie, Stickney and Zeng, 2005; ter Keurs *et al.*, 1993b; Yang and Fu, 2005). Existing .wav files of noise signals were used with permission from E. Hennix[1]. The multi-talker babble has formants in the same frequency area as speech, thus simulating a more realistic hearing environment for cochlear implant users. Three different noise conditions were tested along with the without noise condition. The noise conditions 40dB, 20dB and 0dB SNR. These were chosen to investigate how the speech recognition deteriorates with increasing noise, and whether this predicts findings with listeners.

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

#### 3.5.3.3 Evaluation of Results

Confusion matrices were created from the subjective tests using custom software. These confusion matrices were analyzed through FITA (Feature Information Transmission Analysis) investigation to determine the effect of noise on the transmission of speech cues to the auditory system. FITA analysis is a method that determines the amount of information transmitted (for present purposes, the information contained in the three acoustic cues) and conclusions are drawn as to which characteristics are transmitted most effectively with the acoustic simulation. The process of obtaining the FITA will be explained in the results chapter.

The accuracy of the objective prediction model was assessed by evaluating the same speech files using the model. All the .wav files used in the subjective tests were processed through the new model to create the prediction confusion matrices. In all conditions the clean speech was used as the clean input to the model so as to serve as a reference. Noise was added to the input processed through the CI model. Four different conditions of multi-babble background noise were tested (no noise, 40 dB, 20 dB, 0 dB).

### 3.5.4 Summary

In this chapter, the methodology implemented to develop a new vowel intelligibility prediction model was reported. A step-by-step guide was given as to the steps followed in the development of the model. Considerations made during the development of the model were recorded here, as well as proposed solutions given. This chapter also briefly explained how the results from previous subjective testing were obtained. The results from the subjective and objective experiments are compared using confusion matrices and FITA analysis (as reported in the next chapter). These results are compared separately for each SNR level to evaluate the developed model.

# CHAPTER 4 RESULTS

## 4.1 CHAPTER OBJECTIVES

The vowel prediction model is evaluated in this chapter. Comparison is made between the confusion matrices obtained in the subjective vowel tests and those obtained when running the same speech material through the objective model. These experiments were performed first in the condition of no channel noise added and then with specific increments of added multi-talker babble noise to determine the applicability of the algorithm under various conditions, as explained in the previous chapter. Vowel spaces are presented to give the reader a visual representation of how the confusion matrices were generated. The percentage information correctly transmitted by the major acoustic cues for the various tests is compared and discussed.

The objective model was implemented with two different uncertainty factor calculations. The first implementation used variation in formant frequency and the second implementation used spectral contrast to calculate this uncertainty. These models will be referred to as the Frequency Variation Model and Spectral Contrast Model respectively. Both of these possible factors underlying confusions will be evaluated in this chapter. The chapter will be concluded with a summary of the comparative results of the Frequency Variation Model and the Spectral Contrast Model.

### 4.1.1 Formant Frequencies and Duration

The formant frequency and vowel duration acoustic cues present the model with the information needed to calculate predictions. Table 4.1 shows the formant frequencies for each of the vowels that were used for testing the model. The first and second formant frequencies were determined from the vowels' spectrograms by using the software program PRAAT (Boersma and Weenink, 2001). PRAAT has functionality to determine the mean formant frequencies of a vowel token and allows for visual inspection of the spectrogram of each vowel. The data in the table will serve as a reference to aid in the

analysis of the confusion matrices.

**Table 4.1. Values for vowel duration (ms), F1 frequency (Hz) and F2 frequency (Hz) of selected phonemes as measured with PRAAT**

| | | Original vowels | | | Vowels processed through CI model | | |
|---|---|---|---|---|---|---|---|
| | | Duration (ms) | F1 (Hz) | F2 (Hz) | Duration (ms) | F1 (Hz) | F2 (Hz) |
| pAAt | ɑː | 218 | 765 | 1074 | 226 | 747 | 1266 |
| pIEt | i | 87 | 258 | 2031 | 70 | 540 | 1955 |
| pOEt | u | 84 | 319 | 1057 | 67 | 440 | 1120 |
| pAd | a | 100 | 783 | 1143 | 87 | 740 | 1174 |
| pEt | ɛ | 87 | 508 | 1966 | 88 | 541 | 1941 |
| pOt | ɔ | 102 | 525 | 954 | 118 | 540 | 1100 |
| pIt | ə | 73 | 479 | 1588 | 69 | 598 | 1644 |
| pAt | æ | 135 | 664 | 1506 | 104 | 690 | 1430 |
| pUt | œ | 92 | 508 | 1524 | 84 | 526 | 1502 |
| pEEt | eː | 198 | 337 | 2104 | 132 | 441 | 1955 |
| pêt | ɛː | 274 | 416 | 1904 | 228 | 526 | 1756 |
| pUUt | y | 91 | 285 | 2069 | 71 | 484 | 2034 |

The acoustic model of the cochlear implant shifts the values of the acoustic cues to a certain extent. This is especially evident in the F1 and F2 frequencies. The duration cue is still very similar for most of the vowel sounds. Plotting a two-dimensional vowel space of the F1 and F2 frequencies for both the original vowels and the vowels processed through the CI model provides better insight into how these spectral shifts might affect vowel identification (see Figure 4.1).

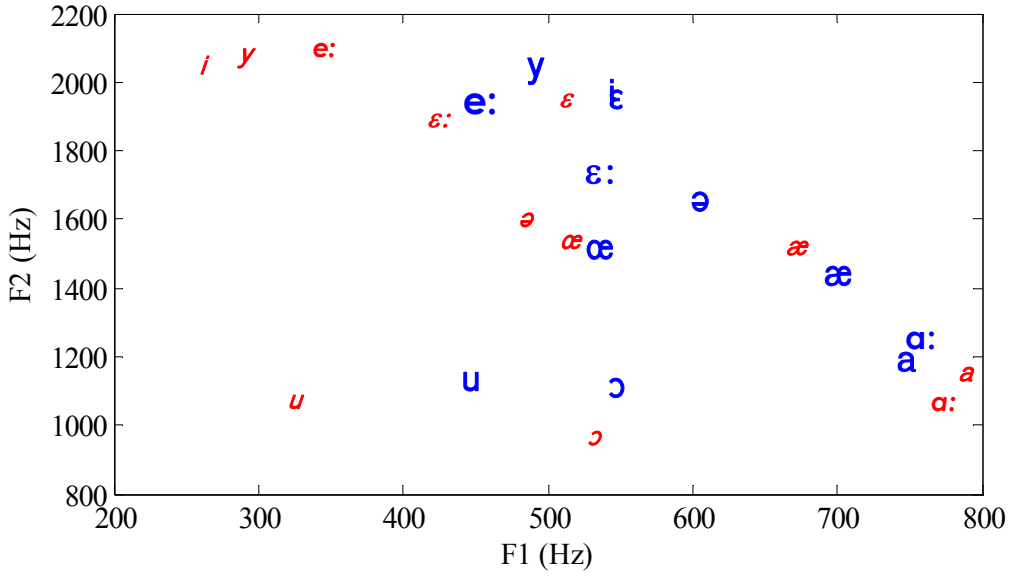


**Figure 4.1. Two-dimensional vowel space of the formant frequencies of original vowels (small italic font) and the same vowels after being processed by the acoustic model (large font).**

The information in Figure 4.1 shows that vowel separation in the vowel space is different for a cochlear implantee compared to a normal hearing listener, at least when using the present acoustic model The most evident difference is that there is no processed vowel which has a F1 frequency lower than 440 Hz. All F1 frequencies are moved up to or past this frequency. It is also evident that the vowel space for the processed vowels lies in a much smaller area than the original vowels. Only looking at the formant frequencies (and ignoring the masking of noise) it is evident that there is already a greater probability of vowel confusion for a cochlear implantee.

When the vowel space is viewed in terms of the F1 and F2 frequencies, 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 similarities, it may be predicted that these vowels will be confused frequently.

What cannot be seen from the vowel space above is the masking effect of noise on the

availability of the acoustic cues to the listener. For the original vowel the formant frequency is stable with a high spectral contrast making it easy to hear. The introduction of noise by the cochlear implant (by, for example, decreasing spectral contrast) plays a major role in the transmission of these acoustic cues to the listener. This factor causes vowel identification to degrade as additional noise is added to the vowel. This can be seen in the confusion matrices in the rest of the chapter. Examples of the spectral effects of the noise introduced by the cochlear implant can be seen in the spectrograms and LPC spectra in the previous chapter.

### 4.1.2 Accuracy of Acoustic Cue Tracking

The predictions of the objective model depend on the correct extraction of the acoustic cues from the vowel sounds, especially in the case where the vowels have been degraded by the CI acoustic model. The following two tables show comparisons between the values extracted by the objective model and the values obtained from inspection (Table 4.1).

**Table 4.2. Comparison of the values of the acoustic cues (vowel duration, F1 frequency and F2 frequency) for clean vowels which have not been processed by the CI model.**

| | Inspection | | | Objective Method | | |
|---|---|---|---|---|---|---|
| | Dur (ms) | F1 (Hz) | F2 (Hz) | Dur (ms) | F1 (Hz) | F2 (Hz) |
| **paat** | 218 | 765 | 1074 | 208 | 609 | 1047 |
| **piet** | 87 | 258 | 2031 | 80 | 297 | 2047 |
| **poet** | 84 | 319 | 1057 | 88 | 328 | 1109 |
| **pad** | 100 | 783 | 1143 | 96 | 609 | 1094 |
| **pet** | 87 | 508 | 1966 | 104 | 453 | 2016 |
| **pot** | 102 | 525 | 954 | 112 | 484 | 844 |
| **pit** | 73 | 479 | 1588 | 72 | 484 | 1578 |
| **pat** | 135 | 664 | 1506 | 128 | 594 | 1531 |
| **put** | 92 | 508 | 1524 | 96 | 484 | 1531 |
| **peet** | 198 | 337 | 2104 | 192 | 313 | 2094 |
| **phet** | 274 | 416 | 1904 | 272 | 422 | 1938 |
| **puut** | 91 | 285 | 2069 | 96 | 297 | 2000 |

Table 4.2 shows the measured values for the three acoustic cues (vowel duration, F1 frequency and F2 frequency) for the clean vowels. On the left are the values as measured by inspection and analysis in PRAAT. On the right are the values as measured by the

objective model. The average absolute differences for the duration, F1 frequency and F2 frequency are 6 ms, 49 Hz, and 31 Hz, respectively. This is less than a one percent difference on average for each of the acoustic cues. The maximum error found was in the word "pad" where there was a 22% difference for the F1 cue. This shows how difficult it is to determine the precise location of the formants in the degraded speech. Apart from this exception, the model performs adequately in cue extraction for clean vowels presented with no additional background noise.

**Table 4.3. Comparison of the values of the acoustic cues (vowel duration, F1 frequency and F2 frequency) for clean vowels that have been processed by the CI model.**

| | Inspection | | | Objective Method | | |
|---|---|---|---|---|---|---|
| | Dur (ms) | F1 (Hz) | F2 (Hz) | Dur (ms) | F1 (Hz) | F2 (Hz) |
| **paat** | 226 | 747 | 1266 | 208 | 563 | 1156 |
| **piet** | 70 | 540 | 1955 | 80 | 500 | 1766 |
| **poet** | 67 | 440 | 1120 | 88 | 547 | 1141 |
| **pad** | 87 | 740 | 1174 | 96 | 656 | 1141 |
| **pet** | 88 | 541 | 1941 | 104 | 531 | 1781 |
| **pot** | 118 | 540 | 1100 | 112 | 531 | 1109 |
| **pit** | 69 | 598 | 1644 | 72 | 516 | 1641 |
| **pat** | 104 | 690 | 1430 | 128 | 688 | 1438 |
| **put** | 84 | 526 | 1502 | 96 | 516 | 1438 |
| **peet** | 132 | 441 | 1955 | 192 | 516 | 1859 |
| **phet** | 228 | 526 | 1756 | 272 | 563 | 1656 |
| **puut** | 71 | 484 | 2034 | 96 | 516 | 1688 |

Table 4.3 shows the values for the three acoustic cues measured for vowels that have been processed by the CI model. Accurate measurement of these vowels provides a challenge because of the amount of noise present in the spectrum of the vowel. Again the values measured by inspection are presented in the left column of the table and the values extracted by the objective model are shown in the right column. The average absolute differences for the vowel duration, F1 frequency and F2 frequency are 21 ms, 56 Hz, and 95 Hz, respectively. This is an average of 99% precision or more for the formant frequencies and more than 98% accuracy for the duration of the vowel. These values were calculated by finding the absolute difference between the cues for the objective extraction and for inspection for each vowel individually and calculating the average for each cue.

Once again, the first and second formant frequencies were determined from the vowels' spectrograms by using the software program PRAAT. The acoustic cues are, therefore, properly tracked by the objective vowel prediction model.

### 4.1.3 FITA Analysis

FITA analysis is used in the rest of the chapter to compare the objective model's confusion matrices to the subjective tests' confusion matrices. The FITA analysis is obtained by using the formant frequencies and the duration of each of the vowels presented to the listener. FITA analysis was done for the vowel recognition confusion matrices to determine which acoustic cues were transmitted most effectively and to determine whether listeners use the acoustic cues to the same extent under the various conditions presented in this study. The output of the FITA analysis is a measure of covariance between input and output. This measure is calculated the following procedure.

If the input variable is x with probability $p_i$, i = 1, 2, ..., k, the mean logarithmic probability (MLP) is defined as

$$MLP(x) = E(-\log p_i) = -\sum_i p_i \log p_i \ . \tag{5.1}$$

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

$$T(x;y) = MLP(x) + MLP(y) - MLP(xy) = \sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}}, \tag{5.2}$$

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 authentic and T(x;y) will be near unity (Miller and Nicely, 1955).

FITA analysis requires that the information (the acoustic cues in this instance) that is to be evaluated is grouped in categories for analysis. The classifications were different for the

processed and original vowels because of the shifting of the acoustic cues as shown in Table 4.3 (Pretorius, Hanekom, Van Wieringen and Wouters, 2005). The acoustic cues are classified as shown in Table 4.4 and Table 4.5. The classifications of the vowels were determined using the guideline summarised in Table 4.6. The vowels were grouped together according to their classifications to determine the percentage information transmitted for a specific characteristic. The confusion matrices were analysed using these classifications to determine whether long vowels could be distinguished from short vowels, vowels with low F1 frequencies could be distinguished from vowels with high F1 frequencies, and so forth.

**Table 4.4. 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 4.5. 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 4.6. Ranges of vowel duration, F1 frequency and F2 frequency used for the classification of processed vowels.**

| | Duration | F1 | F2 |
|---|---|---|---|
| 1 | 0 - 100 | 0 - 540 | 0 - 960 |
| 2 | > 100 | 541 - 900 | 960 - 1700 |
| 3 | | > 900 | >1700 |

## 4.2 FREQUENCY VARIATION MODEL

The results for the Frequency Variation Model are described in this section. The results obtained from the test that was conducted with speech to which no additional noise had been added are presented first. Thereafter, the results from the tests done in the presence of different SNR levels of multi-talker babble noise follow. For each test the confusion matrix from the subjective test is shown along with the prediction confusion matrix generated by the objective model. Section 3.4 explained the procedure that was followed in the particular study and which experimental parameters were set in the subjective tests. The results of the subjective tests serve as a reference against which the new objective algorithm can be evaluated. The confusion matrices are compared using FITA analysis.

### 4.2.1 Speech Without Additional Background Noise

Figure 4.2 gives a representation of the possible confusions between the vowels that were presented to the listener without additional background noise. This figure was generated by the objective evaluation algorithm from the acoustic cues and their respective uncertainty factors. Each ellipse in the figure represents a vowel sound. The centre of each of these lie at the vowel's measured acoustic cues. The size of the ellipse indicates its uncertainty factor. In the Frequency Variation Model the uncertainty factors for F1 and F2 are their standard deviations in terms of frequency.

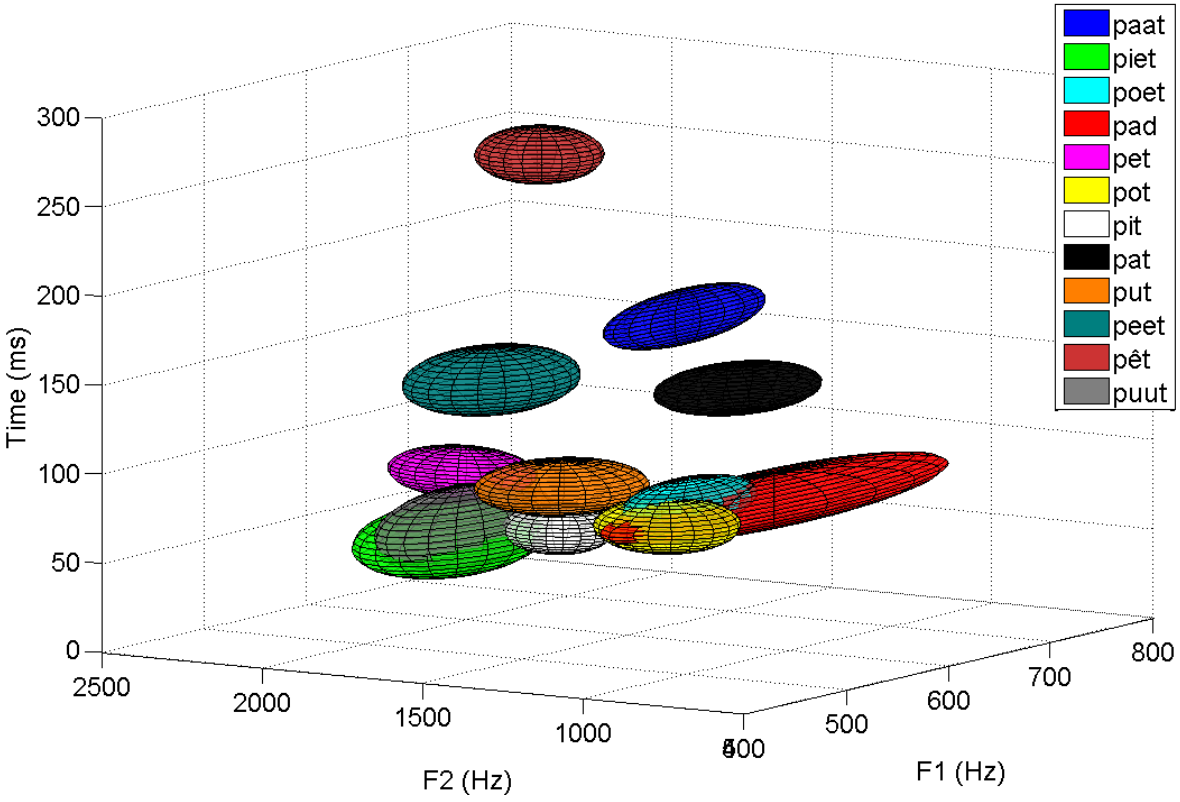


**Figure 4.2. Perceptual vowel space (generated using the Frequency Variation Model) for the processed vowels, with no background noise.**

From the above vowel space it can be predicted that the vowels with a longer duration (above 100 ms), that is, the vowels in the words “paat”, “peet”, “pat”, and “pêt”, have little probability of being confused with other vowels. The vowels of duration between 60 ms and 100 ms all lie close together and are separated only by their formant frequencies. The sizes of all the ellipses for all the vowels are generally the same except for ‘pad’. For the word ‘pad’, the measurement of the F1 frequency has a very large standard deviation. This is so because its first two formant frequencies lie very close together, which causes the CI model to merge them into one in some of the windows, causing the objective algorithm to pick the formants incorrectly. Assumedly, the implantee would do the same.

A confusion matrix was compiled by recording the response of a listener to a specific stimulus. The values lying along the diagonal of the matrix represent the stimuli that were recognised correctly, while incorrect responses are scattered across the matrix. By

examining these matrices, typical confusions between vowels were determined. The matrix shown in Figure 4.3 has been summed from all the listeners in the subjective test. There is a total of 200 answers for each presented stimulus. Following the first matrix, the prediction matrix (Figure 4.4) that was generated by the objective model is shown.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 197 | 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 | eː | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 199 | 0 | 0 |
| pêt | ɛː | 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 |
| | | | | | | | | | | | | Average Correct | **135** |

**Figure 4.3. Confusion matrix obtained from the results of the subjective test for vowels with no background noise.**

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 197 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| pIEt | i | 0 | 45 | 3 | 3 | 31 | 8 | 43 | 1 | 18 | 3 | 0 | 45 |
| pOEt | u | 0 | 1 | 98 | 15 | 1 | 43 | 20 | 3 | 17 | 1 | 0 | 1 |
| pAd | a | 0 | 4 | 50 | 50 | 3 | 50 | 13 | 6 | 18 | 1 | 0 | 4 |
| pEt | ɛ | 0 | 11 | 3 | 1 | 104 | 8 | 29 | 1 | 22 | 6 | 0 | 14 |
| pOt | ɔ | 0 | 4 | 63 | 12 | 8 | 63 | 29 | 1 | 16 | 1 | 0 | 4 |
| pIt | ə | 0 | 9 | 14 | 3 | 18 | 19 | 109 | 0 | 17 | 1 | 0 | 11 |
| pAt | æ | 1 | 2 | 8 | 6 | 3 | 2 | 1 | 152 | 10 | 11 | 0 | 4 |
| pUt | œ | 0 | 8 | 20 | 6 | 30 | 22 | 46 | 4 | 48 | 5 | 0 | 12 |
| pEEt | eː | 2 | 5 | 3 | 1 | 25 | 3 | 6 | 14 | 17 | 113 | 0 | 10 |
| pêt | ɛː | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 200 | 0 |
| pUUt | y | 0 | 31 | 2 | 2 | 37 | 9 | 46 | 2 | 21 | 5 | 0 | 46 |
| | | | | | | | | | | | | | **102** |

**Figure 4.4. Prediction confusion matrix (generated by the Frequency Variation Model) for vowels with no background noise.**

The average percentage of correctly recognised vowels (pooled over all listeners) in the subjective test is 67.5%. This percentage is expected to decrease when background noise is added to the stimuli. The vowels with longer duration are predicted to have a high probability of being correctly recognized. The fact that the confusion matrix is not symmetrical (a given vowel might be confused with another vowel, but the percentage of confusions depend on the presentation order) increases the complexity of predicting the confusions. The results in the confusion matrix are categorized and summarized in Table 4.7. This was be done for all the confusion matrices to help the reader in comparing the matrices.

The matrix in Figure 4.4 was generated by the Frequency Variation Model. The answers have been scaled so that there is a total of 200 answers for each presented stimulus to allow for easier comparison with the subjective model. The overall percentage that was correctly recognised for the vowels presented is 50.5%. This is much lower than the 67.5 % obtained in the subjective test. The confusions in the matrices are summarized in Table 4.8; once again this is done to make the comparisons easier.

**Table 4.7. Summary of the confusion matrix from subjective testing with no background noise.**

| **Best recognized (>80%)** | | **Well recognized (50-80%)** | | | **Poorly recognized (<50%)** | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Paat | 98.5% | Piet | 70.5% | Puut | Poet | 46.5% | Pot, Pad |
| Pad | 86.5% | Pot | 68.5% | Poet | Pet | 42.5% | Piet, Pit |
| Pat | 90.0% | Pit | 58.0% | Put | Put | 30.5% | Pit, Pot, Pet |
| Peet | 99.5% | | | | Puut | 32.0% | Piet |
| Pêt | 85.5% | | | | | | |

From Table 4.7 it is apparent that the vowels that were recognized best are the ones that are of a longer duration than the rest. This is in agreement with the results found in a study by (Van Wieringen and Wouters, 1999). The original vowels for “paat”, “pad”, “pat”, “peet”, and “pêt” all have durations of more than 100 ms. The only vowel with a short duration

that was recognized very well with a shorter duration is the vowel in "pat". The poorly recognized vowels all had similar durations around 90 ms. They were generally confused with vowels that have similar formant frequencies, for instance "poet" with "pot" and "puut" with "piet".

**Table 4.8. Summary of the prediction confusion matrix from the objective model for vowels with no background noise (Frequency Variation Model).**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pêt | 100.0% | Peet | 56.5% | Pet | Poet | 49.0% | Pot, Pit |
| Paat | 98.5% | Pit | 54.5% | Pot, Pet | Pot | 31.5% | Poet |
| Pat | 76.0% | Pet | 52.0% | Pit | Pad | 25.0% | Pot, Poet |
| | | | | | Put | 24.0% | Pit, Pet |
| | | | | | Puut | 23.0% | Pit |
| | | | | | Piet | 22.5% | Puut, Pit |

In Table 4.8 above, compiled from the results of the objective model, three of the five vowels in the subjective test that fell in the best recognized category were correctly predicted to fall in the same category as the subjective test. Again the best recognized vowels were the ones with longer duration. The other two vowels, those in "peet" and "pad", that fell in the best recognized category for the subjective test were incorrectly predicted to have a much lower correct percentage at 56.5% and 25%, respectively. In the well recognized category, only the vowel in "pit" was close to being correctly predicted, although the confusion was incorrectly predicted to be "pot" and "pet" instead of "put" as seen in the subjective test results. Some of the poorly recognized vowels have close correlations to the subjective test's results, for instance "poet", "put", and "puut". Most of the rest of the vowels have lower correct recognition percentages in comparison to the subjective test. This analysis seems to indicate that the Frequency Variation Model produces results which are best described by the duration cue. It seems that the use of frequency variation in the model causes the model to inaccurately predict vowel confusions.

The FITA analyses for the confusion matrices of the subjective test and the objective model are given in Table 4.9. The objective test is referred to as the frequency method in the table to indicate that the Frequency Variation Model was used.

**Table 4.9. Results of FITA analysis for the pooled answers in the subjective test and for the Frequency Variation Model implemented in the objective test (no background noise).**

| % information transmitted | Freq. Method | Subj. Method |
|---|---|---|
| **Duration** | 31% | 60% |
| **F1** | 22% | 43% |
| **F2** | 32% | 57% |

FITA analysis shows that the percentage of information transferred for each acoustic cue is close to double in the subjective test compared to that in the objective model prediction. The vowel duration and F2 frequency are the cues that are transmitted best in both tests. The F1 frequency contains the least information that is transmitted to the listener. It shows that the algorithm follows the trend that the duration and F2 frequency of vowels were transmitted most effectively and F1 information was transmitted poorly. From the above table it is apparent, however, that the Frequency Variation Model produces much lower information transmitted for all three of the acoustic cues for vowels presented with no additional background noise.

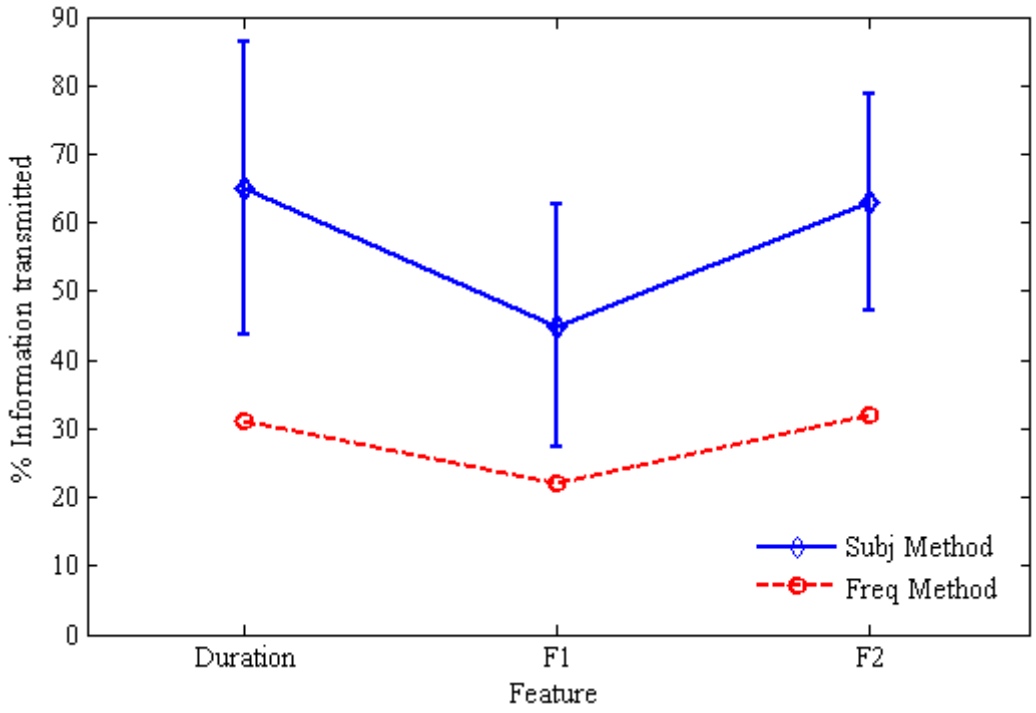


**Figure 4.5. Graph of FITA analysis for the subjective test and the objective test (Frequency Variation Model) performed with no additional background noise. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

Figure 4.5 shows a comparison between the information transmitted in the major acoustic cues in the subjective and objective tests. The FITA analysis for the subjective test was obtained from the individual confusion matrices for each person. The averages for each acoustic cue are calculated and plotted, and the standard deviation (shown as error bars) in the answers between the listeners is given. Since there is only one confusion matrix for the objective model, no standard deviation can be calculated for this model.

Although the percentage information transmitted for the objective model is much lower than it is for the subjective model, from the FITA analysis it can be seen that the trend between the cues is similar. The averages for each of the acoustic cues are between 19% and 24% less for the objective model than for the subjective tests. None of the percentages for any of the cues fall inside the bounds set by the error bars of the subjective model. Therefore, there is not enough information that is transmitted correctly for any of the

acoustic cues in the objective test. This implies either that the model requires more acoustic cues in order to produce correct results, or that the calculated uncertainty factors play too large a role in creating confusions between presented vowels.

### 4.2.2 Speech at 40dB SNR (Multi-Talker Babble Noise)

This test used the same speech tokens that were used in the test without additional background noise, except that background noise was added to the speech at a SNR of 40dB. Multi-talker babble was used as the additional noise since it simulates closely everyday environments that cochlear implantees need to communicate in. Multi-talker babble refers to nonsensical chatter originating from various speakers simultaneously. The multi-talker babble has the same spectral characteristics as normal speech. More information on the experimental set-up can be found in section 3.5.

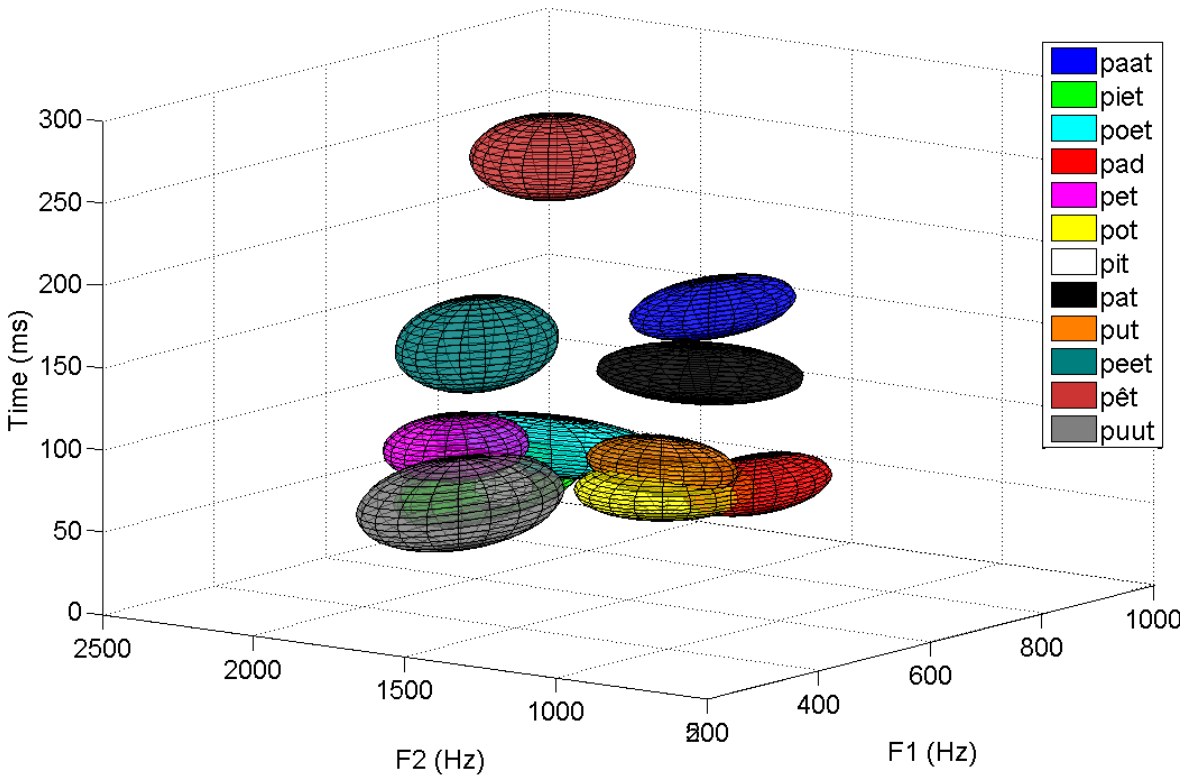


**Figure 4.6. Perceptual vowel space (generated using the Frequency Variation Model) for the processed vowels with added babble background noise at 40 dB SNR.**

Figure 4.6 presents the vowel space generated by the acoustic cues and uncertainty factors as determined by the objective model in the 40dB SNR test. The vowels have moved closer together perceptually when compared to the vowels presented with no background noise. The vowels with longer duration are still separated from each other to a larger extent compared to those with shorter duration. Most of the vowels with shorter duration now intersect with each other. Where the ellipsis for the vowel in the word "pad" was much larger than those for any other vowel in the vowel space for the test without additional noise, in this test it is now of a similar size to the other vowels. From the vowel space it is expected that the words "pot", "put", and "pad" will be confused with each other, as will the words "puut", "pet", and "piet". It is also expected that more confusions will occur and that the percentage information transmitted per acoustic cue will be less due to the intersection of most of the vowels.

The confusion matrix for the subjective test is shown in Figure 4.7.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 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 | eː | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 |
| pêt | ɛː | 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 |
| | | | | | | | | | | | | Average Correct | **59** |

**Figure 4.7. Confusion matrix obtained by pooling the results from the subjective test for vowels with additional multi-talker babble at 40 dB SNR.**

Figure 4.7 was pooled from the confusion matrices for each individual that participated in the test. There was a total of 100 answers for each stimulus. Pooled over all the listeners, the overall percentage of presented vowels that were correctly recognised in the subjective test was 59%. This is 8.5% less than the average obtained in tests with no additional noise.

The confusion matrix shown below in Figure 4.8 was generated by the objective model. The predicted answers in the confusion matrix have been scaled to show 100 answers for each stimulus.

| | | Response | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pAAt | pIEt | pOEt | pAd | pEt | pOt | pIt | pAt | pUt | pEEt | pêt | pUUt |
| | | ɑː | i | u | a | ɛ | ɔ | ə | æ | œ | eː | ɛː | y |
| Stimulus | pAAt ɑː | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 1 | 2 | 0 |
| | pIEt i | 0 | 34 | 12 | 0 | 15 | 0 | 31 | 1 | 1 | 3 | 0 | 4 |
| | pOEt u | 0 | 13 | 25 | 5 | 14 | 8 | 13 | 4 | 10 | 4 | 0 | 4 |
| | pAd a | 0 | 0 | 9 | 49 | 0 | 16 | 1 | 3 | 21 | 0 | 0 | 1 |
| | pEt ɛ | 0 | 12 | 9 | 0 | 32 | 4 | 23 | 2 | 6 | 5 | 0 | 5 |
| | pOt ɔ | 0 | 1 | 8 | 21 | 4 | 26 | 6 | 3 | 26 | 1 | 0 | 4 |
| | pIt ə | 0 | 10 | 5 | 0 | 7 | 3 | 66 | 0 | 6 | 2 | 0 | 1 |
| | pAt æ | 2 | 1 | 10 | 6 | 5 | 6 | 1 | 54 | 5 | 7 | 0 | 2 |
| | pUt œ | 0 | 1 | 7 | 11 | 4 | 18 | 8 | 2 | 46 | 1 | 0 | 2 |
| | pEEt eː | 1 | 7 | 9 | 0 | 15 | 2 | 11 | 7 | 4 | 38 | 2 | 5 |
| | pêt ɛː | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 92 | 0 |
| | pUUt y | 0 | 14 | 10 | 2 | 20 | 8 | 11 | 2 | 6 | 4 | 0 | 22 |
| | | | | | | | | | | | | | **48** |

**Figure 4.8. Prediction confusion matrix (produced by the Frequency Variation Model) for degraded vowels with added multi-talker babble noise at 40dB SNR.**

The average for correct scores (at 48%) was again much lower than in the subjective test. The score has dropped by only 2.5% from the previous test speech with no additional background noise. This shows that, although the Frequency Variation Model does respond to the noise that is added, it does not decline to the extent that it is supposed to. The results for the individual stimuli is grouped and summarized in Table 4.10 and Table 4.11.

**Table 4.10 Summary of the confusion matrix from the subjective test performed with 40 dB SNR multi-talker babble.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pad | 100.0% | Pit | 74.0% | Pot, Put | Put | 27.0% | Pot, Pit |
| Peet | 99.0% | Pot | 57.0% | Pad | Pet | 18.0% | Piet, Pit |
| Paat | 94.0% | | | | Puut | 3.0% | Piet |
| Pêt | 80.0% | | | | Poet | 1.0% | Pit, Pad |
| Pat | 77.0% | | | | | | |
| Piet | 77.0% | | | | | | |

Once again the vowels with longer duration were recognized at an average of more than 75% of the time in the subjective test. The only vowel that has dropped below 80% recognition from the test with no additional noise was “pat” although it was still recognized correctly 77% of the time. The results from this test are very similar to the previous test done with the speech with no additional noise. Most of the vowels still fall into the same category and where confusion of vowels takes place, the confusion still occurs with the same vowels as in the previous test. The poorly recognized vowels have very low percentages. They are mostly confused with one or two other vowels. The confusions are very specific. “Puut”, for example, is recognized very poorly at 3% with almost all incorrect answers being “piet”. This shows that the acoustic cue information available to the listener in the words “piet” and “puut” are very close together and this causes confusion between the two vowels. In general, when comparing the confusion matrices, it is clear that the subjective model has many distinct confusions, where as in the objective test the confusions are spread out more among the vowels (as seen in Figure 4.8).

**Table 4.11. Summary of the prediction confusion matrix from the Frequency Variation Model for vowels with 40 dB SNR multi-talker babble.**

| Best recognized | | Well recognized | | | Poorly recognized | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pêt | 92.0% | Pit | 66.0% | Piet | Pad | 49.0% | Put, Pot |
| Paat | 90.0% | Pat | 54.0% | Poet | Put | 46.0% | Pot, Pad |
| | | | | | Peet | 38.0% | Pet, Pit |
| | | | | | Piet | 34.0% | Pit, Pet |
| | | | | | Pet | 32.0% | Pit |
| | | | | | Pot | 26.0% | Put, Pad |
| | | | | | Poet | 25.0% | Pet,Piet, Pit |
| | | | | | Puut | 22.0% | Pet, Piet |

For the objective model (Table 4.11), only the vowels with the longest duration, namely "pêt" and "paat", still fall in the best recognized category. The other two vowels that were in the best recognized category for the subjective test, "pad" and "peet" were predicted to be recognized much less at 54% and 38%, respectively. The only word that was predicted to fall correctly in the well recognized category is "pit", but where it was confused with "pot" and "put" in the subjective test, the objective model incorrectly predicted "pit" to be confused with "piet". Most of the vowels are predicted to be poorly recognized by the objective model, although all of the predictions are incorrectly identified. This shows that the results of the Frequency Variation Model become less accurate as additional noise is added to the stimuli.

The FITA analysis for the confusion matrices of the subjective test and the objective model is given in Table 4.12. The objective test is referred to as the frequency method in the table to indicate that the Frequency Variation Model was used.

**Table 4.12. Results of FITA analysis for the pooled answers in the subjective test and the Frequency Variation Model implemented in the objective test with added multi-talker babble at 40 dB SNR.**

| % information transmitted | Freq. Method | Subj. Method |
|---|---|---|
| **Duration** | 25% | 42% |
| **F1** | 15% | 38% |
| **F2** | 23% | 58% |

The FITA analysis shows that the information transmitted has dropped predictably for the subjective test. There was also a drop in the percentages for the objective test. The duration cue and the F2 frequency are no longer the better transmitted cues as was the case in the subjective test with no additional background noise. Rather, the percentage transmitted for the F2 frequency is transmitted the best, with the F1 frequency still having the least information successfully transmitted. For the objective test the F1 frequency also still has the lowest FITA score.

Figure 4.9 shows a graph of the average FITA analysis obtained in the subjective tests along with error lines representing the standard deviation between the listeners. The dashed line represents the FITA results from the objective model.

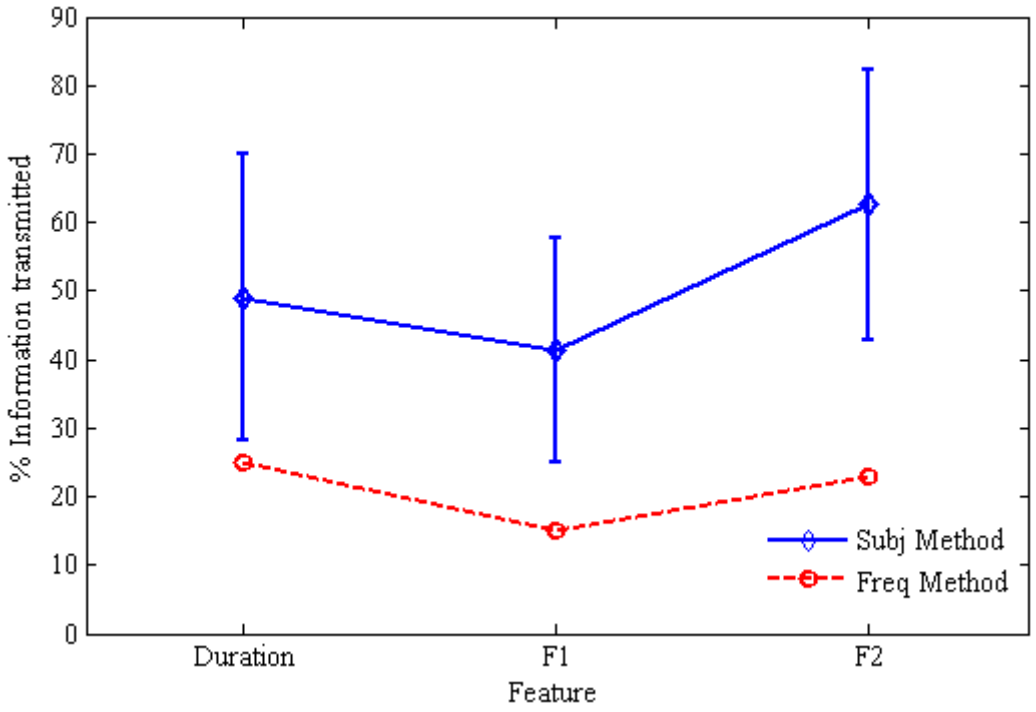


**Figure 4.9. Graph of FITA analysis for the subjective test and the objective test (Frequency Variation Model) performed with multi-talker babble at 40dB SNR. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

For this test the information transmitted by each of the acoustic cues in the objective test were once again below that of the subjective test. Not one of the cues fall inside the error bars. This is consistent with the results obtained in the test without additional noise. The cue that is the closest to the error bar is the duration cue. This shows that, although the average percentage correct scores follow the trend of the subjective tests, the amount of information transmitted in the objective test is by no means accurate.

### 4.2.3 Speech at 20 dB SNR (Multi-Talker Babble Noise)

The results obtained in the 20dB noise test are described next. For this test the listeners were presented with vowels in the presence of multi-talker babble noise at a SNR of 20dB. The accuracy of the answers in the subjective test were expected to decrease once again. If the variation in the frequencies of the formants increases, the objective model is also

expected to predict less correct answers.

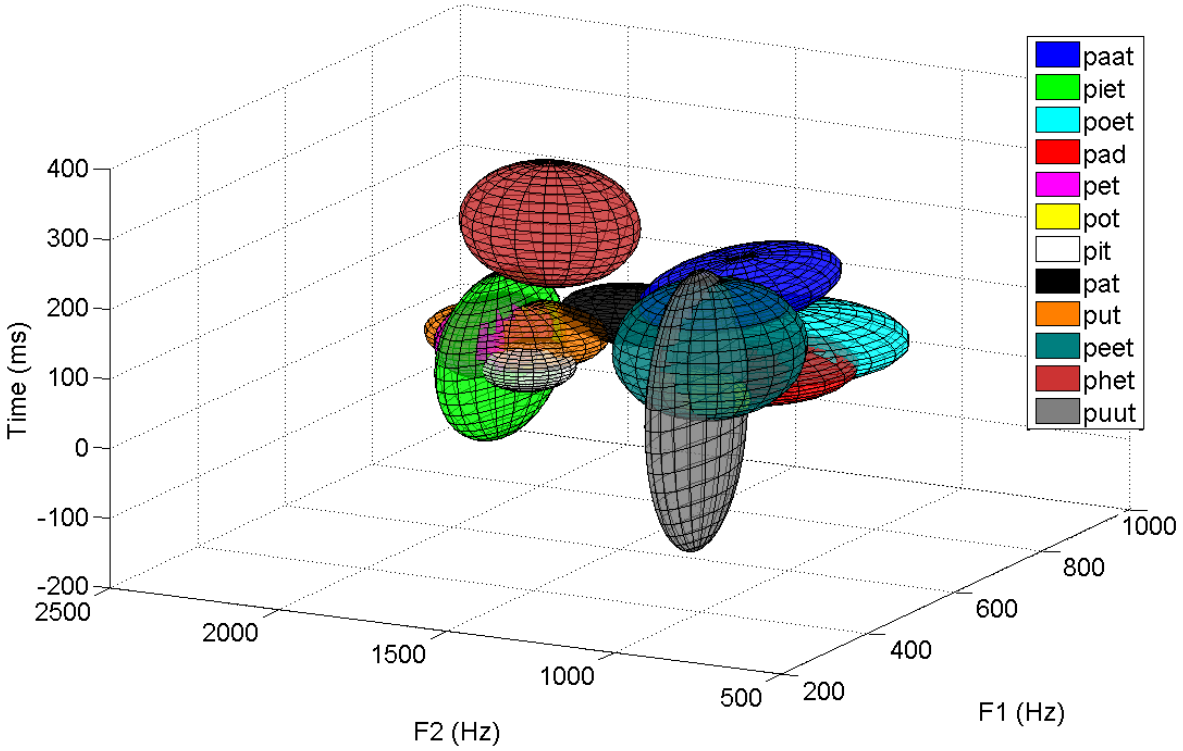


**Figure 4.10. Perceptual vowel space (generated using the Frequency Variation Model) for the processed vowels with added babble background noise at 20 dB SNR.**

The vowel space shown in Figure 4.10 displays the vowels in relation to each other perceptually. The first noticeable difference in this vowel space in comparison to the previous two (Figure 4.2 and Figure 4.6) is that the uncertainty factor of the vowel duration has grown considerably, the largest of these being for the word “puut”. Therefore, according to the objective model, it should be more difficult to distinguish the vowels with longer duration from those with shorter duration. These vowels could easily be distinguished from each other in Figure 4.2 and Figure 4.6. The size of each vowel ellipse has not grown much in terms of its F1 and F2 cues. This shows again that the frequency variation is perhaps not a very good feature to use as an uncertainty factor. The confusion matrices for the subjective test and the objective test are shown in Figure 4.11 and Figure 4.12, respectively.

| | | Response | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pAAt | pIEt | pOEt | pAd | pEt | pOt | pIt | pAt | pUt | pEEt | pêt | pUUt |
| | | ɑː | i | u | a | ɛ | ɔ | ə | æ | œ | eː | ɛː | y |
| pAAt | ɑː | 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 | eː | 0 | 3 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 83 | 0 | 10 |
| pêt | ɛː | 14 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 64 | 18 |
| pUUt | y | 0 | 65 | 0 | 8 | 0 | 2 | 13 | 1 | 11 | 0 | 0 | 0 |
| | | | | | | | | | | | | Average Correct | 45 |

**Figure 4.11. Confusion matrix obtained by pooling the results from the subjective test for vowels with additional multi-talker babble at 20 dB SNR.**

| | | | Response | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | pAAt | pIEt | pOEt | pAd | pEt | pOt | pIt | pAt | pUt | pEEt | pêt | pUUt |
| | | | ɑː | i | u | a | ɛ | ɔ | ə | æ | œ | eː | ɛː | y |
| Stimulus | pAAt | ɑː | 29 | 0 | 7 | 9 | 1 | 5 | 2 | 19 | 4 | 8 | 6 | 9 |
| | pIEt | i | 0 | 21 | 2 | 0 | 21 | 0 | 22 | 7 | 21 | 0 | 5 | 0 |
| | pOEt | u | 8 | 2 | 21 | 21 | 4 | 4 | 5 | 17 | 6 | 5 | 2 | 5 |
| | pAd | a | 5 | 0 | 10 | 27 | 2 | 15 | 5 | 15 | 5 | 4 | 1 | 11 |
| | pEt | ɛ | 0 | 16 | 1 | 0 | 27 | 0 | 27 | 7 | 19 | 0 | 3 | 0 |
| | pOt | ɔ | 1 | 0 | 1 | 1 | 0 | 82 | 2 | 1 | 1 | 7 | 1 | 3 |
| | pIt | ə | 1 | 9 | 1 | 2 | 10 | 1 | 55 | 7 | 12 | 1 | 2 | 0 |
| | pAt | æ | 5 | 4 | 4 | 7 | 9 | 2 | 14 | 41 | 10 | 2 | 3 | 1 |
| | pUt | œ | 2 | 11 | 2 | 4 | 21 | 2 | 21 | 14 | 21 | 1 | 3 | 0 |
| | pEEt | eː | 15 | 0 | 5 | 12 | 1 | 16 | 3 | 7 | 3 | 16 | 4 | 16 |
| | pêt | ɛː | 7 | 9 | 2 | 3 | 11 | 3 | 9 | 12 | 9 | 4 | 31 | 1 |
| | pUUt | y | 6 | 0 | 3 | 8 | 0 | 32 | 0 | 1 | 0 | 17 | 0 | 32 |
| | | | | | | | | | | | | | | 34 |

**Figure 4.12. Prediction confusion matrix (produced by the Frequency Variation Model) for degraded vowels with added multi-talker babble noise at 20dB SNR.**

There is a total of 100 answers for each stimulus for both of the matrices. The overall percentage of presented vowels that were correctly recognised has decreased by only 6% to 45% (from the 59% obtained in the 40dB subjective test). The predicted correct answers

dropped from 58% to 34% in the prediction confusion matrix (as shown in Figure 4.12). This is a drop of 14%, which is more than double that of the subjective test.

Table 4.13 summarizes the confusion matrix for the subjective test for 20dB multi-babble noise and Table 4.14 summarizes the results from the objective test.

**Table 4.13. Summary of the confusion matrix from the subjective test performed with 20 dB SNR multi-talker babble.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Paat | 90.0% | Pêt | 64.0% | Puut | Put | 25.0% | Pit |
| Peet | 83.0% | Pat | 63.0% | Pad | Piet | 11.0% | Pit |
| | | Pad | 59.0% | Pot | Poet | 11.0% | Pad, Pit |
| | | Pit | 59.0% | Put | Pet | 7.0% | Pit |
| | | Pot | 56.0% | Poet | Puut | 0.0% | Piet |

The vowels with the longest durations were still recognized very accurately in the subjective test. “Pêt”, which was correctly recognized more than 80% of the time for the test without additional noise and the 20dB SNR test, has now decreased to 64% recognition. This is interesting since all the vowels with long vowel durations have fallen in the best recognized category up to this point (this showed the superior robustness of the duration cue against noise). “Piet” was recognized correctly 77% of the time in the 40dB SNR, but has now dropped to 11%. Other vowels have relatively the same percentage correct answers as the previous test, that is, “pot”, “put”, and “puut”. This shows that the decrease in percentage is by no means linear. Another observation that can be made is that some of the vowels are now confused with different vowels compared to the 40dB test. “Pot”, for instance, was confused with “pad” most of the time in the 40dB test, but it is now confused most commonly with “poet.” This indicates that the F2 cues of the different vowels have moved closer together in the vowel space, thus causing more confusions.

**Table 4.14. Summary of the prediction confusion matrix from the Frequency Variation Model for vowels with 20 dB SNR multi-talker babble.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pot | 88.0% | Pit | 55.0% | Put, Pet | Pat | 41.0% | Pit, Put |
| | | | | | Puut | 32.0% | Pot |
| | | | | | Pêt | 31.0% | Pat, Pet |
| | | | | | Paat | 29.0% | Pat |
| | | | | | Pat | 27.0% | Pot, Pat |
| | | | | | Pet | 27.0% | Pit |
| | | | | | Piet | 21.0% | Pit, Pet, Put |
| | | | | | Poet | 21.0% | Pad, Pat |
| | | | | | Put | 21.0% | Pet, Pit |
| | | | | | Peet | 16.0% | Pot, Puut |

Table 4.14 shows a summary of the results of the objective model. There is almost no correlation between the percentage correct scores between the objective model and the subjective test. Pit" was the only vowel predicted relatively correctly with 55% (compared to 56% in the subjective test). It was also predicted correctly that it would be confused with "put". The words which were originally recognized very well because of their duration cues, now have a very low recognition percentage. The separation provided by the duration cue is no longer as prominent as it was in previous tests because of the overwhelming effect of the masking factor for the vowel duration cue (as seen in the vowel space in Figure 4.10). The calculation of the uncertainty factor for the duration cue is incorrect; this aspect needs to be reinvestigated. "Peet" is such an example; it was recognized correctly 83% of the time in the subjective test, however, it is now recognized most poorly at 16%. Most of the confusions are also predicted incorrectly, with the exceptions of "pit", "pet", "piet", and "poet".

The lack of correlation between the objective model and the subjective test under exceptionally noisy conditions show that the uncertainty factor used for this implementation (that is, frequency variation) does not provide very good results. This shows that the disguising effect of frequency variation in the model does not properly mimic what happens in subjective tests. It may also mean that the three acoustic cues used

in the objective method need to be supplemented with secondary cues in order to predict accurately any vowel confusions in the presence of this amount of noise.

The FITA analysis for the subjective confusion matrix and the objective confusion matrix in the 20dB SNR test is shown in Table 4.15. The subjective model still has some information transmitted by the cues, with the best acoustic cue being the duration. The FITA analysis for the objective test has dropped to very low percentages in comparison to the previous tests. None of the acoustic cues transmit enough meaningful information for accurate interpretation by the listeners.

**Table 4.15. Results of FITA analysis for the pooled answers in the subjective test and the Frequency Variation Model implemented in the objective test with added multi-talker babble at 20 dB SNR.**

| % information transmitted | Freq. Method | Subj. Method |
|---|---|---|
| **Duration** | 11% | 34% |
| **F1** | 4% | 24% |
| **F2** | 6% | 25% |

The averages and standard deviation of the FITA analysis of the individual confusion matrices in the subjective tests are shown in Figure 4.13. The dashed line represents the FITA results from the objective model.

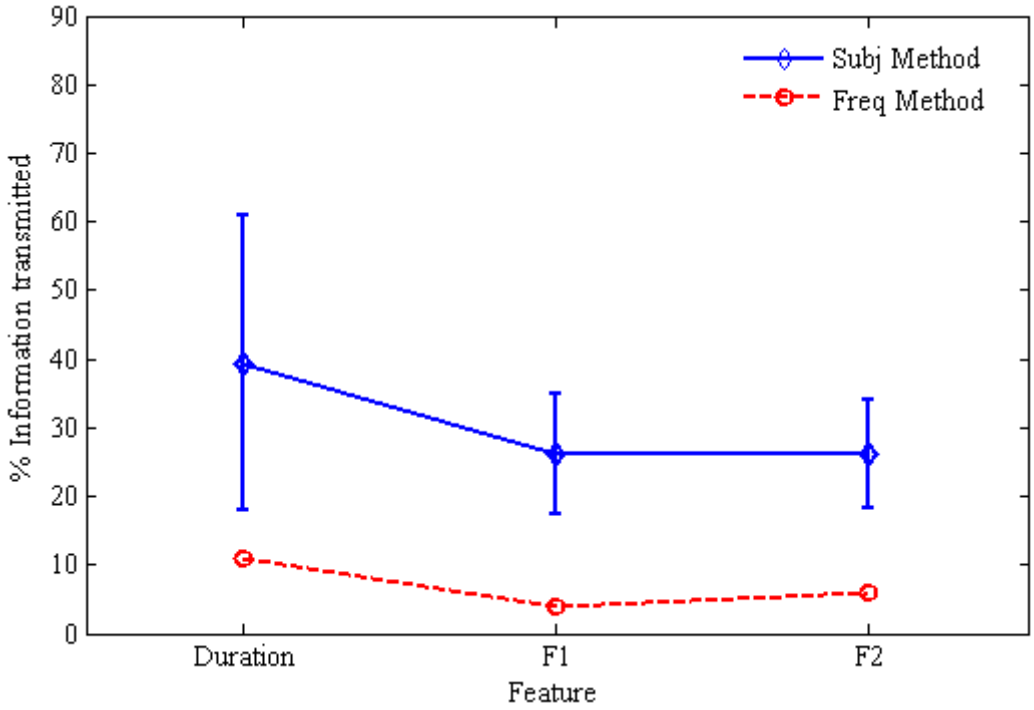


**Figure 4.13. Graph of FITA analysis for the subjective test and the objective test (Frequency Variation Model) performed with multi-talker babble at 20dB SNR. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

The error lines show that the standard deviation is very large for the duration acoustic cue. This shows that the duration cue is used to quite different extents by different listeners. The F1 and F2 cues are very similar in terms of average information transmitted and the standard deviation. The information transmitted for all the acoustic cues in the objective test fall far outside the standard deviation of the subjective model. Once again the objective model under-performed and did not approximate the results of the subjective test.

### 4.2.4 Speech at 0 dB SNR (Multi-Talker Babble Noise)

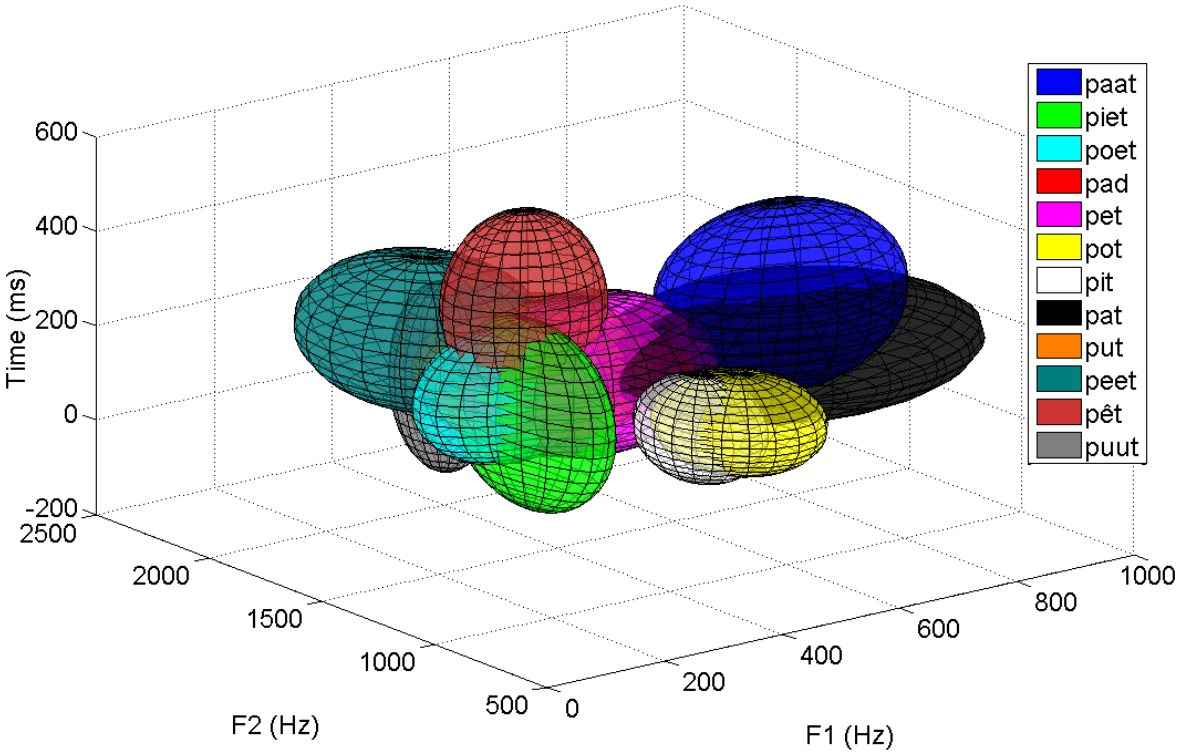


**Figure 4.14. Perceptual vowel space (generated using the Frequency Variation Model) for the processed vowels with added babble background noise at 0 dB SNR.**

Figure 4.14 displays the vowels in relation to each other perceptually for vowels embedded in multi-babble noise at the same level as the speech itself. At this level a person fitted with a CI was expected to have extreme difficulty in distinguishing vowels from each other. Therefore, the vowel space should have all the vowels intersecting each other. In terms of duration this is the case in the figure above, although the formant frequencies do not seem to be large enough to produce the levels of confusion that cochlear implantees are expected to experience in the test.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 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 | eː | 12 | 13 | 5 | 1 | 4 | 4 | 2 | 3 | 5 | 19 | 8 | 24 |
| pêt | ɛː | 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 |
| | | | | | | | | | | | | Average Correct | 13 |

**Figure 4.15. Confusion matrix obtained by pooling the results from the subjective test for vowels with additional multi-talker babble at 0 dB SNR.**

The confusion matrix in Figure 4.15 shows the pooled confusion matrix as recorded in the subjective test. The average correct score is 13% which is so low that the outcome can be attributed to chance alone. The answers are spread widely across the confusion matrix; no presented vowel is recognized correctly or confused incorrectly with any other single vowel. The only vowel that produced some form of recognition was /a:/ in the word "paat", which was identified correctly 35% of the time.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 30 | 2 | 4 | 11 | 9 | 5 | 4 | 21 | 6 | 1 | 5 | 1 |
| pIEt | i | 0 | 20 | 20 | 4 | 4 | 4 | 5 | 0 | 11 | 6 | 8 | 19 |
| pOEt | u | 1 | 8 | 43 | 5 | 8 | 1 | 1 | 1 | 8 | 7 | 9 | 8 |
| pAd | a | 2 | 1 | 9 | 53 | 7 | 5 | 9 | 3 | 5 | 1 | 3 | 1 |
| pEt | ɛ | 3 | 8 | 16 | 15 | 15 | 3 | 3 | 3 | 15 | 4 | 7 | 8 |
| pOt | ɔ | 3 | 5 | 3 | 15 | 5 | 26 | 26 | 6 | 5 | 1 | 3 | 1 |
| pIt | ə | 2 | 6 | 4 | 9 | 7 | 21 | 36 | 4 | 6 | 1 | 4 | 1 |
| pAt | æ | 17 | 2 | 3 | 14 | 9 | 11 | 7 | 27 | 5 | 1 | 4 | 1 |
| pUt | œ | 2 | 10 | 18 | 11 | 9 | 3 | 3 | 2 | 18 | 6 | 7 | 12 |
| pEEt | eː | 1 | 11 | 8 | 3 | 6 | 1 | 2 | 1 | 13 | 23 | 8 | 23 |
| pêt | ɛː | 3 | 10 | 19 | 5 | 9 | 2 | 4 | 2 | 11 | 6 | 19 | 9 |
| pUUt | y | 0 | 6 | 7 | 1 | 7 | 1 | 1 | 0 | 8 | 19 | 7 | 42 |
| | | | | | | | | | | | | | **29** |

**Figure 4.16. Prediction confusion matrix (produced by the Frequency Variation Model) for degraded vowels with added multi-talker babble noise at 0dB SNR.**

The prediction confusion matrix of the objective model in Figure 4.16 also shows low recognition percentages. The average correct score, however, amounts to 29%, which is more than double that of the subjective test. The objective model also does not show any specific confusions with all the confusions spread out among all the vowels. In other words, there is not a specific confusion made for each vowel; the confusions are mostly random. However, it was expected that under extremely noise conditions (SNR of 0 dB) the objective prediction model would no longer be able to predict the confusions correctly due to the variance found in subjective tests under these conditions.

**Table 4.16. Results of FITA analysis for the pooled answers in the subjective test and the Frequency Variation Model implemented in the objective test with added multi-talker babble at 0 dB SNR.**

| % information transmitted | Freq. Method | Subj. Method |
|---|---|---|
| **Duration** | 5% | 1% |
| **F1** | 15% | 0% |
| **F2** | 11% | 1% |

The FITA analysis shown in Table 4.16 shows predictably that there is very little information transmitted by the acoustic cues. The noise drowns out the cues sufficiently so that there is 1% or less information received in the subjective model. The Frequency Variation Model shows that the duration cue is very low at 5%, which is a good approximation. The formants on the other hand provide little information on the identification of the vowels. This outcome correlated well with the vowel space in Figure 4.14, which showed that the vowels did not all intersect in respect to their formant frequencies.

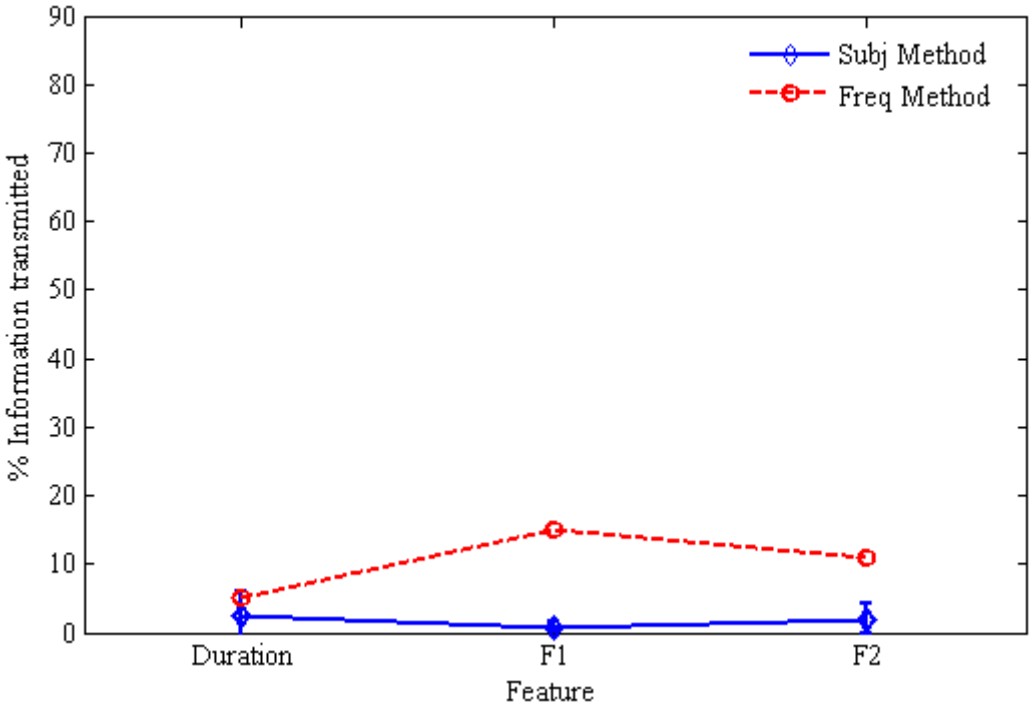


**Figure 4.17. Graph of FITA analysis for the subjective test and the objective test (Frequency Variation Model) performed with multi-talker babble at 0dB SNR. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

The FITA analysis of the individual confusion matrices in the subjective test produced very little deviation as shown by the error bars in Figure 4.17. The duration cue in the objective model falls within the range of the error bars for the subjective test. All the cues are very close to 0% for the subjective test. The F1 and F2 cues in the objective model are too high and, although they are lower than in the better quality speech, they have not dropped as much as in the subjective test.

## 4.3 SPECTRAL CONTRAST MODEL

In this section the Spectral Contrast Model is evaluated. The Spectral Contrast Model was designed to use the spectral contrast of the formants to determine the variation for the probability functions that are used to generate the confusion matrix. The hypothesis for this is: the spectral contrast is the major role player in causing confusion in recognizing vowel sounds. Spectral contrast has been shown in the literature to play an important role in vowel recognition; it is, therefore, expected that this model will perform better than the Frequency Variation Model (Leek *et al.*, 1987; Leek and Summers, 1996a; Loizou and Poroy, 2001a; ter Keurs *et al.*, 1993b). The results obtained in this model are compared once again to the results in the subjective test. The reader should note that the same results from the subjective tests used in the Frequency Variation Model will be used here. The figures and confusion matrices for the subjective tests are repeated here for easier comparison.

### 4.3.1 Speech Without Additional Background Noise

Figure 4.18 gives the vowel space generated by the objective model. It serves as an indication of how the confusion matrix is formed. The centre of each ellipsis lies at the measured acoustic cues for each of the vowels, and is the same as in the Frequency Variation Model. The size of the ellipse in the F1 and F2 dimensions, is calculated differently, however, compared to the Frequency Variation Model. In this instance, the distances are calculated from the spectral contrast for each formant (instead of from the variation in frequency as in the rival model).

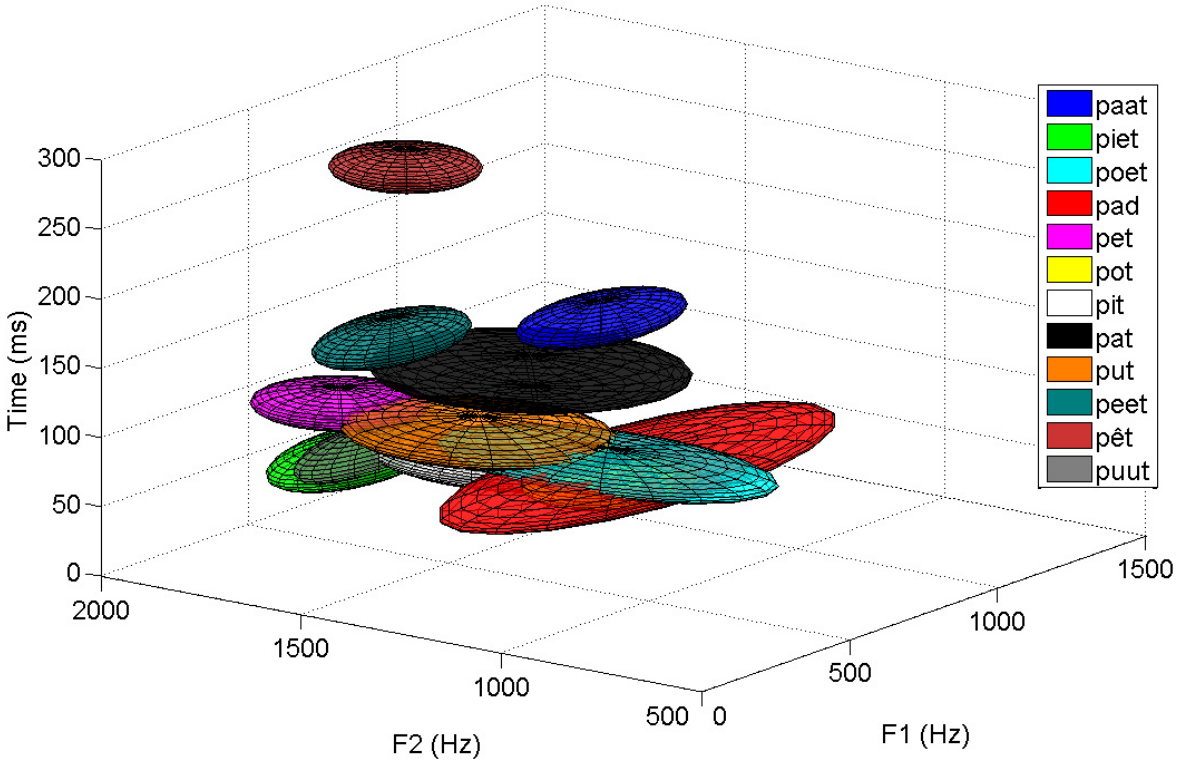


**Figure 4.18. Perceptual vowel space (generated using the Spectral Contrast Model) for the processed vowels, with no background noise.**

The vowel space for the vowels presented without additional background noise (Figure 4.18) shows that the vowels with the longer duration ("paat", "peet", "pat", and "pêt") are well separated from the rest. In the confusion matrix they should be recognized most easily. The vowels of duration between 60 ms and 100 ms all lie close together and, hence, have a greater probability of being confused. In comparison with the Frequency Variation Model the sizes of the ellipses are larger, especially in terms of the F1 dimension. The ellipse for "pad" is the largest once again, although it seems more realistically shaped (in that it is smaller) than in the Frequency Variation Model.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 197 | 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 | eː | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 199 | 0 | 0 |
| pêt | ɛː | 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 |
| | | | | | | | | | | | | Average Correct | 135 |

**Figure 4.19. Confusion matrix obtained from the results of the subjective test for vowels with no background noise.**

The confusion matrix in Figure 4.19 shows the pooled averages from the confusion matrices for all listeners in the subjective test. There is a total of 200 answers for each presented stimulus. The overall percentage for the presented vowels that are correctly recognised is an average of 67.5 %.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 199 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pIEt | i | 0 | 75 | 1 | 0 | 26 | 0 | 15 | 0 | 7 | 2 | 0 | 73 |
| pOEt | u | 0 | 2 | 55 | 21 | 9 | 55 | 27 | 1 | 26 | 0 | 0 | 3 |
| pAd | a | 0 | 0 | 27 | 129 | 1 | 24 | 5 | 4 | 11 | 0 | 0 | 0 |
| pEt | ɛ | 0 | 23 | 6 | 0 | 102 | 1 | 17 | 1 | 15 | 6 | 0 | 29 |
| pOt | ɔ | 0 | 0 | 30 | 20 | 1 | 109 | 22 | 0 | 16 | 0 | 0 | 0 |
| pIt | ə | 0 | 18 | 11 | 2 | 14 | 18 | 96 | 0 | 15 | 1 | 0 | 25 |
| pAt | æ | 1 | 0 | 4 | 9 | 3 | 1 | 0 | 169 | 5 | 7 | 0 | 1 |
| pUt | œ | 0 | 9 | 23 | 6 | 26 | 21 | 42 | 1 | 55 | 4 | 0 | 12 |
| pEEt | eː | 0 | 3 | 0 | 0 | 11 | 0 | 1 | 2 | 5 | 168 | 0 | 9 |
| pêt | ɛː | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 200 | 0 |
| pUUt | y | 0 | 36 | 1 | 0 | 24 | 0 | 27 | 0 | 10 | 5 | 0 | 98 |
| | | | | | | | | | | | | | 121 |

**Figure 4.20. Prediction confusion matrix (generated by the Spectral Contrast Model) for vowels with no background noise.**

The confusion matrix shown in Figure 4.20 was generated by the objective model using the same speech without additional background noise as for the subjective model. The average correct answers predicted by the objective model is 60.5%. This is a much closer approximation to the results of the subjective test than the 51% predicted by the Frequency Variation Model. The two confusion matrices are summarized in the following two tables (Table 4.17 and Table 4.18).

**Table 4.17. Summary of the confusion matrix from subjective testing with no background noise.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Peet | 99.5% | Piet | 70.5% | Puut | Poet | 46.5% | Pot, Pad |
| Paat | 98.5% | Pot | 68.5% | Poet | Pet | 42.5% | Piet, Pit |
| Pat | 90.0% | Pit | 58.0% | Put | Puut | 32.0% | Piet |
| Pad | 86.5% | | | | Put | 30.5% | Pit, Pot |
| Pêt | 85.5% | | | | | | |

**Table 4.18. Summary of the prediction confusion matrix from the objective model for vowels with no background noise (Spectral Contrast Model).**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pêt | 100.0% | Pad | 64.5% | Pot | Puut | 49.0% | Piet |
| Paat | 99.5% | Pot | 54.5% | Poet | Pit | 48.0% | Puut |
| Pat | 84.5% | Pet | 51.0% | Puut | Piet | 37.5% | Puut |
| Peet | 84.0% | | | | Poet | 27.5% | Pot |
| | | | | | Put | 27.5% | Pit |

The vowels that are recognized most accurately in the subjective test are those with a long duration. The objective model predicted this result accurately, with the same vowels lying in the best recognized category except for “pad.” The vowel in the word “pad” has a shorter duration (87 ms once processed through the CI model) than most of the other vowels that lie in the best recognized category. Although “pad” does not lie in the best

recognized category, it is still predicted to have a relatively high percentage of correct answers at 64%. In the well recognized category only "pot" was predicted correctly to lie in this category. The vowel that it is confused with is also predicted accurately. In the poorly recognized category, it was predicted correctly that "puut" would be confused with "piet", "poet" with "pot", "put" with "pit", and "piet" with "puut". This shows a much better prediction to confusions than the Frequency Variation Model can provide; this was expected.

The results of the FITA analyses for the vowels for the above two confusion matrices are given next.

**Table 4.19. Results of FITA analysis for the pooled answers in the subjective test and for the Spectral Contrast Model implemented in the objective test (no background noise).**

| % information transmitted | S. Contrast Method | Subj. Method |
|---|---|---|
| **Duration** | 46% | 60% |
| **F1** | 36% | 43% |
| **F2** | 52% | 57% |

In Table 4.19, the results from the subjective test show that the duration cue presented the most information to the listener in the identification of the vowels. The information of the F2 acoustic cue was transmitted only 3% less than the vowel duration cue. The acoustic cue that carried the least information to the listener was the F1 cue. The FITA analysis of the confusion matrix produced by the objective model showed that each of these cues were lower than they should have been. They are, however, closer in the Spectral Contrast Model than in the Frequency Variation Model (see Table 4.9 for a direct comparison). The F2 acoustic cue is only 5% lower than the subjective test value, and the F1 cue is 7% less than it should be. The duration cue shows the greatest dissimilarity between the two tests – there is a 14% difference.

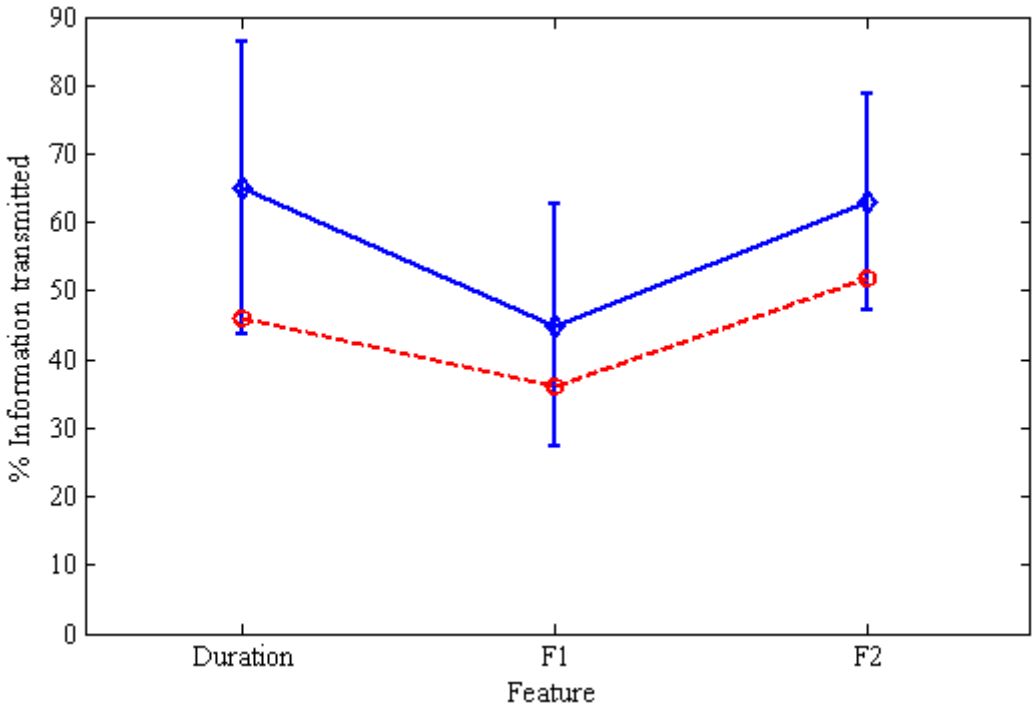


**Figure 4.21 Graph of FITA analysis for the subjective test and the objective test (Spectral Contrast Model) performed with no additional background noise. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

Figure 4.21 shows a comparison between the information transmitted in the major acoustic cues for the subjective and objective models. All the acoustic cues of the objective model fall inside the error bars of the subjective test. The acoustic cue that had the largest difference was duration, but it also lies within the standard deviation of the cues in the subjective test.

### 4.3.2 Speech at 40dB SNR (Multi-Talker Babble Noise)

The vowel space in Figure 4.22 was generated by the objective model. It serves to give the reader an indication of how the confusion matrix is formed.

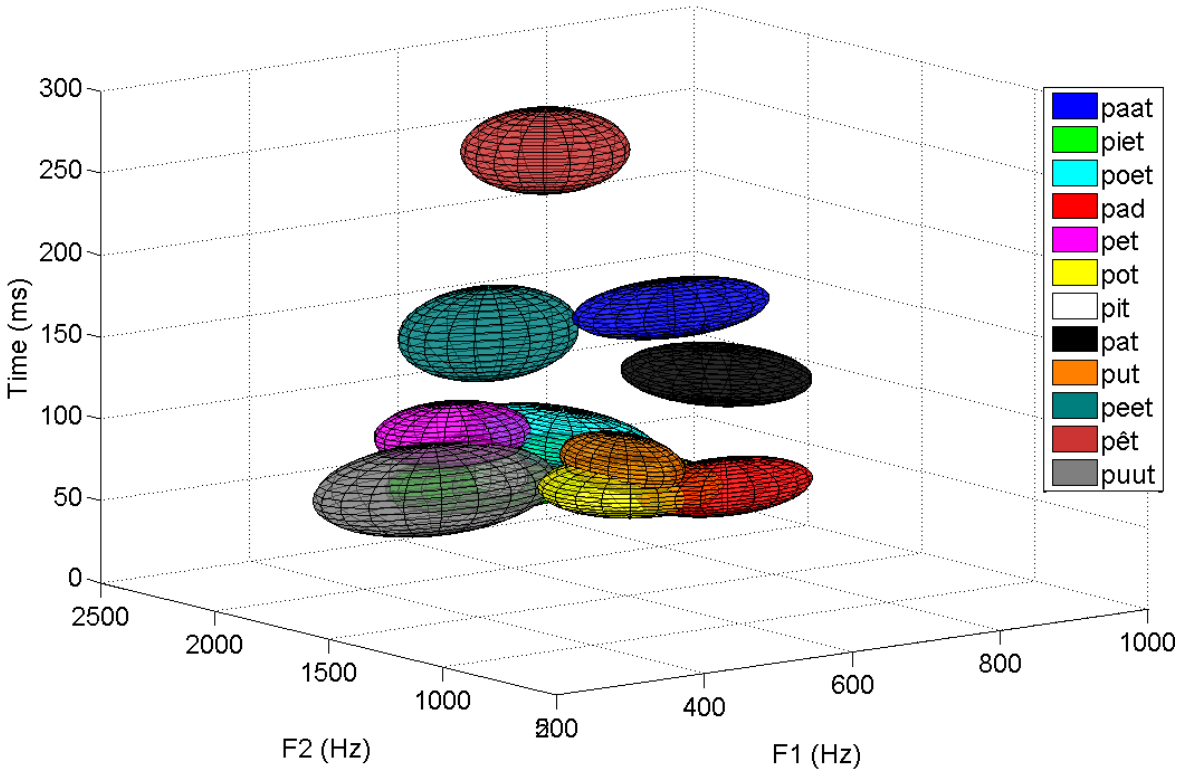


**Figure 4.22. Perceptual vowel space (generated using the Spectral Contrast Model) for the processed vowels with added babble background noise at 40 dB SNR.**

The vowel space above shows that the vowels have moved closer together perceptually in comparison to the vowels with no additional background noise. The vowels with longer duration are still separated from each other and from those vowels with shorter duration. Most of the vowels with shorter duration intersect with each other. The sizes of the ellipses are not much larger than they were for the speech with no additional noise. However, there will be more confusions predicted because of the fact that the formants have moved closer together.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 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 | eː | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 |
| pêt | ɛː | 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 |
| | | | | | | | | | | | | Average Correct | **59** |

**Figure 4.23. Confusion matrix obtained by pooling the results from the subjective test for vowels with additional multi-talker babble at 40 dB SNR.**

The confusion matrix in Figure 4.23 has been summed from all the listeners in the subjective test. The listeners were presented with vowels in the presence of 40dB SNR multi-talker babble. There is a total of 100 answers for each presented stimulus. The overall percentage that is correctly recognised for the vowels presented is 59%. This result is 8.5% less than the average percentage for vowels with no additional background noise.

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 3 | 3 | 0 |
| pIEt | i | 0 | 50 | 11 | 1 | 7 | 1 | 28 | 0 | 1 | 1 | 0 | 1 |
| pOEt | u | 0 | 20 | 41 | 8 | 7 | 5 | 9 | 1 | 4 | 2 | 0 | 2 |
| pAd | a | 0 | 2 | 22 | 60 | 0 | 11 | 1 | 1 | 4 | 0 | 0 | 0 |
| pEt | ɛ | 0 | 13 | 8 | 0 | 42 | 4 | 21 | 1 | 5 | 4 | 0 | 3 |
| pOt | ɔ | 0 | 1 | 9 | 15 | 5 | 43 | 8 | 2 | 14 | 1 | 0 | 2 |
| pIt | ə | 0 | 13 | 7 | 0 | 8 | 3 | 65 | 0 | 3 | 1 | 0 | 1 |
| pAt | æ | 2 | 1 | 10 | 2 | 5 | 6 | 0 | 56 | 5 | 10 | 0 | 3 |
| pUt | œ | 0 | 3 | 9 | 4 | 11 | 28 | 13 | 2 | 28 | 2 | 0 | 2 |
| pEEt | eː | 1 | 7 | 9 | 1 | 14 | 1 | 8 | 6 | 4 | 44 | 1 | 3 |
| pêt | ɛː | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 97 | 0 |
| pUUt | y | 0 | 13 | 13 | 0 | 23 | 6 | 10 | 3 | 4 | 4 | 0 | 25 |
| | | | | | | | | | | | | | **53** |

**Figure 4.24. Prediction confusion matrix (produced by the Spectral Contrast Model) for degraded vowels with added multi-talker babble noise at 40dB SNR.**

Figure 4.24 shows the results when the same speech is processed through the objective algorithm. The average scores correct (at 53%) is very comparable to that of the subjective test. This answer has decreased by 7.5% from the test with speech with no noise added. Although this is less than it should be, the trend is still closely followed. This is again a better result than that obtained by using the Frequency Variation Model.

The results for the individual stimuli are classified in the following two tables. Table 4.20 summarizes the subjective test for 40dB multi-babble noise and Table 4.21 summarizes the objective test.

**Table 4.20. Summary of the confusion matrix from the subjective test performed with 40 dB SNR multi-talker babble.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pad | 100.0% | Pit | 74.0% | Pot, Put | Put | 27.0% | Pot, Pit |
| Peet | 99.0% | Pot | 57.0% | Pad | Pet | 18.0% | Piet, Pit |
| Paat | 94.0% | | | | Puut | 3.0% | Piet |
| Pêt | 80.0% | | | | Poet | 1.0% | Pit |
| Piet | 77.0% | | | | | | |
| Pat | 77.0% | | | | | | |

**Table 4.21. Summary of the prediction confusion matrix from the Spectral Contrast Model for vowels with 40 dB SNR multi-talker babble.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pêt | 97.0% | Pit | 65.0% | Piet | Peet | 44.0% | Pet |
| Paat | 90.0% | Pad | 60.0% | Poet | Pot | 43.0% | Pad, Put |
| | | Pat | 56.0% | Poet, Peet | Pet | 42.0% | Pit |
| | | Piet | 50.0% | Pit | Poet | 41.0% | Piet |
| | | | | | Put | 28.0% | Pot |
| | | | | | Puut | 25.0% | Pet |

Once again, all the vowels recognized correctly more than 80% of the time are those with the longest duration. The two longest vowels were correctly predicted as being well recognized. The predictions of the objective model are not as accurate as for the speech with no extra noise. More than half of the vowels are predicted to be recognized correctly between 40 and 60% and the vowel recognized the worse is “puut” at 25%. The subjective test has a large number of vowels recognized above 75% and below 30% of the time. It would seem then that the predictions become less accurate when noise is added. This could mean that listeners use additional secondary cues to interpret the vowel sounds. Alternatively, this could indicate that spectral contrast as an uncertainty factor needs to be

supplemented by other factors.

**Table 4.22. Results of FITA analysis for the pooled answers in the subjective test and the Spectral Contrast Model implemented in the objective test with added multi-talker babble at 40 dB SNR.**

| % information transmitted | S. Contrast Method | Subj. Method |
|---|---|---|
| **Duration** | 25% | 42% |
| **F1** | 18% | 38% |
| **F2** | 22% | 58% |

The FITA analysis shows that the information transmitted by the acoustic cues has dropped predictably for the subjective test. The acoustic cue that shows the largest decrease is the duration cue, whereas the F2 cue has had no drop in percentage at all. All the acoustic cues for the objective model have decreased; a much larger decline in the percentage of information transmitted is seen.

Figure 4.25 shows the averages of the individual confusion matrices obtained in the subjective tests. It also shows error lines representing the standard deviation between the matrices. The dashed line represents the FITA results from the objective model.

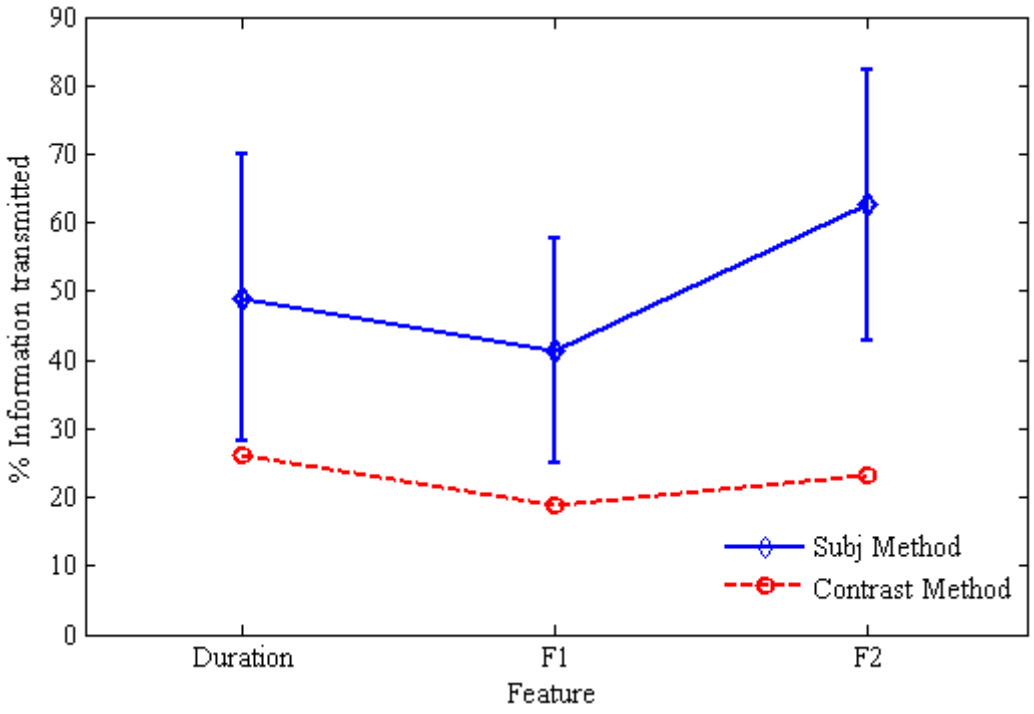


**Figure 4.25. Graph of FITA analysis for the subjective test and the objective test (Spectral Contrast Model) performed with multi-talker babble at 40dB SNR. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

In the test without added noise all three acoustic cues lie inside the error bars for the information transmitted for the subjective test. For this objective test the information transmitted by each of the acoustic cues is lower than it should be. Not one of the cues falls inside the error bars for the subjective test. The cue that is closest to the corresponding error bar is the vowel duration cue. Comparing this graph to Figure 4.9 (which showed the results for the Frequency Variation Model), the results look almost identical. F1 is only a few points lower on the graph than for the Frequency Variation Model.

### 4.3.3 Speech at 20 dB SNR (Multi-Talker Babble Noise)

The results obtained in the 20dB noise test are described below. In this test the listeners were presented with vowels in the presence of multi-talker babble noise at a SNR of 20dB. The decrease in spectral contrast should cause an increase in the number of incorrectly identified vowels in the prediction confusion matrix.

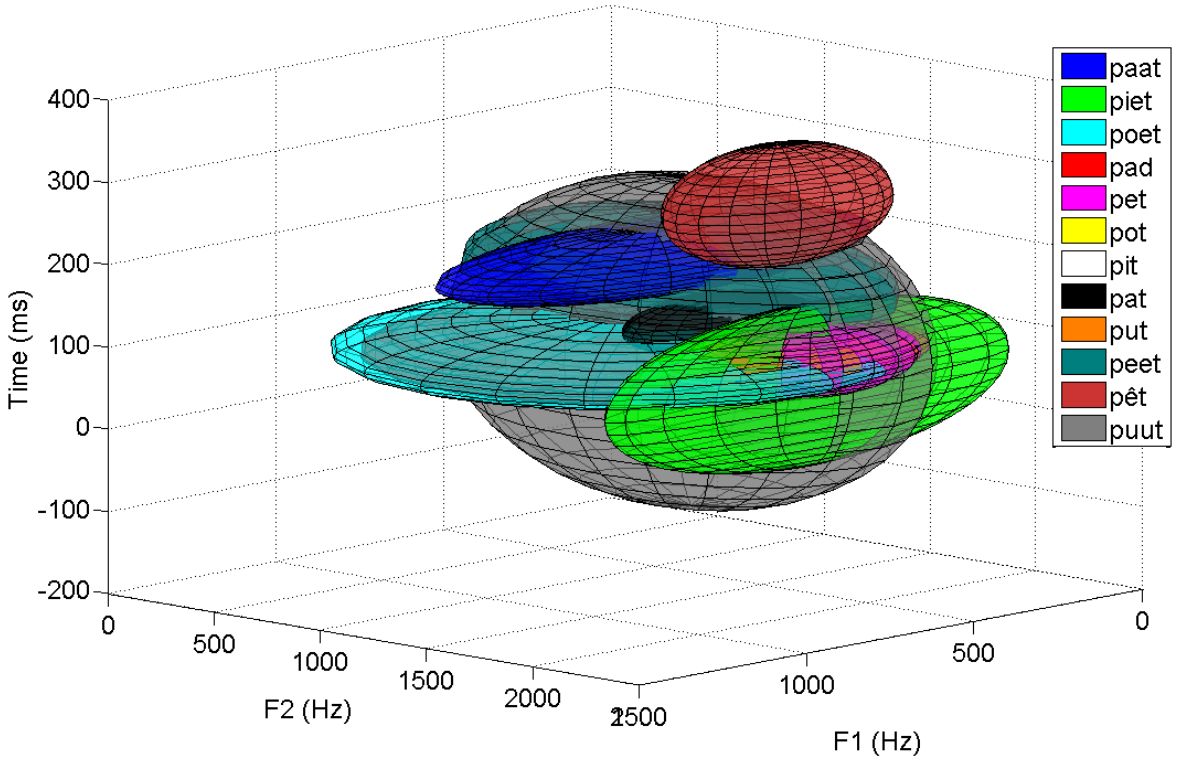


**Figure 4.26. Perceptual vowel space (generated using the Spectral Contrast Model) for the processed vowels with added babble background noise at 20 dB SNR.**

The vowel space in Figure 4.26 shows that the spectral contrast has decreased causing the enlargement in the ellipses that represent the vowels. The increased overlap between the vowels caused the algorithm to predict more confusions between the vowels in comparison with the 40 dB SNR test. From the vowel space it is clear that the only vowel that should be predicted to be recognized relatively well is “pêt”, because it does not intersect with the other vowels.

The confusion matrix from the subjective test is shown in Figure 4.27. Figure 4.28 shows

the prediction confusion matrix generated by the objective algorithm. There is a total of 100 answers for each presented stimulus.

| | | Response | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pAAt | pIEt | pOEt | pAd | pEt | pOt | pIt | pAt | pUt | pEEt | pêt | pUUt |
| | | ɑː | i | u | a | ɛ | ɔ | ə | æ | œ | eː | ɛː | y |
| pAAt | ɑː | 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 | eː | 0 | 3 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 83 | 0 | 10 |
| pêt | ɛː | 14 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 64 | 18 |
| pUUt | y | 0 | 65 | 0 | 8 | 0 | 2 | 13 | 1 | 11 | 0 | 0 | 0 |
| | | | | | | | | | | | | Average Correct | **45** |

**Figure 4.27. Confusion matrix obtained by pooling the results from the subjective test for vowels with additional multi-talker babble at 20 dB SNR.**

| | | Response | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pAAt | pIEt | pOEt | pAd | pEt | pOt | pIt | pAt | pUt | pEEt | pêt | pUUt |
| | | ɑː | i | u | a | ɛ | ɔ | ə | æ | œ | eː | ɛː | y |
| Stimulus | | | | | | | | | | | | | |
| pAAt | ɑː | 23 | 1 | 2 | 3 | 7 | 5 | 3 | 21 | 8 | 13 | 12 | 2 |
| pIEt | i | 0 | 22 | 0 | 0 | 22 | 0 | 20 | 6 | 22 | 0 | 6 | 0 |
| pOEt | u | 6 | 7 | 12 | 12 | 9 | 10 | 7 | 12 | 11 | 5 | 2 | 7 |
| pAd | a | 3 | 5 | 9 | 12 | 10 | 12 | 10 | 12 | 11 | 6 | 2 | 8 |
| pEt | ɛ | 0 | 14 | 0 | 0 | 27 | 1 | 25 | 9 | 19 | 1 | 3 | 0 |
| pOt | ɔ | 0 | 0 | 0 | 0 | 1 | 87 | 2 | 1 | 1 | 4 | 0 | 4 |
| pIt | ə | 0 | 2 | 0 | 0 | 7 | 2 | 77 | 3 | 7 | 1 | 1 | 0 |
| pAt | æ | 1 | 3 | 0 | 0 | 9 | 1 | 6 | 66 | 10 | 1 | 3 | 0 |
| pUt | œ | 0 | 11 | 0 | 0 | 25 | 1 | 19 | 13 | 25 | 1 | 3 | 0 |
| pEEt | eː | 4 | 2 | 1 | 3 | 7 | 23 | 10 | 4 | 7 | 23 | 8 | 7 |
| pêt | ɛː | 2 | 7 | 0 | 0 | 11 | 2 | 8 | 10 | 9 | 3 | 47 | 1 |
| pUUt | y | 3 | 2 | 1 | 8 | 5 | 19 | 8 | 3 | 7 | 18 | 7 | 18 |
| | | | | | | | | | | | | | **37** |

**Figure 4.28. Prediction confusion matrix (produced by the Spectral Contrast Model) for degraded vowels with added multi-talker babble noise at 20dB SNR.**

In the subjective test the overall percentage of presented vowels that are correctly recognised is 45%, pooled over all listeners. This is 14% less than the corresponding average at a 40dB SNR level. The prediction confusion matrix generated by the objective model is shown in Figure 4.28. The average score for correct responses (at 37%) is comparable to that of the subjective test; it is a mere 9% less than the value of the subjective model. The score for the Spectral Contrast Model has decreased by 16% from the prediction made for the speech in the 40dB test. Although the objective model has a lower percentage correct answers than the subjective model, the decrease in both the subjective test and the objective model is exactly the same. This shows that spectral contrast is more promising as an uncertainty factor than frequency variation, the latter decreasing almost double the amount that it should have.

**Table 4.23. Summary of the confusion matrix from the subjective test performed with 20 dB SNR multi-talker babble.**

| **Best recognized (>75%)** | | **Well recognized (50-75%)** | | | **Poorly recognized (<50%)** | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Paat | 90.0% | Pêt | 64.0% | Puut | Put | 25.0% | Pit |
| Peet | 83.0% | Pat | 63.0% | Pad | Piet | 11.0% | Pit |
| | | Pad | 59.0% | Pot | Poet | 11.0% | Pad, Pit |
| | | Pit | 59.0% | Put | Pet | 7.0% | Pit |
| | | Pot | 56.0% | Poet | Puut | 0.0% | Piet |

**Table 4.24. Summary of the prediction confusion matrix from the Spectral Contrast Model for vowels with 20 dB SNR multi-talker babble.**

| Best recognized (>75%) | | Well recognized (50-75%) | | | Poorly recognized (<50%) | | |
|---|---|---|---|---|---|---|---|
| **Stimulus** | **Percentage correct** | **Stimulus** | **Percentage correct** | **Words confused with** | **Stimulus** | **Percentage correct** | **Words confused with** |
| Pot | 87.0% | Pat | 66.0% | Put, Pet | Pêt | 47.0% | Pet, Pat, Put |
| Pit | 77.0% | | | | Pet | 27.0% | Pit, Put |
| | | | | | Put | 25.0% | Pet |
| | | | | | Paat | 23.0% | Pat |
| | | | | | Peet | 23.0% | Pot |
| | | | | | Piet | 22.0% | Pet, Put, Pit |
| | | | | | Puut | 18.0% | Pot, Peet |
| | | | | | Poet | 12.0% | Pad, Pat, Put |
| | | | | | Pad | 12.0% | Pot, Pat, Put |

Table 4.23 and Table 4.24 show a categorized view of the confusion matrices for the subjective test and the objective model, respectively. The vowels with the longest durations are still recognized best in the subjective test. The objective model produced very interesting results. The best recognized vowels are no longer those with longer durations. They are two vowels that have larger spectral contrasts than the others, namely "pot" and "pit". Most of the vowels are predicted to be poorly recognized. This is a wrong prediction as can be seen in the results from the subjective test. Most of the confusions are also incorrectly predicted.

The FITA analysis of the subjective confusion matrix and the objective confusion matrix in the 20dB SNR test is shown in Table 4.25.

**Table 4.25. Results of FITA analysis for the pooled answers in the subjective test and the Spectral Contrast Model implemented in the objective test with added multi-talker babble at 20 dB SNR.**

| % information transmitted | S. Contrast Method | Subj. Method |
|---|---|---|
| **Duration** | 18% | 34% |
| **F1** | 7% | 24% |
| **F2** | 7% | 25% |

In the subjective test some information is still transmitted by the cues, with the best acoustic cue being the vowel duration. The FITA analysis for the objective test has dropped to very low percentages. None of the acoustic cues transmits enough meaningful information for accurate interpretation by the listeners.

The averages and standard deviation of the FITA analyses of the individual confusion matrices in the subjective test are shown in Figure 4.29. The dashed line represents the FITA results from the objective model.

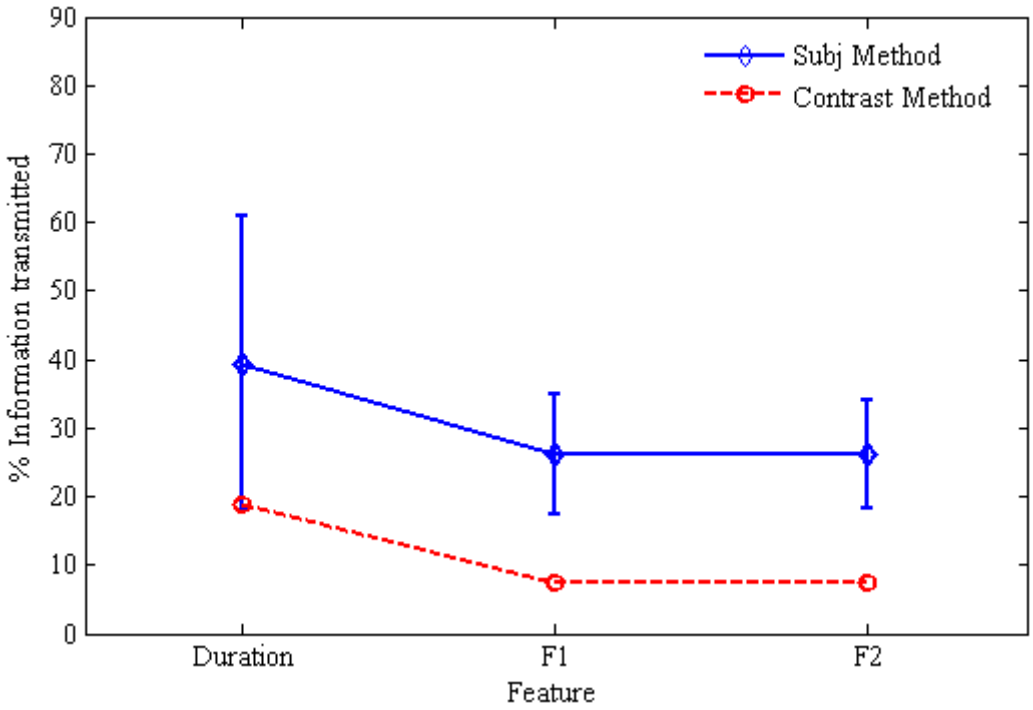


**Figure 4.29. Graph of FITA analysis for the subjective test and the objective test (Spectral Contrast Model) performed with multi-talker babble at 20dB SNR. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

The error lines show that the standard deviation is very large for the duration acoustic cue. The F1 and F2 cues are very similar in average information transmitted and the standard deviations. The information transmitted for the duration cue lies on the boundary of the error bar, and the other two cues fall short of the level they should be. Therefore, the

objective model did not approximate the results of the subjective test well for this test, although the relationship between the cues is still maintained.

### 4.3.4 Speech at 0 dB SNR (Multi-Talker Babble Noise)

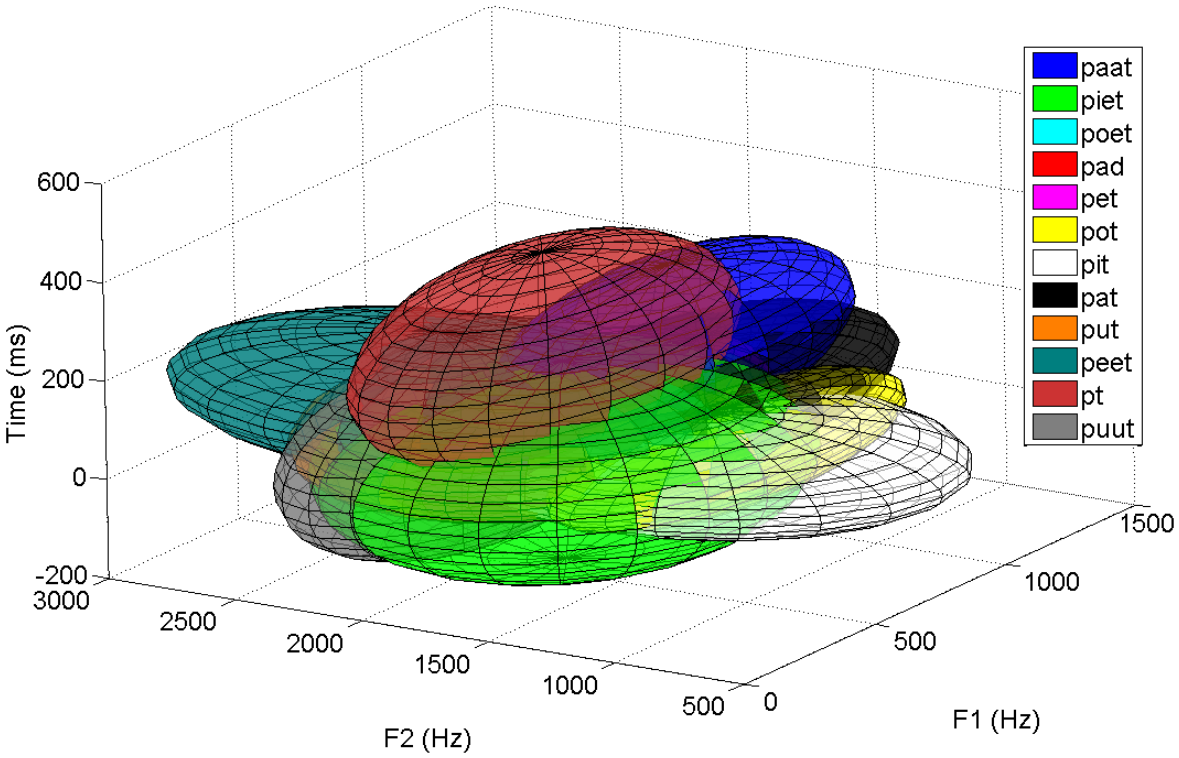


**Figure 4.30. Perceptual vowel space (generated using the Spectral Contrast Model) for the processed vowels with added babble background noise at 0 dB SNR.**

Figure 4.30 displays the vowels in relation to each other perceptually for vowels embedded in multi-babble noise at a 0 dB SNR. At this level a person fitted with a CI was expected to have extreme difficulty in distinguishing vowels from each other. It is, therefore, correct that all the vowels in the vowel space above intersect each other and will cause confusions with each other. The confusion matrices from the subjective test and the objective model are given next in Figure 4.31 and Figure 4.32**,** respectively**.**

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 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 | eː | 12 | 13 | 5 | 1 | 4 | 4 | 2 | 3 | 5 | 19 | 8 | 24 |
| pêt | ɛː | 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 |
| | | | | | | | | | | | | Average Correct | 13 |

**Figure 4.31. Confusion matrix obtained by pooling the results from the subjective test for vowels with additional multi-talker babble at 0 dB SNR.**

| Stimulus | | Response: pAAt ɑː | pIEt i | pOEt u | pAd a | pEt ɛ | pOt ɔ | pIt ə | pAt æ | pUt œ | pEEt eː | pêt ɛː | pUUt y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pAAt | ɑː | 37 | 1 | 0 | 8 | 7 | 10 | 5 | 18 | 4 | 1 | 6 | 2 |
| pIEt | i | 7 | 8 | 8 | 8 | 8 | 8 | 9 | 8 | 8 | 8 | 8 | 9 |
| pOEt | u | 0 | 29 | 38 | 0 | 7 | 1 | 2 | 0 | 0 | 8 | 6 | 9 |
| pAd | a | 1 | 1 | 0 | 80 | 3 | 9 | 2 | 1 | 1 | 0 | 1 | 0 |
| pEt | ɛ | 4 | 4 | 3 | 19 | 20 | 12 | 15 | 3 | 12 | 4 | 2 | 2 |
| pOt | ɔ | 3 | 1 | 3 | 27 | 6 | 32 | 21 | 3 | 1 | 1 | 1 | 1 |
| pIt | ə | 3 | 2 | 5 | 14 | 16 | 21 | 21 | 4 | 6 | 4 | 2 | 1 |
| pAt | æ | 17 | 3 | 4 | 11 | 11 | 14 | 9 | 17 | 3 | 3 | 4 | 2 |
| pUt | œ | 7 | 6 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 6 |
| pEEt | eː | 0 | 9 | 20 | 1 | 11 | 3 | 7 | 3 | 3 | 20 | 7 | 12 |
| pêt | ɛː | 6 | 5 | 13 | 7 | 13 | 7 | 8 | 7 | 5 | 13 | 13 | 4 |
| pUUt | y | 6 | 2 | 11 | 6 | 11 | 3 | 11 | 7 | 11 | 11 | 10 | 11 |
| | | | | | | | | | | | | | 26 |

**Figure 4.32. Prediction confusion matrix (produced by the Spectral Contrast Model) for degraded vowels with added multi-talker babble noise at 0dB SNR.**

The first confusion matrix (Figure 4.31) shows the pooled confusion matrix as recorded in the subjective test. The answers are spread out widely across the confusion matrix; no presented vowel is recognized correctly or confused incorrectly with any other single vowel.

The prediction confusion matrix of the objective model **(**Figure 4.32**)** also shows low recognition percentages. The average correct score, however, amounts to 26%, which is exactly double that of the subjective test. This is a slightly better result than that of the Frequency Variation Model (which predicted that 29% of the answers would be correct). The objective model also does not show any specific confusions with all the vowels spread out among all the other vowels. However, the percentage for correct answers is a little high for a 0dB SNR test. This shows that the model cannot be appropriately used for conditions where speech recognition has decreased to an almost random level.

**Table 4.26. Results of FITA analysis for the pooled answers in the subjective test and the Spectral Contrast Model implemented in the objective test with added multi-talker babble at 0 dB SNR.**

| % information transmitted | S. Contrast Method | Subj. Method |
|---|---|---|
| **Duration** | 3% | 1% |
| **F1** | 10% | 0% |
| **F2** | 3% | 1% |

The FITA analysis in Table 4.26 shows that there is almost no information transmitted by any of the acoustic cues in the subjective test. The objective model predicted this well, except for the F1 acoustic cue. The F1 cue is the only cue that transmitted any information regarding the identity of the vowels (although still at a very low level at 10%).

The FITA analysis of the confusion matrices of the individuals that participated in the testing can be seen in Figure 4.33.

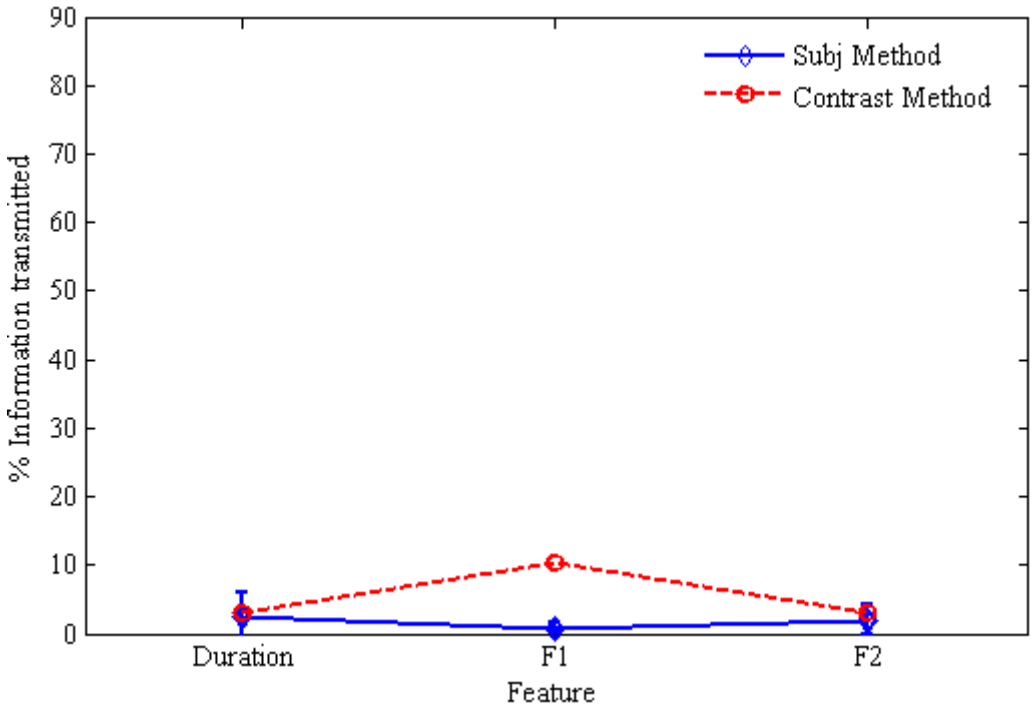


**Figure 4.33. Graph of FITA analysis for the subjective test and the objective test (Spectral Contrast Model) performed with multi-talker babble at 0dB SNR. The average and standard deviation for the percentage of information transmitted is shown for the subjective test.**

The averages for all the acoustic cues in the subjective test are negligible, showing that almost no information is transmitted for each of the three acoustic cues. The standard deviation of all the acoustic cues are very small indicating that none of the individuals performed well in the subjective test. The vowel duration and F2 cues for the objective model compared very well with the subjective model. The F1 acoustic cue in the objective model is the only cue that transmits some information, although the level is still minimal at 10%. Once again, the Spectral Contrast Model outperforms the Frequency Variation Model, although it seems that the model should not be used under such extreme conditions.

## 4.4 SUMMARY

In this chapter, the results of the experimental studies were given. The results from subjective vowel confusion tests were compared to the predictions of the two objective models. The tests were done firstly with speech with no added background noise and then with three levels of SNR using multi-talker babble as additional noise. From the results it can be seen that the Spectral Contrast Model is a good first approximation to the subjective

tests' results. The results presented in this chapter are discussed in the following chapter and compared to those found in the literature.

# CHAPTER 5 DISCUSSION

## 5.1 CHAPTER OBJECTIVES

In this chapter the results from the two implemented methods are discussed and compared with each other. A general discussion then follows to provide interpretation of the results as a whole. The methodology of the model implemented is brought into context with the findings in other studies, although a direct comparison is not possible. The chapter concludes with answers and observations related to the research questions that were posed at the beginning of this study.

## 5.2 EVALUATION OF THE OBJECTIVE MODELS

A comparison of the two objective models with subjective testing data showed that only a first approximation of confusion predictions was possible. The Frequency Variation Model performed poorly in predicting the confusions for the vowels under all noise conditions. The Spectral Contrast Model performed better with some of the most common confusions predicted correctly. The Spectral Contrast Model also followed the trend of average correct answers as the noise level was varied.

The vowel spaces for the Frequency Variation Model showed inconsistencies with regard to the subjective test results. The best example of this is the vowel /ɑ/ (pad). The model showed a high variance for recognition of the vowel (in the no noise test), although it was recognized very well (with 87% accuracy) in the subjective test. This discrepancy may arise because the F1 frequency of /ɑ/ (pad) is very high and lower spectral peaks are introduced which the model then picks incorrectly as the first formant. A very limited number of vowel confusions, for instance between /y/ (puut) and /i/ (piet), were predicted correctly. This could be attributed to the proximity of these vowels in the vowel space. The correlation between the variances in the vowel space and the confusion of vowels generally did not match.

Once background noise was added to the vowel sounds, the confusions increased in the subjective test and in the predictions by the objective model. The average correct answers dropped by 8.5% in the subjective test but only by 2.5% in the objective test. This shows that overall the increase in frequency variation does not correlate with the deterioration of results in the subjective test. The confusion prediction also declined with eight of the twelve vowels predicted to be confused more than 50% of the time compared to the four in the subjective test. Around half of the predictions were correct because of the vowels' proximity in the vowel space.

The 20dB SNR test showed the superior robustness of the duration cue. Only vowels with the longest duration (that is, /ɑː/ (paat) and /e/ (peet)) were recognized above 75% of the time. In the 20dB SNR test, the Frequency Variation Model failed to predict any trend in vowel confusions. For instance /e/ (peet) which was well recognized (at 83%) in the subjective test, was shown to be poorly recognized by the objective model (at 16%). The FITA analysis showed that the F1 and F2 cues only transmitted 4% and 6% of information, respectively. Therefore the model predicted that the formant frequency information is of no use for this test. These results do not correlate with the result of the subjective test. The results show that frequency variation increases at a rate which is not in line with the subjective test.

Tests were also conducted with a 0dB SNR. As discussed in the previous chapter, these results should be ignored since the only random confusions and no information transmitted were shown in the subjective test. The Frequency Variation Model did not predict the extreme deterioration of recognition because of stabilization of the frequency variation at this SNR level. This may occur because new peaks are introduced into the spectrum by the multi-talker background noise. The model's ability to determine vowel confusions does not work at all at 0dB SNR. This is expected since the little vowel recognition that seemingly takes place in the subjective test can be put down to chance. The results obtained for this condition do not provide any useful information concerning the testing and outcome and will be ignored for both models.

The spectral contrast implementation is based on the fact that formants are less discernible when spectral contrast is reduced. The results showed that confusion correlation with the subjective results increased when the spectral contrast method was used. The FITA analysis also showed that the percentage information transmitted fell within the error bars of the subjective test for the test with no additional noise and the 40dB SNR test. This suggests that spectral contrast plays a larger role in the identification of vowels than frequency variation. It also suggests that the use of formants by the human auditory system as identifiers of vowels is related to the level of spectral contrast.

Only specific confusions were correctly predicted by the spectral contrast method (for instance, the confusion between /ə/ (pit) and /œ/ (put), the confusion between /æ/ (pat) and /ɔ/ (pot), and the confusion between /i/ (piet) and /y/ (puut)). This was expected since these vowels lie close to each other in the vowel space, and the increased variance causes them to be confused when the SNR decreases. The confusions which are not so specific in the subjective tests were only predicted to a limited extend (for example, /ɛ/ (pet) being confused with /ə/ (pit), and /ɔ/ (pot) being confused with /a/ (pad)).

In the 40dB SNR test, the average correct score was predicted well (53% predicted for the 59% actually shown in the subjective test.) The subjective test showed that six of the twelve vowels were recognized well (above 75%), but the prediction of the objective model showed that only two vowels met this criterion. Only three of the twelve vowel confusions were predicted correctly (namely, /ɔ/ (pot) and /a/ (pad), /ɛ/ (pet) and /ə/ (pit) and /œ/ (put) and /ə/ (pit)). This shows the potential of the Spectral Contrast Model since these are the most common confusions in the subjective test.

Once the noise level was increased to 20dB, the Spectral Contrast Model failed to predict confusions correctly. The average correct score was predicted well (37% predicted for the 45% actually shown in the subjective test). It seems that when background noise is added the other acoustic cues are used to aid the listener in identifying noises. This suggests that the smaller confusions can be attributed to other secondary acoustic cues which were not included or tested in the model. These cues could include the spectral envelope (Zahorian

and Jagharghi, 1993), formant glides (Hillenbrand and Nearey, 1999), or even vowel-to-consonant transitions (Jenkins *et al.*, 1983; Strange, 1989).

In both tests the vowels separated by their duration were found to be the best recognized. The vowels with the longest durations (that is, /ɛː/ (pêt), /ɑː/ (paat), and /e/ (peet)) were recognized very well up to a 20dB SNR. This is common in the subjective as well as the objective tests – the only test where this is not the case is the 0dB SNR test.

## 5.3 DIRECT COMPARISON BETWEEN THE TWO MODELS

Figure 5.34 shows the average correct scores of the subjective test and the two objective models. The average correct score is the average of the correct responses for each of the presented vowels. For the subjective test the figure includes error bars that represent the standard deviation of the average correct scores between the listeners.

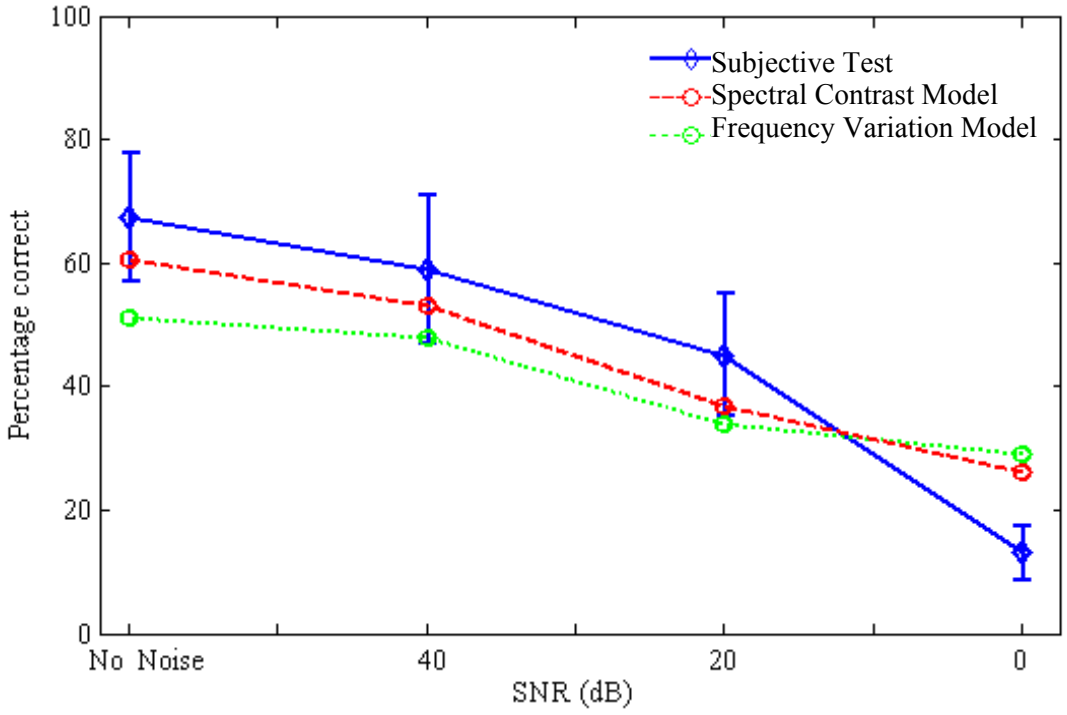


**Figure 5.34. The average correct recognition scores for the subjective method and the two objective models for each SNR level of additional multi-talker babble.**

The graph shows that both of the objective models predicted averages lower than those of

the subjective test. The Frequency Variation Model had only one prediction falling inside the bounds set by the standard deviation of the subjective test. All of the predicted correct answers of the Spectral Contrast Model lay within the standard deviation (error bars in Figure 5.34) of the subjective test with the exception of the 0dB SNR test. Both of the objective tests did not decrease to the measured level at 0dB SNR.

The FITA analyses for all the subjective tests and tests of the Frequency Variation Model and the Spectral Contrast Model are shown in Figure 5.35, Figure 5.36, and Figure 5.37, respectively. Bar graphs are used to show the relationship between the acoustic cues in each level of SNR.

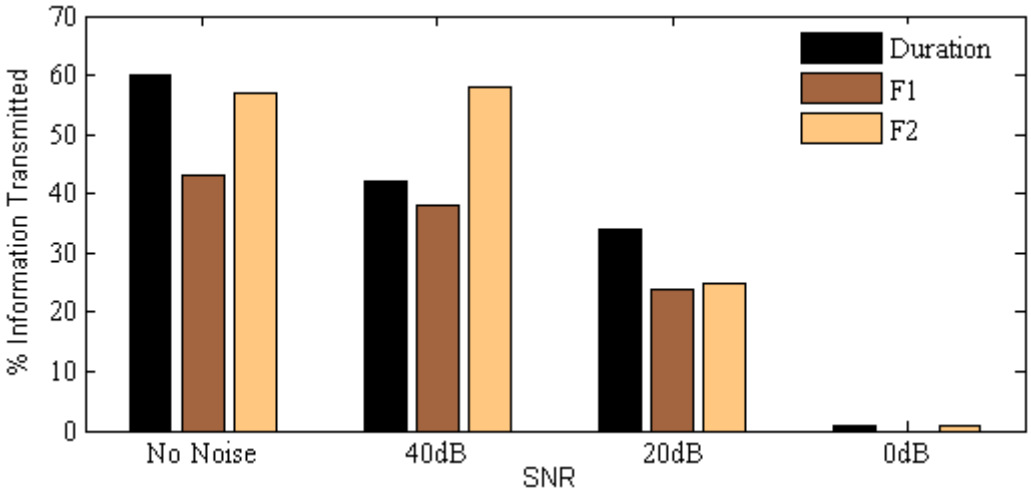


**Figure 5.35. FITA analysis for the subjective tests.**

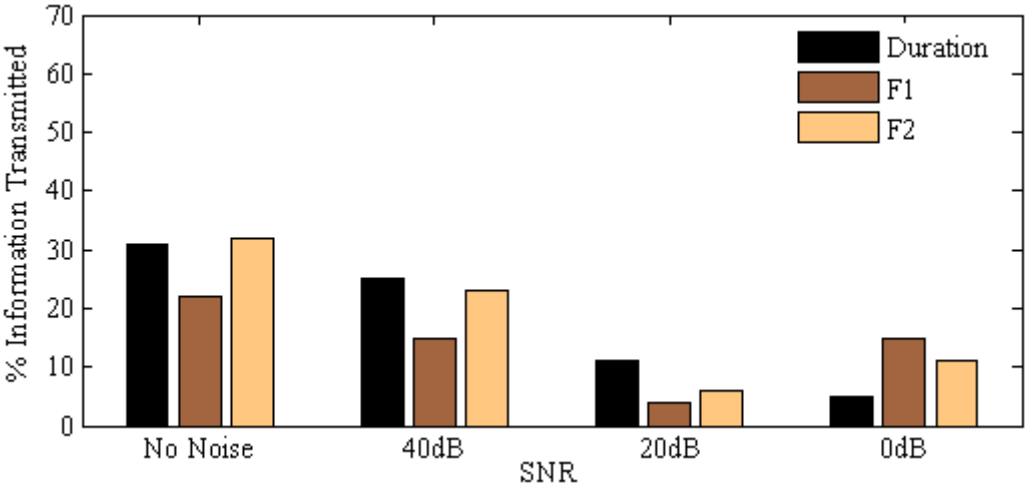


**Figure 5.36. FITA analysis for Frequency Variation Model.**

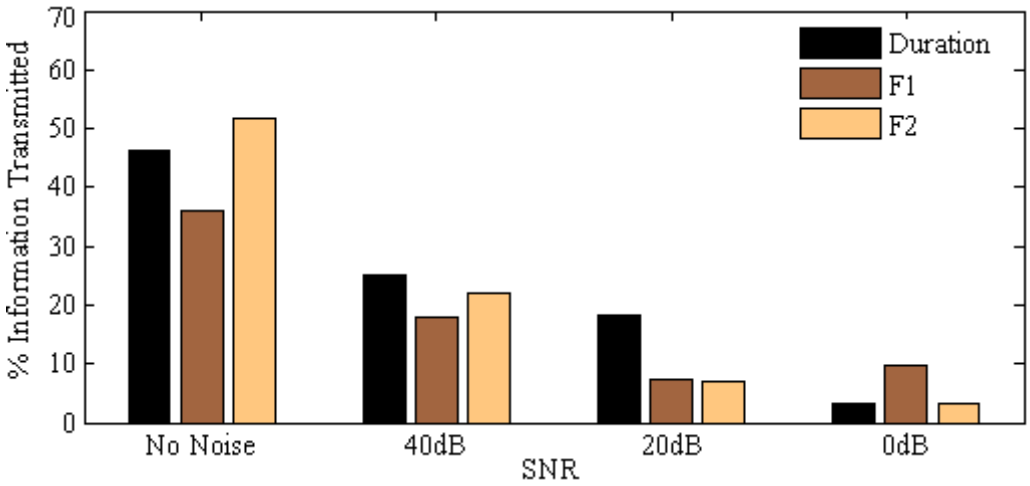


**Figure 5.37. FITA analysis for Spectral Contrast Model.**

The FITA analysis for the subjective tests shows a gradual decrease in the percentage of information transmitted as the signal to noise ratio decreases (with only one exception – see below). From no noise added up to the 20dB level there is a gradual decrease; between 20dB and 0dB the information transmitted drops to almost zero. There seems to be no indication as to which acoustic cue transfers the largest percentage of information throughout. For instance the duration cue has the highest percentage of the three cues for the test with no added noise and the 20dB SNR test; however, in the 40dB test the F2 frequency carries a higher percentage of information correctly transmitted (this is the only exception to the general trend).

The FITA analysis for the Frequency Variation Model also shows a gradual decrease in the percentage of information transmitted up to 20 dB SNR; below this level there is an increase. For each SNR level, however, the information transmitted for each acoustic cue is much lower than in the subjective tests. In fact, for the test with no additional noise, the 40dB SNR test and the 20dB SNR test, the information transmitted for all cues is half the value for the subjective test. The Frequency Variation Model does not utilize the cues well, and the model undershoots the values of the subjective test by a large margin, except for the 0dB SNR test in which the information transmitted actually increases. The Spectral Contrast Model also does not perform well under extremely noisy conditions (as can be seen in the 0dB SNR test results). In general, the F1 frequency transmits the least information in all the tests; the exception occurs in the inflated results for the F1 frequency

in the 0dB SNR tests for both the objective models.

The FITA analysis for the Spectral Contrast Model shows a better correlation with the analysis for the subjective tests. For speech with no additional noise there is a good correlation with the subjective test and the ratio between the cues is similar to the subjective model. However, the results of the objective model decreased too rapidly with respect to the subjective tests. In the 0 dB SNR test the information transmitted for all cues was close to the subjective test, except for the information for the F1 acoustic cue which was much too high.

The FITA analyses for the two objective models and the subjective tests are compared in the following figures to show the downward trend approximation for each acoustic cue separately. This is done to show a more direct comparison for each acoustic cue.

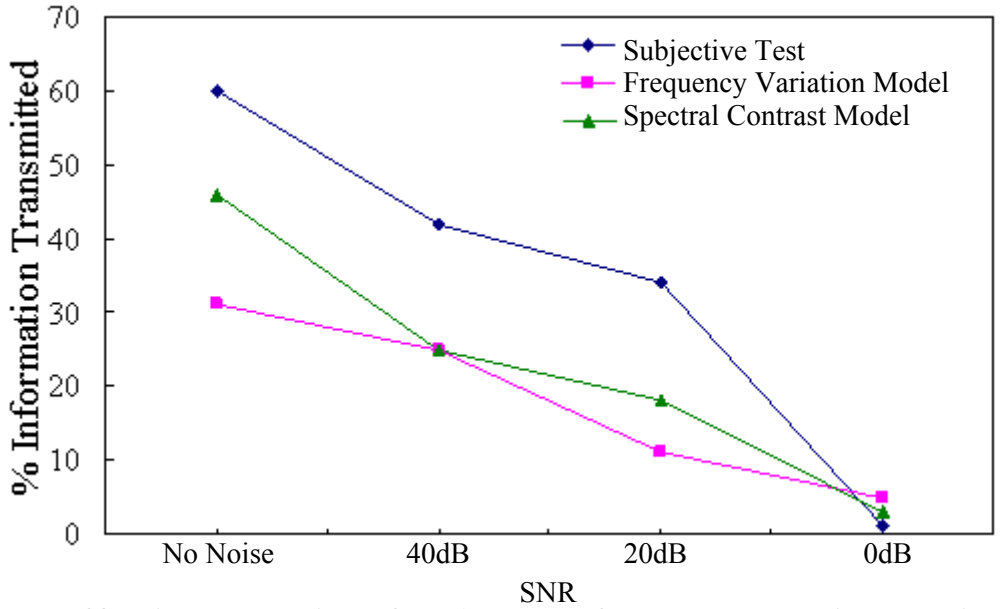


**Figure 5.38. Direct comparison of FITA analyses for the vowel duration acoustic cue.**

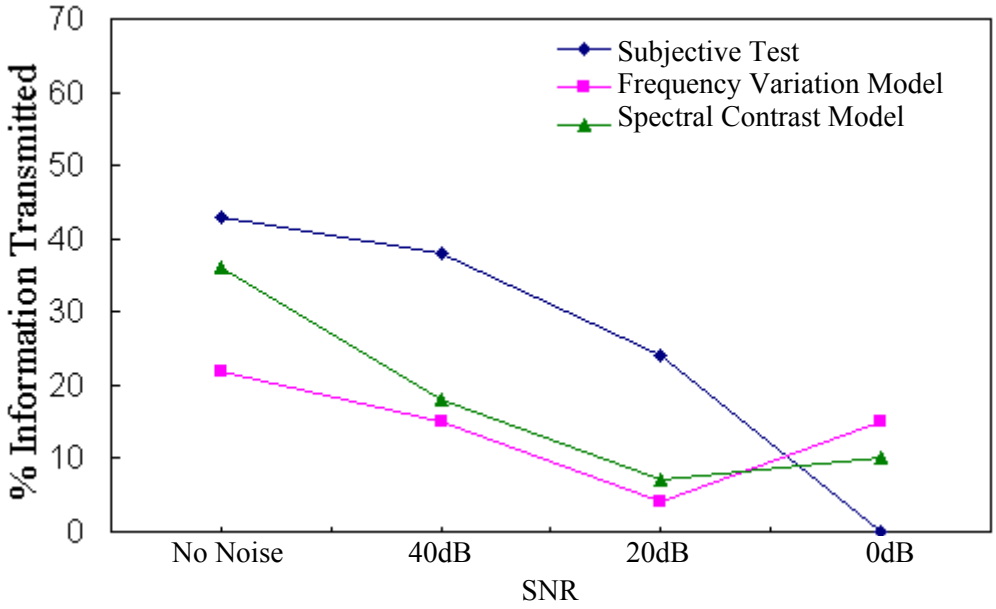


**Figure 5.39. Direct comparison of FITA analyses for the F1 acoustic cue.**

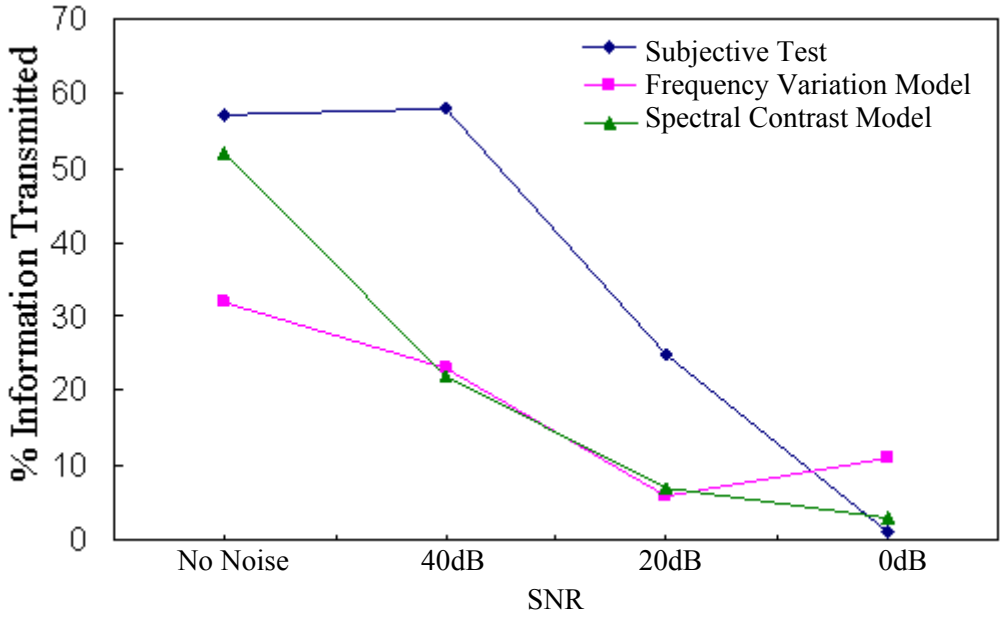


**Figure 5.40. Direct comparison of FITA analyses for the F2 acoustic cue.**

The duration and F1 cue show information transmitted decreases as noise is added (once again ignoring the 0dB SNR test). One anomaly, however, is that in the subjective test the F2 acoustic cue transmitted more information in the 40dB test than in the test with no noise. This was unexpected since the general trend of the correct identification scores and

the other cues shows a constant decrease as noise is added. Both the objective models show a rapid decrease for the F2 cue information transmitted from the no noise to the 40dB SNR test. This suggests that the lack of spectral contrast or increased frequency variation do not necessarily mean that the cue is unusable in identification. There are other factors which still allow the information of cues to be transmitted and used to identify the vowel.

Another interesting observation is that the rate of change for the subjective test increases with the addition of noise. This is not reflected in the results of the two objective models, that show a sharp decrease in information transmitted initially which slows down as more noise is added. This shows that the information transmitted by all of the acoustic cues is not directly related to the measure of spectral contrast and frequency variation, or at least that the relationship is not linear.

From these graphs the Sum of Squared Error (SSE) was calculated to show quantitatively which of the two methods provided the closest approximation to the subjective tests. The following equation is used to calculate the SSE.

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 , \tag{5.3}$$

where $n$ is the number of tests, $\hat{Y}$ is the amount of information transmitted for the specific cue by the prediction model and Y is the information transmitted for the specific cue in the subjective test. The results from the SSE calculation for each acoustic cue are shown in Table 5.27.

**Table 5.27. Sum of Squared Error comparison between the two objective methods and the subjective tests.**

| | Acoustic Cues | | |
|---|---|---|---|
| **Sum of Squared Errors** | **Duration** | **F1** | **F2** |
| Frequency Variation Method | 1675 | 1595 | 2311 |
| Spectral Contrast Method | 745 | 838 | 1649 |

The SSE calculations show clearly that the Spectral Contrast Model provides a better approximation to the subjective method. For the vowel duration cue and the F1 frequency cue the Spectral Contrast Model has only half the error in prediction compared to the Frequency Variation Model. For the F2 frequency cue it performs almost two-thirds more accurately than the Frequency Variation Model. The conclusion can be drawn that spectral contrast plays a larger role in approximating the masking of acoustic cues than the variation of the formant frequency. The frequency variation does not seem to camouflage the identity of the formant frequencies as much as was assumed at the outset.

For further analysis of the role of frequency variation and spectral contrast, the following figures give a comparison between frequency variation and spectral contrast in relation to the percentage correct answers. The graphs were scaled so that they could be plotted on the same set of axes; the y-axis on the left gives the percentage correct answers and the y-axis on the right gives the uncertainty factor measurements in dB.

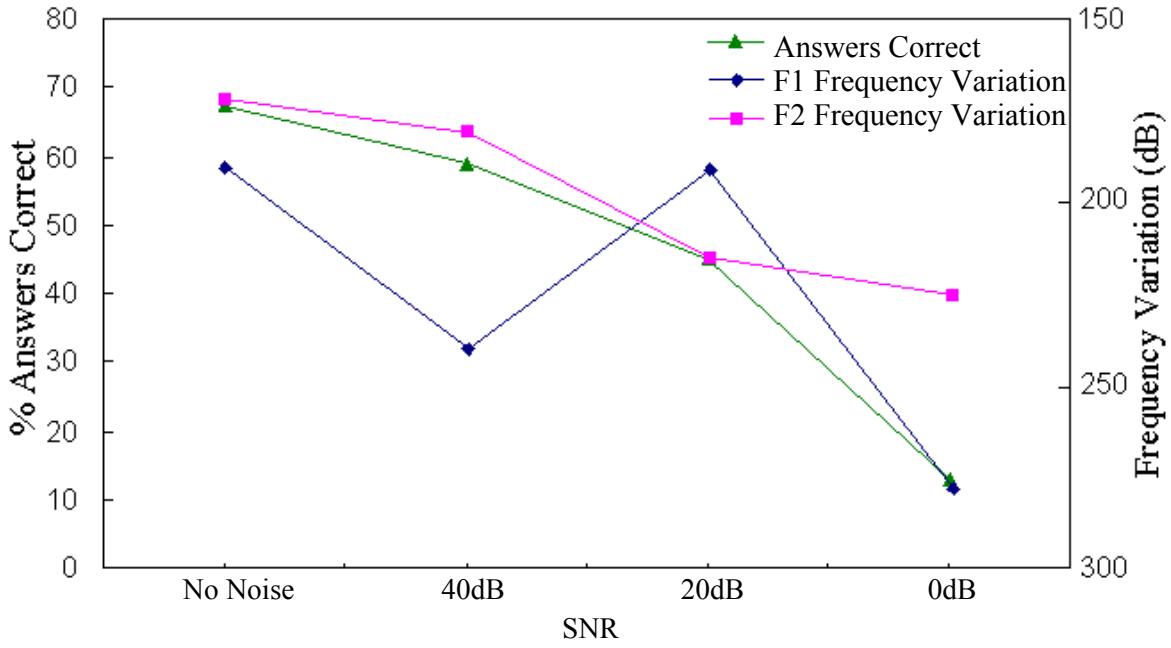


**Figure 5.41. Trend comparison of percentage answers correct for the subjective test and the measured frequency variation of the objective models.**

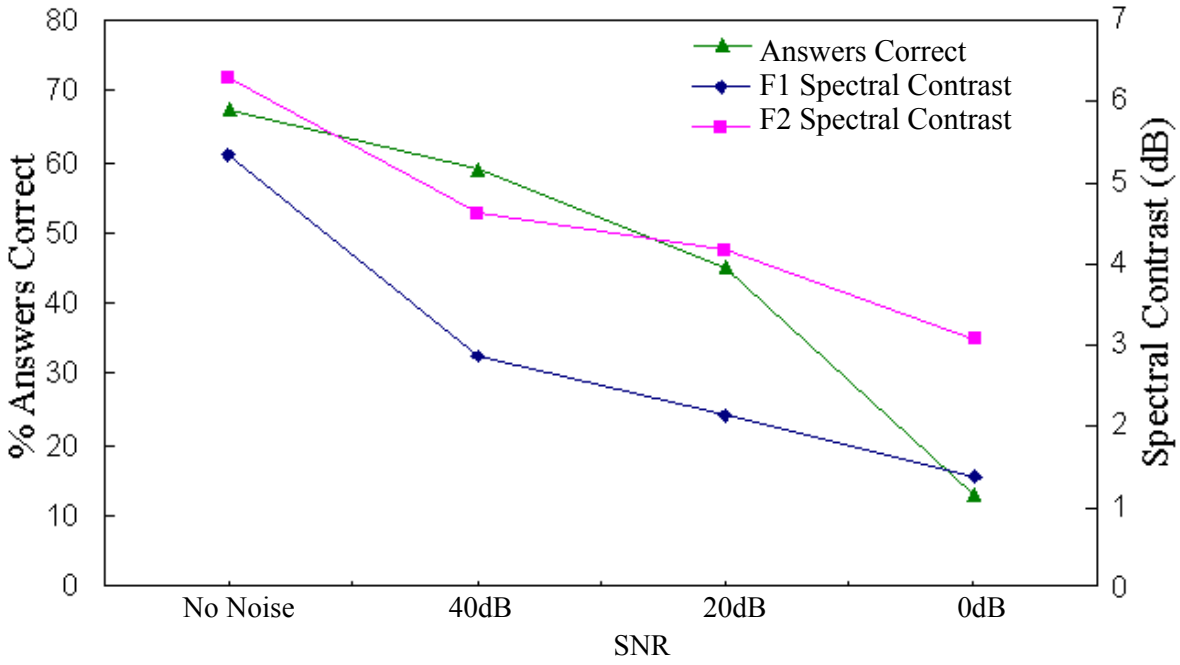


**Figure 5.42. Trend comparison of percentage answers correct for the subjective test and the measured spectral contrast of the objective models.**

Figure 5.41 shows that the variation measured for the F1 frequency on average does not decrease consistently with the addition of noise. The frequency variation of F2 seems to be consistent with the downward trend of the percentage answers correct, except for the 0dB SNR test. On the contrary, the F1 frequency variation decreases from the 40dB to the 20dB SNR test which creates an anomaly in the trends (note that the scale is reversed for frequency variance – higher frequency variance should create less correct answers.) This anomaly could be as a result of new spectral peaks introduced by the background noise, which has a speech-like nature. At low frequency ranges the peak-picking of the model selects new peaks which are introduced around the frequency of the first formant and, therefore, the standard deviation decreases. This does give an indication as to why the Frequency Variation Model did not perform adequately.

In Figure 5.42, spectral contrast for the first and second formant is shown in comparison to the percentage correct scores for each SNR test. Similar to the results for frequency variation, the F2 uncertainty corresponds well to the trend of the percentage correct

answers. The F1 spectral contrast is also consistent with the percentage answers correctly given. The figure suggests that there is a relationship between the spectral contrast in the signal and the intelligibility of vowel sounds. It also supports the methodology of using spectral contrast in the speech intelligibility model as an uncertainty factor allowing for different SNRs to be predicted by the model.

## 5.4 RESEARCH INSIGHTS

The vowel perception model developed in this study allows for various hypotheses to be tested in terms of human perception of noisy speech. The specific cues used for the evaluation of the model were the F1 and F2 formant frequencies. Literature studies indicated vowel duration to be the most important. The results of tests conducted in this study, showed that using only the three acoustic cues is not enough to predict all the vowel confusions properly. The model, however, did give a first approximation to the results of subjective tests suggesting that this is a viable approach to predict confusions. From the less successful results in predicting all the confusions, deductions can be made about the acoustic cues used in the model.

It can be deduced that the acoustic cues did not contain all of the information used by a listener to identify vowel sounds in the presence of background noise. Secondary cues probably aid vowel identification if the primary cues are masked by noise. This accounts for the fact that the percentage correct scores of the objective models were always lower than the results of the subjective test (with the exception of the 0dB SNR test where the results can be attributed to chance alone).

Furthermore, the model was implemented so that each of the three acoustic cues had an equal effect on vowel identification. This is not necessarily the case, and experimentation can be done by applying weights to the acoustic cues. It was also assumed that the acoustic cues are independent from each other and the spectral contrast measurements. From the subjective results it seemed that the duration cue was more important in the presence of noise and it may be beneficial to put a greater weighting on the duration cue.

The unique contribution made with the current study is the addition of variances to the vowel space. The use of individual variances in a vowel space has not been implemented in other models. Usually Euclidean distances between the vowels in a vowel space are used to calculate vowel relationship manually (Pretorius *et al.*, 2005; Remus and Collins, 2005; Van Wieringen and Wouters, 1999). The variances allow factors other than the acoustic cues alone to be incorporated in the decision criteria. It is important to include a variance in the form of an uncertainty factor, since the model would not be able to predict the deterioration of identification if this were not included in the model; that is, the model would have predicted the same level of confusion for all levels of SNR. This would happen since the location of the formants and the duration of the vowel do not change much as noise is added. Although tweaking of the uncertainty factors is still needed, the inclusion of the frequency variation and spectral contrast allowed for different SNR conditions to be tested.

The results in the present study showed that using spectral contrast as a masking factor allowed for an approximation of speech intelligibility. In a broad sense, vowel confusions were able to be predicted especially for vowels that were separated by the duration cue. The model, however, in its current form, still has shortcomings and improvement is needed to properly predict cochlear implant speech perception under conditions of background noise.

The results show that with multi-babble background noise added at a level of 0dB vowel identification can be attributed to chance (13% in the subjective tests). At this level the objective methods do not produce useful results. When analyzing the spectra it is seen that the spectral contrast does not diminish any further and cannot be used as an uncertainty factor. The spectral contrast is still visible, but now the stability of the frequency and amplitude of spectral peaks play a role in creating uncertainty. The variance in formant frequency also is not useful as an uncertainty factor. This is because either the movement in frequency is random at this stage or the algorithm does not pick up any increase of variance. Therefore, it is not recommended that these algorithms be utilized for tests where speech perception is no longer possible and correct results fall below 15%.

In retrospect multi-talker babble background noise was not the best type of noise to initiate experimentation. Multi-talker babble noise has an extreme masking effect on the speech tokens and produces the best real-world simulations. The unpredictable nature of the acoustic model, along with the extremely degraded spectral information, offered the algorithm a realistic, yet difficult challenge. It might have been a better approach to first do experimentation with Gaussian noise (which has less masking effect on speech) and use the results from this testing to improve on the model. Once the model's predictions were acceptable, further experimentation could be done with multi-babble noise. By testing with the multi-talker babble noise, the model was evaluated thoroughly in an appropriate real-world scenario. Pilot experiments performed with white noise (Pollack and Pickett, 1957) also showed that speech recognition in white noise did not deteriorate as it did in the presence of speech-like noise, therefore white noise is not an appropriate type of noise to use to simulate real-world conditions.

Rather than using multi-talker babble noise and the complete range of vowel sounds, this study shows that experimentation should first have been done with a smaller set of vowels. The fact that some confusion could be predicted showed that the basis of the model had merit. It is realistic to say that human speech perception is a highly resilient process and that it is to be expected that only a first approximation is possible with the limited set of elements employed.

## 5.5 COMPARISON WITH OTHER COCHLEAR IMPLANT RESEARCH

The literature shows that using actual listeners as subjects in cochlear implant research is still the most common means of testing specific hypotheses regarding cochlear implants (Parikh and Loizou, 2005; Pretorius *et al.*, 2005; Van Wieringen and Wouters, 1999). This is either done with cochlear implantees or with normal-hearing listeners listening to an acoustic CI model as performed for the subjective test in this study. Confusion matrices have been used extensively in these tests and are still being used regularly in cochlear implant studies to determine cochlear implant performance and to learn more about vowel perception of cochlear implantees. From the confusion matrices, information about

problematic vowel and consonant confusions can be learnt and used to gain insight into primary acoustic cues and the effect of background noise. Some of the results found in these studies can be compared with the findings of the present research.

The choice of acoustic cues used in the present model is based on previous studies done with normal-hearing listeners and cochlear implantees (see literature study in chapter 2). Only using the three cues chosen for this study, along with uncertainty factors, show limited, but promising, results in predicting vowel confusions and general correct answer trends (with changes in background noise levels). Inconsistencies still remain regarding the manner in which cues are used; and which cues are the most important in speech identification. For instance, Hillenbrand and Gayvert (1993) showed that vowels synthesized from the Peterson and Barney (1951) study typically achieved considerably lower identification rates in listening tests. There are also many arguments against formant representation of speech. Zahorian and Jagharghi (1993), for example, favoured a representation based on gross spectral shape. Lindblom and Studdert-Kennedy (1967) suggested that direction and rate of change of formants aid in identifying vowel sounds. Therefore although formant frequencies and vowel duration are found to be the most important acoustic cues there are other factors that need to be included in the model to accurately predict human vowel perception.

The study by Van Wieringen and Wouters (1999) concluded that the vowel duration and the F1 formant frequency are the most important acoustic cues used by Laura cochlear implantees for vowel identification. Van Wieringen and Wouters also commented that duration was especially important for the poorer subjects (cochlear implantees with very limited speech perception). The results are in line with the results of the present study in which the feature analysis showed that the cue which transmitted the most information was the duration cue followed closely by the F2 cue for most of the SNR tests.

The present study showed that the F2 cue was the second-most used acoustic in vowel identification. This also correlated with findings from Van Wieringen and Wouters (1999) and Pretorius *et al.* (2005). Although Van Wieringin and Wouters stated that the F2 and F3 frequencies were hardly used by the poorer subjects.

Contrary to these findings, Hillenbrand *et al.* (2000) established that changing the vowel duration had a small effect on vowel identification (with only a reported 5% decrease in the study). Their tests were done with normal hearing listeners. It seems then that cochlear implantees rely more on the duration cue than normal hearing listeners because of the inherent loss of spectral resolution due to the cochlear implant.

The fact that spectral contrast performed better than frequency variation as an uncertainty factor is to be expected, since various investigations have shown that spectral contrast is an important requirement for formant identification (Loizou and Poroy, 2001b). Studies have shown that hearing-impaired listeners need a larger spectral contrast compared to normal hearing listeners to achieve high vowel recognition performance (Leek *et al.*, 1987). The study by Loizou and Poroy (2001) suggests that a spectral contrast of between 4 – 6% is needed for proper vowel identification. The findings of the present study are in line with these previous findings. The study found that when the spectral contrast for the first two formants was above 5dB the percentage correct vowel recognition was close to 70%. The percentage correct vowel recognition dropped to below 50%, however, when the spectral contrast dropped below 4dB for both formants (see Figure 5.42).

The theory behind the Frequency Variation Model was based on the fact that formant frequencies in vowels are defined as being stable frequency peaks (Hillenbrand *et al.*, 1995; Peterson and Barney, 1952). Inspection of the spectrograms of the vowels in this study however showed erratic fluctuation in formant frequency in the processed vowels. The fact that the variation in frequency increased as the SNR increased supported this theory and an assumption was made that the frequency variation causes vowel confusions. In spite of this, the frequency variation method did not perform well when used in the vowel intelligibility model. No research was found to specifically measure the perceived threshold of unpredictable formant movement. Evidence do however exist which shows that static spectral characteristics, using only the average formant frequencies of vowels, provide very good speech perception (Hillenbrand and Gayvert, 1993; Kirk *et al.*, 1992). It was found that speech perception is however significantly lower than the 94.4% obtained by Peterson and Barney, but still suggests that static formant frequencies are very

important in vowel identification.

Related research has been done into the effect of poor frequency resolution. It has been shown that poor frequency resolution strongly contributes to CI users' difficulty in speech recognition in noisy listening situations (Fu and Nogaki, 2005). A study by Fu and Shannon (1999b) evaluated the recognition of spectrally degraded vowels; the study showed that spectral degradation and frequency shifting influenced the recognition of vowels negatively. Numerous studies have also investigated the minimum number of channels needed for speech recognition for a cochlear implant (Dorman, Loizou and Rainey, 1997; Fu and Shannon, 1999a; Shannon, Zeng, Kamath, Wygonski and Ekelid, 1995). Overall, the results from these studies indicate that a high level of speech recognition is possible with four to six bands of spectral information between 0 Hz and 4 kHz. This works out to between 666 Hz and 1000 Hz cover per cochlear implant channel. The frequency variation measured in all the SNR tests was never above 300 Hz, indicating that it is not beneficial as an uncertainty factor in light of the findings of the above studies.

As stated by Svirsky (2000), investigations into psychophysical variables and speech perception provide important information, but any correlations between psychophysics and speech perception, by their very nature, cannot explain the mechanisms CI users employ to identify speech sounds. Simply stating that speech perception is related to one specific psychophysical variable does not explain how listeners may actually use acoustic information to arrive at a higher-level decision. An objective speech intelligibility model such as the one developed in this study, once refined, can provide a more complete picture into speech perception than separate studies into single variables.

## 5.6 COMPARISON WITH OTHER MODELS

Speech prediction models for cochlear implants are not very common. Only two objective vowel intelligibility models which predict vowel confusions in confusion matrices were found in literature (Remus and Collins, 2005; Svirsky, 2000). Direct comparison of the performance of the model developed in the present study with these other models is compounded by the following:

- Different sets of vowels were evaluated in each of the studies. The present study was done with Afrikaans vowels, where Remus and Collins (2005) performed their experiments with English vowel sounds.
- Different methodologies where utilized to calculate predicted confusions. For instance Remus and Collins applied HMMs which are trained algorithms. The present study used a mathematical model based on signal detection theory.
- The literature studies either tested a subset of vowels or compared the most frequent confusions found. This study predicted confusions for all possible Afrikaans vowel sounds.
- The model developed in the present study was evaluated under different noise conditions. Other authors evaluated their model results under ideal noise conditions.

Rather than comparing the results directly, the methodology followed for the present model is compared with the methodologies used in the other models. The conclusions made in the other studies are also compared with the present study's conclusions.

At the time of writing, Svirsky (2000) was the only study found to use acoustic cues to produce prediction confusion matrices for cochlear implantees. There are a few differences between Svirsky's model and the model used in the present study. Svirsky did not use the acoustic cues from a reconstructed input spectrum as in this study, but rather the output of the channels of the cochlear implant as acoustic cues. The most important cue used was the F1 temporal cue encoded by the first implant channel; the other cue used was the difference between the amplitudes of the implant channel (simulating F1/F2 amplitude ratios). Svirsky used the acoustic cues to form a Cartesian vowel space, the same as was done also in the present study. Signal detection theory was used to calculate the perceptual distance between each vowel in the vowel space and the probabilities calculated were used to produce a confusion matrix.

The model developed in the present study was similarly implemented but instead of using

the output directly from the implant channel (as Svirsky did) to calculate the acoustic cues, the reconstructed speech spectrum was used to calculate the acoustic cue values in order to predict a decision. This was done to emulate human decision-making as a whole on the entire signal. The acoustic cues were then extracted from the signal and also used to produce a perceptual vowel space. The reason the present study based its calculation on the reconstructed speech is that this mimics more closely the vowel sound as a whole. This approach also allows for comparison with subjective testing which has been done in other studies. Svirksy's results where better than the results of the present study. It must follow then that the improvements attempted by the present study does not provide a more accurate model of speech perception.

The present model attempted to improve on Svirsky's methodology in two ways. Firstly, the present study's model was designed to derive an uncertainty factor from the physical signal. Svirsky did not measure an uncertainty factor from the signal but rather used the jnd (just noticeable difference) as measure of uncertainty. The *jnd* had to be estimated for each acoustic cue or it had to be measured by subjective psychophysical testing. The reasons the present study measured the uncertainty factor is to allow time-saving and to give a more quantitative measure for each dimension. In the Svirsky (2000) study, these measures were estimated and manipulated to improve the resulting predictions. Svirsky only showed the performance of the model for one test, where the present study showed that the model could function for various levels of SNR.

The use of spectral contrast as a measure produced results that indicate that this is a plausible method. However, this approach did not necessarily improve on Svirsky's method, since a direct comparison cannot be made because different acoustic cues were used and Svirsky only tested with three vowel sounds. Svirsky only showed that the model provided predictions for speech without background noise. If other SNRs were to be tested, measurements or estimations have to be made to find the *jnd* values. The present study's model calculated a specific measurement for each vowel sound. The accurate results produced by Svirksy's model in comparison with the less accurate results of the present study suggest that it is more accurate to use equal variance throughout.

Secondly, Svirsky implemented a reference vowel space where the acoustic cue information is inserted manually into the model. The model developed in this study extracts the vowel sound automatically from the token and generates the corresponding vowel space. This allowed that the vowel space would be specific to the test conditions (background noise, speaker, and language for instance). The model developed in this study also needs an unprocessed clean token as a guide for the model as to where to start searching for the formant frequencies.

The only other study found to use an objective model to predict the vowel intelligibility for cochlear implantees was by Remus and Collins (2005). These researchers also used an acoustic CI model which summed the separate channels back into a full spectrum (however, it is not apparent that they included the biophysical interface, which our acoustic model does). Remus and Collins developed three models. The first model predicted confusions by finding the Euclidean distances between the cepstrum coefficients of the stimulus and the possible response. The second model used the normalized inner product of the discrete envelopes of two processed speech tokens. The third model used a continuous-observation HMM which was trained for each speech token using a training set of 100 tokens. All training data were collected from a single male speaker in quiet conditions.

Remus and Collins used their model to predict confusions for nine English vowel sounds and fourteen English consonants under various SNR levels from 10dB to -2dB. Remus and Collins used generated speech shaped noise which is a synthetic replica of the multi-babble noise used in the present study's tests. Remus and Collins combined the confusion matrices across noise levels, which were justified by information transmission analyses, which indicated that increasing the amount of additive noise most significantly affected the rate of confusions rather than the pattern of confusions. The results of the present study showed that confusions did vary slightly as noise was added. This suggests that it is necessary to do discrete testing at different noise levels to properly assess the objective model. The present study assessed each of the noise levels separately and found that the relationship of acoustic cue transmission changed as noise was added to the stimulus.

The results from Remus and Collins' study showed that the discrete envelope comparison failed for all their tests. The method which compared cepstrum coefficient distances and the HMM method performed the best (close to 78% accuracy) for both models. Linear regression showed that the HMM method performed very well for vowel recognition ranking (at 96%). These results show much better correlation to the subjective testing than shown in the present study. Hidden Markov Models could have been used to improve the present model; however, this was decided against and will not be considered in future work because of the black-box nature of this approach. HMMs provide no insight into the psychophysical processing and decision-making process of in vowel perception.

In the Remus and Collins model, the cepstrum distance model was the only method that appeared to have any success predicting the correct identification trends for different token sets (for the confusion test). The one method that Remus and Collins implemented using MFCC might have been used to improve the present study's model. Mel-Frequency Cepstrum coefficients approximates the human auditory system's response more accurately than the normal linear frequency scale.

Remus and Collins concluded that their research did prove the concept but that future work was needed to improve the accuracy of the confusion predictions. Remus *et al* took the initial study further by using modifications of the original models to predict impaired channels in a cochlear implant (Remus, Throckmorton and Collins, 2007). This new study concluded that there is still significant improvement needed for all the results, but this method provided potential for expediting the identification of impaired channels of cochlear implants (Remus *et al.*, 2007).

The present model and the models described above go beyond the single answer score produced by traditional speech intelligibility models (Beerends *et al.*, 2002; Kamm, Dirks and Bell, 1985; Rix *et al.*, 2001; Steeneken and Houtgast, 1980; Voran, 1999b). No comparison is made with these models since the results produced is fundamentally different to the results and objectives of the present study.

The prediction of an entire confusion matrix allows for specific predictions about patterns of perceptual behaviour to be tested and also the testing of specific hypotheses. Consequently, the approach represented by the current model may be helpful in advancing our understanding of the role of sensory discrimination abilities and their relation to speech perception by CI users.

## 5.7 RESEARCH QUESTION FINDINGS

With respect to the research questions posed in chapter 1, the following conclusions can be drawn:

- Most speech intelligibility or assessment models found in literature provide single score results and are designed to predict normal-hearing performance under specific conditions. A small number of current speech prediction models were found that predict speech intelligibility for cochlear implant users. Only two studies were found that produced confusion matrices as output and only one of these studies was found to use acoustic cues to determine the outcome. These models are still in development and only give approximations of subjective tests and only for specific groups of confusions.
- The model developed in the present study showed that it is possible to use acoustic cue analysis to approximate the percentage of correct answers by CI users. The acoustic cues were shown to be correctly extracted from the speech token and used to generate a vowel space with variances calculated from uncertainty factors.
- The trends of confusions could be predicted to a certain extent when spectral contrast was used as an uncertainty factor. The prediction of percentage correct scores fell within the standard deviation of all noise tests except at a SNR of 0dB. The frequency variation implementation was unsuccessful in predicting any trends in confusions. It was concluded that frequency variation does not produce confusion in vowel identification, but spectral contrast plays a definitive role.
- The hypotheses that three important acoustic cues, namely the F1 and F2 formant frequencies and vowel duration, are satisfactory to predict vowel perception was

found to be true to a certain degree. The fact that only some confusions could be predicted led to the conclusion that other secondary acoustic cues also aid listeners in identifying vowels in the presence of noise.

- Feature analysis showed that the predictions made by the objective model transmitted less information in comparison to the results in the subjective testing. The relationship between the acoustic cues for the Spectral Contrast Model was found to be similar to the subjective test. For the tests where the SNR was 0 dB, the objective model no longer followed the trend of the subjective test.
- Although the model struggled to predict all confusions made by CI users, it was good enough to prove the concept that the model can be useful for confusion testing. Not all confusions were predicted correctly but trends of group confusions could be predicted. For instance, groups with larger duration differences were predicted well and vowels with specific confusions were predicted well. Other, more random, confusions (for example, vowels which were confused with more than three other vowels) were not predicted correctly.
- The model could predict the deterioration of results in the subjective evaluation test when noise was added to the speech being evaluated. Once again the Spectral Contrast Model performed to a level that fell within the standard deviation of the subjective results. The results of the Frequency Variation Model did deteriorate as noise was added, but not to the same degree as the subjective results.

# CHAPTER 6 CONCLUSION

The goal of the current study was to investigate the methodology of designing a vowel intelligibility perception model that can predict objectively the outcome of a vowel confusion test performed with a cochlear implantee. The output of the model is the calculated probability of a cochlear implant user identifying vowel tokens correctly and the probability of specific vowel confusions occurring. Instead of using training mechanisms to produce more accurate answers, the model attempts to mimic vowel perception mathematically.

The model was developed based on the primary acoustic cues shown in the literature to aid vowel identification. Two versions of the model were developed and tested. The first variant of the model measured an uncertainty factor in terms of formant frequency variation and the second variant measured an uncertainty factor in terms of spectral contrast. This allowed the model to track vowel perception under real-world conditions; it was tested in the presence of various levels of additional multi-talker babble background noise.

The results from the Frequency Variation Model did not provide answers which correlated well with subjective tests done using persons with normal hearing listening to an acoustic cochlear implant model. The measurement of frequency variation did not increase continuously when background noise was added, and the model did not provide a proper representation of the vowel space.

The Spectral Contrast Model showed more promising results. Some of the most frequent confusions could be predicted. Using spectral contrast as a variance in the vowel space showed good correlation with the average answers correctly given for the different SNR tests. The spectral contrast measurement added value in that it showed that the model adapts its answers according to the noise level. This feature allowed for the modest

successes.

In summary, the following conclusions can be drawn from this study.

- The present model successfully extracted acoustic cues and used these to create three-dimensional vowel perception spaces.
- FITA analysis showed that the transmitting of the acoustic cues in the Spectral Contrast Model had similar trends to that in the subjective test. (This does not apply to the Frequency variation Model.)
- The hypotheses that three important acoustic cues in terms of F1 and F2 formant frequencies and vowel duration are satisfactory to predict vowel perception was found to be true to a certain degree, since the most frequent confusions could be predicted in some tests.
- The Spectral Contrast Model produced results which were satisfactory as a first approximation for the most frequent confusions.
- The Frequency Variation Model did not produce accurate predictions for vowel confusions. This shows that the use of variation in frequency is not adequate as an uncertainty factor.

The Spectral Contrast Model could predict the deterioration of results in the subjective evaluation test when noise was added to the speech being evaluated. (Again, this does not apply to the Frequency variation Model.)

Although the objective vowel intelligibility model (using spectral contrast) was, at best, only partially successful at predicting vowel confusions, the methodology followed in the development of the new model forms a framework from which further study is possible.

Once the objective model presented in this study is refined it may be used to speed up research into cochlear implant speech processing strategies. Various scenarios can be set up and the effect of changing specific CI parameters can be determined by analysing the output of the acoustic model. The model may also lead to new means of quantitative research into speech recognition of cochlear implantees. In such a way, the developed

model could make a contribution in the improvement of current cochlear implant processors. Models such as the present one may in future aid cochlear implant research by replacing or complementing subjective listening tests. By experimenting with different acoustic cues, the model can then also be used to gain more insight into the elements in a vowel sound that provide information of the vowel's identity to a cochlear implant listener.

## 6.1 FUTURE WORK

The vowel perception model is essentially based on the measurement and processing of acoustic cues and uncertainty factors using signal detection theory in order to predict vowel confusion probabilities. There are various modifications which can be done to improve on the current implementation.

The model can be improved by taking into account other acoustic cues that have been identified in the literature as aids to human sound perception. For instance, instead of using only the frequency of the first two formants, the direction of the glide of the formant frequency in time can also be taken into account (Hillenbrand and Nearey, 1999). Researchers have also shown that relative amplitude of the first two frequencies in the presence of noise also aid in vowel identification (Ito, Tsuchida and Yano, 2001b). If this suggestion is implemented, the scaling of the vowel space would change, which would, in turn, influence the outcome. Whole spectrum models have also been proposed to be an acoustic cue in vowel identification (Hillenbrand *et al.*, 2006; Ito *et al.*, 2001b). These cues can replace the current three acoustic cues if they are found to be more important under specific testing conditions or they can be added as secondary cues to supplement the current calculation. It must be noted that adding other acoustic cues to the model will increase the vowel space by one dimension for every acoustic cue added. It is the modular nature of the present model that allows for the easy substitution of various types of input.

In the current model the acoustic cues are all assumed to contribute to the same extent to the identification of a vowel sound. Further experimentation should be done to determine if the model's predictions can be improved by changing the weightings for the different acoustic cues. An acoustic cue may also carry a certain impact if it falls in a certain area of

the vowel space. This has been done in previous models, for example, in models using the STI (Speech Transmission Index) and the SII (Speech Intelligibility Index). The STI and SII are based on weighted contributions from a number of frequency bands. For this purpose, the STI uses a fixed bandwidth (octave bands) with a relative weighting summed to provide an intelligibility index between 0 and 1 (Steeneken and Houtgast, 1980; Steeneken, 1987). The same methodology may be implemented to improve the current model.

In the present study it was also assumed that no interdependencies existed between the acoustic cues. The absence of interaction between the acoustic cues causes the dimensions of the vowel space to be orthogonal to each other. This is not necessarily the case, as was concluded by Van Wieringen and Wouters (1999) after multidimensional scaling in their tests did not account for all the confusions in subjective tests. Further research can be carried out into changing the dependency between the acoustic cues and by including co-variances into the probability equations.

The frequency spectrum was calculated from the FFT of the signal in the implemented model. The frequency spectrum, however, is not the best scale to use in calculating human speech recognition. For various speech perception models the Mel-Frequency Cepstrum (as used by Remus and Collins (2005)) has been implemented in calculations instead of the frequency spectrum. This was done because the frequency bands in the Mel-Frequency Cepstrum are positioned logarithmically (on the Mel scale) which approximates the human auditory system's response more closely. This approach may not aid in determining the formant frequencies, but it may give better results for the spectral contrast or other measurements.

In future work, the uncertainty factors can be extended to include other distortions in the signal that mask the acoustic cues. Other effects have also been found to affect human speech recognition. The following signal distortions that affect the variances in the vowel spaces can be investigated in future, for example, spectral resolution (Fu and Nogaki, 2005; Fu, Shannon and Wang, 1998b) and amplitude distortion. There are also uncertainty factors which affect the vowel's position in the vowel space, such as spectral smearing (Fu

and Nogaki, 2005), spectral warping (Fu and Nogaki, 2005) and spectral shifting (Fu and Shannon, 1999b).

Finally, the methods could be extended to include the evaluation of consonant recognition in addition to vowel recognition. This study only looked at vowel perception because of the importance of vowel identification in speech perception. It is important, however, to also extend the study into consonant perception for an overall insight into cochlear implantee speech perception.

Once the vowel intelligibility prediction model functions at an acceptable level, it is envisioned that the model can be used in an automated program which runs through various parameters of the cochlear implant. All combinations of parameter settings of a cochlear implant can then be evaluated with the model and the summed perceptual score for each test can be stored. By ranking the summed results for each of the tests, the best parameter settings for a cochlear implant under certain speech input conditions can be determined. The present model was only tested using a cochlear implant model which uses the SPEAK speech processing strategy. Other speech strategies (for example, CIS) can also be used in series with the model to evaluate speech perception for individuals fitted with those types of cochlear implants.

It is the modular nature of the present model that allows for further fine-tuning. The scope for further research using the present model (particularly the Spectral Contrast Model) is vast, and will culminate eventually in real benefits for the users of CI implants.

# REFERENCES

Assmann, P. F. and Katz, W. F. (2005). Synthesis fidelity and time-varying spectral change in vowels, *Journal of the Acoustical Society of America*, 117(2): 886-895.

Atal, B. S. and Schroeder, M. R. (1978). Linear prediction analysis of speech based on a pole-zero representation, *Journal of the Acoustical Society of America*, 64(5): 1310-1318.

Beddor, P. S. and Hawkins, S. (1990). The influence of spectral prominence on perceived vowel quality, *Journal of the Acoustical Society of America*, 87(6): 2684-2704.

Beerends, J. G., Hekstra, A. P., Rix, A. W. and Hollier, M. P. (2002). Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part II - Psychoacoustic model, *AES: Journal of the Audio Engineering Society*, 50(10): 765-778.

Berouti, M., Schwartz, R. and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise, *IEEE International Conference on Acoustics, Speech, and Signal Processing, April, 1979, Cambridge, MA,* Vol. 4, pp. 208-211.

Blamey, P. J., Dowell, R. C. and Brown, A. M. (1987). Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors, *Journal of the Acoustical Society of America*, 82(1): 48-57.

Boersma, P. and Weenink, D. (2001). PRAAT, a system for doing phonetics by computer, *Glot International*, 5(9/10): 341-345.

Chang, C. H., Anderson, G. T. and Loizou, P. C. (2001). A neural network model for optimizing vowel recognition by cochlear implant listeners, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 9(1): 42-48.

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

Delgutte, B. (1984). Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds, *Journal of the Acoustical Society of America*, 75(3): 879-886.

Dorman, M. F., Loizou, P. C. and Rainey, D. (1997). 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(1): 783-788.

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.

ETSI Standard EG 201 377-1 (2002). Speech Processing, Transmission and Quality Aspects (STQ); Specification and measurement of speech transmission quality; part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks, *ETSI EG 201 377-1*.

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(3): 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.

Frisch, S. A. and Pisoni, D. B. (2000). Modeling spoken word recognition performance by pediatric cochlear implant users using feature identification, *Ear and Hearing*, 21(6): 578-589.

Fu, Q. J. and Nogaki, G. (2005). Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing, *Journal of the Association for Research in Otolaryngology*, 6(1): 19-27.

Fu, Q. J. and Shannon, R. V. (1999b). 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. and Shannon, R. V. (1999a). Effect of acoustic dynamic range on phoneme recognition in quiet and noise by cochlear implant users, *Journal of the Acoustical Society of America*, 106(6): L65-L70.

Fu, Q. J. and Shannon, R. V. (2000). Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners, *Journal of the Acoustical Society of America*, 107(1): 589-597.

Fu, Q. J., Shannon, R. V. and Wang, X. (1998a). 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.

Fu, Q. J., Shannon, R. V. and Wang, X. (1998b). 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.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics,* John Wiley and Sons, New York.

Greenberg, S., Ainsworth, W. A., Popper, A. N. and Fay, R. R. (2004). *Speech processing in the auditory system,* Springer, New York.

Han, W., Chan, C. F., Choy, C. S. and Pun, K. P. (2006). An efficient MFCC extraction method in speech recognition, *Proceedings - IEEE International Symposium on Circuits and Systems, 21 April, 2006, Hong Kong,* Vol. 145-148.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech, *Journal of the Acoustical Society of America*, 87(4): 1738-1752.

Hillenbrand, J. and Gayvert, R. T. (1993). Identification of steady-state vowels synthesized from the Peterson and Barney measurements, *Journal of the Acoustical Society of America*, 94(21): 668-674.

Hillenbrand, J. M., Clark, M. J. and Houde, R. A. (2000). Some effects of duration on vowel recognition, *Journal of the Acoustical Society of America*, 108(6): 3013-3022.

Hillenbrand, J. M., Getty, L. A., Clark, M. J. and Wheeler, K. (1995). Acoustic characteristics of American English vowels, *Journal of the Acoustical Society of America*, 97(51): 3099-3111.

Hillenbrand, J. M., Houde, R. A. and Gayvert, R. T. (2006). Speech perception based on spectral peaks versus spectral shape, *Journal of the Acoustical Society of America*, 119(6): 4041-4054.

Hillenbrand, J. M. and Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour, *Journal of the Acoustical Society of America*, 105(6): 3509-3523.

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

Holden, L. K., Skinner, M. W., Holden, T. A. and Demorest, M. E. (2002). Effects of stimulation rate with the nucleus 24 ACE speech coding strategy, *Ear and Hearing*, 23(5): 463-476.

Ito, M., Tsuchida, J. and Yano, M. (2001a). On the effectiveness of whole spectral shape for vowel perception, *Journal of the Acoustical Society of America*, 110(2): 1141-1149.

Ito, M., Tsuchida, J. and Yano, M. (2001b). On the effectiveness of whole spectral shape for vowel perception, *Journal of the Acoustical Society of America*, 110(2): 1141-1149.

ITU-T Recommendation P.862 (2000). Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, *ITU-T Recommendation P.862*.

Iverson, P., Smith, C. A. and Evans, B. G. (2006). Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration, *Journal of the Acoustical Society of America*, 120(6): 3998-4006.

Jenkins, J. J., Strange, W. and Edman, T. R. (1983). Identification of vowels in "vowelless" syllables, *Perception and Psychophysics*, 34(5): 441-450.

Kamm, C. A., Dirks, D. D. and Bell, T. S. (1985). Speech recognition and the Articulation Index for normal and hearing-impaired listeners, *Journal of the Acoustical Society of America*, 77(1): 281-288.

Kewley-Port, D. and Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels, *Journal of the Acoustical Society of America*, 95(1): 485-494.

Kewley-Port, D. and Zheng, Y. (1999). Vowel formant discrimination: Towards more ordinary listening conditions, *Journal of the Acoustical Society of America*, 106(5): 2945-2958.

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), *Journal for Oto-Rhino-Laryngology and Its Related Specialties*, 58(3): 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.

Kirk, K. I., Tye-Murray, N. and Hurtig, R. R. (1992). The use of static and dynamic vowel cues by multichannel cochlear implant users, *Journal of the Acoustical Society of America*, 91(6): 3487-3498.

Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step, *IEEE International Conference on Acoustics, Speech, and Signal Processing, May, 1982, Cambridge, MA,* Vol. 7, pp. 1278-1281.

Leek, M. R., Dorman, M. F. and Summerfield, Q. (1987). Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners, *Journal of the Acoustical Society of America*, 81(1): 148-154.

Leek, M. R. and Summers, V. (1996a). Reduced frequency selectivity and the preservation

of spectral contrast in noise, *Journal of the Acoustical Society of America*, 100(3): 1796-1806.

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

Lindblom, B. E. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition, *Journal of the Acoustical Society of America*, 42(4): 830-843.

Liu, C. and Fu, Q. J. (2007). Estimation of vowel recognition with cochlear implant simulations, *IEEE Transactions on Biomedical Engineering*, 54(1): 74-81.

Liu, C. and Fu, Q. J. (2005). Relating the acoustic space of vowels to the perceptual space in cochlear implant simulations, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 18 March, 2005, Philadelphia, PA, USA,* Vol. III, pp. III33-III36.

Liu, C. and Kewley-Port, D. (2004b). Vowel formant discrimination for high-fidelity speech, *Journal of the Acoustical Society of America*, 116(2): 1224-1233.

Liu, C. and Kewley-Port, D. (2004a). Formant discrimination in noise for isolated vowels, *Journal of the Acoustical Society of America*, 116(5): 3119-3129.

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

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 Magazine*, 18(1): 32-42.

Loizou, P. C., Dorman, M., Poroy, O. and Spahr, T. (2000). Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution, *Journal of the Acoustical Society of America*, 108(51): 2377-2387.

Loizou, P. C., Dorman, M. F. and Powell, V. (1998). The recognition of vowels produced by men, women, boys, and girls by cochlear implant patients using a six-channel CIS processor, *Journal of the Acoustical Society of America*, 103(2): 1141-1149.

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

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

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model, *Ear and Hearing*, 19(1): 1-36.

Makhoul, J. (1975a). Linear Prediction: A Tutorial Review, *Proceedings of the IEEE*, 63(4): 561-580.

Makhoul, J. (1975b). Spectral linear prediction: properties and applications, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(3): 283-297.

Marieb, E. N. (2004). *Human Anatomy and Physiology,* Benjamin Cummings Publishing Company, Redwood City, CA, USA.

McCandless, S. S. (1974). An algorithm for automatic formant extraction using linear prediction spectra, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(2): 135-141.

Meyer, T. A., Frisch, S. A., Pisoni, D. B., Miyamoto, R. T. and Svirsky, M. A. (2003). Modeling open-set spoken word recognition in postlingually deafened adults after cochlear implantation: Some preliminary results with the neighborhood activation model, *Otology and Neurotology*, 24(4): 612-620.

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.

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.

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.

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.

Parikh, G. and Loizou, P. C. (2005). The influence of noise on vowel and consonant cues, *Journal of the Acoustical Society of America*, 118(6): 3874-3888.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels, *Journal of the Acoustical Society of America*, 24(2): 175-184.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A. and Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon, *Speech Communication*, 4(1-3): 75-95.

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. (2006), 'n Analitiese tegniek om die foneemherkenningsvermoë van Suid-Afrikaanse kogleêre inplantingsgebruikers te bepaal (An analytical technique to determine phoneme recognition capability of South African Cochlear Implant Users), *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie (South African Magazine for Natural*

*Sciences and Technology)*, 25(4): 195-208.

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.

Rabiner, L. R. (1989). Tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77(2): 257-286.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981), Speech perception without traditional speech cues, *Science*, 212(4497): 947-950.

Remus, J. J. and Collins, L. M. (2004). Vowel and consonant confusion in noise by cochlear implant subjects: Predicting performance using signal processing techniques, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, May, 2004, Montreal, Quebec, Canada,* Vol. Vol 4, pp. IV(13)-IV(16).

Remus, J. J. and Collins, L. M. (2005). Expediting the identification of impaired cochlear implant acoustic model channels through confusion matrix analysis, *2nd International IEEE EMBS Conference on Neural Engineering, March, 2005, Washington, DC,* Vol. 418-421.

Remus, J. J., Throckmorton, C. S. and Collins, L. M. (2007). Expediting the identification of impaired channels in cochlear implants via analysis of speech-based confusion matrices, *IEEE Transactions on Biomedical Engineering*, 54(12): 2193-2204.

Rix, A. W., Beerends, J. G., Hollier, M. P. and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs, *IEEE International Conference on Acoustics, Speech and Signal Processing, May, 2001, Salt Lake City, Utah,* Vol. vol 2, pp. 749-752.

Rix, A. W., Hollier, M. P., Hekstra, A. P. and Beerends, J. G. (2002). Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part I - Time-delay compensation, *Journal of the Audio Engineering Society*, 50(10): 755-764.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects, *Philosophical transactions of the Royal Society of London*, 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(5234): 303-304.

Shepherd, R. K., Hatsushika, S. and Clark, G. M. (1993). Electrical stimulation of the auditory nerve: The effect of electrode position on neural excitation, *Hearing Research*, 66(1): 108-120.

Shepherd, R. K. and McCreery, D. B. (2006). Basis of electrical stimulation of the cochlea

and the cochlear nucleus, *Advances in oto-rhino-laryngology*, 64: 186-205.

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., Clark, G. M., Whitford, L. A., Seligman, P. M., Staller, J. S., Shipp, D. B., Shallop, J. K., Everingham, C., Menapace, C. M., Arndt, P. L., Antogenelli, T., Brimacombe, J. A., Pijl, S., Daniels, P., George, C. R., McDermott, H. J. and Beiter, A. L. (1994). Evaluation of a new spectral peak coding strategy for the nucleus 22 channel cochlear implant system, *American Journal of Otology*, 15(2): 15-27.

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.

Skowronski, M. D. and Harris, J. G. (2002). Increased MFCC filter bandwidth for noise-robust phoneme recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing, May, 2002, Orlando, FL,* Vol. 1, pp. 801-804.

Snell, R. C. and Milinazzo, F. (1993). Formant location from LPC analysis data, *IEEE Transactions on Speech and Audio Processing*, 1(2): 129-134.

Steeneken, H. J. M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality, *Journal of the Acoustical Society of America*, 67(1): 318-326.

Steeneken, H. J. M. (1987). Diagnostic information from subjective and objective intelligibility tests, *IEEE International Conference on Acoustics, Speech and Signal Processing, April, 1987, Dallas, Texas,* Vol. 5-8.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context, *Journal of the Acoustical Society of America*, 85(5): 2135-2153.

Summerfield, Q. and Assmann, P. F. (1989). Auditory enhancement and the perception of concurrent vowels, *Perception and Psychophysics*, 45(6): 529-536.

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.

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

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.

Thorpe, L. and Wonho, Y. (1999a). Performance of current perceptual objective speech quality measures, *IEEE Workshop on Speech Coding - Proceedings, June, 1999a, Porvoo, Finland,* Vol. 144-146.

Thorpe, L. and Wonho, Y. (1999b). Performance of current perceptual objective speech quality measures, *Speech Coding Proceedings, 1999 IEEE Workshop on,* 144-146.

Tyler, R. S., Tye-Murray, N. and Otto, S. R. (1989). The recognition of vowels differing by a single formant by cochlear-implant subjects, *Journal of the Acoustical Society of America*, 86(6): 2107-2112.

Vainio, M., Suni, A., Jarvelainen, H., Jarvikivi, J. and Mattila, V. V. (2005). Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish, *Journal of the Acoustical Society of America*, 118(31): 1742-1750.

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

Voran, S. (1998). Simplified version of the ITU algorithm for objective measurement of speech codec quality, *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, May, 1998, Seattle, WA,* Vol. 1, pp. 537-540.

Voran, S. (1999a). Objective estimation of perceived speech quality - Part I: development of the measuring normalizing block technique, *IEEE Transactions on Speech and Audio Processing*, 7(4): 371-382.

Voran, S. (1999b). Objective estimation of perceived speech quality - Part I: development of the measuring normalizing block technique, *IEEE Transactions on Speech and Audio Processing*, 7(4): 371-382.

Voran, S. (1999c). Objective estimation of perceived speech quality - Part II: evaluation of the measuring normalizing block technique, *IEEE Transactions on Speech and Audio Processing*, 7(4): 383-390.

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

Watson, C. I. and Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels, *Journal of the Acoustical Society of America*, 106(1): 458-468.

Werner, M., Kamps, K., Tuisel, U., Beerends, J. G. and Vary, P. (2003). Parameter-based speech quality measures for GSM, *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, September, 2003, Beijing, China,* Vol. 3, pp. 2611-2615.

Whitlon, D. S. (2004). Cochlear development: hair cells don their wigs and get wired,

*Current Opinion in Otolaryngology & Head and Neck Surgery*, 12(5): 449-454.

Wickens, T. D. (2002). *Elementary Signal Detection Theory,* 1st Edition, Oxford University Press, Oxford.

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.

Yost, W. A. (2006). *Fundamentals of Hearing: An Introduction,* Fifth Edition, Academic Press, New York.

Zahorian, S. A. and Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels, *Journal of the Acoustical Society of America*, 94(4): 1966-1982.

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.

Zheng, F., Zhang, G. and Song, Z. (2001). Comparison of different implementations of MFCC, *Journal of Computer Science and Technology*, 16(6): 582-589.

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