CUE ESTIMATION FOR VOWEL PERCEPTION PREDICTION IN LOW SIGNAL-TO-NOISE RATIOS

by

Brian Burmeister

Submitted in partial fulfilment of the requirements for the degree

Master of Engineering (Electronic)

in the

Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

October 2008

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CUE ESTIMATION FOR VOWEL PERCEPTION PREDICTION IN LOW SIGNAL-TO-NOISE RATIOS by

Brian Burmeister

Promotor: Prof. J. J. Hanekom

Department of Electrical, Electronic and Computer Engineering

Master of Engineering (Electronic)

SUMMARY

This study investigates the signal processing required in order to allow for the evaluation of hearing perception prediction models at low signal-to-noise Ratios (SNR). It focusses on speech enhancement and the estimation of the cues from which speech may be recognized, specifically where these cues are estimated from severely degraded speech (SNR ranging from -10 dB to -3 dB). This research has application in the field of cochlear implants (CI), where a listener would hear degraded speech due to several distortions introduced by the biophysical interface (e.g. frequency and amplitude discretization). These difficulties can also be interpreted as a loss in signal quality due to a specific type of noise. The ability to investigate perception in low SNR conditions may have application in the development of CI signal processing algorithms to counter the effects of noise. In the military domain a speech signal may be degraded intentionally by enemy forces or unintentionally owing to engine noise, for example. The ability to analyse and predict perception can be used for algorithm development to counter the unintentional or intentional interference or to predict perception degradation if low SNR conditions cannot be avoided. A previously documented perception model (Svirsky, 2000) is used to illustrate that the proposed signal processing steps can indeed be used to estimate the various cues used by the perception model at SNRs successfully as low as -10 dB.

Keywords: Hearing perception model, speech enhancement, speech cue estimation, low signal-to-noise ratio

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SPRAAKEIENSKAPESTIMASIE VIR KLINKERPERSEPSIE-VOORSPELLING BY LAE SEIN-TOT-RUIS VERHOUDINGS deur

Brian Burmeister

Leier: Prof. J. J. Hanekom

Departement Elektriese, Elektroniese en Rekenaar-Ingenieurswese

Meester van Ingenieurswese (Elektronies)

SAMEVATTING

Hierdie studie ondersoek die seinprosessering wat nodig is om ’n gehoorpersepsievoorspelling-model te evalueer by lae sein-tot-ruis-verhoudings. Hierdie studie fokus op spraakverbetering en die estimasie van spraakeienskappe wat gebruik kan word tydens spraakherkenning, spesifiek waar hierdie eienskappe beraam word vir ernstig gedegradeerde spraak (sein-tot-ruis-verhoudings van -10 dB tot -3 dB). Hierdie navorsing is van toepassing in die veld van kogleêre inplantings, waar die luisteraar degradering van spraak ervaar weens die bio-fisiese koppelvlak (bv. diskrete frekwensie en amplitude). Hierdie degradering kan gesien word as ’n verlies aan seinkwaliteit weens ’n spesifieke tipe ruis. Die vermoë om persepsie te ondersoek by lae sein-tot-ruis kan toegepas word tydens die ontwikkeling van kogleêre inplanting-seinprosesseringalgoritmes om die effekte van ruis teen te werk. In die militêre omgewing kan spraak deur vyandige magte gedegradeer word, of degradering van spraak kan plaasvind as gevolg van bv. enjingeraas. Die vermoë om persepsie te ondersoek en te voorspel in die teenwoordigheid van ruis kan gebruik word vir algoritme-ontwikkeling om die ruis teen te werk of om die verlies aan persepsie te voorspel waar lae sein-tot-ruis verhoudings nie vermy kan word nie. ’n Voorheen gedokumenteerde persepsiemodel (Svirsky, 2000) word gebruik om te demonstreer dat die voorgestelde seinprosesseringstappe wel suksesvol gebruik kan word om die spraakeienskappe te beraam wat deur die persepsiemodel benodig word by sein-tot-ruis verhouding so laag as -10 dB.

Sleutelwoorde: Gehoorpersepsiemodel, spraakkwaliteitverbetering, spraakeienskapberaming, lae sein-tot-ruis

ACKNOWLEDGEMENTS

I wish to thank my study leader, Prof J. J. Hanekom, for his continued support and interest in this field of study. A conscious effort was made to align this study to be of benefit to both the Bio-engineering research group at the University of Pretoria and my employer, the CSIR Defence Peace Safety and Security (DPSS).

Further, I wish thank my employer for continued support in terms of time and finances. Without this support, my part-time studies would have been much more difficult.

LIST OF ABBREVIATIONS

AR	Auto-regressive
AWGN	Additive White Gaussian Noise
BW	Bandwidth
CAGO	Cell Averaging Greater Of
CDF	Cumulative Distribution Function
CFAR	Constant False Alarm Rate
CI	Cochlear Implant
CIS	Continuous Interleaved Sampling
cm	centimetre
CVC	Consonant Vowel Consonant
dB	decibel
DFT	Discrete Fourier Transform
EM	Expectation Maximization
F0	Fundamental Frequency
F1	First Formant Frequency
F2	Second Formant Frequency
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
Hz	Hertz
JND	Just Noticeable Difference
LPC	Linear Predictive Coding
MPI	Multidimensional Phoneme Identification
ms	millisecond
PDF	Probability Density Function
RMS	Root Mean Square
RTI	Relative Transmitted Information
SNR	Signal-to-noise ratio

TABLE OF CONTENTS

LIST OF FIGURES

studies by Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005) (green, black, magenta and cyan lines in Figure 5.4). 96

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Problem Statement

An understanding of how well words are perceived in low signal-to-noise ratio (SNR) conditions (SNR of -3 dB and lower) has application in the military, medical and commercial communications fields. Low SNR conditions during communication can be due to various types of noise sources. Examples of communication under low SNR conditions are the internal communication system in a fighter aircraft (Smith & Lourens, 2006) or helicopter (Acker-Mills, Houtsma, & Ahroon, 2006). In these cases the source of the interference would be unintentional and to some extent under the control of the owner of the respective systems. Examples of such sources of noise are wind and the engine in an aircraft cockpit. Another example would be attempting wireless communication while the level of interference is intentionally increased (Nixon, McKinley, & Moore, 1982), in order to disrupt the effectiveness of the wireless communication. The ability to analyse and predict perception in these conditions can be used for algorithm development to counter unintentional or intentional interference. Perception prediction can also be used to predict perception degradation if low SNR conditions cannot be avoided. Cochlear implants (CI), which are used to restore the hearing of severely deafened people, are particularly susceptible to the effects of background noise (Remus & Collins, 2004b). Moore (2003) presents an overview of the typical difficulties experienced by the hearing impaired, These include reduced audibility, reduced frequency selectivity and loudness recruitment. These difficulties can be caused by regions in the cochlea that have no surviving inner hair cells and/or neurons (dead regions) and can also be interpreted as a loss in signal quality due to a specific type of noise. The ability to investigate perception in low SNR conditions may have application in the development of CI signal-processing algorithms to counter the effects of noise.

Quantitative models for the perception of speech by humans provide important insights for the development of Automatic Speech Recognition (ASR) algorithms (Alwan et al., 1995). Various methods of predicting speech perception are documented (Remus & Collins, 2004a; Remus & Collins, 2004b; Strope & Alwan, 1997a; Strope & Alwan, 1997b; Svirsky, 2000). However, the lowest SNR at which these perception prediction methods are evaluated is -2 dB. This study will refer to speech as severely degraded when the SNR of the speech signal is lower than 0 dB.

These perception predictors use estimates of one or more characteristics of the input speech signal in order to predict human perception of the speech input. According to the Concise Oxford English Dictionary the action of classifying is to "arrange (a group) in classes or categories according to shared qualities or characteristics." The estimated characteristics (also referred to as features in the context of classification and cues in the context of perception) are thus required for the classification method to predict human speech perception, where perception can be interpreted as arrangement into categories. Moreover, perception prediction is also concerned with correctly predicting wrong classifications, for example, if a listener is presented with the word "sat", but perceives the word to be "sit", the perception predictor should make the same classification. The typical processing steps required to solve a classification problem are shown in Figure 1.1, based on Bishop (1995).

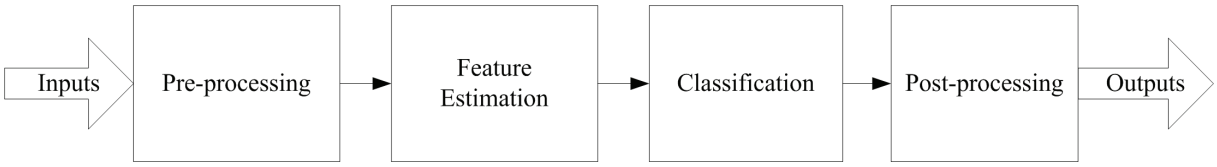


Figure 1.1: Typical processing steps required so solve a classification problem.

For speech perception prediction with the use of a classifier, the input is a speech signal. Pre-processing can be the conditioning of the input into the appropriate format to allow for feature estimation. Feature estimation refers to the calculation of values quantifying the chosen characteristics of the input. Post-processing refers to the final data conditioning that may be required to modify the classifier output into a format appropriate for the specific analysis.

From Figure 1.1 it can be seen that the ability to estimate the features required for the classification successfully, and thus perception prediction, is a crucial step in the classification process. The SNR at which the features can be estimated also determines the SNR at which perception prediction can be performed. As will be discussed in the literature study (chapter 2), the exact set of cues used by human listeners when perceiving severely degraded speech is not well defined. These cues need to be selected before designing algorithms to estimate the cues. The nature of the cue to be estimated dictates the type of mathematical techniques appropriate for estimating its value.

As far as is known, nothing has been published on algorithms for the automatic estimation of speech cues for use in perception prediction models at SNR lower than -2 dB. There are, however, other domains where there are many publications on the processing of small signals in the presence of noise. Examples of these are radar target detection (Hovanessian, 1973; Leung, 1996; Leung & Young, 2000), image processing (Ffrench, Zeidler, & Ku, 1997), remote sensing (Carlotto, 1997) and robotics (Chang & Song, 1997). Techniques from the radar signal processing body of knowledge were applied in this study for the purpose of cue estimation.

If a signal processing algorithm can be developed that can estimate the selected cues of severely degraded speech, predictions of speech intelligibility, and the cues used in these predictions, can be analyzed.

1.2 Approach

This study aimed to expand the work in the field of perception prediction by developing methods to extract the chosen cues from severely degraded speech. This study focussed on speech cue estimation for speech with an SNR of -3 dB and lower, with the aim to allow for perception prediction. The approach was to develop a chain of signal processing techniques which are suitable to (i) enhance the signal of interest (speech) by suppressing noise and (ii) estimate the cues required for speech perception prediction from this signal at the specified low SNRs. In order to evaluate the performance of the speech enhancement and cue estimation, this investigation used previously documented cues (Svirsky, 2000) and a classification method as presented by Svirsky (2000). The cues used for perception prediction are estimated for vowels degraded by additive white Gaussian noise (AWGN). The processing steps presented were not intended as a model of the auditory system. However, the ability to evaluate perception prediction models with severely degraded speech as input may provide some insight as to the auditory mechanisms used during human speech perception in low SNR conditions. The aim of this dissertation was neither to evaluate the particular set of cues or different sets of cues used for classification of vowels, nor to evaluate a specific classifier or different classifiers used for perception prediction, but to allow for the use of existing choices of cues and perception prediction models in very low SNR.

1.3 Research Questions

From the earlier introduction and the literature study (chapter 2), specific research questions can be phrased to address perception prediction for severely degraded speech. From available literature it appears that it is generally accepted in perception prediction models that speech cues are readily available, but this is not the case for severely degraded speech. The primary research question was: using existing signal processing, can an algorithm be developed and successfully applied to severely degraded speech in order to enable the estimation of speech cues as required by perception prediction models? Also, can the signal processing be performed automatically and without any a priori knowledge regarding the input? Given that the selected speech perception prediction model can be evaluated at the low SNRs of interest, a secondary research question would be: do these predictions follow the trends in available published data?

1.4 Objectives

The objective of this study was to develop a signal processing algorithm which would estimate the selected speech cues from severely degraded speech. These estimated cues were used with a pre-existing human perception prediction model (Svirsky, 2000) for SNRs as low as -10 dB. It was expected that as the SNR approached -10 dB the error in the various estimates would increase and thus the recognition performance as predicted by the perception prediction model would decrease. Specific objectives to be achieved during the development of the signal-processing algorithm were:

1. All the signal-processing steps of the algorithm should be automatic. The input to the algorithm would be the degraded speech signal and the output would be the selected speech cues. Such an automated algorithm would allow for repeatable perception predictions without any uncertainty, which may arise if a person performed cue estimation. For example, uncertainties such as decision thresholds would be eliminated
2. In order to make the algorithm as widely applicable as possible the signal-processing techniques should require no a priori knowledge regarding the speech input signal, for example, is the speaker male or female?
3. Owing to the severely degraded nature of the speech input, a form of signal enhancement was required. This signal-processing technique used for the signal

enhancement should be suitable for the enhancement of speech signals, and it was attempted to keep the structure of the signal-enhancement algorithm related to literature on human auditory perception (Watkins & Paus, 2004).

4. For reasons to be outlined in the literature study (chapter 2), this study focussed on the enhancement, cue estimation and perception prediction of vowels. For this reason, a voicing detection technique was required to isolate the vowel in a syllable in order to allow for the cue estimation of the vowel only.

5. To test the performance of the signal enhancement, voicing detection and the cue estimation, a perception prediction model was required. The multidimensional phoneme identification (MPI) model of Svirsky (2000) was implemented and evaluated at SNRs as low as -10 dB. The results of the MPI model evaluation were compared to the listening experiments by Boothroyd and Nittrouer (1988), whose experiments evaluated perception performance of normal hearing listeners at SNRs as low as -10dB.

1.5 Contribution

This study documents the various signal-processing steps required for the cue estimation, the implementation of the classifier and the evaluation of various signal-processing techniques. The study showed that with the proposed signal enhancement, cues can be estimated in low SNR conditions in order to allow for perception prediction with documented classifiers (Svirsky, 2000). The methods described in this dissertation will have the following applicability:

1. The ability to enhance a severely degraded speech signal so that it can be applied to a perception prediction model will allow for the evaluation of a number of already existing perception prediction models (Remus & Collins, 2004a; Svirsky, 2000) for SNRs previously not evaluated by these models.

2. The ability to evaluate perception prediction models at previously unevaluated SNRs may provide insight into the perceptual mechanisms used by a listener in low SNR conditions.

3. The methods used to enhance the noise-degraded signal and to extract cues can be used in the military domain in research to counter the effects of intentional and unintentional noise, typically encountered in an operational environment.

Specific contributions to the body of knowledge regarding speech perception were:

1. A number of independent signal-processing techniques were combined in the development of an algorithm to allow for perception prediction of severely degraded speech.
2. The use of a constant false alarm rate (CFAR) detector (to be defined in section 3.3.1) as a voicing detector. This detector was successfully used to determine the voiced section of a spectrogram for SNRs as low as -10 dB.
3. The perception prediction model of Svirsky (2000) was evaluated at SNRs not previously documented. These results were compared to listening tests performed on normal, hearing listeners (Boothroyd & Nittrouer, 1988; Parikh & Loizou, 2005) with a high degree of correlation (greater than 95%) between the results.

1.6 Dissertation Outline

The dissertation is structured as follows: Chapter 2 is a literature study describing the relevant background literature, to contextualize the work and identify shortcomings in the available literature. Chapter 3 documents the methods used in the study and will start by providing an overview of the entire processing chain used for the cue estimation and classification. This is done to create the context for the subsequent sections, which describe the various processing steps in more detail. Section 3.2 discusses speech enhancement by means of a Kalman filter and expectation maximization, as well as the methods used to characterize the performance of the filter. Section 3.3 discusses the processing required to estimate the cues of interest and section 3.4 documents the classification algorithm used in order to generate a measure of perception prediction performance. Chapter 4 documents the results of the investigation into the algorithm performance, focussing on the contribution of speech-enhancement processing. The performance of the implemented speech enhancement, cue estimation and perception prediction were analyzed over a range of SNRs using confusion matrices and information transmission analysis. Chapter 5 is a discussion of the results (chapter 4) and the algorithm

processing steps in general. Concluding remarks and comments on possible future work are given in chapter 6.

CHAPTER 2 LITERATURE STUDY

This literature study gives the background and motivation for the formulation of the primary research question. As discussed in the introduction, the signal processing required to enable perception prediction of severely degraded vowels was investigated. The motivation for using vowels in perception-prediction models is given, as well as background on speech cues used by perception-prediction models. To enable processing to be performed on a vowel, the vowel has to be extracted from the severely degraded input. Background regarding signal-processing techniques used to isolate the vowel in the input signal is discussed. The signal-processing techniques used to estimate speech cues from severely degraded speech are dictated by the specific cue to be estimated, thus an understanding of these cues is required in the development of a cue estimation algorithm. To evaluate the signal-processing techniques a perception prediction model is required. Existing perception-prediction models are discussed with specific focus on the model selected for the evaluation of the signal-processing techniques used for cue estimation. The selected model is used for vowel perception prediction of CI users. From the presented literature it will become clear that neither the selected model nor other CI perception-prediction models have been evaluated for SNRs below -2 dB. Even though perception-prediction models have not been evaluated lower than -2 dB, listening experiments have been performed on human listeners using vowels, for SNR as low as -10 dB. These data will be important in evaluating the performance of the selected perception prediction model.

As background on perception and the effect of noise on it, work by Fletcher and Galt (1950), as well as Dubbelboer and Houtgast (2007), can be considered. The study of speech perception has various aspects on which one can focus (Fletcher & Galt, 1950). The process that enables a listener to interpret and to repeat sounds that are spoken correctly is the interpretation (intelligibility) aspect. The loudness aspect of speech allows a listener to determine whether a sound which was heard is loud or soft. It can also be determined if the pitch is high or low, which is the pitch aspect, and finally one can determine the quality of the voice of the speaker. The quality can indicate if it is a child's voice, a woman's voice, or a man's voice, or if the voice is harsh or pleasing. Various factors can influence the intelligibility of speech, for example echoes, phase distortion and reverberation (French & Steinberg, 1947). In a much later study Dubbelboer and Houtgast (2007) investigated the effects of noise on speech intelligibility. The effects of noise on speech were divided into three sub-effects. The first was a systematic

lift of the envelope of the speech signal equal to the mean noise intensity. Second was the introduction of stochastic envelope fluctuations and third the corruption of the fine temporal structure. This study deals with the intelligibility aspect of perception, where the sounds are degraded owing to noise, specifically investigating the estimation of certain cues (features) of vowels.

The reason for specifically focussing on vowels is that work by Strange (1989) suggests that problems in explaining the perception of a speakers' intended message arise from variations in vowels as actually produced. Acoustically vowels can be represented using a multi-dimensional acoustic space. The vowels can be represented as coordinates in this multi-dimensional space where the axes of this coordinate system can be the first and second formant frequencies (F1/F2) or the first, second and third formant frequencies (F1/F2/F3). Other possible spaces may exist, based on the cues selected as the primary cues contributing to vowel intelligibility. For example Svirsky (2000) used F1 and the root mean square (RMS) channel amplitude ratios of four bandpass filters (refer to Figure 2.2 for detail regarding the bandpass filters), while Van Wieringen and Wouters (1999) identified F1, F2 and duration as the most important cue used by cochlear implantees.

There is a clear relationship between the recognition of phoneme-like units (Nearey, 2001) and ultimately the recognition of words. This relationship is well documented by Allen (1994) who reviews the comprehensive work of Fletcher (1953). Figure 2.1 illustrates the various recognition layers involved when perceiving words. This is of particular importance since this study specifically investigates the perception-prediction performance of phones, which are evaluated in a consonant-vowel-consonant (CVC) context. Phoneme perception will be investigated by extracting cues thought to be important for the identification of the vowel and then using these cues as inputs to an algorithm to associate the cues with a vowel (a classifier). Consider the syllable "pat"; only the cues of the "a" will be estimated and used as inputs to the classifier.

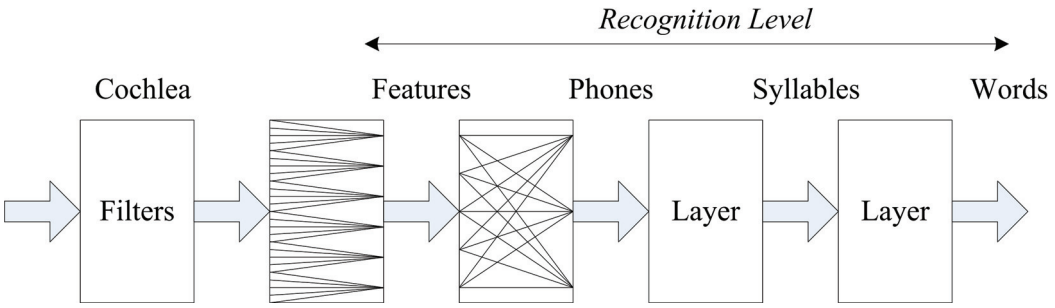


Figure 2.1: Hypothetical cascade of recognition layers, based on Allen (1994).

In order to estimate the various cues required for vowel perception, the vowels have to be isolated within the word. This task may be made more difficult in the presence of noise. The concept of glimpsing (Cooke, 2006; Howard-Jones & Rosen, 1993; Miller & Licklider, 1950), which uses the most energetic regions of speech in a spectrotemporal (spectrogram) representation for the purpose of identifying speech in noise, is particularly well suited to the task of isolating vowels in noise. A survey conducted by Gong (1995) also indicated that essential elements in noisy speech recognition are the incorporation of time and frequency correlations, as well as placing higher emphasis on high SNR portions of speech. The redundant nature of the information conveyed in a spectrogram (Cooke, 2006; Fletcher, 1953; Kasturi et al., 2002) makes this form of speech representation particularly useful for low SNR speech analysis. For the detection of a glimpse, Cooke (2006) used a detection model which assumes that spectrogram elements whose local SNR exceeds 3 dB are a glimpse of the speech signal that can be used for classification.

The exact set or combination of cues which are used by normal hearing and CI users for vowel perception are not clear at this stage, especially in low SNR conditions. The importance of the first and second formant frequencies, as well as duration in the perception of vowels in normal hearing listeners, is well established (Klein, Plomp, & Pols, 1970; Nooteboom & Doodeman, 1980; Peterson & Barney, 1952; Stevens, 1959). The extent to which the various cues are used may also vary; for example, it was found that the importance of the vowel duration increases if the vowel can easily be confused with other vowels by only analysing the first and second formant frequencies (Ainsworth, 1971). Other cues that may be used for vowel perception prediction are the ratio of the formant frequencies (Miller, 1989; Potter & Steinberg, 1950). Van Wieringen and Wouters (1999) showed that the first and second formant frequencies, as

well as duration, are also used as cues by CI listeners for identification. As discussed in the introduction, the MPI model of Svirsky (2000) was used for the evaluation of the proposed algorithm for speech enhancement and cue estimation. In recognition studies on Ineraid multichannel cochlear implant users, Svirsky (2000) used the first formant frequency and the amplitude ratios of four bandpass filters as vowel cues. Figure 2.2 illustrates the compressed analog stimulation strategy of the Ineraid multichannel cochlear implant. The speech signal is filtered into four overlapping frequency bands, with crossover frequencies at roughly 700 Hz, 1.4 kHz and 2.3 kHz.

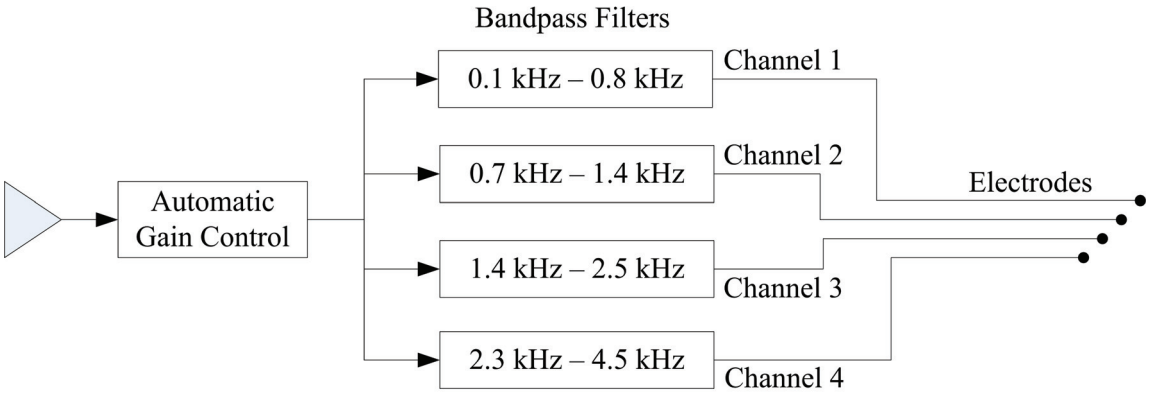


Figure 2.2: Block diagram, based on Svirsky (2000), of the compressed analog stimulation strategy of the Ineraid multichannel cochlear implant.

Formant frequencies can be identified by isolating the spectral peaks from a single cross-section through a steady state portion of an acoustic signal (Peterson, 1952). Peterson and Barney (1952) showed that formant frequencies are not invariant with respect to vowels across men, women and children, and that vowels often overlap in the multi-dimensional acoustic space (Hillenbrand et al., 1995). To address the invariant nature of formant frequencies, the concept of speaker normalization is reviewed by Miller (1989), who states that the use of the ratios of the centre frequencies of the first three formants can reduce and nearly eliminate speaker differences. This approach by Miller (1989) of using ratios of information from the spectral domain is similar to the approach by Svirsky (2000) for the cues selected for the MPI model.

Various forms of classifiers have been investigated for the purpose of perception prediction of CI users. Remus and Collins (2004a; 2004b) evaluated three classification techniques to predict

vowel and consonant confusions. These are envelope correlation, Euclidean distance between Mel-cepstrum coefficients and Hidden Markov Models (HMM, using Mel-cepstrum coefficients). They found that the classifiers using the cepstral representations were better suited for confusion prediction than the classifier using the temporal envelope representation. However, the lowest SNR at which the various models by Remus and Collins were evaluated was -2 dB. Svirsky (2000) used multivariate Gaussian distributions with an Euclidean decision rule as a classifier to predict vowel perception for CI users. Using the MPI model proposed by Svirsky, an entire confusion matrix can be generated, which can easily be compared to results obtained in listener experiments. Svirsky did not, however, evaluate his model at very low SNRs, as his focus was on determining if the proposed MPI model can successfully predict vowel confusions made by CI users.

Boothroyed and Nittrouer (1988) investigated the recognition performance of humans for phones by using CVC words and nonsense CVC syllables in the presence of spectrally shaped noise. The noise was spectrally shaped to have an equal masking effect for all frequencies. Data from their results are shown in Figure 2.3. It was attempted to generate similar results for the same SNR range.

In summary then, there seems to be no literature available on the automatic estimation of speech cues for use in a perception-prediction model when the input speech is severely degraded. Signal-processing techniques were developed to enable the evaluation of the MPI model by Svirsky at SNRs similar to those evaluated by Boothroyed and Nittrouer (1988). The input to the MPI model was CVC vowels for SNRs of -10 dB to 10 dB.

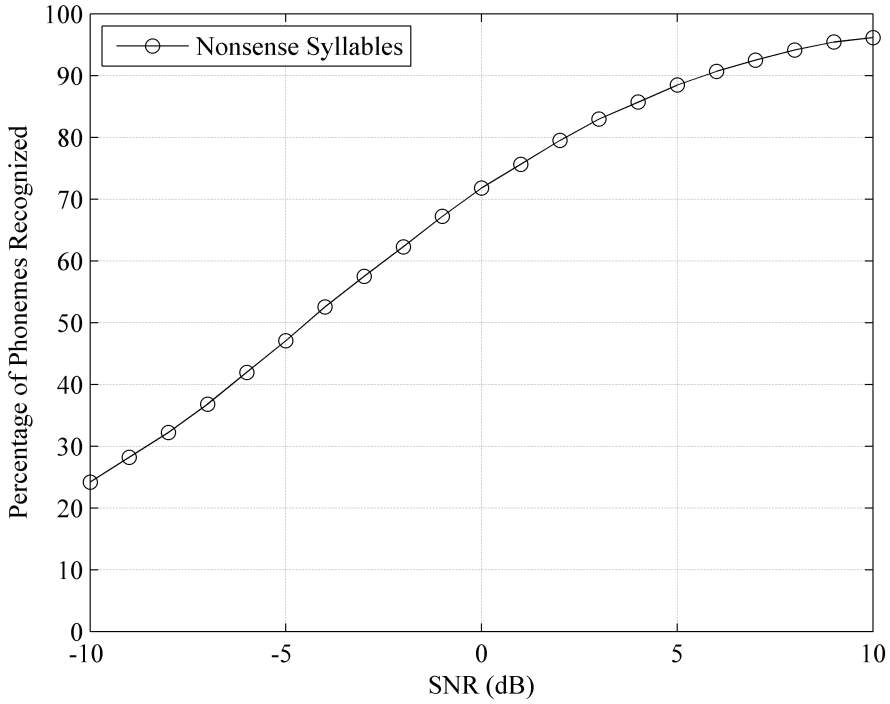


Figure 2.3: Percentage of phonemes correctly recognized by normal listeners for nonsense CVC syllables. Data from Boothroyd & Nittrouer (1988), Figure 2.

CHAPTER 3 METHODS

3.1 Introduction

This chapter describes the various signal-processing techniques that were used for speech enhancement, cue estimation and vowel classification. This section contains a brief overview of the entire algorithm, whereas section 3.2, section 3.3 and section 3.4 will provide more detail regarding the respective signal-processing steps and the motivations for these.

The inputs to perception-prediction models are cues that are estimated from a speech signal, and the lowest SNR at which the various cues can be estimated thus determines the lowest SNR at which a perception prediction model can be used. The accuracy of the various cue estimations degrades with the SNR and thus more vowel confusions (classification errors) are to be expected from the perception model as the SNR decreases. The algorithm proposed in this study has two processing steps, namely speech enhancement or noise suppression and cue estimation. The output of the cue estimation is then used in the perception-prediction model, which is a multivariate Gaussian classifier, proposed by Svirsky (2000). These processing steps are illustrated in Figure 3.1 and each will be explained in detail in the sections that follow, as indicated on the figure. The algorithm assumes no a priori knowledge of the speaker for any of the processing steps.

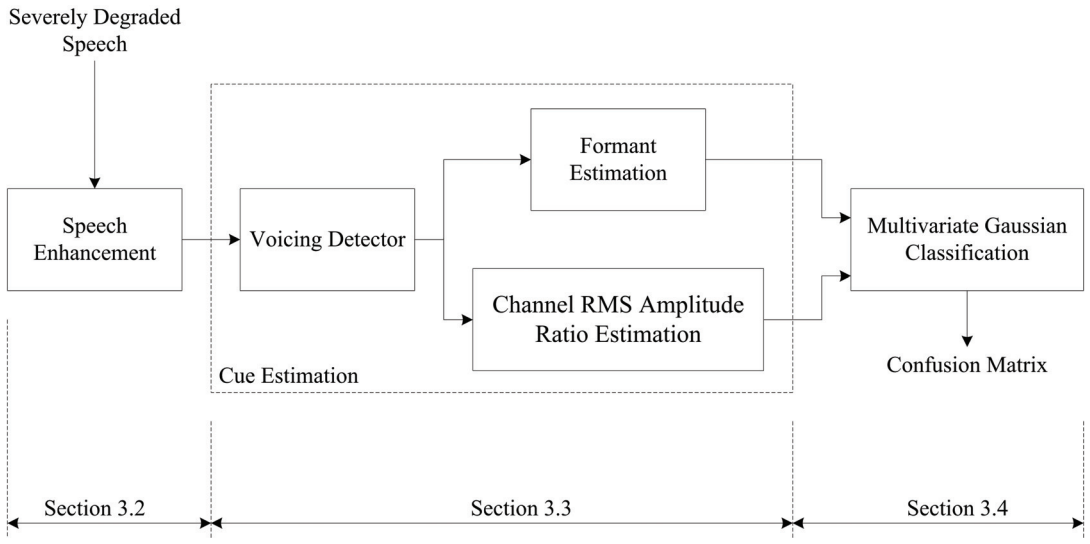


Figure 3.1: Processing steps proposed for perception prediction at SNRs of less than 0 dB. More detail regarding the various processing steps will be given in the respective sections as indicated in the figure.

The speech enhancement is based on a Kalman filter with an auto regressive (AR) model as the internal signal model. The speech enhancement algorithm also uses expectation maximization to suppress unwanted noise further. The voicing detector is used to locate the position of the vowel in the input word. The voicing detector is a CFAR detector, which uses a local estimate of the noise statistics in order to set an adaptive detection threshold. The input to the CFAR detector is the spectrogram of the enhanced speech signal. The spectrogram is a two-dimensional data set with frequency on the vertical axis and time on the horizontal axis (Kopp & Green, 1946). Once the location of the vowel is estimated, the first formant frequency and the channel RMS amplitude ratio estimations are made. The formant frequency estimation is made using an adaptive linear predictive coding (LPC) algorithm. These cues are the inputs to a multivariate Gaussian classification. The output of the classification scheme is a confusion matrix, which can easily be compared to the performance of human listeners when presented with the same inputs as the algorithm. The performance of the signal enhancement is illustrated by investigating the percentage of correct vowel discrimination as a function of SNR. The enhancement of the specific cues used for classification is also investigated using information transmission.

For optimized reading of the various sections of this chapter, the following comments can be considered:

- Section 3.2 discusses the Kalman filter used for the speech-enhancement processing step, and section 3.2.1 presents the results of the performance evaluation of the Kalman filter. Additional information on the techniques used for the Kalman filter performance evaluation is given in 3.2.1.1. The calculation of the LPC coefficients used in the Kalman filter is a critical aspect in the internal calculations done by the Kalman filter. The mathematics regarding the calculation of the LPC coefficients are presented in section 3.2.2. Sections 3.2.1.1 and 3.2.2 can, however, be skipped without loss of continuity if the detail regarding the evaluation for the Kalman filter performance or the LCP coefficients are not of interest.
- Section 3.3 discusses the estimation, specifically voicing detection and the estimation of the selected cues used by the perception prediction model. Section 3.3.1.1 presents the detailed mathematics regarding the design of the voicing detector, and can be skipped without loss of continuity.

3.2 Speech Enhancement

Owing to the degraded nature of the speech signal in the SNR region of interest of this study (-10 dB to -3 dB), some form of signal enhancement is required. Speech enhancement algorithms have attracted a great deal of attention in the past three decades (Gibson, Koo, & Gray, 1991; Hansen & Clements, 1991; Lee & Shirai, 1996; Weinstein et al., 1994), with obvious application in the telecommunications domain. Some speech enhancement algorithms are: the short-time spectral amplitude estimator by Ephraim and Malah (1984) on which they improved with the log spectral amplitude estimator (Ephraim & Malah, 1985), the HMM-based speech enhancement algorithms suggested by Ephraim et al. (1989), the spectral subtraction algorithm suggested by Boll (1979), the Wiener-EM (Expectation Maximization) algorithm (Lim & Oppenheim, 1978), a Kalman filter for speech enhancement when only the degraded speech signal is available for processing by Paliwal and Basu (1987), and the Kalman-EM-iterative and Kalman-gradient-decent-sequential algorithm by Gannot et al. (1998). The use of a model for the analysis of the degraded speech input is based on the work of Wolpert *et al.* (1995). Wolpert argued for the existence of internal models in the central nervous system (Perkell et al., 1997; Sabes, 2000). The decision to use a Kalman filter is based on the work of Watkins and Paus (2004) because the structure of a Kalman filter uses an internal model of the input stimulus to produce an estimate of the input stimulus. Based on the work of Wolpert, it is assumed that biological systems analyse inputs by using these internal models. Conclusions regarding the inputs are made in a process of analysis-by-synthesis. Watkins and Paus (2004) argued for the existence of a link between the biological auditory and speech generation systems. This work suggests that in order to arrive at some perception of an acoustic stimulus, all the steps to reproduce (synthesize) the sound physically are taken, except for the final step of vocalizing the sound. For the auditory system this would imply that an internal model that can produce (synthesize) speech is used to analyse the input speech stimulus in the auditory system.

The structure of a Kalman filter (Kalman & Bucy, 1961) is shown in Figure 3.2, which is based on the work of Bozic (1979). From this figure it can be seen that the structure of the Kalman filter suits the analysis-by-synthesis approach, which Watkins and Paus (2004) suggested exists in the auditory system. The output from the "system parameter" block is a predicted value of the present estimate (filter output) without any additional information. This predicted value is based on the internal model of the input signal and previous estimates.

The type of model used as the internal model will be chosen a priori but the coefficients which govern the model transfer function will be determined adaptively. The output of the "measurement parameter" block is an estimate of the present measurement, which is used to calculate a correction term. In this approach the input signal is estimated/analyzed by synthesizing it and using this synthesized signal to minimize the error between the synthesized signal and the input stimulus.

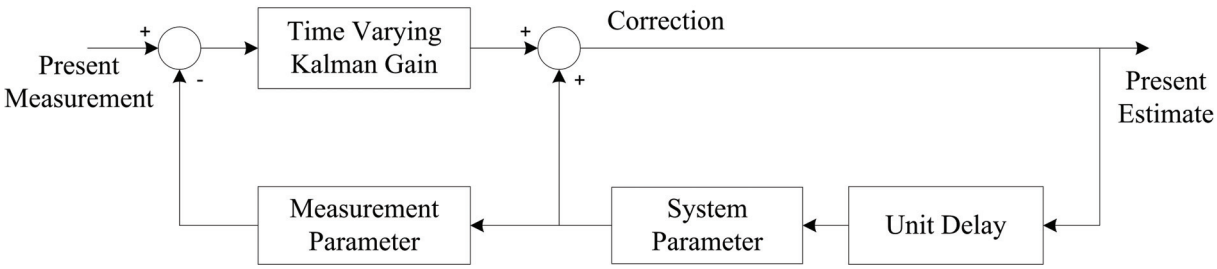


Figure 3.2: Conceptual model for Kalman filter.

The "system parameter" in Figure 3.2 represents the internal model of the system, which is a model of the input stimulus. The complete system in Figure 3.2 represents the speech enhancement model. The implemented Kalman filter speech enhancement algorithm uses a simplified EM approach proposed by Du and Driessen (1991). Assuming that the speech is stationary on a short-time basis and embedded in AWGN, Lim and Oppenheim (1978) suggested using a AR model to model a speech signal. This AR model was used as the internal speech synthesis model of the Kalman filter. LPC can be used to determine the AR model coefficients. The LPC coefficients are determined by using the autocorrelation method as described in Rabiner and Schafer (1978). During the EM the LPC coefficients are iteratively estimated using the previous Kalman filter output. For example, in performing one EM iteration, a second estimation of the LPC coefficients is made using the output of the Kalman filter. The newly estimated LPC coefficients are then used to filter the input signal again. However, because the estimate of the noise is not improved during successive iterations, this is only a partial EM algorithm of which convergence is not guaranteed. The observation noise variance is calculated using a section of the input signal where only noise is present, whereas the excitation noise variance is a byproduct of the LPC analysis procedure (Du & Driessen, 1991).

A Kalman filter (Anderson & Moore, 1979) with a linear AR model as the internal model is used for speech enhancement in this study. LPC is used for the AR model (Makhoul, 1975; Rabiner & Schafer, 1978). The LPC coefficients of a forward linear predictor are determined by minimizing the prediction error in the least squares sense. In order to use the linear model for the speech signal, the speech signal is segmented into short non-overlapping concatenated portions. The length of the segment is 50 ms. On a short time basis the speech signal is assumed to be represented by the following difference equation (Du & Driessen, 1991):

$$s(n) = \sum_{k=1}^{M} a_k s(n-k) + u(k) \tag{3.1}$$

with $s(n)$ the speech signal, $u(k)$ the input white noise (excitation noise), M the order of the LPC model, a_k the LPC coefficients and $s(n-k)$ the k^{th} previous output speech sample. The LPC coefficients are estimated using the autocorrelation method while windowing the input. The coefficients are solved using the Levinson-Durbin algorithm. The LPC coefficient calculation will be described in the following section. In vector-matrix notation (3.1) can be written as:

$$s(n) = \Phi s(n-1) + \Gamma u(n) \tag{3.2}$$

with:

$$s(n) = [s(n-M+1), \ldots, s(n-1), s(n)]^T \tag{3.3}$$

$$\Phi = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_M & a_{M-1} & a_{M-2} & \cdots & a_1 \end{bmatrix} \tag{3.4}$$

$$\Gamma = [0,...,0,1]^T \tag{3.5}$$

where $s(n) \in R^{M\times 1}$, $\Phi \in R^{M\times M}$ and $\Gamma \in R^{M\times 1}$. The signal model (internal model) of the Kalman filter is denoted by (3.4). The observation is assumed to be a speech signal corrupted by AWGN. The observation model of the Kalman filter is given by:

$$x(n) = s(n) + n(n) \tag{3.6}$$

with $x(n)$ the observation (measurement) of the speech signal and $n(n)$ the observation noise. In vector-matrix notation (3.6) can be written as:

$$x(n) = Hs(n) + n(n) \tag{3.7}$$

with:

$$H = [0,...,0,1] \tag{3.8}$$

where $H \in R^{1\times M}$. The excitation noise, $u(n)$, and the observation noise, $n(n)$, are assumed to be uncorrelated, zero mean white Gaussian noise, and the observation noise is assumed to be uncorrelated to the speech signal. The statistical properties of the noise can be expressed as:

$$E[u(n)] = 0 \tag{3.9}$$

$$E[n(n)] = 0 \tag{3.10}$$

$$E[u(n)u(m)] = \sigma_u{}^2 \delta_{mn} \tag{3.11}$$

$$E[n(n)n(m)] = \sigma_n{}^2 \delta_{mn} \tag{3.12}$$

$$E[u(n)n(m)] = 0 \tag{3.13}$$

$$E[s(n)n(m)] = 0 \tag{3.14}$$

with $\sigma_u{}^2$ the excitation noise variance and $\sigma_n{}^2$ the observation noise variance. The Kalman filter recursive algorithm is given by the following equations:

$$s(n \mid n) = \Phi s(n-1 \mid n-1) + K(n)[x(n) - H\Phi s(n-1 \mid n-1)] \tag{3.15}$$

$$K(n) = V_s(n \mid n-1)H^T \left[HV_s(n \mid n-1)H^T + \sigma_n{}^2\right]^{-1} \tag{3.16}$$

$$V_s(n \mid n) = [I - K(n)H]V_s(n \mid n-1) \tag{3.17}$$

$$V_S(n \mid n-1) = \Phi V_s(n-1 \mid n-1)\Phi^T + \Gamma \sigma_u{}^2 \Gamma^T \tag{3.18}$$

with $V_s(n \mid n)$ the error covariance matrix of the estimate of $s(n \mid n)$, $V_s(n \mid n-1)$ the error covariance matrix with respect to the one-step prediction $s(n \mid n-1)$, K the Kalman gain and I an identity matrix with $I \in R^{M \times M}$. The Kalman filter is initialized with:

$$s(0) = E[s(n)] \tag{3.19}$$

$$V_s(0) = E\left[(s(n) - s(0))(s(n) - s(0))^T\right] \tag{3.20}$$

In the subsequent speech segments the parameters are initialized with the values obtained in the previous segment. The variance of the excitation noise is a by-product of the LPC analysis.

The variance of the observation noise can be estimated by finding a portion of the input signal with no speech present. This portion is then used to estimate the variance of the observation noise only. EM can be applied to each of the speech segments. This process entails a second estimation of the LPC coefficients using the output of the Kalman filter for the current segment of speech. The current segment of speech is then used as the input to another Kalman filter stage using the newly estimated LPC coefficients. The noise variances are not estimated again for this filtering stage. The noise variances are thus fixed while the LPC coefficient estimations are improved iteratively. This process can be repeated but convergence is not guaranteed. Figure 3.3 shows a functional block diagram for the Kalman filter processing chain.

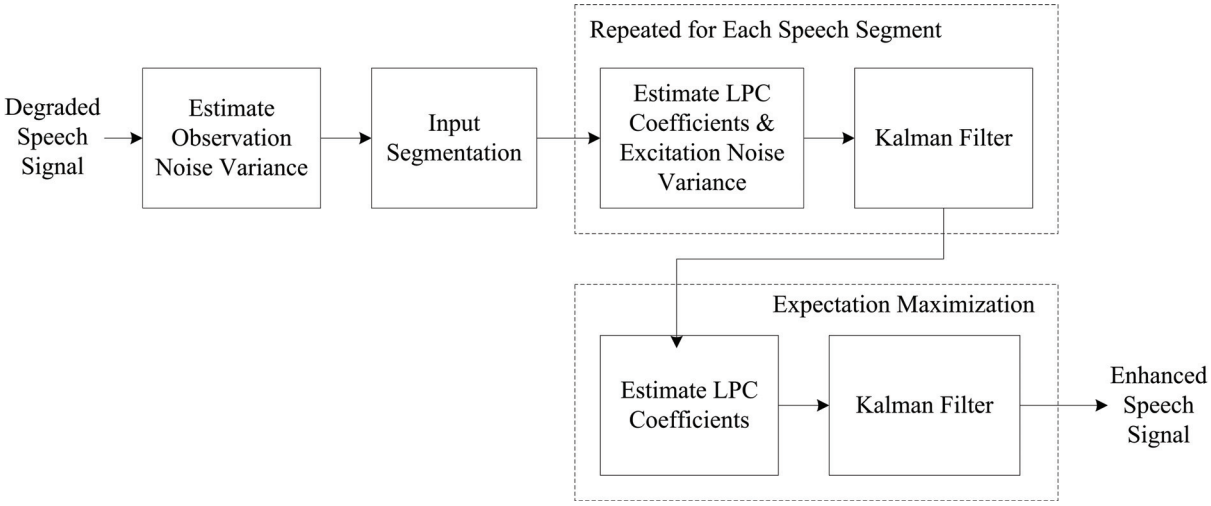


Figure 3.3: Functional block diagram for the Kalman filter processing chain.

3.2.1 Evaluation of Kalman Filter Performance

In order to evaluate the SNR improvement quantitatively, in low SNR (-10 dB to -3 dB) conditions, a fixed and repeatable source was required as input to the filter. SNR improvement is defined as the increase in SNR between the signal at the input of the filter and the signal at the output of the filter. For example, if the signal at the input of the filter had an SNR of -3 dB and the signal at the output of the filter had an SNR of 5 dB, the SNR improvement would be 8 dB. To evaluate the SNR improvement quantitatively a vowel synthesizer was developed, which allowed for full control of formant frequency placement. The vowels were generated with a synthesizer as described by Klatt (1980). The synthesizer implementation is described in Section 3.2.1.1. The Kalman filter gain for low SNR was determined using a synthesized vowel with first and second formant frequencies at 750 Hz and 1100 Hz respectively. Another

parameter of interest was the number of expectation maximization repetitions in order to achieve maximum noise suppression from the filter. The number of EM repetitions to be used in the final algorithm would be determined from these performance simulations. The total output SNR is given by (Gannot, Burshtein, & Weinstein, 1998):

$$SNR = \frac{\sum_t s^2(t)}{\sum_t (s(t) - \hat{s}(t))^2} \tag{3.21}$$

with $s(t)$ the input signal and $\hat{s}(t)$ the Kalman filter estimate of the signal. The time summation is over the entire duration of the signal. Figure 3.4 shows the output SNR of the filter for an input SNR range from -10 dB to 10 dB. The figure also shows the difference in filter performance for different numbers of expectation maximization loops. Figure 3.5 shows the SNR improvement for the same performance evaluation. From Figure 3.4 and Figure 3.5 it can be seen that for none of the number of EM irritations the SNR gain was maximized for the entire SNR region of -10 dB to 10 dB. One EM iteration (green line with diamond marker) did, however, provide maximum SNR improvement for a large proportion of the SNR region of interest (from -6 dB to 10 dB).

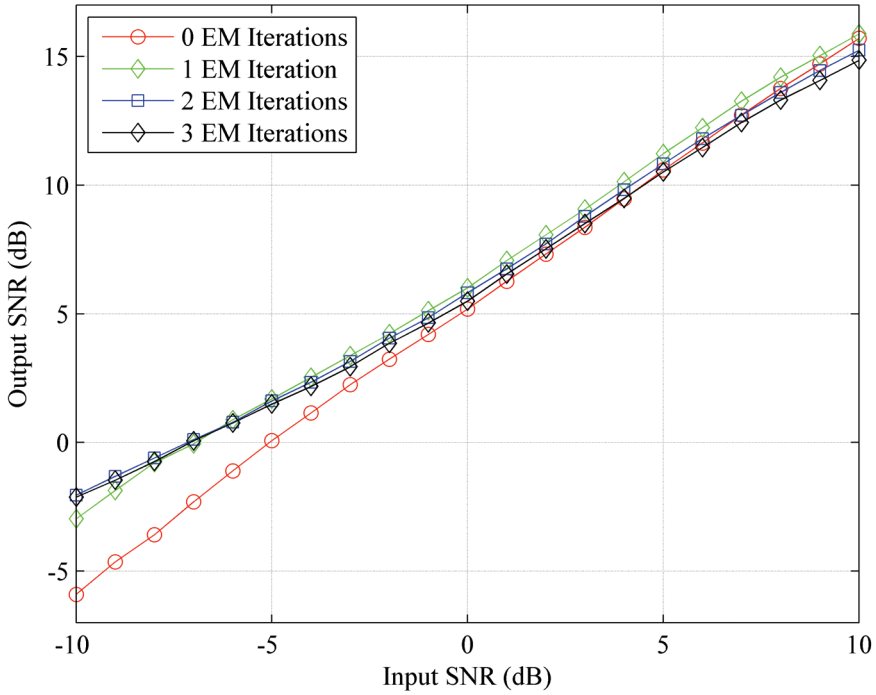


Figure 3.4: Output SNR of the Kalman filter for an input SNR range of -10 dB to 10 dB. EM for no iterations to four iterations.

To illustrate the performance of the Kalman filter, a spectrogram of the syllable 'pAt' is shown in Figure 3.6. This signal shown in Figure 3.6 has no noise added and serves as a reference. Figure 3.7 shows the signal with 5 dB SNR and also -5 dB SNR. Both these signals are used as the input to the Kalman filter and the output of the filter for both signals are shown in Figure 3.8. From this sequence of figures, it is clear that the Kalman filter suppressed the noise while maintaining the characteristics of the speech signal. For example, when comparing Figure 3.7 (b) to Figure 3.8 (b) the duration of the vowel (from roughly 0.15 seconds to 0.28 seconds) is clearly visible in Figure 3.8 (b), but in Figure 3.7 (b) the duration is very difficult to discern. Also, the second formant frequency can be identified in Figure 3.8 (b) at roughly 1900 Hz, whereas it is completely masked by the noise in Figure 3.7 (b).

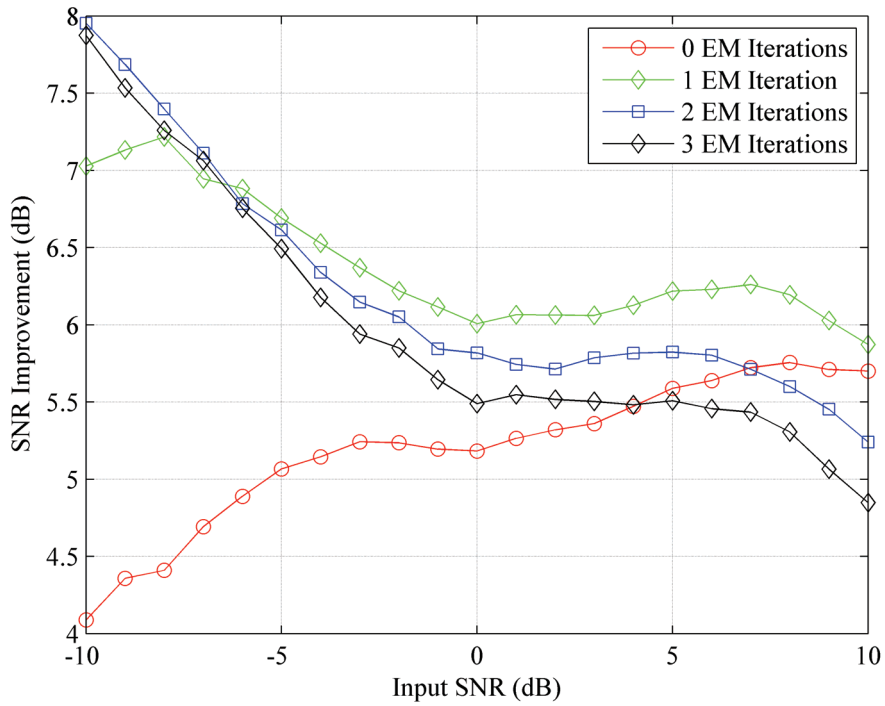


Figure 3.5: SNR improvement of the Kalman filter for an input SNR range of -10 dB to 10 dB. EM for no repetitions to four repetitions.

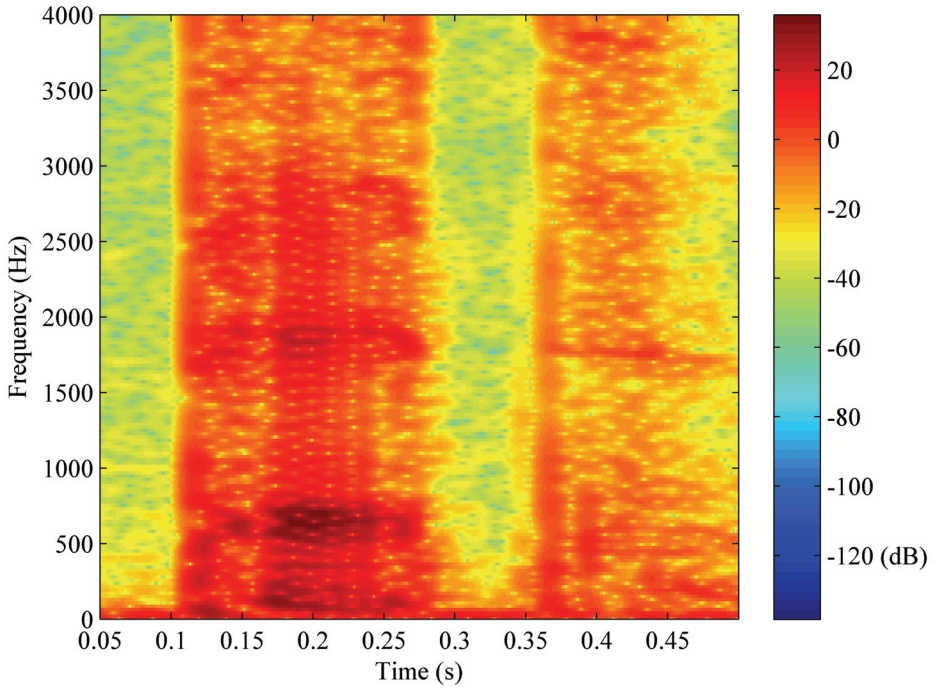
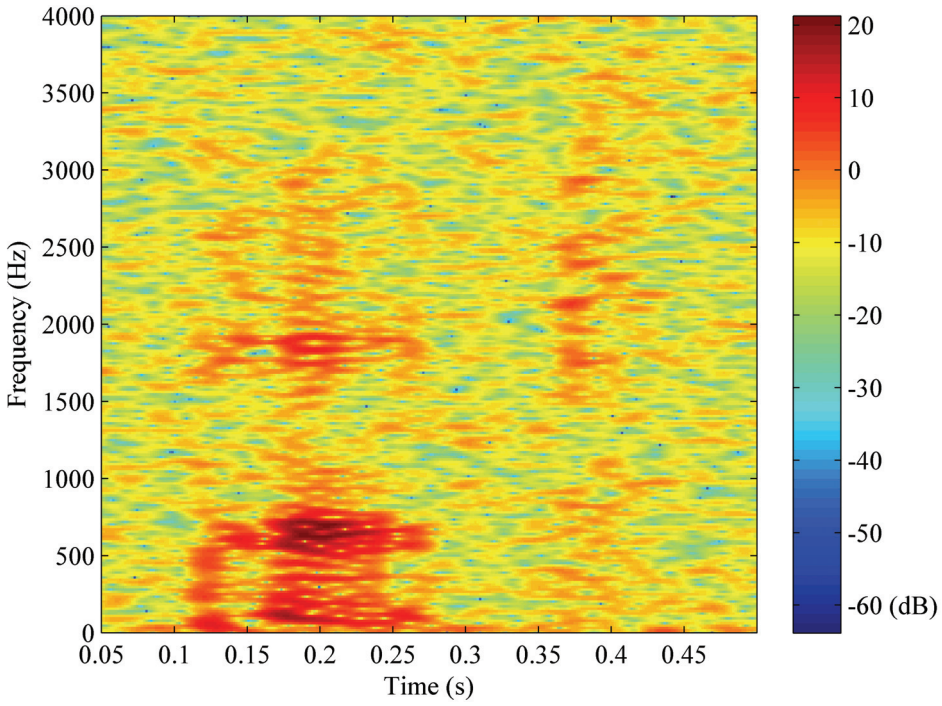
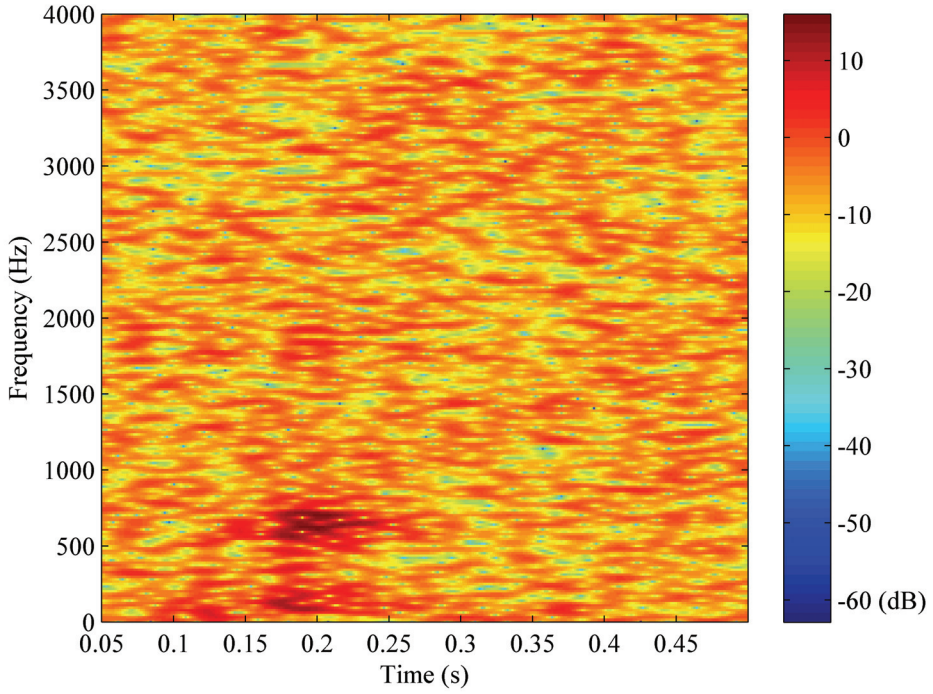


Figure 3.6: Spectrogram of the syllable 'pAt' with no noise added to the signal.

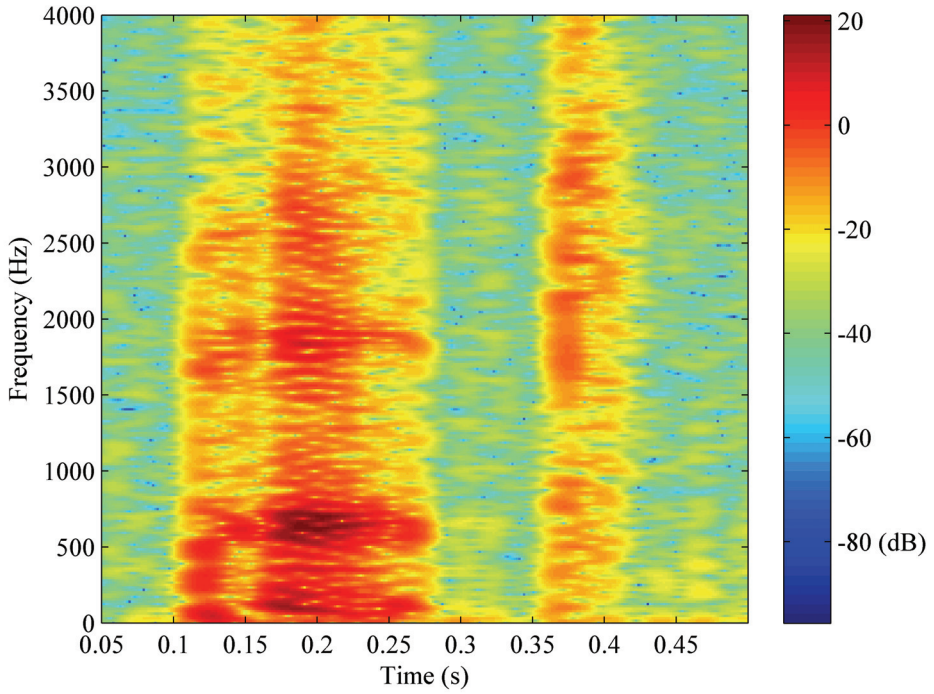


(a)

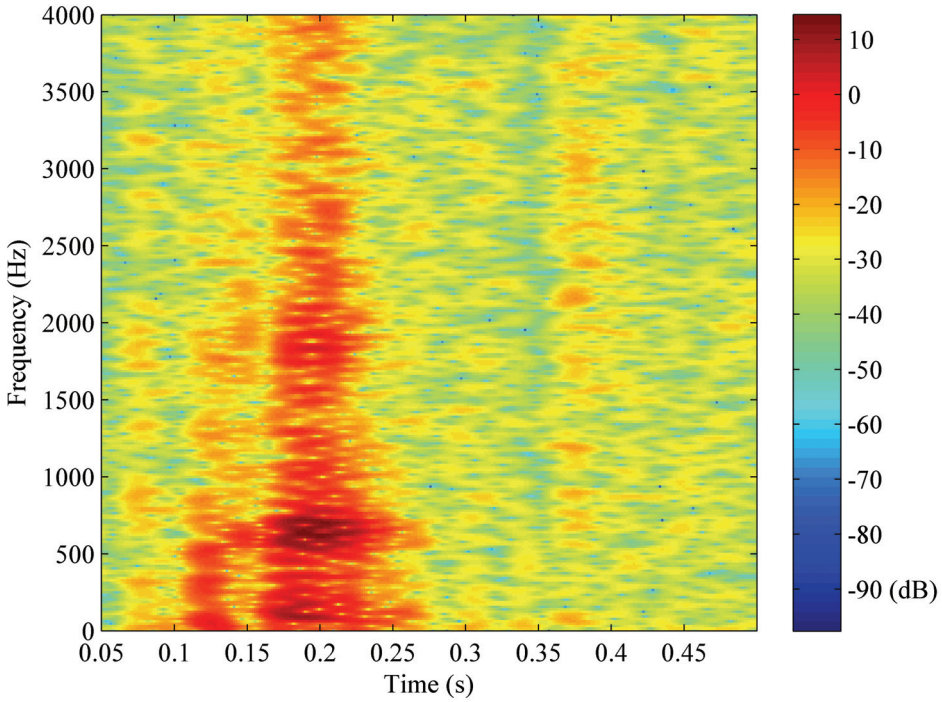


(b)

Figure 3.7: Spectrogram of the syllable 'pAt' before the Kalman filter is applied to the signal, with (a) 5 dB SNR and (b) -5 dB SNR.



(a)



(b)

Figure 3.8: Spectrogram of the syllable 'pAt' after the Kalman filter is applied to the signal, with (a) 5 dB SNR and (b) -5 dB SNR.

3.2.1.1 Formant Synthesizer

A formant is a peak in an acoustic frequency spectrum that results from the resonant frequencies of the vocal tract. The formant synthesizer was developed in order to allow for the software generation of vowel sounds. The controlled and repetitive manner in which these sounds can be generated using the synthesizer makes it ideal for the testing and evaluation of the Kalman filter. The formant synthesizer that was implemented is based on a cascade implementation by Klatt (1980). A block diagram of this implementation is shown in Figure 3.9.

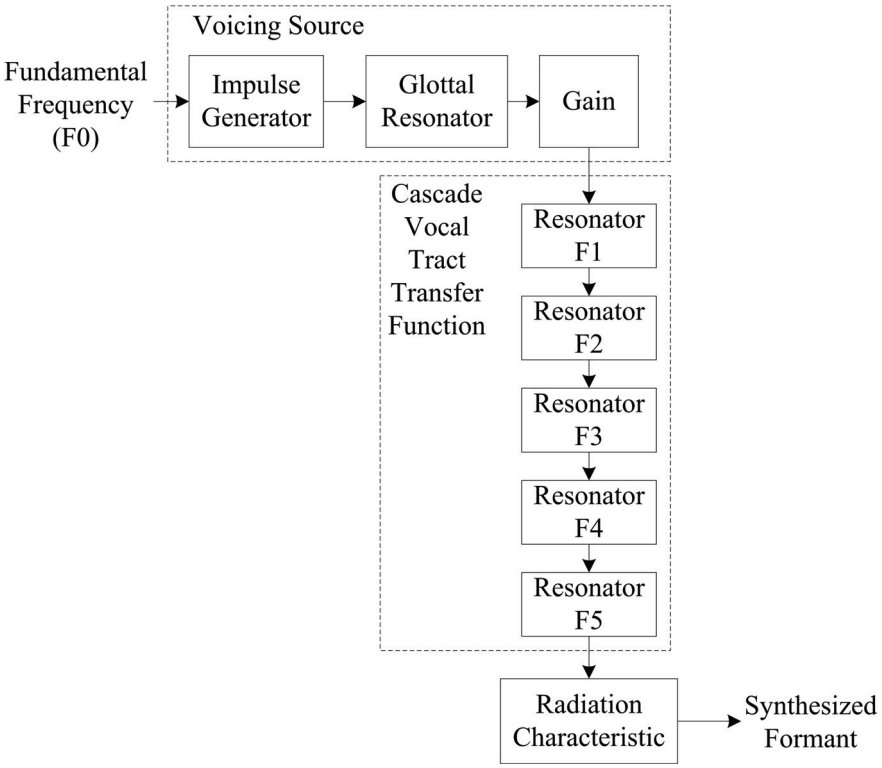


Figure 3.9: Functional block diagram of cascade formant synthesizer.

3.2.1.1.1 Voicing Source

The structure of the voicing source is shown at the top of Figure 3.9. The parameters, which influence the output of the voicing source, are the fundamental frequency (F0) and the gain. The fundamental frequency is specified in hertz, thus a fundamental frequency of 100 Hz would produce a unit impulse train of 100 Hz. The gain can vary from 60 dB for a strong vowel sound to 0 dB when the voicing source is turned off. The function of the glottal resonator is to produce a smoothed waveform that resembles a typical glottal volume velocity waveform. The transfer function of the glottal resonator is:

$$T(f) = \frac{Az^2}{z^2 - Bz - C} \tag{3.22}$$

with:

$$A = 1 - C - B \tag{3.23}$$

$$B = 2\exp(-\pi BWt)\cos(\pi f_R t) \tag{3.24}$$

$$C = -\exp(-2\pi BWt) \tag{3.25}$$

$$z = \exp(j2\pi f_R t) \tag{3.26}$$

and BW the transfer function bandwidth [Hz] and f_R the resonance frequency [Hz]. For the implemented formant synthesizer the bandwidth (BW) was set to 500 Hz and the resonance frequency at 100 Hz. The transfer function of the glottal resonator is shown in Figure 3.10.

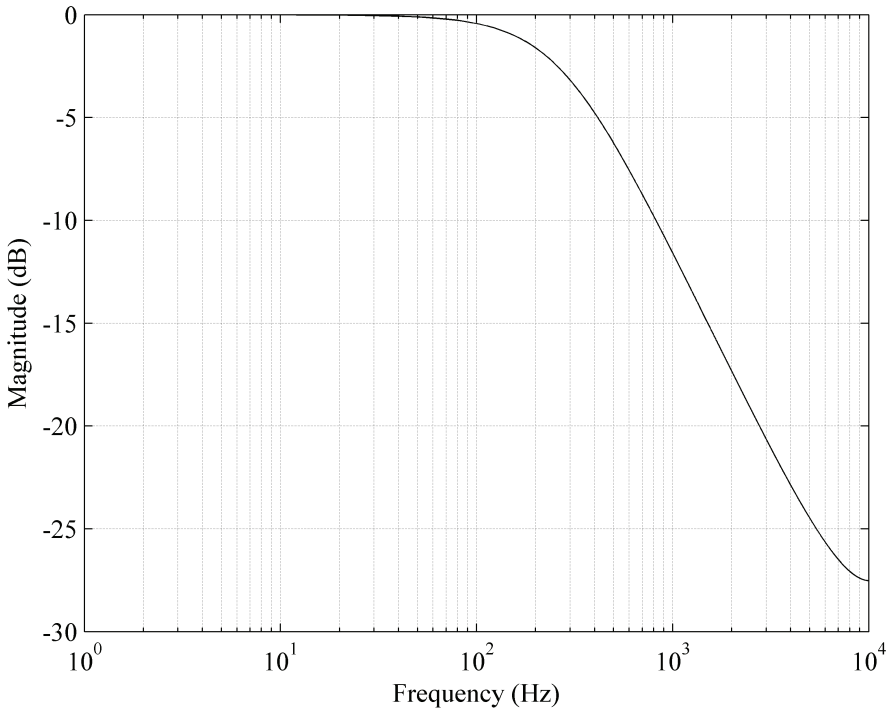


Figure 3.10: Glottal resonator transfer function.

The output of the voicing source with the fundamental frequency (F0) set to 150 Hz and the gain set to 10 dB is shown in Figure 3.11.

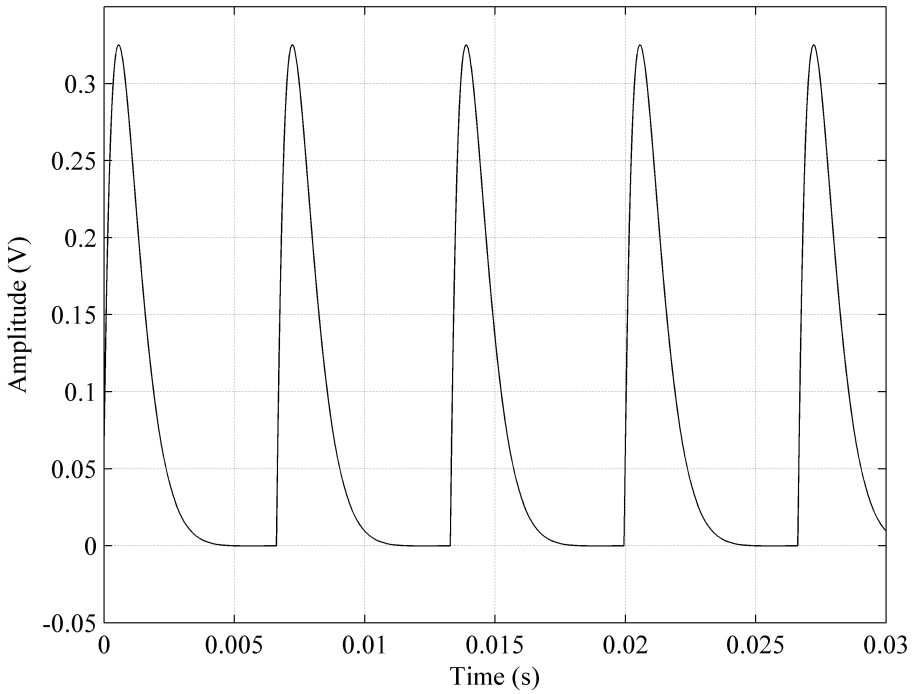


Figure 3.11: Voicing source output.

3.2.1.1.2 Cascade Vocal Tract Transfer Function

The vocal tract transfer function is implemented by cascading multiple resonators (Klatt, 1980). Each of the resonators implements one formant frequency. The transfer functions of the formant resonators are also given by equations (3.29) to (3.33). The formant frequencies (F1 to F5) and BWs for various 'standard' vowels are given in Table 3.1 and Table 3.2 (Gold & Rabiner, 1968). The magnitude of the contribution of each of the formant frequencies decreases sequentially from F1 to F5. The frequencies of the lowest three formants can vary significantly with changes in articulation. The frequencies and BWs of F4 and F5 can be held constant with little decrement in output sound quality (Klatt, 1980). The fourth and fifth formants are included in the cascade of resonators to shape the overall spectrum, but otherwise contribute little to the intelligibility of vowels.

Table 3.1: Formant frequencies for vowels.

Typewritten Symbol for Vowel	Typical Word	F1 (Hz)	F2 (Hz)	F3 (Hz)
IY	Beet	270	2290	3010
I	Bit	390	1990	2550
E	Bet	530	1840	2480
AE	Bat	660	1720	2410
UH	But	520	1190	2390
A	Hot	730	1090	2440
OW	Bought	570	840	2410
U	Foot	440	1020	2240
OO	Boot	30	870	2240
ER	Bird	490	1350	1690

Table 3.2: Formant bandwidths for vowels.

Resonator	Centre Frequency (Hz)	Bandwidth (Hz)	Q
F1	Variable	60	Variable
F2	Variable	100	Variable
F3	Variable	120	Variable
F4	3500	175	20
F5	4500	281	16

Figure 3.12 shows the transfer functions of the cascaded vocal tract for the vowels IY, A , and OO. In Figure 3.12 the various formant frequencies are clearly visible.

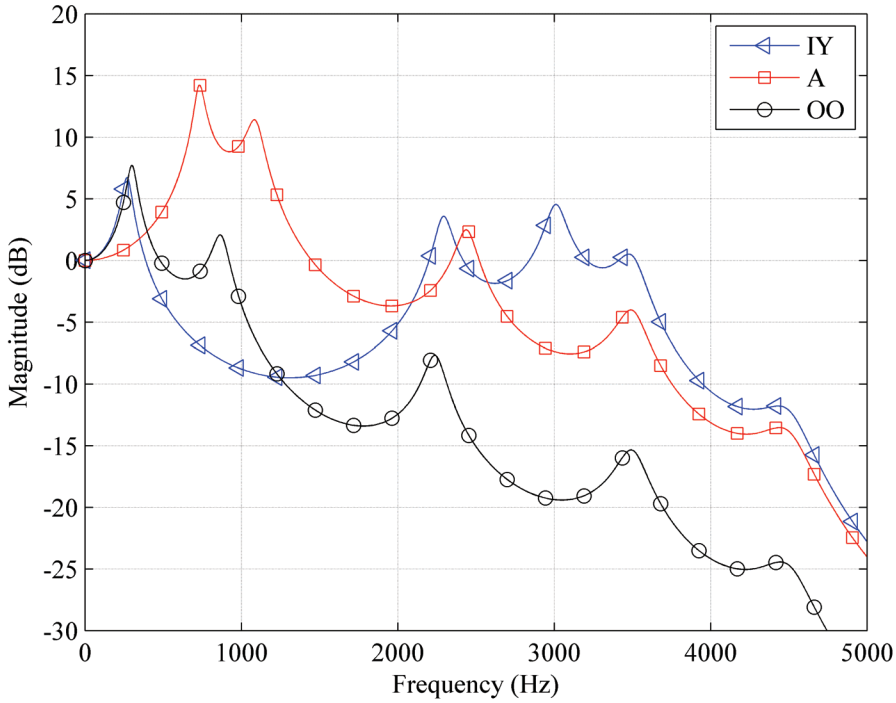


Figure 3.12: Transfer functions for the 5 formant resonators for the vowels IY, A and OO.

3.2.1.1.3 Radiation Characteristic

The radiation characteristic models the effect of directivity patterns of sound radiating from the head as a function of frequency (Klatt, 1980). The transfer function of the radiation characteristic is given by:

$$p(nT) = u(nT) - u((n-1)T) \tag{3.27}$$

with $u(nT)$ the current output of the cascade vocal tract transfer function and $u((n-1)T)$ the previous output of the cascade vocal tract transfer function. Figure 3.13 shows the transfer function of the radiation characteristic, which is the final processing step in the cascade formant synthesiser (refer to Figure 3.9).

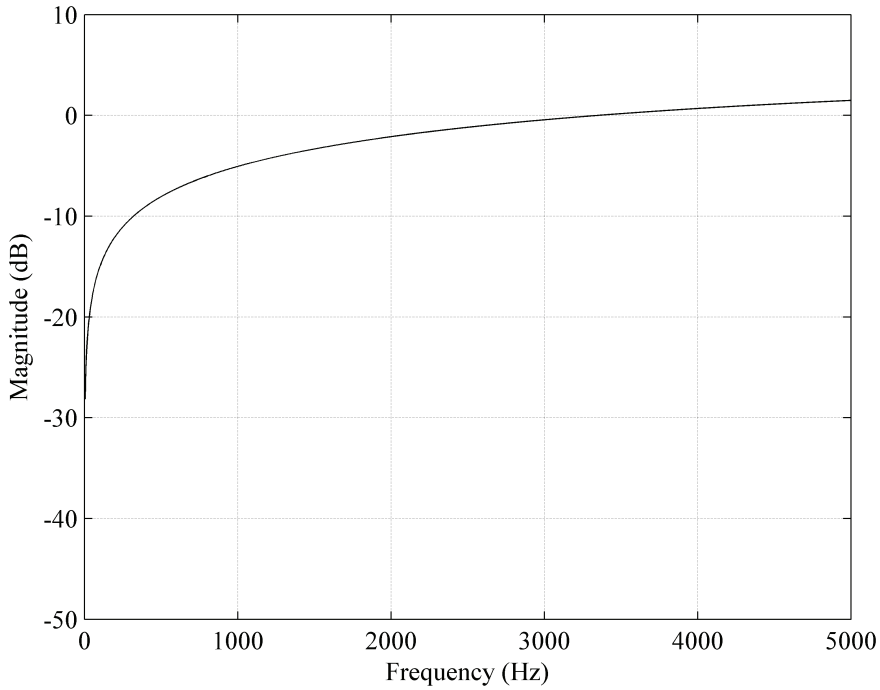


Figure 3.13: Transfer function of the radiation characteristic.

3.2.2 LPC Coefficient Calculation

LPC analysis is an accepted technique in estimating speech parameters such as pitch, formant frequencies, spectra, and vocal tract area functions, as well as representing speech for low bit rate transmission or storage (Rabiner & Schafer, 1978). In using LPC, a single speech sample can be approximated as a linear combination of past speech samples. By analyzing a finite interval, a unique set of predictor coefficients can be obtained by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones. These coefficients are the weighting coefficients used in the linear prediction. Once the predictor coefficients of a system are known, the system can be uniquely identified to the extent that it can be modelled as an all-pole linear system. A linear predictor is defined as a system whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_n s(n-k) \tag{3.28}$$

with n the sample index, p the order of the linear predictor, $\tilde{s}(n)$ the estimation for sample n, $s(n-k)$ the k^{th} previous sample and α_n the prediction coefficients. The prediction error, $e(n)$, is defined as

$$e(n) = s(n) - \tilde{s}(n) \tag{3.29}$$

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_n s(n-k) \tag{3.30}$$

The aim of linear prediction is to determine a set of predictor coefficients directly from the speech signal in order to obtain an estimate of the spectral properties of the input signal. Owing to the locally stationary nature of a speech signal (Makhoul, 1975) the predictor coefficients must be estimated for short segments of the input speech signal. The short-time average prediction error is defined as

$$E_n = \sum_m e_n{}^2(m) \tag{3.31}$$

$$E_n = \sum_m (s_n(m) - \tilde{s}_n(m))^2 \tag{3.32}$$

$$E_n = \sum_m \left(s_n(m) - \sum_{k=1}^{p} \alpha_n s_n(m-k) \right)^2 \tag{3.33}$$

where $s_n(m)$ is a segment of speech that has been selected in the vicinity of sample n. The values of α_n which minimize E_n can be calculated by setting

$$\frac{\partial E_n}{\partial \alpha_i} = 0 \tag{3.34}$$

for i=1,2,…. p. This results in

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{\alpha}_k \sum_m s_n(m-i)s_n(m-k) \tag{3.35}$$

for $1 \le i \le p$, where $\hat{\alpha}_k$ are the values of α_k which minimize E_n. Since $\hat{\alpha}_k$ is unique, the notation of the caret will be dropped to denote the values of α_k which minimize E_n. By defining

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \tag{3.36}$$

Eq. (3.35) can be written as

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi(i,0) \qquad i=1, 2,\ldots. \; p \tag{3.37}$$

This set of p equations in p unknowns can be solved for the unknown predictor coefficients α_k that minimize the average squared prediction error for the segment $s_n(m)$. This is done by first computing $\phi_n(i,k)$ for $1 \le i \le p$ and $0 \le k \le p$ and then solving Eq.(3.37) for α_k. In order to establish the limits on the sums of Eqs. (3.31)-(3.33) and (3.35) it can be assumed that the waveform segment, $s_n(m)$, is zero outside the interval $0 \le m \le N-1$. Using this assumption $s_n(m)$ can be expressed as

$$s_n(m) = s(m+n)w(n) \tag{3.38}$$

with $w(m)$ a finite length window (e.g. Hamming window) which is zero outside the interval $0 \le m \le N-1$. Thus, if $s_n(m)$ is nonzero for the interval $0 \le m \le N-1$, the corresponding predictor error, $e_n(m)$, for a p^{th} order linear predictor will be nonzero for the interval $0 \le m \le N-1+p$. Given the windowing assumption Eq. (3.31) can be written as

$$E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \tag{3.39}$$

Given these limits Eq. (3.36) can be written as

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k) \tag{3.40}$$

for $1 \le i \le p$ and $0 \le k \le p$, which can also be expressed as

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k) \tag{3.41}$$

for $1 \le i \le p$ and $0 \le k \le p$. This formulation allows for the use of the short-time autocorrelation function of $s_n(m)$ evaluated for $(i-k)$. This short-time autocorrelation is given by

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k) \tag{3.42}$$

Given that $R_n(k)$ is an even function, it follows that

$$\phi_n(i,k) = R_n(|i-k|) \tag{3.43}$$

for $1 \le i \le p$ and $0 \le k \le p$. Therefore Eq. (3.37) can be expressed as

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \tag{3.44}$$

for $1 \le i \le p$. The set of equations given by Eq. (3.44) can be written in matrix form as

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \dots \\ \dots \\ R_n(p) \end{bmatrix} \tag{3.45}$$

This $p \times p$ matrix of autocorrelation values (R_n) is a Toeplitz matrix (symmetrical and all elements on a given diagonal are equal). Equation (3.45) can be solved using the Levinson-Durbin recursive solution as shown below (calculated for a segment of speech that has been selected in the vicinity of sample n, subscript omitted on the autocorrelation function):

$$E^{(0)} = R(0) \tag{3.46}$$

$$k_i = \frac{\left(R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-j)} R(i-j) \right)}{E^{(i-1)}} \qquad 1 \le i \le p \tag{3.47}$$

$$\alpha_i^{(i)} = k_i \tag{3.48}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \tag{3.49}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \tag{3.50}$$

Equations (3.47) to (3.50) are solved recursively for $i = 1,2,...,p$, with the final solution given by

$$\alpha_j = \alpha_j^{p} \qquad 1 \le j \le p \tag{3.51}$$

It should also be noted that $E^{(i)}$ in Eq. (3.50) is the prediction error for a prediction of order i.

Section 3.2 described the techniques used for speech enhancement and the methods used for the evaluation of the performance of these techniques. A Kalman filter is the core of the speech-enhancement technique, with LPC and EM as additional processing techniques required to complete the speech-enhancement implementation. The next section will discuss the techniques used for speech cue estimation. The input to the cue estimation techniques is the output from the speech-enhancement techniques.

3.3 Cue Estimation

As discussed in the introduction (chapter 1) and the literature study (chapter 2), this study uses the same classification method as that proposed by Svirsky (2000). The cues used by the MPI model proposed by Svirsky (2000) use the first formant frequency and the amplitude ratio of four overlapping bandpass filters (refer to Figure 2.2) as cues. The filters have crossover frequencies of 700 Hz, 1.4 kHz, and 2.3 kHz and roll-off slopes of 12 dB per octave. The three cues obtained from this processing step are the RMS amplitude ratio (in decibel) of the first channel to the second, third and fourth channels. These cues are estimated for a vowel as part of a syllable and thus the portion of the syllable containing the vowel has

to be identified. To do this, a voicing detector is used. The processing to estimate the vowel location, the first formant frequency and the RMS amplitude ratios are discussed in the following sections. As illustration of the algorithm performance, examples will be given for SNRs of 5 dB and -5dB (see section 3.3.2).

3.3.1 Voicing Detection

Owing to the low SNR conditions and because no a priori knowledge regarding the speaker or the external noise source is assumed, a noise-estimating adaptive threshold detector is suggested. This adaptive threshold detector is suggested as opposed to the voicing detector presented by Mustafa and Bruce (2006), which uses the log ratio of the low-frequency to high-frequency energy of the speech input, as input to a fixed threshold detector with hysteresis. For this study an estimate of the noise statistics is continuously made on the input signal in order to set the detection threshold so as to achieve a constant false alarm rate (CFAR). A CFAR detector is often used in radar signal processing (Gandhi & Kassam, 1988). The detection algorithm is applied to the spectrogram of the input CVC syllable. The output of the detector is an estimation of the duration of the vowel and thus the voiced section of the CVC syllable. The steps in estimating the duration of the vowel are shown in Figure 3.14. The design of the CFAR detector is discussed in detail in section 3.3.1.1. Figure 3.15 illustrates the two-dimensional process of determining the threshold for each test cell in the spectrogram. The shape of the reference cells is selected to search specifically for horizontal lines in the spectrogram, which are typically the shape of a formant persisting for some time.

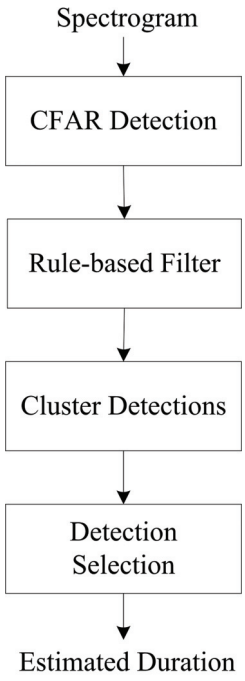


Figure 3.14: Processing steps in estimating vowel duration. The input to the algorithm is the spectrogram of the input CVC syllable.

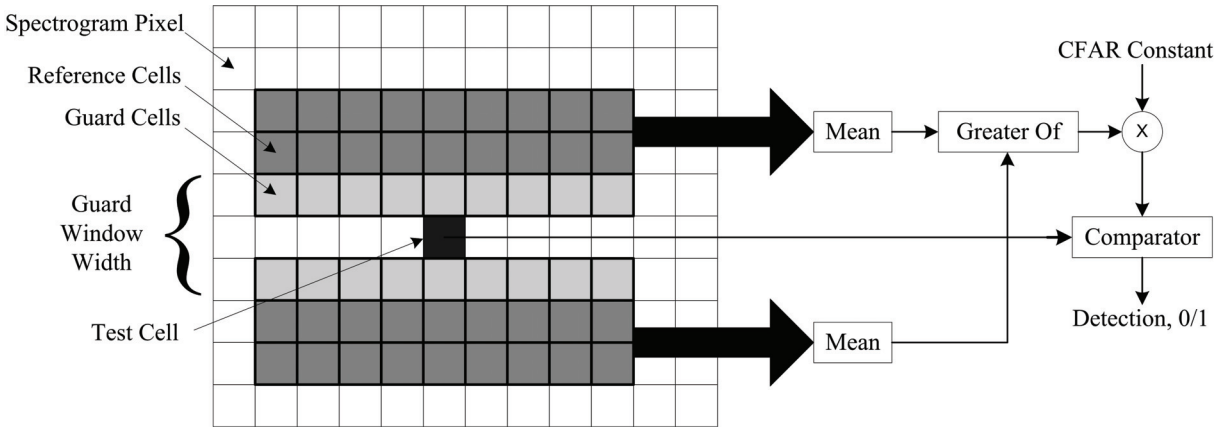


Figure 3.15: Illustration of adaptive threshold calculation for spectrogram detection. The reference cells are used to estimate the noise statistics surrounding the test cell, and the guard cells are required to ensure that the test cell itself does not corrupt the estimation of the noise in the reference cells.

To illustrate the temporal nature of the voiced frequency components, Figure 3.16 shows an example spectrogram for the syllable ‘pAAt’. The spectrogram shows the first formant frequency at roughly 700 Hz between 0.16 seconds and 0.37 seconds and the pitch frequencies

at roughly 200 Hz. Note the broad frequency structure of the formant frequency components compared to the narrow span in frequency of the pitch frequency components.

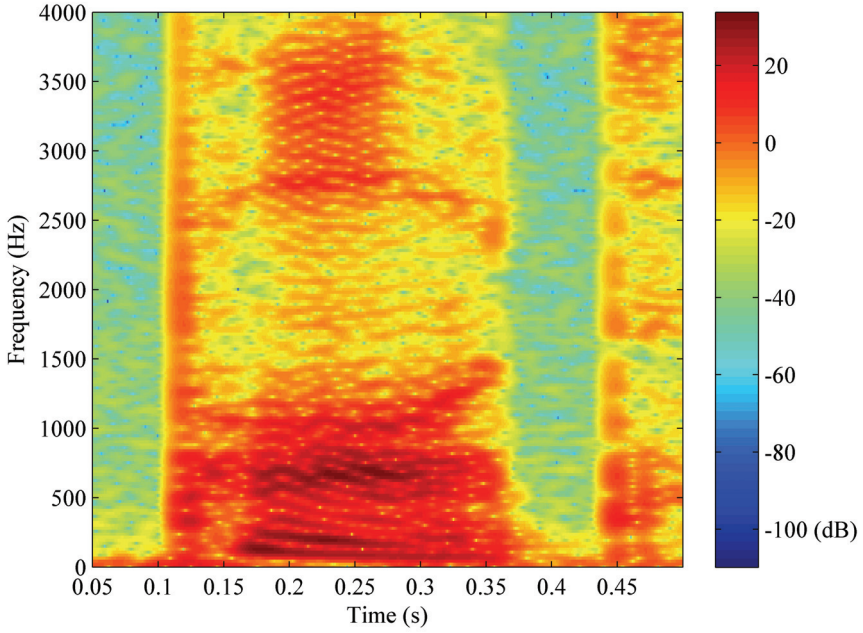


Figure 3.16: Spectrogram of the syllable 'pAAt' with no noise added to illustrate the temporal nature of formant frequencies.

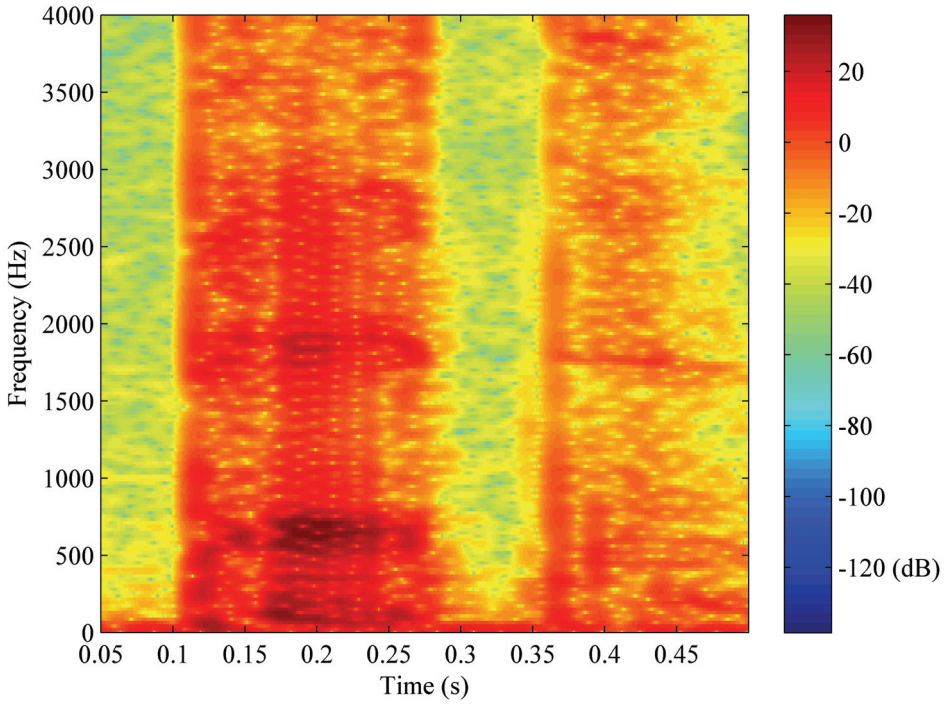
The structure of the frequency components for the voiced section of the spectrogram, in terms of duration and frequency span, is not known a priori. This creates the difficulty that the guard cells in the CFAR detector have to be chosen so that the voiced frequency components do not influence the estimation of the noise in the reference cells. To overcome this problem two CFAR detectors are used in parallel, one with a narrow guard window and another with a broad guard window. Frequency components that span a large frequency range, typically formant frequencies, will mask themselves with respect to a CFAR with a narrow guard window, but should be detected by a CFAR with a broad guard window. In a case where the frequency component span is only a few hertz, typically pitch frequencies, the noise estimation from the CFAR with the broad guard window may not be representative of the noise close to the frequency component and thus the detection threshold will not be set at the correct level. This can only be overcome by using a CFAR with a narrower guard window. The parameters for the guard cells for the broad and narrow windows are shown in Table 3.3. By means of simulation, it was found that these parameters provided good detection

performance at SNRs as low as -10 dB. The narrow guard window spans less time than the broad guard window. This is to allow the narrow window to more accurately follow frequency components that are more dynamic in frequency as time progress.

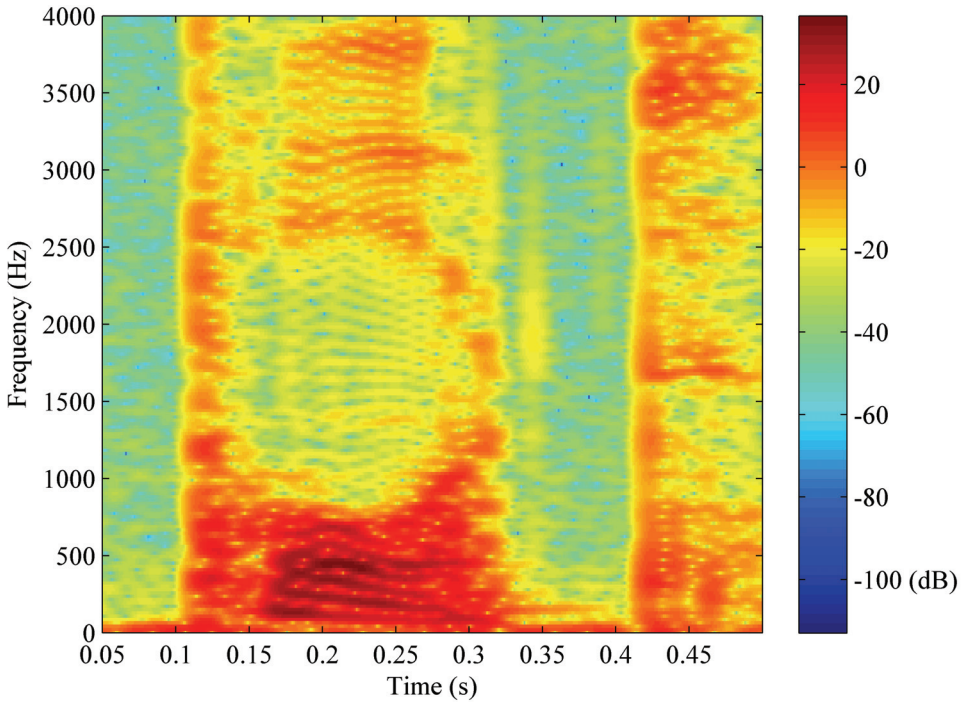
Table 3.3: CFAR guard cell parameters.

	Broad Guard Window	Narrow Guard Window
Time span (ms)	14.6	8.1
Frequency Span (Hz)	390.6	46.9

To illustrate the requirement for the parallel CFAR detectors, the output of the various processing steps shown in Figure 3.14 will be shown in the remainder of this section. The example will be for the syllables ‘pAt’ and ‘pAUt’ at an SNR of -10 dB. The syllables with no noise added are shown in Figure 3.17. From Figure 3.17 the duration of the vowel and the frequency of the various formants can be determined by hand as reference data. These reference data can be compared to the results of the automated cue estimation algorithms (see Figure 3.35). Detections at the output of the CFAR algorithm are shown in Figure 3.18 for the detector with a narrow guard window, and in Figure 3.19 for the detector with a broad guard window. In order to minimize detections that are not due to formant frequencies, a set of rules are applied to the detections at the output of the first processing stage. All detections with a duration of shorter than 10 ms are eliminated. The output of this process is shown in Figure 3.20 for the detector with a narrow guard window and in Figure 3.21 for the detector with a broad guard window. Once valid detections have been determined, these detections need to be clustered together in order to identify the groups of detections belonging to a specific formant frequency. All detections that are within 1 ms and 100 Hz of each other, are grouped together. The output of this processing step is shown in Figure 3.22 for the detector with a narrow guard window and Figure 3.23 for the detector with a broad guard window. Among the various clusters of detections, the cluster with the longest duration is chosen as the cluster representative of the longest formant frequency in the syllable. The selected cluster of detections are shown in Figure 3.24 for the detector with a narrow guard window and Figure 3.25 for the detector with a broad guard window. These examples were specifically chosen to illustrate that for ‘pAt’ the broad guard window provides the most accurate detection, whereas for ‘pAUt’ the narrow guard window is more accurate.

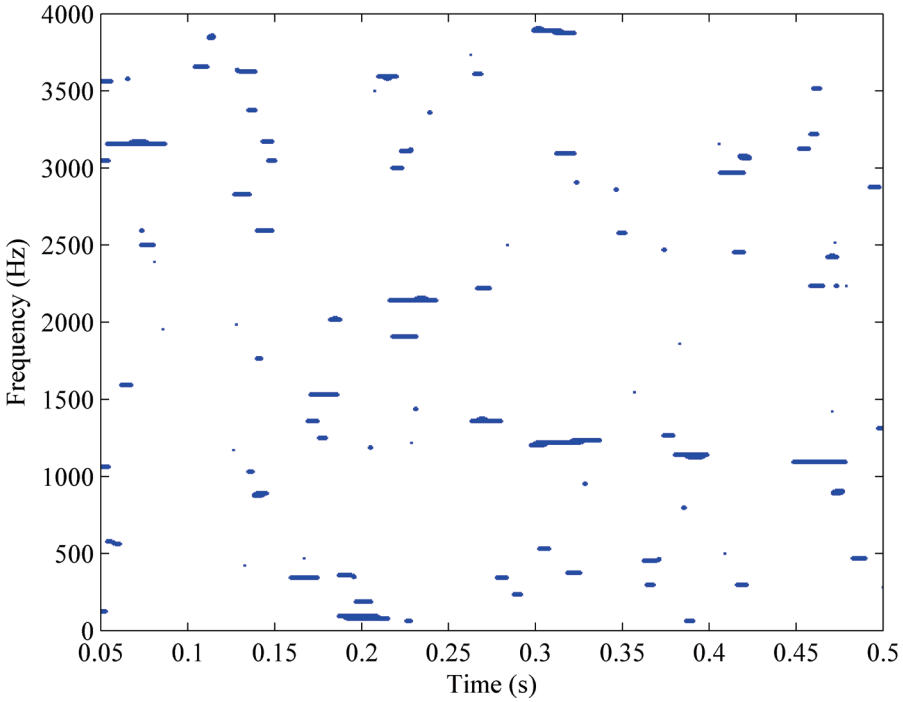


(a)

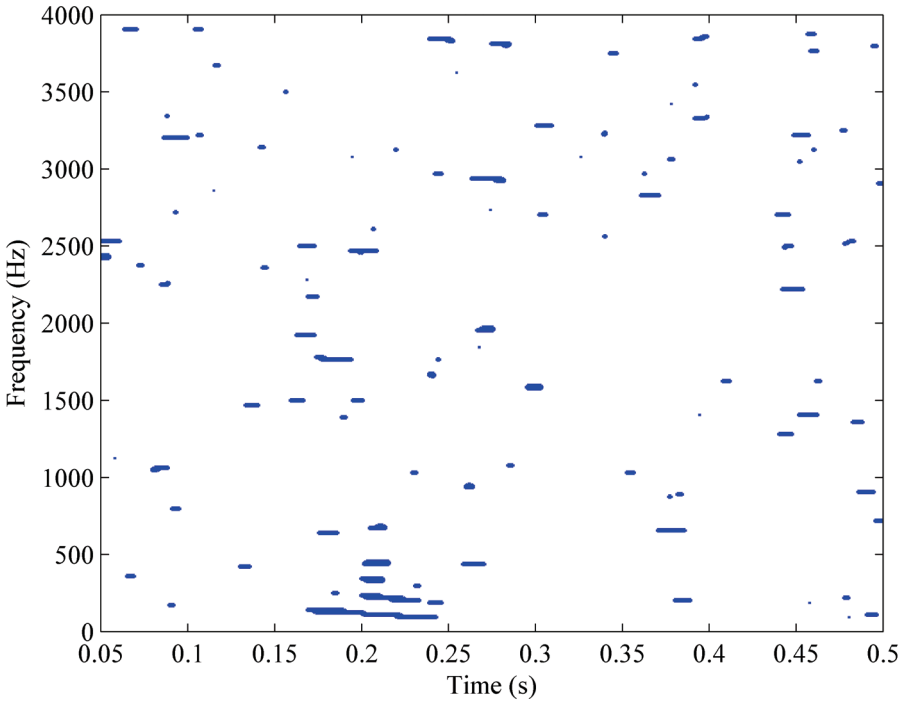


(b)

Figure 3.17: Spectrogram of ‘pAt’ (a), and ‘pAUt’ (b) with no noise added.

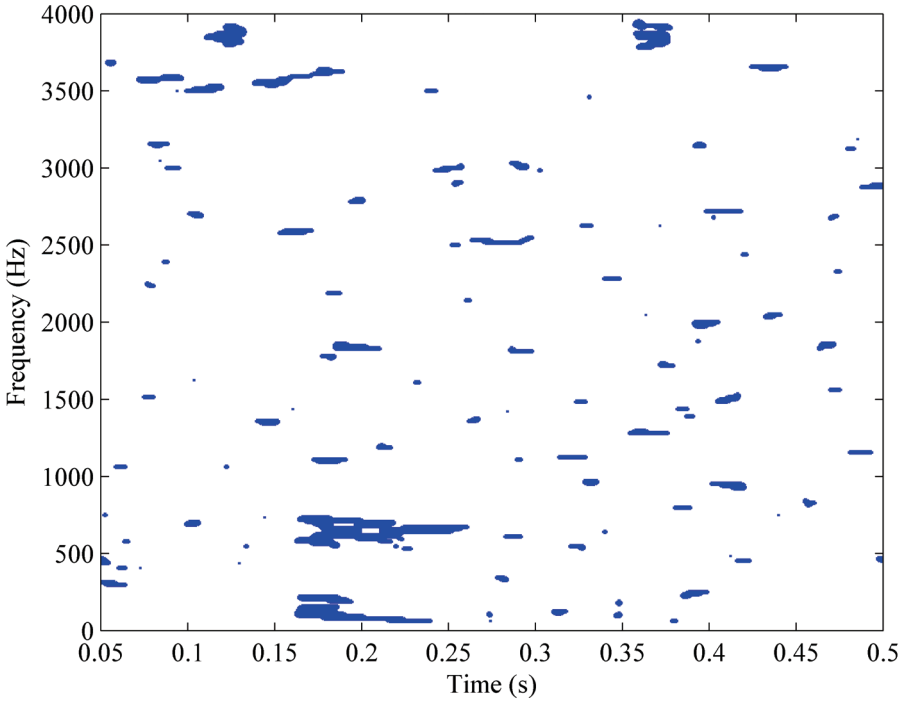


(a)

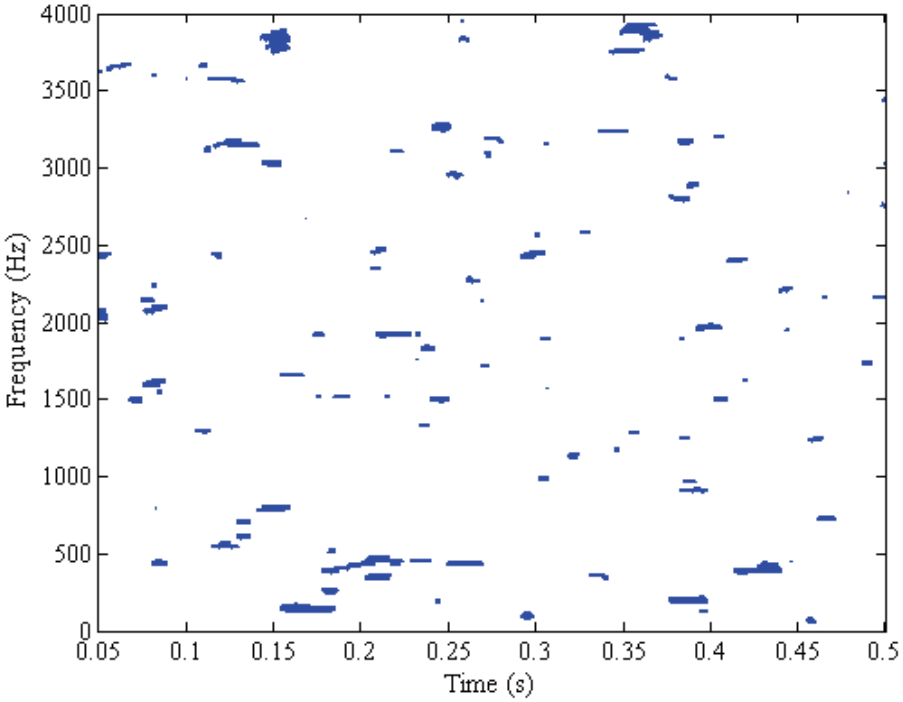


(b)

Figure 3.18: Output of the CFAR detector with a narrow guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

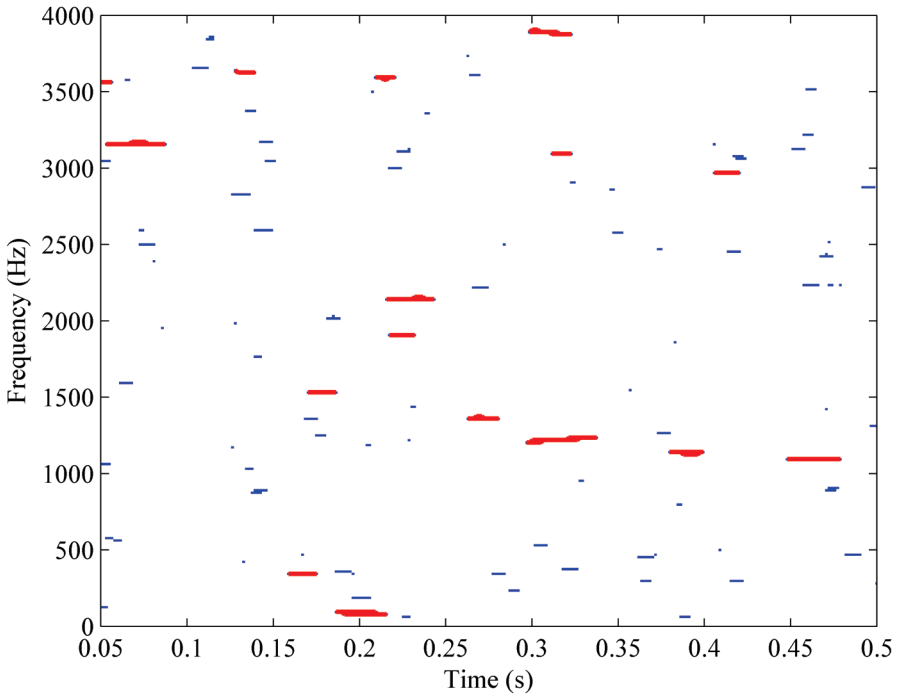


(a)

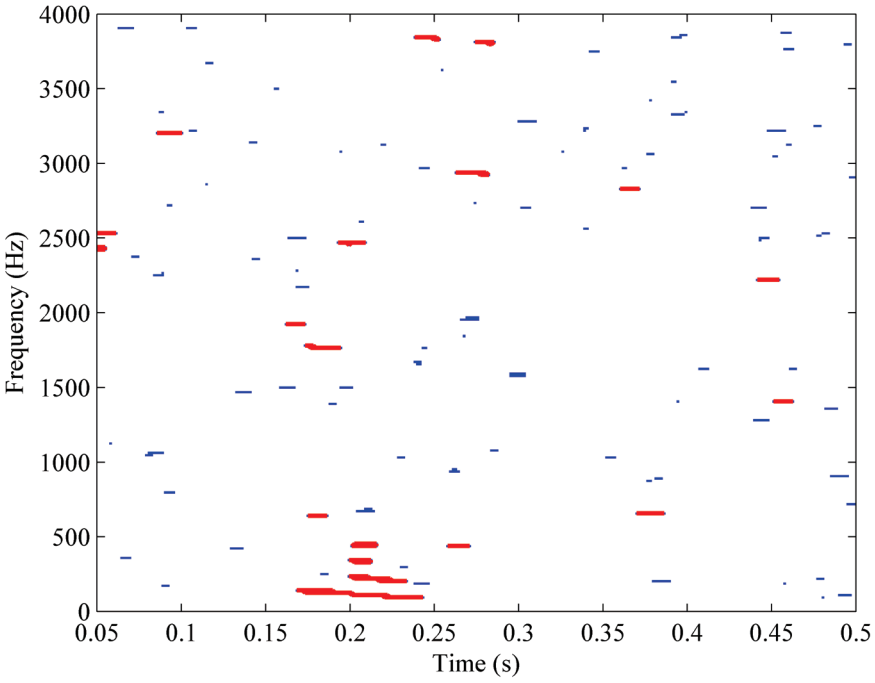


(b)

Figure 3.19: Output of the CFAR detector with a broad guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

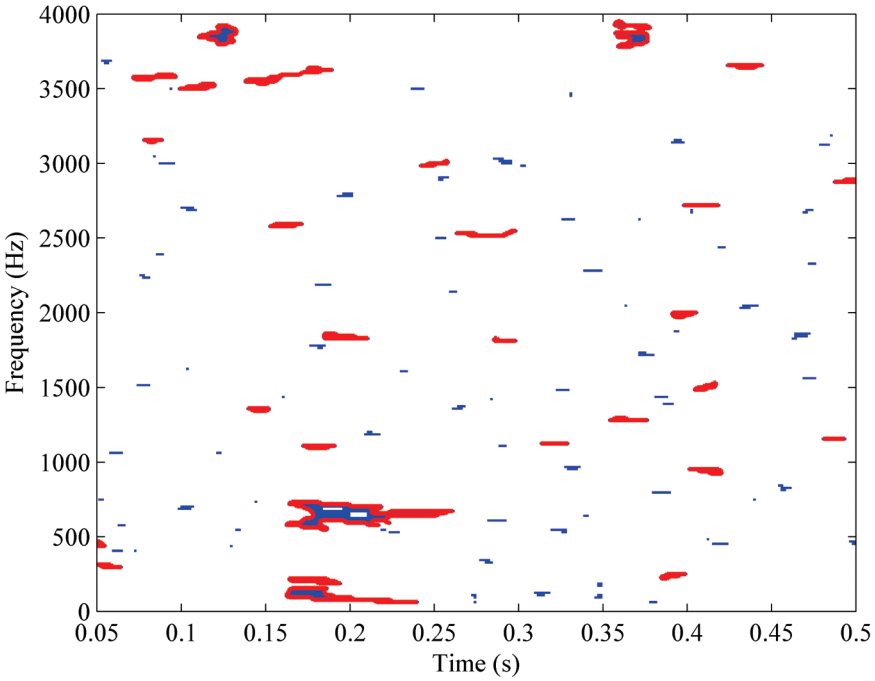


(a)

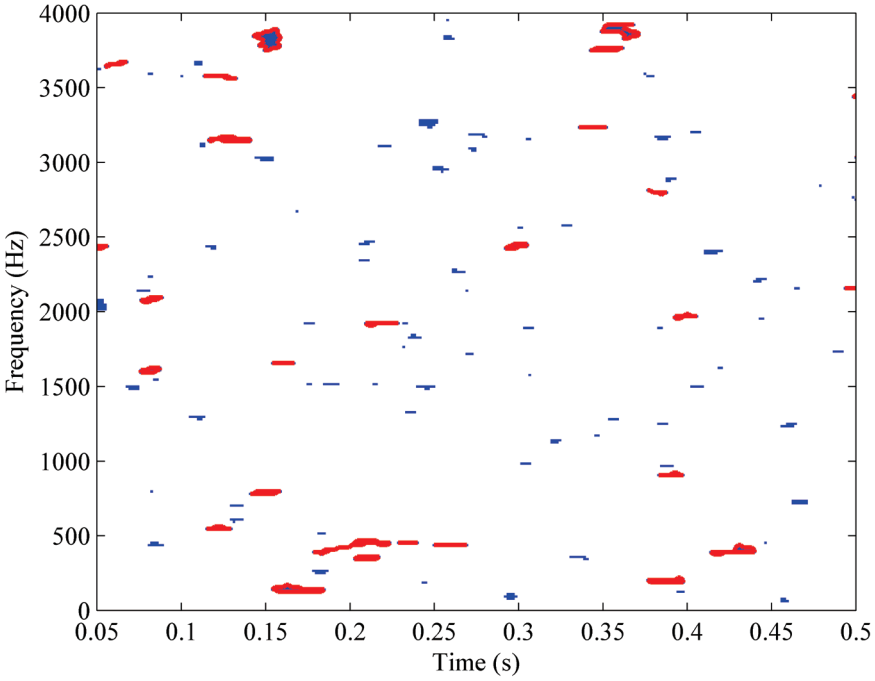


(b)

Figure 3.20: Detections after rule-based false detection elimination for the CFAR detector with a narrow guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

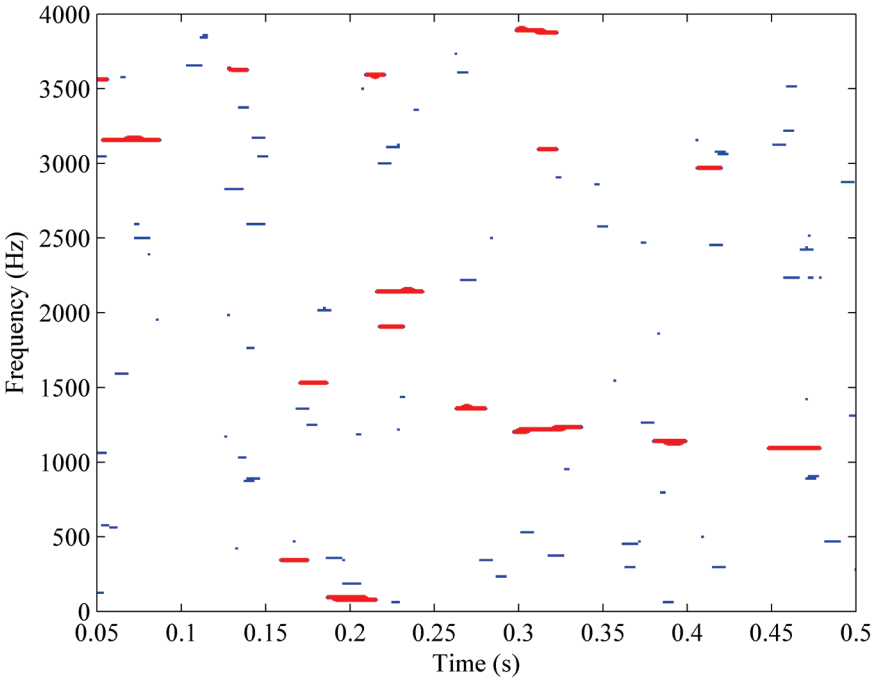


(a)

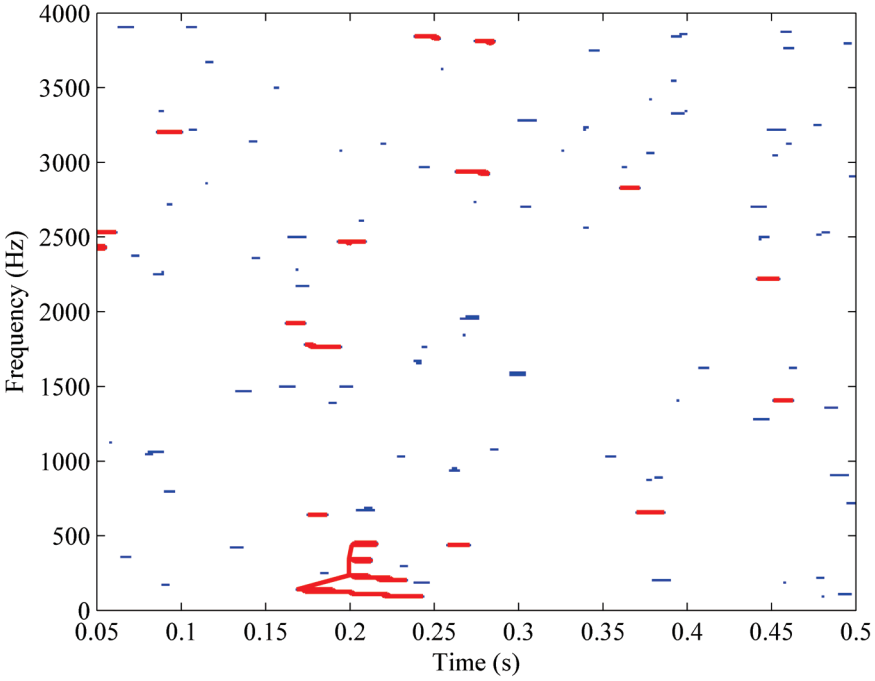


(b)

Figure 3.21: Detections after rule-based false detection elimination for the CFAR detector with a broad guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

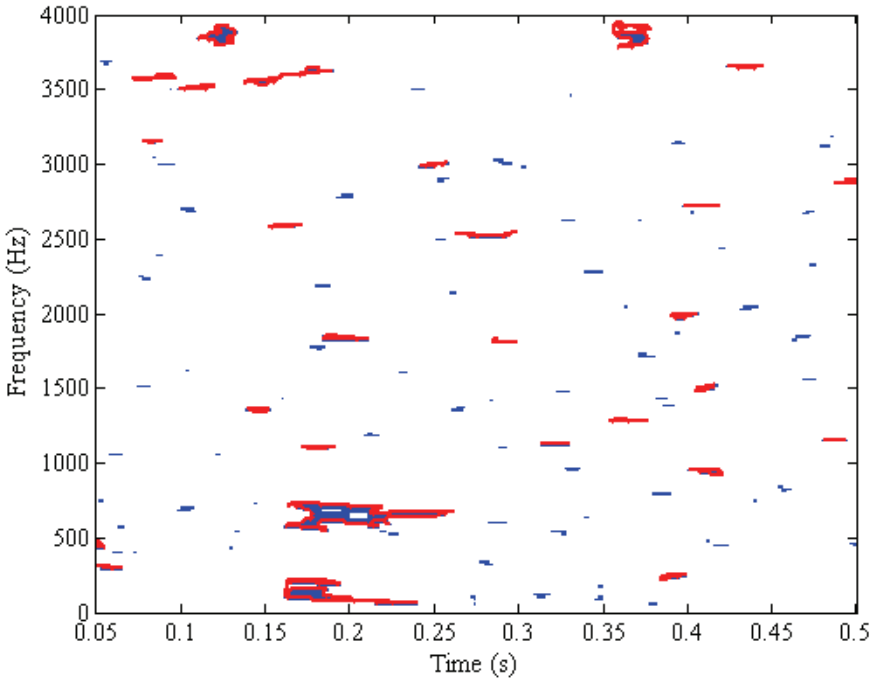


(a)

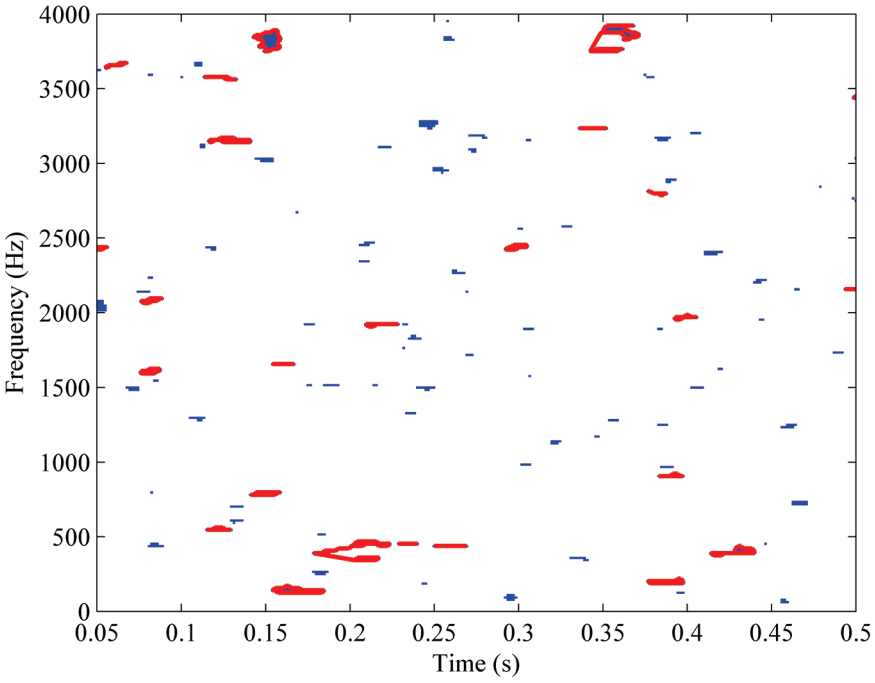


(b)

Figure 3.22: Output of detection clustering for the CFAR detector with a narrow guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

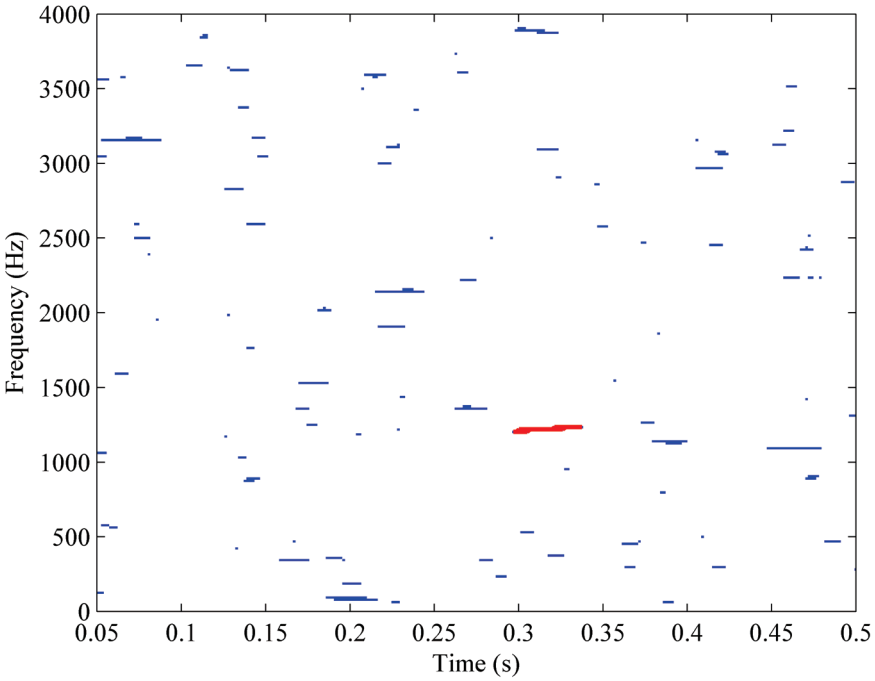


(a)

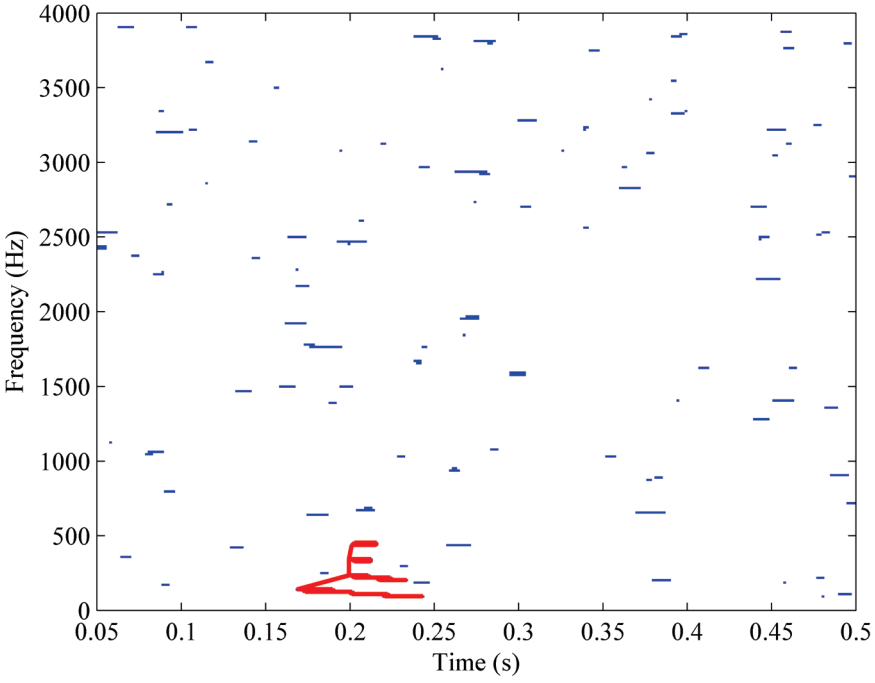


(b)

Figure 3.23: Output of detection clustering for the CFAR detector with a broad guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

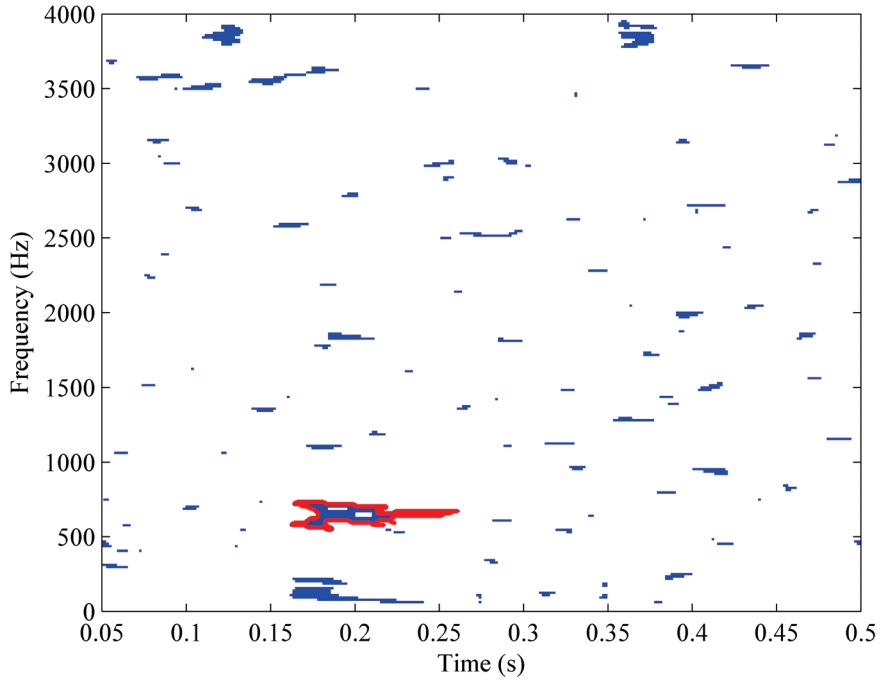


(a)

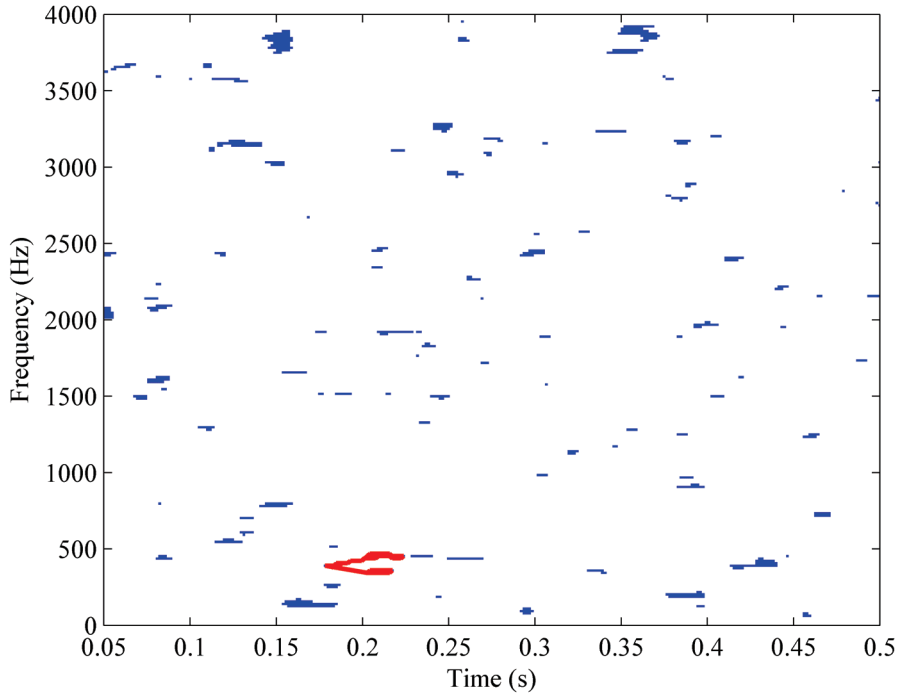


(b)

Figure 3.24: Final detection used for voicing duration estimation for the CFAR detector with a narrow guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.



(a)



(b)

Figure 3.25: Final detection used for voicing duration estimation for the CFAR detector with a broad guard window for 'pAt' (a), and 'pAUt' (b) for an SNR of -10 dB.

3.3.1.1 CFAR Detector Design

Owing to the severely degraded nature of the speech under investigation (negative SNR) and the fact that the signal plus noise power is not known at any given location in the spectrogram, a fixed threshold detector cannot be used for the individual cells in the spectrogram if the false alarm rate is to be controlled (Gandhi & Kassam, 1988). To overcome this problem a CFAR detector can be implemented, which sets the detection threshold adaptively based on the local information (reference cells surrounding the test cell) of the total noise power. The threshold is calculated for each cell in the spectrogram individually. There are essentially two categories of CFAR detectors, non-parametric and parametric (Dillard & Rickard, 1974; Thomas, 1970). Non-parametric detectors, also known as distribution-free detectors, have the characteristic that the probability of false alarm is constant, independent of the probability distribution of the input noise. This property does, however, degrade the performance of the non-parametric detector compared to the parametric detector. The parametric detector assumes that the probability density function (PDF) of the noise is known and estimates only the parameters of the noise distribution in order to determine the required threshold to give a desired probability of false alarm, P_{fa}. For the CFAR detector design discussed in this section, a parametric CFAR detector is proposed. The CFAR detector architecture shown in Figure 3.15 is a cell averaging greater of (CAGO) CFAR detector. This CFAR detector has the characteristic that it is more resistent to false alarms. It can regulate the false alarm rate better in the transition region where the input goes, for example, from noise to signal (Davidson, Griffiths, & Ablett, 2004; Khalighi & Bastani, 2000). Such transitions can also occur during the transition from an unvoiced to a voiced sound, as would be the case in a syllable such as “pAt”. For a detection decision to be made, one of two hypotheses has to be accepted, namely (Chang, Kim, & Mitra, 2006):

$$H_0 : \text{Speech is absent} : x(t) = n(t) \tag{3.52}$$

$$H_1 : \text{Speech is present} : \mathrm{x}(t) = n(t) + s(t) \tag{3.53}$$

where:

$$x(t) = [x_0(t), x_1(t), ..., x_{M-1}(t)] \tag{3.54}$$

$$n(t) = [n_0(t), n_1(t), ..., n_{M-1}(t)] \tag{3.55}$$

$$s(t) = [s_0(t), s_1(t), ..., s_{M-1}(t)] \tag{3.56}$$

are the samples at segment t of the noisy speech, noise and clean speech respectively. With

$$X(t) = [X_0(t), X_1(t), ..., X_{M-1}(t)] \tag{3.57}$$

$$N(t) = [N_0(t), N_1(t), ..., N_{M-1}(t)] \tag{3.58}$$

$$S(t) = [S_0(t), S_1(t), ..., S_{M-1}(t)] \tag{3.59}$$

the discrete Fourier transform (DFT) coefficients at segment t of (3.54), (3.55) and (3.56) respectively. In speech analysis the complex Gaussian PDF can be used to characterize the distribution of the DFT coefficients (Chang, Kim, & Mitra, 2006; Duk & Kondoz, 2001). With this Gaussian PDF assumption, the distributions of the noisy spectral components conditioned on the hypotheses stated in (3.52) and (3.53) are given by

$$p(X_k \mid H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\left(-\frac{|X_k|^2}{\lambda_{n,k}} \right) \tag{3.60}$$

$$p(X_k \mid H_1) = \frac{1}{\pi(\lambda_{n,k} + \lambda_{s,k})} \exp\left(-\frac{|X_k|^2}{\lambda_{n,k} + \lambda_{s,k}} \right) \tag{3.61}$$

with $\lambda_{n,k}$ the variance of N_k and $\lambda_{s,k}$ the variance of S_k. The magnitude spectral components for both of the hypotheses will have Rayleigh probability density functions given by (Proakis & Salehi, 1994):

$$p(v \mid H_0) = \frac{2v}{\lambda_{n,k}} \exp\left(-\frac{v^2}{\lambda_{n,k}}\right) \tag{3.62}$$

$$p(v \mid H_1) = \frac{2v}{\lambda_{n,k} + \lambda_{s,k}} \exp\left(-\frac{v^2}{\lambda_{n,k} + \lambda_{s,k}}\right) \tag{3.63}$$

with:

$$v = \sqrt{\text{Real}(X_k)^2 + \text{Imag}(X_k)^2} \tag{3.64}$$

The Rayleigh distribution given in (3.62) is the basis for the calculation of the required P_{fa} for the CAGO CFAR detector, since the threshold for a detection is calculated using the statistics of the noise. Given that the statistics of the noise are known and that the structure of the CFAR reference cells are fixed, the CFAR constant, as illustrated in Figure 3.15, determines the P_{fa}. The process of calculating the CFAR constant is illustrated in Figure 3.26. For the transformation of the Rayleigh PDF to the PDF of the threshold, a number of transformations need to be applied to the Rayleigh PDF (Papoulis, 1991). To calculate the sum of two samples, given by:

$$Y = X_1 + X_2 \tag{3.65}$$

the following transformation is required:

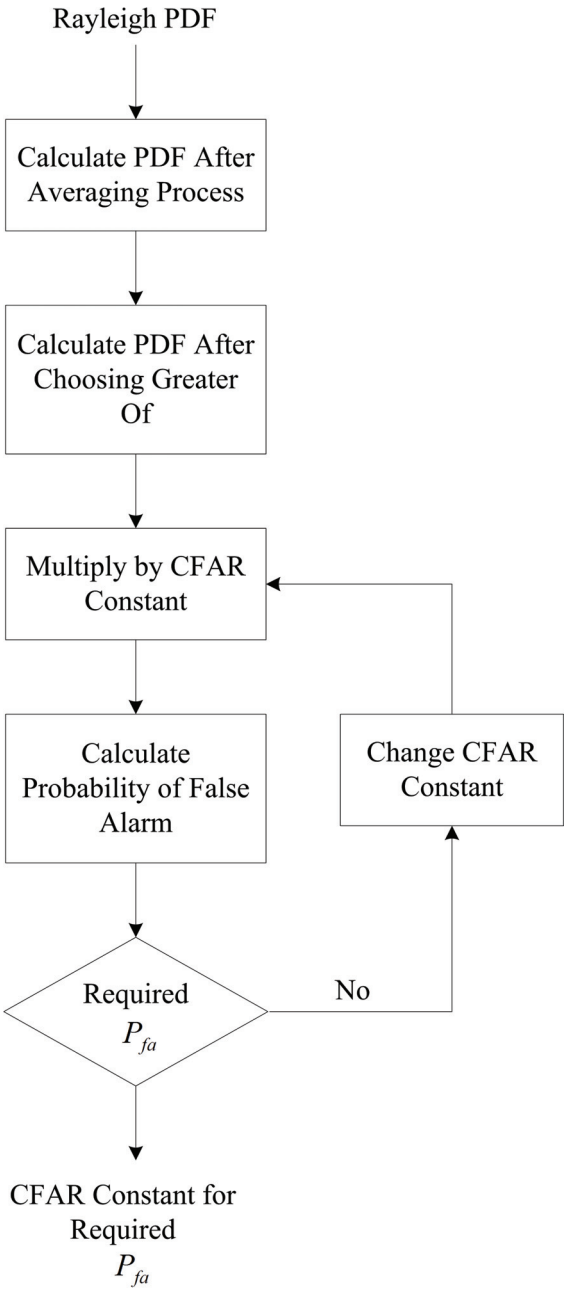


Figure 3.26: Steps in calculating the CFAR constant with the P_{fa} as input to the process.

$$f_Y(y) = f_{X_1}(y) * f_{X_2}(y) \tag{3.66}$$

with $*$ denoting the convolution of $f_{X_1}(y)$ and $f_{X_2}(y)$. Using the property of the Fourier transform that a convolution in the time domain is a multiplication in the frequency domain

(Oppenheim & Schafer, 1999), the Fourier transform can be used to easily calculate the PDF of the sum of k samples given by:

$$Y = \sum_{n=1}^{n=k} X_n \tag{3.67}$$

as

$$f_Y(y) = \Im^{-1}\left(\Im[f_X(y)]^k\right) \tag{3.68}$$

with $\Im(x)$ and $\Im^{-1}(x)$ the Fourier and inverse Fourier transforms respectively, with $f_Y(y)$ the transformed PDF and $f_X(y)$ the PDF to be transformed.

For the multiplication with a constant C, given by

$$Y = C \cdot X \tag{3.69}$$

as would be required during the calculation of an average, the following transformation is required:

$$f_Y(y) = \frac{f_X(y/c)}{|C|} \tag{3.70}$$

For the selection of the greater of the two averaged samples, given by

$$Y = \max(X_1..X_k) \tag{3.71}$$

the following transformation is required:

$$f_Y(y) = k \cdot F_X(y)^{k-1} \cdot f_X(y) \tag{3.72}$$

with $F_X(y)$ the cumulative distribution function (CDF) of $f_X(y)$. Figure 3.27 shows the PDFs after the various transformations required to calculate the P_{fa}, which is the green area indicated in Figure 3.27. This example was calculated for a CFAR constant of 2 and input noise variance of 1.

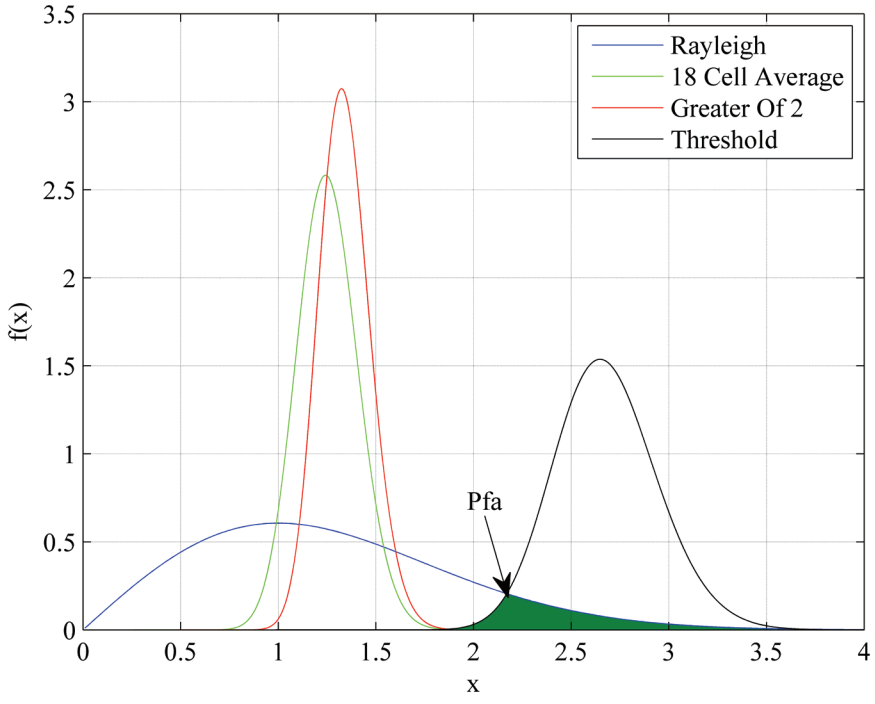


Figure 3.27: PDFs for the various transformations used to calculate the CFAR constant.

A false alarm is declared when the random variable from the Rayleigh distribution is larger than the random variable from the threshold distribution, which can be expressed as

$$\textit{Rayleigh Random Variable} > \textit{Threshold Random Variable} \tag{3.73}$$

In order to calculate the P_{fa}, it is convenient to express (3.73) as

$$\textit{Threshold Random Variable - Rayleigh Random Variable} < 0 \tag{3.74}$$

and evaluate the CDF of (3.74) at zero. Figure 3.28 illustrates the various PDFs involved in the calculation of the P_{fa}. The Rayleigh PDF is multiplied by -1 and convolved with the threshold PDF in order to obtain the PDF from which the P_{fa}, the area under the PDF smaller than zero, is calculated. This area is equal to the value of the CDF at zero.

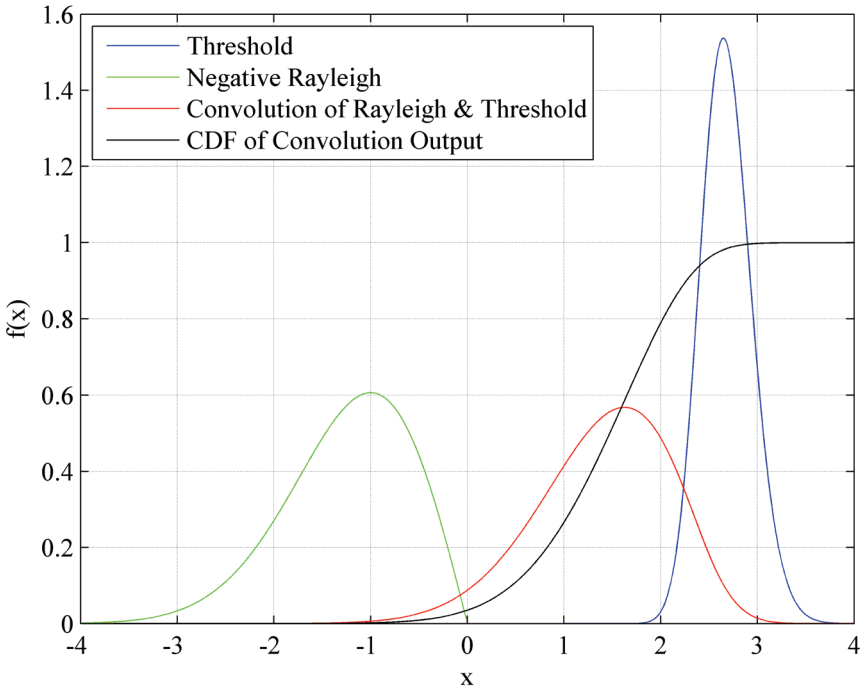


Figure 3.28: PDFs used to calculate the P_{fa}.

For the implemented CFAR detectors, one with a broad guard window and one with a narrow guard window, the number of reference cells are 18 and 15 respectively. Using the process described above a graph can be made of P_{fa} vs. CFAR constant, Figure 3.29. By means of simulation it was determined that a P_{fa} of one false detection in one thousand samples ($P_{fa} = 1e-3$) provides adequate probability of detection at an SNR of -10 dB without placing an excessive processing load on the rule-based filter (refer to Figure 3.14). With this P_{fa} the CFAR constants for the broad guard window and the narrow guard window were 2.94 and 2.96 respectively.

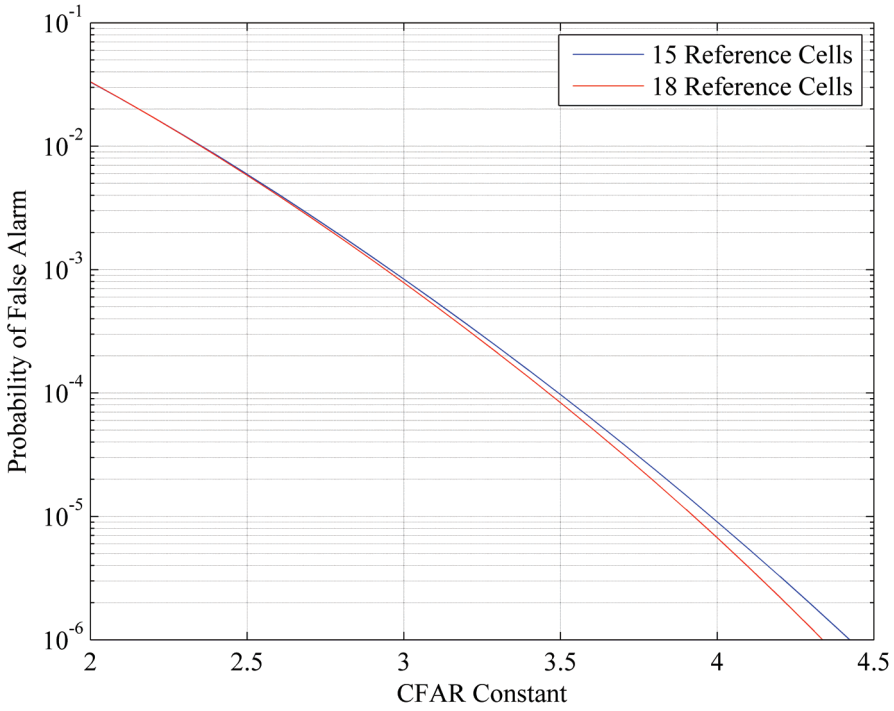


Figure 3.29: Probability of false alarm vs. CFAR constant.

3.3.2 Formant Frequency Estimation

The formant frequency estimation is done by using LPC analysis (Atal & Hanauer, 1971; McCandless, 1974; Snell & Milinazzo, 1993). The duration estimation, as described in the previous sections, is used as an input to the formant frequency estimation in order to estimate

the pre-emphasis filter. The processing steps involved in the formant estimation are shown in Figure 3.30.

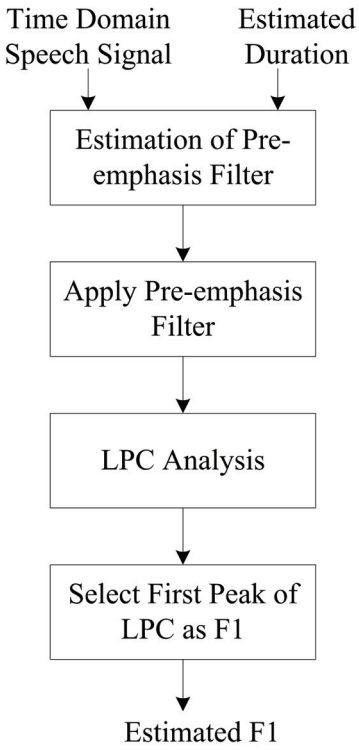


Figure 3.30: Processing steps for formant estimation.

The purpose of the pre-emphasis filter is to re-shape the spectrum of the input syllable in order to increase the probability of detecting the various formant frequencies. This is achieved by suppressing pitch frequencies and lifting/emphasising higher frequency components, which typically roll off/attenuate as frequency increases. A high order (10^{th}) LPC spectral analysis of the input vowel 'A' in 'pAt', with SNR of 5 dB and -5 dB, is shown in Figure 3.31. In the figure the pitch frequencies (near 0 Hz) and the roll-off of the higher frequency components are clearly visible.

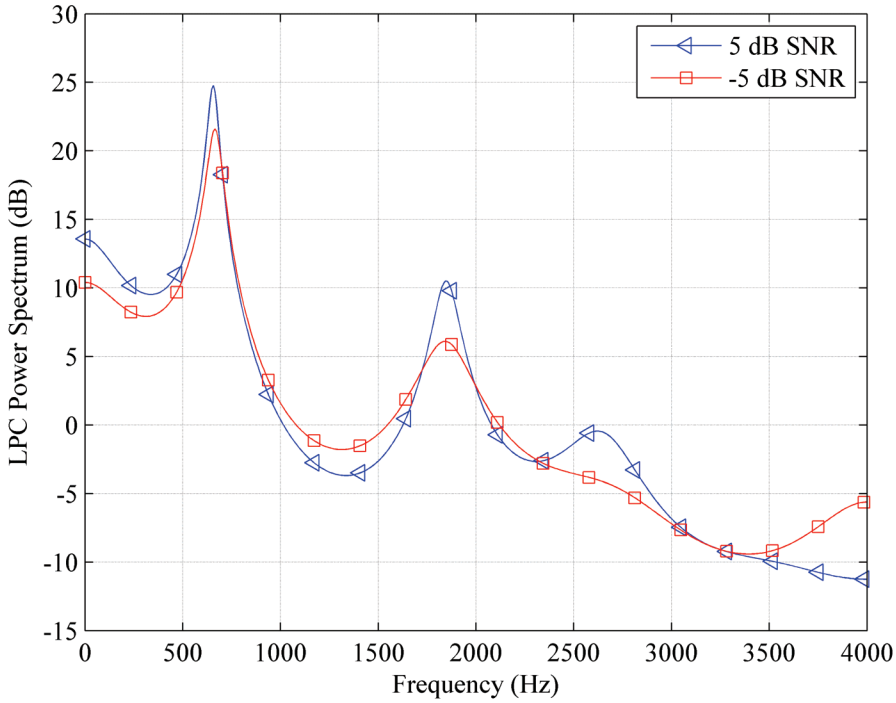


Figure 3.31: 10th order LPC spectral analysis of the vowel in 'pAt' for SNR of 5 dB and -5 dB.

The pre-emphasis filter is estimated using a first order LPC analysis of the input vowel 'A', which is extracted from the input syllable using the identified duration of the vowel. The 1st order LPC analysis for the vowel 'A' is shown in Figure 3.32. The pre-emphasis filter is the inverse of this 1st LPC spectral estimation, which is shown in Figure 3.33. The 10th order LPC spectral estimation of the syllable 'pAt' after the pre-emphasis is shown in Figure 3.34. This 10th order LPC spectral estimation is used to identify the first formant frequency. By using a peak search algorithm, F1 is identified. The final estimation showing the duration and F1 is shown in Figure 3.35.

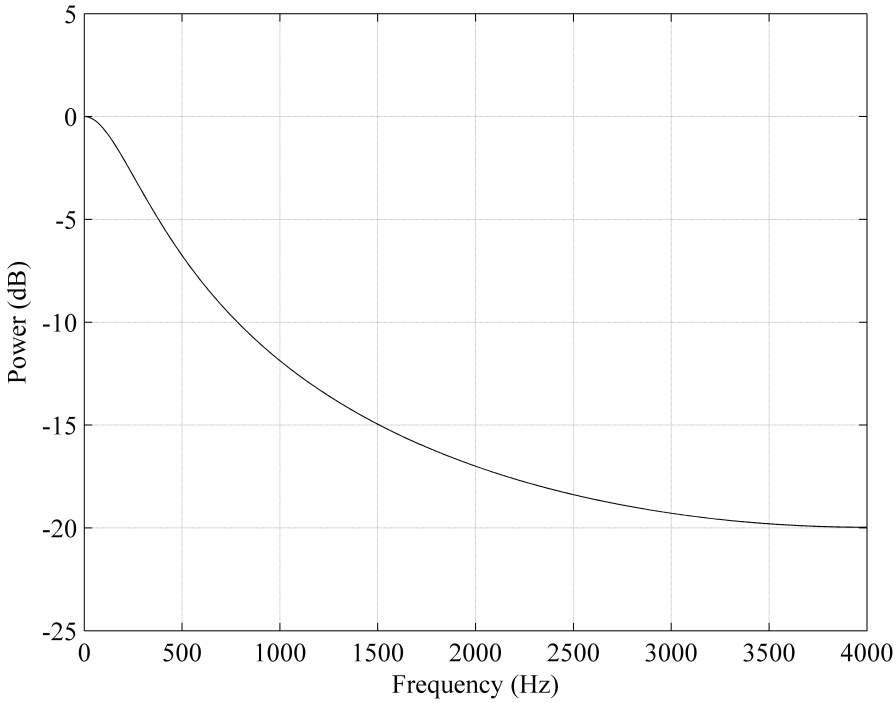


Figure 3.32: First order LPC analysis for input vowel.

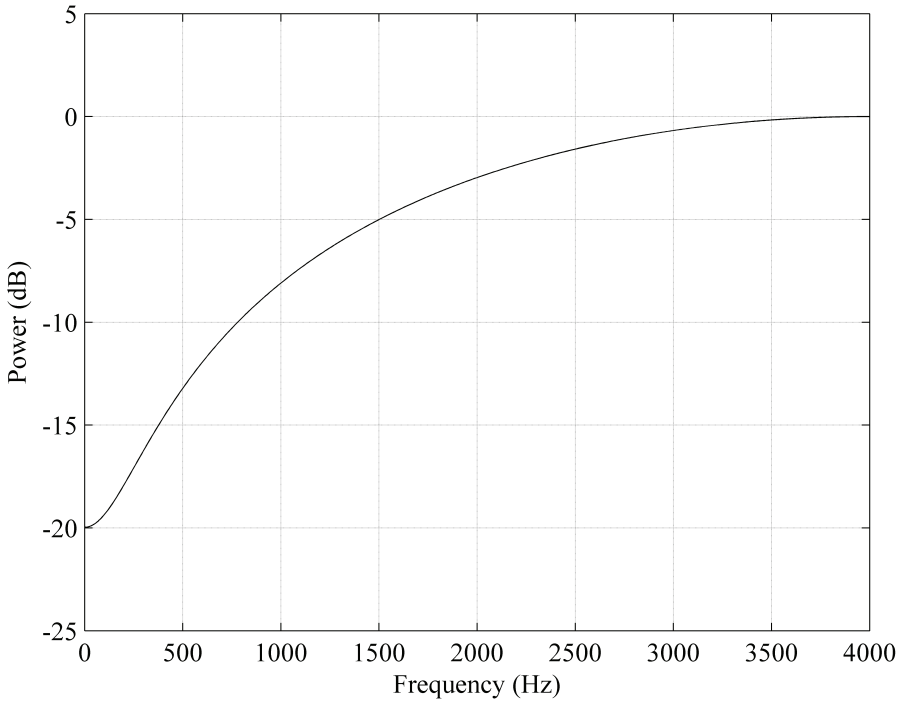


Figure 3.33: Estimated pre-emphasis filter.

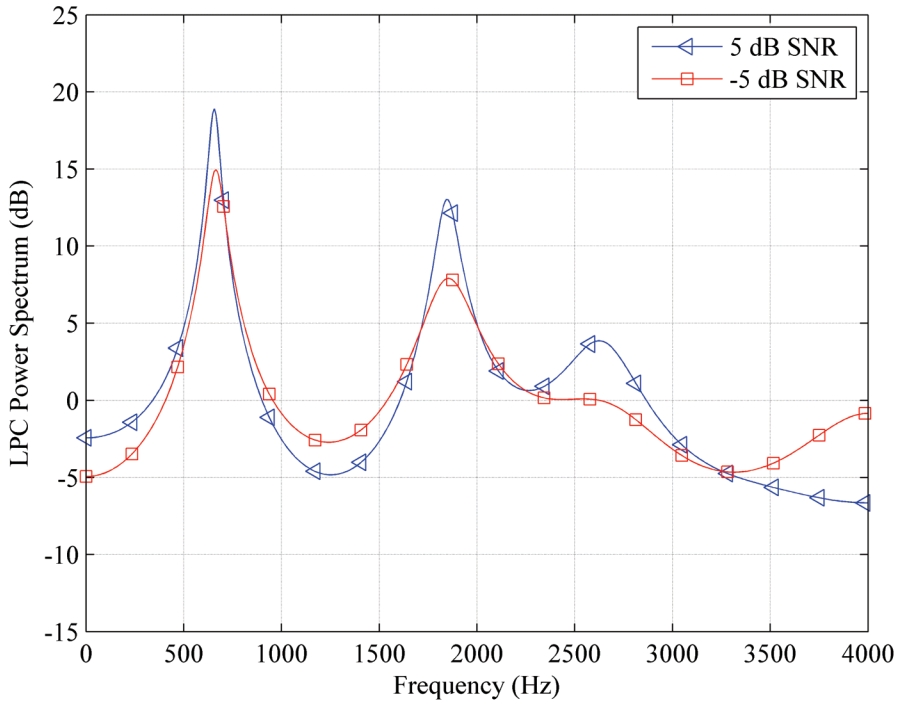
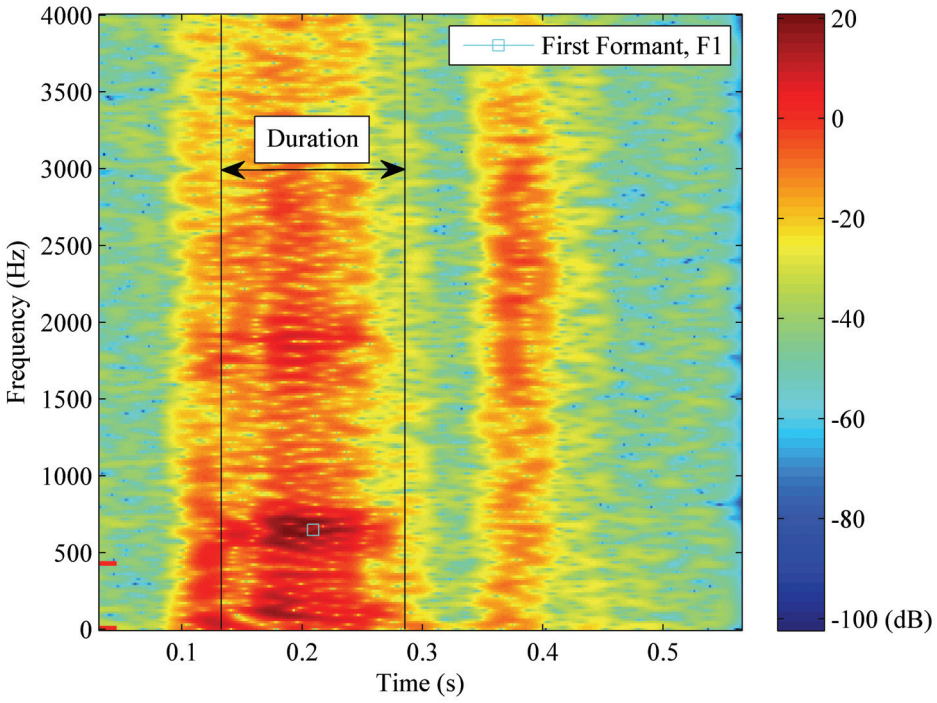
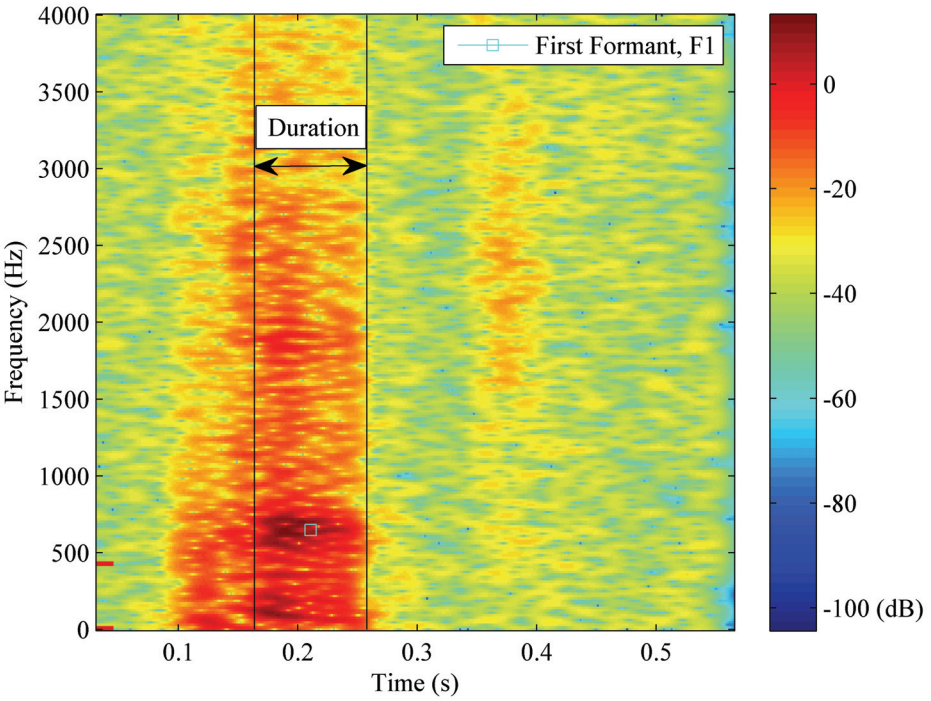


Figure 3.34: 'pAt' LPC spectrum after pre-emphasis for SNR of 5 dB and -5 dB.



(a)



(b)

Figure 3.35: Spectrogram of the syllable 'pAt' with estimated duration and F1 for SNR of (a) 5 dB, and (b) -5 dB.

3.3.3 Channel RMS Amplitude Estimation

Because the classification method proposed by Svirsky (2000) is used in this study, the same cues required for the classification algorithm are estimated. These cues are the first formant frequency and the amplitude ratio of four overlapping bandpass filters as implemented in the stimulation strategy used in the Ineraid multichannel CI (refer to Figure 2.2). The filters have crossover frequencies of 700 Hz, 1.4 kHz, and 2.3 kHz and roll-off slopes of 12 dB per octave. The magnitude transfer functions of the four filters are shown in Figure 3.36. The filters are implemented as a cascade of 2^{nd} order Butterworth filters approximated using 2^{nd} order discrete transfer functions of the form (Oppenheim & Schafer, 1999):

$$H(z) = G\frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{3.75}$$

with b_k the numerator coefficient, a_k the denominator coefficient and G the gain parameter. The coefficient values for the various filters are given in Addendum A.

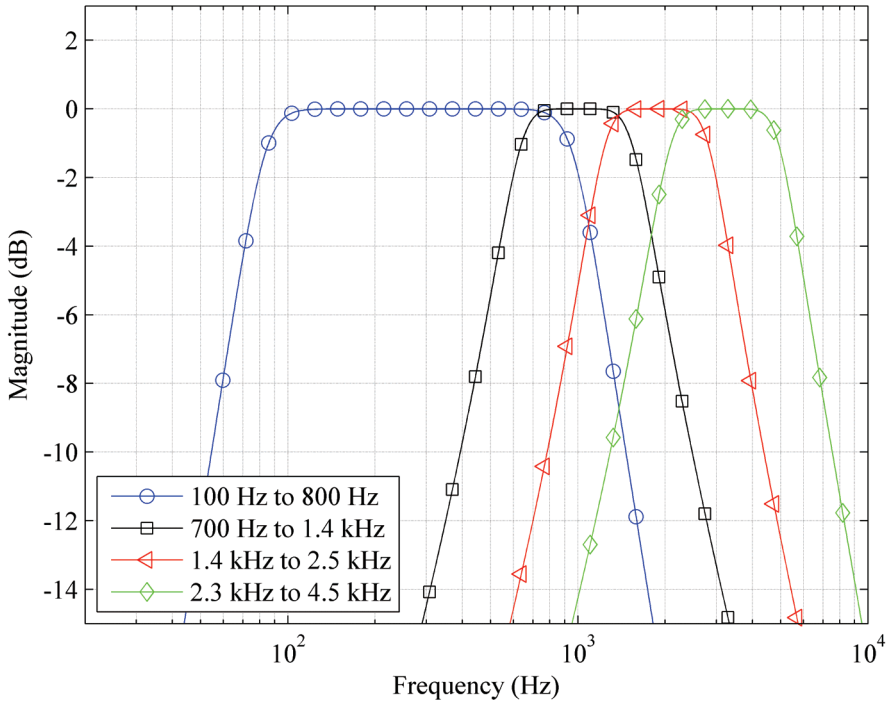


Figure 3.36: Magnitude transfer functions of the four bandpass filters used to calculate the RMS amplitude ratios.

The three cues obtained from this processing step are the RMS amplitude ratio (in decibel) of the first channel (100 Hz to 800 Hz) to the second (700 Hz to 1.4 kHz), third (1.4 kHz to 2.5 kHz) and fourth (2.3 kHz to 4.5 kHz) channels, where each channel is defined as the output of the various bandpass filters. The input to the filter is the vowel in the CVC syllable. With a CVC syllable as input to the algorithm, the voicing detector is used to determine the portion of

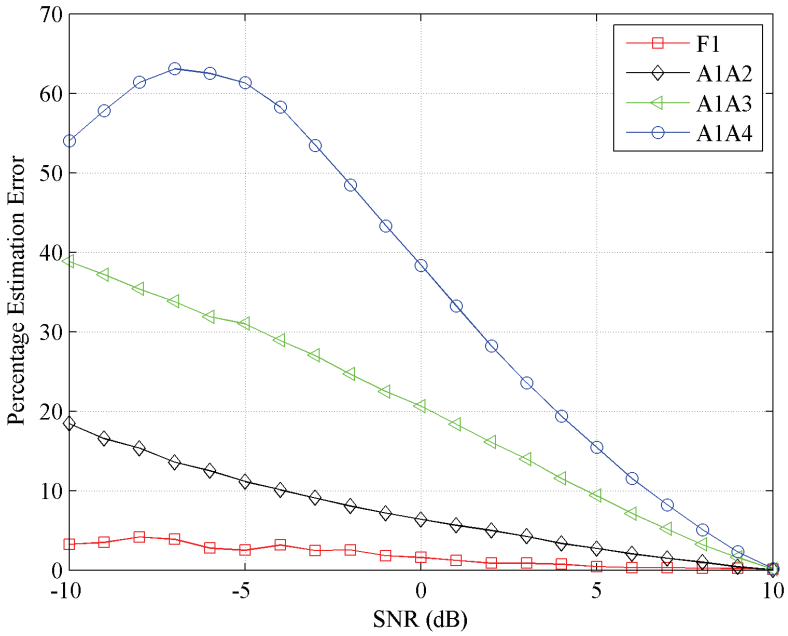
the syllable containing the vowel. The RMS amplitude ratios are calculated using the following equation:

$$RMS = 10\log_{10}\left(\sqrt{\frac{1}{n}\sum_{k=1}^{n} x_k^{\,2}}\right) \tag{3.76}$$

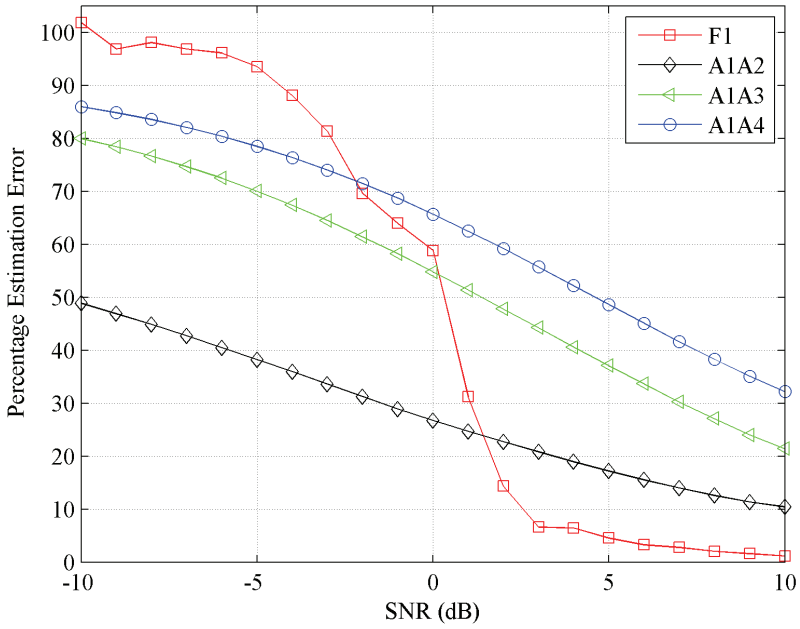
with x_k the time domain sample of the filtered signal. The four cues estimated with the various techniques described in this chapter are the inputs to the classification algorithm used to generate a confusion matrix and ultimately a graph of perception performance vs. SNR.

3.3.4 Evaluation of Cue Estimation Performance

To analyze the benefit of using the speech enhancement technique for the purpose of automated cue estimation, the percentage estimation error for the various SNRs from -10 dB to 10 dB was measured. The reference values for the calculation of the percentage error were the estimated cue values for the respective vowels derived from the high SNR (10 dB) input data. The correctness of these reference values were verified by inspection. The percentage estimation error that was achieved with and without speech enhancement is shown in Figure 3.37. The percentage estimation error of Figure 3.37 was averaged over all vowels. It can be seen that the signal enhancement increases the accuracy of the cue estimation significantly for all of the cues for the entire SNR range from -10 dB to 10 dB. Because of the different spectral and temporal nature of the various vowels, it cannot be expected that that the speech enhancement and cue estimation algorithms will perform equally well for all of the vowels. Figure 3.38 shows the percentage estimation error for each of the vowels when averaged over the SNR range from -10 dB to 10 dB and over the four selected cues. From the figure it is clear that the speech enhancement technique significantly decreases the percentage estimation error. When averaging the data of Figure 3.38 over all of the vowels, the average percentage estimation error when using speech enhancement is 16.2 percent as opposed to 48.1 percent when speech enhancement is not performed.



(a)



(b)

Figure 3.37: Percentage estimation error for the various cues with (a) speech enhancement and (b) no speech enhancement. F1 is the first formant frequency and A1A2, A1A3 and A1A4 are the RMS amplitude ratios of the first channel to the second, third and fourth channels respectively (the channel's definitions are shown in Figure 2.2 and Figure 3.36).

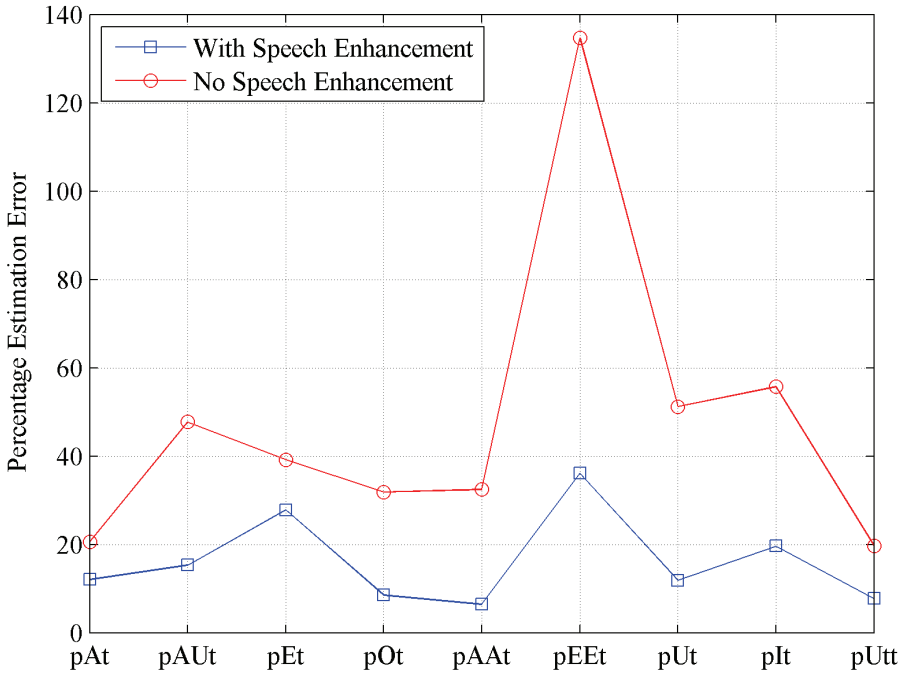


Figure 3.38: Percentage estimation error with and without speech enhancement. The percentage estimation error was averaged over the SNR range from -10 dB to 10 dB, as well as the selected four cues.

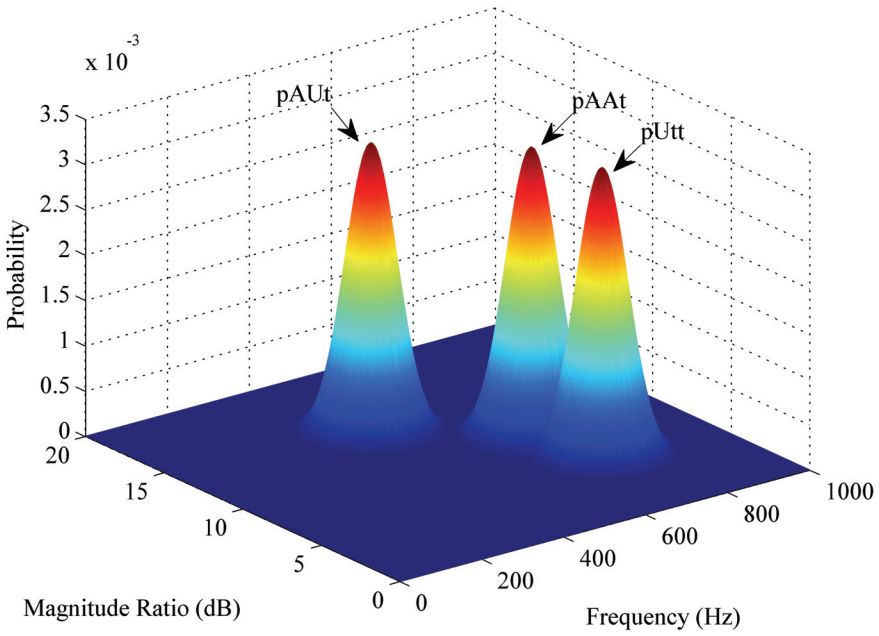
3.4 Vowel Classification

The input parameter for the multidimensional Gaussian classifier as used by Svirsky is the just noticeable difference (JND) for the various cues to be used for classification. The JND for the first formant frequency used in the model presented by Svirsky is 120 Hz. For the channel magnitude ratios the JND is 2.6 dB for each channel. This model assumes that the multidimensional decision space is Euclidean and that the dimensions are orthogonal. The JND is used as the standard deviation of the multidimensional Gaussian distribution. The mean of each vowel for the various cues is derived from the high SNR input data. The mean values of the various vowels are shown in Table 3.4.

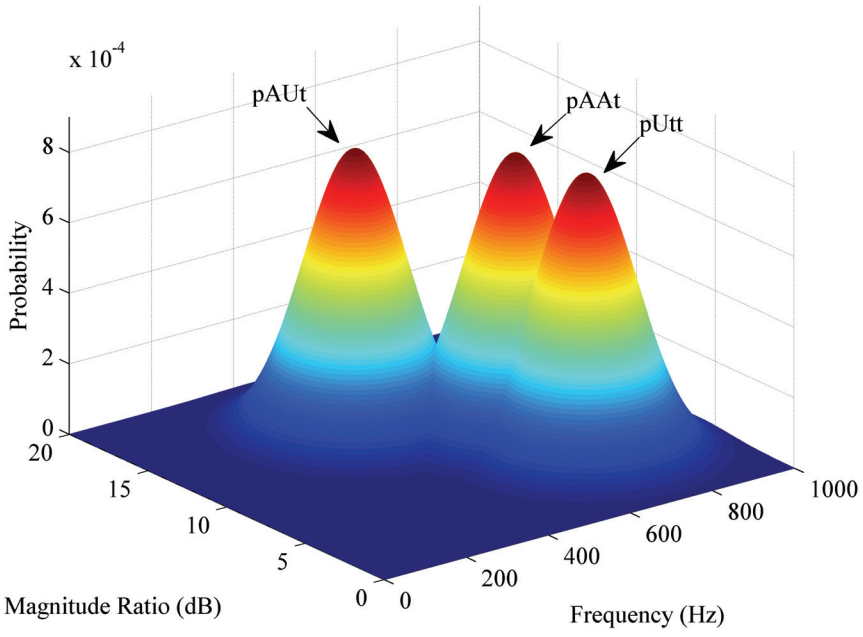
Table 3.4: Mean values used in multidimensional Gaussian classifier for the various vowels.

	First Formant Frequency (Hz)	Channel 1 to Channel 2 Magnitude Ratio (dB)	Channel 1 to Channel 3 Magnitude Ratio (dB)	Channel 1 to Channel 4 Magnitude Ratio (dB)
pAt	648.4	1.494	5.115	6.546
pAUt	454.8	5.909	12.51	13.688
pEt	460.9	7.674	7.224	7.3960
pOt	675.9	1.917	8.67	8.774
pAAt	687.1	1.514	8.787	9.551
pEEt	203.1	13.501	11.391	7.81
pUt	374.1	9.231	11.632	11.119
pIt	422.8	8.057	7.04	7.412
pUtt	727	1.405	5.028	6.11

Figure 3.39 shows an example of a two dimensional perceptual space. Figure 3.39(a) is an example of an exceptional listener with small JNDs in both perceptual dimensions. This would imply that the listener will not make many wrong associations or classifications when listening to the various vowels. Figure 3.39(b) is an example of a listener with larger JNDs. Owing to the overlap in the Gaussian distributions it can be foreseen that this listener will make more wrong associations when listening to the vowels.



(a)



(b)

Figure 3.39: Illustration of a classification space for pAAt, pAUt, and pUtt with a standard deviation of (a) 50 Hz and 1 dB for the 1st formant frequency and the channel 1 to channel 4 magnitude ratios respectively, and (b) 100 Hz and 2 dB for the 1st formant frequency and the channel 1 to channel 4 magnitude ratios respectively.

An example of the classification space used by the model presented by Svirsky is shown in Figure 3.40.

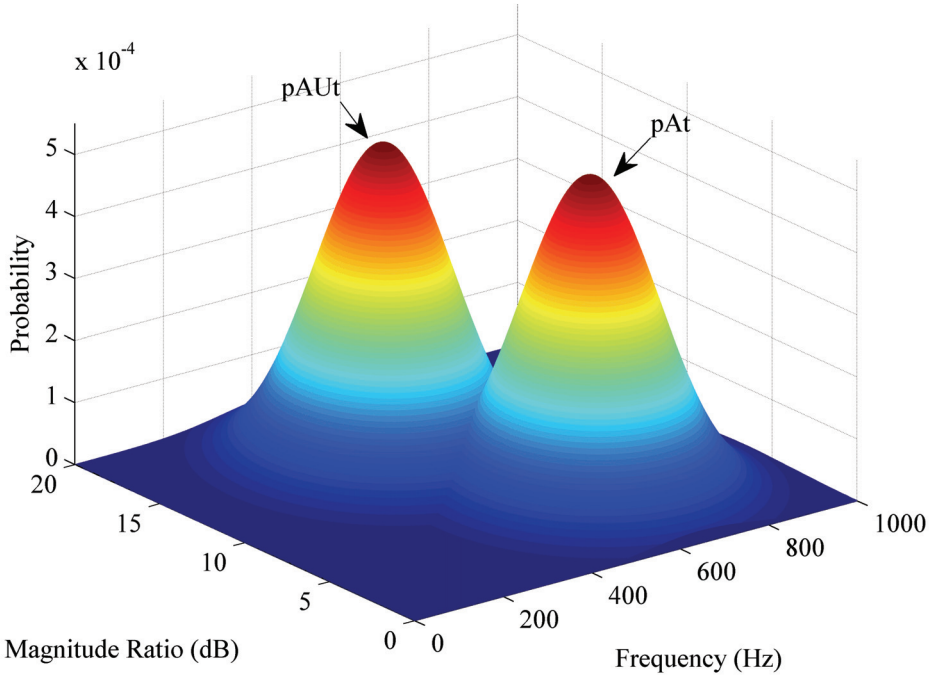
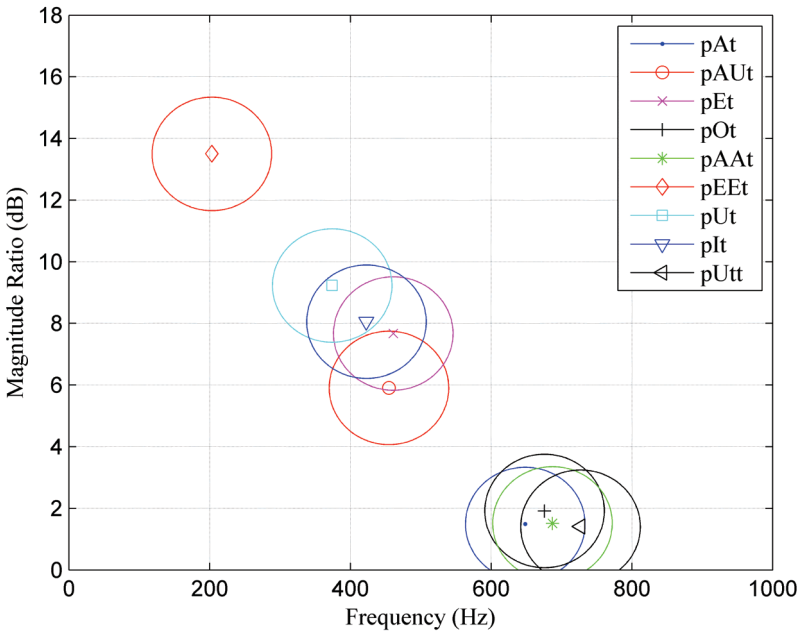


Figure 3.40: Illustration of the MPI model (Svirsky, 2000) classification space for pAt and pAUt, with a standard deviation of 120 Hz and 2.6 dB for the 1st formant frequency and the channel magnitude ratios respectively. The 1st formant frequency and the channel 1 to channel 4 magnitude ratios are displayed on the respective axes as classification features.

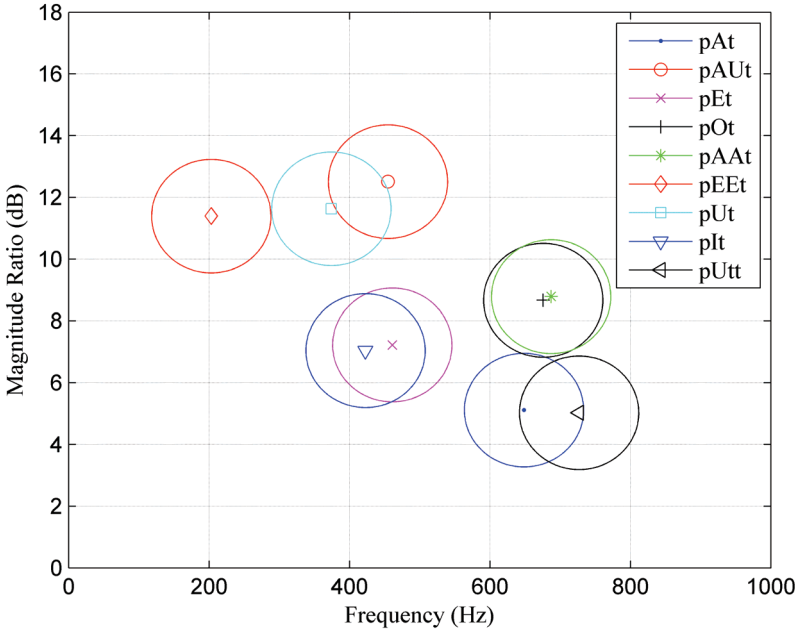
The Gaussian probability function S associated with stimulus E_i is (Svirsky, 2000)

$$
\begin{aligned}
S(E_i) &= S_i(x_1, x_2, ..., x_m) \\
&= \frac{1}{JND_1 JND_2 ... JND_m \left(\sqrt{2\pi}\right)^m} \\
&\exp\left(\frac{-(x_1 - T_{i1})^2}{2JND_1^{\,2}}\right)\exp\left(\frac{-(x_2 - T_{i2})^2}{2JND_2^{\,2}}\right)... \\
&\exp\left(\frac{-(x_m - T_{im})^2}{2JND_m^{\,2}}\right)
\end{aligned}
\tag{3.77}
$$

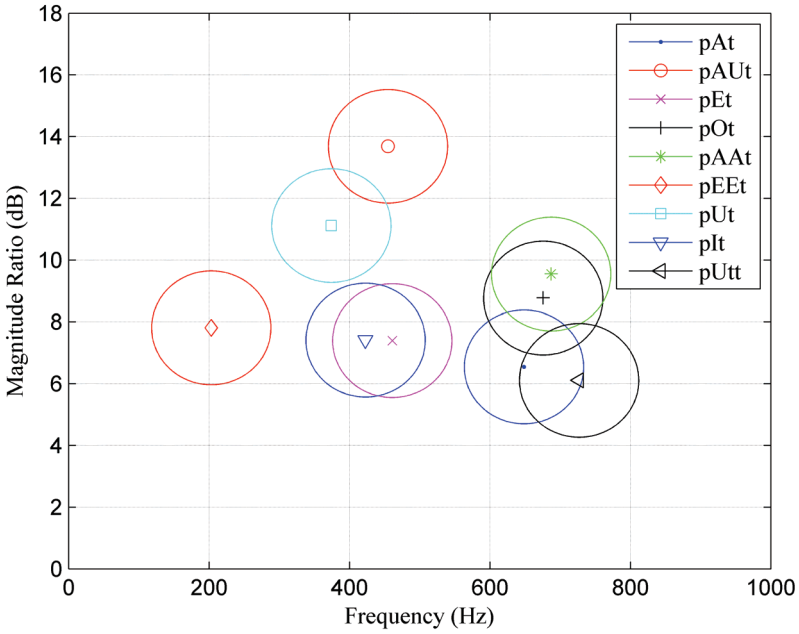
with m the number of dimensions, x_j the value of stimulus E_i along dimension j, T_{ij} the average value of stimulus i over dimension j and JND the JND along dimension j. Because of the multidimensional nature of the classification space it is not possible to show all the features simultaneously graphically. In order to visualize the classification space for the various vowels, Figure 3.41 shows graphs for two features at a time. The marker indicates the mean value of the feature and the contour shows the standard deviation (JND) of that feature for the various vowels.



(a)



(b)



(c)

Figure 3.41: Contour plot of 1st formant frequency vs. channel 1 to channel 2 magnitude ratios (a), 1st formant frequency vs. channel 1 to channel 3 magnitude ratios (b), 1st formant frequency vs. channel 1 to channel 4 magnitude ratios (c). The contour shows the standard deviation and the marker the mean value for the various vowels.

The Euclidean decision rule used for the classifier states that the vowel PDF, which provides the highest probability for a given sample, is the PDF chosen to be associated with the given sample. For example, when considering the two dimensional case represented in Figure 3.41 (a), a sample with a first formant frequency of 200 Hz and a channel 1 to channel 2 amplitude ratio of 14 would be associated with the vowel pEEt. Figure 3.42 illustrates the functional flow of the classification process. The output of the classification process is a decision on which vowel was presented to the algorithm, and this vowel is used as input to the confusion matrix.

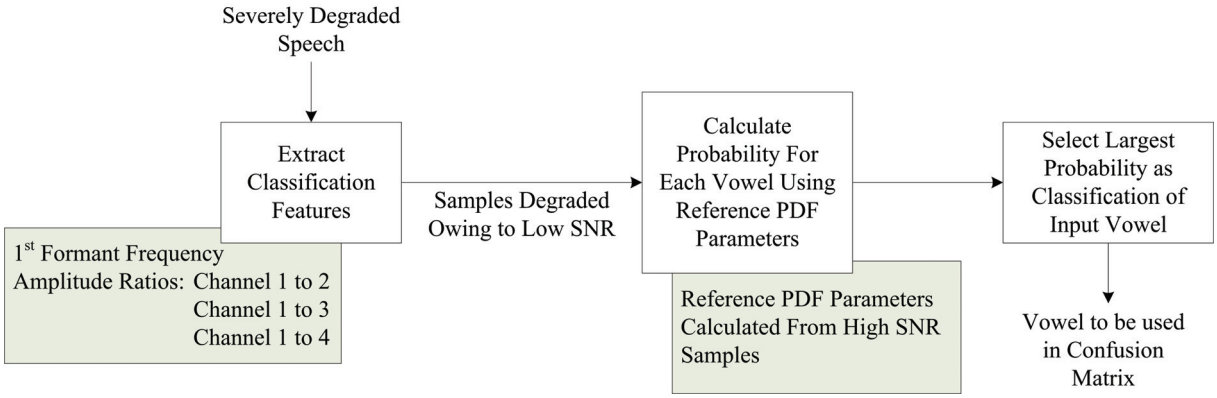


Figure 3.42: Functional block diagram illustrating the classification process.

3.4.1 Evaluation of MPI Model Implementation

This section discusses the results obtained in verifying the implementation of the MPI model by Svirsky (2000). The data used by Svirsky to validate the MPI model came from identification experiments conducted by Dorman et al. (1992) on Ineraid cochlear implantees. The stimuli which Dorman et al. used were /u/, /ɛ/, /æ/, and six conflicting-cue vowels. The conflicting-cue vowels were generated by letting the first formant frequency specify one vowel but the RMS channel amplitudes specify another vowel (refer to Figure 2.2 and Figure 3.36 for information on the four Ineraid channels). The exact values of the first formant frequency and the channel amplitude values are shown in Table 3.5. These values were used by Dorman et al. (1992) for the conflicting-cue experiment, as well as by Svirsky (2000) to validate the MPI model. Svirsky showed that the greatest error predicted by the MPI model was one of 20 percent and that for only four cells (as shown in Table 3.6) the error was between 10 and 20 percent of the confusions as measured by Dorman et al. (1992). The verification of the implemented MPI model attempted to generate the same confusions as those predicted by the MPI model, using

the data as specified by Dorman et al. (1992) presented in Table 3.5. The results of the MPI model validation done by Svirsky, as well as the results of the MPI model implementation verification (of this study) are shown in Table 3.6. In comparing the data from Table 3.6, it can be seen that the largest error is 8 percent, with the correlation between the data by Svirsky and the data of this study at 99.64 percent.

Table 3.5: The values of the cues used by Dorman et al. (1992) to specify the vowels and conflicting-cue vowels used in the experiments on Ineraid cochlear implantees. Svirsky (2000) used these values for the validation of the MPI model. These values were also used in this study to verify the implementation of the MPI model. For the entries with the form / / → / /, the first member indicates the vowel whose formant frequency was used and the second member of the expression indicates the vowel whose channel amplitude profile was used as conflicting-cue.

	Channel RMS Values (dB)				First Formant Frequency, F1 (Hz)
Stimulus	Channel 1	Channel 2	Channel 3	Channel 4	
/u/	15	7	6	1	350
/ɛ/	10	6	12	9	500
/æ/	11	10	9	6	700
/u/ → /ɛ/	10	6	12	9	350
/u/ → /æ/	11	10	9	6	350
/ɛ/ →/u/	15	7	6	1	500
/ɛ/ → /æ/	11	10	9	6	500
/æ/ → /u/	15	7	6	1	700
/æ/ → /ɛ/	10	6	12	9	700

Table 3.6: The table shows the percentage response as a function of channel amplitude profile. The data obtained by Svirsky (2000) in the evaluation of the MPI model, and the evaluation of the implemented MPI model of this study are shown. Differences in comparing the data may be due to a different number of Montecarlo tokens used to generate the data.

		Response		
	Stimulus	u	ε	æ
Data obtained by Svirsky (2000) in the evaluation of the MPI model.	/u/	99	0	1
	/ε/	0	95	5
	/æ/	0	5	95
	/u/ → /ε/	0	98	2
	/u/ → /æ/	9	58	33
	/ε/ →/u/	98	0	2
	/ε/ → /æ/	2	31	67
	/æ/ → /u/	88	0	12
	/æ/ → /ε/	0	64	36
Data obtained in this study by evaluating the implemented MPI model by Svirsky (2000).	/u/	100	0	0
	/ε/	0	93	7
	/æ/	0	7	93
	/u/ → /ε/	0	99	1
	/u/ → /æ/	5	54	41
	/ε/ →/u/	99	0	1
	/ε/ → /æ/	1	29	70
	/æ/ → /u/	93	0	7
	/æ/ → /ε/	0	72	28

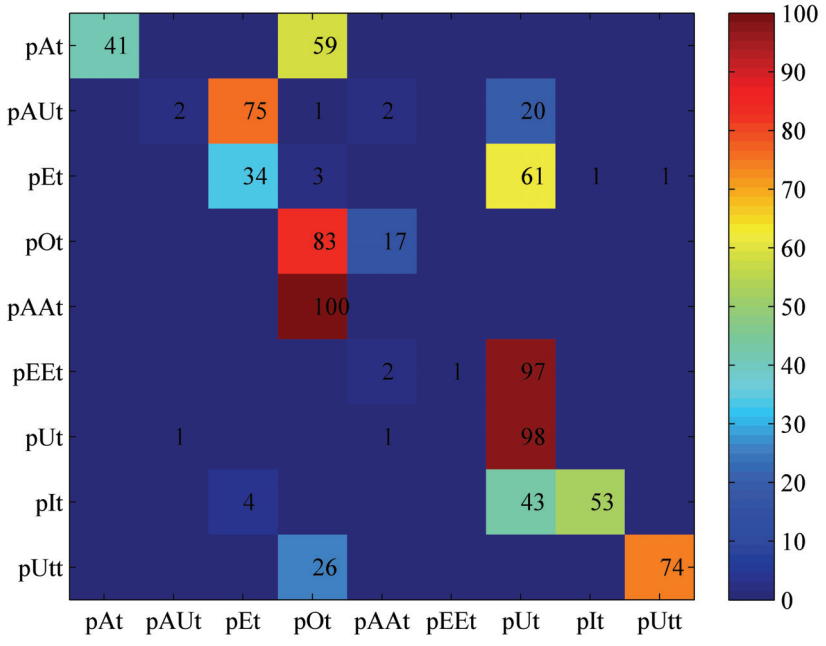
3.5 Summary

This chapter documented the details of the various processing steps which were used for the speech enhancement, the cue estimation and finally the classification technique. The aim of the speech enhancement was to suppress the noise in the degraded speech signal sufficiently to allow for successful speech cue estimation. The speech enhancement was achieved using a Kalman filter. The respective speech cues were estimated for a vowel in a CVC syllable. To identify the location of the vowel in the syllable a CFAR detector was used. This detector could successfully locate the vowel down to a SNR of -10 dB. Once the vowel was located, the same cues and classification method were used as those proposed by Svirsky (2000). The output of the classifier can now be used to produce a confusion matrix for the SNR of interest, which allows for the generation of a graph of recognition performance vs. SNR. These results will be discussed in the following chapter.

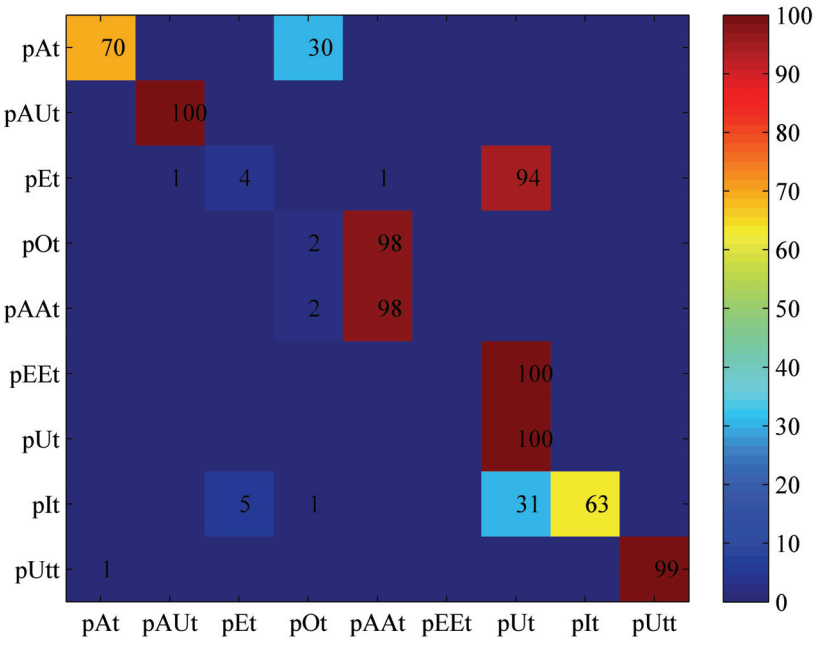
CHAPTER 4 RESULTS

Outputs from the various signal-processing stages of the speech enhancement and cue estimation algorithm were shown in chapter 3. This chapter will show the results generated by using the speech enhancement and cue estimation algorithm together with the MPI model by Svirsky (2000). The output of the MPI model was measured in the form of confusion matrices, which were used to generate a graph of percentage correct recognition vs. SNR. This chapter also discusses the information transmission analysis that was performed on the confusion matrices in order to quantify the robustness of the selected cues to noise.

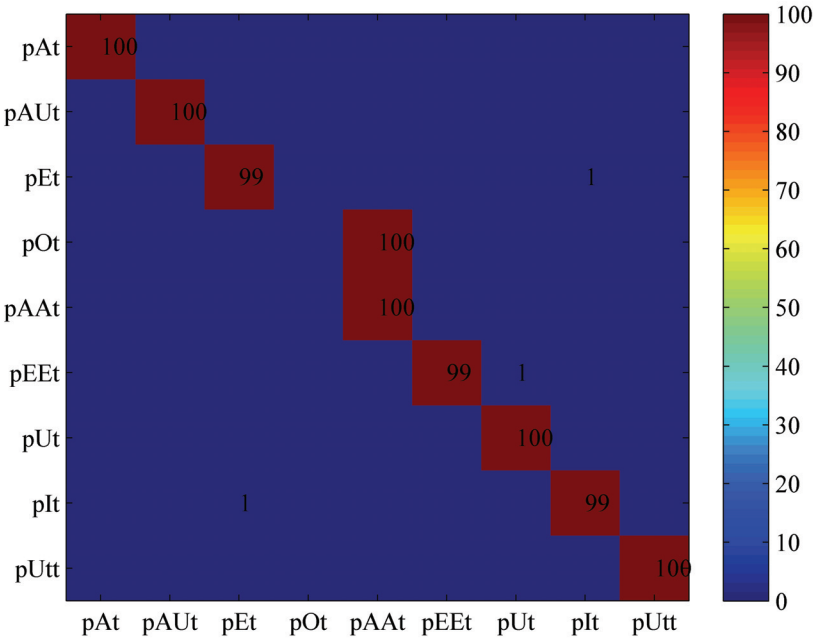
For a specific SNR the classifier can be expected to make a certain amount of incorrect classifications, which will result in apparent confusions. It can be expected that the percentage of correct classifications will increase as the SNR increases. Figure 4.1 shows examples of three confusion matrices for SNRs of -10 dB, -5 dB and 0 dB. For these examples each vowel was presented to the algorithm 100 times. The diagonal of the graph is associated with correct classifications and the number in any specific block (and the colourbar) is the number of times a specific vowel was selected by the model.



(a)



(b)



(c)

Figure 4.1: Confusion matrix for (a) -10 dB SNR with 43% correct classification, (b) -5 dB SNR with 60% correct classification, and (c) 0 dB SNR with 89% correct classification.

The performance of the model in terms of the percentage of vowels correctly discerned, given a specific SNR, is used as validation of the model when compared to human listener performance. It is also of interest to investigate the gain in performance due to the signal enhancement achieved by the Kalman filter. Figure 4.2 shows the percentage of vowels discerned correctly for various SNRs, with and without signal enhancement.

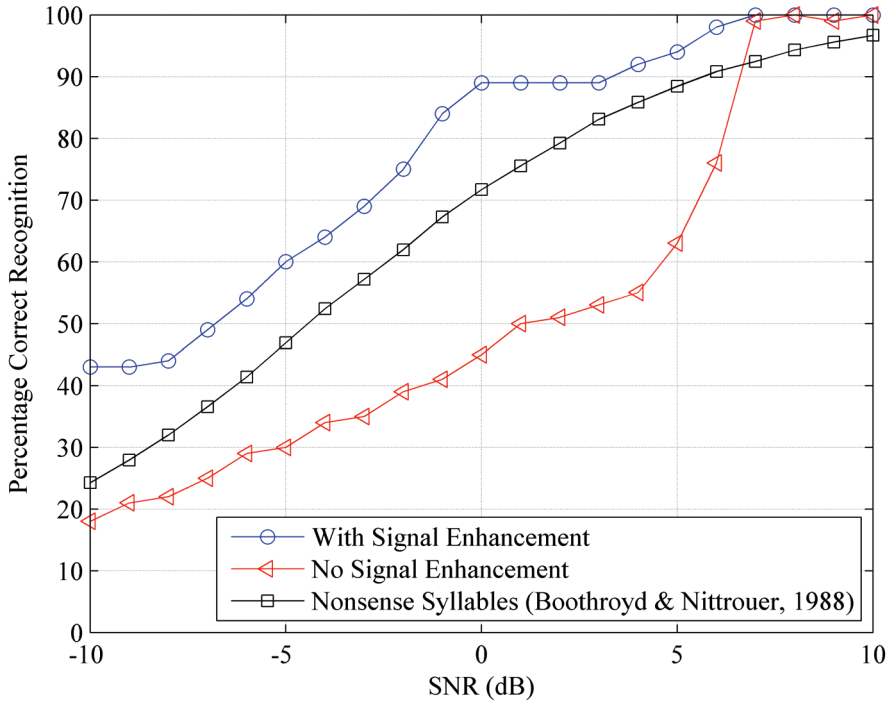


Figure 4.2: Algorithm classification performance vs. SNR with and without signal enhancement.

Figure 4.2 also shows the data from Boothroyd and Nittrouer (1988), which was obtained during listening experiments on normal hearing listeners for nonsense syllables. The reason for using nonsense syllables is to remove the added perceptual advantage of context. Nonsense syllables are the best direct comparison with the vowel classification used in this study, since the vowels were specifically extracted from the syllables by the voicing detection processing. The SNR range that they investigated was the same SNR range as used for this study during cue estimation and classification. However, the noise used by Boothroyd was speech-shaped noise and not white noise. From Figure 4.2 is can be seen that although the results are not exactly the same, the trend of the Boothroyd data and the data from this study (with signal enhancement) is similar.

4.1 Information Transmission Analysis of Confusion Matrix

To investigate the robustness of the selected cues to additive noise, an information transmission analysis was performed on the predicted confusion matrices. This analysis also investigated which cues were predominantly responsible for the recognition performance degradation, as shown in Figure 4.2. Furthermore, this analysis serves as an example of the type of analysis that is now possible at SNRs as low as -10 dB using the speech-enhancement and cue-estimation algorithm.

An information transmission analysis measures how well the respective cues were conveyed (transmitted) to the listener by measuring which vowels are recognised, as captured in a confusion matrix. An example would be that if speech should be severely degraded owing to the intentional or unintentional injection of noise in a transmission channel, cues that are normally used to recognize vowels may be masked. Analytical tests may be used to characterize to what extent a specific cue was degraded or enhanced owing to a specific process. These tests are often used in the context of hearing impairment and CIs, where it is of value to know which cues do not reach the auditory system of the listener (Van Wieringen & Wouters, 1999). In order to investigate the nature of vowel confusions further, techniques first introduced by Miller and Nicely (1955) can be used. Using information transmission analysis (Miller & Nicely, 1955; Wang & Bilger, 1973) (also referred to as feature information transmission analysis) confusion matrices are analysed in order to calculate the amount of information transmitted by a specific cue. To this end, each vowel is first classified into one of several categories for each cue. An example is shown in Table 4.1 (Van Wieringen & Wouters, 1999) for three cues: vowel duration, first formant frequency (F1) and the second formant frequency (F2).

Table 4.1: Example of classification of vowel features from (Van Wieringen & Wouters, 1999). "Duration" was classified into two categories (shorter and longer than 200 ms). Both F1 and F2 were divided into three categories. For example, F2: category 1 was less than 1000 Hz, category 2 was 1 kHz to 2 kHz and category 3 was more than 2 kHz.

	u	y	i	o	e	a	ə	ɪ	ɔ	ɑ
Duration	1	1	2	1	1	1	2	2	2	2
F1	1	1	1	2	2	3	2	2	3	3
F2	1	3	3	1	3	2	2	3	1	2

Once classification has been done, information transmission analysis proceeds using the confusion matrix as input. For each cue separately, the confusion matrix is collapsed into the number of categories available for that cue. For example, to determine the information transmitted on F2 in the example above, the confusion matrix will be collapsed into a stimulus-response matrix with three categories (category 1 <=1000 Hz, category 2 = 1000-2000 Hz, category 3 >=2000 Hz), and all the vowels that fit into a particular category are pooled together.

The total transmitted information (in bits) is calculated as follows (Van Tasell et al., 1987).

$$r = -\sum_{ij} \frac{n_{ij}}{n} \log_2 \frac{(n_i / n)(n_j / n)}{n_{ij} / n} \tag{4.1}$$

with i an index ranging from 1 to the number of categories for a specific cue (i range from 1 to 2 for duration and 1 to 3 for F1 in the example), j an index ranging from 1 to the number of categories for a specific cue (j range from 1 to 2 for duration and 1 to 3 for F1 in the example), n_i the frequency of the stimulus (sum of all the times a specific category was presented), n_j the frequency of the response (sum of all the times a specific category was selected, irrespective of

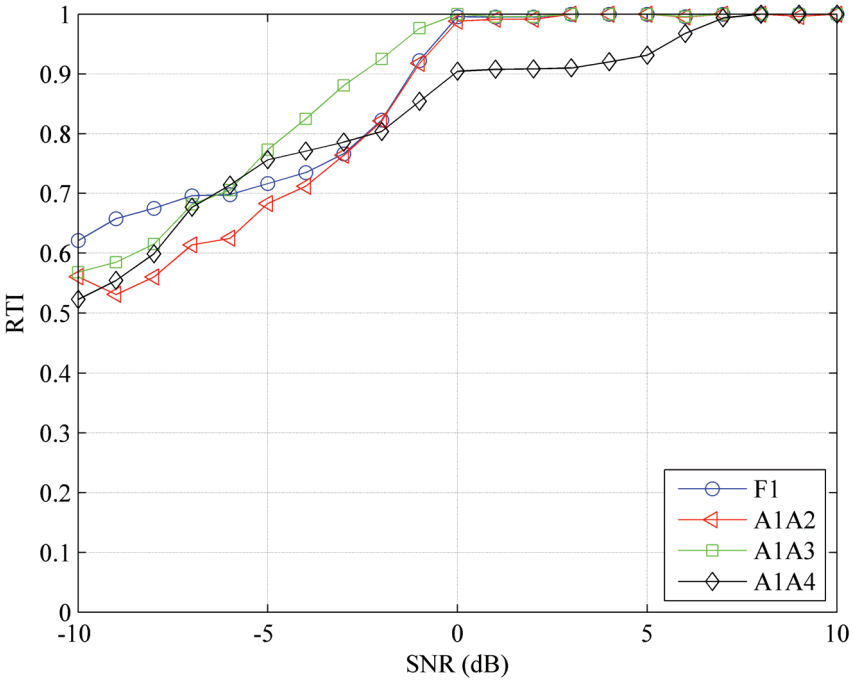
the stimulus), n_{ij} the frequency of the joint occurrence of a particular stimulus-response pair (sum of all the times a specific category was presented and that same category was selected) and n the total number of stimuli presented. If the confusion matrix is collapsed for each feature or cue, the information transmitted about that cue is calculated as follows (Van Tasell, Soli, Kirby, & Widin, 1987):

$$r_{\max} = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n} \tag{4.2}$$

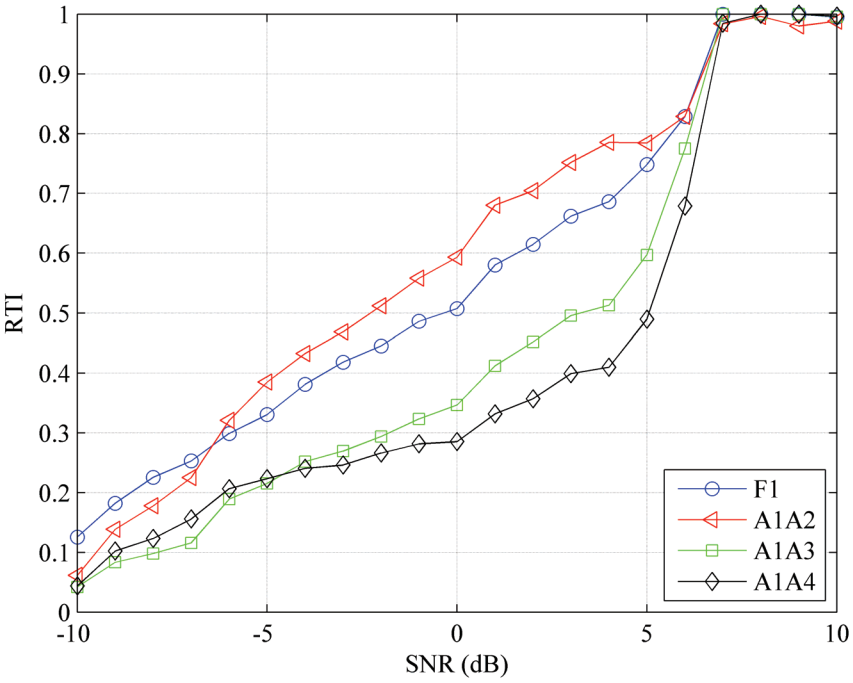
with $r_{\max}$ the maximum available information. Using (4.2) the relative transmitted information (RTI) for a specific cue can be calculated as:

$$r_{rel} = RTI = \frac{r}{r_{\max}} \tag{4.3}$$

with r_{rel} the relative transmitted information. If the RTI for a specific cue is 0, the particular cue could not be perceived by the listener and was thus lost/masked owing to intentional or unintentional noise. Using the process described above, an RTI analysis was performed on the confusion matrices used to generate Figure 4.2. For the first formant frequency ten categories were chosen, starting at zero Hz and incrementing in 100 Hz steps. All of the channel amplitude ratios had the same 15 categories, starting at zero dB and incrementing in one dB steps. The results of the RTI analysis, with and without speech enhancement, are shown in Figure 4.3.



(a)



(b)

Figure 4.3: Relative transmitted information for each of the cues used in the multivariate Gaussian classification, for (a) speech enhancement using a Kalman filter and (b) no signal enhancement.

In comparing the results of Figure 4.3 with the recognition performance shown in Figure 4.2, it can be seen that the masking of specific cues contributes to lower perception performance at specific SNRs.

From the various results shown in this section, it is clear that automated speech cue estimation can be achieved. The estimated cues can be used as inputs to a perception-prediction model with the aim of generating recognition performance graphs. The recognition performance graphs are derived from confusion matrices. An additional advantage of having the confusion matrices available is that information transmission analysis can be done on the confusion matrices in order to investigate the robustness of specific cues to additive noise further. Insight can also be gained into which cues contribute most to vowel recognition when speech is severely degraded. For example, it can be seen from Figure 4.3 that the amplitude ratio of the first bandpass channel to the fourth bandpass channel (A1A4) is more susceptible to the masking effect of the noise than the other three cues.

CHAPTER 5 DISCUSSION

The focus and area of contribution for this study was speech enhancement and cue estimation of severely degraded speech signals. The evaluation of the proposed techniques did, however, require the implementation of a speech perception model and the use of the perception model led to some interesting insights that will be discussed in this section. This study classifies a signal as severely degraded when the power of the additive noise is larger than the power of the speech, resulting in a negative SNR. The SNR region from -10 dB to -3 dB was investigated. Work on perception prediction by Remus and Collins (2004a; 2004b) did show perception prediction performance for SNRs of -2 dB, although they did not address the estimation of the features required for classification. Figure 5.1 shows the processing steps for the proposed signal enhancement and cue estimation algorithm, which will be discussed in this section. The severely degraded speech was the input to the algorithm, with the focus on the first and second processing steps of speech enhancement and cue estimation, as shown in Figure 5.1.

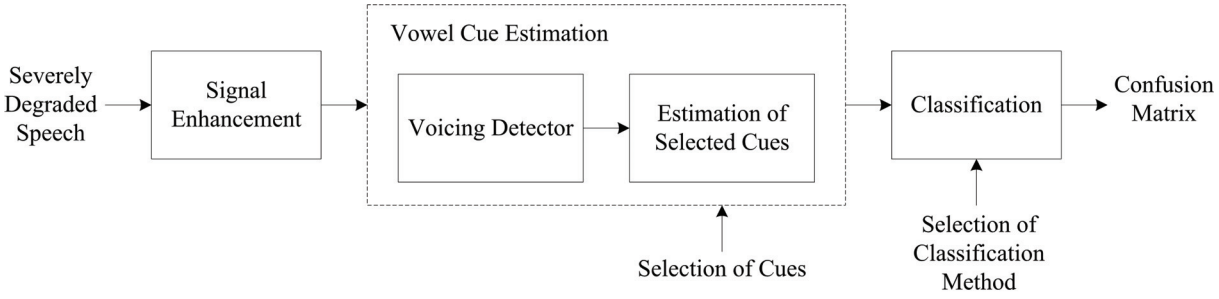


Figure 5.1: Processing steps required for automatic speech enhancement and cue estimation to enable prediction of perception performance.

The classification method and the set of cues (Svirsky, 2000) which were used were an example of the application of the first two processing steps. Other selections are possible and future work may entail investigating the estimation of other cues and the evaluation of other classification methods for normal hearing listeners and CI users. For example, Ainsworth (1971) found that the importance of the vowel duration increases if the vowel can easily be confused with other vowels when only the first and second formant frequencies are analysed. The first two processing steps of Figure 5.1 can be used to enable an investigation into whether

this observation also holds for severely degraded speech. An example of another classifier, which can be used in conjunction with the speech enhancement algorithms, is HMMs (Rabiner, 1989). HMMs are used for classification in ASR (Alwan, Narayanan, Shen, & Strope, 1995) and perception prediction (Remus & Collins, 2004a; Remus & Collins, 2004b). Cepstral coefficients (the inverse Fourier transform of the logarithm of the power spectrum (Rabiner & Schafer, 1978)) are often used as the classification features required by the HMMs (Alwan, Narayanan, Shen, & Strope, 1995). The investigated speech-enhancement technique would allow for the evaluation of HMMs used in ASR or perception prediction for severely degraded speech by suppressing the noise in the power spectrum of the signal and thus enhancing the cepstral coefficients. Another example of a perception prediction classification method that can be used with the proposed speech-enhancement techniques is the token envelope correlation method of Remus and Collins (2004a; 2004b). This method uses the discrete envelope of speech signals as input to a correlator, which is used as the classification method. The proposed speech-enhancement technique would make the use of this classification method possible at negative SNRs, since the envelope of the speech signal would be recovered by suppressing the additive noise.

5.1 Speech Enhancement

Speech enhancement was investigated in the presence of white Gaussian noise. To make this work more applicable to the various types of noise that may be encountered in everyday circumstances, additional work can be done to implement a Kalman filter that estimates the speech signal in the presence of noise that is not white Gaussian. Work by Gannot (1998) and Gibson, Koo and Gray (1991) can be used to extend the Kalman filter to allow for an input disrupted by coloured noise.

A drawback of the implemented Kalman filter is that a section of the input signal should not contain any speech in order to allow for the estimation of the noise statistics required by the filter. If such a segment cannot be identified, the quality of the noise suppression will be degraded. The non-linear noise suppression that the Kalman filter provides (refer to Figure 3.5) may be due to increased difficulty with estimating the noise statistics, as the noise is increasingly masked by the speech signal as the SNR increases. Parameters in the Kalman filter, which can influence the performance of the filter, are: the length of the speech segment

(frame) used for the LPC coefficient estimation, the order of the LPC estimation, the method used to estimate the variance of the noise added to the speech (the observation noise) and the number of EM repetitions. The number of EM repetitions used in the Kalman filter was optimized to give maximum signal enhancement (see Figure 3.4). The frame length and the LPC order were, however, not optimized. The values for these parameters were based on values used in literature (Du & Driessen, 1991; Lim & Oppenheim, 1978; Makhoul, 1975). The perception prediction performance can, however, be influenced by changing the order of the LPC used in the Kalman filter, while changes in the frame length do not have a significant influence on the signal-processing gain of the Kalman filter. Figure 5.2 and Figure 5.3 show the SNR improvement that can be achieved by the Kalman filter for various LPC orders and frame lengths respectively (refer to section 3.2.1 regarding the method used to calculate the SNR improvement). From Figure 5.2 it can be seen no particular LPC order provided the greatest SNR gain over the entire SNR region from -10 dB to 10 dB, and thus the choice of LPC order depends on the SNR region of interest. The SNR region of interest may be determined by the requirement to fit perception prediction model data to data from listening experiments. For example, if a higher recognition performance was required for the SNR region of -10 dB to -2 dB, an LPC order of 18 may be more appropriate than an LPC order of 12. No significant SNR improvement was achieved by increasing the LPC order beyond 30, which resulted in an SNR improvement of 12 dB for an input signal SNR of -10 dB (refer to Figure 5.2). In terms of maximizing the SNR improvement over the entire SNR region of -10 dB to 10 dB, there is no outright optimum choice of LPC order, but an LPC order of 14 may be a better choice than the 10^{th} order LPC used for the proposed algorithm. Figure 5.3 shows that the frame length does not significantly influence the SNR improvement and that the choice of a 50 ms frame length is acceptable (note that the y-axis scale for Figure 5.3 is not the same as for Figure 5.2).

The performance of the implemented speech-enhancement processing in terms of the technique and the parameters of that technique was not optimised, according to some criteria. However, from the results in Figure 4.2 it can be seen that a form of speech enhancement is required for perception prediction of severely degraded speech. The perception-prediction performance of the MPI model by Svirsky (2000) is closer to the data from Boothroyd and Nittrouer (1988) when speech enhancement is performed (this observation will be discussed in section 5.3). There are various techniques for the enhancement of speech signals (refer to section 3.2); however, based on the work of Wolpert, Ghahramani and Jordan (1995), as well as Watkins and

Paus (2004), a Kalman filter was used for the speech enhancement. This was done in an attempted to keep the structure of the speech-enhancement algorithm related to literature on human auditory perception (Watkins & Paus, 2004).

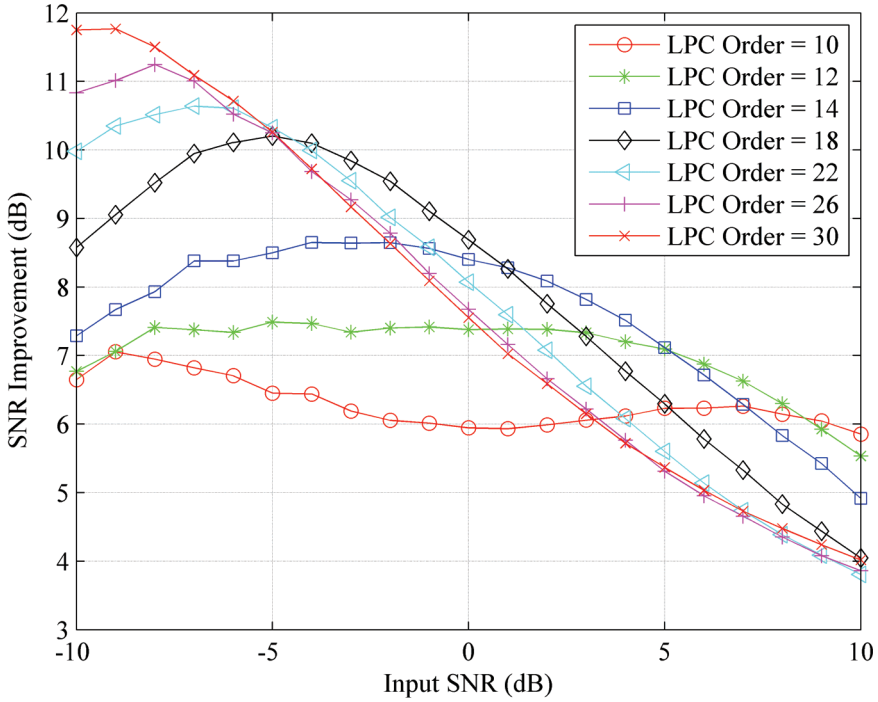


Figure 5.2: SNR improvement (defined in section 3.2.1) using a Kalman filter for various LPC orders. The number of EM repetitions was set to 1 and the frame length was 50 ms. A synthesized vowel input was used with F1 at 750 Hz and F2 at 1050 Hz. White Gaussian noise was added to the input.

5.2 Cue Estimation

The advantage of using a constant false alarm detector for the purpose of voice detection is that the detector can be tailored to be more sensitive by allowing more false alarms. This does, however, place an additional processing burden on the techniques used to filter through all the detections to decide which detections are valid. The sensitivity of the detector and thus the specific probability of a false alarm will have an influence on the perception prediction performance of the classification algorithm. The detection of a voiced section in the

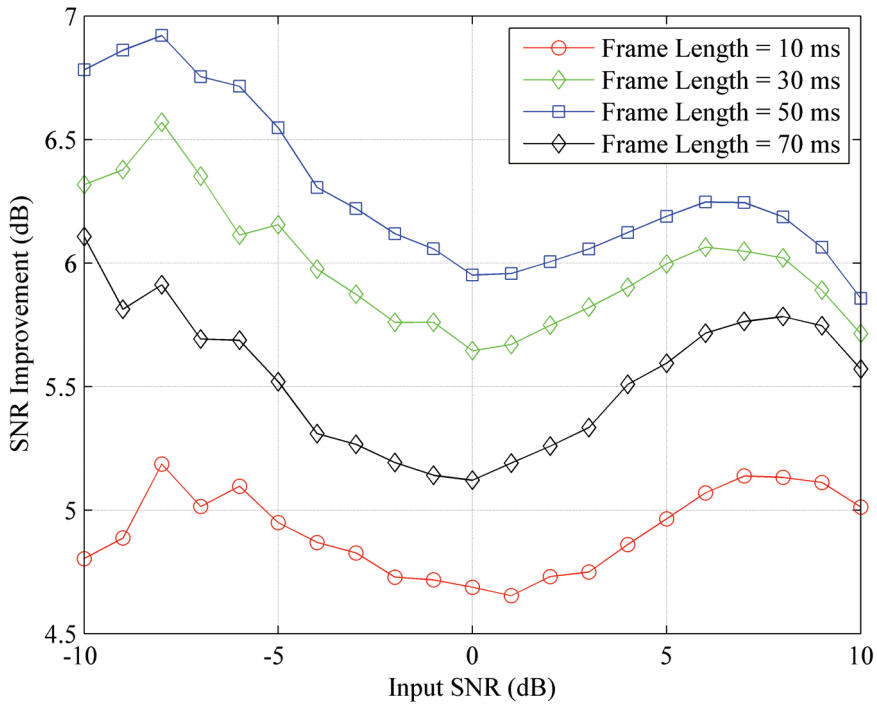


Figure 5.3: SNR improvement using a Kalman filter for various frame lengths. The number of EM repetitions was set to 1 and the LPC order was 10. A synthesized input was used with F1 at 750 Hz and F2 at 1050 Hz. White Gaussian noise was added to the input.

syllable is a requirement in order to allow for the correct feature estimation of the vowel. If the correct voiced section of the syllable cannot be detected, a random section (due to a false alarm) will be selected for the feature estimation. This will result in incorrect feature estimation and consequently wrong classification of the vowel. It was not attempted to modify the probability of a false alarm in order to adjust the prediction of recognition performance. The focus was on investigating the usability of a CFAR detector as a voicing detector.

The input to the CFAR detector is a spectrogram and the resolution with which a detection can be made is determined by the parameters used to generate the spectrogram. If greater frequency resolution is required the length of the fast Fourier transform (FFT) must be increased, and if the time domain resolution has to increase, the overlap between successive FFT frames must be increased. The increased resolution does, however, come at the price of an increased processing load and therefore a trade-off must be made with regard to algorithm execution time and the required detection resolution. An additional advantage of the CFAR detector is that no a priori

knowledge of the formant frequencies are required in order to isolate the frequency components of the spectrogram where the formant frequencies occur. For example, the formant frequency tracking filters of Mustafa and Bruce (2006) require an initial formant frequency estimate, which enables the formant tracking filters to lock onto the formant frequency. However, if this initial formant frequency estimate is not near to the respective formant frequency, as a result of speaker variation, the tracking filter will not acquire and track the formant frequency. The CFAR technique is far more processing-intensive than, for example, a constant threshold detector. The two parallel CFAR detectors do, however, increase the processing load further. Two CFAR detectors were deemed necessary because no assumptions about the input formant frequency structure were made.

In summary then, the CFAR detector was successfully applied as a voicing detector for severely degraded speech, at an SNR as low as -10 dB, to estimate the location of the vowel in a CVC syllable. The combination of speech enhancement and voicing detection techniques forms an algorithm that enables the estimation of speech cues as required by perception prediction models. Both these processing steps are automatic and require no a priori knowledge of the input speech signal. These remarks relate to the primary research question.

5.3 Vowel Classification

The specific set of speech cues and the chosen classifier were used to provide an example of the application of the speech enhancement and cue estimation for severely degraded speech. Further work can, for example, be done to evaluate the classifiers described by Remus and Collins (2004a; 2004b) at SNRs as low as -10 dB, using the signal enhancement technique presented. It is of interest to note the similarities between the perception-prediction results obtained with the multivariate Gaussian classifier as suggested by Svirsky (2000) and the results obtained by Boothroyd and Nittrouer (1988). Boothroyd and Nittrouer performed recognition tests using CVC words and nonsense CVC syllables on normal hearing listeners. When analyzing the results of Boothroyd and Nittrouer, it is most appropriate to consider the nonsense CVC syllables. The nonsense syllables were used to remove the added perceptual advantage of context, which would be applicable when listening to CVC words. The context that is provided by the lexicon from which words are drawn is more subtle (Boothroyd & Nittrouer, 1988). For example, it has been shown that real words that are presented in isolation are recognized more

easily than nonsense syllables, and that words with a high frequency of occurrence are more easily recognized than words with a low frequency of occurrence (Hirsh, Reynolds, & Joseph, 1954; Howes, 1957; Pollack, Rubenstein, & Decker, 1959; Savin, 1963). Boothroyd and Nittrouer investigated recognition performance for SNRs as low as -10 dB (Figure 5.4). The additive noise used in the Boothroyd and Nittrouer study was, however, not white noise but spectrally shaped noise with the intent to have an equal masking effect for all frequencies. Pickett (1957) showed that shifts in vowel confusions can occur if the spectrum of the noise changes. This observation by Pickett makes a direct comparison with the results of Boothroyd and Nittrouer difficult. For example, even with shifts in vowel confusions resulting in a specific vowel being confused with two different vowels in the respective studies, the percentage of vowels correctly discerned may still be very similar. Without knowing the particular confusions in the Boothroyd study, the cues that are transmitted cannot be compared. At best then, only the trends in vowel recognition performance (as a function of SNR) are similar in the present study and the Boothroyd and Nittrouer study. The high degree of correlation (greater than 99 % as shown in Figure 5.5) between the results does suggest that the MPI model of Svirsky can possibly be used as perception-prediction model for normal hearing listeners. This observation should be investigated in future research. If the percentage recognition performance had to be modified for the MPI model of Svirsky (2000), additional cues, such as duration and the second formant frequency, could be considered. Changes to the JNDs of the respective cues would also adjust the model's performance. Table 5.1 shows a summary of the various parameters of the proposed algorithm that can be adjusted in order to modify the performance of a perception-prediction model.

Table 5.1: Parameters that can modify the performance of a perception-prediction model.

Processing Step	Parameter	Comments
Speech Enhancement	LPC order	No particular LPC order provides the greatest SNR improvement for the entire SNR range from -10 dB to 10 dB. The specific choice of LPC order will be determined by the modification required to the predicted recognition performance. Refer to Figure 5.2.
	Frame length	Referring to Figure 5.3, it can be seen that the frame length does not significantly influence the SNR improvement that can be achieved with the Kalman filter. It is, however, assumed that the speech is stationary on a short-time basis and the frame length should not be so long that this assumption does not hold.
	EM repetitions	Repetition of the EM processing step does not guarantee continued increases in SNR improvement. One to two repetitions are recommended. Refer to Figure 3.5.
Cue Estimation	Probability of false alarm for CFAR detector	As the probability of false alarm decreases, the CFAR detector also becomes less sensitive to possible voiced sections of the input speech signal.
	Spectrogram resolution	If the time resolution is too coarse, the voiced section of the syllable may not be estimated correctly and unvoiced sections of the syllable or noise may influence the cue estimation accuracy. If spectral resolution is too coarse, spectral differences between vowels may not be identified.
Perception Prediction	Choice of cues	The robustness of speech cues against the degrading effects of additive noise differs from cue to cue. Refer to Figure 4.3. The choice of cues will thus influence the predicted perception for severely degraded speech.
	Choice of perception prediction model	A specific perception prediction model will have its own set of parameters that will influence the perception prediction. For the MPI model of Svirsky (2000) this parameter is the JND for the respective cues.

In a study by Parikh and Loizou (2005), in which they investigated the influence of noise on vowel and consonant cues, they also reported on listening tests they performed on normal hearing listeners. They focussed on the first and second formant frequencies as cues and found that listeners must rely on relatively accurate first formant frequency information, along with partial second formant frequency information. This result supports the choice made by Svirsky (2000) to use the first formant frequency as one of the classifier features as well. The SNRs evaluated by Parikh and Loizou were -5 dB, 0 dB, 5 dB and 10 dB, for speech-shaped noise and multi-talker babble. The perception results of the experiment by Parikh and Loizou, together with the results from this study and that of Boothroyd and Nittrouer, are shown in Figure 5.4. Figure 5.5 shows the linear correlation between the perception prediction data of the various studies. The x-axis is the predicted recognition performance using the presented automated speech enhancement and cue estimation algorithm with the MPI model of Svirsky. The y-axis shows the recognition performance data of the Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005) studies. The high degree of correlation between the data (higher than 95 % in all cases) suggests that normal hearing listeners, as used for the Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005) studies, may use the same cues as those used by the MPI model of Svirsky (2000) (which uses the cues available to Ineraid CI users), in low SNR conditions. This observation should be investigated in future research. The large difference between the data using no signal enhancement and the other data sets does suggest that in order to do automatic feature estimation of noise-degraded speech, some form of signal enhancement is required.

When considering the results shown in the RTI analysis (refer to Figure 4.3) it is important to note that the respective speech cues have different robustness to the degrading affects of the additive noise. For example, more information is removed owing to the masking effect of the noise for any of the amplitude ratio cues than for the first formant frequency cue at -10 dB SNR. It can also be seen that the amplitude ratio of the first channel to the fourth channel is the cue that is most susceptible to noise. This may be due to the natural roll-off of a speech spectrum. As the SNR decreases, the higher frequency spectral components would be masked by the noise before the lower frequency spectral components. For example, the amplitude of the fourth channel would be distorted at higher SNRs compared to the amplitude of the second channel. From Figure 4.3 it can be seen that as the SNR increases, the amplitude ratio of the first channel to the fourth channel is largely responsible for the sharp increase in recognition performance,

around 6 dB, when no signal enhancement is used. This observation is consistent with the findings of Loizou and Poroy (2001), who determined that CI listeners using a six-channel CIS (Continuous Interleaved Sampling) processor needed at least 4 dB of spectral contrast to identify vowels.

In summary then, although the main focus was automatic signal enhancement and cue estimation, it is interesting to note that the perception prediction model of Svirsky (2000), evaluated on Ineraid cochlear implant users, may provide some insight into the cues used by normal hearing listeners when speech is severely degraded by noise. Amplitude ratios of spectral regions do seem to be used by normal hearing listeners when almost all formant frequency information is masked by noise. In performing automatic cue estimation for severely degraded speech, speech enhancement is required to improve spectral contrast. The parameters dictating the performance of the speech-enhancement technique can be adjusted (section 5.1) in order to modify the predicted recognition performance of a selected model. The manner is which these parameters are modified will, however, be iterative since the parameters of the signal enhancement influencing the final predicted recognition performance will depend on the selected set of cues and the perception prediction model. The use of a Kalman filter and a CFAR detector, used for isolating a vowel in the spectrogram of severely degraded speech, did enable the automatic estimation of speech cues as required by the perception-prediction model, without any a priori knowledge of the input signal. This statement relates to the primary research question. The predicted perception performance also showed a high correlation with available published data, which relates to the secondary research question.

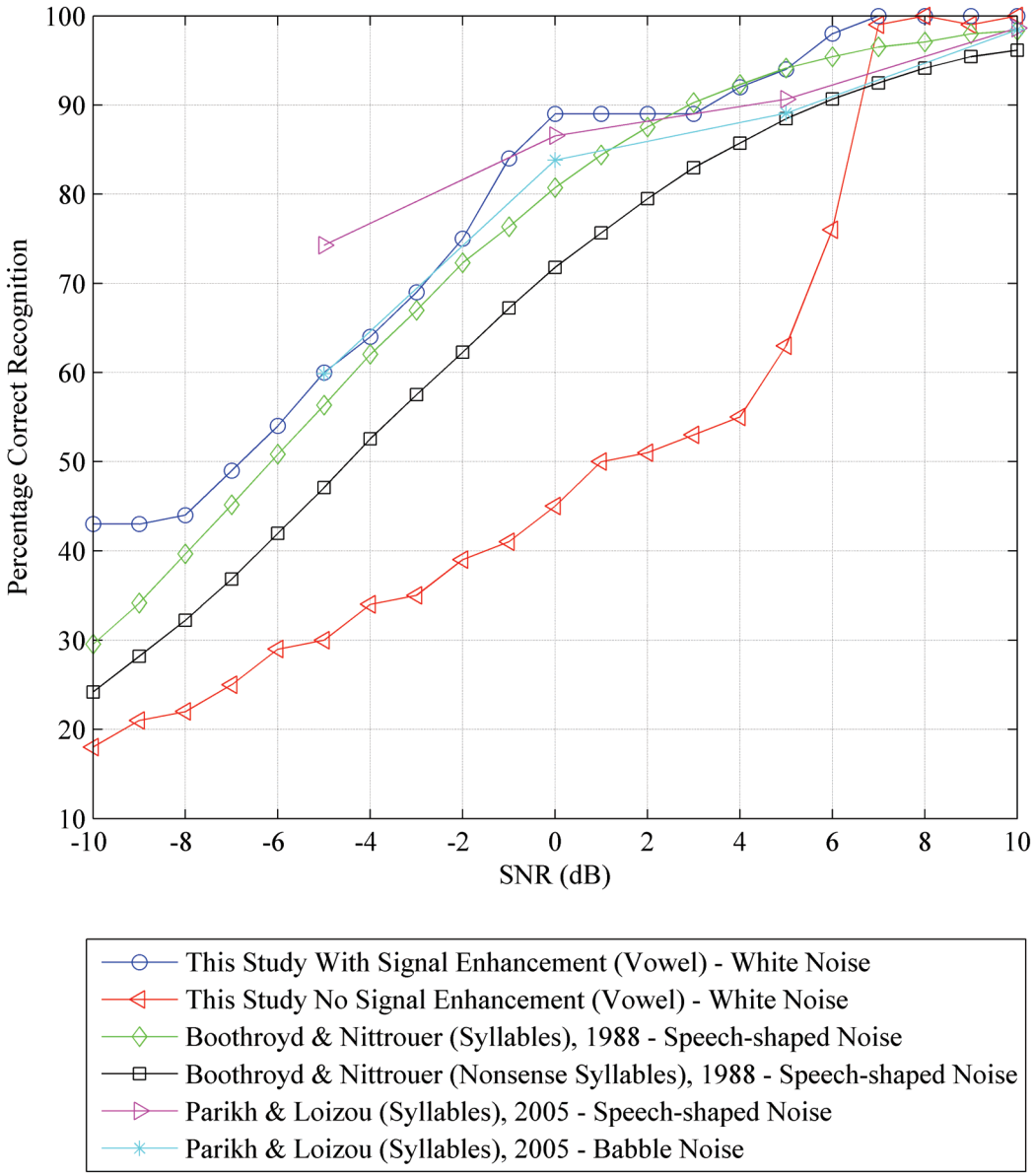


Figure 5.4: Recognition performance obtained using the speech enhancement and cue estimation with Svirsky's (2000) perception prediction model. The figure also shows data from Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005) for vowel recognition experiments on normal hearing listeners.

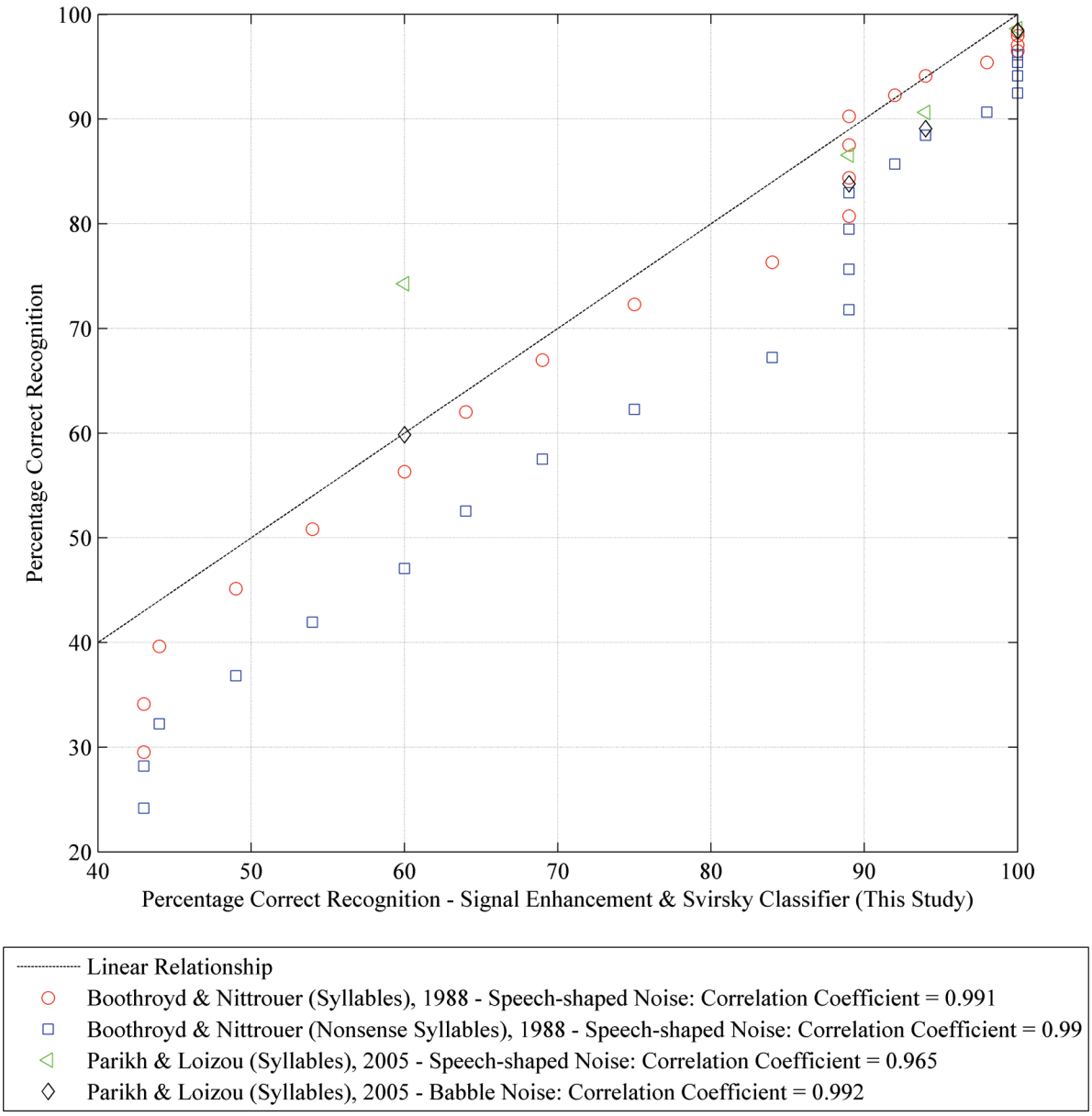


Figure 5.5: A scatter plot showing the degree of linear correlation between the data of this study and those of Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005). The x-axis is the percentage recognition achieved using the MPI model by Svirsky and the proposed algorithm (blue line in Figure 5.4). The y-axis is the percentage recognition of normal hearing listeners for the studies by Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005) (green, black, magenta and cyan lines in Figure 5.4).

CHAPTER 6 CONCLUSION

The aim of this study was to develop a signal-processing algorithm that would enable the evaluation of speech-perception models to be used for severely degraded speech, typically at negative SNRs. No attempt was made to investigate the cues used for speech perception prediction or to develop a new speech-perception model. The focus of this research was successful estimation of the required vowel cues for CVC syllables degraded by white Gaussian noise, in order to allow for the use of existing perception models. The set of cues and the perception model were proposed by Svirsky (2000). The cues used for vowel perception are well documented, and thus it was attempted to contribute to the body of knowledge regarding vowel perception, specifically at low SNRs. The proposed signal-processing algorithm consists of three processing steps. Namely:

1) Speech enhancement
2) Cue estimation
3) Generation of confusion matrices by means of classification.

The speech enhancement is done by means of a Kalman filter using EM. The signal-processing gain that can be achieved with the use of the filter is SNR-dependent, but is in the order of 6 dB. Only AWGN was investigated, but future research can be conducted on cue estimation for speech degraded by other types of noise. To enable the cue estimation of the specific vowel, the location of the vowel in the syllable has to be identified. This is accomplished by means of a CFAR detector used for voicing detection, with the spectrogram of the degraded syllable as an input. The vowel duration is not one of the cues used for classification in this study, but this cue is a by-product of this algorithm. The cues that are estimated are the first formant frequency and the RMS channel amplitude ratios of four bandpass filters. For the estimation of the first formant frequency, a process is proposed which does not assume any a priori knowledge regarding the input CVC syllable. The transfer function of the pre-emphasis filter is adapted according to the characteristics of the CVC syllable. The formant frequency is estimated using a high-order LPC analysis of the isolated vowel in the CVC syllable.

The channel amplitude ratios for the four frequency bands used by Svirsky are calculated using Butterworth bandpass filters. The vowel classification is done using a multivariate Gaussian classifier. The model proposed by Svirsky uses the JNDs of each of the cues as the standard

deviations of the multivariate Gaussian distributions. Each of the estimated cues are used as vowel features in order to generate a confusion matrix from the various degraded CVC syllables, which were the input to the algorithm. By generating a confusion matrix for various SNRs a perception performance graph can be generated. The proposed algorithm allows this perception performance graph to be generated from an SNR as low as -10 dB. By performing an RTI analysis, it can also be seen which of the cues used in the classification contributed to the degradation of perception performance as the SNR deteriorates.

The primary research question was: using existing signal processing, can an algorithm be developed and successfully applied to severely degraded speech to enable the estimation of speech cues as required by perception prediction models? Also, can the signal processing be performed automatically and without any a priori knowledge regarding the input? This study shows that, with the utilization of the appropriate signal-processing techniques (section 3.2 and section 3.3), hearing perception can be investigated (chapter 4), with the use of perception models (section 3.4), at lower SNRs than previously investigated in the literature. The secondary research question was: do these perception predictions follow the trends in available published data? As discussed in chapter 5, no published data could be found for a direct comparison with the data generated using the MPI model by Svirsky (2000). Data were, however, available (Boothroyd & Nittrouer, 1988; Parikh & Loizou, 2005) for listening experiments, performed on normal hearing people, for an SNR as low as -10 dB. In comparing the data of Boothroyd and Nittrouer as well as Parikh and Loizou to the data generated using the proposed algorithm, similar trends can be observed (chapter 5).

6.1 Future Work

The following investigations can be done to build on the presented work:

1. An investigation can be done on whether the same signal-processing techniques presented in this study, can be applied to consonant cue estimation. The cues required for consonant perception prediction would be different from those presented in this study, but some of the processing techniques may also be applicable to the enhancement and estimation of consonant cues.
2. The additive noise used to generate the severely degraded speech was white noise, and this determined the specific implementation of the Kalman filter. The speech

enhancement (based on a Kalman filter) can be extended to allow for coloured noise suppression. This work would be of particular importance in the military domain where techniques used to disrupt speech communication intentionally may have a wide variety of spectral, temporal and statistical properties. Furthermore, the noise experienced by CI users has distinct spectral and temporal characteristics. Again, the Kalman filter implementation required to enhance the speech signal as heard by CI users would require specific adjustments.

3. The parameters that influence the amount of signal enhancement, as mentioned in the discussion, can be adjusted to modify the perception prediction performance to match that of listening experiment results on CI users or normal hearing listeners. The perception-prediction performance is also linked to the choice of cues used in the specific perception-prediction model.
4. Experiments can be done on listeners with CIs to investigate perception performance at negative SNRs, when the input is disrupted by AWGN. Using the techniques described in this dissertation, these listening experiment results can be used to evaluate the prediction performance of existing models (and the associated cues) at negative SNRs.
5. The various CI signal-processing strategies (Conning, 2005) degrade the various speech cues of the input speech in different ways (for example, discretization of frequency and amplitude). It can be investigated if the degrading effects of the CI signal-processing strategy can be related to an equivalent SNR of white Gaussian noise degradation.
6. Based on the high correlation between the predicted recognition of vowels (using the MPI model) and the data from Boothroyd and Nittrouer (1988) and Parikh and Loizou (2005), a listening experiment can be done to investigate the observation (as discussed in section 5.3) that normal hearing listeners may use amplitude ratios of spectral regions as cues when listening to severely degraded speech.

REFERENCES

Acker-Mills, B. E., Houtsma, A. J. M., & Ahroon, M. A. (2006), "Speech Intelligibility in Noise Using Throat and Acoustic Microphones", *Aviation Space and Environmental Medicine*, vol. 77, no. 1, pp. 26-31.

Ainsworth, W. A. (1971), "Duration as a Cue in the Recognition of Synthetic Vowels", *Journal of the Acoustical Society of America*, vol. 51, no. 2, pp. 648-651.

Allen, J. B. (1994), "How do Humans Process and Recognize Speech?", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567-577.

Alwan, A., Narayanan, S., Shen, A., & Strope, B. (1995), "Speech Production and Perception Models and Their Applications to Synthesis, Recognition, and Coding", URSI International Symposium on Signals, Systems, and Electronics, Proceedings, 25-27 Oct. 1995, San Francisco, pp. 367-372.

Anderson, B. D. O. & Moore, J. B. (1979), *Optimal Filtering*, Prentice-Hall: Englewood Cliffs.

Atal, B. S. & Hanauer, S. L. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637-655.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press: Oxford.

Boll, S. F. (1979), "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120.

Boothroyd, A. & Nittrouer, S. (1988), "Mathematical Treatment of Context Effects in Phoneme and Word Recognition", *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 101-114.

Bozic, S. M. (1979), *Digital Kalman Filtering*, Edward Arnold: London.

Carlotto, M. J. (1997), "Detection and Analysis of Change in Remotely Sensed Imagery with Application to Wide Area Surveillance", *IEEE Transactions on Image Processing*, vol. 6, no. 1, pp. 189-202.

Chang, C. C. & Song, K. (1997), "Environment Prediction for a Mobile Robot in a Dynamic Environment", *IEEE Transactions on Robotics and Automation*, vol. 13, no. 6, pp. 862-872.

Chang, J., Kim, N. S., & Mitra, K. (2006), "Voice Activity Detection Based on Multiple Statistical Models", *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965-1976.

Conning, M. (2005), *Acoustic Modelling of Cochlear Implants,* Masters of Engineering (Bio-Engineering), University of Pretoria.

Cooke, M. (2006), "A Glimpsing Model of Speech Perception in Noise", *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562-1573.

Davidson, G., Griffiths, H. D., & Ablett, S. (2004), "Analysis of High-Resolution Land Clutter", *IEE Proceedings - Vision, Image and Signal Processing,* vol. 151, no. 1, pp. 86-91.

Dillard, G. M. & Rickard, J. T. (1974), "A Distribution-Free Doppler Processor", *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-10, no. 4, pp. 479-486.

Dorman, M. F., Smith, L., Smith, M., & Parkin, J. (1992), "The Coding of Vowel Identity by Patients who use the Ineraid Cochlear Implant", *Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3428-3431.

Du, W. & Driessen, P. (1991), "Speech Enhancement Based on Kalman Filtering and EM Algorithm", IEEE Pacific Rim Conference on Communication, Computers and Signal Processing, 9-10 May 1991, Victoria, pp. 142-145.

Dubbelboer, F. & Houtgast, T. (2007), "A Detail Study on the Effects of Noise on Speech Intelligibility", *Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. 2865-2871.

Duk, Y. & Kondoz, A. (2001), "Analysis and Improvement of a Statistical Model-Based Voice Activity Detector", *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276-278.

Ephraim, Y. & Malah, D. (1984), "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121.

Ephraim, Y. & Malah, D. (1985), "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. ASSP-33, no. 2, pp. 443-445.

Ephraim, Y., Malah, D., & Juang, B. H. (1989), "On the Application of Hidden Markov Models for Enhancing Noisy Speech", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 37, no. 12, pp. 1846-1856.

Ffrench, P. A., Zeidler, J. H., & Ku, W. H. (1997), "Enhanced Detectability of Small Objects in Correlated Clutter Using an Improved 2-D Adaptive Lattice Algorithm", *IEEE Transactions on Image Processing*, vol. 6, no. 3, pp. 383-397.

Fletcher, H. (1953), *Speech and Hearing in Communication*, Krieger: New York.

Fletcher, H. & Galt, R. H. (1950), "The Perception of Speech and Its Relation to Telephony", *Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 89-151.

French, N. R. & Steinberg, J. C. (1947), "Factors Governing the Intelligibility of Speech Sounds", *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90-119.

Gandhi, P. P. & Kassam, S. A. (1988), "Analysis of CFAR Processors on Nonhomogeneous Background", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 427-445.

Gannot, S., Burshtein, D., & Weinstein, E. (1998), "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms", *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373-385.

Gibson, J. D., Koo, B., & Gray, S. D. (1991), "Filtering of Colored Noise for Speech Enhancement and Coding", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 39, pp. 1732-1742.

Gold, B. & Rabiner, L. R. (1968), "Analysis of Digital and Analog Formant Synthesizers", *IEEE Transactions of Audio and Electroacoustics*, vol. AU-16, no. 1, pp. 81-94.

Gong, Y. (1995), "Speech Recognition in Noisy Environments: A Survey", *Speech Communication*, vol. 16, no. 3, pp. 261-291.

Hansen, J. H. L. & Clements, M. A. (1991), "Constrained Iterative Speech Enhancement with Application to Speech Recognition", *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795-805.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995), "Acoustic Characteristics of American English Vowels", *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099-3111.

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954), "Intelligibility of Different Speech Materials", *Journal of the Acoustical Society of America*, vol. 26, no. 4, pp. 530-538.

Hovanessian, S. A. (1973), *Radar Detection and Tracking Systems*, Artech House: Dedham.

Howard-Jones, P. A. & Rosen, S. (1993), "Uncomodulated Glimpsing in 'Checkerboard' Noise", *Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2915-2922.

Howes, D. (1957), "On the Relation Between the Intelligibility and Frequency of Occurrence of English Words", *Journal of the Acoustical Society of America*, vol. 29, no. 2, pp. 296-305.

Kalman, R. E. & Bucy, R. S. (1961), "New Results in Linear Filtering and Prediction Theory", *Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering*, vol. 83D, pp. 95-108.

Kasturi, K., Loizou, P. C., Dorman, M., & Spahr, T. (2002), "The Intelligibility of Speech with 'Holes' in the Spectrum", *Journal of the Acoustical Society of America*, vol. 112, no. 3, pp. 1102-1111.

Khalighi, M. A. & Bastani, M. H. (2000), "Adaptive CFAR Processors For Nonhomogeneous Environments", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 36, no. 3, pp. 889-897.

Klatt, D. H. (1980), "Software for a Cascade/Parallel Formant Synthesizer", *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971-995.

Klein, W., Plomp, R., & Pols, L. C. W. (1970), "Vowel Spectra, Vowel Spaces, and Vowel Identification", *Journal of the Acoustical Society of America*, vol. 48, no. 4B, pp. 999-1009.

Kopp, G. A. & Green, H. C. (1946), "Basic Phonetic Principles of Visible Speech", *Journal of the Acoustical Society of America*, vol. 18, no. 1, pp. 74-89.

Lee, K. Y. & Shirai, K. (1996), "Efficient Recursive Estimation for Speech Enhancement in Colored Noise", *IEEE Signal Processing Letters*, vol. 3, pp. 196-199.

Leung, H. (1996), "Nonlinear Clutter Cancellation and Detection Using a Memory-Based Predictor", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 4, pp. 1249-1256.

Leung, H. & Young, A. (2000), "Small Target Detection in Clutter Using Recursive Nonlinear Prediction", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 36, no. 2, pp. 713-718.

Lim, J. S. & Oppenheim, A. V. (1978), "All-Pole Modelling of Degraded Speech", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. ASSP-26, no. 3, pp. 197-210.

Loizou, P. C. & Poroy, O. (2001), "Minimum Spectral Contrast Needed for Vowel Identification by Normal Hearing and Cochlear Implant Listeners", *Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1619-1627.

Makhoul, J. (1975), "Linear Prediction: A Tutorial Review", *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561-580.

McCandless, S. (1974), "An Algorithm for Automatic Formant Extraction using Linear Prediction Spectra", *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. ASSP-22, no. 2, pp. 135-141.

Miller, G. A. & Licklider, J. C. R. (1950), "The Intelligibility of Interrupted Speech", *Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 167-173.

Miller, G. A. & Nicely, P. E. (1955), "An Analysis of Perceptual Confusions among some English Consonants", *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338-352.

Miller, J. D. (1989), "Auditory-Perceptual Interpretation of the Vowel", *Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2114-2134.

Moore, B. C. J. (2003), "Speech Processing for the Hearing-Impaired: Successes, Failures, and Implications for Speech Mechanisms", *Speech Communication*, vol. 41, no. 1, pp. 81-91.

Mustafa, K. & Bruce, I. C. (2006), "Robust Formant Tracking for Continuous Speech with Speaker Variability", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 435-444.

Nearey, T. M. (2001), "Phoneme-Like Units and Speech Perception", *Language and Cognitive Processes*, vol. 16, no. 5, pp. 673-681.

Nixon, C. W., McKinley, R. L., & Moore, T. J. (1982), "Increase in Jammed Word Intelligibility due to Training of Listeners", *Aviation Space and Environmental Medicine*, vol. 53, no. 3, pp. 239-244.

Nooteboom, S. G. & Doodeman, G. J. N. (1980), "Production and Perception of Vowel Length in Spoken Sentences", *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 276-287.

Oppenheim, A. V. & Schafer, R. W. (1999), *Discrete-Time Signal Processing*, 2 edn, Prentice Hall: Upper Saddle River.

Paliwal, K. K. & Basu, A. (1987), "A Speech Enhancement Method Based on Kalman Filtering", IEEE International Conference on Acoustics, Speech, and Signal Processing, April 1987, Dallas, pp. 177-180.

Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes*, 3 edn, McGraw-Hill: New York.

Parikh, G. & Loizou, P. C. (2005), "The Influence of Noise on Vowel and Consonant Cues", *Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3874-3888.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997), "Speech Motor Control: Acoustic Goals, Saturation Effects, Auditory Feedback and Internal Models", *Speech Communication*, vol. 22, pp. 227-250.

Peterson, G. E. (1952), "The Information-Bearing Elements of Speech", *Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 629-637.

Peterson, G. E. & Barney, H. L. (1952), "Control Methods Used in a Study of the Vowels", *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175-183.

Pickett, J. M. (1957), "Perception of Vowels Heard in Noises of Various Spectra", *Journal of the Acoustical Society of America*, vol. 29, no. 5, pp. 613-620.

Pollack, I., Rubenstein, H., & Decker, L. (1959), "Intelligibility of Known and Unknown Message Sets", *Journal of the Acoustical Society of America*, vol. 31, no. 3, pp. 273-279.

Potter, R. K. & Steinberg, J. C. (1950), "Toward the Specification of Speech", *Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 807-820.

Proakis, J. G. & Salehi, M. (1994), *Communication Systems Engineering*, Prentice-Hall: Englewood Cliffs.

Rabiner, L. R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286.

Rabiner, L. R. & Schafer, R. W. (1978), *Digital Processing of Speech Signals*, Prentice-Hall: Englewood Cliffs.

Remus, J. J. & Collins, L. M. (2004a), "Predicting Vowel and Consonant Confusions Using Signal Processing Techniques", Proceedings of the VIII International Cochlear Implant Conference, May 2004, Indianapolis, pp. 15-18.

Remus, J. J. & Collins, L. M. (2004b), "Vowel and Consonant Confusion in Noise by Cochlear Implant Subjects: Predicting Performance using Signal Processing Techniques", IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, 17-21 May 2004 , Montreal, pp. IV-13-16.

Sabes, P. N. (2000), "The Planning and Control of Reaching Movements", *Current Opinion in Neurobiology*, vol. 10, no. 6, pp. 740-746.

Savin, H. B. (1963), "Word-Frequency Effect and Errors in the Perception of Speech", *Journal of the Acoustical Society of America*, vol. 35, no. 2, pp. 200-206.

Smith, J. C. & Lourens, J. G. (2006), "Optimization of a Feedforward Active Noise Reduction (ANR) System for Broadband Noise Cancellation", Conference Proceedings of the Institute of Noise Control Engineering of the USA, 3-6 December 2006, Honolulu, Hawaii.

Snell, R. C. & Milinazzo, F. (1993), "Formant Location from LPC Analysis Data", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129-134.

Stevens, K. N. (1959), "Effect of Duration on Identification", *Journal of the Acoustical Society of America*, vol. 31, no. 1, p. 109.

Strange, W. (1989), "Evolving Theories of Vowel Perception", *Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2081-2087.

Strope, B. & Alwan, A. (1997a), "A Model of Dynamic Auditory Perception and its Application to Robust Word Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 451-464.

Strope, B. & Alwan, A. (1997b), "Modeling Auditory Perception to Improve Robust Speech Recognition", Conference Record of the Thirty-First Asilomar Conference on Signals, Systems & Computers, 2-5 November 1997, Pacific Grove, pp. 1056-1060.

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*, vol. 107, no. 3, pp. 1521-1529.

Thomas, J. B. (1970), "Nonparametric Detection", *Proceedings of the IEEE*, vol. 58, no. 5, pp. 623-631.

Van Tasell, D. J., Soli, S., Kirby, V. M., & Widin, G. P. (1987), "Speech Waveform Envelope Cues for Consonant Recognition", *Journal of the Acoustical Society of America*, vol. 82, no. 4, pp. 1152-1161.

Van Wieringen, A. & Wouters, J. (1999), "Natural Vowel and Consonant Recognition by Laura Cochlear Implantees", *Ear & Hearing*, vol. 20, no. 2, pp. 89-103.

Wang, M. D. & Bilger, R. C. (1973), "Consonant Confusions in Noise: A Study of Perceptual Features", *Journal of the Acoustical Society of America*, vol. 54, no. 5, pp. 1248-1266.

Watkins, K. & Paus, T. (2004), "Modulation of Motor Excitability during Speech Perception: The Role of Broca's Area", *Journal of Cognitive Neuroscience*, vol. 16, no. 6, pp. 978-987.

Weinstein, E., Oppenheim, A. V., Feder, M., & Buck, J. R. (1994), "Iterative and Sequential Algorithms for Multisensor Signal Enhancement", *IEEE Transactions on Signal Processing*, vol. 42, no. 4, pp. 846-859.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. (1995), "An Internal Model for Sensorimotor Integration", *Science*, vol. 269, no. 5232, pp. 1880-1882.

ADDENDUM A BANDPASS FILTER IMPLEMENTATION

Butterworth Bandpass Filter: Bandwidth 100 Hz to 800 Hz

Cascade Section 1	
Numerator	1, 0, -1
Denominator	1, -1.9123143660445407, 0.93043073090650597
Gain	0.062275036716503571
Cascade Section 2	
Numerator	1,0,-1
Denominator	1, -1.9936368797188968, 0.99377659460037082
Gain	0.062275036716503571
Cascade Section 3	
Numerator	1,0,-1
Denominator	1, -1.9806620771653949, 0.98082912389638044
Gain	0.060459283298695238
Cascade Section 4	
Numerator	1,0,-1
Denominator	1, -1.8156724417618664, 0.829953939290627
Gain	0.060459283298695238

Cascade Section 5	
Numerator	1,0,-1
Denominator	1, -1.8788591287436636, 0.88038701562255961
Gain	0.05980649218872023
Output Gain	1

Butterworth Bandpass Filter: Bandwidth 700 Hz to 1400 Hz

Cascade Section 1	
Numerator	1, 0, -1
Denominator	1, -1.8634916998334847, 0.90766979815107174
Gain	0.067047097302036923
Cascade Section 2	
Numerator	1,0,-1
Denominator	1, -1.9508650770587062, 0.95918951975318745
Gain	0.067047097302036923
Cascade Section 3	
Numerator	1,0,-1
Denominator	1, -1.8514578319945982, 0.87004316108222313
Gain	0.064978419458888506
Output Gain	1

Butterworth Bandpass Filter: Bandwidth 1400 Hz to 2500 Hz

Cascade Section 1	
Numerator	1, 0, -1
Denominator	1, -1.7061556700802505, 0.84822037302878328
Gain	0.11415697136652779
Cascade Section 2	
Numerator	1,0,-1
Denominator	1, -1.8954307470808738, 0.92645306620885304
Gain	0.11415697136652779
Cascade Section 3	
Numerator	1,0,-1
Denominator	1, -1.7199142387663993, 0.78300857347910702
Gain	0.10849571326044652
Output Gain	1

Butterworth Bandpass Filter: Bandwidth 2300 Hz to 4500 Hz

Cascade Section 1	
Numerator	1, 0, -1
Denominator	1, -1.3451751549044562, 0.74624943878835648
Gain	0.19647994473530325

Cascade Section 2	
Numerator	1,0,-1
Denominator	1, -1.7878422898224349, 0.87286497991222567
Gain	0.19647994473530325
Cascade Section 3	
Numerator	1,0,-1
Denominator	1, -1.4669689176208813, 0.63737613606067089
Gain	0.18131193196966458
Output Gain	1