UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

VOWEL PERCEPTION IN SEVERE NOISE

by

Rikus Swanepoel

Submitted in partial fulfillment of the requirements for the degree

Master of Engineering (Bio-Engineering)

in the

Faculty of Engineering, the Built Environment and Information Technology

UNIVERSITY OF PRETORIA

December 2010

VOWEL PERCEPTION IN SEVERE NOISE by

Rikus Swanepoel

Supervisor: Prof JJ Hanekom

Department of Electrical, Electronic and Computer Engineering

Master of Engineering (Bio-Engineering)

Summary

A model that can accurately predict speech recognition for cochlear implant (CI) listeners is essential for the optimal fitting of cochlear implants. By implementing a CI acoustic model that mimics CI speech processing, the challenge of predicting speech perception in cochlear implants can be simplified. As a first step in predicting the recognition of speech processed through an acoustic model, vowel perception in severe speech-shaped noise was investigated in the current study. The aim was to determine the acoustic cues that listeners use to recognize vowels in severe noise and make suggestions regarding a vowel perception predictor. It is known that formants play an important role in quiet, while in severe noise the role of formants is still unknown. The relative importance of F1 and F2 is also of interest, since the masking of noise is not always evenly distributed over the vowel spectrum. The problem was addressed by synthesizing vowels consisting of either detailed spectral shape or formant information. F1 and F2 were also suppressed to examine the effect in severe noise. The synthetic stimuli were presented to listeners in quiet and signal-to-noise ratios of 0 dB, -5 dB and -10 dB. Results showed that in severe noise, vowels synthesized according to the whole-spectrum were recognized significantly better than vowels containing only formants. Multidimensional scaling and FITA analysis indicated that formants were still perceived and extracted by the human auditory system in severe noise, especially when the vowel spectrum consisted of the whole spectral shape. Although F1 and F2 vary in importance in listening conditions of quiet and less noisy conditions, the role of the two cues appears to be similar in severe noise. It was suggested that not only the availability formants, but also details of the vowel spectral shape can help to predict vowel recognition in severe noise to a certain degree.

Keywords: Cochlear implant, acoustic model, speech-shaped noise, acoustic cues, formant, multidimensional scaling

VOKAALPERSEPSIE IN SWAAR RUIS deur

Rikus Swanepoel

Studieleier: Prof JJ Hanekom

Departement Elektriese, Elektroniese en Rekenaar-Ingenieurswese

Meester van Ingenieurswese (Bio-Ingenieurswese)

Opsomming

`n Model wat spraakherkenning vir gebruikers van kogleêre inplantings kan voorspel, is belangrik vir die optimale instelling van kogleêre inplantings. Die gebruik van `n akoestiese model, wat die spraakverwerking in `n kogleêre inplanting naboots, kan die uitdaging van spraakherkenning vir gebruikers van kogleêre inplantings vereenvoudig. `n Eerste stap om die persepsie van spraak afkomstig vanaf `n akoestiese model te voorspel, is om vokaalpersepsie in swaar spraakgevormde ruis te ondersoek. Die doel van hierdie studie is dus om te bepaal watter vokaaleienskappe belangrik is vir vokaalherkenning in swaar ruis, en om voorstelle te maak oor die wyse waarop hierdie vokaaleienskappe gebruik kan word om vokaalpersepsie in swaar ruis te voorspel. Formante speel `n belangrike rol in vokaalherkenning in stilte, maar die rol in swaar ruis is nog onbekend. Die relatiewe belangrikheid van F1 en F2 is ook van belang omdat ruis nie altyd egalig versprei is oor die vokaalspektrum nie. Die probleem rakende die rol van formante en heelspektrum is aangespreek deur vokale te sintetiseer op so `n wyse dat die vokaalspektrum slegs inligting vanaf die heelspektrum, of slegs formant-inligting bevat. F1 en F2 is ook onderdruk om die effek daarvan waar te neem op vokaalherkenning in swaar ruis. Die sintetiese vokale is aangebied aan luisteraars in stilte en sein-tot-ruis verhoudings van 0 dB, -5 dB en -10 dB. Resultate het bewys dat in swaar ruis, vokale met heelspektrumeienskappe beter herken is as vokale wat slegs uit formante bestaan. Multidimensionele skalering en FITA-analise het aangedui dat formante steeds gehoor is in diep ruis, alhoewel die formante beter oorgedra word indien heelspektrum-inligting ook beskikbaar is. Al verskil die belangrikheid van F1 en F2 in stilte en ligte ruis, speel dié

twee vokaaleienskappe `n eenderse rol in swaar ruis. Daar word voorgestel dat vokaalherkenning in diep ruis tot `n mate voorspel kan word indien die beskikbaarheid van beide die heelspektrum en formante bepaal kan word.

Sleutelwoorde: Kogleêre inplanting, akoestiese model, spraakagtige ruis, akoestiese eienskappe, formant, multidimensionele skalering

ACKNOWLEDGEMENTS

It is a pleasure to thank those who played a role in my life during my Master's studies:

- My wife Albe, whose love, support, patience and persistent confidence in me kept me going until the end. I could not have asked for a better wife.
- My parents for their encouragement and believe in me from the beginning.
- My parents-in-law for their continuous interest and support during the past few years.
- My promoter for his guidance that enabled me to develop an understanding of the research field.
- Family, friends and everyone else who supported me in any respect during my studies. I am a richer person knowing people like you.
- God, without Whom this dissertation would not have been possible.

"Be still, and know that I am God" – Ps 46:10 (NIV)

List of abbreviations

AI	Articulation index
ANOVA	Analysis of variance
CI	Cochlear implant
CIS	Continuous interleaved sampling
dB	Decibel
DCT	Discrete cosine transform
DCTCs	Discrete cosine transform coefficients
DSS	Damped sinewave synthesizer
DTW	Dynamic time warping
F0	Fundamental frequency
F1	Formant 1
F2	Formant 2
F3	Formant 3
F4	Formant 4
F5	Formant 5
FFT	Fast Fourier Transform
FITA	Feature information transmission analysis
HMM	Hidden Markov models

Hz	Hertz
LCP	Linear Predictive Coding
MDS	Multidimensional Scaling
MPI	Multidimensional Phoneme Identification
ms	Millisecond
PLP	Perceptual Linear Prediction
rms	Root Mean Square
SIFT	Simplified Inverse Filter Tracking
SM	Spectral Manipulation
SNR	Signal-to-Noise Ratio
SNRs	Signal-to-Noise Ratios
SPL	Sound Pressure Level
ST	Synthesis Type
TEC	Token Envelope Correlation
VAF	Variance Accounted For
MLP	Mean Logarithmic Probability
INDSCAL	Individual Difference Scaling
FIR	Finite Impulse Response

Table of contents

CHAPTER 1 INTRODUCTION

1.1 CHAPTER OBJECTIVES

This study investigates the perception of vowels in severely degraded speech. Insight into this subject is important for future research on cochlear implants since the sound perceived by cochlear implant (CI) users appear to be severely degraded when compared with the speech perception of normal-hearing listeners. CI acoustic models provide a method by which speech can be simulated to ensure that it is perceived similarly by normal-hearing listeners and CI users alike (Liu and Fu, 2007; Remus and Collins, 2005). Vowel perceptions of normal-hearing listeners listening to an acoustic model will therefore be investigated by analyzing vowel confusions in severe noise, which will provide more insight into the problem of predicting the speech perception of CI users.

1.2 BACKGROUND AND SCOPE OF WORK

A CI is a prosthetic device that can be inserted in the inner ear to restore partial hearing to persons with sensorineural hearing loss. It functions in such a way that the outer, middle and part of the inner ear are bypassed in order to stimulate the auditory neurons directly (Loizou, 1999). The success of cochlear implants depends on factors such as the period of deafness and the location of the remaining auditory neurons, amongst others (Liu and Fu, 2007). Sufficient speech perception depends on tuned parameters for each individual user's implant, such as frequency and amplitude mapping (Hoth, 2007; Van Hoesel, Böhm, Battmer, Beckschebe and Lenarz, 2005), while design parameters such as the number of electrodes (Fishman, Shannon and Slattery, 1997) and electrode insertion depth (Dorman, Loizou and Rainey, 1997a) also contribute to the success of the implant.

The process involved in optimizing speech perception of cochlear implants is time-consuming since requirements differ from one user to another. The importance of optimizing speech processors becomes even more evident in a clinical environment, as time becomes a factor. It would be convenient to objectively determine the various parameters in the clinical fitting of cochlear implants in order to avoid extensive tests with the implant user (Liu and Fu, 2007). This can be done with the use of a speech perception model that objectively determines the outcome of CI speech perception tests.

A reliable model of speech perception for CI users will consequently aid in the development of better strategies for speech processing, as well as to fit cochlear implants in an optimal way without the need for tedious listening tests (Chang, Anderson and Loizou, 2001). Accurate predictors of speech perception will also help in the further improvement of speech processing strategies and noise reduction methods for cochlear implants (Remus and Collins, 2005). Attempts to design objective models for speech recognition in cochlear implants have been made by utilizing listeners' discrimination according to certain perceptual dimensions and locating phonemes in a multidimensional space according to psychophysical measurements (Svirsky, 2000). Neural networks have also been found to be a good predictor of CI perception, as it allows one to automatically optimize speech processing parameters for implants, without the need of listening experiments (Chang et al., 2001).

The development of objective CI perception models have been simplified by the use of CI acoustic models (Blamey, Dowell and Tong, 1984; Dorman, Loizou, Fitzke and Tu, 1998b; Liu and Fu, 2007). These models enable the use of normal-hearing listeners in speech intelligibility tests rather than CI users. When normal-hearing listeners are presented with speech tokens processed through an acoustic model, the sounds will appear to be severely degraded and corrupted by noise. Factors influencing vowel perception for CI users become apparent when vowels are represented through acoustic models to normal-hearing listeners. Research has shown that predictions of vowel and consonant confusions for CI users become apparent through acoustic models by means of prediction

metrics using spectral and temporal signal properties (Remus and Collins, 2005), Mel-frequency cepstrum coefficients (Liu and Fu, 2007), and maps of perceptual vowel spaces according to vowel cues (Iverson, Smith and Evans, 2006). CI perception can, however, still be improved (Remus and Collins, 2005) and future work is needed to develop an automatic model depicting speech recognition results with a very high degree of accuracy (Chang et al., 2001).

The problem of explaining speech perception for CI users and normal hearing listeners of CI acoustic models can be extended to the perception of degraded speech by normal hearing listeners. Speech recognition becomes extremely difficult in noisy environmental conditions like, for example, factory and helicopter noise (Cooke, Green, Josifovski and Vizinho, 2001), external aircraft noise (Williams, Pearsons and Hecker, 1971) and aircraft cockpit noise (Chan and Simpson, 1990). Common forms of speech perception interference like talker babble or reverberation mask information that would normally be available to listeners, leaving them with access to only certain parts of the spectral and temporal information in the speech signal (Nabelek, Czyzewski and Krishnan, 1992). Therefore, similar to CI perception, the poor intelligibility of distorted or masked speech is related to the limited information that is transferred to the auditory system.

The objective of this study is to contribute to the solution of predicting and explaining the perception of vowels and consonants of CI users. Since vowels are generally assumed to convey more information than consonants in normal conversational speech (Kewley-Port, Burkle and Lee, 2007), the present study focuses on the recognition of vowels in degraded conditions. The sound perceived by both a CI user and a normal-hearing listener listening to an acoustic model is usually degraded to a much greater extent compared with the sound heard by a normal-hearing person (Dorman, Loizou and Fitzke, 1998a; Ter Keurs, Festen and Plomp, 1992; Zwolan and Kileny, 1993). For this reason the CI or acoustic model can be seen as a degradation channel altering the quality of a speech signal. To simplify this degradation channel and to control the degree of degradation in speech signals, vowel confusions will be analyzed in noise.

Figure 1.1 summarizes the solution of predicting speech recognition results in severe noise. To predict vowel confusions presented to normal-hearing persons in a quiet environment, it is necessary to know the various cues that aid these listeners in making vowel judgments. These cues are then used to design a model that will predict vowel confusions based on the availability of cues in severe noise. The main aim will be to determine the specific cues that listeners use to recognize vowels in severe noise.

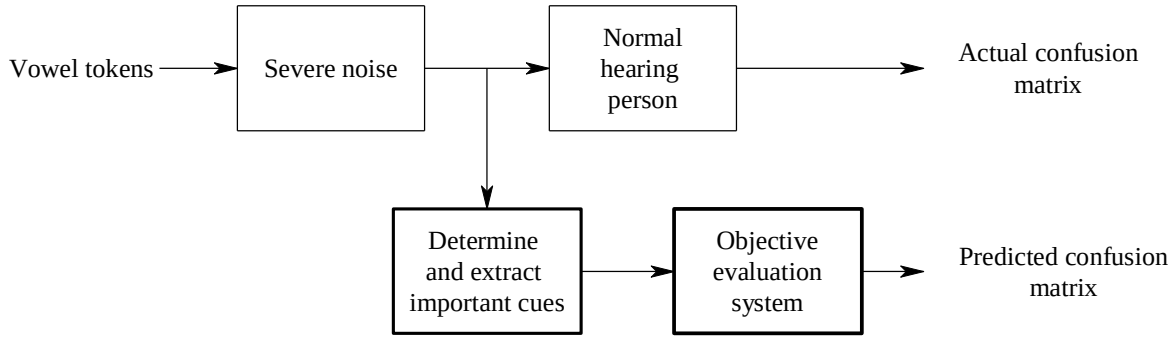

Figure 1.1. Proposed solution for predicting vowel recognition results in severe noise.

1.3 HYPOTHESIS AND RESEARCH QUESTIONS

The intent of this study is firstly to determine whether the cues known to be important for vowel recognition in quiet are still available to listeners in degraded hearing conditions, and secondly, whether or not listeners use these cues when making vowel judgments. Formants, defined as the resonant peaks in the spectral envelope of a vowel sound (Borden and Harris, 1985), are believed to be the main cues used by listeners in recognizing vowels in quiet (Delattre, Liberman, Cooper and Gerstman, 1952; Hillenbrand, Getty, Clark and Wheeler, 1995). Predicting vowel confusions for normal-hearing listeners in quiet is often done by locating vowels in a vowel space with F1 (Formant 1) and F2 (Formant 2) as

parameters (Iverson et al., 2006; Peterson and Barney, 1952). The question remains whether this is still the case in severe noise.

The hypothesis of this study is that listeners still use formant cues to identify vowels in severe noise. Formants are known to be important cues used by listeners for vowel identification in moderate speech-shaped noise (Gong, 1995). Despite the fact that noise severely distorts the speech spectrum, the frequency locations of formants are usually well preserved, even when the ratios of the formant peak amplitudes are lost (Assmann and Summerfield, 2004). Clear F1 and vague F2 cues have been found to still be available in vowels at signal-to-noise ratio (SNR) levels of -5 dB (decibel) (Parikh and Loizou, 2005). Knowing this, it is of interest to investigate whether formants are still available to listeners at SNR levels lower than -5 dB, and if listeners do indeed use these available formants.

Research questions to be investigated in this study are therefore:

- Do listeners only use formant cues to identify vowels in severe noise?
- If formants are indeed used, does the importance of F1 and F2 differ?
- What additional cues, other than formants, lead to vowel recognition in severe noise?
- Will the importance of certain vowel cues depend on the level of noise that is added to the signal?
- In which manner can the cues found to be important for vowel recognition in severe noise be used to predict vowel perception in noise?

1.4 APPROACH

There are at least two methods by which the relative importance of certain cues for vowel perception can be investigated. With the first method, natural vowels are manipulated to

observe the effect of the deletion of a cue on listening test results (Nearey and Assmann, 1986; Sakayori, Kitama, Chimoto, Qin and Sato, 2002). The second method involves the implementation of vowel synthesis, making it possible to control certain vowel cues in the process, while disregarding others (Delattre et al., 1952; Hillenbrand and Gayvert, 1993; Ito, Tsuchida and Yano, 2001; Leek, Dorman and Summerfield, 1987; Sakayori et al., 2002). The second method is followed in the present study.

The hypothesis of the present study is tested by synthesizing vowel sounds and presenting it to listeners in quiet and in noisy conditions. Deterioration in identification scores occurs as a result of the noise masking important information that is usually available to listeners in quiet. When preserving only certain vowel cues in the synthesis process, the identification scores of the quiet and noisy vowels can be compared, and conclusions regarding the importance of cues in noise can be drawn..

The hypothesis is further explored by investigating confusions among vowels. When the confusion between two vowels increase with an increase in noise, the noise will enhance certain cues that are similar to the two vowels, while at the same time the noise will also suppress important cues that should aid the listener in correctly identifying the vowel.

The study is outlined in Figure 1.2. Recorded vowels are analyzed for certain cues that are reported in literature to be important for vowel perception. These cues are used to synthesize different representations of each vowel, with each synthesized vowel preserving only certain cues. The synthesized vowels are then presented to normal-hearing listeners in different levels of noise.

Methods such as Feature Information Transmission Analysis (FITA) and Multidimensional Scaling (MDS) are used to analyze listening test results when investigating which vowel cues are robust in noise and which are not perceived by listeners. MDS analysis presents vowel tokens as a spatial representation according to the experimental outcomes of the listening tests. The multidimensional arrangement can then be compared to different vowel

cue spaces from literature in order to determine the specific cues that explain the vowel confusions best.

Confusion patterns between vowels are also examined, while the percentage of correct scores for each vowel is noted. Listening tests as a whole will give a measure of importance for each cue, which provides a platform to suggest a model for vowel recognition in severe noise.

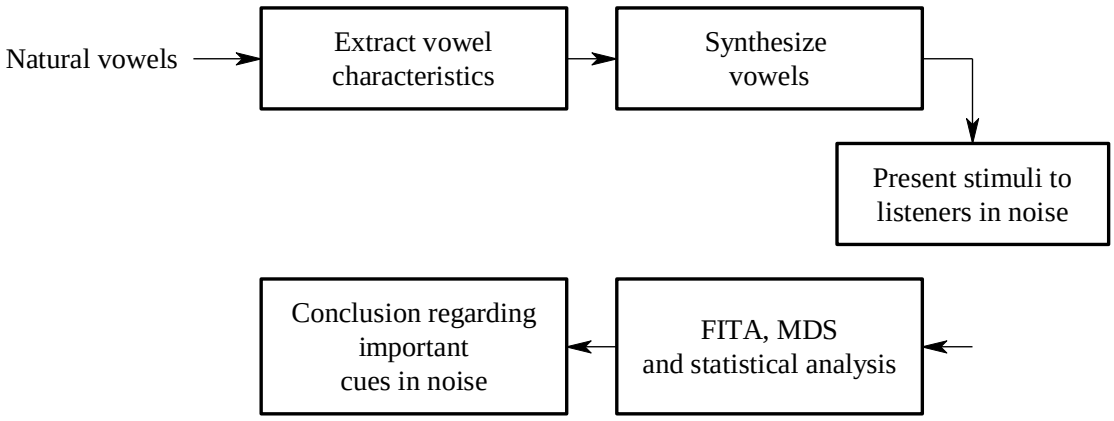

Figure 1.2. Outline of the approach followed in the present study.

1.5 OBJECTIVES

The main objective of this study was to determine the cues that are important for vowel recognition in severely degraded conditions. To achieve this, vowels were firstly synthesized according to certain cues, after which it was presented to listeners in severe noise. Statistical and analytical tests were done on the vowel confusion results to draw final conclusions regarding the important cues in severe noise. These three objectives are discussed next.

The first objective was to analyze vowel tokens for specific cues that should be preserved in the vowel synthesis process. These cues were found in literature to be used by listeners in quiet to make vowel judgments. Vowels were synthesized in different manners in order to retain certain cues while disregarding others. By controlling the specific cues present in vowel signals, the importance of cues could be compared directly by analyzing the recognition scores for each synthetic vowel. It also allows the systematical suppression of cues, which could give an indication of whether a specific cue is still used in vowel recognition at a certain degradation level.

The second objective was to present the synthetic vowels to normal-hearing listeners in quiet and severe noise conditions, which gave an indication of the influence of each cue on the perception of vowels at different noise levels. Allowing the SNR to be lowered systematically can reveal how certain cues become redundant and others important for vowel perception in the specific listening conditions.

The last objective was to analyze vowel confusion results. FITA and MDS analysis gave a measure of the availability of certain cues in severe noise, while percentage correct results were indicative of whether these cues were indeed used by listeners to recognize vowels correctly. MDS results provided an indication of how vowel perception in noise can be modelled in a multidimensional space. The best fit of two different vowel cue spaces to the MDS results was found, while thereafter the correlation of cues to each dimension of the MDS vowel space gave an indication of the ranking of importance for each cue.

1.6 OUTLINE OF THE PRESENT STUDY

The present study describes the process by which the perception of vowels in noise is investigated. Chapter 2 provides information from the literature regarding vowel recognition in quiet, as well as in degraded conditions. Vowel recognition for CI users is also discussed in order to become familiar with the cues that these listeners use in

identifying vowels. Gaps in the literature are discussed, after which the most important cues are identified for further testing in severe noise.

The method with which the vowel stimuli are produced and presented to listening subjects is described in Chapter 3. Acoustic analysis of natural vowels is done to extract cues that are identified from literature. The mimicking of these natural vowels is described where different vowel synthesis techniques are used, with each enhancing or suppressing certain cues. Synthetic consonants are added to the stimuli, after which it is presented to normal-hearing listeners. Chapter 4 provides information regarding the experimental procedure where the synthetic vowels were presented to normal-hearing listeners at different noise levels. It is also described how FITA and MDS methods are used to determine the effective transmission of certain vowel cues to the listener, as well as the relationship between the listening test results and certain vowel cue spaces.

Chapter 5 evaluates the effectiveness of the vowel synthesis techniques. It is evaluated if cues that were intended to be present in the synthetic vowels were indeed there. Predictions regarding vowel confusions are made by means of simple vowel space analysis. The results of listening experiments are statistically analysed, while confusions among vowels are also examined to determine the importance of certain cues for vowel recognition in noise.

A discussion of the results given in Chapter 5 follows in Chapter 6. Results are compared to data from literature, while the most important cues for vowel perception in noise which became apparent in the results section, are outlined. Chapter 7 summarizes the accomplishments reached in this study, while research questions raised in Chapter 1 are answered. The chapter concludes with suggestions for future work needed on the subject.

CHAPTER 2 CONTEXT OF THE WORK

2.1 CHAPTER OBJECTIVES

In this chapter, a thorough discussion of the literature is done to identify gaps in the knowledge regarding vowel recognition. The main focus falls on human perception of vowel sounds in quiet and degraded conditions. The essential cues that listeners extract from vowel sounds during vowel identification are investigated, while present vowel perception models for normal-hearing listeners and CI users are also discussed.

Literature provides clarity regarding the specific cues important for vowel perception for CI listeners and normal-hearing listeners in quiet and in noise. While knowing these specific cues are useful, it is also of value to take note of the methods implemented by the authors to determine these cues. The question of whether the cues that are important for vowel recognition in quiet are similar for noisy conditions is investigated by comparing results of several studies exploring this issue. Firstly, perception for normal-hearing listeners in ideal conditions is investigated, after which the focus moves to cochlear implants, acoustic modelling and finally to vowel perception in noisy environments.

2.2 VOWEL PERCEPTION FOR NORMAL-HEARING LISTENERS IN QUIET

Vowel perception in quiet for normal-hearing listeners can be modelled by locating the vowels in a vowel space, with the position of each vowel determined by a set of parameters that are related to the properties of the vowel (Delattre et al., 1952; Hillenbrand et al., 1995; Liu and Fu, 2005; Peterson and Barney, 1952). Vowel confusions are usually predicted by calculating the Euclidian distance between the vowels in the vowel space, with short distances depicting an increase in the probability of confusions, and longer distances indicating a decrease in the chance of confusions (Klatt, 1982; Liu and Fu,

2005). From this literature study, the specific cues that can be used to locate a vowel in its vowel space will become apparent, as well as the methods that are used to determine these cues.

2.2.1 Cues important for vowel recognition in quiet

2.2.1.1 Formant frequencies

Formants are the vocal tract resonant frequencies that appear as peaks in the spectra of vowels (McCandless, 1974). The frequencies of the formants in a vowel signal, and particularly the first two formants, have been found to be important cues for identifying vowel sounds (Borden and Harris, 1985; Dorman, Loizou, Spahr and Maloff, 2002; Iverson et al., 2006; Peterson and Barney, 1952; Sawusch, 1996). Figure 2.1 shows an example of the formant frequencies describing the phoneme in the English word "put".

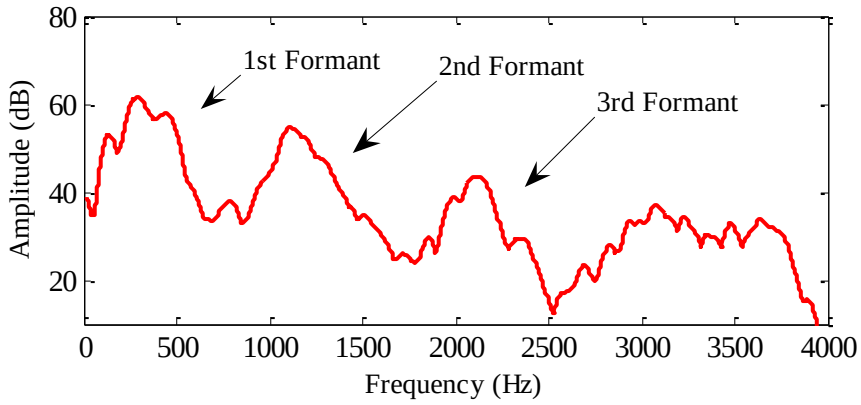

Figure 2.1. Example of formant frequencies depicted by the FFT (Fast Fourier Transform) of the vowel part of a recording of the word "put".

Delattre et al. (1952) confirmed the importance of formants for vowel recognition by synthesizing 16 vowels with the use of a pattern playback instrument that converts spectrograms to sound. The aim was to find the number of formants and the best frequency locations of each formant for optimum vowel recognition. In the first set of experimental vowels, the frequency of F1 was fixed at four different values while F2 was varied

systematically for each of the four different F1 values. In the second set, F2 was fixed at four frequencies while F1 was varied. The authors selected the vowels that, to their judgment, sounded similar to 16 well-known vowels. These vowels were then synthesized and presented at random order to listeners. The results indicated that the synthetic vowels were recognized to a large extent and that formants were adequate for representing vowel sounds.

In line with the study of Delattre et al. (1952), other researchers also found both F1 and F2 to be important for recognizing vowels (Carlson, Fant and Granstrom, 1975; Hillenbrand et al., 1995; Miller and Nicely, 1955; Peterson and Barney, 1952). More recent evidence exists, however, that the F1 and F2 do not contribute equally to vowel perception.

Kasturi, Loizou, Dorman and Spahr (2002) found a greater influence of F2 on vowel recognition than F1. Vowels were synthesized as the sum of sine waves, while gaps were created in the spectra of the vowels. A significant reduction in vowel recognition results was found when the spectra of vowels in the F1 or F2 region were deleted. The deletion of F2 spectral regions, however, led to a greater reduction in vowel recognition than deletion of the other regions. The conclusion was made that listeners need access to proper F2 information, while only partial information regarding the F1 spectral region is sufficient for proper vowel recognition.

Ito et al. (2001) suppressed F1 and F2 in synthetic vowels, while keeping the original spectral shape intact. The suppression of F1 was found to have a greater effect on recognition than the suppression of F2. Listeners were therefore more reliant on F1 than F2 when spectral shape information was still available

The importance of formants also depends on the formant frequency locations, which differ between different vowels. The human auditory system processes vowel formants in different ways, depending on the position of a vowel's formant frequencies. Vowels with the first two formants closely spaced are perceived similar to vowels containing one broad

spectral peak located at the average frequency of the two formants (Chistovich and Lublinskaya, 1979; Nossair and Zahorian, 1991; Zahorian and Jagharghi, 1993).

The perception of two closely spaced formants as one spectral peak was also tested in the study of Delattre et al. (1952). The intensities of F1 and F2 for synthetic vowels were reduced separately to establish the effects of intensity variations on vowel perception. When the F1 intensity was reduced for vowels with considerable differences between the F1 and F2 frequencies, the identity of the vowel gradually disappeared. This eventually produced a non-vowel sound characterized by a pitch equal to that of the higher formant. The reduction of F1 intensity for vowels with small differences between the F1 and F2 resulted in a different outcome. The vowel sound did not disappear, but was perceived as a different vowel found to be in close proximity to the original vowel in the vowel space. This indicated the tendency of listeners to perceive two closely related formants as a single formant located at the average of the two.

It is also known that the auditory system averages closely spaced F2 and F3 peaks into one broad spectral peak, indicating F3 as an important cue for front vowels (Carlson et al., 1975; Johnson, 1989) but not for back vowels (Klatt, 1982).

2.2.1.2 Studies considering spectral shape versus formants as cues

It is evident from the literature that formants are important for vowel recognition, and that in some cases only one formant is sufficient to identify a vowel. It still remains unclear whether the human auditory system recognizes a vowel sound only by the frequencies of the F1 and F2 peaks, or if the spectral detail, which would usually include the formant peaks, also contributes to vowel perception. The study by Delattre et al. (1952) proved that formant frequencies determined the identity of vowels, but when the intensities of the formants on the pattern playback instrument were altered, identification errors became apparent.

Some studies suggest that formants are better in explaining vowel perception than the whole-spectrum. Klatt (1982) conducted three experiments to investigate the effect of formant frequency changes opposed to acoustic changes on the perception of the phonetic distance between vowels. The vowels /æ/ (containing substantial formant frequency differences) and /ɑ/ (with F1 and F2 frequencies closely related) were manipulated by altering formant frequencies, amplitudes and spectral tilt. The vowels were also subjected to high-pass, low-pass and notch filtering. For both vowels, phonetic distance measures among vowels were only influenced by formant frequency changes and not by the other spectral manipulations.

Other studies found the whole-spectrum to be more influential on the perception of vowels than formants. In the study of Zahorian and Jagharghi (1993), the influence of spectral shape and formants were investigated by using automatic vowel classifications for static and dynamic spectral features. The relative importance of certain vowel parts, the fundamental frequency as well as duration and context effects were also investigated.

Firstly, the frequencies, amplitudes and bandwidths of the first five formants were obtained for each frame of each vowel by means of the 10^{th} order linear predictive coding (LPC) spectrum. A formant tracking routine was developed and the results verified by visual inspection of the formant trajectories. For the spectral shape, the discrete cosine transform coefficients (DCTCs) were used to represent the smoothed spectrum. DCTCs differ from cepstral coefficients since it represents the coefficients obtained from the cosine expansion of the spectral magnitude of the signal.

Listening experiments were carried out with the vowel database to see which segments of the vowel signals played significant roles in vowel perception. Results were also compared to the automatic classification experiments. The vowel database consisted of 99 consonant-vowel-consonant tokens, recorded by 30 speakers. Five vowel types, each representing a certain acoustic segment of the vowel, were presented to the listeners. Automatic classification experiments were carried out for the formants and DCTCs. Four classifiers were used to obtain the automatic recognition results for time-varying and static spectra.

The results indicated that both formant and spectral shape information are needed for vowel perception, although certain information appeared to be redundant. Formant frequencies contain most of the information required for recognition, but the spectral envelope provides a more comprehensive description of the vowel. The overall automatic classification results obtained from the formants and DCTCs indicated that DCTC features define vowel identity to a better degree than formant cues.

The whole-spectrum was found to be sufficient for recognition in the three experiments done by Ito et al. (2001). In the first experiment, F1 and F2 for synthesized vowels were alternatively suppressed while the spectrum remained intact. This resulted in three types of synthesized vowels; one type contained all formants, while F1 was suppressed in the second type and F2 suppressed in the third. Vowels were synthesized by varying F1 between 250 Hz (hertz) and 1250 Hz in 125 Hz steps and F2 between 750 Hz and 2250 Hz in 125 Hz steps. These complete spectrum, suppressed F1 and suppressed F2 stimuli were presented to listeners. The results showed that the suppression of the formants did not have a substantial effect on vowel recognition since the suppressed stimuli results did not differ substantially from the control group results (containing all formants). The outcome revealed that F1 and F2 are not exclusively used as cues for vowel recognition.

In the second experiment the first formant frequency and amplitude were varied independently. The amplitude difference between F1 and F2 was varied between 0 dB and 48 dB in 6 dB steps while F1 was varied between 250 Hz and 1250 Hz in 125 Hz steps. It was found that the amplitude ratio of the high to low frequency components influence vowel perception to a certain degree.

In the third experiment the formant amplitude ratios of F1 to F3 as well as F4 to F5 were varied, while the F2 frequency was held constant and the F1 frequency was varied similar to the first experiment. Amplitude ratio did affect vowel perception, even with all the formants available. High nominal F2 stimuli were perceived as front vowels, while low nominal F2 were perceived as back vowels.

Overall from the results of the three experiments, it was found that formant frequencies alone are not sufficient for vowel perception and that the whole spectral shape can define vowels, even if formant frequencies are not available.

To extend the results from Ito et al. (2001), Kiefte and Kluender (2005) investigated the relative significance of detailed and global spectral shape on vowel perception. Detailed formant peaks, as well as spectral tilt were manipulated independently to see if any relationship exists between these cues.

In the first experiment, vowels were synthesized to range from the /i/ to the /u/ vowel. This was done by fixing F1, F4 and F5 and varying F2 between 2400 Hz and 3000 Hz in steps of 100 Hz, while also varying F3 as a linear function of F2. The spectral tilt for each vowel was altered to correspond to that of another vowel. Spectral tilt and formant frequencies jointly and separately, had a substantial influence on the perception of the vowels /i/ and /u/.

In the second experiment, the influence of formants and spectral tilt was investigated for diphthongs. Diphthongs were synthesized to range from the /aI/ diphthong to the /au/ diphthong. F1, F2 and F3 were fixed at the onset, while these values changed linearly to the formant values at the vowel end. Formants at the vowel end were changed in the same way as in the first experiment. Spectral tilt was matched for each 5 ms (millisecond) frame with the spectral tilt of each of the other vowels. Spectral tilt was shown not be an effective perception cue for vowels with spectral change through time.

In the third experiment, stimuli from both the first and second experiments were combined by using the same vowel end values as in the second experiment and onset values obtained from the /u/ vowel as in experiment one. Stimuli therefore ranged from the /u/ monophthong to the /uI/ diphthong with regards to their formant frequencies and spectral tilt. Results showed that listeners identified the stimuli based on spectral change, even though steady state spectral tilt characteristics were included in the stimuli.

From the three experiments it was concluded that both a change in spectral tilt and formants frequencies have a substantial effect on vowel perception for steady state vowels, while this is not the case for vowels with spectral change through time.

Some studies found that listeners use F1 and F2 in conjunction with spectral shape cues to identify vowels. Sakayori et al. (2002) conducted tests to evaluate the importance of F1, F2, F3, spectral envelope and F0 (fundamental frequency) as cues for vowel recognition. Certain spectral regions of naturally spoken vowels were eliminated to observe if the deletion of these regions would lead to an increase in identification errors. Five different vowels from 10 speakers were used in the study. The removal of F0 and F3 did not lead to any errors, suggesting that these cues are not essential for recognition. It was found that the regions containing F1 and F2 are sufficient and also critical for successful vowel recognition.

An interesting finding was for the female voice /u/ vowel and the male voice /e/ vowel, which had similar F1 and F2 frequencies. When these vowels were presented consisting only of the F1 and F2 spectral regions, both were still recognized as different vowels. It was assumed that listeners used information provided by the spectral envelope to correctly identify the vowels. It was found that spectral shape information from the critical spectral region (F1 and F2 region) is primarily used as a cue for recognition, while the remaining spectral region is not essential. From the overall results in the study, it was concluded that listeners use F1 and F2, as well as information regarding the spectral shape in the critical region to successfully recognize vowels.

No major difference between the influence of formants and whole-spectrum cues on vowel perception was found by Hillenbrand, Houde and Gayvert (2006). They examined the relative contributions of spectral shape and formants to speech recognition by asking listeners to identify synthesized speech sounds. Vowels, consonants and sentences were synthesized by implementing the source-filter synthesizer. The original signal was first

classified as voiced or unvoiced, after which the synthesis filter was derived from either the spectral peaks of the signal, or directly from the entire spectral shape. Listeners were asked to identify two types of synthesized speech sounds, one where the whole-spectrum was used to produce it and the other type where peaks in the envelope were used. The outcome revealed no clear evidence that either spectral shape or spectral envelope peaks has the greatest influence on perception.

2.2.1.3 Spectral change through time

Formant movement over time has a significant influence on vowel perception (Assmann and Katz, 2005; Assmann, Nearey and Hogan, 1982; Ferguson and Kewley-Port, 2002; Hillenbrand and Nearey, 1999; Sawusch, 1996). A study was done by Hillenbrand and Gayvert (1993) to investigate how well vowels could be recognized if only static spectral information was present. Listeners had to identify synthesized steady state versions of vowels tokens. Approximately 75% of the vowels were identified correctly. These results were compared to results from tests done with the original vowels, where the error was much smaller. It was concluded that certain information used by listeners to identify vowels were missing in steady state signals.

In Nearey and Assmann (1986), formant changes were modelled over time by measuring the F1 and F2 frequencies at the initial and final part of the vowel. These frequencies were changed around in the time progress of the vowels, which led to an increase in recognition errors. Neel (2004) confirmed the influence of formant change over time by presenting synthetic stimuli with five different stages of formant information over time to listeners. It was found that listeners did not perform well in tokens where the formants were only represented at one and two time locations. Maximal recognition was found for tokens having formant frequencies present at either three or five time locations. An improvement in vowel recognition was therefore obtained with a thorough representation of formant movement.

Assmann and Katz (2005) used vowel synthesis to test the effect of dynamic and static F0 and formant frequencies on vowel perception. The STRAIGHT synthesis system was used to create two types of synthetic vowels. For the first type, the F0 and formant frequencies were varied through time, while these cues were kept constant for the second vowel type. It was found that vowel identification was not influenced significantly by a flattened F0, while the static formants led to a substantial reduction in identification.

Hillenbrand and Nearey (1999) synthesized vowels in /h/-vowel-/d/ context. Two synthesized versions of the vowels were created, one type with dynamic formant movement through time and the other where the formants remained constant. It was found that the importance of formant frequency change varied from one vowel to another. The effect of formant flattening on vowel identification was less in vowels known to originally have minor formant frequency movement.

2.2.1.4 Spectral contrast

Spectral contrast, which can be defined as the difference between the amplitude peaks at a formant frequency and its adjacent valley, are known to be more important for CI users, but definitely also needs to be taken into account for normal-hearing listeners (Ter Keurs et al., 1992).

Leek et al. (1987) synthesized vowels by adding 30 harmonics from a tone of 100 Hz. The amplitudes of each harmonic were identical, except for the components in the region of the first three formants. These components differed in steps of 1 dB (from 1 – 8 dB) from the normal harmonics. The formants were chosen to match those for the vowels /æ/, /ɑ/, /i/ and /u/. The vowels were presented to normal-hearing listeners in quiet and in low-pass filtered white noise (at an appropriate SNR that would simulate the hearing-impaired) and for hearing-impaired listeners only in quiet. The results showed that normal-hearing listeners in noise needed only 1 dB to 2 dB spectral contrast (amplitude differences

between formants and normal harmonics), while 4 dB was needed in noise. For the hearing-impaired, 6 dB to 7 dB spectral contrast was needed.

2.2.1.5 Vowel duration

The duration of a vowel is a characteristic that influences the perception of vowels (Ainsworth, 1972; Bennett, 1968; Hillenbrand, Clark and Houde, 2000; Sawusch, 1996). In Hillenbrand et al. (2000), vowel tokens in /h/-vowel-/d/ context were presented to listeners. Four different duration values for each vowel were generated to determine the effect on vowel recognition. The outcome revealed that duration only had a minor influence on recognition of the average vowel token, although it did significantly influence vowel perception for a small group of vowels.

Sawusch (1996) conducted two experiments to investigate the effect of duration on vowel recognition. The two vowels /ɛː/ and /æ/, which contain identical formants at the midpoint of the vowels, were used in the study. In the first experiment, the duration of the two vowels were interchanged, while recognition results indicated no effect of duration on the perception of the vowels. In the second experiment, the first three formants of the synthetic vowels were generated to change from those appropriate for /æ/ to those appropriate for /ɛː/. Duration was also altered by creating two tokens that reflect the longer duration and two tokens reflecting the shorter duration. It was found that duration did play a major role in vowel recognition in the second experiment. The overall conclusion was made that duration is more important for vowel perception when other important cues are unavailable.

2.2.1.6 Consonantal and sentence context

The consonantal context of vowels is important for vowel identification (Nearey, 1989). The dynamic spectral content of a vowel in the initial and final transitional part of the token (in utterances of consonant-vowel-consonant context) is used as a cue for vowel perception. Experiments were carried out where the center of the vowel tokens was

attenuated to silence while the initial and final transitional parts of the token were kept intact. It was found that listeners could still accurately identify the vowel tokens. When only the initial or final parts of the transition were presented apart from each other, identification errors increased significantly. It was concluded that the initial and final transitional parts of vowel token taken together, contain sufficient information needed by listeners to identify vowels (Strange, Jenkins and Johnson, 1983).

When listeners are provided with vowels in sentence context, better vowel identification is obtained (Verbrugge, Strange, Shankweiler and Edman, 1976). With different speakers, better vowel recognition in sentence context is found than without it. This is due to the fact that the sentences allow listeners to adjust to the speaker speech rate, rather than to compensate for different speakers.

2.2.2 Modelling vowel perception in quiet

Modelling vowel perception requires the knowledge of vowel cues important for recognition. Usually a distance metric is computed to define a difference in perception between two vowels, or the probability of the vowels being confused. Some studies focus on the modelling of confusions among vowels, while others only aim to predict the probability of correctly recognizing a vowel.

In the classic study of Peterson and Barney (1952) vowel perception was modelled by locating the vowels in a simple vowel space. The authors used 76 speakers and 70 listeners in an experiment where random vowel tokens were recorded on a magnetic tape. The speech signals were then played back to the listeners prompting them for a response. The signals were also analyzed with a spectrogram, where the first three formant frequencies were determined and analyzed. The vowels were plotted in a vowel space described by F1 and F2. It was seen that phonologically similar vowels were grouped together in this vowel space, while dissimilar vowels were separated more.

Several authors have cited Peterson and Barney (1952) in studies of vowel recognition models, where F1 and F2 were shown to predict vowel perception. Hillenbrand et al. (1995) attempted to replicate and extend the Petersons and Barney study by recording vowels in the context of /h/-vowel-/d/ using 45 male, 48 female and 46 children speakers. Formant frequency measurements were made similar to the Peterson and Barney study, with additional measurements of formant frequency contours and duration. Vowel tokens were presented to listeners and results were found to be similar to the results of Peterson and Barney (1952). It was also found, however, that vowel perception could not be predicted well enough by static F1 and F2 information, and that the classification accuracy can be improved by the addition of the duration cue and a coarse representation of spectral change.

Molis (2005) evaluated two models of vowel perception, one based on formants and one on whole-spectrum cues. 54 five-formant vowels were synthesized so that the F2 and F3 frequencies varied orthogonally, while the other formants were held constant. The formants were modified in such a way that the three vowels present in the words “hid”, “hood” and “heard” were included. The stimuli were presented to listeners, who were asked to distinguish between the three vowels mentioned. The results of the listening test were used as the input to logistic regression analysis for evaluation of the models for vowel perception. For each model, a set of parameters were developed as the prediction variables for the logistic regression. For the formant-based model, the first three formant frequencies and amplitudes were measured from the synthetic vowels. Linear and non-linear combinations of these measurements were used as parameters for the different models.

For evaluation of the whole-spectrum model, three different approaches were followed. In the first method, excitation patterns derived from the frequency analysis in the human auditory system were used. A bank of overlapping critical band filters were used to filter the magnitude spectrum of the 1024-point FFT of the input signal. The excitation that occurs in the cochlea was represented by the summed output of the filter bank. The second method measured the spectral slope from the excitation pattern, while the third derived the

perceptual linear prediction (PLP) cepstral coefficients. These coefficients describe auditory processing by the method of critical band filtering that includes the intensity to loudness conversion done by the auditory system.

The results of the study indicated that formant peaks alone did not provide a good fit to the data, but improved when formant amplitudes were added. The set of input parameters of F2, F3, $F2^2$, $F3^2$ and $F2 \times F3$ provided the best representation for the probabilities of correct vowel identification. The excitation pattern model was found to marginally describe vowel perception. It was concluded that the formant frequencies, with included quadratic terms, provide the best representation of vowel perception. Whole-spectrum models should, however, not be abandoned.

The findings of Ito et al. (2001) implicating the whole spectral shape as an important cue for vowel recognition led them to propose a simple perception model based on the vowel spectral shape. The model consists of a database of different spectral shapes for each vowel. It accepts the spectral shape of a vowel as input and compares it to a database of different spectral shapes to determine the identity of the vowel. This model is far from perfect, as it cannot explain a human's adaptability to different speakers since spectral shapes of vowels differ for different speakers.

Overall, vowel perception models in quiet are mainly based on either spectral shape, formant frequencies or some combination of the two. Spectral properties are preferred while temporal characteristics like duration are only incorporated in some instances.

2.3 VOWEL PERCEPTION FOR COCHLEAR IMPLANT USERS

The aim of the present study is to define vowel cues that are important in the perception of vowels in severely degraded conditions. Knowledge of the cues that listeners extract from degraded vowels would be valuable in predicting a listener's response for vowels

processed through cochlear implant acoustic models, which would ultimately provide a first step towards the prediction of vowel perception for CI users. Vowel perception cues and present perception models for CI users will be discussed next.

The success of cochlear implants is mostly user specific. Depending on factors such as insertion depth of the electrodes, period of deafness and the location of the remaining auditory neurons, recognition results can vary between users. Different CI processors also have a substantial effect on recognition because of the variations in rate of stimulation (Liu and Fu, 2007).

In the light of this information, several studies have investigated the factors that could cause inaccurate vowel identification in cochlear implants (Friesen, Shannon, Baskent and Wang, 2001; Fu and Shannon, 1999; Loizou and Poroy, 2001; Van Wieringen and Wouters, 1999). Other studies used CI acoustic models to simulate the speech processing of a typical CI (Dorman et al., 1998b; Liu and Fu, 2007; Remus and Collins, 2005). These models enable the involvement of normal-hearing listeners in speech intelligibility tests rather than cochlear implant users, who are not always available. When normal-hearing listeners are presented with speech tokens processed through an acoustic model, the sounds would appear to be severely degraded and noise corrupted. Factors influencing vowel perception for CI users become apparent when vowels are represented through acoustic models to normal-hearing listeners.

2.3.1 Cues used by CI users in recognizing vowels

The common factor that influence perception of vowels for CI users is the number of spectral channels (Dorman, Loizou and Rainey, 1997b; Fishman et al., 1997; Shannon, Zeng, Kamath, Wygonski and Ekelid, 1995; Xu, Thompson and Pfingst, 2005). In Fu, Shannon and Wang (1998), the loss of spectral resolution in cochlear implants for vowel and consonants was simulated by producing these tokens to normal-hearing listeners through an acoustic model. Different numbers of frequency bands were simulated by

altering the frequency resolution of the model. It was concluded that better recognition of vowels and consonants in noise for CI users could be obtained by increasing the number of implant channels for better spectral resolution.

Spectral contrast is the difference between the amplitude peak at a formant frequency and the valley adjacent. Loizou and Poroy (2001) tested the relationship between spectral resolution and spectral contrast. It was found that both spectral resolution and spectral contrast have a substantial influence on recognition accuracy. The main outcome of the study suggested that for sufficient vowel identification, a larger spectral contrast is needed when the spectral resolution of vowels is poor. Since the spectral resolution for people fitted with cochlear implants is limited, they require a larger spectral contrast than normal-hearing listeners.

Some of the cues that are important for normal-hearing listeners are also of importance for CI users (Dorman et al., 1998a). Important cues for normal-hearing listeners were found to be the first two formant frequencies, with duration and formant movement acting as secondary cues (Hillenbrand et al., 1995). A study that investigated the importance of the four abovementioned cues for CI listeners were carried out by Iverson et al. (2006). The effect of formant movement and duration on vowel recognition was tested for CI users and normal-hearing listeners listening to a simulation of cochlear implants. Vowels in /h/-vowel-/d/ context were recorded and resynthesized to remove formant movement and equate duration. For the normal-hearing subjects, the stimuli were processed through an eight, four and two channel noise-vocoder simulating the continuous interleaved sampling (CIS) processing of a CI.

The results showed that the removal of formant movement lowered recognition accuracy by an average of 13%, the altering of the duration cue lowered it by 14% and a combination of the manipulated cues lowered the recognition scores by 29.4%. Recognition scores also decreased when the number of channels in the noise-vocoder decreased. The results showed that formant movement and duration had a substantial

influence on vowel recognition for both CI users and listeners of noise-vocoder simulations.

In a second experiment, a goodness-optimization experiment was performed to map vowels in their vowel space using F1, F2, formant movement and duration as parameters. Listeners were asked to search along the different vectors in the vowel space until the specific vowel becomes a good representation of the vowel presented on the screen. It was found that there were no substantial differences between the vowel spaces of CI users and normal-hearing listeners. F1 and F2 were the consistent parameters chosen by most listeners, while formant movement and duration were also preferred, but to a lesser extent.

Kirk, Tye-Murray and Hurtig (1992) found that both normal-hearing listeners and CI users could recognize vowels above average, based only on the vowel formant transitions between the consonants and the vowel (where the centre vowel portion of the consonant-vowel-consonant token was deleted).

2.3.2 Modelling vowel perception for CI users

Knowing the cues that are important for vowel recognition in degraded conditions will ultimately contribute to the improvement of vowel perception models for CI users. To explain the context of these models in the current study, existing vowel perception predictors of acoustic model and CI recognition will be discussed.

2.3.2.1 Predicting CI perception through acoustic models

In the study of Remus and Collins (2004), vowels and consonants were presented to normal-hearing listeners through acoustic models of two CI processors. The main objective was to measure similarities between speech tokens for prediction of confusion patterns for vowels and consonants. Three signal processing methods based on prediction metrics were used to predict the results of confusions tests, namely token envelope correlation (TEC),

dynamic time warping (DTW) and hidden Markov models (HMM). From these three, the cepstral-based DTW and HMM methods performed better than the strictly temporal TEC method. It was concluded that signal processing methods involving the Mel-cepstrum description of the vowel and consonant signal are better predictors of confusions than techniques that incorporate the temporal envelope.

2.3.2.2 The multidimensional phoneme identification (MPI) model

The MPI model (Svirsky, 2000) is used to predict the perception of vowels for CI users. The model predicts phoneme identification based on a listener's discrimination according to certain perceptual dimensions. It utilizes an internal noise model for basic sensitivity, a decision model for response bias and a multidimensional perceptual space. It was found that vowel perception for CI users can be accurately explained by including one spectral cue (vowel amplitude spreading among channels) and one temporal cue (time domain representation of the signal) in the identification model.

2.3.2.3 Neural networks

Chang et al. (2001) developed a neural network that could predict a CI user's performance in a vowel identification test. An adaptive neural network structure was used, which is based on system identification applications where the input and output of an unknown system (in this case the auditory system of the user) are known. The inputs to the unknown system were the vowel stimuli processed through the CIS strategy while the outputs were the responses of the CI listeners.

The proposed model is believed to be simpler than the model explained previously in Svirsky (2000) as no psychophysical measurements are needed. This model only requires the input-output matrix depicting the response to vowel stimuli of each user, and has the benefit of optimizing the processing of a CI. It was found that the model could closely predict the vowel confusion results for each user for which it was trained to model. The

results of the neural network were used to adjust the CI channel amplitudes for the users, with a significant improvement found for vowel recognition.

2.4 VOWEL RECOGNITION IN NOISE

It is hypothesized in the present study that listeners may still use formant cues to identify vowels in severe noise. Vowel recognition in noise will be discussed in further detail.

2.4.1 Factors influencing vowel perception in severely degraded conditions

2.4.1.1 Effects of listening condition, speaker and listener type

The noisy environment in everyday conversation is normally simulated by adding either speech-shaped noise (Parikh and Loizou, 2005; Phatak and Allen, 2007; Summers and Cord, 2007; Wang, Kjems, Pedersen, Boldt and Lunner, 2009), multi-talker babble (Ferguson and Kewley-Port, 2002; Liu and Kewley-Port, 2004b; Parikh and Loizou, 2005) or reverberation (Hedrick and Nabelek, 2004; Nabelek et al., 1992) to speech signals. The type of listening condition influences vowel perception due to the different ways in which it reduces the information transferred to the listener. Noise reduces information due to formant peaks being masked, while reverberation smears the formant peaks through time. Any segment along the duration of the vowel that is lower in intensity than other parts will be masked by noise to a greater extent (Nabelek et al., 1992).

The perception of vowels in degraded speech can be influenced by the speaker. The same vowels of different speakers have different locations in the F1 and F2 vowel space, and would therefore be perceived differently. Another reason for the influence of the speaker is that the energy distribution along the duration of a vowel differs for different speakers (Nabelek et al., 1992).

Perception in noise will also vary among different listening groups like the hearing-impaired or normal-hearing listener types. The perceptual limitations of hearing-impaired listeners influence their ability to differentiate between vowels (Ferguson and Kewley-Port, 2002).

2.4.1.2 Acoustic cues in noise

The previously discussed literature indicated that the formant frequencies and spectral shape of a vowel are important cues for vowel recognition in quiet. The question remains whether these cues are still important in noise, and if there is a difference between the relative importance of F1, F2 and the spectral shape.

Frequency bands related to the cochlea have been used to evaluate the recognition of speech in noise by means of the articulation index (AI). The AI provides a measure to determine speech audibility by taking into account the effect of noise masking in the spectral regions of the critical frequency bands of the cochlea, with preference given to the regions that are known to contain speech information (Allen, 2005).

Formants are known to be the main cues used by listeners for vowel identification in speech-shaped noise (Fattah, Zhu and Ahmad, 2009; Gong, 1995; Nabelek et al., 1992). Despite the fact that noise severely distorts the speech spectrum, the frequency locations of formants are usually well preserved, even when the ratios of the formant peak amplitudes are lost (Assmann and Summerfield, 2004).

Not many published studies have investigated the relative importance of formants compared to spectral shape for vowels in noisy conditions. A first step was taken in the study of Parikh and Loizou (2005), where acoustic analysis was done to investigate the effects of speech-shaped noise and multi-talker babble on the spectrum of vowels. The main focus was to investigate acoustic measurements of formant frequencies and spectral envelopes when masked by noise. Spectral differences and reliable F1 and F2 information

in noise were examined to determine if either formants or spectral shape cues are available to listeners. Results were compared to vowel confusion test results to determine the effect of noise on listeners' perception of these cues.

Two types of acoustic analysis were done on the vowel signals in the study of Parikh and Loizou (2005). Firstly, the critical band spectral differences between the clean and noise-corrupted vowels were measured. A 21-channel filter bank was implemented with center frequencies chosen according to critical-band spacing. The vowel spectra were estimated by calculating the root mean square (rms) energy for each filter bank in each of the 10 ms windows. The spectral differences between the clean and noisy vowels were determined by calculating the Euclidian distance of the filter bank energies at the two spectral regions where F1 and F2 normally occur.

The critical-band spectral differences decreased as the SNR increased. The Euclidian distance measures for the F2 spectral region were significantly higher than the F1 region only at -5 dB SNR when masked by multi-talker babble. For speech-shaped noise the spectral distance measures for the F1 and F2 regions were similar for all SNR levels.

The second analysis focused on the reliable detection of F1 and F2 in the noisy signals. The formants of the signals were measured by using 22-pole LPC spectra, after which the clean LPC spectra were overlaid on the noise-corrupted spectra for an indication of the locations of F1 and F2. Each speech frame was classified as either one of the following: having no F1, having no F2, neither F1 nor F2 present, or F1 and F2 reliably present. For the frames in which F1 and F2 were reliably detected, the absolute difference between the formant frequencies of the clean and noisy vowels were calculated.

F1 was detected significantly more often than F2 for both types of noise. F1 and F2 were only reliably detected more than 50% of the time for SNRs above 5 dB. At -5 dB SNR, F1 and F2 detection were significantly higher in speech-shaped noise than in multi-talker babble. No difference was however found between the two types of noise for higher SNRs.

The acoustic measurement results indicated that multi-talker babble and speech-shaped noise affected the spectra in the same way. For the vowel spectral difference measurements it was found that the F2 region was affected most by noise at -5 dB SNR, while F2 was not detected as frequently as F1 for the formant count data. The spectral differences between noisy and clean vowels in the F1 region were relatively small, while F1 was detected more reliably in the formant detection analysis. It was suggested that F1 is more resistant to noise than F2.

The same noise-corrupted vowels were presented to a panel of listeners to determine whether there exists a relationship between the acoustic analysis findings and vowel perception. The listening test results were correlated with parameters obtained from the acoustic analysis of the vowels to obtain a measure of similarity.

From the results it was concluded that F1 and F2 were not exclusively used as cues in severe noise since F2 was severely masked by noise and could not be reliably detected with the formant detection analysis. From the confusion matrices of the listening tests in severe noise, it was concluded that listeners did not rely exclusively on spectral shape cues, as vowel pairs not having the same spectral shape were still confused with one another. This was also confirmed by the insignificant correlation between the whole-spectrum difference metric and the vowel identification scores.

From the overall acoustic analysis, it was suggested that listeners rely primarily on F1 for vowel recognition in noise, since a reasonably clear F1 together with a poor F2 were present in the vowels. This was confirmed from the confusion matrix results, which showed that vowels were confused with vowels containing similar F1 values.

Although F2 was unavailable in noise, there was still some evidence that information from the F2 region was important. Some vowels with similar F1 but vastly different F2 frequencies were not confused with each other, suggesting that some information in the F2

spectral region was used. The conclusion was made that in the absence of sufficient formant frequency information, listeners probably rely on other external cues such as duration, spectral change and formant movement.

The final findings of Parikh and Loizou (2005) regarding cues important for vowel identification in noise were the following:

- The second formant of vowels is heavily masked by noise at very low SNRs (-5 dB).
- Listeners had access to reliable F1 information but unclear F2 information in noise.
- From the correlation of the acoustic analysis and identification scores, it was found that listeners relied on clear F1 information, together with partial F2 information to identify vowels in noise.

In agreement with the finding of Parikh and Loizou (2005) regarding the greater relevance of F1 than F2 in noise, Hedrick and Nabelek (2004) found that a change in the intensity of F2 was not perceived by listeners along the /ʊ/ to /i/ vowel continuum in noise. Synthetic tokens of the vowel /ʊ/ were generated, with the F2 amplitude intensity varied in 5 dB steps from 15 dB to 55 dB. The stimuli were presented to normal-hearing and hearing-impaired listeners in quiet, reverberation and speech-shaped noise. Subjects were asked whether either the vowels /i/ or /ʊ/ was perceived. The aim was to investigate the perception of spectral shape in the F2 region for the two different noise conditions.

It was found that, for the normal-hearing listeners in quiet and reverberation, the /ʊ/ stimuli were gradually interpreted more as the /i/ vowel when the F2 intensity was progressively decreased. For the vowels masked with speech-shaped noise however, the percentage of /ʊ/ responses did not decrease nearly as much as in the quiet and reverberation conditions. It was concluded that spectral shape alterations (in this case a change in the F2 intensity

along the /u/ to /i/ continuum) are perceived in quiet and reverberant conditions but not so clearly in speech-shaped noise.

Although literature indicates that listeners do not rely strongly on F2 in noise, this cue was found to be important for diphthongs. Nabelek, Ovchinnikov, Czyzewski and Crowley (1996) found that the perception of diphthongs in noise and reverberation is related to the intensity of the F2 transition segments. In the first experiment, 16 versions of the /ɑɪ/ stimulus were synthesized, each with the transition segment linearly increasing from 0 dB to the maximum value at the end of the transition segment. The maximum value differed for each of the 16 vowels in 1 dB steps. Synthesized vowels were presented to normal-hearing and hearing-impaired listeners in quiet, noise and reverberation conditions. The outcome revealed that, when the intensity of the transition formants were less than the steady state parts, the stimuli were identified as /ɑɪ/ in silence, while in noise or reverberation the stimuli were mostly identified as /a/.

In the second experiment, 30 vowel utterances for each of the diphthongs were analyzed spectrally. It was found that the intensities of the F2 transition segment were directly related to identification errors of listeners. It was therefore concluded that for diphthongs, the intensity of the F2 transition segment does play a key role in noise and reverberation.

While the two studies above commented on the intensity variations of the formants, Liu and Kewley-Port (2004a) investigated the discrimination of vowel formant frequencies in speech-shaped noise and multi-talker babble. The aim was to find whether discrimination thresholds decreased for a decrease in SNR. Seven different vowels were synthesized with variations in F1 and F2 frequencies. Thresholds for F1 and F2 were measured by a three-interval, two-alternative forced choice procedure that determined the frequency increment that was required for a 71% correct response score.

Results showed a significant influence of the SNR, formant frequencies and noise type on formant discrimination. For SNRs lower than 0 dB, thresholds increased significantly

when compared to thresholds in quiet. Formant discrimination was better in the F1 region due to speech-shaped noise severely masking the F2 region.

2.5 GAPS IN THE KNOWLEDGE OF THIS SUBJECT

The literature revealed various vowel characteristics important for perception of vowels in quiet. Perception models that aimed to predict vowel recognition were also mentioned. The question remains whether these characteristics and perception models are still relevant in noise.

Literature indicated diverse results regarding the question of whether formant frequencies or spectral shape cues are essential for vowel recognition in quiet. Some studies found that formants are more important than the whole-spectrum (Klatt, 1982; Molis, 2005), while other studies found the whole-spectrum to be the more important cue (Zahorian and Jagharghi, 1993). No major difference between the formants or whole-spectrum importance was also suggested (Hillenbrand et al., 2006), while a combination of formants and spectral shape were also suggested to be important (Sakayori et al., 2002). The fact that the abovementioned studies do not portray similar results show that there is still uncertainty regarding this issue.

Not many studies exist in literature regarding the importance of formants opposed to spectral shape in noisy conditions. A first step was taken in the study of Parikh and Loizou (2005) where it was found that listeners neither make use of only formants nor spectral shape information for vowel recognition. The findings were obtained by acoustic analysis where the presence of formants in noise, and the spectral shape differences between the clean and noisy vowels were investigated.

A few questions arise from this study of Parikh and Loizou (2005). It should be asked whether listeners can perceive formants or spectral shape cues in noise, even if it can’t be

exposed by acoustic analysis. An example is given by Parikh and Loizou (2005) where acoustic analysis showed F2 to be totally masked by noise, but that F2 information was indeed perceived by listeners. It is suggested that other cues play a role when formants are completely masked. The possibility should however be considered that formants are indeed perceived when it is totally masked by noise. Original recorded vowels were used in their study, containing both formant and spectral shape data. An alternative to this approach would be to synthesize vowels containing only formant data or spectral shape information.

The relative importance of F1 and F2 should be further investigated in noise. The literature revealed that F1 and F2 cannot always be given the same weight of importance in quiet. In noise, listeners depend on clear F1 but vague F2 information to recognize vowels, due to the F2 region being heavily masked by noise (Parikh and Loizou, 2005). No spectral manipulation was done to confirm this finding.

Hedrick and Nabelek (2004) emphasized the inability of listeners to perceive an intensity change of F2 along the /u/ to /i/ vowel continuum in noise. No tests were done to investigate other vowels or intensity changes for F1. F2 transition segments in noise were found to be important in diphthongs (Nabelek et al., 1996). The question remains whether the role of F2 in noise would depend on the formant structure. F2 may be more important for back vowels since noise do not mask the lower frequencies as heavily as the higher frequencies.

In quiet conditions, some studies have shown that F1 alone could be sufficient for vowel perception in front vowels (Dubno and Dorman, 1987), while back vowels could be represented by a single peak representing the average between F1 and F2 (Chistovich and Lublinskaya, 1979; Delattre et al., 1952; Nossair and Zahorian, 1991; Zahorian and Jagharghi, 1993). It would be of interest to study these findings in the presence of noise. Ito et al. (2001) showed that neither the suppression of F1 nor F2 in synthesized vowels significantly changed recognition in quiet if the spectral shape was kept intact. It is of

importance to find if this is also true in noisy environments, or if recognition scores will decrease significantly when formant peaks are suppressed.

It was suggested that other cues, like duration, spectral change and formant contour, could possibly play a role in noise (Parikh and Loizou, 2005). The influence of duration and spectral change is well known for recognition in quiet. It is however unknown if these cues play an important role in noise and if these cues would gradually become irrelevant with a decrease in SNR.

2.6 OBJECTIVE OF THE PRESENT STUDY

A key objective of the present study is to extend the work of Parikh and Loizou (2005) regarding the influence of noise on vowel cues. Parikh and Loizou (2005) reported vowel recognition scores as high as 75% at SNR levels of -5 dB, where it was found that listeners still depend partially on formants for vowel recognition in noise. However, at noise levels lower than -5 dB, the role of formants is still unknown. Presenting vowel tokens consisting of only formant information to listeners in severe noise will give an indication of formant importance in these conditions. As alternative to strict formant information, vowel stimuli consisting of the detailed spectral shape will indicate if listeners require a more complete description of the vowel spectrum in noise than is provided by formants.

An efficient method to test for formant cues will be to perform psychophysical experiments with synthesized vowels. Vowel synthesis allows control over acoustic parameters, permitting the manipulation of certain cues in the process. If vowels can be synthesized by using either strict formant information or by taking only the whole-spectrum into consideration, a comparison of the vowel identification scores for the two types of vowels will indicate whether formants are still important in severe noise. If this is not the case, listeners would probably prefer the more complete description of the vowel

spectra provided by the whole-spectrum vowels. To date, a similar experiment has only been done in quiet (Hillenbrand et al., 2006).

An effective way to determine whether listeners still perceive formant information from heavily masked formant regions in noise will be to monitor the response to vowels having a shortage of either F1 or F2 information. If the conclusion of Parikh and Loizou (2005) that listeners have access to accurate F1 and partial F2 information in noise is correct, recognition scores should decline rapidly when F1 is suppressed, while suppression of F2 should not influence vowel recognition to the same extent. Formant suppression experiments have been performed in quiet conditions (Ito et al., 2001; Kasturi et al., 2002; Sakayori et al., 2002) but as far as is known, have not been conducted in noise. The inability of listeners to perceive a formant intensity manipulation in noise is known (Hedrick and Nabelek, 2004), but is limited only to the second formant and was tested only for one vowel at a relatively high SNR.

In the present study, the relative importance of formants compared to spectral shape was investigated under conditions of severe noise. Vowel stimuli were generated in a way similar to the study of Hillenbrand et al. (2006), where two source-filter synthesizers were used to synthesize vowels either according to the first three formants, or by utilizing the entire spectral envelope. The synthesized vowels were also spectrally manipulated so that F1 and F2 were suppressed alternatively. Stimuli were presented to listeners in different levels of speech-shaped noise. Results gave an indication of the importance of formants compared to spectral shape in severe noise, as well as the individual importance of F1 and F2.

CHAPTER 3 METHODS 1: SIGNAL PROCESSING

3.1 CHAPTER OBJECTIVES

The objective of the present chapter is to expand on the signal processing techniques implemented to synthesize vowels according to specific vowel cues. It is explained how the source-filter model of vowel synthesis is used to synthesize vowels either according to detailed spectral shape information or formant peaks. The method by which F1 and F2 for both types of synthetic vowels are suppressed is also described, after which the process of adding noise to the vowel tokens is explained.

3.2 INTRODUCTION

Chapter 2 showed that insight into whether listeners rely on formant or whole-spectrum information for vowel recognition in noise may be obtained by developing two types of stimuli. The first type should contain whole-spectrum information, while the second type should contain only formant information. The most advantageous way to isolate cues for vowels is to synthesize vowels, so as to control the spectral information that the vowel contains.

Vowel synthesis has been used in numerous studies investigating speech perception (Assmann and Katz, 2005; Ito et al., 2001; Kiefte and Kluender, 2005; Klatt and Klatt, 1990). Vowel synthesis methods simplify the control of frequencies, amplitudes and bandwidths of formants, detailed spectral shape, fundamental frequency and spectral movement through time. Hillenbrand et al. (2006) made use of speech synthesis to compare the importance of spectral shape and spectral envelope for speech recognition. Two synthesis methods, one preserving only formant information, and another utilizing the spectral envelope, were used to produce the stimuli. A similar approach was followed in

the current study to investigate the importance of formants versus spectral shape in noise.

Vowel synthesis is a process where synthetic vowels are produced by an approximation of the way the human vocal system operates. In the human vocal system, voiced sounds are produced by vocal cords forcing sounds through the glottis. The vocal cords are adjusted to oscillate in such a manner to produce quasi-periodic pulses of air. The spectral properties of the sounds from the vocal cords are determined by the shape and dimension of the vocal tract, which defines the type of phoneme that is produced. For each shape of the vocal tract, different formant frequencies are produced.

The mechanism of speech production described above can be reproduced by the linear source-filter model depicted in Figure 3.1 (Assmann and Summerfield, 2004; Rabiner and Schafer, 1978). The source consists of a series of single-sample pulses (representing the quasi-periodic pulses of the vocal cords), with period equal to the instantaneous fundamental period of the vowel that is synthesized. For the production of an unvoiced sound, the amplitudes of each pulse must either be zero or non-zero with a probability of 0.5 at each sample point. A_v represents the overall amplitude of the excitation.

The filter (vocal tract) corresponds to the spectral envelope of the vowel that needs to be reproduced. From the spectral envelope, a finite impulse response (FIR) filter is obtained to filter the flat spectrum of the source signal. For the production of a naturally spoken vowel, the value of the fundamental frequency and the shape of the filter should be adjusted to change over time (Hillenbrand et al., 2006).

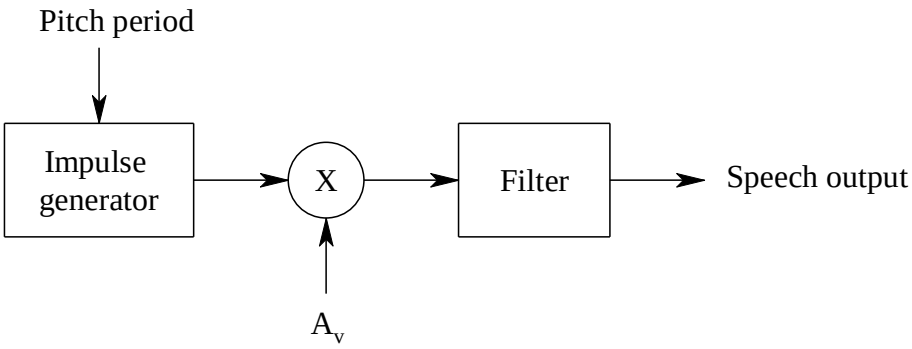

Figure 3.1. The source-filter model for vowel synthesis (Rabiner and Schafer, 1978).

In this study, the spectral shape of the filter is determined by the two different ways in which spectral information had to be represented (whole-spectrum or formants-only).

3.3 VOWEL SYNTHESIS METHOD

Figure 3.2 provides a detailed description of the method used to synthesize the vowel sounds. Naturally spoken vowels are recorded to serve as a reference to the synthesized vowels. The vowel part of the original recorded vowels in consonant-vowel-consonant context is extracted, after which synthesis is carried out according to either the spectral shape or formants of the vowels. For each synthesis method, spectral manipulations are also carried out to suppress either F1 or F2. After the synthesis process, synthetic consonants are appended to the vowels and speech-shaped noise is added at different SNRs.

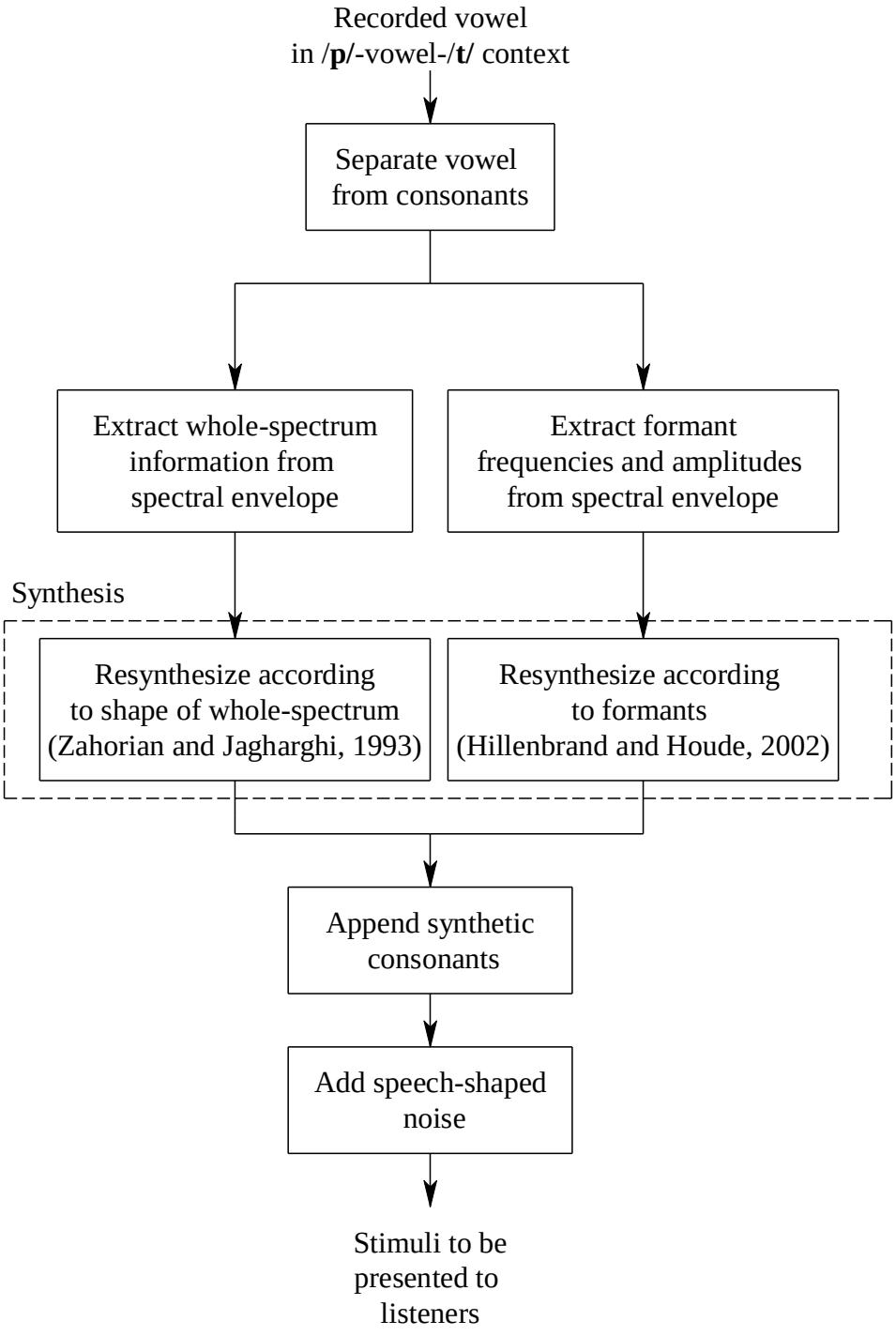

Figure 3.2. Block diagram describing the steps used in the vowel synthesizing process.

3.3.1 Recorded vowels

The synthesized vowels were based on 12 originally recorded Afrikaans vowels in /p/-vowel-/t/ context. Two speakers were used, one male and one female. Each vowel was repeated three times and spoken directly after a word containing the same phoneme as the token being recorded. For each of the three vowel tokens, the vowel with the clearest formants was selected in Praat (Boersma and Weenink, 2004) for use in further signal processing. The vowels, with explanations in brackets depicting the recorded Afrikaans tokens and descriptive English words are: /ɑː/ (paat, f**a**ther), /ɑ/ (pad, c**u**t), /æ/ (pat, c**a**t), /ɛ/ (pet, g**e**t), /e/ (peet, b**ee**r), /ɛː/ (pêt, h**ea**d), /i/ (piet, k**i**t), /ə/ (pit, **a**way), /u/ (poet, c**ou**ld), /ɔ/ (pot, n**au**ght), /œ/ (put, f**u**r) and /yː/ (puut, b**ea**d).

3.3.2 Separation of vowels from the consonants

Each vowel was separated from its initial /p/ and ending /t/ consonant with Praat. Figure 3.3 depicts the time domain and spectrogram representation of a word containing the vowel /e/ plotted with Praat. The vowel (or voiced) part of the token is clearly distinguished from the rest of the token by the pitch indicator (bottom blue line), as well as the pulses in the top waveform indicating the periodicity of the token. The fundamental frequency gave an indication of the initial and final part of the vowel.

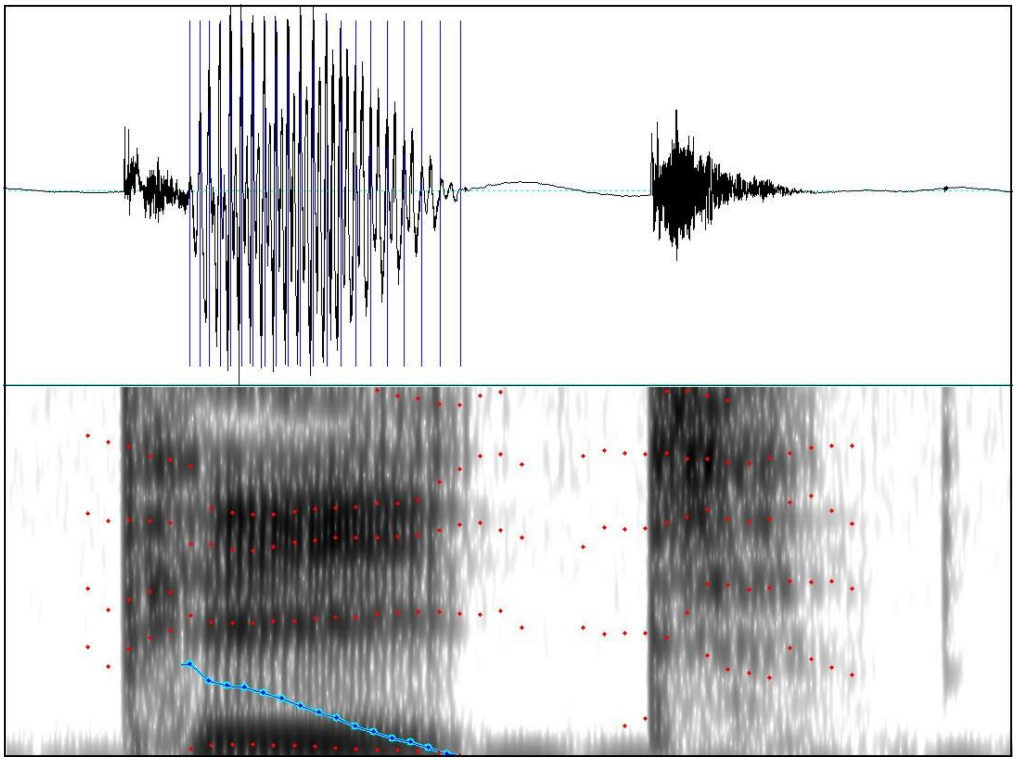

Figure 3.3. Time domain representation (top) and spectrogram (bottom) of the vowel token /ɐ/ spoken in /p/-vowel-/t/ context and plotted with Praat.

3.3.3 Down sampling and segmentation

The original recorded vowels were sampled at 44 kHz. Because most information in speech falls beneath 3800 Hz (Zwicker and Fastl, 1999), it was adequate to sample the vowel part of the token down to 8000 Hz for a Nyquist frequency of 4000 Hz. A standard build-in linear-phase low-pass FIR filter with a cutoff frequency of 4000 Hz was therefore implemented by Matlab prior to down sampling.

The analysis and synthesis of vowel signals require the signals to be segmented into overlapping blocks of an even number of samples. Overlapping of speech segments is necessary to avoid rapid adjustments from one block to another. Segmentation models the nature of human perception and typical segments in a speech signal are 15 ms to 40 ms in

length (ETSI EG 201 377-1, 2002). The vowel tokens were segmented into blocks of 256 samples (32 ms) with 128 (50%) samples of overlap for each segment. It was found that a frame size of 256 samples was adequate, since the lowest pitched vowel still contained adequate information about the fundamental frequency in each segment.

Segmentation can be seen as the multiplication of each segment by a rectangular window. Because the frequency response of a rectangular window contains high side lobes, it is normally not recommended to make use of this type of window. Each segmental block was therefore multiplied by a Hamming window, which reduces the endpoints of each segmental block to zero to prevent spectral leakage. The Hamming window is defined by

$$h(k) = 0.54 - 0.46\cos\left(\frac{2\pi k}{N}\right) \tag{3.1}$$

where N is the number of samples in each window and k the specific sample being multiplied.

3.3.4 Synthesis of vowels using spectral shape features

To reproduce vowels by taking the entire detailed spectrum of the vowel into consideration, the smoothed overall spectral shape approximating the original vowel spectrum was used for the FIR filter in the source-filter model. Cepstral coefficients were calculated in a different way than it is usually done to provide such a smooth approximation. In contrast to the usual method, the coefficients were calculated by means of the cosine expansion of the magnitude spectrum of the signal being analyzed. Since the discrete cosine transform (DCT) was used to calculate the coefficients, this method of computing the smoothed approximation of the spectrum is also referred to as "the calculation of the DCT coefficients (DCTCs)". The calculation of these coefficients relates closely to cepstrum coefficients, except for the way in which frequency range selection, frequency scaling and amplitude scaling are done (Nossair and Zahorian, 1991). DCT coefficients were used instead of an LPC spectral representation. The latter approximates the speech spectrum by emphasizing the spectral resonances, while no effort is made to

provide a good estimate of the spectral detail between formants. The DCTCs provides a smoothed representation of the spectral shape, similar to the LPC spectrum, but more accurate. For the remainder of this document, this type of synthetic vowels will be referred to either as the “whole-spectrum vowels” or the "DCT vowels". Figure 3.4 shows the consecutive steps that were followed in the calculation of the DCTs.

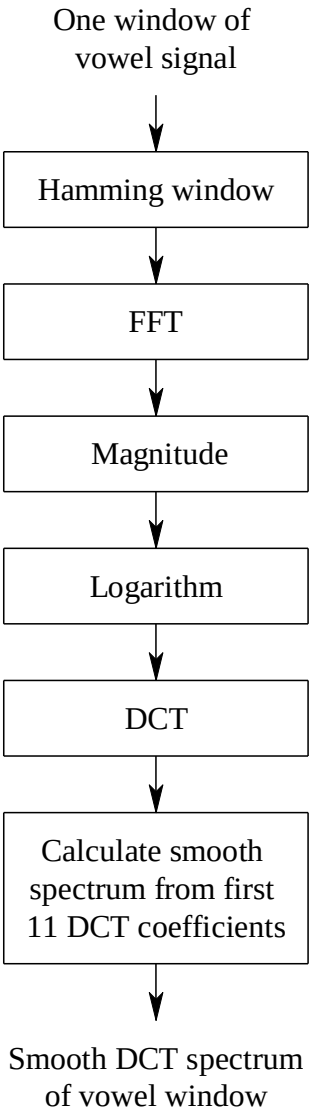

Figure 3.4. Sequential steps followed in the calculation of the DCTCs (Assmann and Summerfield, 2004; Molau, Pitz, Schluter and Ney, 2001).

As was explained in 3.3.3, the speech signal was first segmented into different windows of 32 ms each, after which each window was multiplied by a Hamming window. This was necessary as the windowing (and abrupt cutoff) of the signal introduced aliasing artifacts to the high frequency components of the signal.

A 512 point Fast Fourier Transform (FFT) was computed for each Hamming-windowed speech frame of the original recorded vowel sound. The logarithm of the magnitude of this complex-valued FFT was then computed, after which the DCT coefficients were calculated as in the equation

$$a_m = \frac{2}{N} k_m \sum_{n=0}^{N-1} X(n) \cos\left[\frac{(2n+1)(m-1)\pi}{2N}\right], \qquad m = 1, \cdots N, \tag{3.2}$$

with

$$k_m = \begin{cases} \frac{1}{\sqrt{2}}, & m = 1 \\ 1, & m \neq 1 \end{cases}. \tag{3.3}$$

The coefficients are depicted by a_m while N represents the number of coefficients needed (Watson and Harrington, 1999). $X(n)$, with n =1 up to half the number of FFT taken, are the logarithm of the magnitude spectrum. The final smoothed spectrum was obtained by using the first 12 coefficients in the DCT expansion of

$$[H'(f)] = \sum_{m=1}^{m=N} a_m \cos[(m-1).\pi.f] \tag{3.4}$$

where $[H'(f)]$ represents the frequency response with logarithmic scaled amplitude over a selected frequency range (0-8000 Hz), and N the order of the DCT (Nossair and Zahorian, 1991). The smoothness of the spectrum depends on the number of DCT coefficients used. Figure 3.5 depicts the FFT, LPC and DCT spectrum of one window for the vowel /e/.

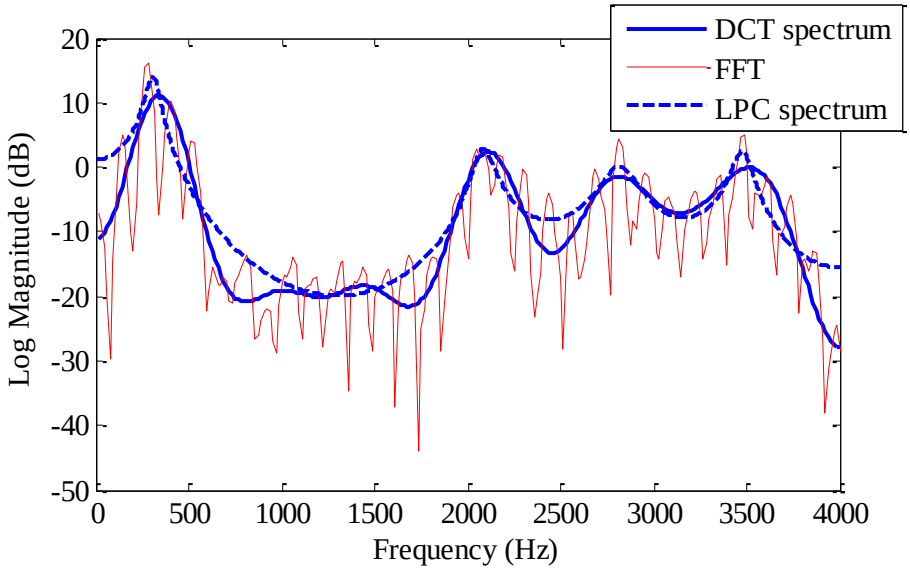

Figure 3.5. FFT, DCT and LPC spectrum of the vowel /ɐ/.

3.3.5 Synthesis of vowels based on formants

In the second method of synthesis, the filter was constructed based on the first three formants of each vowel. This differs to the first method, where the whole-spectrum was taken into account. The impulse response of the filter was obtained by summing three exponentially damped sinusoids at frequencies and amplitudes corresponding to the three formants in each window of each vowel. This method of synthesis is called the damped sine wave synthesizer (DSS) (Hillenbrand and Houde, 2002; Hillenbrand et al., 2006). This type of synthetic vowels will be referred to as either the “formants-only vowels” or the "DSS vowels" for the remaining of this document.

Figure 3.6 describes the steps that were followed in the calculation of the impulse response for the DSS.

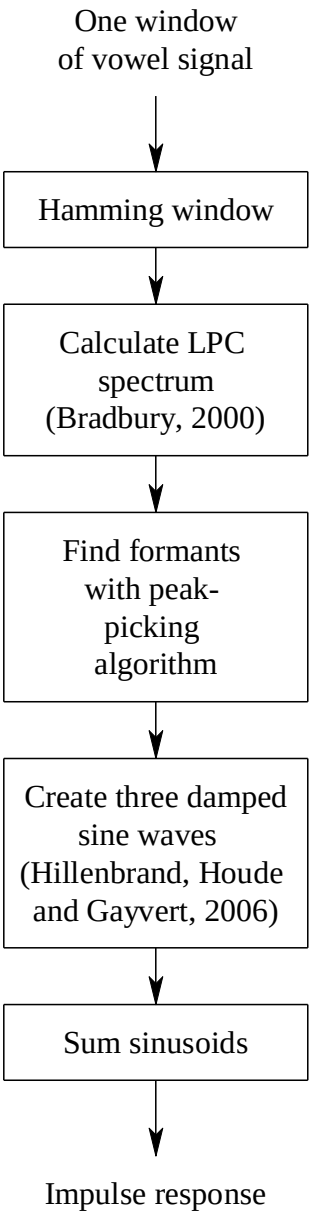

Figure 3.6. Steps implemented for the DSS.

For each window of speech (32 ms), a Hamming window was multiplied with the signal to remove aliasing artifacts. The smoothed LPC spectrum for each window was used to obtain the frequencies and amplitudes of the formants for each vowel.

3.3.5.1 Calculation of the LPC spectrum

Linear Predictive Coding is described as a way to encode analogue human speech in a digital manner. It relies on the fact that speech is produced as a result of excitation of the vocal tract by the vocal cords. This mechanism can be represented by a time-varying filter (the vocal tract) and a signal generator (the vocal cords). The linear predictive model is therefore a mathematical estimation of the vocal tract, and is defined as a method where, at a particular time *t*, the speech sample *s(t)* is represented as a linear sum of the *p* previous samples (Makhoul, 1975).

Figure 3.1 depicts the source-filter model for encoding a vowel signal. LPC calculates an all-pole filter as part of the model in order to mimic the vocal tract filtering of glottal pulses. The smoothed spectrum of the all-pole filter was used to determine the formant frequencies of vowel signals.

An automatic peak-picking algorithm was implemented in Matlab to obtain the formants for each time window from the LPC smoothed spectrum. Peaks were detected by sequentially comparing the amplitude value at each frequency of the spectrum with the amplitude of the preceding and following frequency as is depicted by the equation of

$$Y(f-1) < Y(f) > Y(f+1) \tag{3.5}$$

where $Y(f)$ depicts the LPC amplitude response at frequency *f*. A frequency value *f* is classified as a formant frequency if the amplitude response $Y(f)$ at the specific frequency exceeds the amplitude response at the previous and following frequency.

If a spectral peak was detected for a specific time-window, the entire vowel signal was manually inspected to see if the peak formed part of the formant trajectory through the time windows of the specific vowel (Zahorian and Jagharghi, 1993). Figure 3.7 shows the formants for each window of the vowel /æ/. Peaks that could wrongly be labeled as formants are encircled. The formant values obtained from the peak-picking algorithm were

confirmed by also analysing the vowels in Praat.

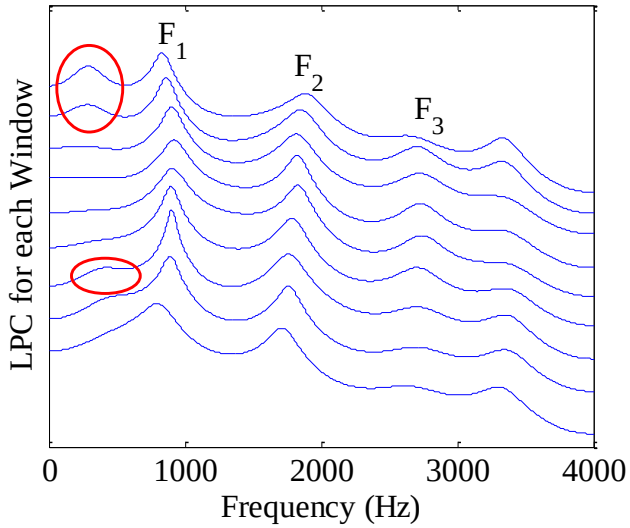

Figure 3.7. LPC spectrum for different time windows of the vowel /æ/ (female voice). Spectral peaks that may be wrongly labeled as formants are encircled.

Having determined the formant frequencies and amplitudes, three sinusoids were calculated for each window of each vowel. The sinusoids in the time domain are described by

$$d(t) = ae^{-bt\pi}\sin(2\pi ft) \qquad for\, t >= 0 \tag{3.6}$$

where a depicts the amplitude of each formant, b the bandwidth and f the frequency of the formant. For this study, b was held constant at 80 Hz (Hillenbrand et al., 2006).

The top three graphs in Figure 3.8 show an example of the time and frequency domain of three sinusoids that were summed to create the final impulse response shown in the two bottom figures. The spectrum of each sinusoid shows a prominent peak at the formant frequency. The final sum of the sinusoids is the finite impulse response of the filter that was used for synthesis.

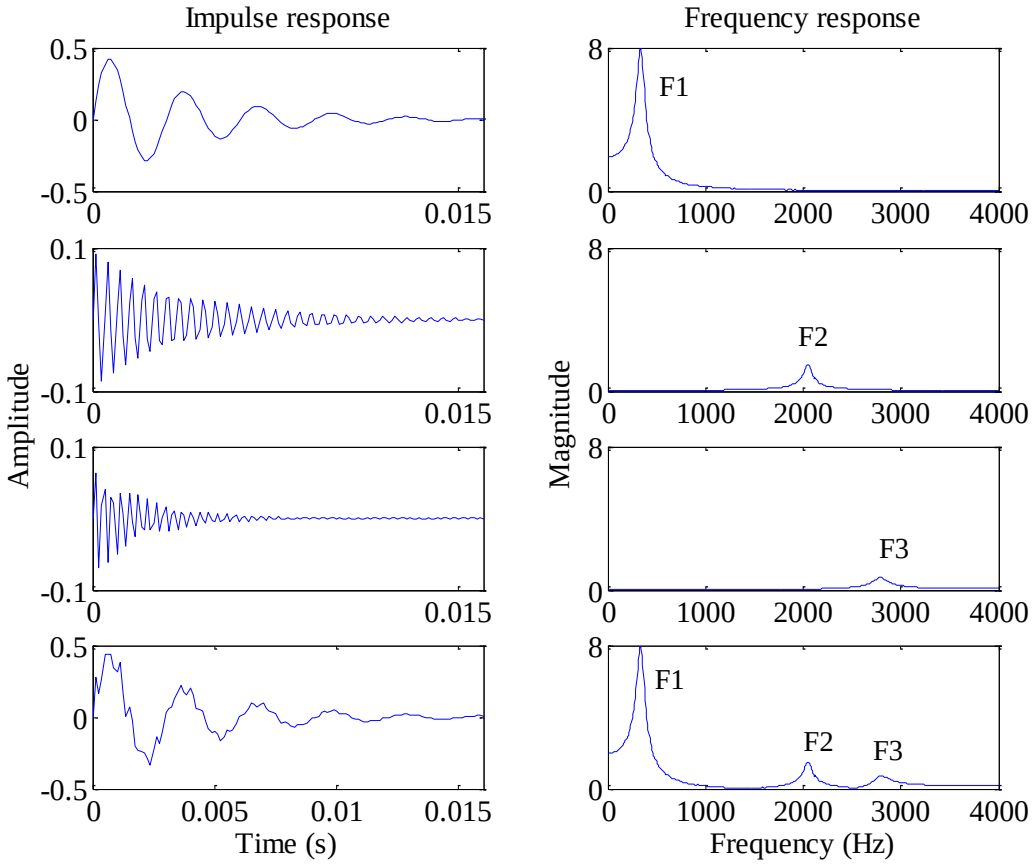

Figure 3.8. Time and frequency domain of the three damped sinusoids (top six graphs) and the sum of the three sinusoids (bottom two) of the vowel /i/.

Figure 3.9 shows the spectrum for the sum of damped sine waves, together with the FFT and LPC spectrum for one time window of the vowel /ɛ/. It can be seen that the DSS describes the filter function only by the first three formants, and that the general spectral envelope according to the FFT is not followed closely.

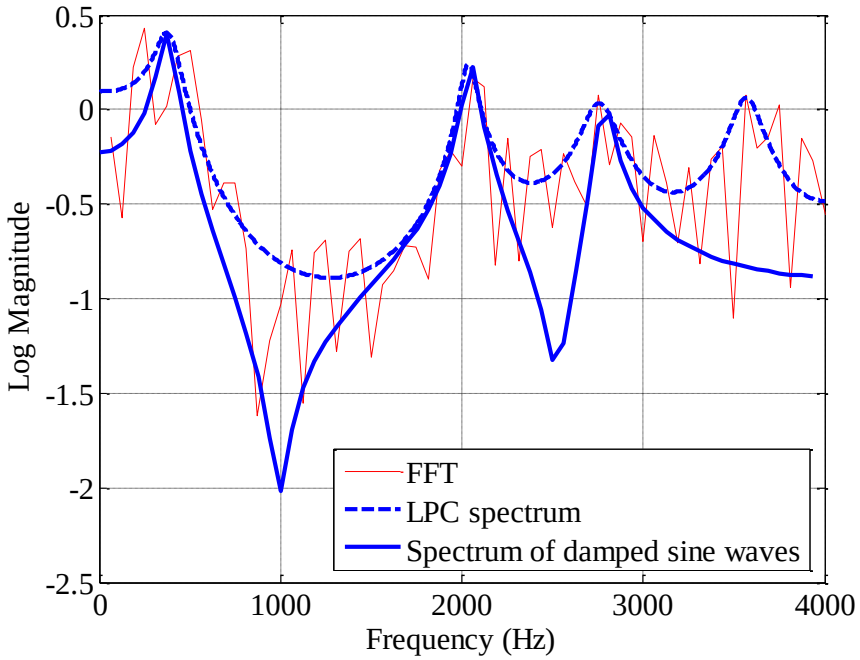

Figure 3.9. FFT and spectrum of the LPC and damped sine waves for the vowel /ɛ/.

3.3.6 Extracting the fundamental frequency

As mentioned previously, synthesis was carried out by filtering a series of pulses with a frequency equal to the instantaneous F0 of the natural spoken vowel. F0 was determined from the vowel part of the original vowel signal by the Simplified Inverse Filter Tracking (SIFT) algorithm (Markel, 1972). Figure 3.10 depicts the sequential signal processing steps required for the SIFT algorithm.

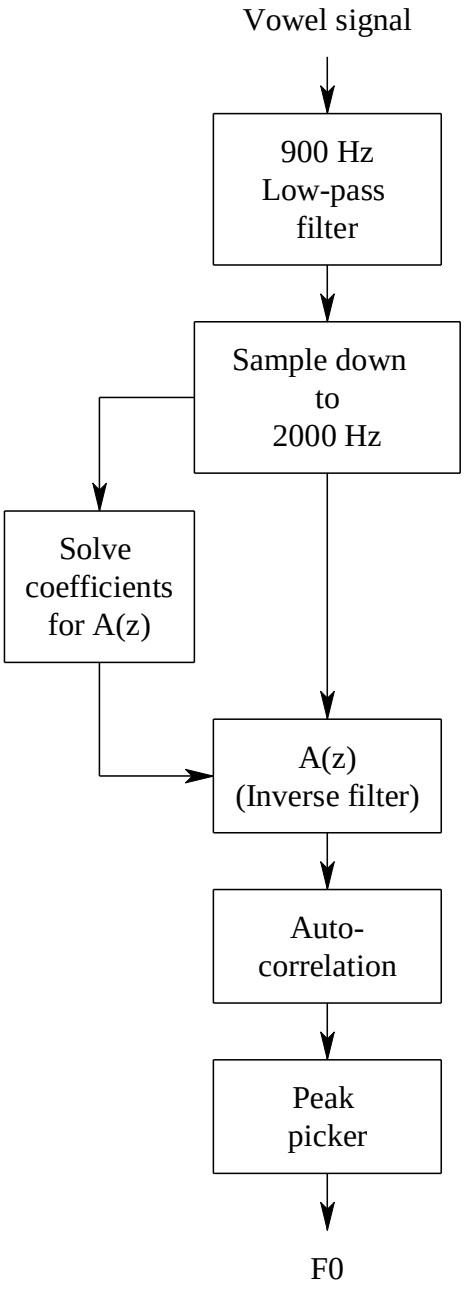

Figure 3.10. Sequential signal processing steps used in the SIFT algorithm (Rabiner and Schafer, 1978).

An anti-aliasing filter was used to band limit the signal, after which it was sampled down from 8000 Hz to 2000 Hz. This was done to reduce the total number of operations required. The fourth order inverse LPC filter of $A(z)$ was implemented to whiten the

spectrum of the input signal and to eliminate the spectral shape of the vowel. A fourth order filter was sufficient in the range of 0 - 1000 Hz, since either one or two formants was expected in this range (Rabiner and Schafer, 1978).

The final pitch period was obtained by the autocorrelation of the inverse filtered signal. The largest peak in the autocorrelation function was chosen as the fundamental period. These calculations were done for window lengths of 20 ms with an overlap of 10 ms (Markel, 1972; Zahorian and Jagharghi, 1993).

Figure 3.11 to Figure 3.15 depict the signal waveforms at different points in the processing of the SIFT algorithm for the vowel /ʊ/. A 40 ms segment from the middle of the vowel was used in this example. Figure 3.11 shows the input waveform down sampled to 2000 Hz. In Figure 3.12 the frequency response of the input signal (Figure 3.11) is shown, together with the fourth order LPC filter response that was calculated for this signal. From Figure 3.12 a single formant at 300 Hz can be seen. The inverse of the smoothed LPC spectrum was used to filter the input signal to obtain a relatively flat spectrum, as is depicted in Figure 3.13. The time domain of the filtered signal is shown in Figure 3.14. The inverse LPC filter removes the formant structure and leaves sharp pulses at each pitch period in the time domain. In Figure 3.15 the pitch period is visible in the autocorrelation waveform as the largest peak.

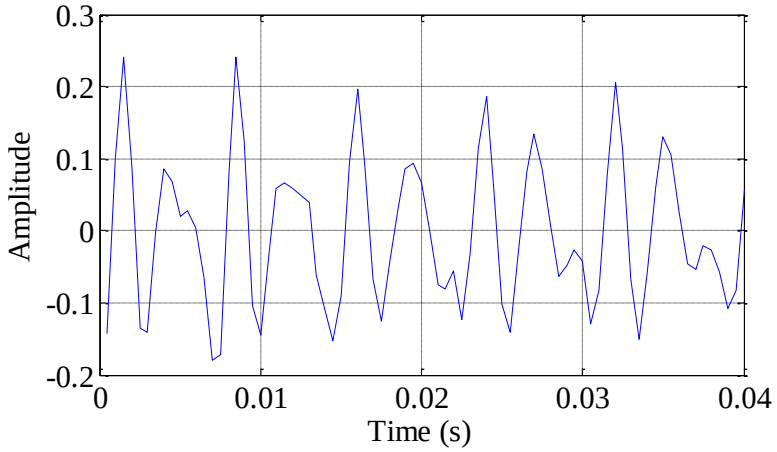

Figure 3.11. Input signal for a 40ms window of the vowel /ʊ/.

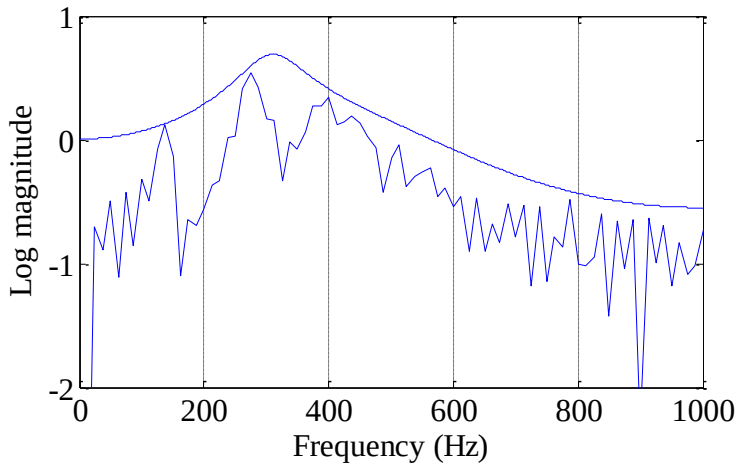

Figure 3.12. FFT and smoothed LPC spectrum for the input signal.

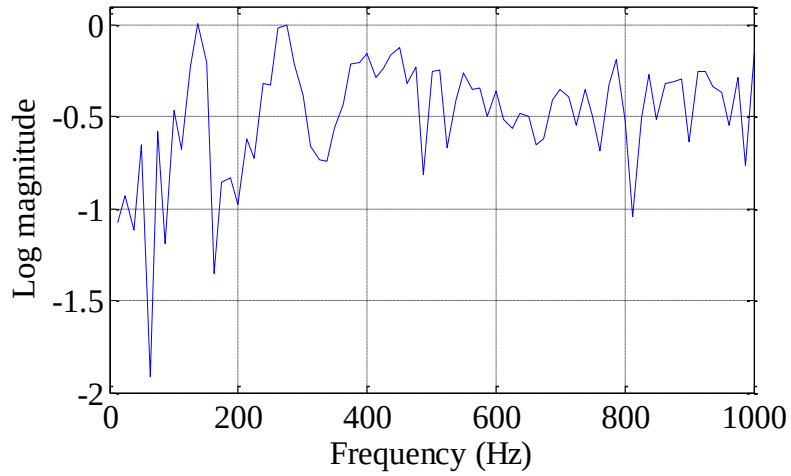

Figure 3.13. FFT of the input signal after inverse LPC filtering.

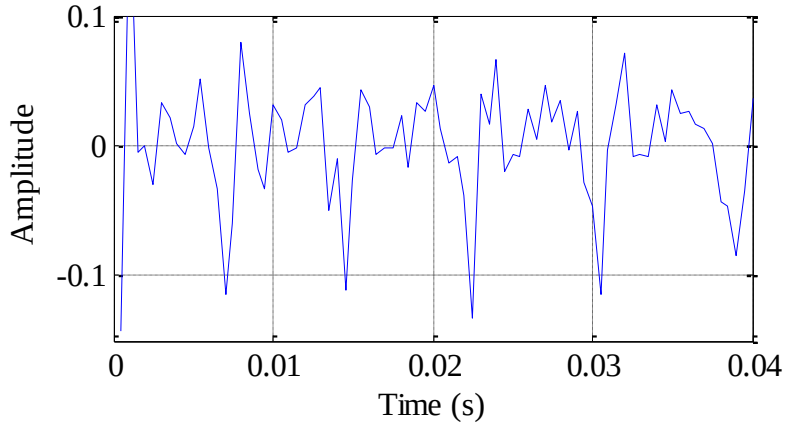

Figure 3.14. Time domain waveform after inverse LPC filtering.

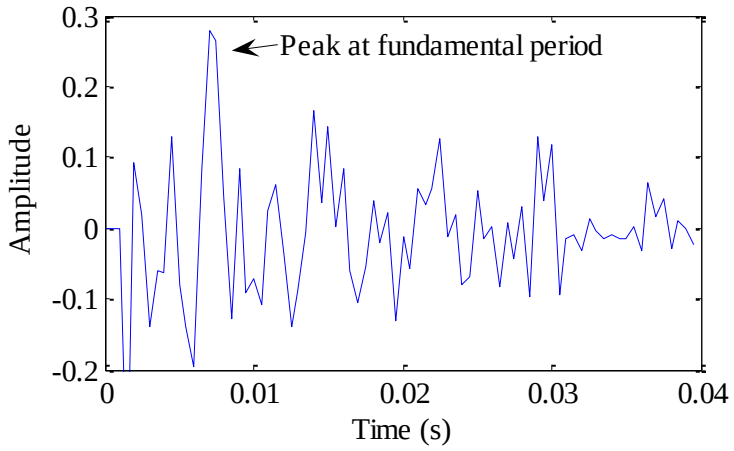

Figure 3.15. Autocorrelation function. Notice the large peak at the fundamental period.

3.3.7 Source model for vowel synthesis

The source-filter model for speech synthesis was described in Figure 3.2. A source signal, mimicking the signal produced by the vocal cords, is filtered by a FIR filter to produce the synthesized speech sound. A simple and efficient way to represent the periodic (voiced) excitation of the vocal cords is by means of an impulse train, with a period equal to the instantaneous fundamental period of the signal being synthesized. For unvoiced sounds, the impulse train source is replaced by white noise (Hillenbrand et al., 2006; Paul, 1981).

Klatt (1987) developed a voicing source somewhat different to the simplified source of an impulse train, where the impulse is substituted by a pulse described by a third order equation. This source, which is used in the KLSYN88 formant synthesizer, leads to more natural synthesized speech sounds and was therefore used in this study. Figure 3.16 describes the voicing source used in the KLSYN88 formant synthesizer.

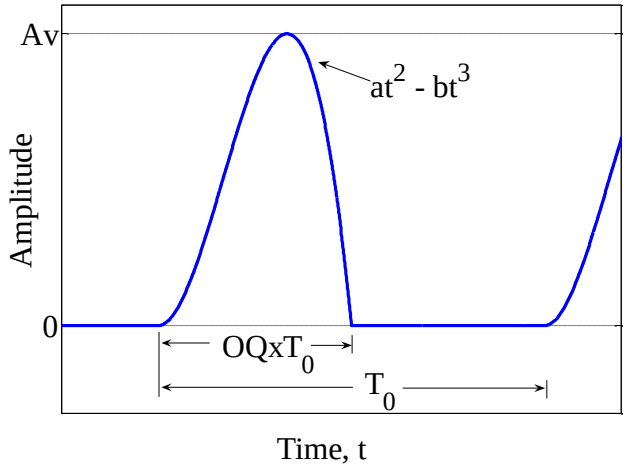

Figure 3.16. Voicing source suggested by Klatt (1987).

The time domain waveform for the open phase of the glottal cycle is described by

$$U_g(t) = at^2 - bt^3. \tag{3.7}$$

The variables *a* and *b* are dependent on the amplitude of voicing and the duration of the open period, as is described in the equations

$$a = \frac{6.75Av}{R^2} \tag{3.8}$$

and

$$b = \frac{6.75Av}{R^3} \tag{3.9}$$

with

$$R = OQ \times T_o \tag{3.10}$$

and *Av* the amplitude of voicing. *OQ* is the open quotient of the glottal waveform, describing the part of the fundamental period T_o for which the source is non-zero.

The top graph of Figure 3.17 shows the source waveform that was used to synthesize the vowel /a/, while the fundamental frequency, plotted as a function of time, is depicted in the bottom figure. The period of the source waveform is equal to the instantaneous fundamental period of the original vowel obtained from the SIFT algorithm.

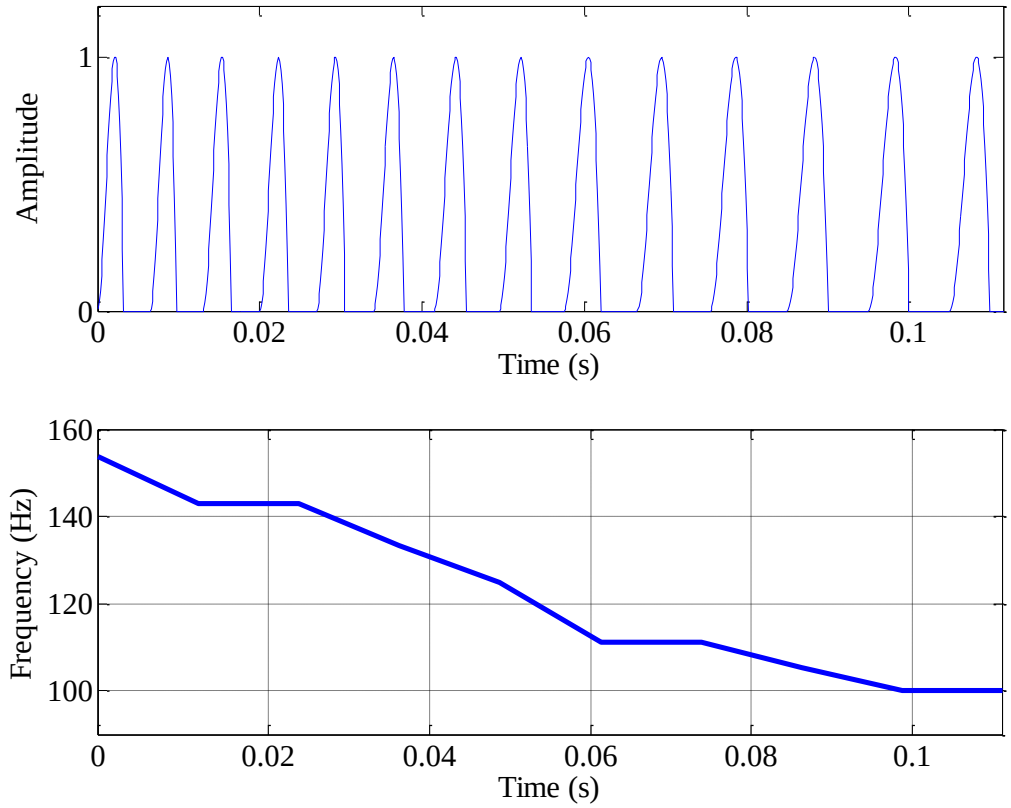

Figure 3.17. Source signal (top) that was used to synthesize the vowel /a/. The varying fundamental frequency is shown on the bottom graph.

Prior to synthesis, the source signal was filtered with a high-pass filter reflecting the radiation characteristics of the lips, resulting in a relatively flat spectrum (Klatt and Klatt, 1990). Similar to the original vowel signal, the source waveform was divided into 32 ms windows with 16 ms overlap. Each window was multiplied by a Hamming window to minimize spectral leakage. Final synthesis was done by filtering each window of the source

signal with either the formants-only or whole-spectrum filter corresponding to the same window of the original vowel signal.

The original /p/ and /t/ consonants of the recorded vowels were not concatenated to the synthesized vowels, as the consonants contain transitional formant cues that would aid listeners in identifying the vowels (Kirk et al., 1992; Strange et al., 1983). To ensure that no transitional cues were transmitted by the consonants, neutral consonants containing no vowel-specific cues were synthesized and added to the synthesized vowel signals.

The /p/ consonant was created by using the method described in Hillenbrand et al. (2006) to synthesize unvoiced speech. A source signal was generated, consisting of a sequence of pulses with a probability of 0.5 to have an amplitude of zero or non-zero at each sampling point. This source signal, that was spectrally indistinguishable from white Gaussian noise, was filtered by the average spectral envelope of the /p/ consonants of the original vowels. The /t/ consonants were created by averaging the time-domain waveforms of all the /t/ consonants, thereby eliminating any vowel cues contained in the consonants.

3.3.8 Suppression of F1 and F2

To assess the relative importance of individual formant peaks for vowel recognition in noise, F1 and F2 were suppressed alternatively for each synthetic vowel signal. To suppress these formants, the damped sine wave impulse response of the formants-only vowels was generated either without F1 or F2 frequency inputs, while the whole-spectrum filter transform function was manually manipulated so that a linear amplitude transition from the start to the end of the suppressed region existed.

Figure 3.18 shows the filter frequency responses for the whole-spectrum vowels (left column) and formants-only vowels (right column) for the vowel /a/ (male speaker). The first row shows the response for the complete spectrum vowels (containing both F1 and F2), while the second and third row depict the suppression of F1 and F2 respectively. It can

be seen that the DCT spectrum follows the FFT while the DSS spectrum only match the FFT of the original vowel in terms of formant amplitudes and frequencies.

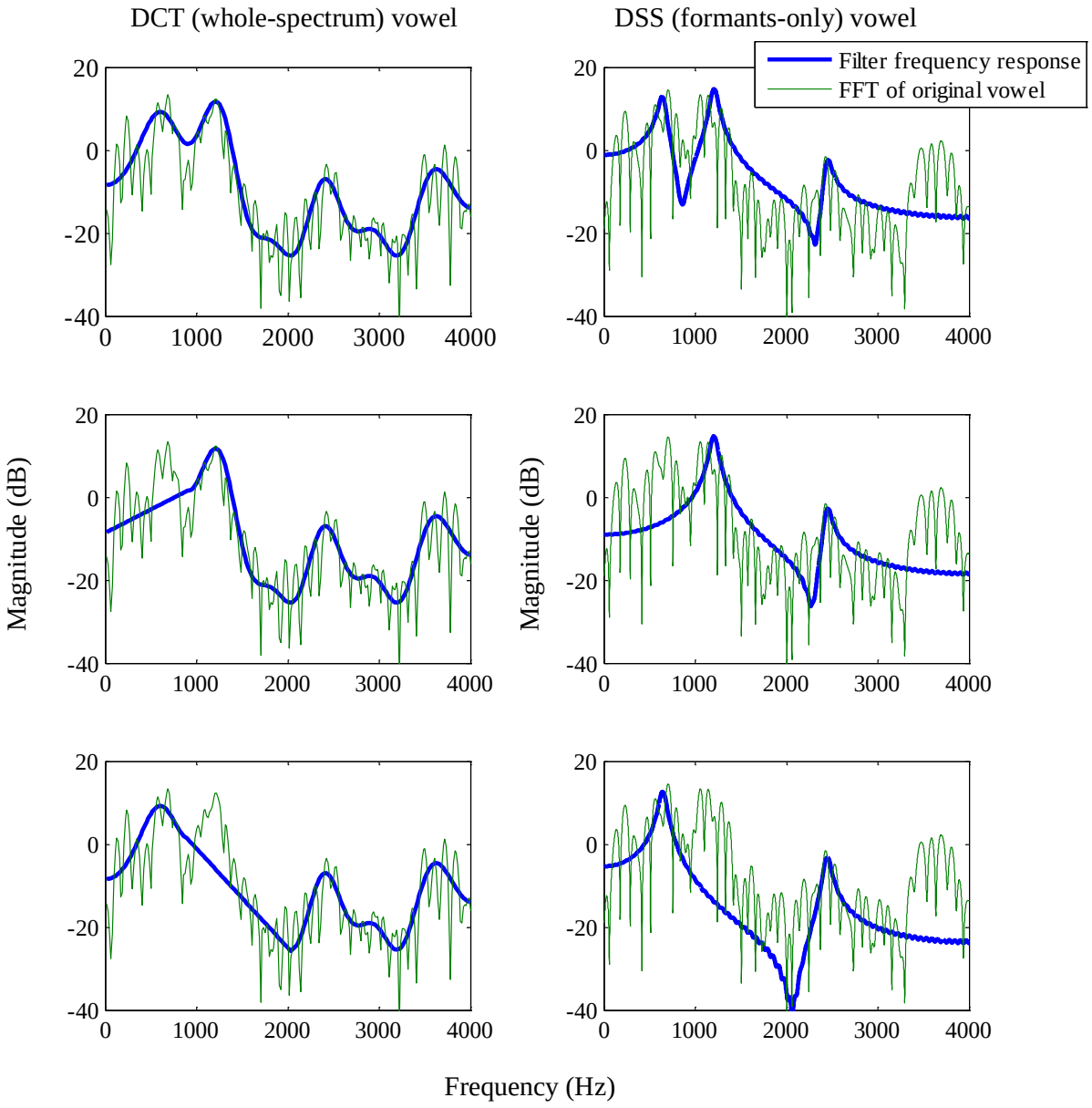

Figure 3.18. Filter frequency responses for the whole-spectrum vowels (left column) and formants-only vowels (right column) of the vowel /ɑ/ (male speaker). The first row depicts the response for no formant suppression, while the second and third show the suppression of F1 and F2 respectively.

3.3.9 Addition of speech-shaped noise

To determine the cues that listeners depend on in noise, it was necessary to present the synthesized signals to the listeners in noisy conditions. According to Phatak and Allen (2007) the power spectrum of average speech has a roll-off of about –29 dB/decade above 500 Hz. If a speech signal is masked by white noise, the SNR will not be the same at all frequencies as a result of this roll-off. To achieve an equal SNR throughout the frequency band, speech-shaped noise should be added to each vowel (Hedrick and Nabelek, 2004; Phatak and Allen, 2007; Xu and Zheng, 2007). Speech-shaped noise is a noise signal with a spectrum similar to the average spectrum of speech.

Adding only a single speech-shaped noise signal to all the stimuli was considered. It would, however have introduced a problem when comparing vowel scores across the vowel groups of the complete spectrum, F1- and F2-supressed vowels. Masking the formant-suppressed vowels with identical spectral noise shapes than the complete-spectrum vowels would have created regions at the suppressed formant areas where the SNR is considerably lower than the other spectral regions.

Another approach was considered where noise containing a spectrum analogues to each individual vowel would be added, ensuring a constant SNR at all frequencies for every vowel. The problem with this approach was that the spectral definition of each vowel would have been (to some extent) embedded into the filtered white noise signal, leading to possible vowel cues being transferred to the listener through the noise signal.

It was therefore decided that three types of noise signals would be added to the vowels, with each type depicting the average vowel spectra of the three spectrally manipulated vowel groups. The first noise signal was generated for the vowels containing both the first and second formants, while the second and third noise signals were generated for the F1- and F2-suppressed vowels respectively. A FIR filter, which was derived from the average LPC spectrum related to the specific group of vowels, was used to filter white noise, resulting in speech-shaped noise. This method ensured that (i) the SNR was almost

constant throughout the vowel spectra for all vowels; (ii) no vowel-specific spectral characteristics were contained in the added noise signal. Figure 3.19 depicts the spectra of the three types of noise signals.

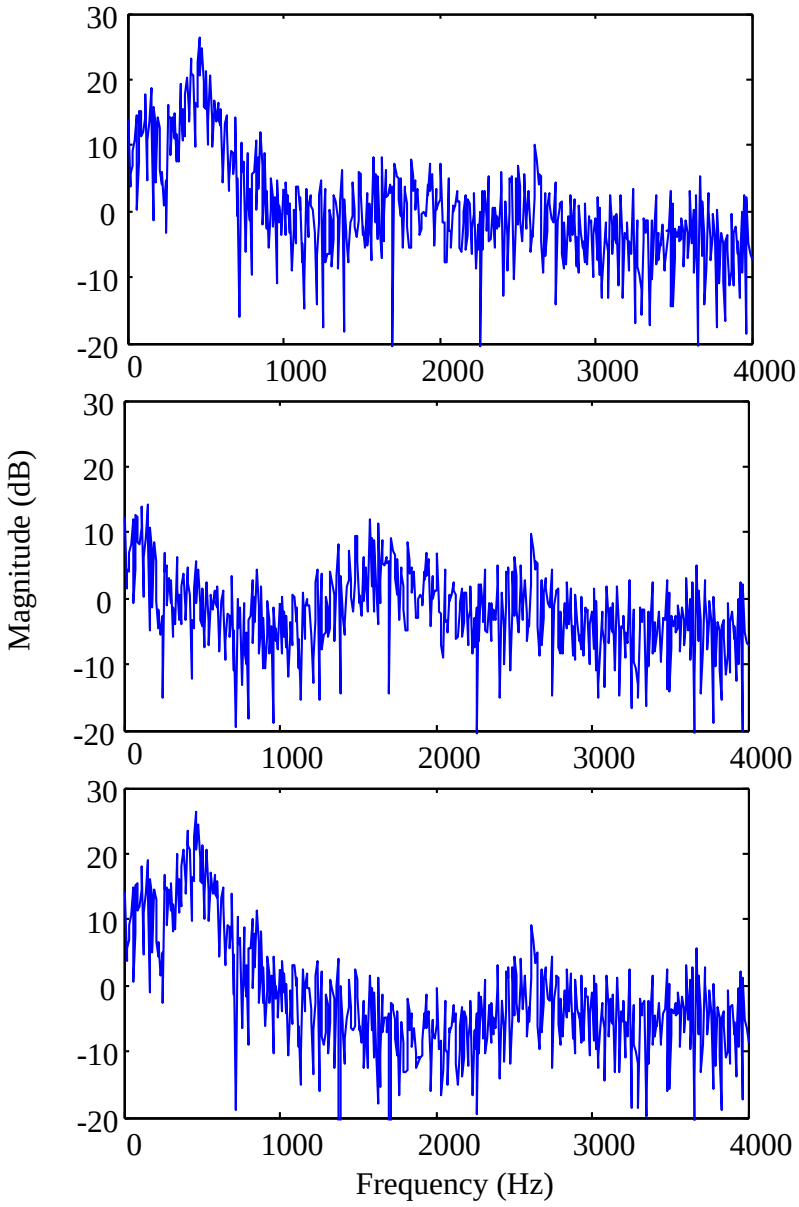

Figure 3.19. Spectra of the three types of noise signals used to mask the vowel stimuli. From the top down, the figures depict the noise signals for the complete spectrum, F1-suppressed and F2-suppressed vowels.

Noise was added to each vowel at SNRs of -10 dB, -5 dB and 0 dB. An algorithm was developed to generate the noise-corrupted vowels. As input, the algorithm accepted the SNR value and the clean synthetic vowels to produce the noisy vowels as output. The desired SNR (in dB) was first converted to a linear intensity value *g* by the equation

$$g = 10^{SNR/20}. \tag{3.11}$$

A white noise signal was created by a normally distributed random number generator with zero mean and a standard deviation of one. For both the speech and the white noise, the rms level was normalized to one by dividing each signal by its own rms value. The final noise corrupted signal *S* was obtained by the equation

$$S = s + g \times n \tag{3.12}$$

where *s* and *n* were the scaled vowel and noise signals respectively (Regnier and Allen, 2008). The final noisy vowel signal was scaled to an intensity level of 70 dB SPL (estimated intensity in Praat) to ensure that all the tokens were presented at equal intensity to the listeners. Each vowel was centered in an equal duration noise burst of 1000 ms.

All signal processing was implemented in Matlab version 7.0.1 on a computer with an Intel Pentium (R) D 2.8 GHz processor and 1 GB RAM.

3.4 SUMMARY

In this Chapter, the whole-spectrum and formants-only methods of vowel synthesis were described. Analysis of the original recorded vowels for formant and whole-spectrum cues was explained, as well as the integration of these cues into the vowel synthesis process. Lastly, the addition of speech-shaped noise to the synthesized vowel tokens in different SNRs was described.

CHAPTER 4 METHODS 2: EXPERIMENTAL WORK

4.1 CHAPTER OBJECTIVES

In this chapter, the methods that were used to conduct experimental listening tests are described, as well as techniques implemented to analyze the listening test results.

4.2 EXPERIMENTAL STUDY

4.2.1 Subjects

A group of 12 listeners, six males and six females, took part in the experimental study. Listeners were native Afrikaans-speaking, had normal hearing determined through screening by an audiologist, and were between the ages of 20 and 28.

4.2.2 Stimuli

Figure 4.1. shows a layout of the stimuli used in the experiment. Each synthetic vowel was divided into three groups. The first group consisted of vowels with no suppressed formants, while the second and third group represented the F1-suppressed and F2-suppressed vowels respectively. Vowels were presented in four listening conditions namely -10 dB, -5 dB, 0 dB SNR and in quiet. Each vowel was repeated 10 times (five male voice and five female voice), resulting in a total number of 2880 vowels presented to each listener.

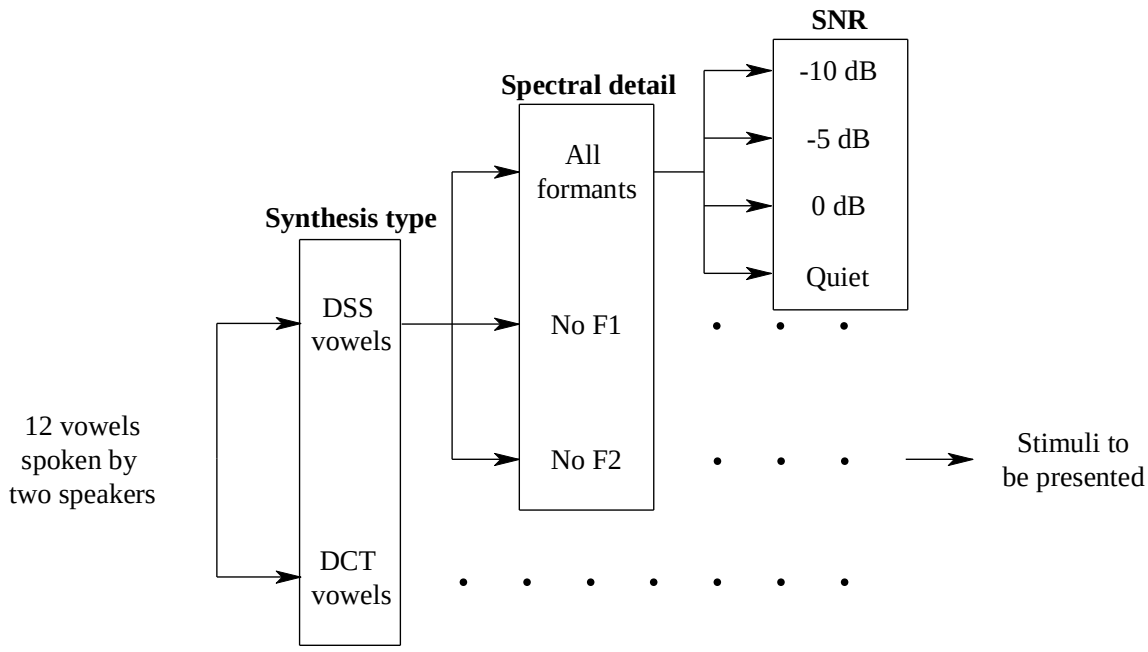

Figure 4.1. Layout of the different stimuli types presented to listeners.

A sound proof booth was used as location for the tests to minimize external noise or interference. A computer equipped with an M-Audio Fast Track Pro external soundcard was used in the experiments. The speech samples were presented through an M-Audio EX66 loudspeaker having a frequency response of 37 Hz to 22 kHz with a passband flatness of ±1 dB.

A software application was used to track the response of the listeners. Vowel sounds were presented in random order, and the listener had to select the sound description (on a set of buttons) that matched what was heard from the speaker. A representation of the software user interface is shown in Figure 4.2. After each experiment a confusion matrix was generated for each stimulus type depicting the vowel that was presented, as well as the

response by the listener.

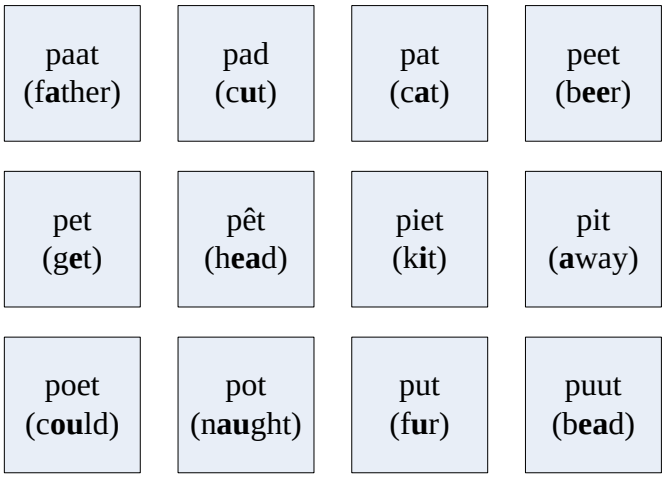

Figure 4.2. Representation of the user interface for the software used in the listening tests. Each button depicted the vowels in /p/-vowel-/t/ context, together with an Afrikaans word containing the vowel. In the figure, the vowels are described by English words for clarity to the reader.

Stimuli were presented in random order (Hedrick and Nabelek, 2004; Parikh and Loizou, 2005) to eliminate any predictability. Prior to the main test, each listener completed a practice round where the vowels containing all formants were presented in quiet. The objective of the practice round was to familiarize listeners with each vowel sound and its description that appeared on the computer the screen. Each test took approximately 2 hours to complete for each listener.

4.3 ANALYSIS OF THE DATA

Results from the listening experiments were analyzed to gain an understanding of important cues used by listeners to perceive vowels in the different listening conditions. Two of the analysis methods are described in the next two sections.

4.3.1 FITA analysis

FITA analysis was done to determine the cues that were most effectively transmitted to listeners in severe noise. A measure of the transmission of perceptual cues is defined as the covariance between the input and output (stimuli and response) of a listening test. It is described as the mean logarithmic probability (MLP) by

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

where each discrete value of i = 1,2,..,k for the input variable x, has a probability of p_i. A similar expression holds for the output y with values of i = 1,2,..,m. The covariance between the input and output is defined as

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

where MLP(xy) defines the measure of the joint stimulus-response pair and p_{ij} the probability of the combined occurrence of the input i and the output j. The relative transmission is defined by

$$T_{rel}(x;y) = \frac{T(x;y)}{H(x)} \tag{4.3}$$

with $H(x)$ the maximum available information (Miller and Nicely, 1955). The value of $T_{rel}(x;y)$ ranges between zero and one. If a specific cue was not transferred to (or perceived by) a listener, the relative transmission will be zero, while the value will be unity if the specific cue was successfully received.

The percentage information transmission of the cues F1, F2 and duration was tested with FITA analysis. Three groups for each cue were created, and each vowel was divided into one group for each cue according to its F1, F2 and duration value. Division of cues into groups allows one to determine a measure by which the listeners perceived vowels containing cues in the same group as the stimuli vowels. The group division of the vowels is shown in Table 4.1. Each cue for each vowel is classified as a number between one and three according to its cue value.

Table 4.1. Classification of cues into groups for FITA analysis

Group	**Duration (ms)**		**F1 (Hz)**		**F2 (Hz)**	
	Male voice	Female voice	Male voice	Female voice	Male voice	Female voice
1	75 - 145	100 - 200	270 - 390	270 - 470	1000 - 1350	1000 - 1500
2	145 - 215	200 - 270	390 - 500	470 - 670	1350 - 1700	1500 - 2050
3	> 215	> 270	> 500	> 670	> 1700	> 2050

Table 4.2 shows the division of each vowel into the groups defined in Table 4.1 for each speaker.

Table 4.2. Allocation of cue category for each vowel and speaker

		Male voice			Female voice		
		Duration	F1	F2	Duration	F1	F2
paat	/ɑː/	3	3	1	3	3	1
pad	/a/	1	3	1	1	3	1
pat	/æ/	1	3	2	1	3	2
peet	/e/	2	1	3	2	1	3
pet	/ɛ/	1	2	3	1	1	3
pêt	/ɛː/	3	2	3	3	1	3
piet	/i/	1	1	3	1	1	3
pit	/ə/	1	2	2	1	2	2
poet	/u/	1	1	2	1	1	1
pot	/ɔ/	1	2	1	1	2	1
put	/œ/	1	2	2	1	2	2
puut	/yː/	3	1	3	3	1	3

4.3.2 Multidimensional scaling

Multidimensional scaling (MDS) provides a means of reducing a large set of data into an uncomplicated spatial representation, consisting of only a few dimensions (Mugavin, 2008). It is a method by which the underlying cues that influence a subject's perception can be found from confusion matrix data. Input to an MDS analysis consists of the measure of similarity or dissimilarity between stimuli, whilst the output portrays a spatial representation where the stimuli are positioned in a manner where the distance between two stimuli corresponds to the similarity of the two. A large distance depicts a high degree of dissimilarity, while stimuli close to each other can be interpreted as perceptually similar.

One can distinguish between metric and non-metric MDS implementations. When performing metric MDS, it is assumed that the dissimilarity matrix used as input to the analysis consists of metric properties or measurable distances between elements. The MDS spatial representation of the data therefore represents the ratios and intervals between dissimilarities very accurately. In non-metric MDS, it is assumed that the ratios and intervals between dissimilarities are not of importance, while only the order of the dissimilarities are preserved in the multidimensional space (Wickelmaier, 2003). The plotted MDS coordinates are generally not suitable for direct interpretation, as the MDS coordinates only represents the projection of each object on the different axes. When the configuration is rotated, projections will change according to the degree of rotation, but the distances between objects will stay exactly the same. For this reason it is often allowable for the MDS results to be rotated to find a match between the MDS data and features that possibly influenced the input data (Kruskal and Wish, 1978).

Individual difference scaling (INDSCAL) is an MDS technique which provides the option of using more than one dissimilarity matrix as input. It uses the assumption that different subjects use the same dimensions when interpreting stimuli, but that different weights are applied to each dimension. INDSCAL can be classified as a metric analysis with the results being unique, implicating that the resulting spatial representation need not be rotated to find an optimal fit to features that influenced the input data. Results can therefore be directly interpreted (Carroll and Chang, 1970).

For the reasons stated above, vowel confusion matrices determined in the vowel experiments (see section 4.1) were subjected to INDSCAL analysis. The aim was to determine if vowel recognition and confusion patterns can be explained by a spatial vowel cue space. Figure 4.3 provides an overview of the procedure followed in analyzing the confusion matrices with INDSCAL.

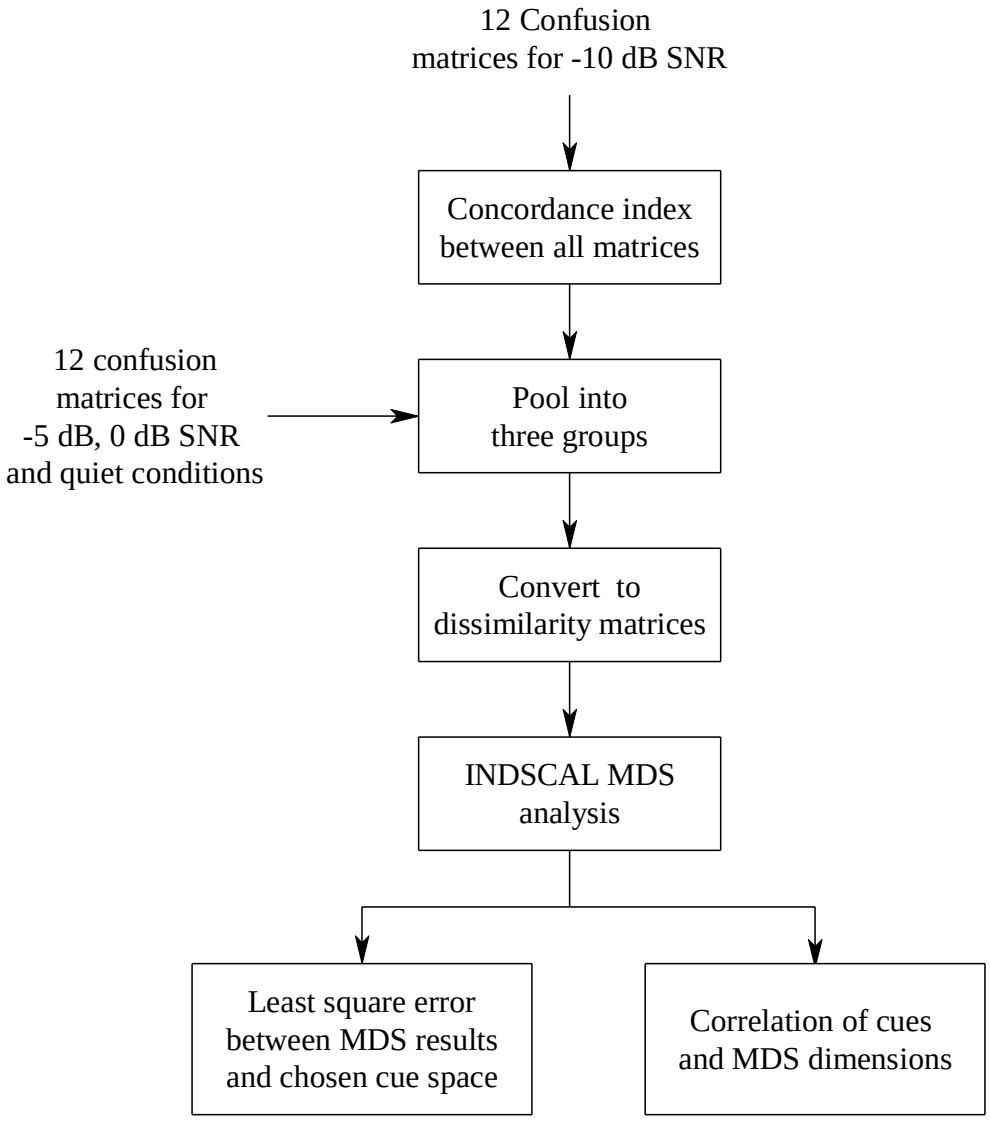

Figure 4.3. Procedure followed in the INDSCAL MDS analysis of confusion matrices. The confusion matrices for the -10 dB SNR condition are pooled, while this pooling was also used for the other conditions.

4.3.2.1 Pooling into groups using the concordance index

It was necessary to group confusion matrices of listeners together due to the differences in vowel confusion patterns found between subjects in severe noise. The 12 confusion matrices representing results at -10 dB SNR for all vowel groups were pooled into three groups for input to the INDSCAL analysis (Van Wieringen and Wouters, 1999). This grouping was also used for all the other conditions.

Prior to pooling, a measure of correlation between the 12 confusion matrices was determined. This was done by calculating the within-stimulus concordance index between the individual confusion matrices (Brusco, 2004). Knowing the similarity between the vowel confusion results of each listener allowed for pooling of subjects in groups where similar vowel perception results were obtained, which also ensured that important information was not lost when each pool was averaged. Between any pair of confusion matrices, the relationship

$$b_{hijq} = \begin{Bmatrix} sign(a_{hjq} - a_{hiq}) & for\, distinct\, h,i,j \\ 0 & otherwise \end{Bmatrix} \tag{4.4}$$

was calculated where

$$sign(x) = \begin{Bmatrix} +1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{Bmatrix}. \tag{4.5}$$

The value of b_{hijq} is equal to one if the value of column *h*, row *j* for confusion matrix *q* is bigger than the value in column *h*, row *i* for confusion matrix *q*. If these two values are the same, b_{hijq} is equal to zero, while it equals -1 if the value is smaller. The concordance index Ω_{qr} between matrix *q* and *r* was finally calculated as

$$\Omega_{qr} = \frac{\sum_{h,i,j} b_{hijq} b_{hijr}}{\sum_{h,i,j} \left| b_{hijq} b_{hijr} \right|} \qquad for\, q = 1,\ldots,Q-1, r = q+1,\ldots,Q. \tag{4.6}$$

The concordance index between all 12 matrices for the -10 dB SNR conditions was calculated for the male and female voice for the whole-spectrum and formants-only vowels. This provided four matrices depicting the concordance index between all the listeners for the four conditions. Each of these matrices were processed through an initial MDS analysis to obtain a measure of the distance between listeners' results. The MDS results for each listening condition were plotted in an appropriate three-dimensional space, after which subjects were pooled by visual inspection of the MDS space. The entire space was divided into three distinct groups of subjects according to the proximity amongst the plotted subjects. The same groups obtained from the -10 dB SNR conditions were used in the FITA and MDS analysis for the other SNR and quiet conditions.

4.3.2.2 Conversion to dissimilarity matrices

Prior to INDSCAL analysis, each pooled confusion matrix was first converted to a symmetrical dissimilarity matrix. To convert the asymmetrical confusion matrix to a symmetrical one the equation

$$s(i,j)=s(j,i)=\frac{1}{2}\sum_{k=1}^{12}\left[c(i,k)+c(j,k)-\left|c(i,k)-c(j,k)\right|\right] \tag{4.7}$$

was used where element $s(i,j)$ in row *i* and column *j* depicts the entry in the new matrix, with *c* representing confusion elements in the original confusion matrix. From the above equation, the term

$$\sum_{k=1}^{12}\left|c(i,k)-c(j,k)\right| \tag{4.8}$$

was used to obtain the dissimilarity matrix (Klein, Plomp and Pols, 1970).

4.3.2.3 INDSCAL analysis and interpretation

After subject pooling and conversion to dissimilarity matrices, INDSCAL analysis was performed for all listening conditions. The input to the analysis was the dissimilarity matrices of the three groups for each speaker (male and female) and synthesis type (formants-only or whole-spectrum) at each SNR condition. INDSCAL analysis was performed using NewMDSX (Coxon, Brier and Hawkins, 2005). The INDSCAL results for each condition were evaluated in terms of the squared correlation index (r^2) which gave the proportion of variance of the input data that was accounted for by the results. r^2, which is also defined as the "variance accounted for" (VAF) value, increases with an increase in dimensions. An acceptable fit to the input data is obtained for $r^2 \geq 0.6$ (Jaworska and Chupetlovska-Anastasova, 2009).

INDSCAL analysis was used to determine the underlying cues that listeners use in making vowel judgments for the two synthesized vowels (whole-spectrum and formants-only). To determine the degree to which listeners utilized formants or spectral shape cues in noise, vowels were located in two different vowel spaces that comprised of either the F1, F2 and duration cues, or five spectral band cues.

Formants and duration (hereafter referred to as the formants space) were determined as was described in section 3.3.5.1, while whole-spectrum information was extracted by filtering the spectrum of the synthesized whole-spectrum vowels into five spectral bands. In Klein et al. (1970), vowels were band-pass filtered according to a number spectral bands in the spectrum. For each vowel, the output of each band-pass filter was summed energetically and presented to listeners. The outcome revealed that when vowels were represented by five frequency bands spaced between 500 Hz and 4000 Hz and located two-thirds of an octave from each other, an identification score of 94% was still obtained. Similar spectral bands were therefore used in the present study to locate each vowel in a five-dimensional spectral-band space. The dB Sound Pressure Level (SPL) value for each frequency band (with centre frequencies at 445 Hz, 680 Hz, 1120 Hz, 1780 Hz and 2800 Hz) was computed by band-pass filtering the vowel spectrum at each defined spectral band region. The five dB SPL values for each vowel were used in this study to define a whole-

spectrum vowel space. Plots of the spectral band vowel space are presented in Chapter 5.

INDSCAL analysis was done for both three dimensions (for fit to the formants space) and five dimensions (for fit to the spectral bands space). Two criteria were used to evaluate the MDS results. In the first criterion, a measure of similarity between the vowel placement in the cue vowel space and MDS vowel space was determined. The least squares error between the MDS results and the vowel cues for each individual vowel was calculated and summed for all vowels. The squared error was calculated as

$$S = \sum_{i=1}^{n} d_i(p,q)^2 \tag{4.9}$$

where *S* is the sum of squared distances between each vowel *i*, stipulated in the MDS space and the vowel cue space for *n* vowels. Points *p* and *q* represent the vowel coordinate for the MDS and vowel cue space respectively. The Euclidian distance between the two vowel representations in *m* dimensions for each vowel is defined as

$$d_i(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_m - p_m)^2} \ . \tag{4.10}$$

The squared error between the vowel cue space and the MDS data was calculated for different fits of the cues to the MDS data. Therefore for the formant and duration space, MDS dimensions one, two and three were fitted to the cues F1, F2 and duration, after which all possible combinations of the three MDS dimensions were fitted to the three vowel cues. The combination of MDS and vowel cue pairs that provided the smallest fitting error were logged.

In the second (and alternative) criterion, the MDS dimensions and vowel cue pairs that provided the smallest fitting error in criterion 1 were correlated with each other. This provided a measure of similarity between the specific vowel cues and the MDS data. Pearson's linear correlation was used for calculating the correlation coefficient between the two data sets (Fox, Flege and Munro, 1995). Results for these analyses are given in Chapter 5.

CHAPTER 5 RESULTS

5.1 CHAPTER OBJECTIVES

In this Chapter, results of the vowel synthesis methods and listening tests are presented. Acoustic analysis is used to determine whether the intended cues are indeed present in the synthetic vowels. The original vowels are compared to the synthetic ones to evaluate the accuracy of the synthesis techniques. Predictions regarding the experimental outcomes are made by locating the stimuli in two different vowel spaces. Finally, listening test results are presented and analyzed.

5.2 RESULTS OF VOWEL SYNTHESIS

In the Chapter 4 the vowel synthesis techniques were described. Two types of vowels were synthesized, one depicting the whole-spectrum, and one containing only formant frequency and amplitude information. For each vowel token, F1 and F2 were also suppressed to investigate the single effect of these cues on recognition.

To investigate whether the two synthesis techniques used in this study represent the whole-spectrum and the formants accurately, 22-pole LPC spectra were taken of the synthesized vowels. Figure 5.1 and Figure 5.2 show the FFT of the original recorded vowels /ɛː/ and /æ/ overlaid on the 22-pole LPC smoothed spectrum of the synthesized whole-spectrum and formants-only vowels respectively. In both figures it can be seen that the spectrum of the whole-spectrum vowel closely follows the original vowel spectrum, while the formants-only vowel spectrum only roughly estimates the spectrum in terms of the first three formant frequencies and amplitudes. If listeners should rely on the detailed spectrum between two formants, the whole-spectrum vowels will provide adequate information, while listeners would find the formants-only vowels hard to recognize.

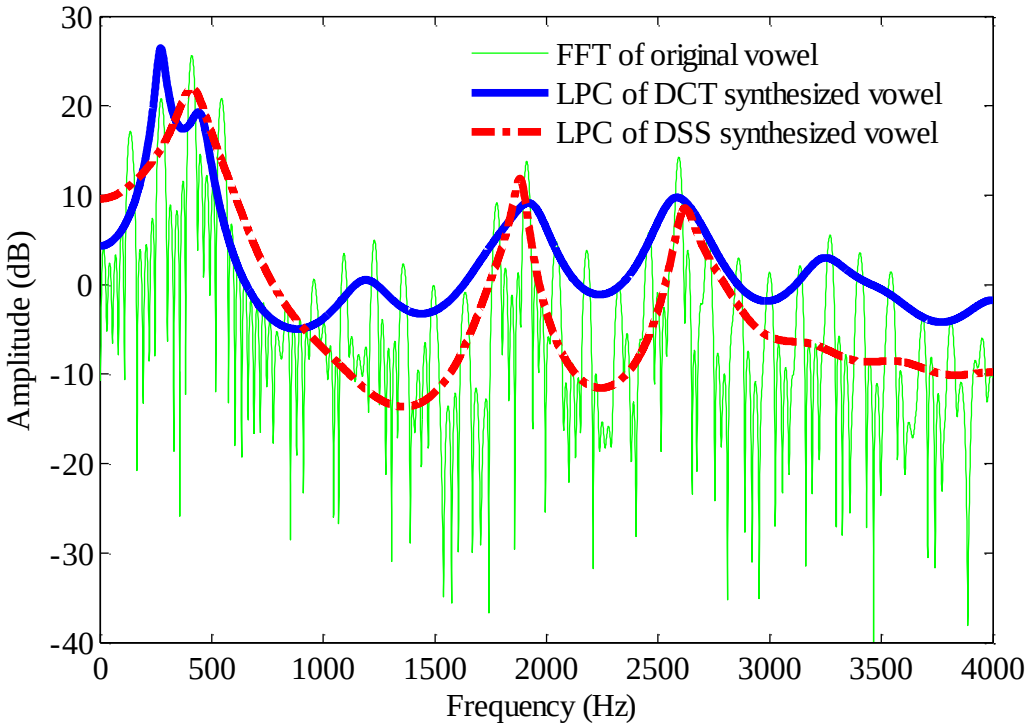

Figure 5.1. 22-pole LPC smoothed spectra of the whole-spectrum and formants-only synthesized vowel /ɛː/ (male speaker) overlaid on the FFT of the original recorded vowel.

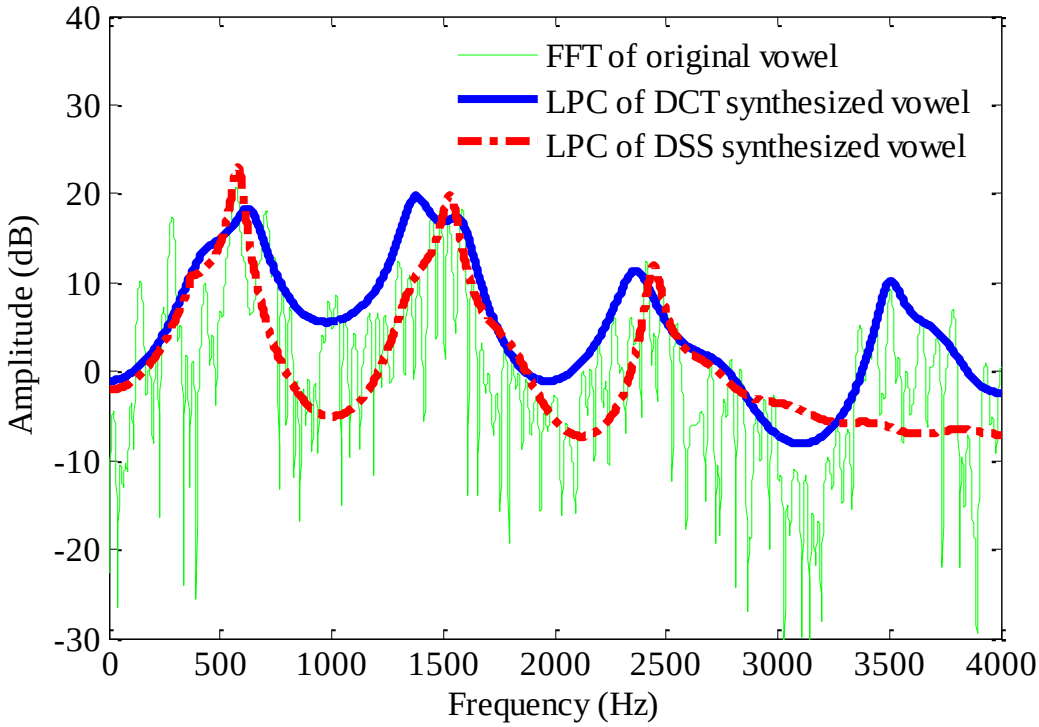

Figure 5.2. 22-pole LPC smoothed spectra of the whole-spectrum and formants-only synthesized vowel /æ/ (male speaker) overlaid on the FFT of the original recorded vowel.

For all the vowels synthesized, either F1 or F2 was suppressed to examine the effect of these cues on vowel perception in noise. To investigate the absence of these formants on the final synthesized vowels, an 11-pole LPC smoothed spectrum was taken for the complete spectrum, F1-suppressed and F2-suppressed vowels and overlaid on each other. Figure 5.3 shows these spectra for the vowel /ɑ/ (male speaker) for both the whole-spectrum and formants-only synthesized vowels. Figure 5.4 shows the same information for the vowel /ɔ/ (female speaker). From the figures it can be seen that, for both the F1- and F2-suppressed vowels the whole-spectrum synthesis type, an almost linear connection between the valley to the left and the right of the formant exists. The rest of the non-manipulated spectrum follows the LPC spectrum of the complete spectrum vowel closely. The LPC spectrum for the formants-only synthesis also shows F1 and F2 to be fully suppressed for the respective vowels.

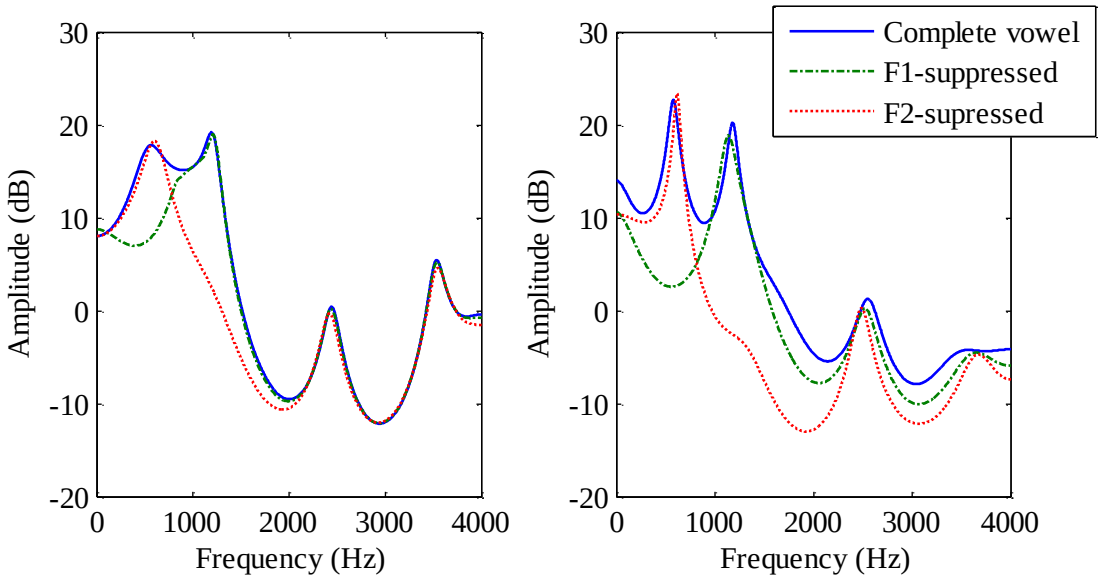

Figure 5.3. LPC spectra for the male voice vowel /ɑ/. The whole-spectrum vowels (left) and formants-only vowels (right) are depicted for the complete spectrum, F1-suppressed and F2-suppressed vowels.

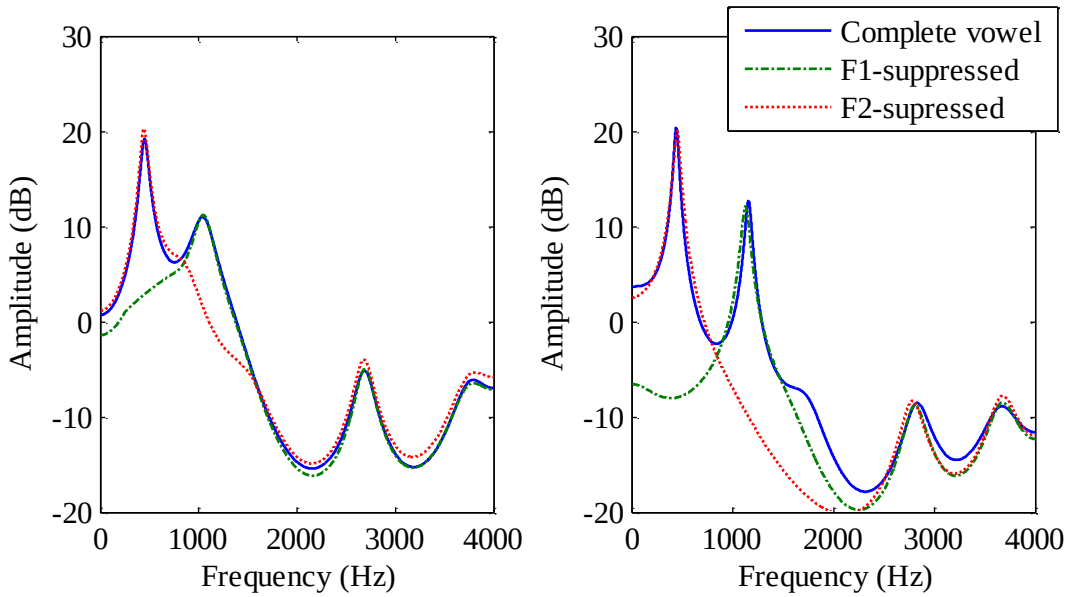

Figure 5.4. LPC spectra for the female voice vowel /ɔ/. The whole-spectrum vowels (left) and formants-only vowels (right) are depicted for the complete spectrum, F1-suppressed and F2-suppressed vowels.

5.3 VOWEL SPACE PARAMETERS

Predicting confusions for vowels in quiet conditions is usually done by locating each vowel in a vowel space. Specific cues that are regarded as essential for vowel recognition are usually defined in a vowel space (Delattre et al., 1952; Hillenbrand et al., 1995; Liu and Fu, 2005). In this study, two different vowel space representations were considered; one is defined by F1, F2 and duration values, and the other is defined by five spectral bands (representing the whole-spectrum) for each vowel. The two vowel spaces will be described and analysed in the following sections.

5.3.1 F1, F2 and duration

The predominant way to define a vowel space is by utilizing F1, F2 and duration characteristics of the vowels as the axes (referred to as the formant space in this study). In certain degraded vowels (similar to vowels processed through a CI acoustic model) formant frequencies may differ from the original vowel. Locating vowels in a new vowel space and comparing locations to those of the original vowels allows one to predict which vowels would be confused with other vowels.

To predict possible confusions among the vowels in this study, the average F1 and F2 frequencies as well as duration were determined for each of the originally recorded vowels (male and female voice) in the manner described in Chapter 3. These values are displayed in Table 5.1

Table 5.1. F1, F2 and duration information for the original recorded vowels

		Male voice			Female voice		
		Duration (ms)	F1 (Hz)	F2 (Hz)	Duration (ms)	F1 (Hz)	F2 (Hz)
paat	/ɑː/	224	619	1048	334	799	1181
pad	/a/	112	626	1145	194	841	1493
pat	/æ/	138	596	1525	186	878	1775
peet	/e/	211	360	2043	257	307	2578
pet	/ɛ/	94	463	1929	174	441	2399
pêt	/ɛː/	280	427	1924	290	436	2563
piet	/i/	98	292	2043	130	284	2513
pit	/ə/	76	509	1646	131	544	1810
poet	/u/	90	331	1218	123	292	1396
pot	/ɔ/	102	484	1001	171	491	1011
put	/œ/	101	474	1558	174	470	1673
puut	/yː/	217	272	1979	293	270	2525

Figure 5.5 shows the vowels located in the vowel space for the male and female voice vowels. The F1, F2 and duration cues are plotted against each other to provide a description of the vowel space.

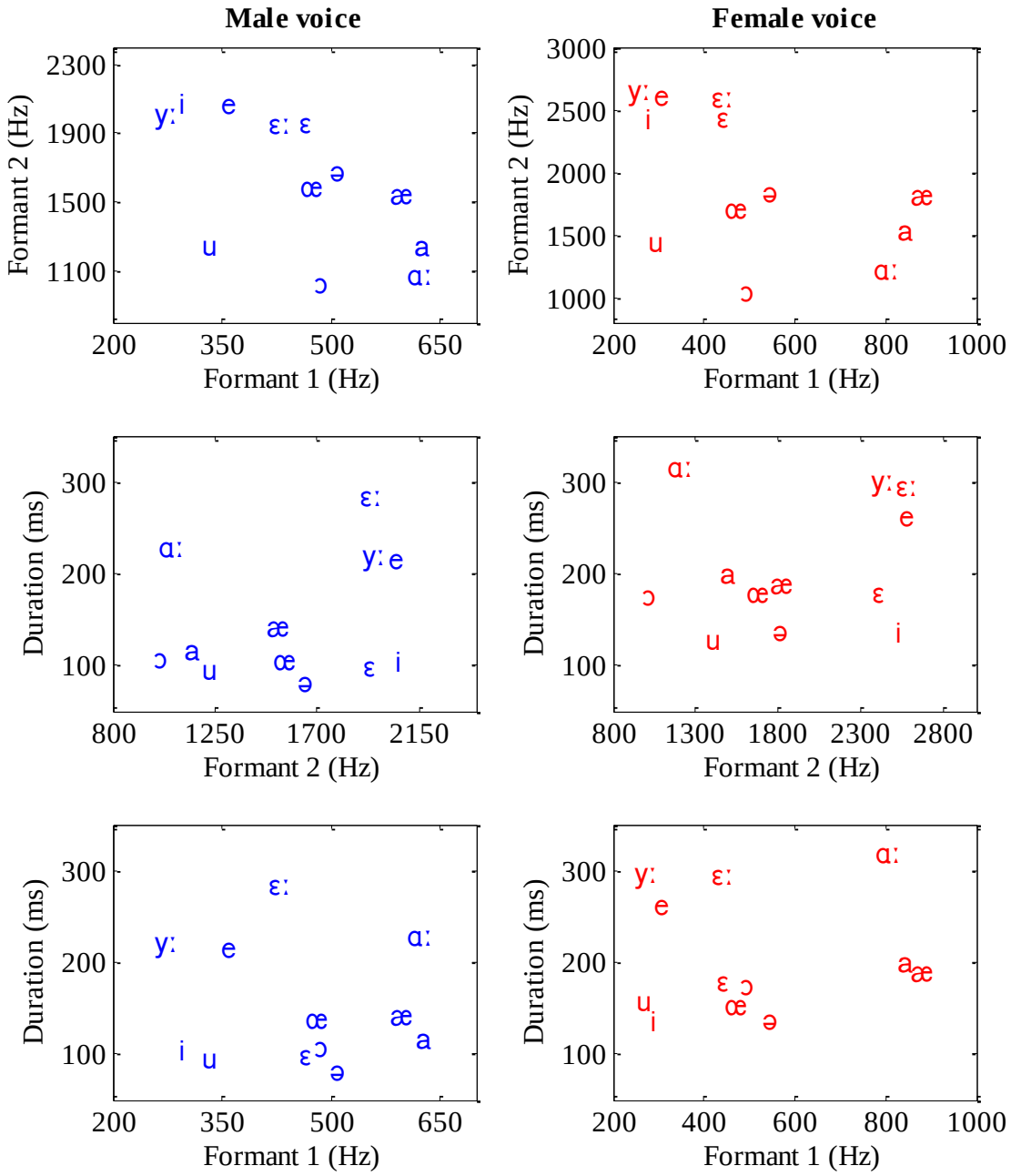

Figure 5.5. Vowels located in the F1, F2 and duration vowel space for the male voice (left column) and female voice (right column).

For both the male and female voice the vowel pairs /yː/ and /i/, /ɛː/ and /ɛ/ as well as /a/ and /ɑː/ are closely related in the F1 and F2 space although they are separated by different duration values. If confusions occur between these vowels, one can assume that the duration cue is absent. Since F1 and F2 are closely situated to each other, confusions could occur when only these cues are present.

For both voices the vowels /œ/ and /ə/ are likely to be confused with each other, as all three the vowel space cues for these two vowels lie in close proximity. The vowels /u/ and /ɔ/ are unlikely to be confused with any other vowels when all cues are available to the listener, since neither of these two are located near any other vowels. If the duration information for the vowel /ɑː/ is preserved, it may be expected that this vowel would not be confused with another vowel, since it's duration value exceeds the duration for any other vowel in the proximity of its vowel space. Confusions between the longer duration vowels /yː/, /ɛː/ and /e/ are however a possibility, since these are the three longest duration vowels in close proximity in the F1 and F2 vowel space.

One can expect more confusions between the vowels when either F1 or F2 is suppressed. An example would be when F1 is suppressed for the vowel /ɔ/. Confusions can be expected between this vowel and the /u/ or /a/ vowels since the F2 and duration values for these vowels are similar. When F2 is suppressed, /ɔ/ is expected to be confused with /ɛ/, /œ/ or /ə/ since the F1 and duration cues of these four vowels are related.

Although the F1 and F2 values of the same vowels for the male and female speakers differ somewhat, the overall distribution of the vowels in the vowel spaces are quite similar.

5.3.2 Spectral bands space

For each vowel that was synthesized by the whole-spectrum method, the dB SPL values were determined for five spectral bands, spaced two-thirds of an octave from each other (Klein et al., 1970), as is described in section 4.3.2.3. The spectral band values for both the male and female voice vowels are given in Table 5.2

Table 5.2. Spectral band values, given in dB SPL

		Male voice					Female voice				
		Band 1	Band 2	Band 3	Band 4	Band 5	Band 1	Band 2	Band 3	Band 4	Band 5
paat	/ɑː/	68	71	70	64	59	63	70	72	65	48
pad	/a/	69	72	73	70	59	61	70	73	69	60
pat	/æ/	69	70	70	71	67	58	70	72	69	67
peet	/e/	68	64	52	64	70	69	67	56	61	68
pet	/ɛ/	72	69	58	70	71	72	71	60	64	66
pêt	/ɛː/	70	64	52	61	63	72	70	55	54	62
piet	/i/	70	64	51	65	69	71	69	57	63	68
pit	/ə/	74	72	64	68	66	71	72	66	65	70
poet	/u/	73	69	69	67	57	72	70	58	58	55
pot	/ɔ/	73	73	65	51	51	72	72	69	63	52
put	/œ/	73	71	65	68	65	71	70	65	68	67
puut	/yː/	68	63	49	63	66	68	67	53	63	69

Figure 5.6 and Figure 5.7 show the five spectral band values for the vowels /ɑː/ and /ɛ/ respectively.

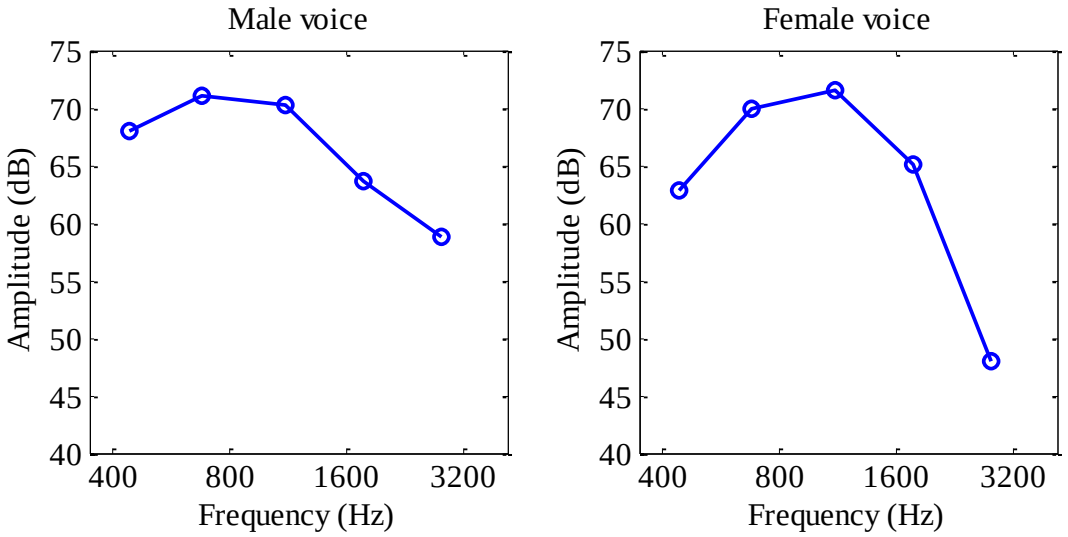

Figure 5.6. Five spectral band values (male and female speaker) for the vowel /ɑː/.

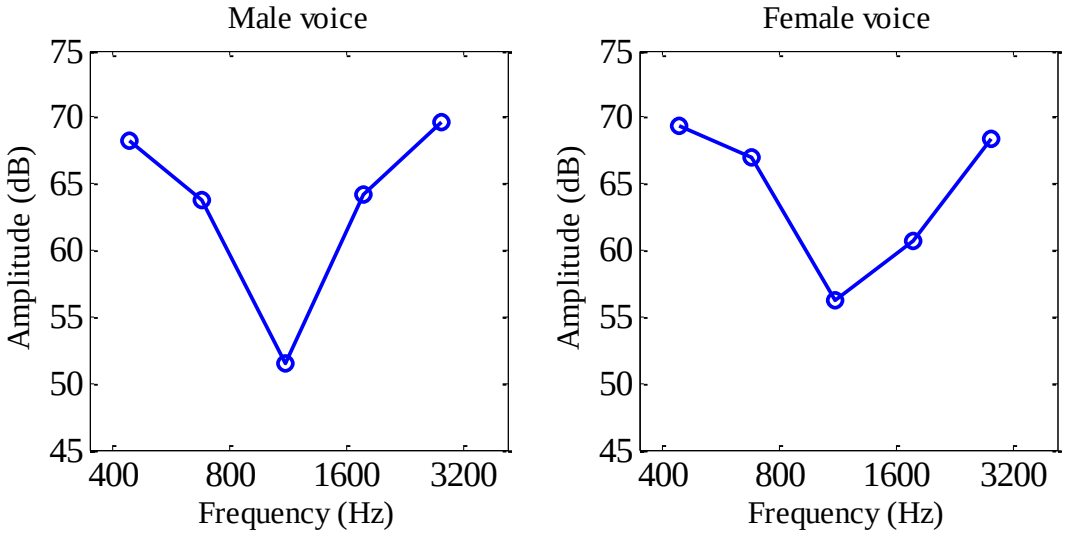

Figure 5.7 Five spectral band values (male and female speaker) for the vowel /ɛ/.

The five-dimensional vowel space derived from the spectral bands is plotted in Figure 5.8 and Figure 5.9. For the male voice, the three vowels /æ/, /ɑ/ and /ɑː/ are clearly distinguished from the rest of the four vowels in the four plots of band 1 against bands 2 to 5. In the plots of band 2 against band 3 and band 4, the two vowels /ɵ/ and /yː/ are separated from the rest of the vowel group. Two distinct vowel groups are formed in the plots of band 3 against band 4 and band 5. In the plots of band 3 against band 4 and band 5, as well as band 4 against band 5, the vowels are separated from each other more than in the other plots.

For the spectral band plots of the female voice, the three vowels /æ/, /ɑ/ and /ɑː/ are again located in close proximity to each other for the majority of all plots, while it is clearly distinguished from the other vowels in the plot of band 3 against band 1. Similar to the plots for the male voice, the two vowels /ɵ/ and /yː/ is in close proximity to each other, and also forms a distinct group with /ɨ/ and /ɛː/ in the plots of band 2 against bands 3 to 5 and band 3 against bands 4 and band 5.

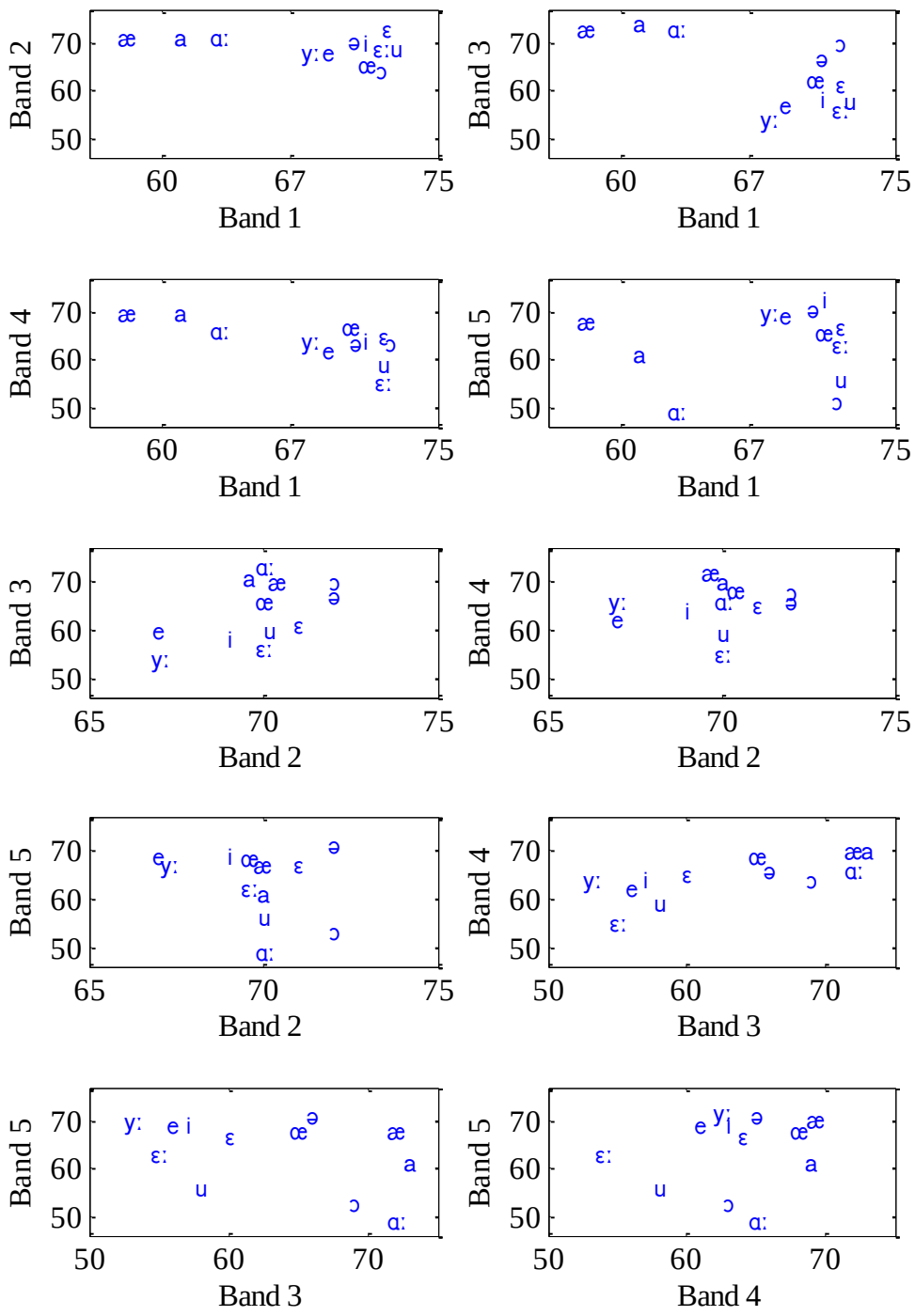

Figure 5.8 Plots of five-dimensional spectral space for the male voice vowel.

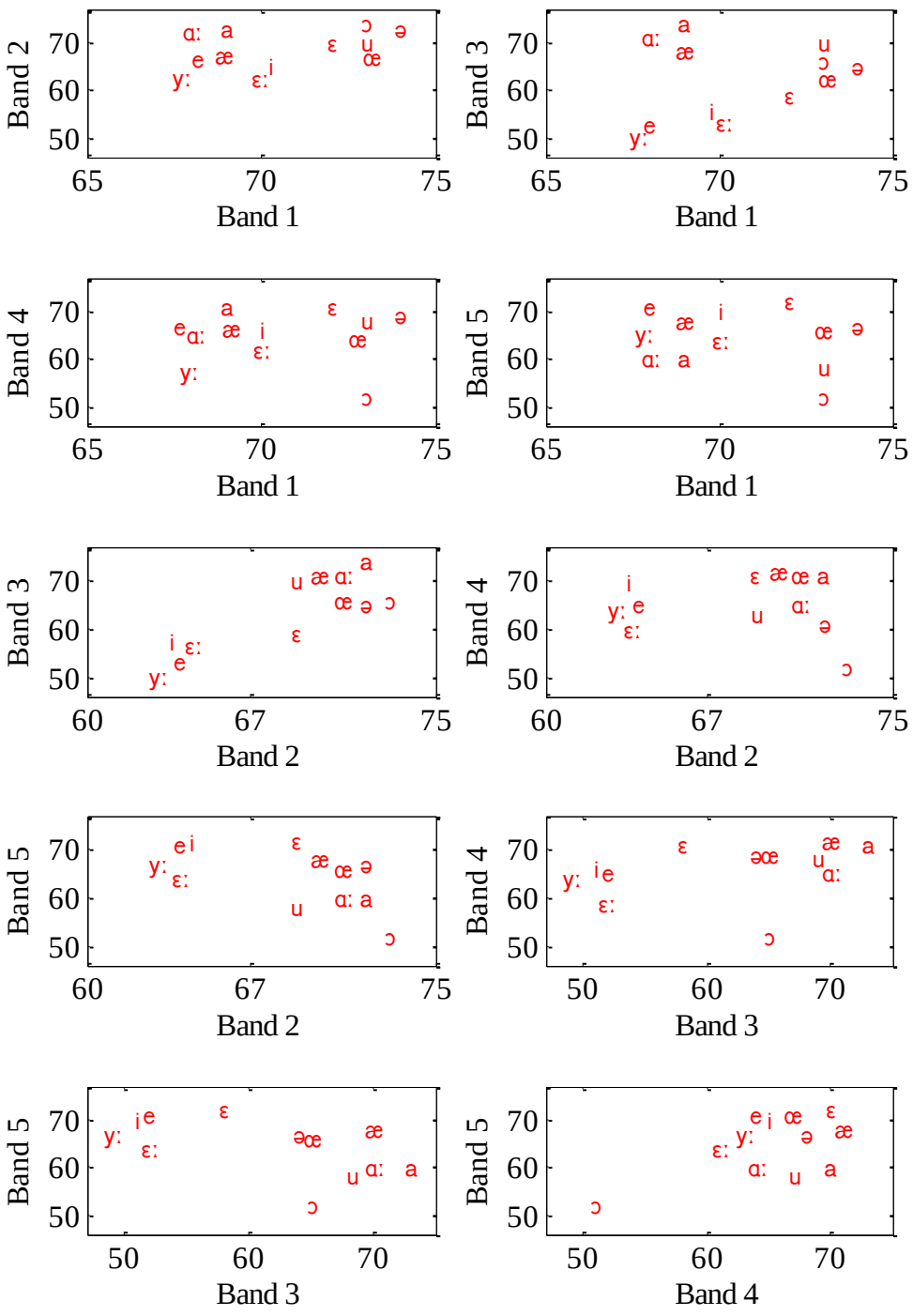

Figure 5.9 Plots of five-dimensional spectral space for the female voice vowel.

To accurately analyze the vowels in a five-dimensional space is a difficult task. Table 5.3 and Table 5.4 show the normalized Euclidian distances between each vowel pair for the spectral band space of the male and female voices respectively.

Overall, for both the male and female voice spectral band space, the following vowel groups can be identified and would be likely confused with each other, under the assumption that whole-spectrum information is used by listeners: group one as /œ/, /ɛ/ and /ə/, group two as /yː/, /i/, /e/, /ɛː/ and /ɛ/, and group three as /ɑː/, /a/ and /æ/.

Table 5.3. Normalized Euclidian distances between vowels for the spectral band vowel space (male voice). Shaded blocks indicate distances smaller than 0.5 (shown in bold) that correspond between the male and female voice.

		paat /ɑː/	pad /a/	pat /æ/	peet /e/	pet /ɛ/	pêt /ɛː/	piet /i/	pit /ə/	poet /u/	pot /ɔ/	put /œ/
pad	/a/	**0.24**	**0**									
pat	/æ/	**0.39**	**0.30**	**0**								
peet	/e/	0.80	0.91	0.75	**0**							
pet	/ɛ/	0.66	0.68	**0.47**	**0.38**	**0**						
pêt	/ɛː/	0.71	0.86	0.77	**0.26**	0.50	**0**					
piet	/i/	0.82	0.92	0.76	**0.07**	**0.37**	**0.26**	**0**				
pit	/ə/	**0.43**	**0.45**	**0.31**	0.59	**0.29**	0.59	0.58	**0**			
poet	/u/	**0.24**	**0.26**	**0.40**	0.79	0.62	0.69	0.79	**0.38**	**0**		
pot	/ɔ/	0.57	0.77	0.92	0.99	0.99	0.78	1.00	0.80	0.61	**0**	
put	/œ/	**0.37**	**0.39**	**0.27**	0.60	**0.33**	0.59	0.60	**0.07**	**0.32**	0.77	**0**
puut	/yː/	0.84	0.96	0.84	**0.16**	**0.50**	**0.18**	**0.16**	0.67	0.81	0.96	0.67

Table 5.4. Normalized Euclidian distances between vowels for the spectral band vowel space (female voice). Shaded blocks indicate distances smaller than 0.5 (shown in bold) that correspond between the male and female voice.

		paat /ɑː/	pad /a/	pat /æ/	peet /e/	pet /ɛ/	pêt /ɛː/	piet /i/	pit /ə/	poet /u/	pot /ɔ/	put /œ/
pad	/a/	**0.44**	**0**									
pat	/æ/	0.70	**0.29**	**0**								
peet	/e/	0.93	0.77	0.74	**0**							
pet	/ɛ/	0.82	0.66	0.67	**0.24**	**0**						
pêt	/ɛː/	0.89	0.89	0.93	**0.34**	**0.40**	**0**					
piet	/i/	0.91	0.74	0.73	**0.13**	**0.14**	**0.37**	**0**				
pit	/ə/	0.82	0.55	0.52	**0.43**	**0.27**	0.62	**0.36**	**0**			
poet	/u/	0.68	0.79	0.92	0.52	**0.46**	**0.29**	**0.50**	0.66	**0**		
pot	/ɔ/	**0.36**	0.52	0.74	0.74	0.58	0.66	0.69	0.61	**0.44**	**0**	
put	/œ/	0.78	0.51	0.51	**0.42**	**0.25**	0.63	**0.34**	**0.16**	0.64	0.58	**0**
puut	/yː/	1.00	0.83	0.78	**0.15**	**0.33**	**0.44**	**0.20**	0.52	0.60	0.83	**0.48**

In the MDS analysis of the experimental results, both the F1, F2 and duration, as well as the spectral bands vowel space was evaluated against MDS results to determine which of the two vowel spaces provided the best prediction of the experimental outcomes.

5.3.3 Effect of noise on formants and spectral shape

Since the aim of this study is to evaluate the availability of vowel cues in severe noise, the effect of speech-shaped noise on the vowel spectrum is investigated. Figure 5.10 and Figure 5.11 depict the effect of speech-shaped noise on the spectrum of the whole-spectrum synthesized vowel /e/ and the formants-only synthesized vowel /æ/. From the figures it is noted that a decreasing SNR has a substantial effect on the spectral shape and amplitudes of the formants. Spectral contrast is also severely decreased as a result of the noise. Spectral contrast has a substantial effect on vowel intelligibility (Leek et al., 1987;

Kiefte, Enright and Marshall, 2007), while a change in the detailed spectral shape may also cause an increase to vowel identification errors (Ito et al., 2001; Zahorian and Jagharghi, 1993). One can therefore expect an increase in identification errors in severe noise as a result of the low spectral contrast and altered spectral shape..

Noise does not affect the vowel spectrum uniformly. Figure 5.10 and Figure 5.11 show a larger effect of noise on the spectrum in the regions of F2 and F3 than F1. This correlates with the findings of Parikh and Loizou (2005) who found that in noise (-5 dB SNR) the spectral shape in the F2 region (1 – 2 kHz) was manipulated by noise to a higher extent than the F1 region.

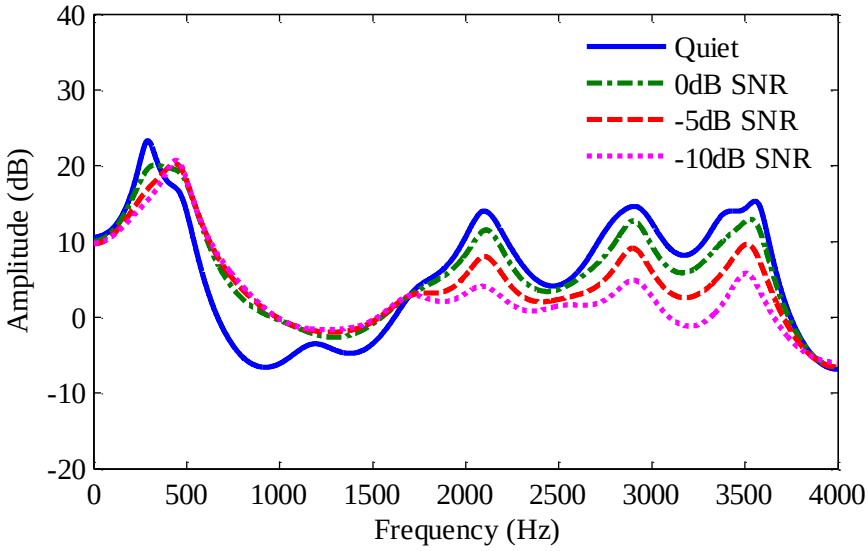

Figure 5.10. LPC spectra of the whole-spectrum vowel /e/ for different SNRs.

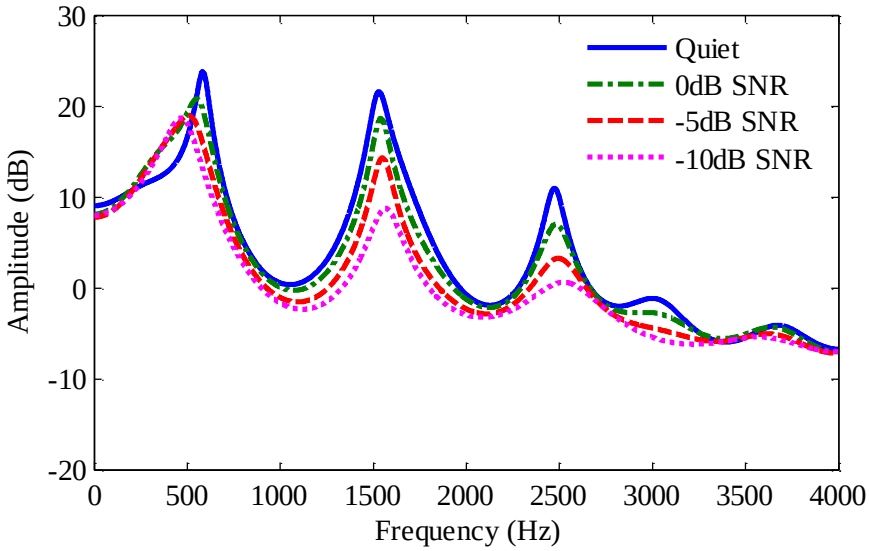

Figure 5.11. LPC spectra for formants-only vowel /æ/ for different SNRs.

Figure 5.12 and Figure 5.13 show the effect of noise on the formant contours of the whole-spectrum synthesized vowel /yː/ and the formants-only synthesized vowel /ɐ/. It can be seen how noise affects the overall spectral shape and spectral contrast through time. The effect of noise is not constant through time at low SNR levels, which would make the task of identifying the vowel much more difficult. In severe noise at -10 dB SNR, formant movement through time is almost completely obscured, while only partial formant and spectral detail information are still available. The diphthong of /ɐ/ would be highly affected by the distortion as formant movement plays an important role in the identifications of diphthongs in noise (Nabelek et al., 1996).

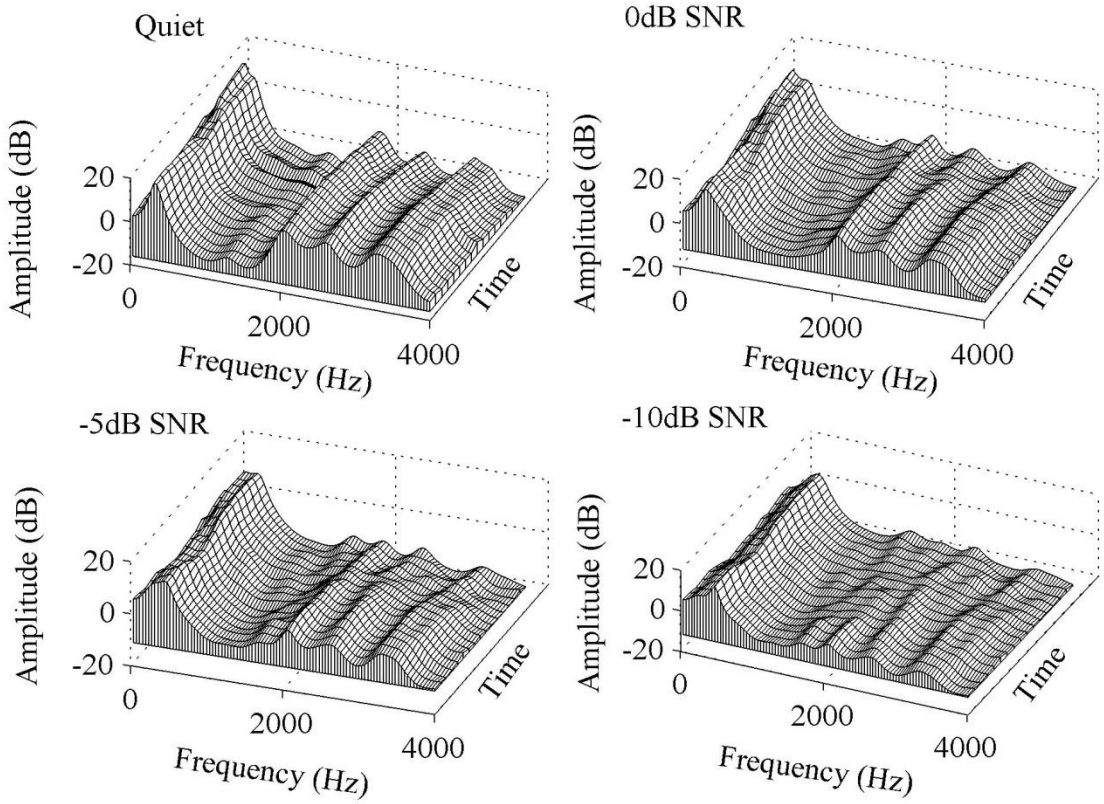

Figure 5.12. Whole-spectrum synthesized vowel /yː/ (male voice). As the SNR decreases, spectral contrast decreases while uneven effects of noise are seen through the duration of the vowel.

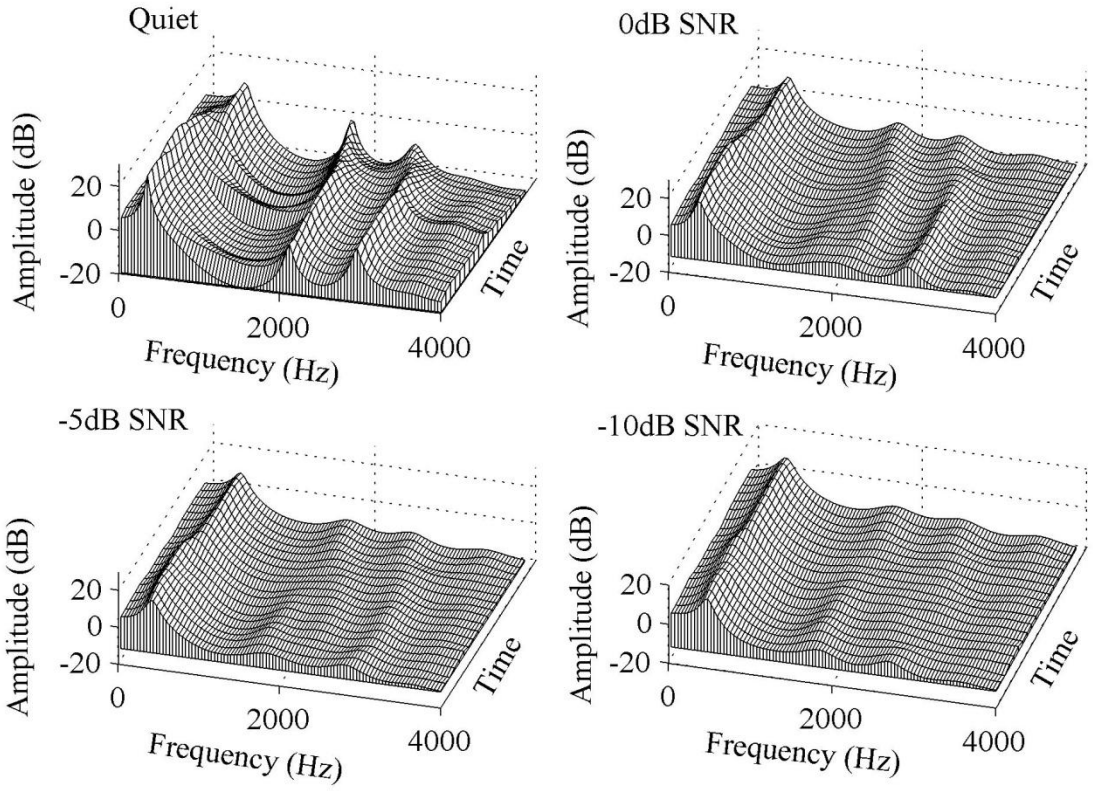

Figure 5.13. Formants-only synthesized vowel /ɐ/ (male voice). As the SNR decreases, less information about F2 and F3 remains, while clear F1 information is still seen.

5.4 RESULTS OF LISTENING TESTS

The results of the speech recognition experiments described in Chapter 4 are presented next. The percentage correct scores are presented, as well as the influence of noise, formant suppression and synthesis type (whole-spectrum or formants-only) thereupon. The percentage correct scores give an indication of the specific cues that assist listeners to successfully recognise vowels. Vowel confusions are analysed by the inspection of the raw confusion matrices, and also by FITA and MDS analyses. These results give an indication of the specific vowel cues that is perceived by listeners in noise.

A summary of the results are provided in 5.4.1, after which detailed analysis of all the results are given in the sections that follow. The reader may disregard sections 5.4.2 to 5.4.4 if an in-depth analysis of the results is not required.

5.4.1 Summary of results

Mean percentage correct scores for the 12 listeners were calculated for the two synthesis methods (formants-only and whole-spectrum), three spectral manipulation types (complete spectrum vowel, F1-suppressed and F2-suppressed) and SNR (-10 dB, -5 dB, 0 dB and quiet). The mean and standard deviation values for this data are shown in Table 5.5. A three-way repeated measures ANOVA was used to evaluate the effects of synthesis type (ST), spectral manipulation (SM) and SNR. Mauchly's test indicated that the assumption of spherity had been violated only for the effect of the interaction

$$\mathrm{ST}\times\mathrm{SM}\times\mathrm{SNR}[\ \chi^2(20)=35.359, p<0.05\ ].$$

Therefore, the degrees of freedom are corrected using Greenhouse-Geisser estimates of spherity with $\varepsilon = 0.48$ (Field, 2009).

All factors as well as interactions are found to be statistically significant: ST [F(1,11) = 131.4; p < 0.0001], SM [F(2,22) = 242.5; p < 0.0001], SNR [F(3,33) = 336; p < 0.0001], ST× SM [F(2,22) = 63; p < 0.0001], ST×SNR [F(3,33) = 15.4; p < 0.0001], SM×SNR [F(6,66) = 6.2; p < 0.0001], ST×SM×SNR [F(2.9,31.7) = 24; p < 0.0001]. Post hoc tests using Bonferroni adjustment show a significant main effect of the SNR between all combinations of SNRs (p < 0.0001) except between 0 dB and quiet (p > 0.05). Vowel perception in noise at a SNR of 0 dB therefore remained essentially at the same level as in quiet.

Figure 5.18 (top panel) depicts the percentage correct scores as bar plots for the vowels without any spectral manipulation, comparing the scores for the whole-spectrum and formants-only synthesized vowels. The whole-spectrum vowels yielded recognition scores

significantly higher than the formants-only vowels at -10 dB SNR ($F(1,22) = 14.83$; $p = 0.001$), while the formants-only vowel scores were higher than the whole-spectrum vowel scores at 0 dB SNR ($F(1,22) = 8.2$; $p < 0.01$) and in quiet ($F(1,22) = 5.52$; $p < 0.05$). When comparing vowel score differences between the quiet and -10 dB SNR condition, the formants-only vowel scores showed a 39.4% decrease in vowel recognition while the whole-spectrum vowel scores decreased by 21.7%.

Apart from these group differences between whole-spectrum and formants-only synthesized vowels, individual differences are also evident, with particular vowels being more robust against noise. Consideration of individual vowel scores reveals that the better overall recognition of the formants-only vowels over the whole-spectrum vowels at low noise levels (quiet and 0 dB SNR conditions) is primarily due to the scores of the vowels of /ɛ:/ and /ɛ/. In severe noise, however, vowels /æ/, /e/, /i/, /u/, /ɛ/, /œ/ and /a/ yield better recognition for the whole-spectrum vowels than for the formants-only vowels. The two back vowels /ɑ:/ and /a/ prove to be the most robust in severe noise.

Tukey multiple comparison tests indicate that the scores for the whole-spectrum vowels are significantly higher than the F1- and F2-suppressed vowels at all SNRs ($p < 0.05$). Vowel scores for the F1-suppressed vowels are significantly higher than the F2-suppressed vowels in low noise conditions (0 dB SNR and quiet, $p < 0.05$), while in severe noise, the suppression of F1 or F2 has a similar detrimental effect. The effect of either F1 or F2 suppression therefore leads to a significant reduction in vowel recognition scores in severe noise, while in the quieter conditions, F2 is more important in conveying vowel information than F1.

For the F1-suppressed vowels, the whole-spectrum vowel scores are significantly higher than the formants-only vowel scores at -10 dB ($F(1,22) = 5.25$; $p < 0.05$) and -5 dB ($F(1,22) = 7.48$; $p < 0.05$), while no significant differences are found for the quieter listening conditions. For the F2-suppressed vowels, the whole-spectrum vowel scores are

significantly higher than the formants-only vowel scores at all SNRs ($p < 0.05$ for 10 dB, -5 dB, 0 dB SNR and quiet) [$F(1,22) = 12.52, 9.41, 4.6, 73.76$]. These results suggest that (i) more robust cues for vowel identity are embedded in the whole-spectrum representation, as opposed to the depleted formants-only spectrum; (ii) the importance of having more complete spectral information of vowels available grows relative to the formants-only representation of vowels as noise increases.

Although the results above show the value of having more complete spectral information available when perceiving vowels in noise, this does not resolve the question of whether the auditory system extracts formant information from the available spectral information (in these experiments, either the whole-spectrum or the formants-only representation), or whether it relies on spectral shape information to recognize vowels. With the objective of determining whether vowel recognition and confusion patterns can be explained best by a vowel space spanned by formant axes (formant space) or by axes that characterize the complete spectrum (spectral-band space), an MDS analysis was carried out.

INDSCAL analysis was performed for the two synthesized vowel conditions (whole-spectrum and formants-only). Vowels were located in two different vowel cue spaces (or vowel spaces) that had either F1, F2 and duration cues as dimensions (formant space), or five spectral band cues (spectral-band space). The objective was not to search for cues that would match the MDS dimensions, but rather to compare the importance of formants to that of spectral shape when perceiving vowels in noise. A detailed explanation of the method followed in conducting MDS analysis is given in 4.3.2.

Figure 5.40 shows the least square error of the fit between the MDS dimensions and the two types of synthesized vowels represented either by a formant space or spectral-bands space. Results show that a fit between five MDS dimensions and the whole-spectrum vowels represented in spectral-bands space is generally inferior to a three-dimensional fit between MDS dimensions and either type of synthesized vowel represented in formant space. I.e., under all noise conditions tested, formant-space cues (F1, F2 and duration)

appear to be more important for recognizing vowels in noise than whole-spectrum cues.

Figure 5.43 shows the correlation between a three-dimensional MDS and the formant space cues (F1, F2 and duration) for the two vowel synthesis types. Each set of three bars (in each panel) shows which of the three cues correlated best with each of the three MDS dimensions. Variance accounted for (VAF) by the three MDS dimensions is also shown.

Some general trends may be observed. For either type of synthesized vowel, F1 appears to be the most salient cue for vowel identity for the male speaker under all noise conditions (i.e., highest correlation with the first MDS dimension). Using a VAF value of 0.6 as yardstick (Wickelmaier, 2003), it appears that three or more dimensions are needed to explain the experimental data in the quieter conditions (quiet and 0 dB SNR), while two dimensions can explain the data in the noisiest condition. This may be because some of the cue redundancy that listeners have available at good SNR wanes as cues are masked by noise at increasing noise levels, leaving fewer available cues for listeners to use.

Sections 5.4.2 to 5.4.4 provide more detail on all the results obtained from the current study. The reader may disregard these sections if a full description of the results is not required.

5.4.2 Percentage correct scores

The mean and standard deviation values for the vowel recognition scores are shown in Table 5.5. Statistical analysis on the data was provided in 5.4.1. In severe noise (-10 dB SNR) the average vowel recognition scores for the whole-spectrum vowels are about 13% better than the formants-only vowel scores. For the F1- or F2-suppressed stimuli, the whole-spectrum vowels are also recognized better than the formant-synthesized vowels (with a mean difference of 8% and 9% respectively). The main effects and interactions will be discussed next.

Table 5.5. Mean and standard deviation values for vowel recognition scores for all conditions.

		Whole-spectrum (DCT synthesis)				Formants-only (DSS synthesis)			
		-10dB	-5dB	0dB	Quiet	-10dB	-5dB	0dB	Quiet
Complete vowel	mean	62.5	76.5	78.3	84.2	48.9	71.9	84.7	89.3
	s.d.	9.3	5.6	5.5	5.0	8.0	9.2	5.4	5.7
F1-suppressed	mean	39.6	56.2	65.3	69.6	31.7	47.3	61.2	64.8
	s.d.	9.0	8.3	8.5	10.9	7.9	7.6	8.5	7.6
F2-suppressed	mean	34.2	51.6	57.8	65.4	25.3	40.8	50.7	39.9
	s.d.	7.3	7.6	6.7	6.7	4.8	9.6	9.3	7.8

5.4.2.1 Main effect of SNR

The effect of SNR is significant. This is not surprising since the significant influence of speech-shaped noise on speech recognition has been proven in literature (Gong, 1995; Parikh and Loizou, 2005; Phatak and Allen, 2007). Post hoc tests using Bonferroni adjustment prove a significant main effect of the SNR between all combinations of SNRs ($p < 0.0001$), except between 0 dB and quiet ($p > 0.05$) where the main effect is insignificant. The addition of noise at 0 dB SNR is therefore not as effective in changing vowel perception scores in comparison with the other SNRs.

5.4.2.2 Main effect of synthesis type

The effect of synthesis type is found to be significant. This shows that vowel scores are influenced differently by a formants-only and a spectral shape representation.

5.4.2.3 Main effect of spectral manipulation

Suppressing F1 and F2 has a significant effect on vowel perception. Post hoc tests using Bonferroni adjustment reveal a significant difference in vowel scores between all combinations of the three spectral manipulation conditions ($p < 0.05$). One can conclude in advance that the absence of F1 or F2 leads to a change in vowel recognition, and is subsequently important for vowel recognition in noise.

5.4.2.4 Synthesis type and SNR

The significant interaction between synthesis type and SNR (ST×SM) indicates that the difference in identification scores for the two synthesis type vowels (whole-spectrum vowels or formants-only vowels) depends on the SNR. Contrast tests were done to further investigate these interactions, with the quiet condition chosen as reference for comparison of each SNR to this condition. Contrasts reveal that the interaction effect of SNR and synthesis type is only significant between 0 dB SNR and quiet ($F(1,11) = 34.2$; $p < 0.0001$). Tukey multiple comparison tests indicate that a significant difference between the whole-spectrum and formants-only vowel scores are obtained at both -10 dB SNR and in quiet ($p < 0.05$).

5.4.2.5 Spectral manipulation and SNR

The difference between the vowel scores of different spectral manipulations depends on the SNR. Figure 5.14 shows the overall percentage correct scores for all the synthesized vowels. With the complete spectrum as reference factor, contrast tests reveal that the interaction effect is significant between the change between the complete spectrum to the F2-suppressed vowels and the change between quiet and -10 dB SNR ($F(1,11) = 19.111$; $p < 0.01$). A significant effect is also seen between the complete spectrum and the F2-suppressed vowels and the quiet and -5 dB SNR vowels ($F(1,11) = 11.693$; $p < 0.01$), as well as the quiet and 0 dB SNR vowels ($F(1,11) = 26.179$; $p < 0.0001$). No significant contrast effects are found for the change from the complete spectrum to the F1-suppressed vowels.

Tukey multiple comparison tests indicate that the scores for the complete spectrum vowels are significantly higher than the F1- and F2-suppressed vowels at all SNRs ($p < 0.05$). Vowel scores for the F1-suppressed vowels are significantly higher than the F2-suppressed vowels only at 0 dB SNR and in quiet ($p < 0.05$).

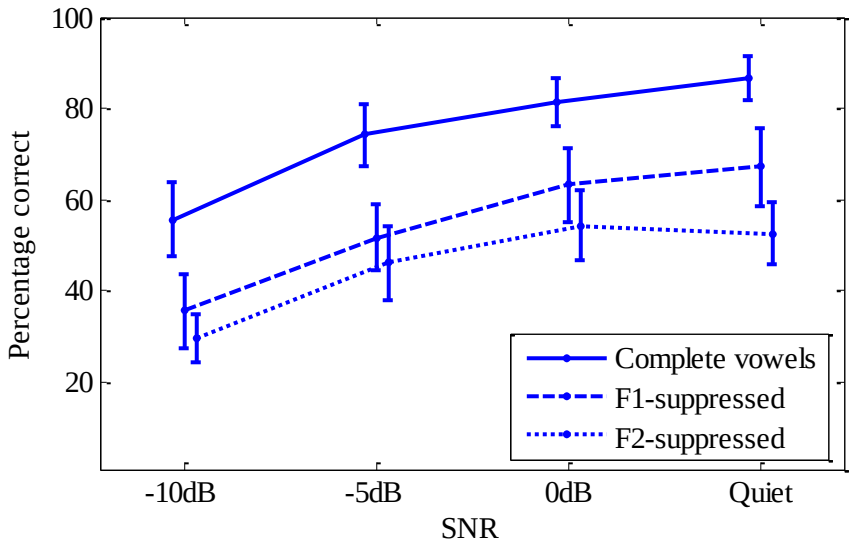

Figure 5.14. Overall percentage correct scores for all the synthesized vowels.

5.4.2.6 Spectral manipulation and synthesis type

The significant interaction between spectral manipulation and synthesis type signifies that the difference in vowel scores for the two synthesis types is dependent on the type of spectral manipulation. Contrasts indicate significant interaction specifically between synthesis type and the differences between the complete spectrum vowel with both F1-suppressed ($F(1,11) = 28.4$; $p < 0.0001$) and F2-suppressed vowels scores $F(1,11) = 160$; $p < 0.0001$) respectively. When disregarding the SNR, a significant difference is found between the whole-spectrum and formants-only vowels only for the F2-suppressed vowels, where the formants-only vowels are recognized significantly less than the whole-spectrum vowels ($p < 0.05$).

5.4.2.7 Interaction between spectral manipulation, synthesis type and SNR

The difference between the scores for the two synthesis types also depends on both the

type of spectral manipulation and on the SNR. Contrast tests reveal significant effects for all combinations except between -5 dB and quiet and between 0 dB and quiet. These insignificant effects can be observed in Figure 5.15 and Figure 5.16 where the gradient of the four connections (F1-suppressed and complete spectrum vowel for both whole-spectrum and formants-only types) between the 0 dB SNR and the quiet condition are similar. The same applies between the -5 dB SNR and quiet condition for the previously mentioned cases. The whole-spectrum and formants-only vowels are therefore similarly affected when F1 is suppressed, but only if noise is added at 0 dB and -5 dB SNR. In severe noise, different effects are found. The interaction between spectral manipulation, synthesis type and SNR is more significant for the F2-suppressed vowels than for the F1-suppressed vowels.

Figure 5.15 and Figure 5.16 show the scores for the whole-spectrum and formants-only synthesized vowels. The effect of speech-shaped noise for all SNRs can be seen. An increase in vowel recognition can be observed as the SNR is increased, with the only exception between 0 dB and the quiet condition for the formants-only synthesized F2-suppressed vowels where the vowel scores decrease.

Tukey multiple comparison tests indicate that for the scores of the whole-spectrum type vowels, the complete spectrum vowels are recognized significantly better than the F1- and F2-suppressed vowels at all SNRs ($p < 0.05$, Figure 5.15). Vowel scores for the F1-suppressed vowels are significantly higher than the F2-suppressed vowels only at 0 dB SNR ($p < 0.05$).

Tukey multiple comparison tests also indicated that the recognition scores for the formants-only complete spectrum vowels are significantly higher than the F1- and F2-suppressed vowels at all SNRs ($p < 0.05$) (Figure 5.16). Vowel scores for the F1-suppressed vowels are significantly higher than the F2-suppressed vowels only at 0 dB SNR and in quiet ($p < 0.05$).

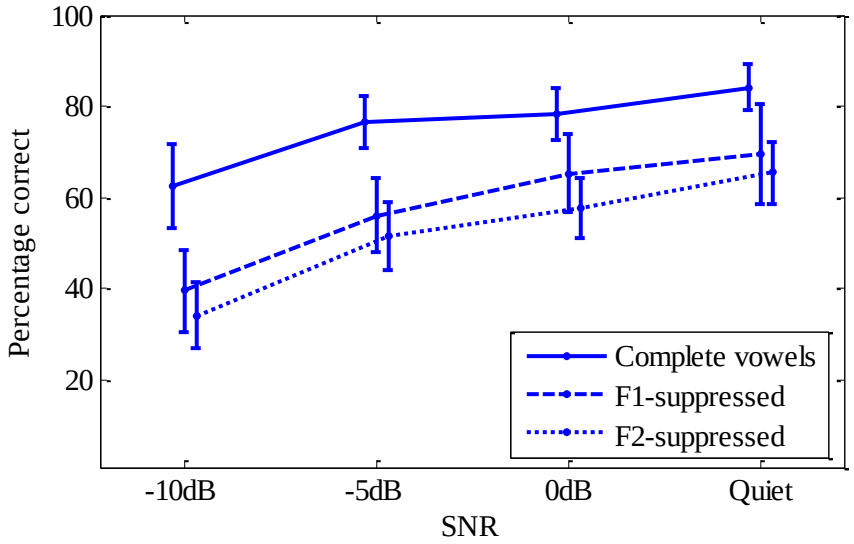

Figure 5.15 Percentage correct scores for whole-spectrum vowels for all noise conditions

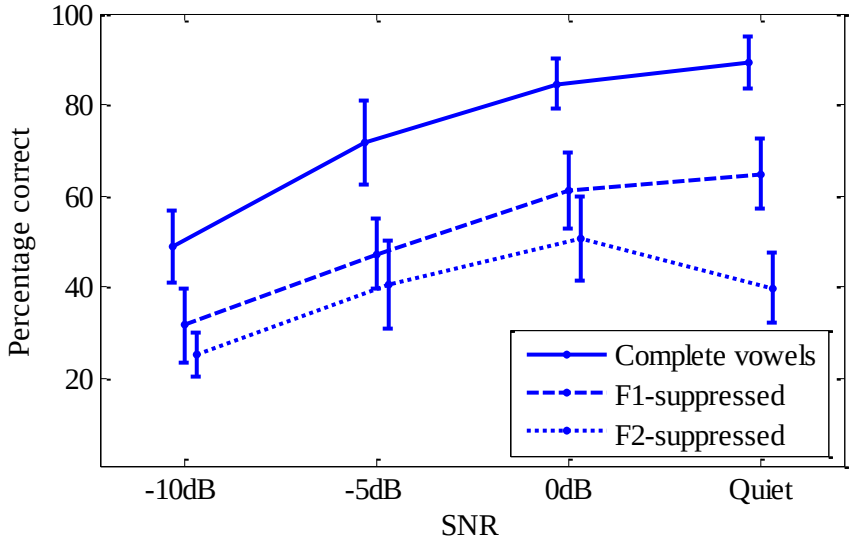

Figure 5.16 Percentage correct scores for formants-only vowels for all noise conditions.

Thus far, the percentage correct vowel scores were presented as the average scores between the male and female voice. Figure 5.17 shows the percentage correct scores for the formants-only and whole-spectrum vowels for all spectral manipulations, plotted for the male and female voice separately. The suppression of either F1 or F2 shows a

significant decrease in percentage correct score ($p < 0.05$), except for the whole-spectrum vowels of the male voice in quiet. For the formants-only vowels (male voice), a significant difference between the scores for the F1 and F2 suppressed vowels is noticed for the -5 dB SNR, 0 dB SNR and quiet conditions, while for the female voice a significant difference is only seen in quiet ($p < 0.05$). For the whole-spectrum vowels (male voice), a significant difference between the F1- and F2-suppressed vowel scores is only seen at -10 dB SNR, while for the female voice, the only significant difference is seen at 0 dB SNR ($p < 0.05$).

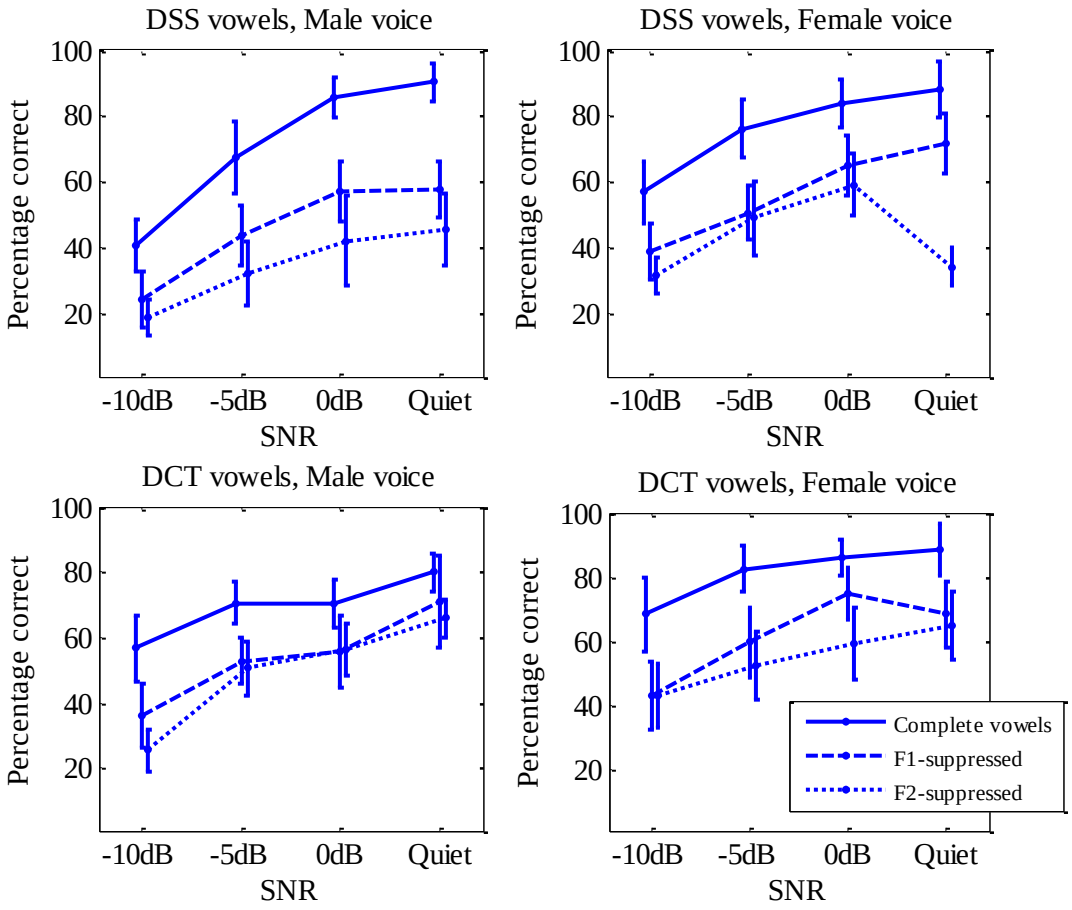

Figure 5.17. Percentage correct scores for the formants-only (DSS) and whole-spectrum (DCT) vowels for all spectral manipulations, plotted for the male and female voice separately.

Figure 5.18 depicts the percentage correct scores as bar plots for the complete spectrum, F1-suppressed and F2-suppressed vowels while directly comparing the scores for the whole-spectrum and formants-only synthesized vowels. For the complete spectrum vowels, the whole-spectrum vowels yield scores significantly higher than the formants-only vowels at -10 dB SNR ($F(1,22) = 14.83$; $p = 0.001$), while the formants-only vowel scores are significantly higher than the whole-spectrum scores at 0 dB SNR ($F(1,22) = 8.2$; $p < 0.01$) and in quiet ($F(1,22) = 5.52$; $p < 0.05$).

For the F1-suppressed vowels, whole-spectrum vowel scores are significantly higher than the formants-only vowels at -10 dB SNR($F(1,22) = 5.25$; $p < 0.05$) and -5 dB SNR ($F(1,22) = 7.48$; $p < 0.05$). For the F2-suppressed vowels, the whole-spectrum vowel scores are significantly higher than the formants-only vowels at all SNRs: -10 dB ($F(1,22) = 12.52$; $p < 0.01$), -5 dB ($F(1,22) = 9.41$; $p < 0.05$), 0 dB ($F(1,22) = 4.6$; $p < 0.05$) and in quiet ($F(1,22) = 73.76$; $p < 0.0001$).

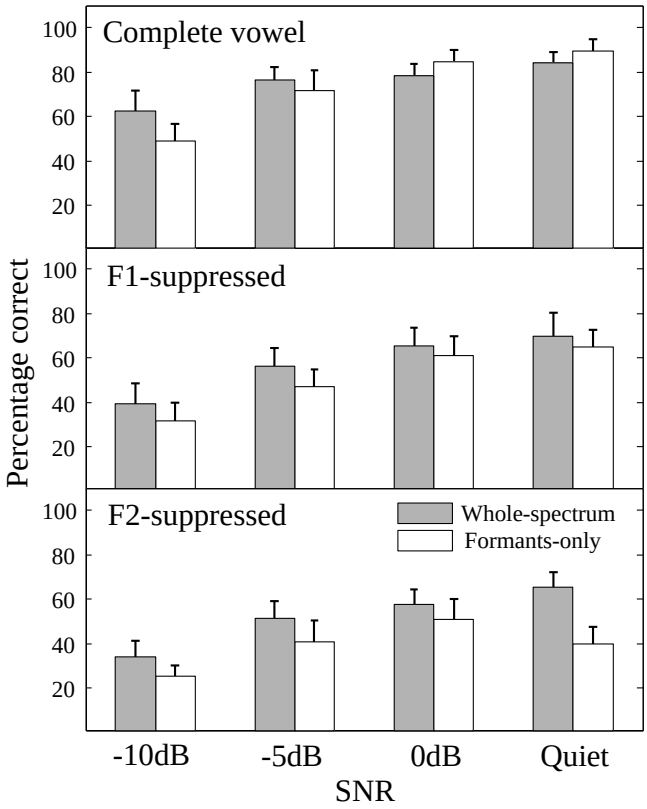

Figure 5.18. Percentage correct scores for the complete spectrum, F1- and F2-suppressed vowels while directly comparing the whole-spectrum vowels with the formants-only vowels.

5.4.2.8 Effect of speaker

Nabelek et al. (1992) found that vowel identity in noise does also depend on speaker differences. In the current study, the utterances of two speakers, one male and one female, were used for the vowel experiments. Although it is not the main focus in the current study, it is of interest to analyze the difference in vowel perception scores for the two speakers. Figure 5.19 shows the difference between vowel scores for the two speakers at each SNR for the complete spectrum vowels, F1-suppressed and F2-suppressed vowels. Except for the complete spectrum vowels and F2-suppressed vowels in quiet, the vowel scores for the female speaker is significantly higher than for the male speaker ($p < 0.05$) for all conditions. For the F2-suppressed vowels in quiet the male voice results are significantly higher than the female voice results ($p < 0.05$) while for the complete spectrum quiet condition no significant difference in vowel identification scores between

the two voices are found.

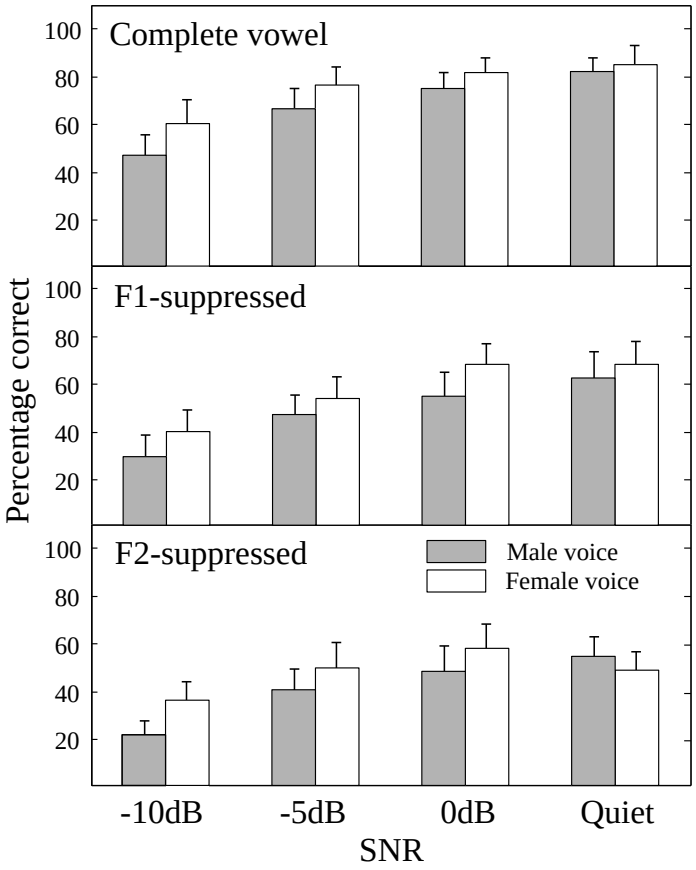

Figure 5.19 Percentage correct scores for the complete spectrum, F1- and F2-suppressed vowels with direct comparison of the male and female voice.

5.4.2.9 Individual vowel results

The mean recognition scores between all the vowels were analyzed as a whole in the previous section. It is necessary to investigate the effects of noise, loss in formants and the importance of whole-spectrum versus formant representations for individual vowels. Vowels can be classified in three categories depending on the frequency of its second formant. Back vowels have low F2 values (/ɑː/, /a/, /ɔ/ and /u/), front vowels have high F2 values (/ɛː/, /ɛ/, /yː/, /e/ and /i/), while central vowels have neither very high nor low F2 values (/æ/, /ə/ and /œ/).

With regards to the duration cue, it would be of interest to compare the vowels with a long duration (/ɛː/, /yː/, /ɐ/, /ɑː/) to vowels with a shorter duration. The perception of the only diphthong (/ɐ/) in the vowel group would also be of interest since spectral or formant movement through time would play an important role in its identification. To compare the importance of the spectral shape and formants directly, Figure 5.20, Figure 5.21 and Figure 5.22 show the direct difference between the whole-spectrum and formants-only vowel scores for each individual vowel for all SNRs and spectral manipulations.

Figure 5.20 depicts the difference between the whole-spectrum and formants-only vowel scores for each individual vowel at all SNRs. The reason for the better overall recognition of the formants-only vowels over the whole-spectrum vowels in the quiet and 0 dB SNR conditions is only due to the scores of the vowels /ɛː/ and /ɛ/, while no significant differences between the two types of synthetic vowels are found for the remaining vowels in these conditions. In severe noise, the vowels /æ/, /ɐ/, /i/, /u/, /ɛ/, /œ/ and /a/ show better recognition for the whole-spectrum vowels than for the formants-only vowels. The vowel with the poorest performance at -10 dB SNR is /ɛː/ while the two back vowels /ɑː/ and /a/ prove to be the most robust in severe noise. For the formants-only stimuli, the best recognized vowels at -10 dB SNR are /a/, /ɑː/, /ɔ/ and /u/. These four vowels all have an F1 to F2 separation of less than 1100 Hz. For the remaining formants-only vowels, recognition scores are all below 50%, while the F1 to F2 separation is larger than in the first group. For the whole-spectrum vowels, no relationship between the frequency distance of F1 and F2 and vowel recognition is found in severe noise. For the vowels /i/ and /ɐ/, which are both characterized by an F1 and F2 difference of more than 1500 Hz, recognition scores for the whole-spectrum vowels are found to be twice as high as those of the formants-only vowels.

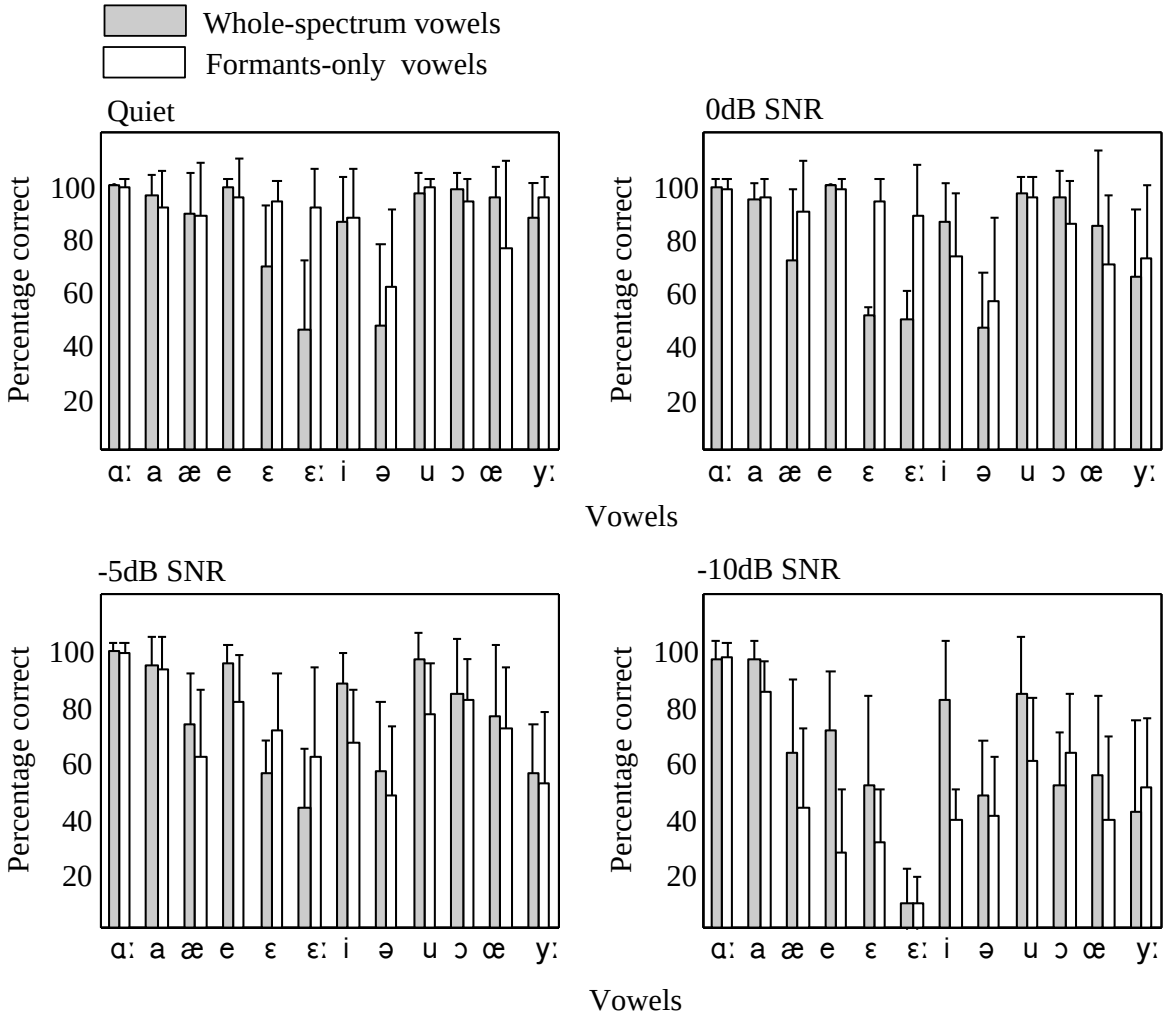

Figure 5.20 Individual vowel results for the complete spectrum vowels.

For the F1-suppressed tokens (Figure 5.21), the vowels /ɔ/ and /ʊ/ show better recognition for the whole-spectrum vowels in 0 dB SNR and quiet. In severe noise, the diphthong of /ɐ/ shows better recognition for the whole-spectrum than the formants-only vowels. The vowels /ɐ/, /ɑː/, /ɪ/, /a/ and /ə/ show no decrease in percentage correct scores due to the F1 suppression in quiet. These vowels all have either very high or very low F1 frequencies. In noise (-5 dB SNR), only /ɐ/, /ɑː/ and /ɪ/ show no significant decrease in recognition score due to the F1 suppression.

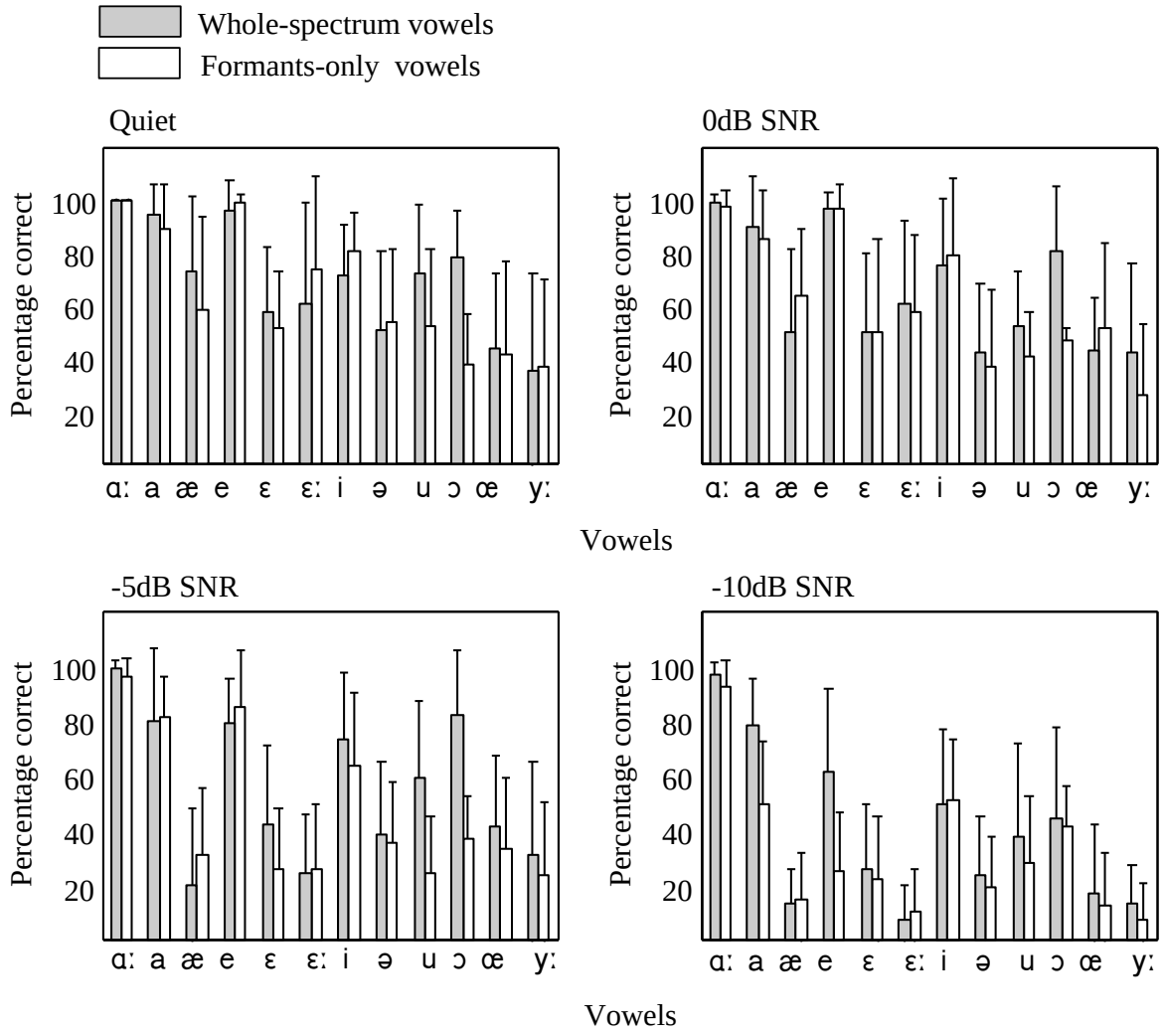

Figure 5.21 Individual vowel results for the F1-suppressed vowels.

For the F2-suppressed vowels (Figure 5.22), the majority of vowels show superior recognition for the whole-spectrum compared to the formants-only vowels in quiet. In noise at -5 dB SNR, mainly the front vowels show better recognition for the whole-spectrum than the formants-only vowels. The vowels /e/, /ɑː/, /i/, /ɔ/ and /yː/ show no decrease in percentage correct score due to F2 suppression. These vowels all have either very high or very low F2 frequencies. Scores for these vowels also did not decrease with added noise (at -5 dB SNR).

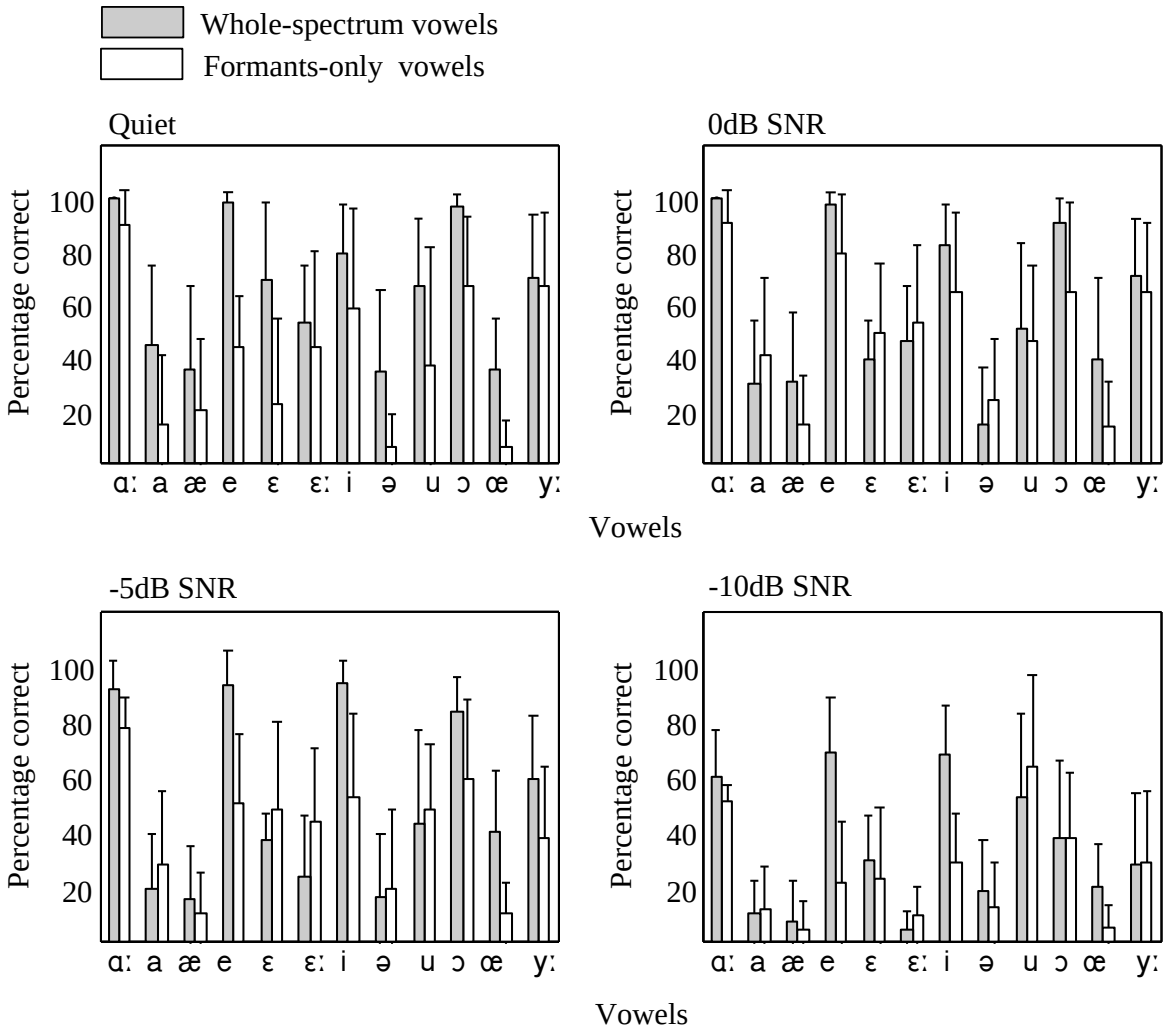

Figure 5.22 Individual vowel results for the F2-suppressed vowels.

Figure 5.23 depicts the drop in recognition scores due to the suppression of F1 and F2 in the two lowest SNRs. Each graph shows the difference between recognition scores for the vowels containing all formants and vowels with either F1 or F2 suppressed for each individual vowel. For the whole-spectrum vowels at -5 dB SNR, the largest drop in identification scores due to F1 suppression is for the vowels /ʊ/, /æ/ and /œ/, while vowel scores for the F2-suppressed vowels /ɑ/, /ʊ/, /æ/ and /œ/ show the largest decrease. In terms of vowel backness, these vowels are central vowels with the location F2 some distance away from F3 and F1. For the remaining whole-spectrum vowels, classified as either extreme back vowels or front vowels, only a small drop in recognition scores are found for the suppression of either F1 or F2. For the formants-only vowels, the same relationship between vowel backness and recognition scores are found, except for the F1-suppressed vowels /ɔ/ and /ɛ/, and for the F2-suppressed vowel /ɵ/. In severe noise at -10 dB SNR, the vowels /ɛ/, /ɵ/, /ɨ/, /ɛː/ and /yː/ show the least decrease in percentage correct scores due to F2-suppression for both the formants-only and whole-spectrum vowels. These vowels are all front vowels. For the F1 suppressed vowels, no relationship between vowel backness and a decrease in vowel scores is found.

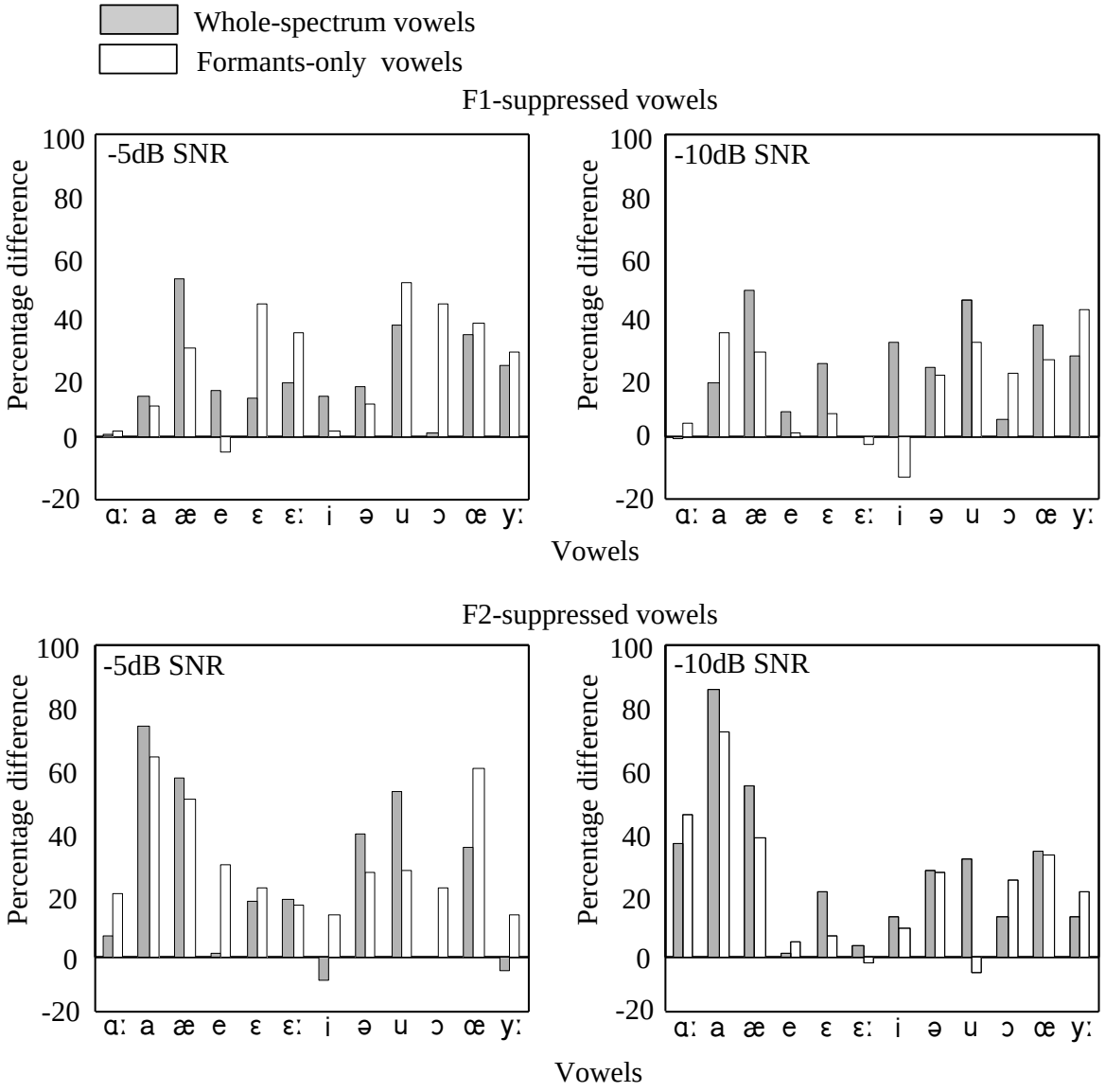

Figure 5.23. Effect of F1- and F2-suppression in noise. Each graph shows the difference between the complete spectrum and formant suppressed vowel for SNRs of -5 dB and -10 dB respectively. For each vowel, the recognition scores for the whole-spectrum and formants-only vowels are given.

5.4.3 Analysis of confusion matrices

From the vowel space analysis it was seen that some vowel tokens are likely to be confused with other vowels with the addition of uncertainty factors like noise. For further investigation regarding the importance of certain vowel cues in noise, it is necessary to analyze the actual vowel confusions from the raw confusion matrix data.

As was observed in 5.3.3 that noise decreases the spectral contrast of the spectrum and manipulates the spectral detail, which will increase the uncertainty of the listener. Noise can also cause diphthongs to be confused with monothongs when the transition segment of the vowel is less intense than the steady state part. It is important to note that the whole-spectrum vowels also contain formant information, but at a lower spectral contrast.

The confusion matrices obtained from the experimental study are presented in Figure 5.25 to Figure 5.36. Vowels tokens that were presented to the listeners are shown on the left column of the matrix, while the vowels that was perceived are shown in the top row of the matrix. Each entry in the matrix represents the average percentage of times the vowel corresponding to the stimulus is confused with vowels shown in the top row. For the shaded areas on the matrix diagonal, the value given is the percentage for which the vowel was identified correctly. For each case where the percentage confusion or recognition exceeds 10% for either the whole-spectrum or formants-only vowel, a breakdown of the whole-spectrum, formants-only and average between the two synthesis type vowel percentages is given. For each entry, the left number gives the percentage confusion or recognition of the whole-spectrum synthesized vowels while the number on the right depicts the percentage confusion or recognition of the formants-only synthesized vowels (see Figure 5.24). This would display major differences between confusions for the whole-spectrum and formants-only vowels.

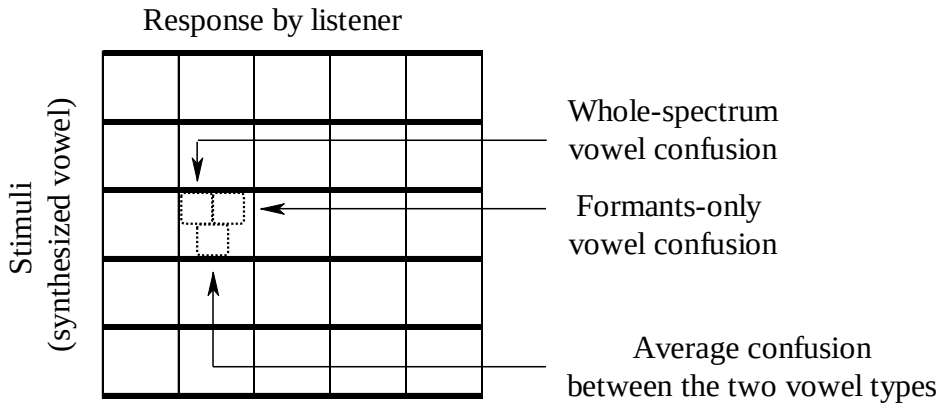

Figure 5.24. Description of the layout of the confusion matrix results.

For the confusion matrices of the complete spectrum and F2-suppressed vowels, the vowels are arranged according to increasing F1 values, while for the F1-suppressed vowels it is arranged according to increasing F2. If the major confusions are clustered around the diagonal for these matrices, it would provide evidence regarding the perception of F1 or F2 in the various conditions.

All confusion matrices are discussed in detail in Addendum A in terms of the locations of the confused vowel pairs in both the formants-only and whole-spectrum vowel space.

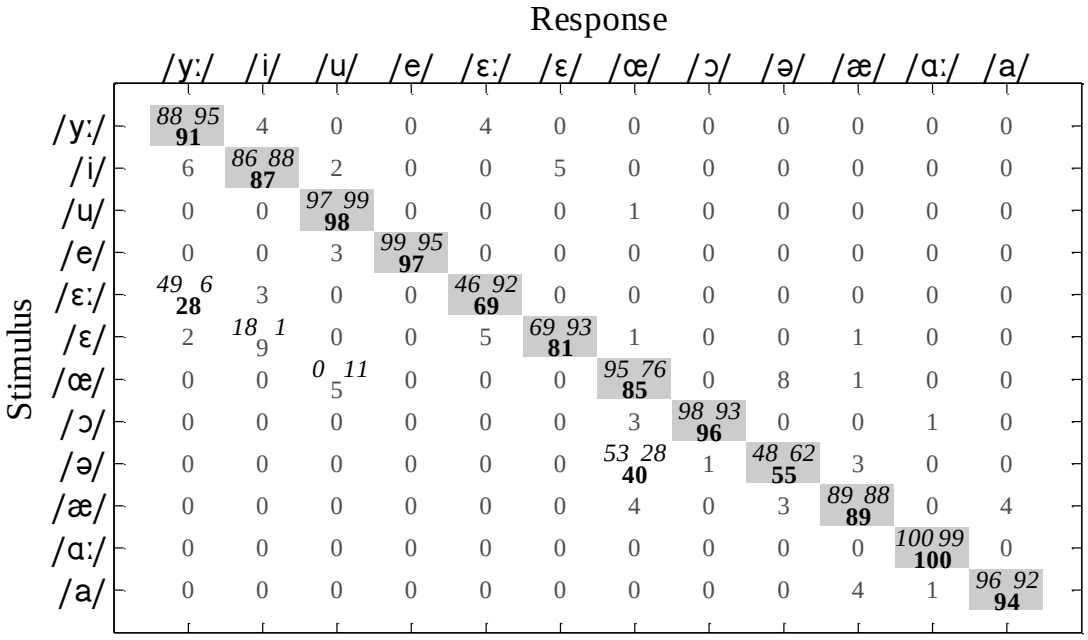

Figure 5.25. Confusion matrix results for the complete spectrum vowels in the quiet condition.

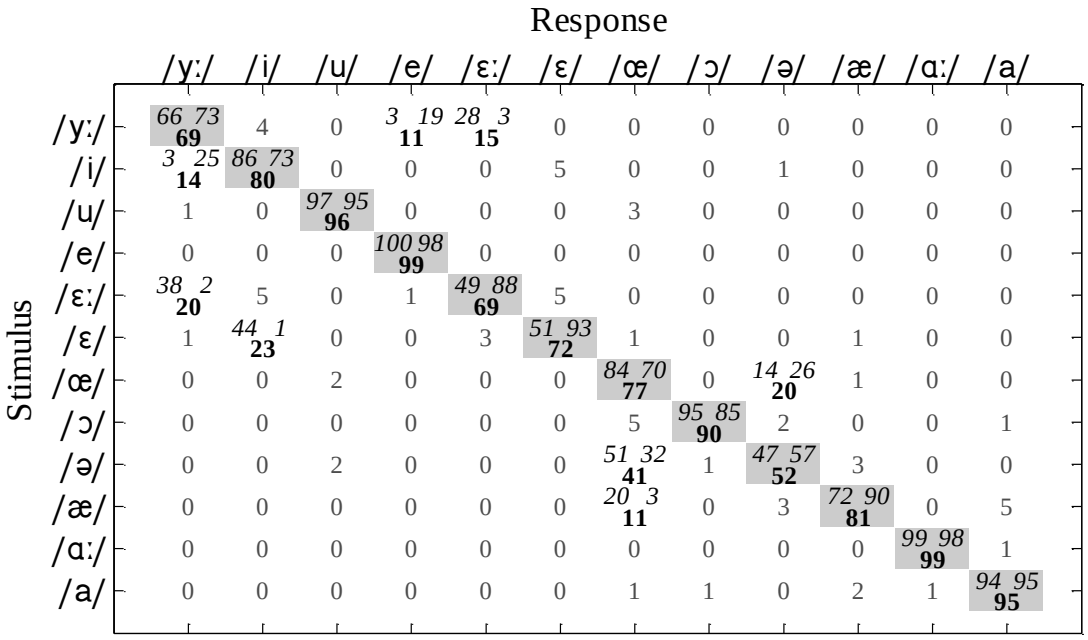

Figure 5.26. Confusion matrix results for the complete spectrum vowels in the 0 dB SNR condition (arranged by F1).

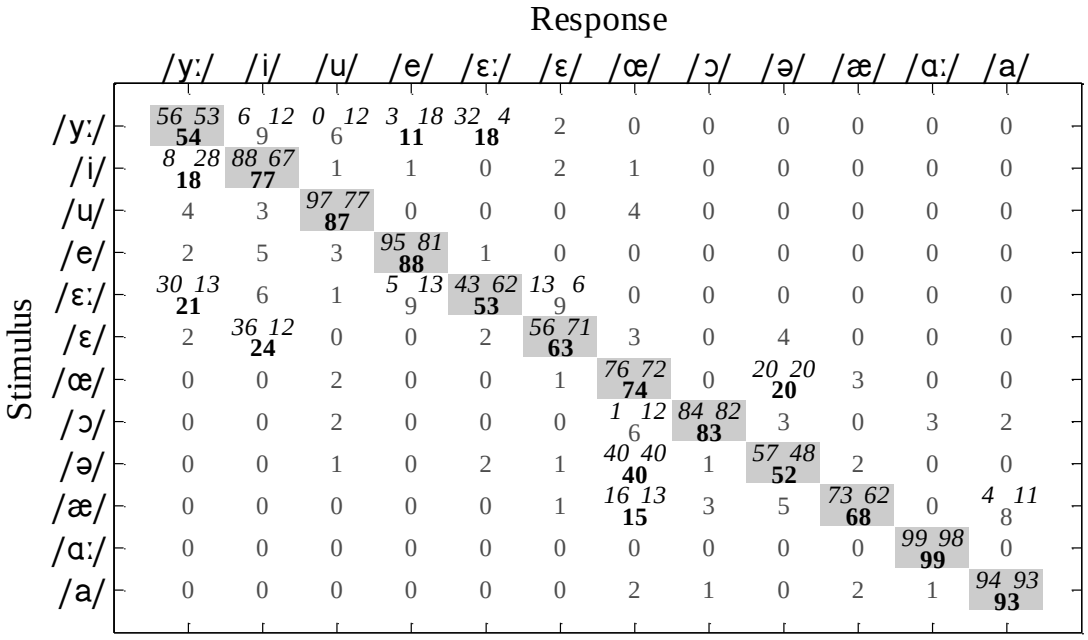

Figure 5.27. Confusion matrix results for the complete spectrum vowels in the -5 dB SNR condition (arranged by F1).

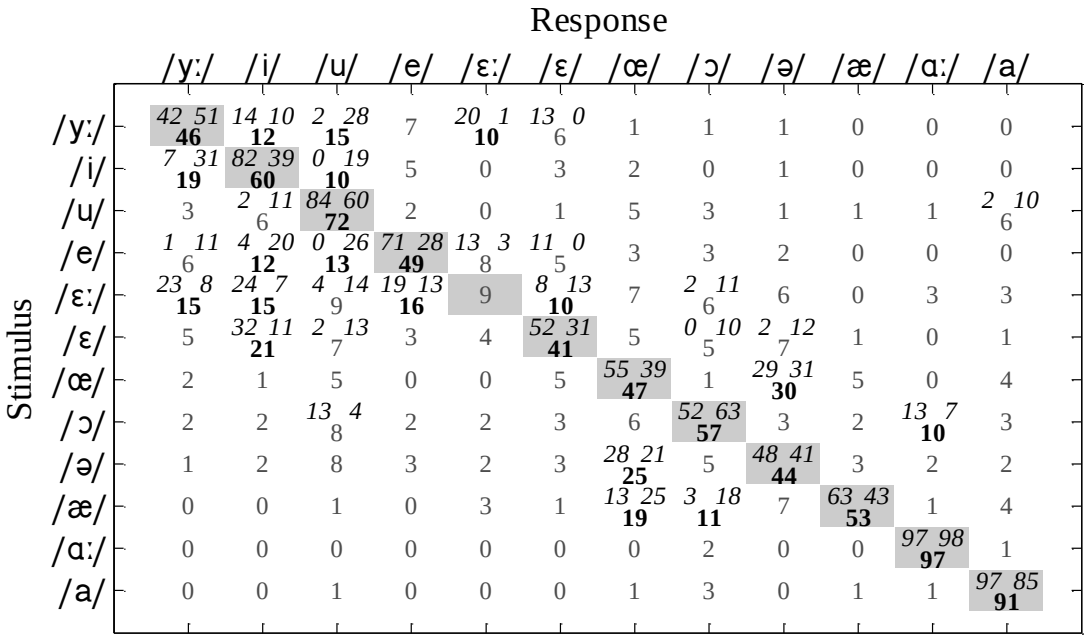

Figure 5.28. Confusion matrix results for the complete spectrum vowels in the -10 dB SNR condition (arranged by F1).

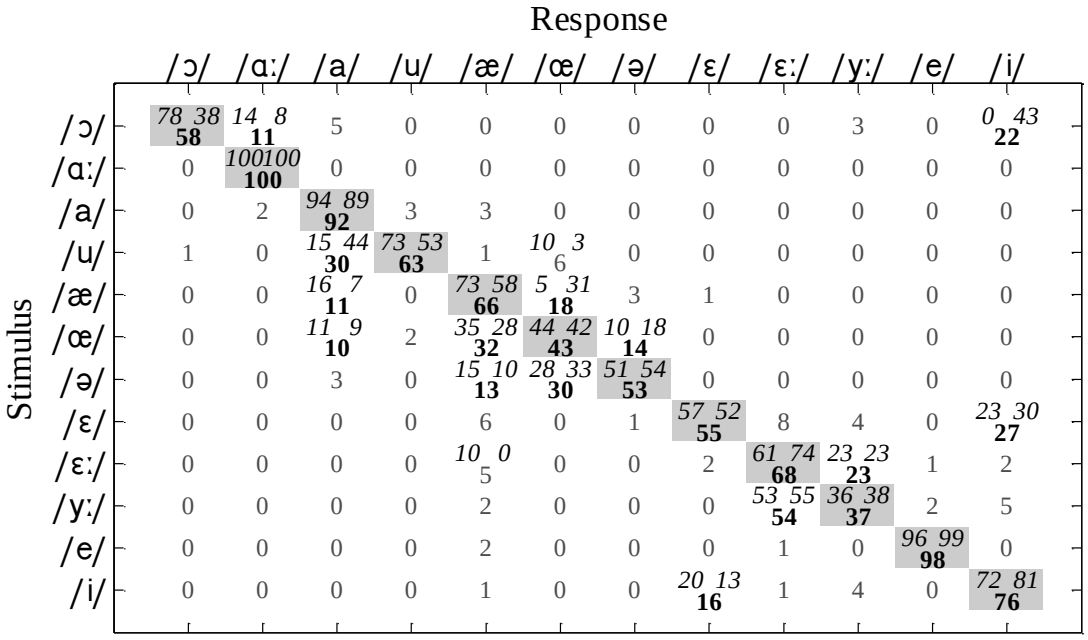

Figure 5.29. Confusion matrix results for the F1 suppressed vowels in the quiet condition (arranged by F2).

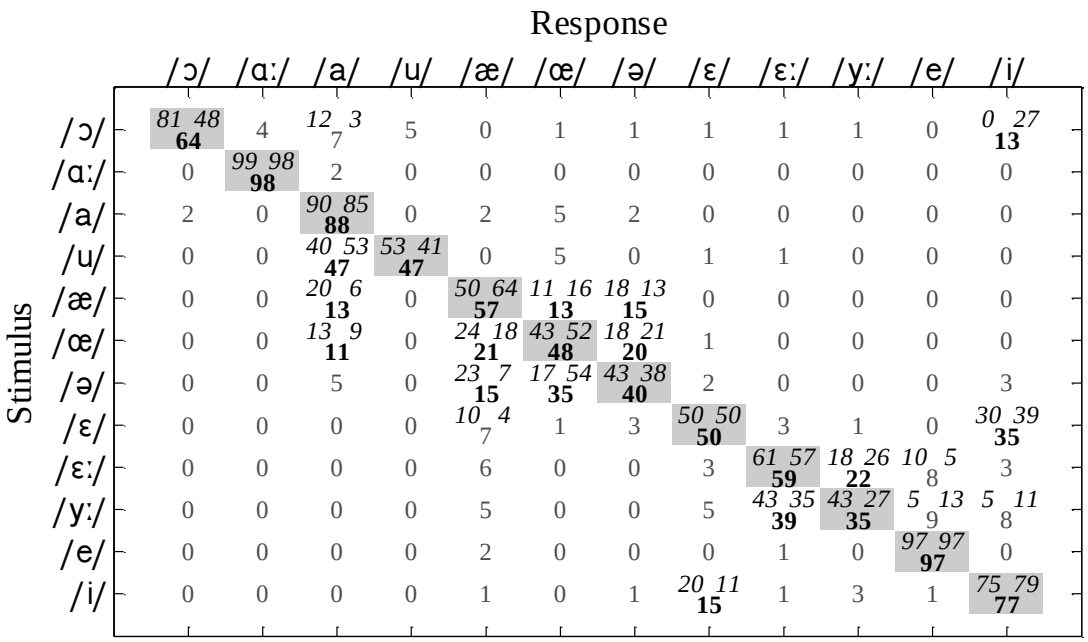

Figure 5.30. Confusion matrix results for the F1 suppressed vowels in the 0 dB SNR condition (arranged by F2).

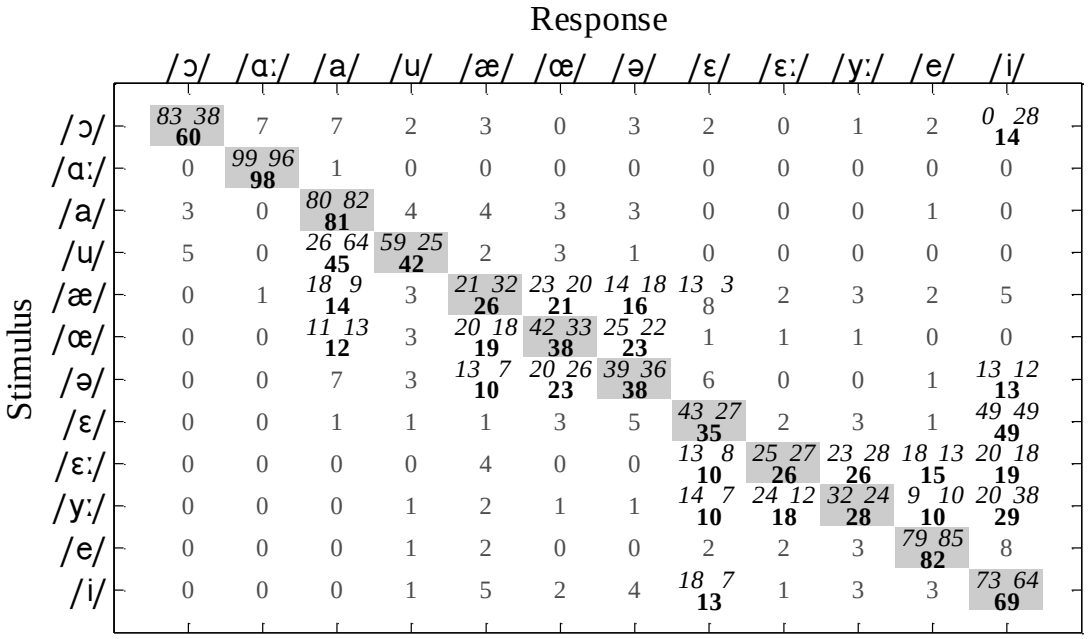

Figure 5.31. Confusion matrix results for the F1 suppressed vowels in the -5 dB SNR condition (arranged by F2).

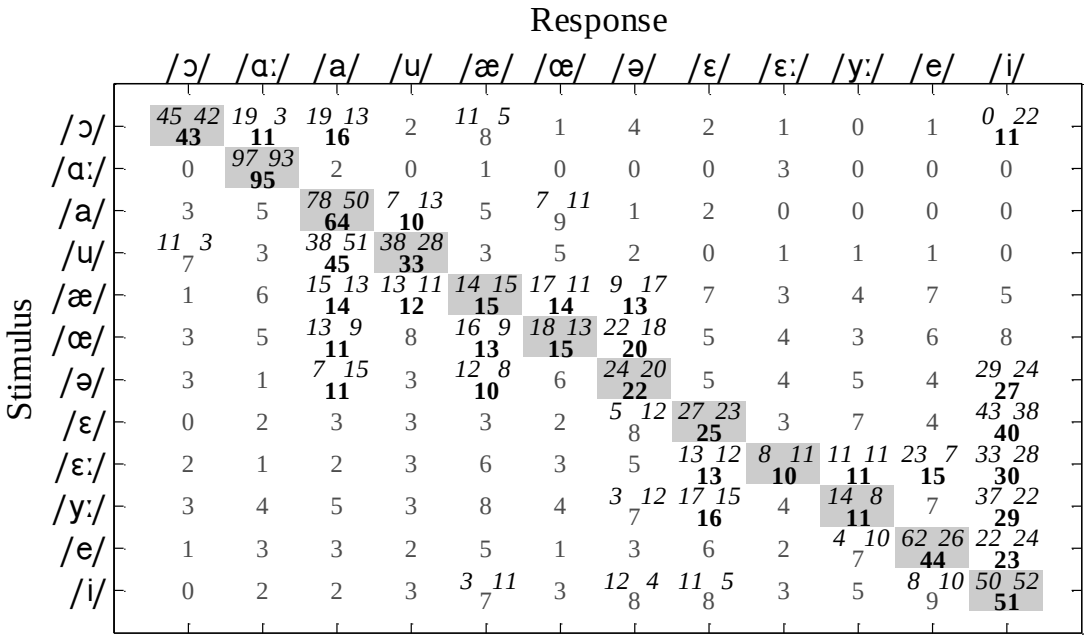

Figure 5.32. Confusion matrix results for the F1 suppressed vowels in the -10 dB SNR condition (arranged by F2).

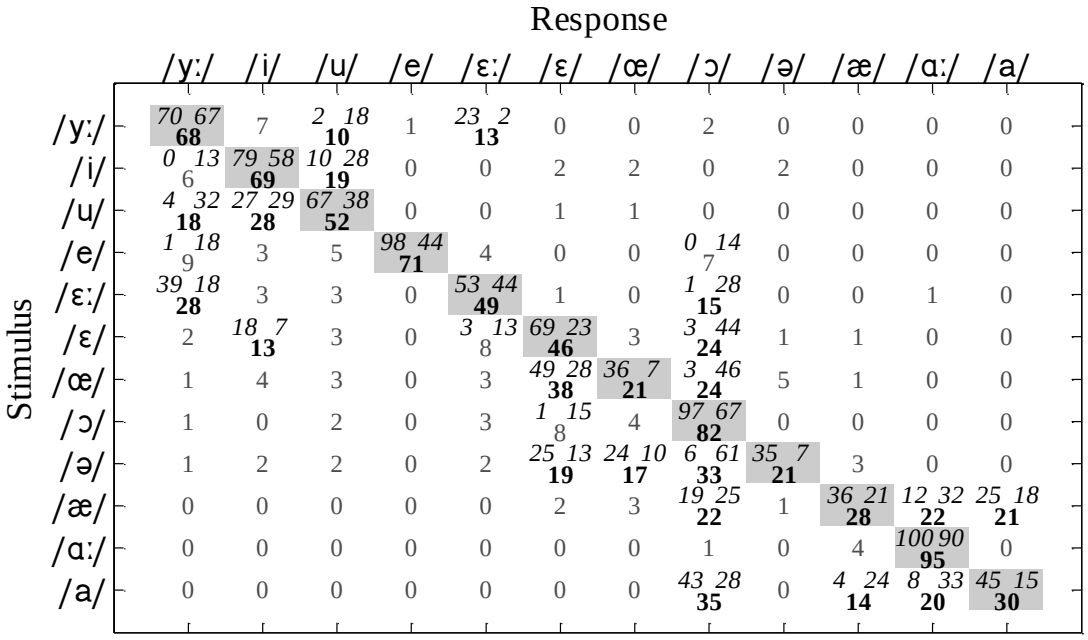

Figure 5.33. Confusion matrix results for the F2 suppressed vowels in the quiet condition (arranged by F1).

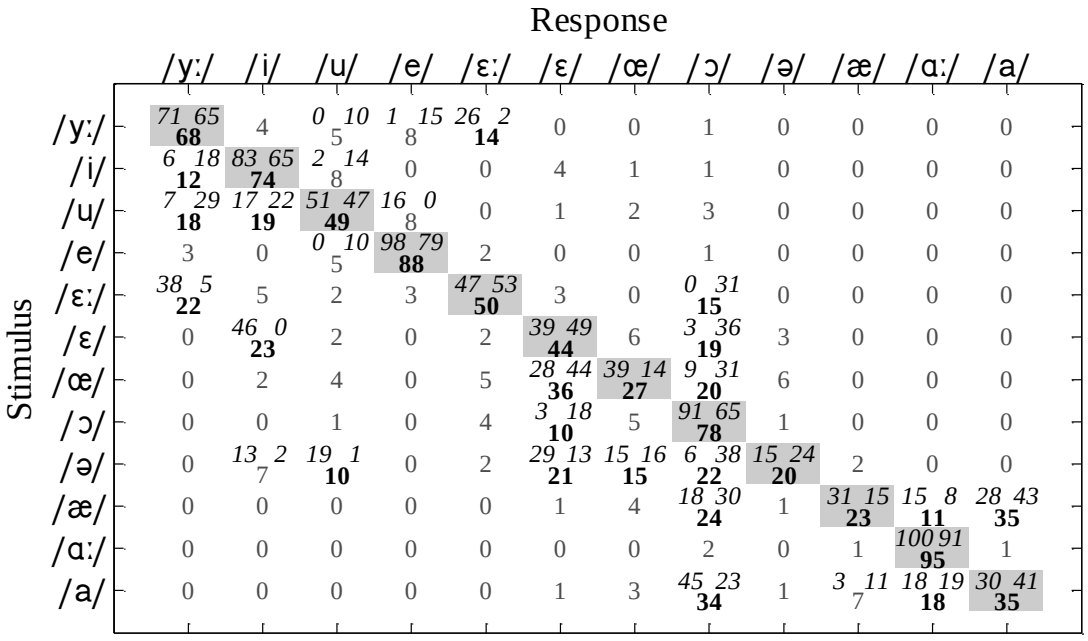

Figure 5.34. Confusion matrix results for the F2 suppressed vowels in the 0 dB SNR condition (arranged by F1).

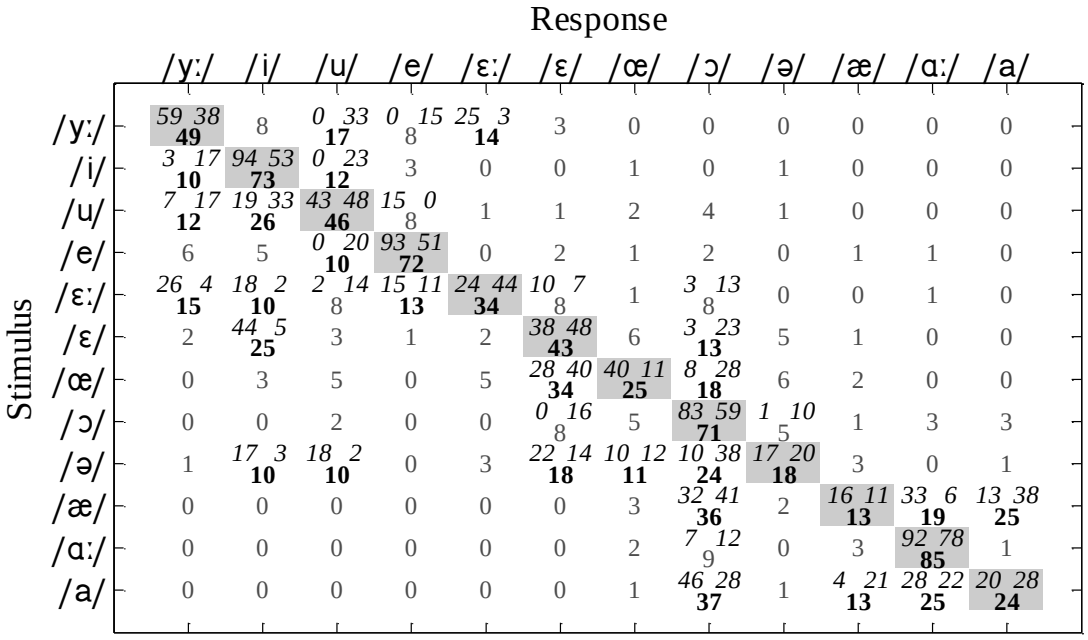

Figure 5.35. Confusion matrix results for the F2 suppressed vowels in the -5 dB SNR condition (arranged by F1).

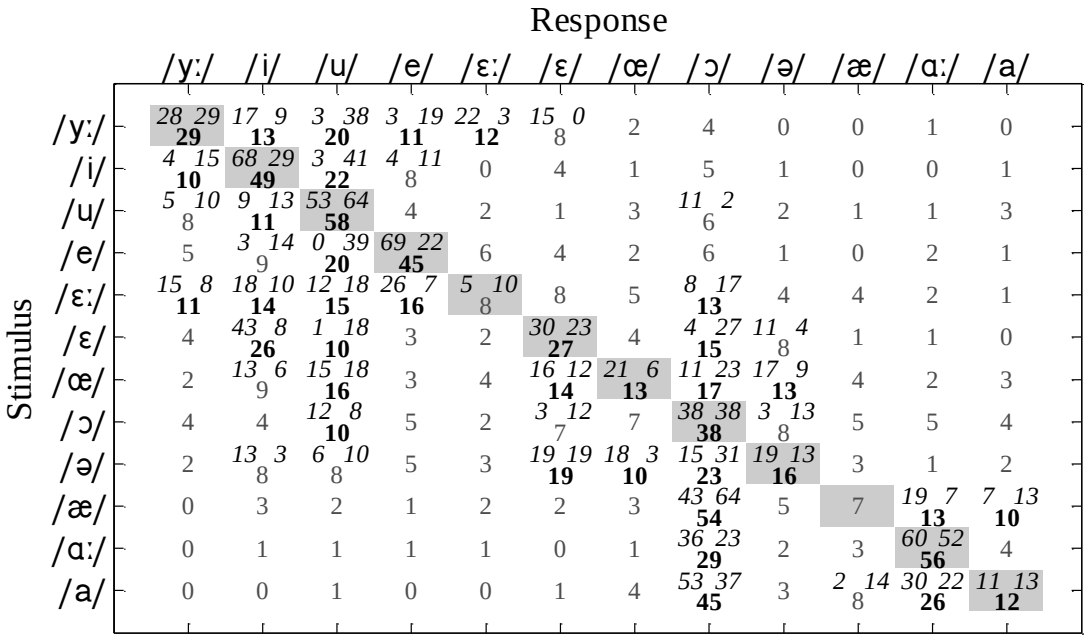

Figure 5.36. Confusion matrix results for the F2 suppressed vowels in the -10 dB SNR condition (arranged by F1).

Confusion matrices for the vowels with no formant suppression show that in quiet and 0 dB conditions, confusions mostly occur between vowels that share similar F1, F2, duration and spectral band cues. An example is the confusions of the vowel /ɛ/ with /i/ (only for whole-spectrum vowels), and the vowel /ə/ with /œ/ (both whole-spectrum and formants-only vowels). In severe noise at -10 dB SNR, substantially more confusions is seen between all vowels compared to the higher SNR conditions, demonstrating that noise adds to the uncertainty of listeners when making vowel judgments. The vowels /a/ and /ɑː/ do not show any confusion percentage exceeding 10% for both types of vowels.

For the whole-spectrum vowels, the two vowels /a/ and /ɑː/, together with /i/, /u/ and /ɔ/ show no confusions in severe noise. In all other confusions for the whole-spectrum vowels, the majority of confusions are due to either a similar F1 or F2 value, or due to proximity in the spectral bands space, while stimuli are confused with vowels sharing similar duration values to a lesser extent than the other cues. The vowels /e/, /ɛː/, /ɛ/, /ə/, /yː/ and /œ/ are confused with vowels situated close together in the spectral bands space. All these vowels also share either a similar F1, F2 or both formants with the vowels being confused, showing that whole-spectrum information is used together with formants in making vowels judgments.

Regarding the duration cue, confusion matrices show that in severe noise, the longer-duration vowels /ɑː/, /ɛː/, /e/ and /yː/ are confused with vowels having shorter and longer duration values, while shorter-duration stimuli are only confused with other shorter-duration vowels. Noise therefore eliminates duration cues mostly for longer-duration vowels containing only formant information, while the duration cues in shorter-duration vowels containing detailed spectral information are more robust in severe noise.

Confusion matrices for the vowels with F1 and F2 suppressed show that confusions occur between vowels sharing similar cues that is still available after formant suppression. For the F1-suppressed vowels in quiet and 0 dB SNR, confusions occur near the diagonal, showing that vowels are confused due to similar F2 cues, while near-diagonal confusions for the F2-suppressed vowels show that vowels are confused due to similar F1 cues. With

the addition of noise, confusions along the diagonal increases for both the F1-suppressed an F2-suppressed vowels. In severe noise, more confusions other than along the matrix diagonal occur for the F2-suppressed vowels than for the F1-suppressed vowels. A more thorough discussion regarding all vowel confusions can be found in Addendum A .

5.4.4 Feature information transmission analysis (FITA)

FITA analysis was performed on the confusion matrix results to obtain a measure of how effective the F1, F2 and duration cues were transmitted to the listeners.

5.4.4.1 Data Pooling

Prior to MDS (see section 4.3.2) and FITA analysis, the data was pooled in the manner described in section 4.3.2.1. The concordance index of the confusions matrices between all the subjects for each condition was taken, while the pooling of data was done by using only the concordance index data for the experimental results in the -10 dB SNR condition. These pooling outcomes were used for all other listening conditions. The pooled confusion matrices were used as input to both FITA and MDS analysis. FITA analysis was done to analyse the overall importance of F1, F2 and duration in different listening conditions and for several spectral manipulations. The results are presented as bar plots in Figure 5.37 to Figure 5.39.

Figure 5.37 depicts the FITA results for the vowels without any formant suppression. Comparing the male and female voice results for the formants-only vowels, the main difference is seen at -10 dB SNR where similar but poor information transmission is seen for F1 and F2 (31% and 29%) for the male voice, while 57% for F1 and 38% for F2 is seen for the female voice. At -5 dB SNR for both voices, F1 shows the best information transmission, followed by F2 and duration. Overall for the two voices, the effect of noise on the three cues can be described by a decrease in the percentage information that is transferred, with the main drop in cue transmission when the SNR is lowered from 0 dB to

-5 dB.

For the whole-spectrum vowels, the results for the male and female voice seem similar. At -10 dB SNR, 55% and 64% transmission is seen for F1 and F2 respectively for the male voice, while 70% for both formants is found for the female voice. For the male voice, F2 is transferred slightly better than F2 for the three noise conditions, while the F1 and F2 transmission appears almost identical for the female voice.

A comparison between the percentage information transmission between the formants-only and whole-spectrum vowels reveals that at -10 dB SNR, F1 and F2 as well as duration are perceived more successfully by listeners for the whole-spectrum vowels than for the formants-only vowels. For the formants-only vowels, F1 and F2 transmission start to decrease at -5 dB SNR, while for the whole-spectrum vowels, a large decrease in transmission for the two cues is only found at -10 dB SNR.

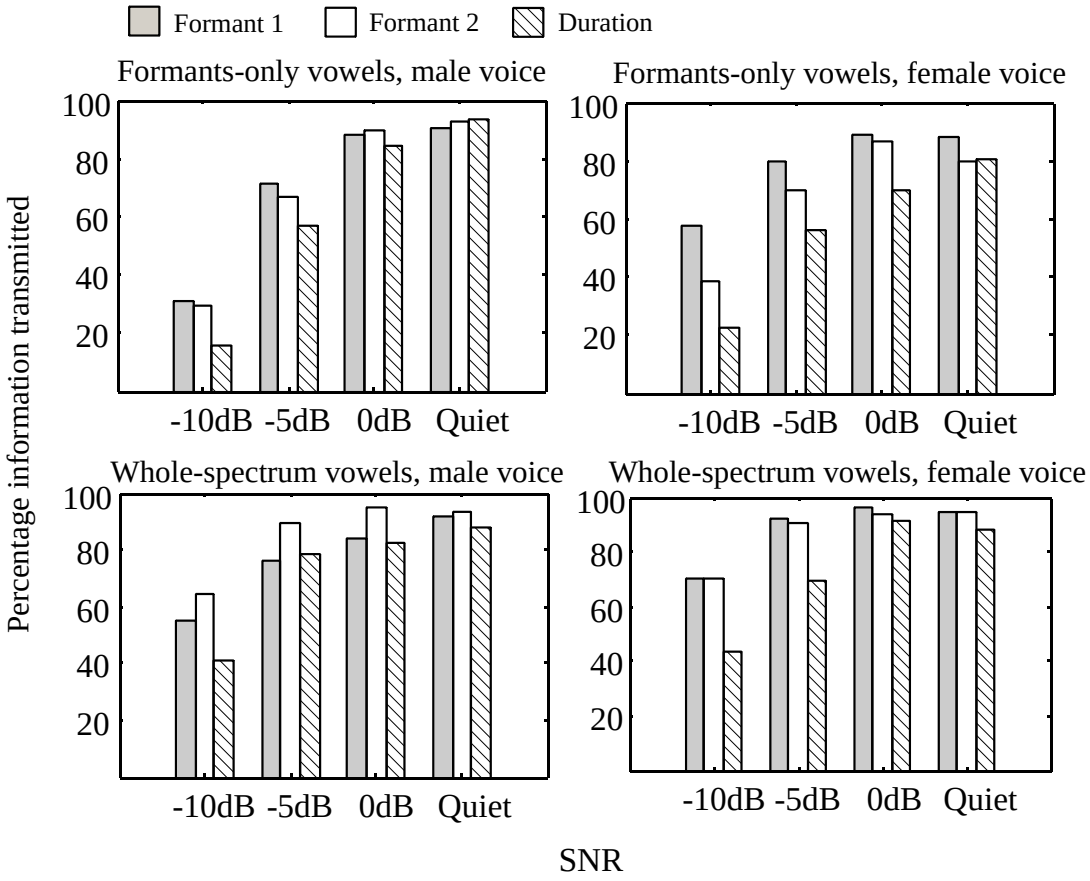

Figure 5.37. FITA analysis for the vowels with no formant suppression.

Figure 5.38 depicts the FITA results for the F1-suppressed vowels. Apparent from the results is the higher information transmission of F2 compared to F1 for all conditions, except for the -10 dB SNR condition for the formants-only male voice vowels, where transmission of the three cues is equally poor. For the formants-only vowels it appears that, for only the male voice, duration has a higher information transmission than F1 and F2 in quiet, 0 dB and -5 dB SNRs. Comparing these results to the results from the complete spectrum vowels in Figure 5.37, the absence of F1 resulted in a decrease in information transmitted for the F1, F2 and duration cues.

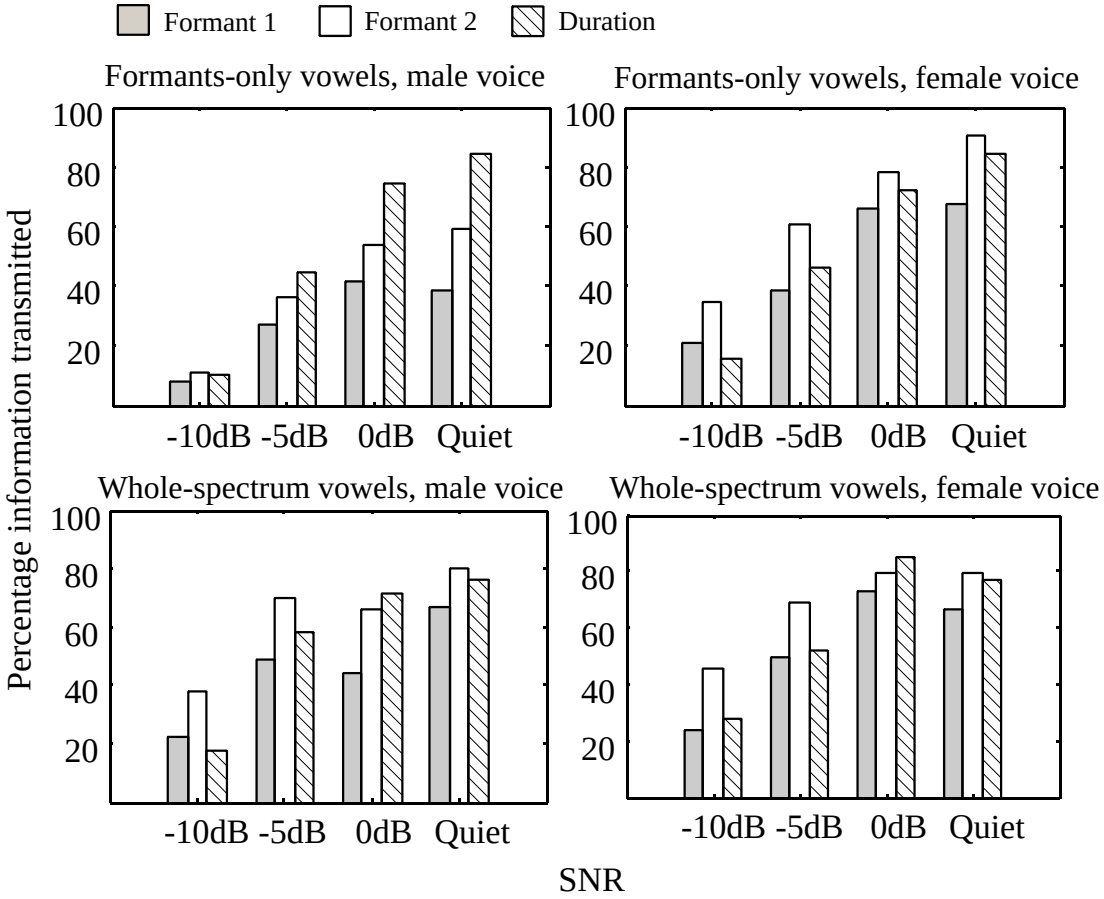

Figure 5.38. FITA analysis for the vowels with F1 suppressed.

Figure 5.39 depicts the FITA results for the F2-suppressed vowels. It can be seen that the whole-spectrum vowels for the male voice are least affected by the suppression of F2, since the transmission of F2 is lower than F1 for SNRs above -10 dB. In severe noise at -10 dB SNR, the highest percentage transmission is found for the female voice whole-spectrum vowels. The duration cue in quiet is transmitted better for the whole-spectrum vowels than the formants-only vowels (91% and 75% compared to 52% and 14%). Comparing the FITA analysis to Figure 5.37 where no formants were suppressed, it becomes apparent that a greater decrease in F2 transmission occurred for the formants-only vowels than the whole-spectrum vowels when F2 is suppressed. The information transmission of F2 is higher at all conditions for the whole-spectrum vowels compared to the formants-only vowels.

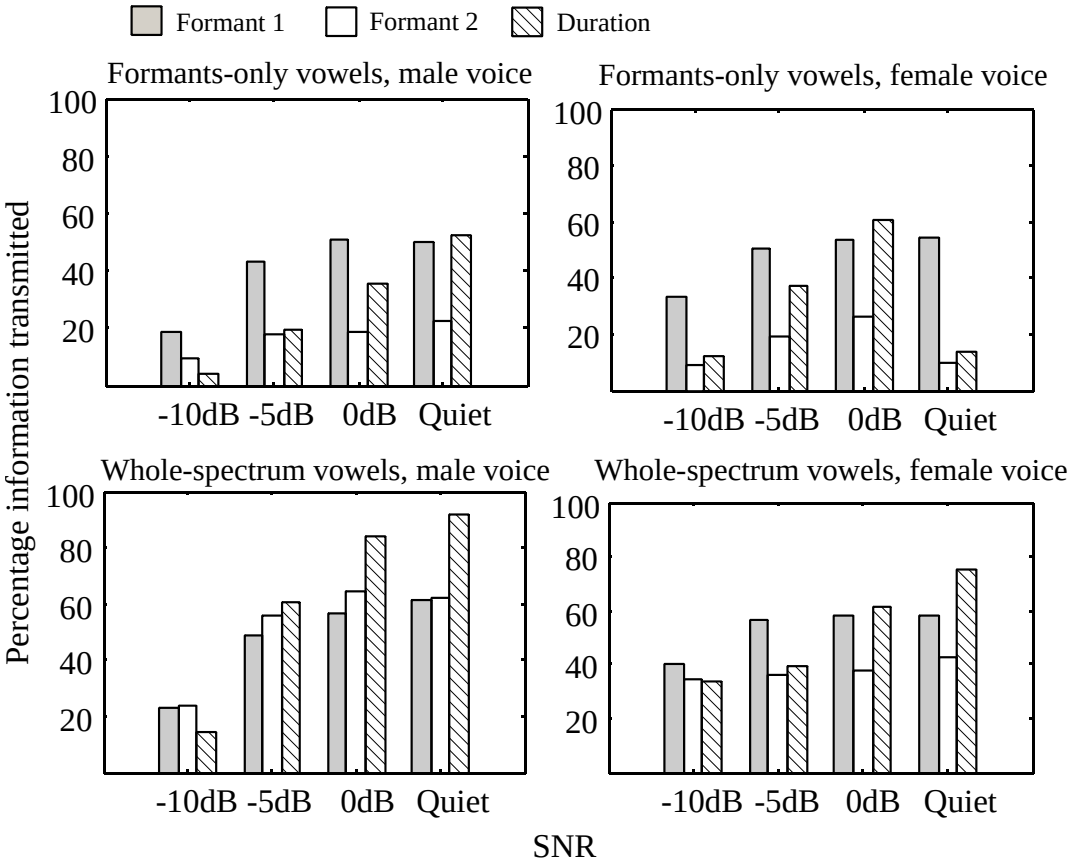

Figure 5.39. FITA analysis for the vowels with F2 suppressed.

5.4.5 Multidimensional scaling

In Chapter 3 the method by which INDSCAL MDS analysis was performed is explained. The results are given in this section for both criterion 1 and 2. Prior to MDS analysis, the vowel cues are extracted from the vowels for correlation with the MDS results. The F1, F2 and duration values as well as the dB SPL values for each spectral band are given in Table 5.1 and Table 5.2 respectively.

5.4.5.1 MDS analysis results – criterion 1

The results of INDSCAL MDS analysis for three- and five-dimensions were fitted correspondingly to the vowel cue results of F1, F2 and duration (hereafter only referred to

as the formants fit), and the dB SPL values for the five spectral bands of each vowel. Fitting was done by normalizing both the MDS and cue values to a number between 0 and 1, after which the summed least square error was obtained for each experimental condition. Figure 5.40 to Figure 5.42 depict the results for this fitting. The formants-only type vowels are fitted to the formant space, while the whole-spectrum type vowels are fitted to both the formants and the spectral band space.

In Figure 5.40, results show that a fit between the five-dimensional MDS space and the five spectral bands representation for the whole-spectrum vowels seems to be weaker than a three-dimensional fit between the MDS space and the formant cue space for both the formants-only and whole-spectrum vowels. Only for the female speaker in the quiet condition does the least square error for the spectral bands come close to a formants fit. It does however appear that the fit for the spectral bands improves as the SNR level decreases. For both the speakers, the formant vowel space fit does not differ substantially between the whole-spectrum and formants-only synthesized vowels. In the quiet condition for both voices, the formants-only vowels show slightly better fits to the formants space than the whole-spectrum vowels. In the -5 dB SNR condition however, a better fit of formants is seen for the whole-spectrum vowels than the formants-only vowels. This is also the case at -10 dB SNR for the female voice results.

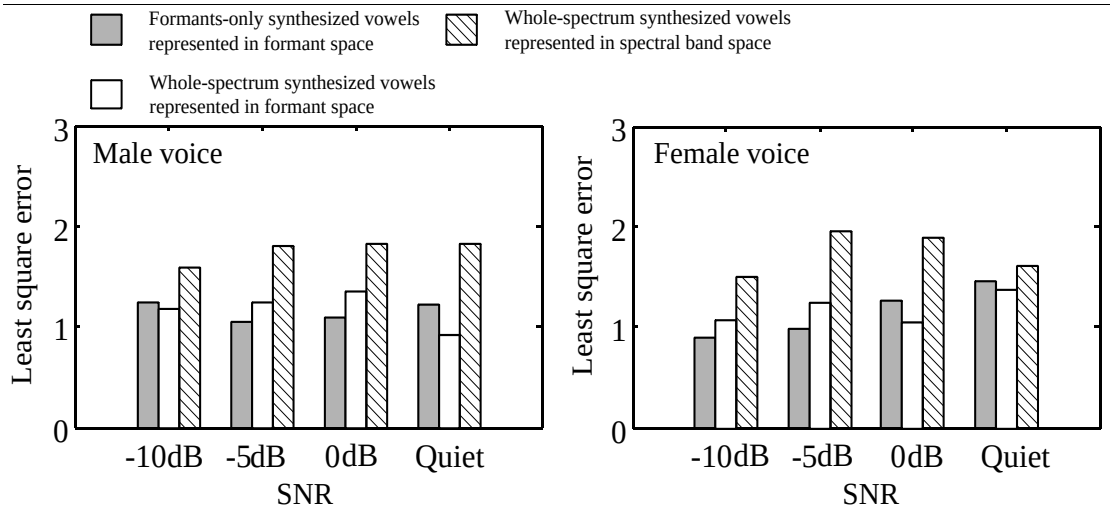

Figure 5.40. Fitting of MDS results and vowel cues for the complete spectrum vowels, male voice (left) and female voice (right).

Figure 5.41 shows the least square error values for the F1-suppressed vowels. The overall observation is that a spectral bands fit is much more similar to the formants fit of both the formants-only and whole-spectrum vowels than was the case for the complete spectrum vowels. Only for the male voice results do the spectral bands fit seem weaker than the formants fit (0 dB SNR and quiet condition).

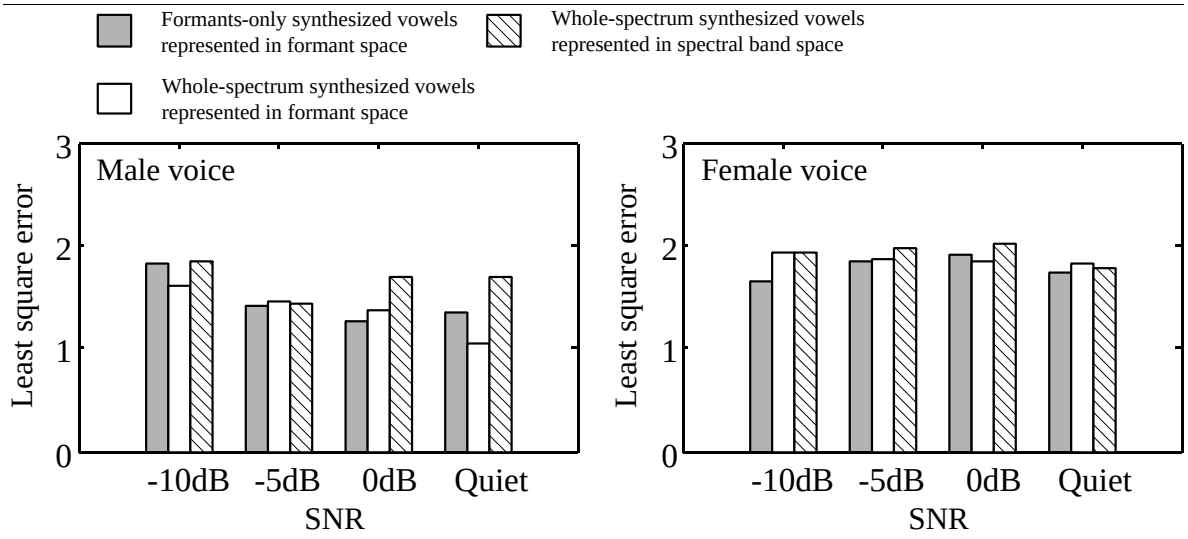

Figure 5.41. Fitting of MDS results and vowel cues for the F1-suppressed vowels, male voice (left) and female voice (right).

The least square errors in Figure 5.42 for the F2-suppressed vowels are similar to the results for the complete spectrum vowels where the spectral bands fit is weaker than the formants fit of the formants-only and whole-spectrum vowels for all experimental conditions.

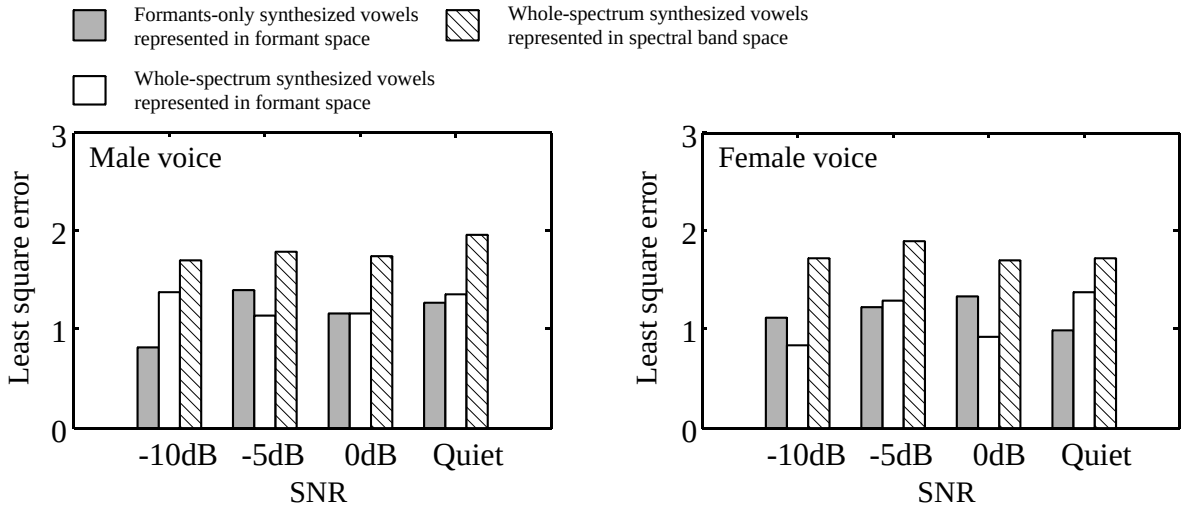

Figure 5.42. Fitting of MDS results and vowel cues for the F2-suppressed vowels, male voice (left) and female voice (right).

5.4.5.2 MDS results - criterion 2

Criterion 2 implies that the specific cues that fitted the respective MDS dimensions best are correlated with the corresponding MDS dimensions to obtain an idea of the similarity of the cue and MDS data. The correlation of the MDS and formant and duration data is given in Figure 5.43 to Figure 5.45 for the formants-only and whole-spectrum vowels, while Figure 5.46 to Figure 5.48 depict the correlation values between MDS and the spectral bands data (whole-spectrum vowels). The variance accounted for (r^2 or VAF) by the MDS results for each dimension is also given. Any specific VAF value depicts the accumulated VAF by the specific dimension and the lower dimensions. Therefore, the VAF value at Dimension 2 depicts the variance that the MDS data accounts for by both dimensions 1 and 2.

In Figure 5.43 for the male voiced formants-only vowels, the best correlations for both F1 and F2 are found at -5 dB and -10 dB SNR. Correlations for F1 and F2 are fairly equal for all conditions and increase rapidly from the 0 dB SNR to the -5 dB SNR. The correlation for duration increases with the addition of noise at 0 dB but decrease again as noise becomes more severe. For the female voiced formants-only vowels, the best correlation for F1 is found at -10 dB SNR while F2 obtained the best correlation at -5 dB SNR. The correlation for duration decrease from quiet to -5 dB SNR and improved slightly for -10 dB SNR. Overall for the formants-only vowels, F1 appears to be the most important cue for both the male and female voice in severe noise as it fits dimension 1 with a good correlation coefficient. F2 seems to be the secondary cue in severe noise with correlations of 0.82 and 0.74 and a contributing VAF value of 0.24 and 0.15 for the male and female voice respectively. For both the two voices, the three VAF values increase as the SNR decreases, showing that with the addition of noise, fewer dimensions are needed to explain the confusion data compared to the quiet condition.

In the left bottom plot of Figure 5.43, the formant and duration correlations for the whole-spectrum vowels (male voice) is shown. The highest correlation is obtained at -10 dB SNR for F1 (0.80). F1 is fitted with dimension 1 for all noise conditions, showing its

importance for the whole-spectrum vowels. The correlation for F2 stays equal for the three noise conditions. The correlation for duration decreases as noise is added and is the lowest at -10 dB SNR (0.46). For all noise conditions, the VAF value is larger than for the quiet condition. The bottom right bar plots of Figure 5.43 show the correlation values for the whole-spectrum female voice vowels. Both F1 and F2 show good correlation in severe noise. F2 is fitted to dimension 1 (with a contributing VAF of 0.42) and F1 to dimension 2 (with a contributing VAF of 0.22) at -10 dB SNR. Overall for the whole-spectrum vowels it seems that F1 plays an important role in severe noise, while F2 is also important, especially for the female voice results. An overall decrease in correlation for duration can be seen with a decrease in SNR for both voices. By taking the VAF measure of 0.6 as a benchmark for an accurate representation of the input data (Wickelmaier, 2003), one can conclude that between three and four dimensions are needed to explain the experimental results in the quiet and 0 dB SNR conditions. Three dimensions are needed for the -5 dB SNR condition while two dimensions would explain the results for the -10 dB SNR condition.

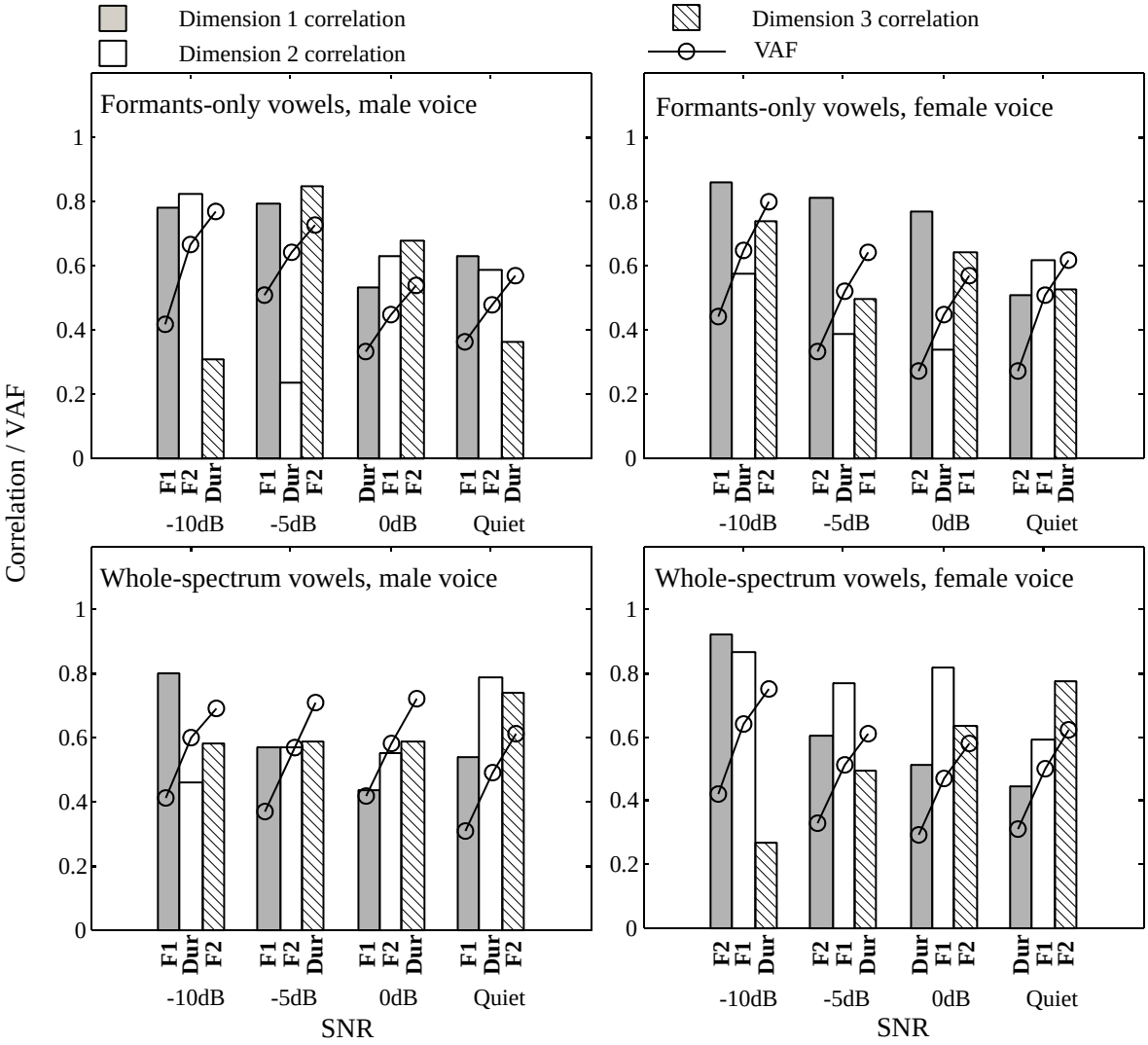

Figure 5.43. Complete spectrum vowels: Correlation and VAF values for the formants vowel space.

Figure 5.44 shows the MDS correlation results for the F1-suppressed vowels. A noticeable decrease in correlation for F1 can be seen when compared to its correlation values in Figure 5.43 (vowels with no formant suppression). For all conditions, F2 is allocated to dimension 1, except for the formants-only male voice results at 0 dB SNR and in quiet. In severe noise at -10 dB SNR, all four graphs indicate a fit of F2 with dimension 1, duration with dimension 2 and F1 with dimension 3. Similarly, a high VAF value (VAF > 0.6) at -10 dB SNR for dimension 1 is seen in all four figures. Consequently in the absence of F1,

for both formants-only and whole-spectrum vowels, listeners rely almost solely on F2 for vowel recognition in severe noise. An increase in the VAF values can again be seen as noise becomes more severe. Using the VAF benchmark of 0.6, two dimensions seem adequate to represent the formants-only vowel data for the quiet, 0 dB and -5 dB SNR conditions. In severe noise at -10 dB SNR, only one dimension would be sufficient. For the whole-spectrum vowels, three dimensions are needed for the quiet, 0 dB and -5 dB SNR, while one dimension at -10 dB SNR would explain the confusion data.

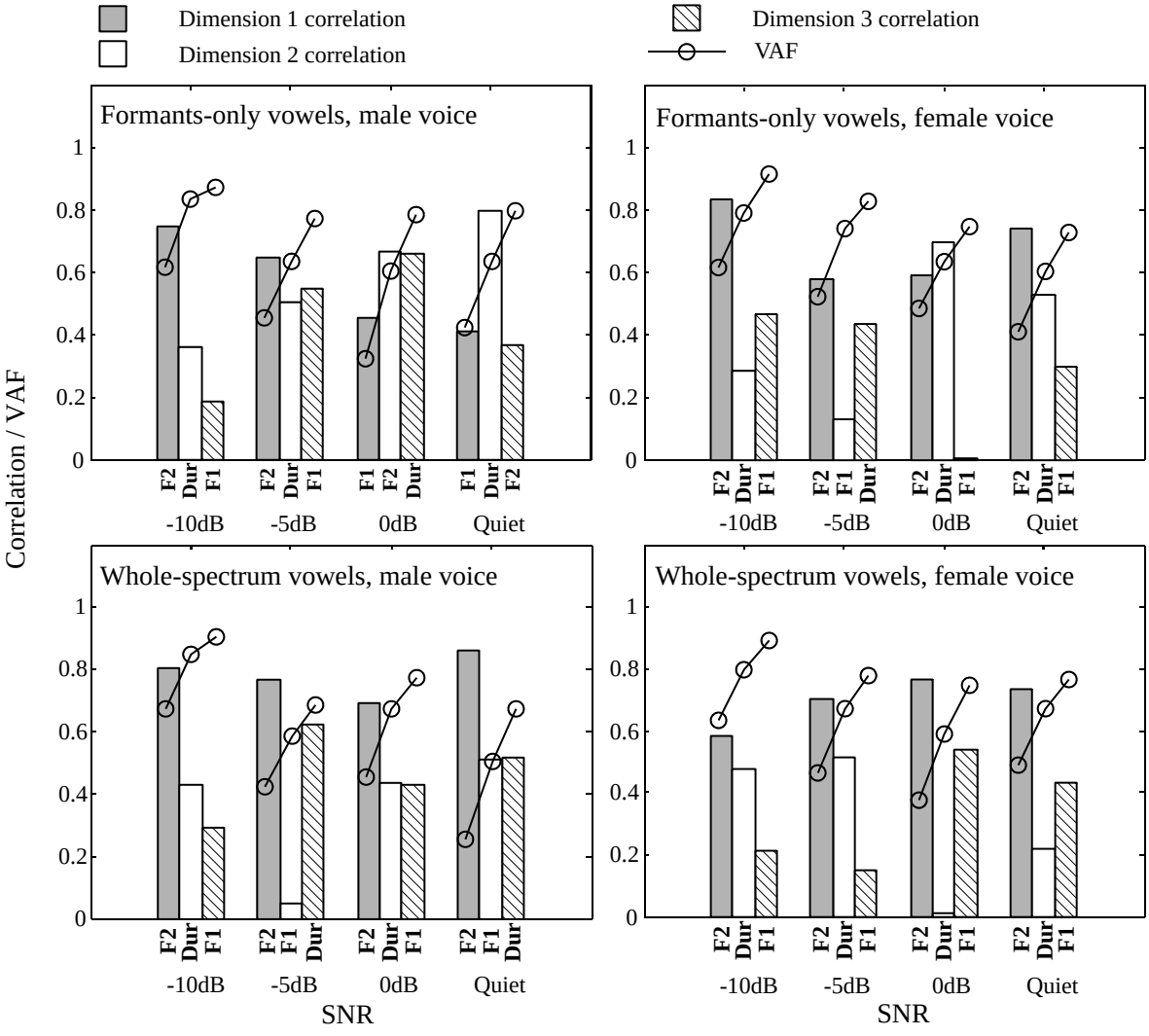

Figure 5.44. F1-suppressed vowels: Correlation and VAF values for the formants vowel space.

Figure 5.45 depicts the correlation results for the F2-suppressed vowels. When comparing the formants-only correlations with those of the whole-spectrum vowels, it becomes apparent that the suppressed F2 cue still plays a role in the whole-spectrum vowels, but not for the formants-only vowels at 0 dB and -5 dB SNRs. This can be observed for the male voice results, where for the formants-only vowels, F2 is fitted best to the second and third dimension, compared to the whole-spectrum vowels where F2 is fitted to the first and second dimension of the two noise conditions. For the formants-only female voice results, F2 shows very low correlation with the third dimension at -5 dB and 0 dB SNR, while for the whole-spectrum counterpart, F2 is fitted to the second dimension, with a correlation of 0.6 at 0 dB SNR. F1 shows better correlations for the formants-only vowels than for the whole-spectrum vowels at SNRs of -5 dB, 0 dB and in quiet. At -10 dB SNR, equal correlation for F1 is seen for the two vowel synthesis types. When taking the VAF values into consideration, the formants-only vowel confusion results can be explained by less dimensions (only F1), while for the whole-spectrum vowels the confusion results can be explained by two or three dimensions.

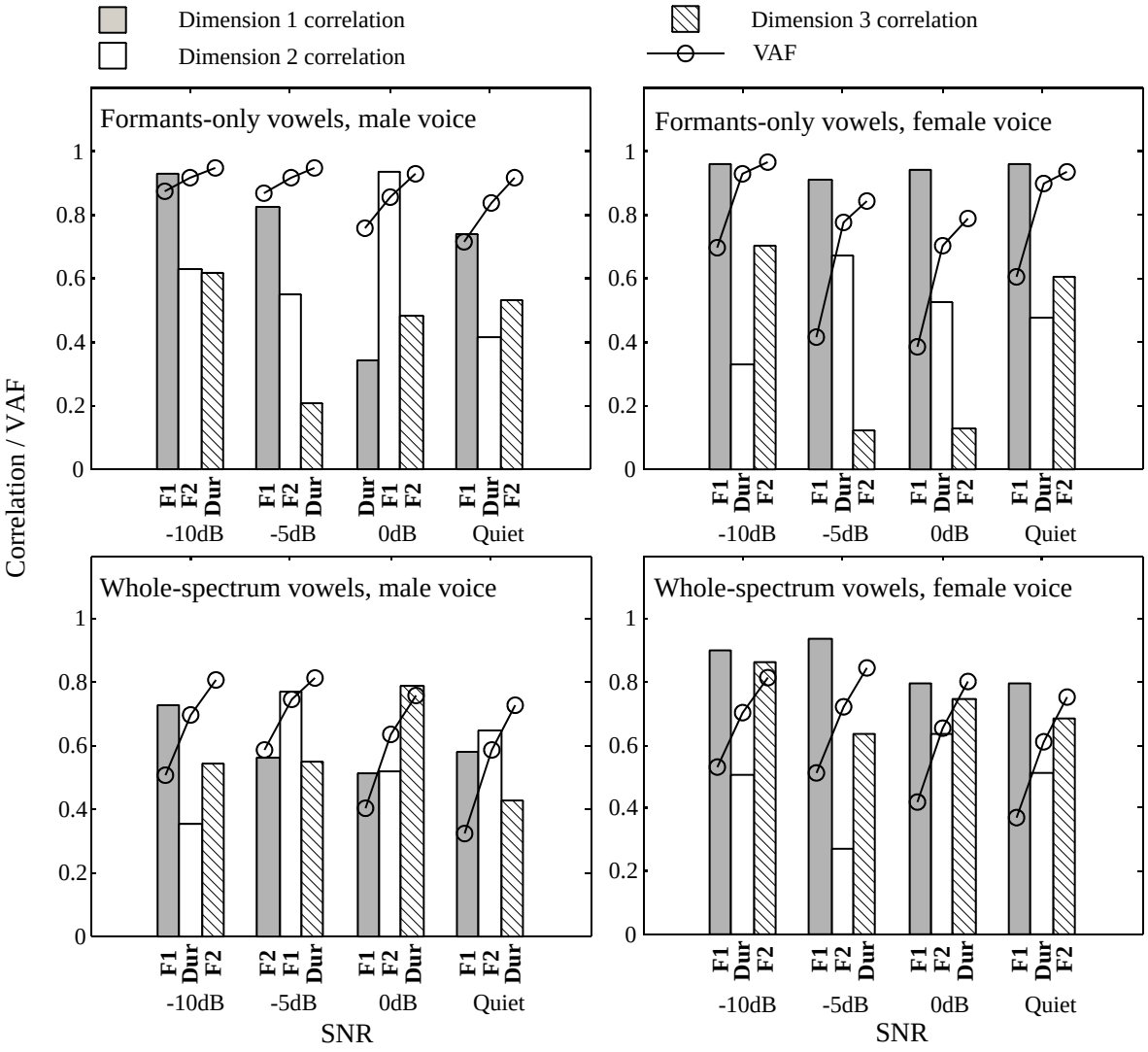

Figure 5.45. F2-suppressed vowels: Correlation and VAF values for the formants vowel space.

The results in Figure 5.46 (top) depict the correlation of the spectral bands with the five-dimensional MDS results for the male voice. Band 3 is fitted to dimension 1 for all the noise conditions and increases its correlation with a decrease in SNR. Band 5 shows the second best correlations and shows good correlation throughout all conditions. Band 1 and band 4 do not show good correlations at any listening condition.

In Figure 5.46 (bottom, whole-spectrum vowels for the female voice), band 3 shows the best correlation at -10 dB SNR while fitted to dimension 2. Band 5 shows good correlation values in quiet and -10 dB SNR, while band 1 correlates well at all conditions except at -10 dB SNR. Overall for the two voices, band 3 and band 5 show good correlations at -10 dB SNR, while band 2 appears also to be important.

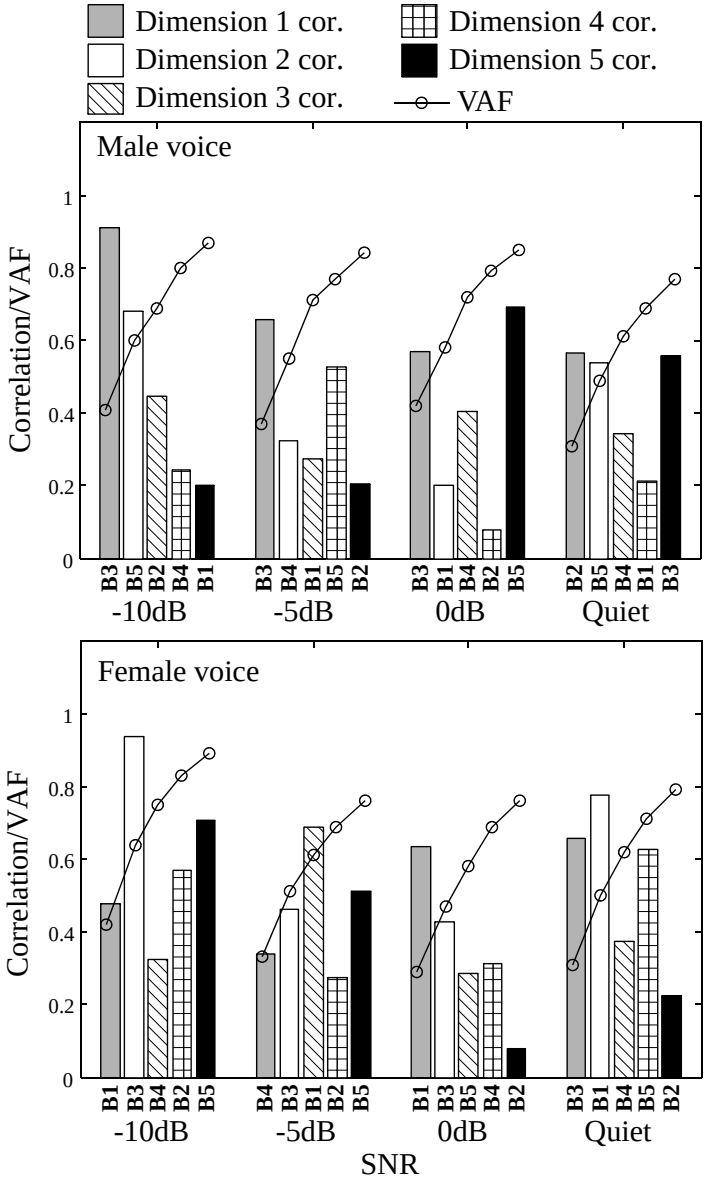

Figure 5.46. Complete spectrum vowels: Correlation and VAF values for the spectral band vowel space for the male and female voice vowels.

Figure 5.47 (top) shows the spectral bands correlations for the F1-suppressed vowels, male voice. Band 3 shows good correlation for the complete spectrum vowels (fitting to dimension 1 for all noise conditions and a correlation coefficient of 0.9 at -10 dB SNR). No such correlation is found for the F1-suppressed vowels. Band 3 is only fitted to dimension 1 at -10 dB SNR, but with a correlation of only 0.64. Band 5 seems to be the spectral band with the best correlation in quiet and 0 dB SNR (with correlation of 0.79 and 0.55 respectively, fitting to dimension 1), while also fitting to dimension 2 at -5 dB and -10 dB with correlation coefficient of 0.45 and 0.54 respectively.

Figure 5.47 (bottom) depicts the spectral bands correlations for the F1-suppressed vowels of the female voice. A noticeable difference between the F1-suppressed and complete-spectrum vowels is the correlation of band 1. In Figure 5.46, band 1 shows good correlation throughout the SNR levels, especially at -10 dB and 0 dB, where it was fitted to dimension 1. In the F1-suppressed vowel results, band 1 does not show any significant correlations, except at -10 dB SNR. Band 5 and band 3 show the best correlation values at -10 dB and -5 dB SNR compared to the other bands, but with low correlation values.

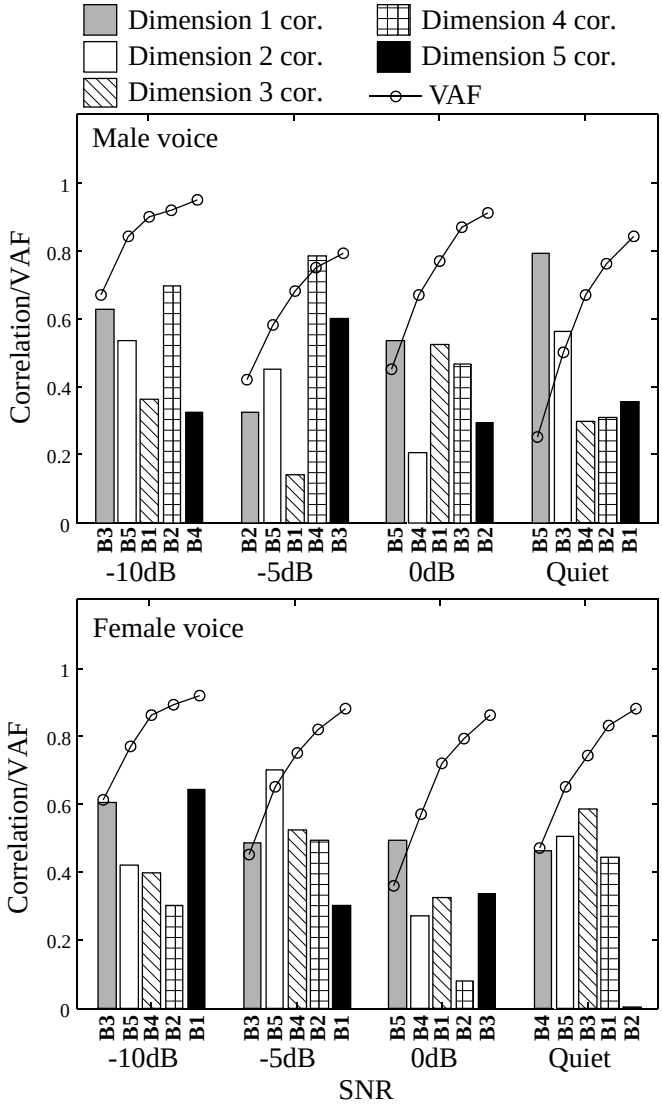

Figure 5.47. F1-suppressed vowels: Correlation and VAF values for the spectral band vowel space for the male and female voice vowels.

Figure 5.48 (top) shows the male voice results for the F2-suppressed vowels. Band 5 is fitted to dimension 1 for the quiet and -5 dB SNR condition, and to dimension 2 for the -10 dB SNR condition. The best overall correlation is seen for band 3 at -10 dB SNR, where it is fitted to dimension 1. Band 3 and band 5 are found to provide the two best correlations at -10 dB in the complete spectrum vowels (fitted to dimensions 1 and 2), and a similar fitting is found for the F2-suppressed vowels at this SNR, but only with smaller correlation

values.

Figure 5.48 (bottom) shows the correlation values for the female voice F2-suppressed vowels. Band 1 is fitted to dimension 1 for all conditions and has the best overall correlation at -10 dB and 0 dB SNR. This is similar to the results of the complete spectrum correlations where band 1 is fitted to dimension 1 at -10 dB and 0 dB SNRs. Correlations exceeding 0.7 can also be seen for band 3 at -5 dB, 0 dB and quiet conditions, although only fitted to dimension 2 or 3.

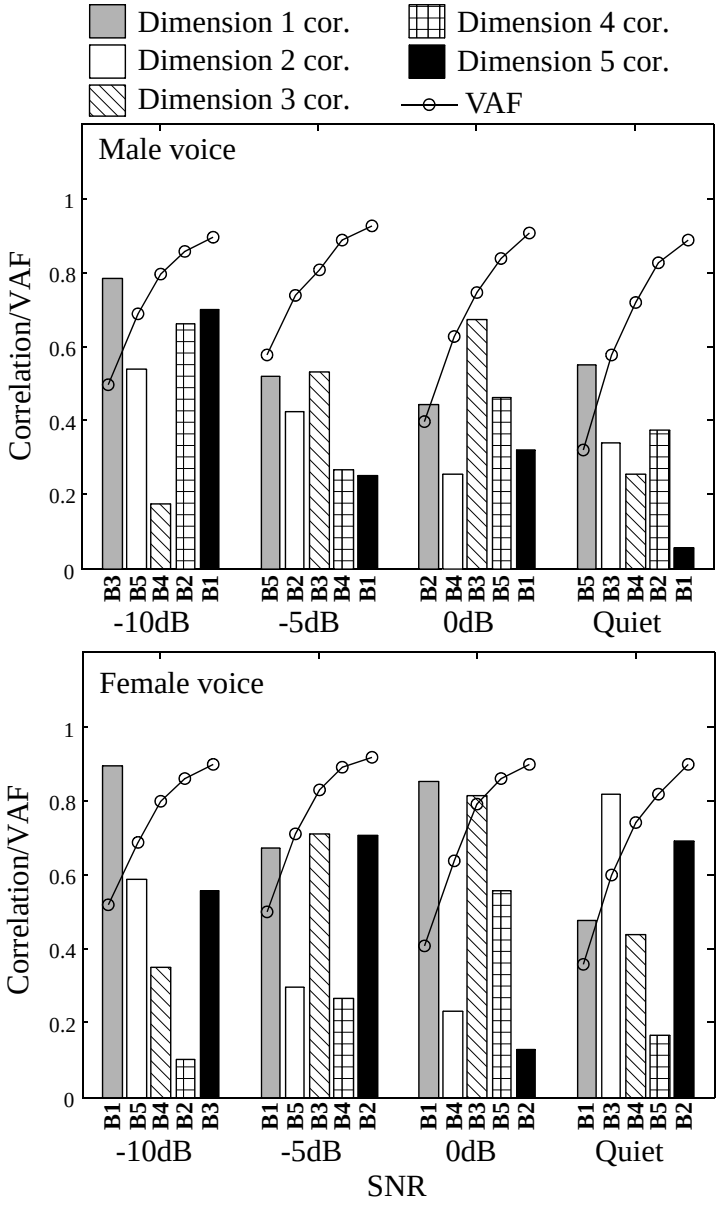

Figure 5.48. F2-suppressed vowels: Correlation and VAF values for the spectral band vowel space for the male and female voice vowels.

5.5 SUMMARY

In this chapter, results were presented for the vowel synthesis methods, as well as experimental listening tests done with synthesized vowels. Plots were analyzed to investigate the spectra for the whole-spectrum and formants-only synthesized vowels. Two vowel spaces were examined to predict vowel confusions according to the F1, F2 and duration space, as well as a space consisting of five spectral bands for each vowel. The influence of speech-shaped noise on these spatial representations was also considered. Experimental listening test results were analyzed using four analysis techniques namely percentage correct scores, confusion matrix analysis, FITA and MDS analysis. The first technique gave insight into the cues that listeners use to successfully recognize vowels, whilst the last three techniques gave results about the specific cues that are transferred to listeners and that may subsequently be used to predict vowel confusions in noise.

CHAPTER 6 DISCUSSION

6.1 CHAPTER OBJECTIVES

In this chapter the experimental outcomes of the present study are discussed. The findings regarding the importance of detailed spectral shape versus formants in severe noise are analysed, after which the overall importance of F1 versus F2 is discussed. The difference in relevance of F1 and F2 for the two types of synthetic vowels (formants-only and whole-spectrum) is also analysed, after which the findings are compared to literature. Attention is given to the importance of the specific cues, particularly pertaining to the way it leads to the recognition of vowels and the availability of these cues in noise.

6.2 FORMANTS AND SPECTRAL SHAPE

The importance of the vowel whole-spectrum and vowel formants were explored in severe noise by means of vowel perception tasks. The aim was to determine if F1 and F2 still play a role in severe noise to the same extent as in quiet conditions.

The effect of noise on the whole-spectrum and formants-only synthetic vowels were examined prior to analysis of the experimental listening test results. The main effect of noise was found to be a distortion of the original spectral shape for the whole-spectrum vowels, and a decrease in spectral contrast for the formants-only vowels. Despite the effect of noise on spectral detail, inspection of noise-corrupted vowel spectra showed that formant peaks (though with low spectral contrast) were still visible in both the whole-spectrum and formants-only spectra. This suggested that formant peaks could possibly play a role in both the recognition of the whole-spectrum and formants-only vowels.

6.2.1 Formants versus whole-spectrum as cues to recognize vowels

Overall, the main effect of noise only had a significant influence on vowel recognition scores at -5 dB and -10 dB SNR. In severe noise at -10 dB SNR, the identification of the whole-spectrum vowels were significantly better than the formants-only vowels, which can be attributed to additional spectral information that is absent for the formants-only vowels. However, the recognition of the formants-only vowels at 0 dB SNR and in quiet was better than the whole-spectrum vowels, while no significant differences in vowel correct scores could be found at -5 dB SNR. The better recognition of the whole-spectrum vowels was only due to the recognition scores of seven vowels. Spectral detail therefore appears to have a greater effect on vowel recognition than formants for the majority of vowels in severe noise, but exceptions exist depending on the specific individual vowels.

6.2.2 Formants versus whole-spectrum information perceived by listeners

Although the percentage correct results indicated better recognition of the majority of vowels in severe noise when it contained detailed spectral information, it did not specify the information that was actually carried over to the listener. The relationship between the two vowel spaces (five spectral band space and F1, F2 and duration space) and the vowel confusion data was explored for further analysis of formants and whole-spectrum perception in noise.

MDS analysis showed that, with the addition of noise, the vowel perception results could be explained by fewer dimensions than in quiet. For both male and female voice vowels of the complete spectrum vowels, between three to four dimensions were needed for the quiet to -5 dB SNR conditions, while two dimensions were required for the -10 dB SNR condition. The addition of noise therefore decreased the number of cues that were relevant in noise. Formants and duration were found to be the best representation of the vowel confusion data compared to a whole-spectrum vowel space at all SNRs. This was found through the calculation of the least square errors in terms of the Euclidian distances between the MDS and cue data. However, with a decrease in the SNR the whole-spectrum

cues play a growing role, as the whole-spectrum fit to the MDS data improved with added noise.

The specific correlations between the vowel cues and the MDS results revealed that both F1 and F2 explained vowel perception in severe noise, while no clear evidence was found that listeners perceived the duration cue. Although the whole-spectrum cues of band 3 and band 5 showed correlation values exceeding 0.6, it generally did not match those of the F1 and F2 correlations, showing that the whole-spectrum vowel space was not as accurate in explaining vowel perception as the formant vowel space.

Analysis of confusion matrices revealed information regarding the confusion of individual vowels and the possible linking of these confusions to either the formants or whole-spectrum vowel space. Substantially more confusions were observed between vowels without formant suppression in severe noise at -10 dB SNR compared with the higher SNR conditions. The majority of confusions were due to either a similar F1 or F2 value, or due to proximity in the spectral bands space, while stimuli were confused with vowels sharing similar duration values to a lesser extent than the other cues. Vowels containing similar F1, F2 or both formants, and situated in close proximity in the spectral band space, were found to be confused with each other, showing that whole-spectrum information was also used together with formants in making vowel judgments.

It is of interest to note the confusions of /ʏː/ with /ɛː/ (20% whole-spectrum vowels, 1% formants-only vowels), /ɵ/ with /ɛː/ (13% whole-spectrum vowels, 3% formants-only vowels), and /ɵ/ with /ɛ/ (11% whole-spectrum vowels, 0% formants-only vowels) in severe noise. For these three confusions, the vowel pairs are grouped close together in the spectral bands space, while it also has F1 and F2 cues in common. Since the formants-only vowels did not show the same confusion scores than the whole-spectrum vowels, it can be assumed that listeners rather perceived information regarding the whole-spectrum in severe noise, than information regarding formants.

The confusions of /yː/ with /iː/ (14% whole-spectrum vowels, 10% formants-only vowels), /ɛː/ with /e/ (19% whole-spectrum vowels, 13% formants-only vowels), /œ/ with /ə/ (29% whole-spectrum vowels, 31% formants-only vowels) and /ə/ with /œ/ (28% whole-spectrum vowels, 21% formants-only vowels) in severe noise are also of interest. These vowel pairs are also closely spaced in the spectral band vowel space and share similar F1 and F2 cues, but this time the formants-only vowel confusion scores are similar to the whole-spectrum confusion scores. These confusions show that formants were also extracted from the whole-spectrum vowels, while spectral-shape cues were either redundant, or not available in noise.

FITA analysis of the listening results for vowels with no formant suppression revealed that, in severe noise at -10 dB SNR, the F1 and F2 cues for both voice types of the whole-spectrum vowels were better transmitted to the listener than for the formants-only vowels. The presence of spectral detail in the vowel spectrum is therefore supportive in providing listeners access to formants in severe noise conditions.

6.3 IMPORTANCE OF F1 AND F2 IN SEVERE NOISE

The previous section dealt with the influence of formants compared to the detailed spectrum on vowel perception. Another matter that was investigated in this study is the importance of either F1 or F2 for vowel perception in noise. For both the types of synthetic vowels, F1 and F2 were alternatively suppressed for all noise conditions and compared to vowel recognition results when no formants were suppressed.

6.3.1 Importance of F1 and F2 in the recognition of vowels

When assessing the results of the male and female voice separately for the formants-only vowels, the suppression of F1 and F2 led to a significant reduction in vowel recognition for all conditions. The absence of F2 had a greater influence on vowel recognition than the absence of F1 in quiet and moderate noise conditions, but in severe noise at -10 dB SNR,

the lack of both formants had the same influence on vowel recognition. Although the analysis of the spectrum at -10 dB SNR showed that F1 was less affected by noise than the other formants, the suppression of either F1 or F2 had the same significant effect on vowel recognition. This suggests that in severe noise, the combination of F1 and F2 is critical for a listener when making vowel decisions, and that neither F1 nor F2 is redundant in these circumstances.

Considering the results of the individual vowels for both the whole-spectrum and formants-only vowels, it was found that F1 and F2 played more important roles for central vowels (containing substantial frequency differences between F1, F2 and F3) in less noisy conditions, compared to vowels classified as back or front vowels. Listeners therefore use nearby formants to still perceive a vowel in the absence of F1 or F2. If the formant that is suppressed is not situated close to another formant, listeners find it hard to recognize the vowel. In severe noise at -10 dB SNR, F2 plays a more important role in back and central vowels, as its absence is not perceived for front vowels.

6.3.1.1 F1 and F2 cues perceived by listeners

A comparison between the influence of F1 and F2 on vowel confusions was essential to predict vowel confusions in noise. FITA analysis for the formants-only vowels for both male and female voices in the -5 dB SNR condition showed that F1 was transferred to the listener more effectively than F2. At -10 dB SNR, FITA results showed that listeners used F1 cues more often than F2 cues when making vowel judgments for the female voice, while both formants played a role when vowels were confused for the male voice.

In the MDS analysis, two measures were used to determine the significance of each formant. The VAF value provided a means by which one can determine the least number of dimensions for which the vowel confusions are spatially represented, while the correlations of each cue with the MDS data depicted a measure of similarity of a specific cue and MDS dimension. F1 and F2 showed good correlations with the MDS data in noise

for the vowels without formant suppression, with F1 being the most important cue for the formants-only vowels, as it was fitted to dimension one. For the whole-spectrum vowels in severe noise, F2 appeared to be the most important cue for the female voice data, while F1 was most important for the male voice.

An analysis of the MDS results when F1 was suppressed for both the formants-only and whole-spectrum vowels showed that the F2 cue fitted to dimension one for the majority of the correlation fits. It was evident that at -10 dB SNR, the VAF for dimension one was the highest compared to the other SNRs for both synthesis type vowels and speakers, showing the increased importance of dimension one (correlated with F2) as the SNR was lowered. The same effect was seen for the formants-only F2-suppressed vowels, where F1 showed the best correlations compared to the other cues and generally correlated well with dimension one. These results therefore revealed that in noise, F1 was more likely to explain vowel confusions in the absence of F2, than F2 explaining these confusions when F1 was suppressed.

When FITA results for the formants-only vowels are compared for the complete spectrum vowels and formant-suppressed vowels in terms of the cues transferred to listeners, similar conclusions than for the MDS results can be made. Listeners made vowel judgments mainly on the basis of F1 information in severe noise when F2 was absent, while in the absence of F1 listeners did not base vowel confusions solely on the F2 cue.

The effect of F1 and F2 suppression on vowel confusions was also analyzed by inspection of vowel confusion matrices. In confusion matrices for the F1-suppressed vowels, the stimuli was arranged in the order of ascending F2 frequency values, while for the F2-suppressed vowels, stimuli was arranged according to the F1 frequency values. For the formant suppressed vowel confusion matrices at all noise conditions, confusions were aligned close to the diagonal, showing that vowels were mainly confused with other vowels sharing the same F1 cue for the F2-suppressed vowels, and sharing the same F2 cue for the F1-suppressed vowels. Less confusions near the diagonal were found for the F2-

suppressed vowels than for the F1-suppressed vowels. Listeners therefore had more difficulty in consistently perceiving the same vowels and were more uncertain of vowel decisions when F2 was absent.

6.3.1.2 Difference in F1 and F2 importance for formants-only and whole-spectrum vowels

The importance of the whole-spectrum versus formants, as well as the F1 versus F2 cues on vowel perception in noise was discussed. When the vowel spectrum consists of only formants or detailed spectral information, it is important to distinguish the difference in significance between F1 and F2. The significant interaction between spectral manipulation, vowel synthesis type and SNR are noteworthy in terms of the percentage correct scores, showing that the SNR and spectral manipulation has an effect on the difference between the whole-spectrum and formants-only vowel identification.

For both the formants-only and whole-spectrum vowels of both speakers, the suppression of either F1 or F2 led to a significant reduction in vowel recognition scores. On average in severe noise at -10 dB SNR, the whole-spectrum vowels were recognized significantly better than the formants-only vowels for both F1- and F2- suppressed vowels. This showed again that the whole-spectrum vowels contain additional spectral information that aids listeners in correctly perceiving vowels, since with the absence of formants still does not lead to the same recognition results than the formants-only vowels.

An important observation was the perception of F2 in severe noise for the F2-suppressed vowels. When comparing FITA analysis for the whole-spectrum and formants-only vowels in severe noise, it was seen that a larger decrease in F2 information occurred for the formants-only vowels than for the whole-spectrum vowels when F2 was suppressed, while no substantial differences was seen for the F1 information in the F1-suppressed vowels. Listeners therefore still perceived F2 information for the whole-spectrum vowels when F2 was suppressed, while for the formants-only vowels poor F2 information was perceived.

Analysis of the confusion matrices gave more insight into this. The following vowel confusions at -10 dB SNR for the F2-suppressed vowels were found: /yː/ with /ɛː/ (22% whole-spectrum vowels, 3% formants-only vowels), /yː/ with /ɛ/ (15% whole-spectrum vowels, 0% formants-only vowels) and /ɛ/ with /i/ (43% whole-spectrum vowels, 8% formants-only vowels). For all these confusions the vowel pairs contained similar F2 values even though F2 was suppressed. The vowel pairs are also situated close together in the spectral band space. Listeners therefore perceived whole-spectrum cues, which led them to derive F2 information from the F2-suppressed spectrum to confuse these vowel pairs.

6.4 OTHER INFLUENCES ON RESULTS

Information from the literature revealed influences other than formants and spectral shape to also play a role in vowel recognition. Some of these will be discussed in the context of the present study.

6.4.1 Effect of duration

FITA results showed that the duration cue was perceived more clearly in the whole-spectrum vowels than the formants-only vowels in severe noise. Noise eliminated duration cues mostly for longer-duration vowels containing only formant information, while the duration cues in shorter-duration vowels containing detailed spectral information were more robust in severe noise. It can be assumed that the whole-spectrum vowels provided enough information in noise to ensure that, when spectral cues were masked in one time window, listeners still had access to redundant information from another window.

6.4.2 Effect of speaker

Two speakers were used in the present study, one male and one female. Some of the results were presented by pooling the listening results for the two speakers together, while for

certain analysis methods, the male and female voice results were kept apart. Results showed that, although the utterances of the two speakers had similar formant distribution patterns, the female voice vowels were recognized significantly better than the male voice vowels for both the whole-spectrum and formants-only vowels in low SNR conditions. It is interesting to note that, for the female voice, the percentage correct scores increased as noise was added from quiet to 0 dB SNR. This was seen for the formants-only vowels (F2-suppressed) and for the whole-spectrum vowels (F1-suppressed).

Between the quiet and 0 dB SNR condition, FITA results showed an increase in F2 and duration information for the F2-suppressed formants-only female voice vowels, while an increase in F1 and duration was seen for the F1-suppressed whole-spectrum vowels. For the female voice vowels, the addition of noise at 0 dB SNR therefore made the specific suppressed cues more perceivable to the listener.

6.5 COMPARISON OF FINDINGS WITH LITERATURE

The purpose of the present study was to investigate the cues that are important for vowel recognition in severe noise. Vowels containing only formant information were synthesized using the DSS method, a technique also implemented by Hillenbrand et al. (2006), who presented DSS synthesized vowels to listeners in quiet. Results from the current study, when the DSS synthesized vowels were presented to listeners in quiet, matched the results from the study of Hillenbrand et al. (2006), where on average, listeners obtained a recognition score of 89%.

On the other hand, vowels containing detailed information of the vowel spectrum were synthesized using coefficients in the DCT of the vowel spectrum, described in the study of Zahorian and Jagharghi (1993). The application of the DCTCs in the current study differed from that of Zahorian and Jagharghi (1993), where automatic vowel classification was done based on features from the whole-spectrum. Spectral plots of the DCT spectrum from the study of Zahorian and Jagharghi (1993) did, however, compare with the DCT

spectrum plots in Chapter 4.

It is known that normal-hearing listeners only struggle with speech recognition in SNRs lower than 0 dB (Assmann and Summerfield, 2004). Results from the present study showed a similar effect of noise on vowel recognition, where listeners had little difficulty in recognizing vowels in quiet and 0 dB SNR, while a significant decrease in vowel recognition was found at -5 dB and -10 dB SNR.

Presenting vowels in noise at -5 dB SNR showed no significant difference between the vowel identification scores for the whole-spectrum and formants-only vowels, signifying that both types of synthetic vowels contained cues that were necessary for vowel recognition in noise. Knowing that the whole-spectrum vowels also contained formant information, it could be said that only the formant peaks were used at -5 dB SNR to recognize vowels. For the same noise condition at -5 dB SNR, Parikh and Loizou (2005) concluded that listeners do not exclusively utilize formant cues, nor exclusive spectral shape information to recognize vowels. The findings were based on the fact that F2 was found to be heavily masked, as well as the observation that vowels with different spectral shapes were still confused with each other. While Parikh and Loizou (2005) based their findings on the spectral analysis of the noisy vowels, the findings in the current study were made by knowing the cues that are present in the vowel spectrum of the undegraded vowels. Listeners were therefore able to extract vowel cues in severe noise, even when it was known by inspection that it were masked by severe noise.

Although the whole-spectrum provided better information in severe noise than only formants, MDS analysis gave indications that an F1, F2 and duration space fitted the experimental results better than a whole-spectrum space, but that the difference in fits decreased as the SNR was lowered. MDS and confusion data correlations revealed that only specific spectral band cues correlated with the MDS analysis at -10 dB SNR. F1 and F2, together with partial spectral detail therefore seem to explain vowel confusions. Sakayori et al. (2002) investigated formants and spectral shape combination cues. Even

though experimental tests were only done in quiet, it was found that listeners utilize the spectral regions at the locations of F1 and F2 for vowel recognition. Two phonetically different vowels containing similar F1 and F2 information were found to be easily distinguished, even when the vowels consisted of only spectral information at the F1 and F2 locations. It was argued that the human auditory system identifies formants, but uses spectral shape information in these regions to make a final decision regarding the vowel identity. It seems that this theory is more applicable to severe noise conditions.

In the formant discrimination experiments in noise, Liu and Kewley-Port (2004a) concluded that vowel identification depends on the ability to distinguish between differences in the spectrum, especially formant changes. Additional spectral shape information therefore provides listeners with the ability to better discriminate between the formants of two vowels, leading to less vowel confusions and an increase in vowel recognition.

Inspection of the stimuli in severe noise showed that the vowel spectra were distorted by noise, which confirmed information from the literature (Cooke et al., 2001; Gong, 1995; Nabelek et al., 1992). In quite and in noise at 0 dB SNR, F2 was found to be the most influential cue to listeners due to the suppression of F2, leading to greater recognition errors than the suppression of F1 for the formants-only vowels. Similar results were obtained in the study by Kasturi et al (2002), where spectral holes were created for each vowel, after which it was established that holes in the vicinity of F2 led to a greater decline in recognition scores than in the F1 region.

At a SNR of -10 dB for the formants-only vowels, the suppression of F1 and F2 led, on average, to the same significant reduction in recognition scores for both types of synthesized vowels. Although noise masked the vowel spectrum differently according to the intensity of the formants (Liu and Kewley-Port, 2004a; Nabelek et al., 1992), results showed an overall trend that the F1 and F2 peaks were of equal importance in severe noise for the formants-only vowels. By analysing the percentage correct scores for individual

vowels, it was found that the importance of formants depended on vowel backness. At -5 dB SNR the suppression of F1 or F2 showed the smallest effect on front and back vowels, while central vowels were affected the most. Confirmation of this is seen in the study of Hedrick and Nabelek (2004), where the F2 intensity of the back vowel /u/ was gradually reduced and presented to listeners in speech-shaped noise. Listening results revealed that listeners were unable to perceive the formant intensity change in noise.

In severe noise at -10 dB SNR, it was found that front vowels were least affected by the suppression of F2. However, no relationship between vowel backness and recognition scores was found for the F1-suppressed vowels. It is likely that the centre of gravity theory (Chistovich and Lublinskaya, 1979) is relevant in severe noise, which states that the auditory system will average two closely spaced formants to one broad energy concentration. The absence of one formant is perceived as a loss of partial information regarding the broad spectral peak, but would not be enough to change vowel perception.

Identification of vowel utterances from the female speaker was found to be significantly better than for the male speaker in severe noise. In the study of Marguiles (1979) it was established that the speech of female speakers are significantly more intelligible than male speakers at a SNR level of 0 dB SNR. For the same SNR, Nabelek et al. (1992) found that vowel recognition is significantly influenced by speaker differences.

6.6 GENERAL DISCUSSION

The main research question stipulated in Chapter 1 was whether listeners still use formants to perceive vowels in severe noise. The findings at -10 dB SNR differed from the results at -5 dB SNR, where it was found that vowels containing detailed information on the spectral shape were recognized significantly better than vowels consisting only of formant information. In the listening tests of natural vowels in noise, Parikh and Loizou (2005) found that certain vowels containing similar F1 frequencies were not confused with each

other, although F2 was found to be masked by noise. It was suggested that, in addition to F1 information, listeners possibly still obtained information regarding F2 by means of spectral shape cues in the F1 and F2 regions. Despite the fact that formant cues alone have been found to be adequate for vowel perception in quiet (Delattre et al., 1952; Klatt, 1982; Molis, 2005), a more complete description of the vowel spectrum in severe noise conditions therefore led to better recognition. Zahorian and Jagharghi (1993) argued that a whole-spectrum representation provides some relevant spectral information that are not provided by formants. Listeners therefore still use formant information to make vowel judgments in severe noise, but adequate information regarding the detailed spectral shape promotes the perception of formants.

Considering the importance of F1 and F2, the current study found a general trend that both these formants are of similar importance in severe noise. Parikh and Loizou (2005) measured the difference between the clean and noisy spectrum of vowels in the two spectral regions of F1 and F2 and found that, at a SNR of -5 dB, the spectral shape in the F2 region was severely altered by noise, more than in the F1 region. While the deformation of the spectral shape would normally lead to a decrease in vowel recognition scores, listeners were able to compensate for these distortions to a certain extent (Van Buuren, Festen and Houtgast, 1996; Watkins, 1991) and still perceive equal F1 and F2 information. It was also found that the importance of F1 and F2 in severe noise may differ for individual vowels depending on vowel backness.

Chapter 2 discussed several cues that have been found to be important for vowel perception in quiet. Results from the current study suggest that there exist both similarities and differences in the perception of vowels in quiet and severe noise. Formants alone have been found to be adequate for vowel recognition in quiet (Klatt, 1982), as well as duration and spectral change through time (Hillenbrand et al., 1995). Recognition scores in the current study showed that, with the exception of two vowels, similar vowel recognition scores were obtained for the whole-spectrum and formants-only vowels in quiet. In severe noise, listeners needed both whole-spectrum and formant cues to recognize vowels. MDS

analysis showed that the three-dimensional formant space explains the vowel confusions best in severe noise, but that the five-dimensional whole-spectrum vowel space becomes more relevant with a decrease in SNR.

Although no clear evidence was found from the literature to suggest that duration influences vowel perception in severe noise, it is known that duration may become relevant as other cues are masked by noise (Assmann and Summerfield, 2004). It was decided not to control the duration cue in the vowel synthesis process in the way that was done with the formants and spectral shape cues. Introducing duration as an additional factor in the current study would have led to an increase in the complexity of the statistical analysis, and would have shifted the focus away from the main research questions. Duration was however included in the FITA and MDS analysis as part of the F1, F2 and duration vowel space, which gave some indications on the relevance of duration in noise. In future studies, the duration cue may be controlled by altering the time duration of all synthesized vowels and monitoring the effect on recognition in severe noise.

The reasoning behind finding important perceptual cues in severe noise was to develop a model that could predict vowel recognition in degraded conditions. Based on the results in this study, information regarding the availability of both spectral detail and formants is needed to predict vowel confusions. Although formant frequencies should predict vowel confusions to a certain degree, knowledge about the availability of spectral detail would provide a measure of how well the vowel would be recognized. It was found that the importance of F1 and F2 depended on vowel backness. Therefore, a method of predicting confusions should not only rely on the similarity of formant frequencies between vowels, but also on the frequency difference between formants for each vowel.

The method by which cues important for vowel recognition in noise was determined, had some strong and weak points that merit discussion. The use of synthetic vowels to isolate certain cues in noise was successful. Knowing the exact cues that were present in the undegraded vowels assisted in finding the vowel characteristics that were important in

severe noise. Not using natural vowel utterances did, however, exclude the possibility that a combination of various cues was used for identifying vowels in severe noise. An alternative method would be to eliminate all possible cues, one at a time, in natural vowels and present the manipulated stimuli to listeners in noise. Recognition scores of such experiment would provide a more complete understanding of cues that are essential in degraded conditions.

Implementing MDS analysis successfully exposed the relevant vowel cues that are used by listeners in noise. It was considered to use general MDS techniques like classical MDS (Jaworska and Chupetlovska-Anastasova, 2009) or nonmetric MDS (Kruskal and Wish, 1978) rather than INDSCAL MDS. However, these techniques only allow for one dissimilarity matrix as input for each analysis. Due to the need for correlating the physical vowel cues to the MDS dimensions, these general MDS techniques were not adequate for the application in the current study, since it did not allow the results to be directly interpreted. In view of the fact that the coordinates merely represents the projection of each data point on the specific axes, rotation of the MDS results were needed to find the best correlation between each vowel cue and MDS dimension. The use of INDSCAL analysis provided two benefits in this regard, namely that it allowed for more than one dissimilarity matrix as input, while the results could also be directly interpreted without rotation.

6.7 SUMMARY

A discussion of the overall findings of the present study was done. It was shown that recognition scores and vowel confusions in noise occur as a result of spectral details that are available to the listener. While spectral detail would not necessarily be relevant in quiet, it does influence recognition in severely degraded conditions significantly. The human auditory system still uses formants to recognize vowels in severe noise, but adequate whole-spectrum information assists the listener in perceiving the formants. Depending on the type of vowel, either F1 or F2 is of importance, but the combination of

the two formants generally contributes the most to vowel identification in severe noise.

CHAPTER 7 CONCLUSION

This dissertation outlined the influence of different perceptual cues on vowel recognition in severe noise. A literature study was done to investigate vowel perception in different listening conditions for normal-hearing and CI listeners. Current vowel perception models were briefly discussed, after which studies on vowel recognition in degraded conditions gave insight into factors influencing vowel recognition in noise. Gaps in the literature became apparent, which suggested that the influence of formants versus spectral shape and the individual importance of F1 and F2 on vowel perception in severe noise should be investigated.

Vowel utterances from two speakers were approximated by using different synthesis methods to preserve certain cues and suppress others. Vowels containing either formants or spectral shape features were presented to listeners, as well as vowels where either F1 or F2 was suppressed. Results were analyzed in terms of the recognition rates of the different types of synthetic vowels, as well as the vowel confusions that became apparent. Experimental outcomes were matched to two different vowel spaces to find a measure by which vowel recognition in noise could be predicted.

Analysis of the results gave a final indication of the specific importance of spectral shape and vowel formants, as well as the individual significance of F1 and F2 in different noise conditions.

7.1 DISCUSSION OF RESEARCH QUESTIONS

The research questions raised in Chapter 1 are answered as follows:

- Formants still play a role in vowel recognition in severe noise, but formant cues

alone are not enough to yield good vowel recognition. If the undegraded vowel contains detailed spectral information additional to formant peaks, vowel recognition in severe noise improves significantly. Vowel perception errors in noise occur mostly between vowels sharing similar formant information, while vowel perception can be explained best by a formant vowel space.

- Overall, the combination of F1 and F2 is found to be important in noise only at -10 dB SNR, while in quiet and less severe noise conditions, F2 plays a bigger role than F1. The importance of F1 and F2 is also found to differ, depending on the frequency locations of the formants. For a SNR of -5 dB, both F1 and F2 are essential for central vowels, while at -10 dB SNR, F2 is less important for front vowels.
- Although formants are found to be robust in severe noise, a significant improvement in identification scores are seen when information regarding the detailed spectral shape is provided. Formant cues are perceived by the listeners in severe noise to a greater extent when the undegraded vowel contains whole-spectrum information. Vowel errors in noise are firstly related to formants, and then to detailed spectral information.
- Different SNR levels have a significant effect on the importance of vowel cues. In quiet and 0 dB SNR, F1, F2, and duration are the most important cues, although vowels consisting of whole-spectrum information (which also contain formants) give similar recognition scores. At -5 dB and -10 dB SNR, whole-spectrum cues are required for vowel recognition, even though the vowel confusion results are best explained by F1 and F2 cues.
- Vowel perception in severe noise can be predicted if the availability of both formants and spectral shape in the noise-corrupted vowels can be determined. Confusions between vowels can be predicted by noting similar F1 and F2 values, while a measure of availability of detailed spectral information will provide a probability of perceiving the formants. Knowledge regarding the vowel-specific attributes like vowel backness and diphthong or monothong categorization is also needed.

7.2 FINAL CONCLUSIONS

The following conclusions can be drawn from this study.

- Additional spectral shape cues lead to superior vowel recognition, compared to only formant cues in severe noise.
- Detailed spectral shape is used in conjunction with formants to recognize vowels in severe noise.
- Although it was shown that noise lowered the spectral contrast of formants and altered the spectral shape, listeners still perceived formant information and whole-spectrum information (but to a lesser extent) in severe noise.
- Vowels consisting of detailed spectral properties carry more formant information in severe noise than vowels consisting of only formants.
- The number of cues relevant for the perception of vowels decreases with a decrease in SNR.
- A three-dimensional formant and duration vowel space explains vowel confusions better than a five-dimensional whole-spectrum space in quiet and in noise.
- F2 plays a more significant role in vowel recognition in quiet, while in severe noise a combination of F1 and F2 cues is needed for good vowel recognition.
- The different contributions of F1 and F2 to vowel identification in noise depend on the frequency locations of the formants, as well as the level of noise.
- Vowel recognition in noise may differ according to the gender difference in talkers.

7.3 FUTURE WORK

In this study the intention was to determine the cues that are used by listeners to recognize vowels in noise. These results are required to develop a model that will predict vowel perception in degraded conditions.

The main focus fell on the importance of formant and whole-spectrum cues in severe noise. While various other cues exist that are known to be important for vowel recognition in quiet, these cues could also possibly play a role in severe noise. Additional cues that should be considered in noise include spectral change through time, spectral contrast, F3 and the fundamental frequency.

The synthetic vowels in the current study were generated by keeping the spectral change through time similar to the original vowel. The effect of spectral change could be explored by synthesizing vowels with constant formants through time, as was done in quiet conditions by Hillenbrand and Gayvert (1993) and Neel (2004). The influence of spectral contrast should be investigated by repeating the study of Hedrick and Nabelek (2004) where the intensity of F2 was gradually lowered in noise for the vowel /ʊ/. Additional vowels, as well as F1 and F3 cues, should be incorporated in the study to gain a better understanding of the influence of spectral contrast in noise for F1 and F2.

Hillenbrand et al. (2006) found that synthetic vowels that consist only of spectral peaks (that does not necessarily correspond to formants), were not recognized significantly poorer than formant synthesized vowels in quiet. This should be explored to establish if listeners rather attend to spectral peaks in severe noise, regardless of its formant label.

Although the majority of the whole-spectrum synthesized vowels were recognized better in severe noise than the formants-only vowels, recognition still varied between the different vowels. All vowels were scaled to an intensity level of 70 dB SPL before being presented to listeners, but a small group of vowels still showed similar recognition scores for the two

types of synthetic vowels. Significant differences between the vowel recognition for different speakers were also identified. In future studies the difference in vowel recognition between individual vowels should be investigated, while vowel utterances from various speakers should also be used to find a general trend for vowel perception over a pool of different voiced vowels.

The concordance index between confusion matrices of subjects showed that in severe noise, vowel perception differed considerably among listeners. More subjects are therefore needed to participate in experimental listening tests, while additional repetitions of vowels will ensure that experimental results are more accurate and interpretable. Similar results between subjects will also eliminate the need to group subjects together when conducting FITA and MDS analysis.

Finally, the findings in this dissertation were based on the recognition of vowels for normal-hearing listeners. Nabelek et al. (1992) found significant differences between normal-hearing and hearing-impaired listeners pertaining to vowel recognition in speech-shaped noise. It would be of interest to investigate these differences in severe noise. It should also be considered whether results from the current study can be used in modelling vowel perception for CI users.

REFERENCES

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels, *Journal of the Acoustical Society of America*, **51**(2 Part 2): 648-651.

Allen, J. B. (2005). Consonant recognition and the articulation index, *Journal of the Acoustical Society of America*, **117**(4): 2212-2223.

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

Assmann, P. F., Nearey, T. M. and Hogan, J. T. (1982). Vowel identification: Orthographic, and acoustic aspects, *Journal of the Acoustical Society of America*, **71**(4): 975-989.

Assmann, P. F. & Summerfield, Q. 2004, "Perception of speech under adverse conditions," in *Speech Processing in the Auditory system*, 1 edn, S. Greenberg et al., eds., Springer, pp. 231-308.

Bennett, D. C. (1968). Spectral form and duration as cues in the recognition of English and German vowels, *Language and speech*, **11**(2): 65-85.

Blamey, P. J., Dowell, R. C. and Tong, Y. C. (1984). Speech processing studies using an acoustic model of a multiple-channel cochlear implant, *Journal of the Acoustical Society of America*, **76**(1): 104-110.

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

Borden, G. and Harris, K. S. (1985). *Speech Science Primer* Williams and Wilkins, Baltimore.

Brusco, M. J. (2004). On the concordance among empirical confusion matrices for visual and tactual letter recognition, *Perception and Psychophysics*, **66**(3): 392-397.

Carlson, R., Fant, G. and Granstrom, B. (1975). Two-formant models, pitch and vowel perception, *Auditory Analysis and Perception of Speech* 55-82.

Carroll, J. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition, *Psychometrika*, **35**(3): 283-319.

Chan, J. and Simpson, C. (1990). Comparison of speech intelligibility in cockpit noise using SPH-4 flight helmet with and without active noise reduction, *Moffett Field, CA: U.S.Army Aviation Research and Technology*, ADA-227153.

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

Chistovich, L. A. and Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli, *Hearing Research*, **1**(3): 185-195.

Cooke, M., Green, P., Josifovski, L. and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Communication*, **34**(3): 267-285.

Coxon, A. P. M., Brier, A. P. and Hawkins, P. K. The newMDSX program series, Version 5. 2005. Edinburgh: NewMDSX Project.

Delattre, P., Liberman, A. M., Cooper, F. S. and Gerstman, L. J. (1952). An Experimental study of the acoustic determinants of vowel color; Observations on one- and two-formant vowels synthesized from spectrographic patterns, *Word*, **8**(3): 195-210.

Dorman, M. F., Loizou, P. C. and Fitzke, J. (1998a). The identification of speech in noise by cochlear implant patients and normal-hearing listeners using 6-channel signal processors, *Ear and Hearing*, **19**(6): 481-484.

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

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

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

Dorman, M. F., Loizou, P. C., Spahr, A. J. and Maloff, E. (2002). Factors that allow a high level of speech understanding by patients fit with cochlear implants, *American Journal of Audiology*, **11**(2): 119-123.

Dubno, J. R. and Dorman, M. F. (1987). Effects of spectral flattening on vowel identification, *Journal of the Acoustical Society of America*, **82**(5): 1503-1511.

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

Fattah, S. A., Zhu, W. P. and Ahmad, M. O. (2009). A time-frequency domain formant frequency estimation scheme for noisy speech signals, pp. 1201-1204.

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

Field, A. (2009). *Discovering Statistics using SPSS*, 3 edn, Sage Publications Ltd, Los Angeles.

Fishman, K. E., Shannon, R. V. and Slattery, W. H. (1997). Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor, *Journal of Speech, Language, and Hearing Research*, **40**(5): 1201-1215.

Fox, R. A., Flege, J. E. and Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis, *Journal of the Acoustical Society of America*, **97**(4): 2540-2551.

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

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

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

Gong, Y. (1995). Speech recognition in noisy environments: A survey, *Speech Communication*, **16**(3): 261-291.

Hedrick, M. S. and Nabelek, A. K. (2004). Effect of F2 intensity on identity of /u/ in degraded listening conditions, *Journal of Speech, Language and Hearing Research*, **47**(5): 1012-1021.

Hillenbrand, J. and Gayvert, R. T. (1993). Vowel classification based on fundamental frequency and formant frequencies, *Journal of Speech and Hearing Research*, **36**(4): 694-700.

Hillenbrand, J., Getty, L. A., Clark, M. J. and Wheeler, K. (1995). Acoustic characteristics of American english vowels, *Journal of the Acoustical Society of America*, **97**(5 I): 3099-3111.

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

Hillenbrand, J. M. and Houde, R. A. (2002). Speech synthesis using damped sinusoids, *Journal of Speech, Language, and Hearing Research*, **45**(4): 639-650.

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

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

Hoth, S. (2007). Indication for the need of flexible and frequency specific mapping functions in cochlear implant speech processors, *European Archives of Oto-Rhino-Laryngology*, **264**(2): 129-138.

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

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

Jaworska, N. and Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains, *Tutorials in Quantitative Methods for Psychology*, **5**(1): 1-10.

Johnson, K. (1989). Higher formant normalization results from auditory integration of F2 and F3, *Perception and Psychophysics*, **46**(2): 174-180.

Kasturi, K., Loizou, P. C., Dorman, M. and Spahr, T. (2002). The intelligibility of speech with "holes" in the spectrum, *Journal of the Acoustical Society of America*, **112**(3): 1102-1111.

Kewley-Port, D., Burkle, T. Z. and Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners, *Journal of the Acoustical Society of America*, **122**(4): 2365-2375.

Kiefte, M., Enright, T. and Marshall, L. (2007). Formant Amplitude in the Perception of /i/ and /u/, *Journal of the Acoustical Society of America*, **121**(5): 3189.

Kiefte, M. and Kluender, K. R. (2005). The relative importance of spectral tilt in monophthongs and diphthongs, *Journal of the Acoustical Society of America*, **117**(3): 1395-1404.

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

Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step, *Acoustics, Speech and Signal Processing, IEEE International Conference*.

Klatt, D. H. (1987). Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America*, **82**(3): 737-793.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America*, **87**(2): 820-857.

Klein, W., Plomp, R. and Pols, L. C. W. (1970). Vowel spectra, vowel spaces and vowel identification, *Journal of the Acoustical Society of America*, **48**(4 pt 2): 995-1009.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*, 1 edn, Sage Publications, Newbury Park, London, New Delhi.

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

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

Liu, C. and Fu, Q. J. (2005). Relating the acoustic space of vowels to the perceptual space in cochlear implant simulations, *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, 18 March, 2005, Philadelphia, PA* Vol. III.

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

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

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

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

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

Marguiles, M. K. (1979). Male-female differences in speaker intelligibility; normal and hearing-impaired listeners, *Journal of the Acoustical Society of America*, **65**(S1): S99.

Markel, J. D. (1972). Sift algorithm for fundamental frequency estimation, *IEEE Transactions on Audio and Electroacoustics*, **AU-20**(5): 367-377.

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

Miller, G. A. and Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants, *Journal of the Acoustical Society of America*, **27**(2): 338-352.

Molau, S., Pitz, M., Schluter, R. and Ney, H. (2001). Computing mel-frequency cepstral coefficients on the power spectrum, *2001 IEEE Interntional Conference on Acoustics, Speech, and Signal Processing, 7 May, 2001, Salt Lake, UT* Vol. 1, pp. 73-76.

Molis, M. R. (2005). Evaluating models of vowel perception, *Journal of the Acoustical Society of America*, **118**(2): 1062-1071.

Mugavin, M. E. (2008). Multidimensional scaling: A brief overview, *Nursing Research*, **57**(1): 64-68.

Nabelek, A. K., Czyzewski, Z. and Krishnan, L. A. (1992). The influence of talker differences on vowel identification by normal- hearing and hearing-impaired listeners, *Journal of the Acoustical Society of America*, **92**(3): 1228-1246.

Nabelek, A. K., Ovchinnikov, A., Czyzewski, Z. and Crowley, H. J. (1996). Cues for perception of synthetic and natural diphthongs in either noise or reverberation, *Journal of the Acoustical Society of America*, **99**(3): 1742-1753.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception, *Journal of the Acoustical Society of America*, **85**(5): 2088-2113.

Nearey, T. M. and Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification, *Journal of the Acoustical Society of America*, **80**(5): 1297-1308.

Neel, A. T. (2004). Formant detail needed for vowel identification, *Acoustic Research Letters Online*, **5**: 125-131.

Nossair, Z. B. and Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants, *Journal of the Acoustical Society of America*, **89**(6): 2978-2991.

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

Paul, D. B. (1981). Spectral envelope estimator vocoder, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-29**(4): 786-794.

Peterson, G. E. and Barney, H. L. (1952). Control Methods Used in a Study of the Vowels, *Journal of the Acoustical Society of America*, **24**(2): 175-184.

Phatak, S. A. and Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise, *Journal of the Acoustical Society of America*, **121**(4): 2312-2326.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of speech signals*, 1 edn, Prentice-Hall inc..

Regnier, M. S. and Allen, J. B. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/, *Journal of the Acoustical Society of America*, **123**(5): 2801-2814.

Remus, J. J. and Collins, L. M. (2004). Vowel and consonant confusion in noise by cochlear implant subjects: Predicting performance using signal processing techniques, *Proceedings - IEEE International Conference on Acoustics, Speech, and Signal Processing, 17 May, 2004, Montreal, Que* Vol. 4.

Remus, J. J. and Collins, L. M. (2005). The effects of noise on speech recognition in cochlear implant subjects: Predictions and analysis using acoustic models, *Eurasip Journal on Applied Signal Processing*, **2005**(18): 2979-2990.

Sakayori, S., Kitama, T., Chimoto, S., Qin, L. and Sato, Y. (2002). Critical spectral regions for vowel identification, *Neuroscience Research*, **43**(2): 155-162.

Sawusch, J. R. (1996). Effects of duration and formant movement on vowel perception, *Proceedings of the 1996 International Conference on Spoken Language Processing, ICSLP. Part 1 (of 4), 3 October, 1996, Philadelphia, PA, USA* Vol. 4, IEEE, pp. 2482-2485.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science*, **270**(5234): 303-304.

Strange, W., Jenkins, J. J. and Johnson, T. L. (1983). Dynamic specification of coarticulated vowels, *Journal of the Acoustical Society of America*, **74**(3): 695-705.

Summers, V. and Cord, M. T. (2007). Intelligibility of speech in noise at high presentation levels: Effects of hearing loss and frequency region, *Journal of the Acoustical Society of America*, **122**(2): 1130-1137.

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

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

Van Buuren, R. A., Festen, J. M. and Houtgast, T. (1996). Peaks in the frequency response of hearing aids: Evaluation of the effects on speech intelligibility and sound quality, *Journal of Speech, Language, and Hearing Research*, **39**(2): 239-250.

Van Hoesel, R., Böhm, M., Battmer, R. D., Beckschebe, J. and Lenarz, T. (2005). Amplitude-mapping effects on speech intelligibility with unilateral and bilateral cochlear implants, *Ear and Hearing*, **26**(4): 381-388.

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

Verbrugge, R. R., Strange, W., Shankweiler, D. P. and Edman, T. R. (1976). What information enables a listener to map a talker's vowel space?, *Journal of the Acoustical Society of America*, **60**(1): 198-212.

Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B. and Lunner, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking, *Journal of the Acoustical Society of America*, **125**(4): 2336-2347.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion, *Journal of the Acoustical Society of America*, **90**(6): 2942-2955.

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

Wickelmaier, F. (2003). An introduction to MDS, *Sound Quality Research Unit, Aalborg University, Denmark.*

Williams, C. E., Pearsons, K. S. and Hecker, M. H. L. (1971). Speech Intelligibility in the Presence of Time-Varying Aircraft Noise, *Journal of the Acoustical Society of America*, **50**(2A): 426-434.

Xu, L., Thompson, C. S. and Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition, *Journal of the Acoustical Society of America*, **117**(5): 3255-3267.

Xu, L. and Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise, *Journal of the Acoustical Society of America*, **122**(3): 1758-1764.

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

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

Zwolan, T. A. and Kileny, P. R. (1993). Cochlear implants for the profoundly deaf, *Proceedings of the 6th Annual IEEE Symposium on Computer-Based Medical Systems, 13 June, 1993, Ann Arbor, MI, USA* Publ by IEEE, pp. 241-246.

ADDENDUM A

Confusions matrices presented in Chapter 5 (Figure 5.25 to Figure 5.36) are analyzed in Addendum A for each individual vowel.

A.1. /yː/, puut (front, long vowel)

Complete spectrum vowel

The vowel /yː/ is well recognized for both types of synthesized vowels in quiet. With a decrease in SNR, the formants-only vowel is confused with /ɐ/ (situated close to /yː/ in the formant space), while the whole-spectrum type vowel is confused with /ɛː/ (sharing similar F2, duration and spectral band cues). In severe noise, the formants-only vowel is confused with /i/ (sharing similar F1 and F2 cues with /yː/) and /u/ (sharing similar F1 cues). The whole-spectrum vowel /yː/ is confused with /i/ (sharing similar F1, F2 and spectral band cues), /ɛː/ and /ɛ/ (sharing F1, F2 and spectral band cues).

F1-suppressed

The vowel /yː/ is confused with /ɛː/ (sharing similar F2, duration and spectral band cues). In noise, confusions with /i/ (sharing similar F1, F2 and spectral band cues) and /ɛ/ (sharing similar F2 and spectral band cues) are observed for both whole-spectrum and formants-only synthesis types. The confusion with /ɛ/ can suggest that no F1 is perceived in noise, while the confusion with /i/ indicates that whole-spectrum cues are perceived.

F2-suppressed

An 18% confusion with /u/ for the formants-only vowels is the only sign of the suppression of F2 in quiet. The whole-spectrum vowel is confused with /ɛː/ (sharing similar F2 and spectral band cues with /yː/). With a decrease in SNR, confusions for the formants-only vowels with /u/ increase, showing that the F2 cue is not perceived. For the whole-spectrum vowels, confusions with /i/, /ɛ/ and /ɛː/ (which is in close proximity in the spectral bands space) show no sign of a suppressed F2, while the /i/ and /ɛ/ confusions show that noise eliminates the duration cue.

A.2. /ɪ/, piet (front short vowel)

Complete spectrum vowel

In noise, the formants-only vowel /ɪ/ is confused with /yː/ containing similar F1 and F2 cues. In severe noise, /ɪ/ is confused with /ʊ/ (sharing similar duration and F1 cues). For the whole-spectrum vowels, no other major confusions are observed. Major confusions in severe noise occur only for the formants-only vowels.

F1-suppressed

For the F1-suppressed vowel for both types of synthesized vowels, confusions in quiet occur with the vowel /ɛ/ (sharing similar F2 and duration cues). For a decrease in SNR, the whole-spectrum vowels show more confusions with /ɛ/ than the formants-only vowels, while in severe noise, minor confusions are distributed over the matrix. The effect of F1 suppression created more uncertainty among listeners, as the confusion distribution in severe noise is less than for the same SNR for the complete spectrum vowels.

F2-suppressed

For the formants-only vowels in quiet, vowels are confused with /ʊ/ containing only similar F1 and duration cues. There is also a 13% confusion with /yː/ sharing only similar F1 and F2 values. For the whole-spectrum vowels in quiet, only a small confusion with /ʊ/ is seen. With the SNR decreasing, more confusions with /ʊ/ occur for the formants-only vowels, while no major confusions occur for the whole-spectrum vowels.

A.3. /ʊ/, poet (back, short vowel)

Complete spectrum vowel

The vowel /ʊ/ is recognized relatively well for both whole-spectrum and formants-only vowels in quiet and 0 dB SNR. In severe noise, at -10 dB SNR, the formants-only vowel shows confusions with the vowels /ɑ/ (sharing similar F2 and duration values) and /ɪ/ (sharing similar F1 and duration values). No major confusions are detected for the whole-spectrum vowel in severe noise.

F1-suppressed

No major differences between the whole-spectrum and formants-only vowel confusions are observed for this vowel. In the quiet condition, some confusions with /ɑ/ occur (sharing similar duration and F2 cues). As the SNR decreases, more confusions with the /ɑ/ vowel occur, while less correct responses for /u/ are seen.

F2-suppressed

In quiet, the whole-spectrum vowel shows better recognition results (67%) than the formants-only vowels (38%). The formants-only and whole-spectrum vowel is confused with /i/ (sharing similar F1 and duration cues than /u/), but the formants-only vowel also shows a 32% confusion with /yː/ (sharing only a similar F1 cue). For the whole-spectrum vowels, the duration cue is perceived better than for the formants-only vowels. The same confusions that are observed in quiet also occur in noise, except for the confusion with /e/ (sharing a similar F1 cue than /u/) for the whole-spectrum vowels.

A.4. /e/, peet (front, long vowel)

Complete spectrum vowel

No difference between the confusions for the complete spectrum and exclusive formant vowels is seen in quiet. In severe noise, the whole-spectrum vowels are recognized better than the formants-only vowels. The whole-spectrum vowels are confused with /ɛː/ (sharing similar formants and whole-spectrum cues) and with /ɛ/ (sharing similar F1, F2 and spectral band cues). For the formants-only vowels, confusions occur with /u/ (sharing a similar F1 cue), /i/ (sharing similar F1 and F2 cues) and /yː/ (sharing F1, F2 and duration cues). Noise therefore eliminated the F2 and duration cues, and also increased the uncertainty under listeners for the formants-only vowels.

F1-suppressed

The suppression of F1 does not show any effect on vowel recognition in quiet for this vowel. In severe noise, the whole-spectrum vowels are recognized significantly better than

the exclusive formant vowels. Confusions are made with /i/ and /y:/, vowels that share the same F2 cue than the steady state part of the vowel /e/.

F2-suppressed

The whole-spectrum vowel is recognized better than the formants-only vowel in quiet and in noise for the F2-suppressed vowels. As noise is added, more confusions with /u/ (sharing only the same F1 as /e/) become apparent for the formants-only vowels.

A.5. /ɛ:/, pêt (front, long vowel)

Complete spectrum vowel

The formants-only vowels are recognized better than the whole-spectrum vowels in quiet. As noise is added, both synthesis type vowels show poor recognition performance. For the whole-spectrum vowels, more confusions with /y:/ (sharing similar F2, duration and whole-spectrum cues) and /i/ (sharing similar F2 and whole-spectrum cues) occur in quiet and noise. A lack of F1 information can therefore be the reason for the poor recognition performance of the whole-spectrum vowels. A 19% confusion is also seen with /e/ (sharing similar F2, duration and whole-spectrum cues).

F1-suppressed

With no F1 present in the vowel spectrum, the recognition of the whole-spectrum and formants-only vowels are similar in quiet and noise. Both types of vowels are confused with /y:/, showing that the F1 cue is not perceived. Similar to the complete spectrum vowels, poor performance is seen in severe noise.

F2-suppressed

Again the whole-spectrum and formants-only vowels show similar recognition scores in quiet and in noise. More confusions with /y:/ (sharing similar F2, duration and spectral shape cues) are found for the whole-spectrum vowels at 0dB SNR. In severe noise, the formants-only vowels are confused with /ɔ/ (sharing only a similar F1 cue with /ɛ:/).

A.6. /ɛ/, pet (front, short vowel)

Complete spectrum vowel

In quiet, the formants-only vowels are recognized better than the whole-spectrum vowels. With noise added, recognition scores decrease and confusions occur mainly with /i/ for the whole-spectrum vowels. The vowel /i/ shares similar duration and F2 values than /ɛ/ and also contains an F1 value within 200 Hz from the F1 value of /i/. In addition, it is grouped together in the spectral band space. No difference in confusion patterns are found between the two types of synthesized vowels in severe noise.

F1-suppressed

For the F1-suppressed vowel tokens, the whole-spectrum and formants-only vowels show similar confusion patterns. The percentage recognition score decrease in quiet when F1 is suppressed, while confusions with /i/ (sharing similar F2 and duration cues) also occur.

F2-suppressed

The whole-spectrum vowel /ɛ/ is not affected by the F2 suppression in quiet, as the same recognition score is obtained for complete spectrum and F2-suppressed vowel. However, the formants-only vowel recognition score decreased significantly from 93% to 23%. Proof that F2 is still perceived by the listeners for the whole-spectrum vowels is found by the confusion with the vowel /i/ (sharing similar F2 and duration values than /ɛ/). With a decrease in SNR, confusions with /i/ increase. The formants-only vowels do not show the same confusions as the whole-spectrum vowels; it is mainly confused with /ɔ/, sharing the same F1 and duration value than /ɛ/.

A.7. /œ/, put (central, short)

Complete spectrum vowel

The whole-spectrum and formants-only vowel /œ/ show the same percentage correct scores in quiet. When noise is added, more confusions occur with the vowel /ə/ for both types of synthesized vowels (located close to /œ/ in the formant and spectral band vowel

space).

F1-suppressed

When F1 is suppressed for this vowel, both the whole-spectrum and formants-only vowels show the same confusions. For both types of synthesized vowels, confusions are found with /æ/ sharing similar F2 and duration cues as /œ/. Confusions with /ə/ (close to /œ/ in both the formant and whole-spectrum space) also occur and increase as the SNR is decreased.

F2-suppressed

The whole-spectrum F2-suppressed vowels show better recognition scores than the formants-only vowels in quiet and in noise. The two main confusions that occur is with /ɔ/ and /ɛ/ (sharing similar F1 and duration cues).

A.8. /ɔ/, pot (back, short)

Complete spectrum vowel

The vowel /ɔ/ is not situated near any other vowels in the formant vowel space. It is therefore not surprising that the recognition score for both the whole-spectrum and formants-only vowels are relatively high. Confusions that become apparent in severe noise are with the vowels /œ/, /ʊ/ and /ɑː/, all which are situated in the greater vicinity of /ɔ/ in the formant space.

F1-suppressed

When F1 is suppressed, the whole-spectrum vowels still show good recognition in quiet, while the recognition for the formants-only vowels decreases. With the addition of noise however, the whole-spectrum vowel recognition score decreases, while the formants-only vowel score stays constant. In severe noise, the whole-spectrum vowel /ɔ/ is confused with /ɑː/ (sharing a similar F2 value than /ɔ/) and /a/ (sharing a similar F2 and duration value).

F2-suppressed

The whole-spectrum and formants-only vowels are recognized relatively well without the F2 cue. Confusions with /ɛ/ and /ə/ become apparent (sharing similar F1 values than /ɔ/).

A.9. /ə/, pit (central, short)

Complete spectrum vowel

No significant difference in confusions is seen between the whole-spectrum and formants-only vowels for the vowel /ə/. Confusions with the vowel /œ/ (situated close to /ə/ in both the formant and whole-spectrum vowel space) become apparent and decreases as the SNR is decreased.

F1-suppressed

No substantial difference is observed between the percentage correct scores for the complete spectrum vowel and the F1-suppressed vowel /ə/. Similar to the complete spectrum vowel results, confusions with /œ/ initially become apparent, but decreases with the addition of noise. In quiet and noise, confusions with the vowel /æ/ are found (sharing similar F2 and duration cues as /ə/). In severe noise, both types of synthesized vowels show confusions with /i/ (having neither F1 nor F2 related to the formant cues of /ə/).

F2-suppressed

Poor recognition performance is seen when F2 is suppressed for /ə/. In quiet, better recognition is found for the whole-spectrum vowels (35%) compared to the formants-only vowels (7%), while similar percentage correct scores are seen in noise. In quiet, the F2-suppressed vowel /ə/ is confused with /ɛ/ (sharing similar F2 and whole-spectrum cues) and /œ/ (situated very close to /ə/ in the formant and whole-spectrum vowel space). For the formants-only vowel, a confusion score of 61% is found with the vowel /ɔ/ (sharing a similar F1 value).

A.10. /æ/, pat (central, short)

Complete spectrum vowel

For the vowel /æ/ both types of vowels (formants-only and whole-spectrum) are well recognized in quiet. Confusions with /œ/ for both types of vowels are explained by the similar F2 and F3 values. In severe noise for the formants-only vowels, confusions with /ɔ/ occurs.

F1- suppressed

The suppression of F1 influences only the percentage correct scores for the formants-only vowels compared to the whole-spectrum vowels. When noise is added, confusions with /œ/ and /ə/ (both with similar F2 and duration values than /æ/) as well as with /a/ (close to /æ/ in both the formant and whole-spectrum vowel space) become apparent. Confusions with more vowels become apparent with a reduction in the SNR.

F2-suppressed

Similar confusions in noise are seen for the two types of synthetic vowels. In quiet, confusions with /ɔ/ become apparent. Confusions with /a/ (sharing similar duration and F1 cues) show that listeners rely on F1 in the absence of the F2 cue, while confusions with /ɑː/ (sharing similar F1 and F2 cues) also become apparent.

A.11. /ɑː/, paat (back, long)

The vowel /ɑː/ is robust in noise, even when F1 is suppressed. When F2 is suppressed in severe noise, confusions with /ɔ/ are seen. Other than /a/ and /ɑː/, the vowel /ɔ/ is the only other back vowel in the vowel space. In low SNRs, the higher frequency detail is totally masked by noise, leaving only the first formant available to the listener. For both speakers and with duration eliminated by noise, the remaining first formant of the vowel /ɑː/ represents the average between F1 and F2 of the vowel /ɔ/, creating a possibility for the listener to perceive this vowel as the vowel /ɔ/.

A.12. /ɑ/, pad (back, short)

Similar to /ɑː/, the vowel /ɑ/ is robust in noise, even with F1 suppressed. When F2 is suppressed in quiet, the formants-only vowels show a percentage correct score of only 15% compared to the whole-spectrum vowel score of 45%. The percentage correct score for the formants-only vowels increases at 0 dB SNR, after which it decreases to match the same score as the whole-spectrum vowel at -10 dB SNR. Confusions with /æ/ can be explained by the comparable F1 and duration values. For the quiet condition, the formants-only vowels are confused with /ɑː/.