PERCEPTION OF PROSODY BY COCHLEAR IMPLANT RECIPIENTS

by

Marianne van Zyl

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor (Biosystems)

in the

Department of Electrical, Electronic and Computer Engineering

Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

April 2014

ACKNOWLEDGEMENTS

- To the Almighty God, our Father in Jesus Christ, the source of all truth and wisdom, and the Creator of all things – my eternal gratitude for his grace and provision which enabled me to complete this project.
- My husband, André, for his constant support and help in all the different aspects of the project.
- Our daughter, Annabel, for patiently playing with grandma while I was working on the thesis.
- My parents and parents-in-law for their prayers and support.
- Prof Johan Hanekom for his support, wisdom and guidance throughout this journey.
- My Biogroup friends, specifically Fritz, Farao, Pieter, Dirk, Ilse, Larry and Stephen for their moral and technical support, and for the hours of wonderful and stimulating conversation we shared.
- All the participants (normal-hearing and implant recipients) who generously and patiently “lent me their ears”.
- Andy Field, for making statistics easy and even amusing.
- The National Research Foundation for financial support.

SUMMARY

PERCEPTION OF PROSODY BY COCHLEAR IMPLANT RECIPIENTS

by

Marianne van Zyl

Supervisor:	Prof J.J. Hanekom
Department:	Electrical, Electronic and Computer engineering
University:	University of Pretoria
Degree:	Philosophiae Doctor (Biosystems)

Keywords: cochlear implants, prosody, suprasegmental cues, noise, speech recognition in noise, phoneme recognition, intonation, vocal emotion, speech-weighted noise

Recipients of present-day cochlear implants (CIs) display remarkable success with speech recognition in quiet, but not with speech recognition in noise. Normal-hearing (NH) listeners, in contrast, perform relatively well with speech recognition in noise. Understanding which speech features support successful perception in noise in NH listeners could provide insight into the difficulty that CI listeners experience in background noise. One set of speech features that has not been thoroughly investigated with regard to its noise immunity is prosody. Existing reports show that CI users have difficulty with prosody perception. The present study endeavoured to determine if prosody is particularly noise-immune in NH listeners and whether the difficulty that CI users experience in noise can be partly explained by poor prosody perception. This was done through the use of three listening experiments.

The first listening experiment examined the noise immunity of prosody in NH listeners by comparing perception of a prosodic pattern to word recognition in speech-weighted noise (SWN). Prosody perception was tested in a two-alternatives forced-choice (2AFC) test paradigm using sentences conveying either conditional or unconditional permission, agreement or approval. Word recognition was measured in an open set test paradigm using meaningful sentences. Results indicated that the deterioration slope of prosody recognition (corrected for guessing) was significantly shallower than that of word recognition. At the lowest signal-to-noise ratio (SNR) tested, prosody recognition was significantly better than word recognition.

The second experiment compared recognition of prosody and phonemes in SWN by testing perception of both in a 2AFC test paradigm. NH and CI listeners were tested using single words as stimuli. Two prosody recognition tasks were used; the first task required discrimination between questions and statements, while the second task required discrimination between a certain and a hesitant attitude. Phoneme recognition was measured with three vowel pairs selected according to specific acoustic cues. Contrary to the first experiment, the results of this experiment indicated that vowel recognition was significantly better than prosody recognition in noise in both listener groups.

The difference between the results of the first and second experiments was thought to have been due to either the test paradigm difference in the first experiment (closed set versus open set), or a difference in stimuli between the experiments (single words versus sentences). The third experiment tested emotional prosody and phoneme perception of NH and CI listeners in SWN using sentence stimuli and a 4AFC test paradigm for both tasks. In NH listeners, deterioration slopes of prosody and phonemes (vowels and consonants) did not differ significantly, and at the lowest SNR tested there was no significant difference in recognition of the different types of speech material. In the CI group, prosody and vowel perception deteriorated with a similar slope, while consonant recognition showed a steeper slope than prosody recognition. It is concluded that while prosody might support speech recognition in noise in NH listeners, explicit recognition of prosodic patterns is not particularly noise-immune and does not account for the difficulty that CI users experience in noise.

OPSOMMING

PERSEPSIE VAN PROSODIE DEUR KOGLEÊRE-INPLANTINGGEBRUIKERS

deur

Marianne van Zyl

Studieleier: Prof J.J. Hanekom

Departement: Elektriese, Elektroniese en Rekenaaringenieurswese

Universiteit: Universiteit van Pretoria

Graad: Philosophiae Doctor (Biosisteme)

Sleutelwoorde: kogleêre inplantings, prosodie, suprasegmentele leidrade, geraas, spraakherkenning in geraas, foneemherkenning, intonasie, stem-emosie, spraakgeweegde ruis

Ontvangers van hedendaagse kogleêre inplantings (KI's) behaal merkwaardige sukses met spraakherkenning in stilte, maar nie met spraakherkenning in geraas nie. Normaalhorende (NH) luisteraars, aan die ander kant, vaar relatief goed met spraakherkenning in geraas. Begrip van die spraakeienskappe wat suksesvolle persepsie in geraas ondersteun in NH luisteraars, kan lei tot insig in die probleme wat KI-gebruikers in agtergrondgeraas ervaar. Een stel spraakeienskappe wat nog nie deeglik ondersoek is met betrekking tot ruisimmuniteit nie, is prosodie. Bestaande navorsing wys dat KI-gebruikers sukkel met persepsie van prosodie. Die huidige studie is onderneem om te bepaal of prosodie besonder ruisimmuun is in NH luisteraars en of die probleme wat KI-gebruikers in geraas ondervind, deels verklaar kan word deur swak prosodie-persepsie. Dit is gedoen deur middel van drie luistereksperimente.

Die eerste luistereksperiment het die ruisimmuniteit van prosodie in NH luisteraars ondersoek deur die persepsie van ’n prosodiese patroon te vergelyk met woordherkenning in spraakgeweegde ruis (SGR). Prosodie-persepsie is getoets in ’n twee-alternatiewe-gedwonge-keuse- (2AGK) toetsparadigma met sinne wat voorwaardelike of onvoorwaardelike toestemming, instemming of goedkeuring oordra. Woordherkenning is gemeet in ’n oopstel-toetsparadigma met betekenisvolle

sinne. Resultate het aangedui dat die helling van agteruitgang van prosodieherkenning (gekorrigeer vir raai) betekenisvol platter was as dié van woordherkenning, en dat by die laagste sein-tot-ruiswaarde (STR) wat getoets is, prosodieherkenning betekenisvol beter was as woordherkenning.

Die tweede eksperiment het prosodie- en foneemherkenning in SGR vergelyk deur die persepsie van beide te toets in ’n 2AGK-toetsparadigma. NH en KI-luisteraars is getoets met enkelwoorde as stimuli. Twee prosodieherkenningstake is gebruik; die eerste taak het diskriminasie tussen vrae en stellings vereis, terwyl die tweede taak diskriminasie tussen ’n seker en onseker houding vereis het. Foneemherkenning is gemeet met drie vokaalpare wat geselekteer is na aanleiding van spesifieke akoestiese eienskappe. In teenstelling met die eerste eksperiment, het resultate van hierdie eksperiment aangedui dat vokaalherkenning betekenisvol beter was as prosodieherkenning in geraas in beide luisteraarsgroepe.

Die verskil tussen die resultate van die eerste en tweede eksperimente kon moontlik die gevolg wees van óf die verskil in toetsparadigma in die eerste eksperiment (geslote- teenoor oop-stel), óf ’n verskil in stimuli tussen die eksperimente (enkelwoorde teenoor sinne). Die derde eksperiment het emosionele-prosodie- en foneempersepsie van NH en KI-luisteraars getoets in SGR met sinstimuli en ’n 4AGK-toetsparadigma vir beide take. In NH luisteraars het die helling van agteruitgang van die persepsie van prosodie en foneme (vokale en konsonante) nie betekenisvol verskil nie, en by die laagste STR wat getoets is, was daar nie ’n betekenisvolle verskil in die herkenning van die twee tipes spraakmateriaal nie. In die KI-groep het prosodie- en vokaalpersepsie met soortgelyke hellings agteruitgegaan, terwyl konsonantherkenning ’n steiler helling as prosodieherkenning vertoon het. Die gevolgtrekking was dat alhoewel prosodie spraakherkenning in geraas in NH luisteraars mag ondersteun, die eksplisiete herkenning van prosodiese patrone nie besonder ruisimmuun is nie en dus nie ’n verklaring bied vir die probleme wat KI-gebruikers in geraas ervaar nie.

LIST OF ABBREVIATIONS

ACE	Advanced Combination Encoder
AFC	Alternatives forced-choice
ANOVA	Analysis of variance
CI	Cochlear implant
CVC	Consonant-vowel-consonant
DL	Difference limen
F0	Fundamental frequency
F1	Formant one
F2	Formant two
FS	Female speaker
GUI	Graphic user interface
HINT	Hearing in Noise Test
HNR	Harmonics-to-noise ratio
MS	Male speaker
NH	Normal-hearing
rms	Root-mean-square
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SRT	Speech recognition threshold
SWN	Speech-weighted noise
VCV	Vowel-consonant-vowel

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 PROBLEM STATEMENT

1.1.1 Context of the problem

Speech recognition in background noise poses a great challenge to all listeners, especially individuals with hearing loss who rely on hearing aids or cochlear implants (CIs) to aid their communication. A great deal of daily spoken communication occurs in at least some degree of background noise, and the ability to perceive speech successfully in the presence of noise is therefore an important skill that has a significant impact on an individual's quality of life. Present-day CIs provide listeners who had very little or no residual hearing with access to sufficient acoustic cues for successful perception of many auditory stimuli, and many CI recipients display remarkable success with open set speech recognition in quiet (e.g. Caposecco, Hickson and Pedley, 2012). However, speech recognition in noise remains a problem for these listeners, who require a much more favourable signal-to-noise ratio (SNR) than normal-hearing (NH) listeners to obtain the same degree of success with speech recognition in noise (Gifford & Revit, 2010).

Investigations into the relative noise immunity and importance of different speech cues and speech features[1] for recognising speech in noise play an important role in aiding researchers' understanding of how NH listeners perceive speech in noise. Understanding which speech features support successful perception in noise in NH listeners could provide insight into the difficulty that listeners with CIs experience in background noise, and might offer some solutions to this challenge. Extensive work has been done on the cues underlying the recognition of segmental speech features (vowels and consonants) in quiet and the availability of these cues in noise. For example, a number of studies have investigated the acoustic cues that enable NH listeners to identify vowels even in severe background noise (e.g. Ferguson, 2004; Ferguson and Kewley-Port, 2002), while other studies have examined how speech-

[1] The term "speech cues" or "cues", as used in this study, refers to underlying acoustic cues of speech. The term "speech features" is used to refer to segmental (phonetic) and/or suprasegmental (prosodic) elements of speech.

weighted noise (SWN) affects consonant identification (Phatak and Allen, 2007; Woods, Yund, Herron and Ua Cruadhlaoich, 2010).

1.1.2 Research gap

One set of speech features that has not been thoroughly investigated to date in noise is suprasegmental features or prosody. Prosody fulfils a variety of important functions in spoken communication. Reports in existing literature suggest that some prosodic cues are important for speech perception in noise in NH listeners (see Chapter 2 for details), and may be more immune to the effects of noise than segmental information, although the evidence for this is still very limited. Anecdotally, it seems that in difficult listening situations it is often easier to hear *how* a person said something (i.e. the prosody of the utterance) than to hear *what* exactly he or she said (i.e. the content of the utterance, consisting of phonemes and words). However, this observation has, to the best of the researcher's knowledge, not been directly investigated in existing literature. Further investigation into the noise immunity of prosody is therefore needed, along with comparisons between the relative noise immunity of prosody and other important speech features such as vowels and consonants.

Evidence from listeners with CIs indicates that these listeners have difficulty in perceiving a number of important prosodic cues (Meister, Landwehr, Pyschny, Walger and Wedel, 2009; Most and Aviner, 2009; Most, Gaon-Sivan, Shpak and Luntz, 2012). Given the importance of prosody and the role that it plays in speech recognition in noise in NH listeners, CI recipients' difficulty with prosody perception might provide a partial explanation for their difficulty to perceive speech in noise. The present study was undertaken with the goal to gain better understanding of the perception of prosody in noise, and the role that prosody plays in speech recognition in noise, as a means to provide deeper insight into the difficulties that CI recipients experience in background noise. Chapter 2 provides more details on existing literature on the topic as a theoretical background for the experimental work described in Chapters 3, 4, and 5.

1.2 HYPOTHESIS

In light of the reported difficulty that CI recipients have with the perception of prosody (Meister *et al.*, 2009; Most and Aviner, 2009; Most *et al.*, 2012), it was hypothesised that these listeners would perform more poorly on the recognition of prosody in comparison to the recognition of segmental speech information in noise. If this hypothesis could be supported by experimental data, it would indicate the importance of conveying prosodic cues more accurately to CI users and future efforts in improving speech processors should specifically attempt to improve access to these cues. Fundamental voice frequency (F0), for example, is an important cue to a variety of prosodic functions (see Chapter 2 for details), and various ways of conveying F0 to CI users have been reported over the years, although much work remains to be done in evaluating and improving these techniques (Brown and Bacon, 2010). Besides providing support for ongoing efforts to improve the encoding of F0 and other prosodic cues in speech processors, the present work could demonstrate useful techniques and guidelines on how to assess these improvements through specific evaluation of both prosody and phoneme perception.

Furthermore, it was hypothesised that for NH listeners, the perception of prosody in noise would be better than the perception of phonemes or words, given anecdotal observations from everyday experience as described in Section 1.1.2, and the evidence in existing literature (albeit limited) (see e.g. Mattys, 2004; Smith, Cutler, Butterfield and Nimmo-Smith, 1989 for evidence on the noise immunity of prosodic cues). Sentence-level prosody was expected to be especially noise immune because of its redundancy (information spread out across a number of words or even all the words in the sentence), and in light of the literature mentioned above. Lexical (word-level) prosody was also expected to be more noise robust than phoneme recognition, as the supporting cues for prosody are often spread out across more than one segment/phoneme. Prosody might support speech recognition in noise in NH listeners in a variety of ways, especially if prosodic features turn out to be more noise immune than segmental speech features. On a lexical level, prosody provides cues to word identity by marking syllable stress (Fry, 1955; Fry, 1958). This could be helpful in a situation where noise prevents accurate perception of all phonemes constituting a word as it might enable a listener to identify a word according to its stress pattern (Binns and Culling 2007). On a sentence level, the intonation contour can improve speech perception by stressing important content words and thereby increasing the processing priority of these words (Laures & Weismer 1999). Stress (a prosodic cue) can also help listeners to identify word boundaries (Mattys, 2004), which might aid word recognition within a sentence. Research by Nygaard, Herold and Namy (2009) has also suggested that prosody can even help a listener to guess the meaning of a novel word. If prosody supports semantics in this manner, it might also improve a listener's chances of correctly identifying a word in adverse listening conditions. Finally, if a listener can identify a speaker's emotion or attitude based on prosodic cues, it might make it easier to fill in the parts of a sentence that was missed, as certain words may be more strongly associated with a specific emotion than others (e.g. a speaker with a happy intonation might be more likely to use the word "glad" than the word "bad", and if noise interfered with the identification of the initial consonants, the emotion might help the listener to guess which of the two words was produced).

If the hypothesis regarding the noise immunity of prosody in NH listeners should be supported by the data of the present work, it could mean that the difficulty that CI users have with prosody perception (see Tables 2.3 and 2.4) may play a part in the difficulty they experience with speech perception in noise. Should this be the case, increasing CI users' access to prosodic cues might be a useful way of improving their perception of speech in noise, and new developments in speech processing algorithms should take this into account.

1.3 RESEARCH QUESTIONS

A number of research questions were formulated to address the problem expressed in the problem statement. Firstly, are NH listeners better at perceiving prosody on sentence level than at recognising words in a sentence in background noise? Although some evidence in literature (e.g. Mattys, 2004; Smith *et al.*, 1989) seems to suggest that prosody is a relatively robust and redundant speech feature, a direct comparison between the perception of prosody and other speech elements (such as phonemes or words) is needed. The perception of words in meaningful sentences is frequently used as a measure of speech recognition in both quiet and noise (Mendel and Danhauer, 1997; Soli and Wong, 2008), and comparing prosody perception with this basic speech recognition measure provided a useful starting point to indicate the noise immunity of prosody. If it turned out that NH listeners are better at perceiving prosody than words in noise, it would underscore the importance of giving CI users better access to prosodic cues in order to improve speech recognition in noise.

Secondly, are NH listeners better at perceiving prosody on a single-word level than at recognising vowels in single words in background noise? Single-word utterances provide a means of testing perception of segmental information (vowels, in this case) without including many of the built-in prosodic cues that occur on sentence level even in neutral utterances (such as rhythm, word stress, and juncture). Vowels were considered particularly important to test, given the work of Kewley-Port, Burkle and Lee (2007), which showed that vowels have a greater impact on speech intelligibility than consonants. Answering this question for NH listeners provided a baseline against which the performance of CI users could be compared.

Thirdly, are CI listeners better at perceiving prosody on a single-word level than at recognising vowels in single words in background noise? Direct comparisons between perception of prosody and perception of segmental information in CI users in existing literature are scarce, and a comparison between these two skills in noise has, to the best of the researcher’s knowledge, not been documented. Understanding which of these two tasks is more difficult for CI users in noise could provide further insight into the difficulty these listeners have in noise, and which speech cues they need to access better in order to improve their recognition of speech in noise.

Fourthly, is the slope of deterioration of prosody recognition shallower with increasing levels of background noise than the slope of phoneme recognition in NH listeners? If prosody recognition is less affected by increasing levels of noise than phonemes, it could indicate that prosody may be an important speech feature with a high degree of noise immunity, which assists NH listeners with speech recognition in noise. If this is the case, it underscores the importance of conveying prosodic cues to CI users in order to enhance their speech perception in noise.

Fifthly, is the slope of deterioration of prosody recognition shallower with increasing levels of background noise than the slope of phoneme recognition in CI recipients? Comparing the outcomes of CI users on prosody and phoneme recognition in noise to that of NH listeners could help to provide an explanation for the difficulty that CI users have with speech recognition in noise, and could indicate which speech cues are especially vulnerable to noise effects in this population.

1.4 APPROACH AND OBJECTIVES

In order to address the research questions as specified in the previous section, a number of perceptual (listening) experiments were conducted. Listening experiments were considered an appropriate approach to address these questions, as it provided a direct means to assess the perception of the speech features under investigation in the populations of interest (NH and CI listeners). A possible alternative to listening experiments as an approach is to conduct acoustic analyses of recorded speech, embed the recorded speech in different levels of noise and compare the resulting spectra, as was done by Parikh and Loizou (2005) with consonants and vowels. This method allows a visualisation of the effects of noise on the spectra of the speech recordings. However, while this method may be useful to see, for instance, the effect of noise on formant frequencies, it is not a direct measure of perceptual effects. Listening experiments, on the other hand, provide a more direct measure of the perceptual effects of noise on recorded speech and were therefore selected as the method of choice in the present work.

Three listening experiments were conducted. Since there are inadequate data in existing literature on the perception of prosody in noise by NH listeners, the first experiment was conducted using NH listeners as participants. In this experiment, word recognition in sentence context was compared to the recognition of prosody on sentence level. The prosodic contrast that was selected for this experiment was a linguistic prosody pattern, which offered participants two possible options to choose from in the listening task (a two-alternatives forced-choice or 2AFC test paradigm) while the word recognition task was an open set speech recognition task.

The second experiment included CI recipients as listeners, and compared the data measured from these participants to those of NH listeners. This experiment included both linguistic and attitudinal prosodic patterns, also cast in a 2AFC test paradigm. In this experiment, however, prosody recognition was compared to vowel recognition, which was also presented in a 2AFC task to equate the difficulty of the prosody and segmental recognition tasks. An adaptive test procedure was used to measure recognition in noise in order to avoid floor and ceiling effects.

The third experiment also included both NH listeners and CI users as participants. In this experiment, vowel and consonant recognition was compared to the recognition of emotional prosody in sentence-level utterances. This time, a 4AFC test paradigm was used for both the phoneme and prosody tasks and fixed SNRs were used instead of an adaptive noise procedure. The use of fixed SNRs reduced testing time per listener, as measurements did not need to be repeated as many times as with the adaptive procedure.

The research approach followed to answer the research questions posed in section 1.3 was guided by a main objective and a number of sub-aims. The main objective of the present study was to compare the relative noise immunity of prosody and segmental information in NH and CI listeners. To achieve this objective, the following sub-aims were formulated.

1. Suitable pre-recorded speech materials were developed to explore each of the research questions. Acoustic analyses of the developed materials were conducted and SWN noise was added to materials that were used in noise experiments.

2. Perception of prosody was tested on sentence level in noise in NH listeners and compared to their perception of words in a sentence at the same noise levels.
3. Perception of prosody on a single-word level was tested in noise in NH listeners and compared to their perception of vowels in noise.
4. Perception of prosody on a single-word level was tested in noise in CI recipients and compared to their perception of vowels in noise.
5. Perception of emotional prosody on sentence level was tested in noise in NH listeners and compared to their perception of phonemes (vowels and consonants).
6. Perception of emotional prosody on sentence level was tested in noise in CI recipients and compared to their perception of phonemes (vowels and consonants).

1.5 CONTRIBUTION

On the whole, this study aimed to expand existing knowledge on the perception of prosody, specifically with regard to how well prosodic cues are perceived by NH and CI listeners in the presence of interfering noise. The main contributions are summarised below. Details about these contributions are included in further chapters.

Firstly, the study demonstrated that in SWN, NH listeners perform significantly better with the perception of a prosodic contrast that occurs on sentence level (in a 2AFC test paradigm) than with open set recognition of words in a sentence. However, the second experiment showed that, on single-word level, NH listeners perform significantly better with vowel recognition in a 2AFC paradigm than with prosody recognition in the same test paradigm in SWN. The same finding was made in CI recipients. Results from the third experiment indicated that when phoneme and prosody perception are tested on sentence level in quiet, without semantic clues and in the same test paradigm, emotional prosody perception is significantly more difficult than vowel and consonant perception for both NH and CI listeners. It was found that in increasingly poor SNRs (using SWN), the recognition of emotional prosody and phonemes shows a similar deterioration slope in NH listeners, while in CI listeners, consonant recognition shows a steeper deterioration slope than vowel and prosody recognition.

In addition to these primary contributions, the development of the following speech materials needed for the perceptual experiments yielded valuable resources for future investigations.

1. Sentence materials with identical semantic content, but differing in terms of prosodic realisation, were developed. The sentences communicated permission or agreement, and prosodic cues were used to indicate whether the permission was unconditional (without reservations) or conditional. These sentences were recorded from eight speakers (four female), and the acoustic characteristics thereof were analysed and documented.
2. Single-word (bi-syllabic) materials with identical semantic content, but differing in terms of prosody, were developed to convey certain (baseline) versus hesitant (reluctant) permission, reflecting an attitudinal function of prosody. These were recorded from eight speakers (four female), and the acoustic characteristics of the recorded materials were analysed and documented.
3. Single-word (bi-syllabic) materials with identical content but with prosodic differences defining each utterance as either a question or a statement were recorded from four speakers (two female), reflecting a linguistic function of prosody, and the acoustic characteristics thereof were analysed and documented.
4. Consonant-vowel-consonant materials (pVOWELt) were developed, including 15 vowels commonly used in the test language (Afrikaans). These were recorded from four speakers (two female), and acoustically analysed.
5. "Jabberwocky" sentences that can be used to evaluate the perception of emotional prosody without interference from semantic content were developed and recorded from two speakers (one female) with acting experience. Jabberwocky refers to sentences of which content words (nouns, verbs, adjectives and adverbs) are replaced with words that consist of

phonemes and phoneme combinations that occur in the test language, but that do not have any meaning, while function words (e.g. "the", "a", "in", "is" etc.) are preserved (Pannekamp, Toepel, Alter, Hahne and Friederici, 2005; Silva-Pereyra, Conboy, Klarman and Kuhl, 2007; Yamada and Neville, 2007).

6. Sentences that can be used to assess vowel and consonant perception in a pVOWELt context, without semantic clues in the sentence as to the identity of the target phoneme/word, were developed and recorded from two speakers (one female).

1.6 OVERVIEW OF THE STUDY

The study is described in the following chapters. A short overview of each chapter's main content is provided here.

Chapter 2 consists of a review of the relevant literature. It discusses the challenge that speech perception in noise poses to CI listeners, the speech cues that support speech recognition in noise in NH listeners, as well as the shortage of available literature on the perception of prosody in noise. From existing literature a definition of prosody is provided and the importance of prosody in spoken communication is illustrated. This chapter also reviews findings from existing literature that suggest the redundancy and noise immunity of prosodic cues and existing data showing the difficulty that CI recipients experience with prosody recognition are presented and discussed. The test language of the present work is Afrikaans, and Chapter 2 ends with background information on this language.

Chapter 3 describes the first listening experiment and its outcome. In this listening experiment, the ability of NH listeners to perceive a specific prosodic pattern[2] that occurs on sentence level was measured at different SNRs and compared to their ability to recognise words in a sentence at the same SNRs. Results indicated that while prosody recognition remained virtually unchanged with increasingly poorer SNRs, word recognition deteriorated rapidly. A limitation of this experiment was that

[2] The term "prosodic pattern" is used in this study to refer to a combination of prosodic cues that are produced to convey a specific meaning, such as the emotion of speakers, or whether their utterance is a question or statement.

the prosody recognition task offered listeners only two alternatives to choose from (a 2AFC test paradigm), while word recognition was tested in an open set format, which might have had a confounding influence on the findings. For this reason, the following experiments were designed in such a way that prosody and phoneme recognition could be compared in identical test paradigms.

Chapters 4 and 5 focus on the second and third listening experiments that were conducted. **Chapter 4** describes the development of test materials for the second experiment, as well as the method and results for this experiment. During this listening experiment, two prosodic contrasts (certain versus hesitant and question versus statement) were used, both occurring on a single-word level. This was compared to vowel recognition, also on a single-word level. Both NH and CI listeners participated in this experiment, and perception was tested in quiet and in an adaptive noise condition using SWN; vowel perception was shown to be significantly better than prosody perception in both listener groups and listening conditions. Since this experiment compared perception on a single-word level, the next experiment (described in **Chapter 5**) was designed to again compare prosody and phoneme perception, this time using sentence-length utterances. The recognition of emotional prosody was compared to vowel and consonant recognition in NH and CI listeners, in quiet and at fixed SNRs using SWN. The findings of this experiment confirmed that in quiet, prosody perception is a more difficult task than phoneme perception (for both listener groups), and showed that in NH listeners, prosody perception and phoneme perception showed similar slopes of deterioration with deteriorating SNR. In CI recipients, consonant recognition had a significantly steeper deterioration slope than both vowel and prosody recognition.

Chapter 6 is a unified discussion of the findings of all the listening experiments. Conclusions are drawn from these findings and these are placed into context with existing literature. The chapter also discusses the limitations of the study and possible future work that could follow from the present study.

CHAPTER 2 LITERATURE STUDY

2.1 CHAPTER OBJECTIVES

This chapter provides an overview of the literature that constitutes the theoretical framework for the study. Firstly, speech recognition in noise in NH and CI listeners is discussed with specific reference to the underlying cues that support speech perception in noise. A discussion on the definition of prosody and the important role that prosody plays in spoken communication is included. The lack of research on prosody in noise is discussed, along with the limited existing evidence suggesting the relative noise robustness of prosody as compared to other speech features. The difficulties of CI recipients with the perception of prosody are illustrated from the literature. Finally, background information on the language that was used as a vehicle for the study (Afrikaans) is provided.

2.2 SPEECH PERCEPTION IN NOISE: CI RECIPIENTS AND NH LISTENERS

CIs provide listeners with permanent hearing loss who are unable to get sufficient benefit from conventional hearing aids with remarkably restored hearing. A CI is a prosthetic device that produces auditory sensations in the implanted individual through electrical stimulation of the auditory nerve. Auditory stimuli are picked up by a microphone (usually worn behind the ear) and converted to electrical signals by a signal processor. These signals are transmitted through the skin of the recipient via radio waves to an array of electrodes implanted in the cochlea, which then stimulates the auditory nerve fibres (Clark, 2003; Loizou, 1999). With current implant technology and speech processing algorithms, CI users have excellent speech perception in quiet, and many CI recipients are able to attain 100% on an open set sentence recognition task in quiet (Caposecco *et al.*, 2012; Gifford, Dorman, Shallop and Sydlowski, 2010).

Unfortunately, a great deal of verbal communication occurs not in quiet but in the presence of varying degrees of background noise. Noise, which can be defined as unwanted sound, can disrupt verbal communication (Moudon, 2009). For successful perception of complicated speech messages by listeners with normal hearing, an SNR of +15 dB (A-weighted) or better has been recommended (World Health Organization, 2000), with the implication that indoors, depending on the size of the room, maximum background noise levels of only 27-34 dBA are acceptable (Bradley, 1986). However, recent findings from 73 000 person-hours of noise exposure (measured using noise dosimeters worn by 286 individuals in a variety of occupational and social settings) have shown that in a typical day, a listener's average eight-hour noise exposure level (in equivalent continuous levels) was 76 dBA (Flamme, Stephenson, Deiters, Tatro, VanGessel, Geda, Wyllys and McGregor, 2012). It should be added that there is a great deal of variability both within and between subjects in terms of typical daily exposure, and that the reported noise levels included different types of noise, not all of which would have caused the same degree of interference with communication. Despite these considerations, it still seems from these findings that a great deal of daily communication occurs in situations where the noise level does not allow perfect speech intelligibility, even for NH listeners.

CI users have significant difficulty with speech perception in noise. A number of studies show that these listeners require a much more favourable SNR to reach the same level of speech recognition as NH listeners. A recent report on listeners using commercially available CI processors shows that CI recipients reach 50% speech recognition at -0.6 dB SNR (standard deviation = 3.7 dB) on a sentences-in-noise test for which NH listeners reach 50% speech recognition at -8.5 dB SNR (standard deviation = 1.5 dB) (Qazi, van Dijk, Moonen and Wouters, 2013). Another study using results from the Hearing in Noise Test (HINT) indicated that CI recipients (using the Nucleus Freedom implant) obtained an average of 64% recognition in noise at an SNR of +10 dB (Balkany, Hodges, Menapace, Hazard, Driscoll, Gantz, Kelsall, Luxford, McMenomy, Neely, Peters, Pillsbury, Roberson, Schramm, Telian, Waltzman, Westerberg and Payne, 2007), which compares poorly to NH listeners who are able to reach 50% recognition at -2.6 dB SNR in the same test (Soli and Wong, 2008).

In contrast to CI users, NH listeners are remarkably successful at speech perception, even in adverse listening conditions such as background noise (Assmann and

Summerfield, 2004). Measurements of sentence recognition in SWN using the HINT and adaptations thereof in different languages have shown that NH listeners achieve 50% correct recognition at an average SNR of -3.9 dB (standard deviation 0.8 dB) when the speech and noise originate from the same direction (Soli and Wong, 2008). Listening to single words in multi-talker babble noise, NH listeners are able to achieve an average of 92.5% correct at an SNR of 9 dB as measured with the Speech Recognition in Noise Test and 50% correct at an SNR of 2.7 dB as measured with the Words in Noise test (Wilson and Cates, 2008).

Understanding which speech features underlie the success of NH listeners' speech perception in noise could provide important insights to those who seek to improve speech recognition in noise in CI recipients. Knowing which speech features remain available to NH listeners in background noise could help to indicate which cues should be provided to CI users in order to improve their speech perception in background noise. To date, a fair amount of research has been conducted to investigate the effects of noise on specific speech cues as perceived by NH listeners. Segmental features (vowels and consonants) in particular have received a great deal of attention in the literature. A number of studies have investigated the acoustic cues that enable NH listeners to identify vowels even in mild to severe background noise (see for example Ferguson, 2004; Ferguson and Kewley-Port, 2002; Swanepoel, Oosthuizen and Hanekom, 2012). Formant frequencies, especially of the first two formants (F1 and F2) (Liu and Kewley-Port, 2004; Nearey, 1989; Peterson and Barney, 1952), the properties of the spectral shape as a whole (Parikh and Loizou, 2005), as well as formant movement and duration (Iverson, Smith and Evans, 2006), have all been shown to contribute to successful perception of vowels. Other studies have explored the underlying cues of consonants and how these cues are affected by noise. For instance, a classic study by Miller and Nicely (1955) reported that voicing, nasality, affrication, duration, and place of articulation are all important distinctive features of consonants and that some of these features (voicing and nasality) are more resistant to the effects of white noise than others. These results have since been reproduced using computerised measures (Phatak, Lovitt and Allen, 2008), while other studies have examined how SWN affects consonant identification (Phatak and

Allen, 2007; Woods *et al.*, 2010) or compared the effects of different noise types on consonant perception (Broersma and Scharenborg, 2010). Some researchers have hypothesised that the noise immunity of consonants may be related to the rapid spectral changes that characterise them, because of a form of auditory enhancement of such changes in the peripheral or central auditory system (Assmann and Summerfield, 2004; Summerfield, Sidwell and Nelson, 1987).

Phonemes, however, are not the only important pieces of information that listeners require to perceive a speaker's message accurately. Suprasegmental features, also called prosody, also play a vital role in spoken communication and fulfil a variety of important communicative functions (as described in the next section). Despite its important role in spoken communication, the effects of noise on the perception of prosody have not been thoroughly explored in existing literature. To broaden researchers' understanding of the robustness of NH listeners' speech perception in noise, it is important that the availability of prosodic cues in noise be investigated. A comparison between the effects of noise on the recognition of different speech features, such as vowels, consonants and prosody, could provide insight into which cues NH listeners rely on to obtain the remarkable degree of success they do with speech perception in noise, which in turn could be used to inform efforts to improve speech recognition in noise in CI recipients. The following sections will provide a definition of prosody and an overview of its many communicative functions, and subsequently discuss the limited amount of data that is currently available on the perception of prosody in noise.

2.3 PROSODY: DEFINITIONS AND FUNCTIONS

The term "prosodic quality of speech" was coined in 1947 by a physician in an attempt to describe the symptoms of a particular patient who had suffered a traumatic brain injury (Monrad-Krohn, 1947). The patient exhibited abnormalities in speech intonation (the rising and falling of voice pitch) and stress or emphasis, which resulted in her speech sounding as if she had a foreign accent. Since the patient's musical faculties appeared to be intact, Dr Monrad-Krohn felt that "melody of language" was not the right terminology to use, preferring the term "prosodic quality of speech" and defining the "prosodic faculty" as the faculty regulating the correct use of pitch and stress (emphasis).

The precise definition of prosody and terms considered to be "synonymous" with it remains a complicated issue. Ladd and Cutler (1983) identify two distinct approaches to the definition and investigation of prosody, namely a concrete approach and an abstract approach. According to the concrete approach to its definition, prosody refers to phenomena related to the acoustic parameters of pitch (roughly correlated with fundamental voice frequency or F0), duration or tempo, loudness or intensity, and pauses (Cutler, Dahan and Van Donselaar, 1997; Ladd and Cutler, 1983). These features of speech co-occur with the segmental features of speech that mark the differences between different phonemes, and could be seen as "a secondary, overlaid function" of segmental features (Lehiste, 1970). For example, while voicing on a single segment serves to distinguish one segment from another (for instance, the difference between /p/ and /b/), changes in the F0 of the voice across one or more segments are perceived as a particular tone or intonation pattern, which could influence the perceived content of the message (Cruttenden, 1997; Lehiste, 1970). The abstract approach to the definition of prosody views prosody as "phenomena that involve phonological organization at levels above the segment" (Ladd and Cutler, 1983). Although the abstract approach certainly has value, the present study used the concrete approach, since the focus of the study was on acoustic characteristics of prosody and the effects of noise and CIs on these characteristics. In line with the concrete approach, the term prosody can be considered synonymous to "suprasegmental features" of speech (Cutler *et al.*, 1997). The present study uses the term prosody throughout, and considers it a synonym to suprasegmental features.

A number of speech features are considered prosodic features. Different authors use different terms for similar speech features, as illustrated in Table 2.1. As illustrated in this table, terms such as rhythm, length, tempo, pause, quantity and juncture are all used to refer to durational aspects of speech (i.e. can be measured in the time domain). Intonation, tone and pitch are terms used to describe changes that occur in the frequency domain. Perceived changes in voice pitch can be roughly correlated to the F0 of the voice (Borden, Harris and Raphael, 2007). Features that are related to the amplitude or intensity of the speech signal are sometimes called loudness, and sometimes stress. However, what is perceived by listeners as stress (emphasis of or accent on a particular syllable or word) is marked by changes in not only intensity, but also in duration and voice pitch (Cruttenden, 1997).

Table 2.1: Prosodic features of speech

Duration features	**Frequency features**	**Intensity/ prominence features**	**Authors**
Rhythm	Intonation	Stress	(Grant and Walden, 1996)
Length, tempo, pause	Pitch	Loudness	(Cruttenden, 1997)
Duration, quantity, tempo	Pitch, tone, intonation	Stress and emphasis	(Lehiste, 1976)
Juncture	Intonation	Stress	(Borden *et al.*, 2007)

The different prosodic speech features are used to fulfil a great number of important communicative functions. Table 2.2 provides a summary of the communicative functions of prosody. As shown in Table 2.2, prosody plays a very important role in speech communication by fulfilling both linguistic functions (such as marking boundaries and sentence focus) and non-linguistic functions (such as communicating a speaker's emotion). Many of the prosodic features listed here function only on the level of multi-word utterances, i.e. phrases or sentences (e.g. word or phrase boundaries, resolution of ambiguous sentences, sentence accent). However, some prosodic features can also function on the level of a single word, e.g. marking a stressed syllable (Fry, 1958), lexical tone distinctions (Lehiste, 1976), question/statement contrasts (Chatterjee and Peng, 2008) and marking the emotion or attitude of the speaker (Hammerschmidt and Jürgens, 2007; Van Zyl and Hanekom, 2013b).

Table 2.2: Communicative functions of prosody

Communicative function	**References**
Linguistic functions	
Provides structure to spoken language by marking boundaries between words and phrases	(De Pijper and Sanderman, 1994; Watson and Gibson, 2005)
Facilitates resolution of ambiguity in sentences or phrases	(Millotte, Wales and Christophe, 2007; Price, Ostendorf, Shattuck-Hufnagel and Fong, 1991)
Helps listeners to predict upcoming information in the sentence and/or length of the utterance	(Grosjean, 1983; Grosjean and Hirt, 1996; Snedeker and Trueswell, 2003)
Facilitates turn-taking by indicating finality or continuity of an utterance	(Berkovits, 1984; Caspers, 1998; Thorsen, 1980)
Indicates which syllable in a word is stressed, thereby disambiguating words such as 'object and ob'ject	(Fry, 1955; Fry, 1958)
Indicates the focus of a sentence by marking accented words	(Carlson, 2009; Pell, 2001)
May be used to indicate the meaning of novel words in infant-directed speech	(Nygaard, Herold and Namy, 2009)
Marks the difference between questions and statements in sentences and single-word utterances	(Chatterjee and Peng, 2008; Grant and Walden, 1996)
Natural intonation contour assists with speech recognition in noise	(Binns and Culling, 2007; Laures and Bunton, 2003)
Marks lexical tone distinctions in tonal languages	(Botinis, Granström and Möbius, 2001)
Non-linguistic functions	
Communicates the attitude or emotion of the speaker	(Berckmoes and Vingerhoets, 2004; Mozziconacci, 2001; Pell, 2001; Tomlinson and Fox Tree, 2011; Williams and Stevens, 1972)

2.4 PROSODY IN NOISE

As mentioned briefly in section 2.2, the effects of background noise on the recognition of prosody have not been reported often in existing literature, while the noise immunity of phonemes has been extensively investigated. The limited amount of data that are available on the effects of noise on prosodic features (as reported below) seems to suggest that prosody may be quite robust and immune to the effects of noise, possibly more so than segmental features of speech. Anecdotally, it seems that in some difficult communication situations (such as in noise, or when the speaker is in the next room), it is sometimes easier to hear "how" something is said than to hear exactly what is said. For example, a listener might be able to hear that a question has been asked, but not what the content of the question was, or it might be clear that the speaker is excited or angry about something without being able to hear the content of his or her utterance.

A few findings reported in existing literature seem to support this anecdotal observation to some degree. In a study by Smith *et al*. (1989), NH listeners (n = 39) were presented with English phrases in white noise at an SNR of -10 dB, an SNR that did not allow recognition of segmental features. Listeners were required to choose from four possible options which sentence they most likely heard. In actual fact, none of the options they could choose from matched the segmental (phonetic) content of the utterance presented. However, each of the options the listener could choose from was specifically compiled either to match or mismatch the rhythm (sequence of weak and strong syllables) and/or word boundaries (number of syllables per word) of the target utterance. The findings indicated that listeners were highly sensitive to the stress rhythm of utterances, and selected the option that was not a rhythmic match to the target only 18% of the time. They were also able to derive some word boundary information, selecting the right locations of boundaries significantly more often than expected by chance. It seems from these findings that in adverse listening conditions where segmental information was no longer available, the prosodic features related to word boundaries and rhythm were still available and useful to NH listeners. More recent work (Mattys, 2004; Mattys, White and Melhorn, 2005) has also indicated that in adverse listening conditions, stress (a prosodic feature) was a more important cue to word boundaries than acoustic-phonetic cues, because of the degradation of acoustic-phonetic cues by noise. These researchers suggested that stress may be particularly tolerant to signal degradation.

In addition to the work on word boundaries, a number of studies have shown that the natural intonation contour (which correlates with the contour of the speaker's voice F0) plays an important role in speech recognition in noise. Laures and Weismer (1999) investigated the effect of a flattened F0 contour on speech intelligibility in an attempt to explain the poor intelligibility of speakers with dysarthria. Sentences from the Speech Perception in Noise test (Kalikow, Stevens and Elliot, 1977) were presented to NH listeners in white noise, at an SNR of 4-5 dB. There were two versions of each sentence – one with a normal (original) intonation contour, and the other with a flattened F0 contour. Results indicated that flattening the F0 contour had a significant effect on the intelligibility of the sentence materials. A similar study by

Laures and Bunton (2003), this time including both a white noise and a multi-talker babble noise condition, also found that flattening the F0 contour significantly decreased intelligibility. In contrast, Binns and Culling (2007) found that a flattened F0 contour did not have a significant effect on intelligibility in speech-shaped noise, while an inverted F0 contour did affect intelligibility in this listening condition significantly. The same finding was made when the background noise was a single interfering talker. Although none of these studies compared the effect of background noise on the F0 contour and its effect on other speech cues, the findings suggest that the F0 contour has enough noise immunity to enable it to play an important role in speech recognition in different types of noise.

It appears from all the findings referred to above that prosody is supported by an acoustically rich set of cues that may have a high degree of noise immunity. However, it cannot be assumed that prosody is more immune to the effects of noise than segmental speech features until the two types of speech features have been directly compared in perceptual experiments on the same group of listeners. The present study attempted to address this question by comparing the relative noise immunity of a number of different prosodic patterns to that of words and phonemes. It should be noted that the existing literature suggesting the noise robustness of prosody as discussed above all refer to sentence-level prosody. As mentioned in section 2.3, many prosodic features can also occur on a lexical (single-word) level. To ensure comprehensive investigation of the issue, the present work included both sentence-level and word-level prosodic realisations in the assessment of noise immunity. Ultimately, the goal of answering this question would be to have a better understanding of which speech cues and features support NH listeners' successful speech perception in noise in order to motivate future efforts to provide these cues to CI users.

Besides the possibility that prosody might be particularly noise-immune, the investigation of prosody in noise and the comparison of its perception to the perception of phonemes and words was motivated by another consideration. This second consideration was that CI recipients not only have difficulty with speech recognition in noise (as discussed earlier in this chapter) but are also reported to have problems with the perception of prosody, although this ability is rarely compared directly to their perception of other speech features, which makes it difficult to determine the degree of the problem. If prosody consists of a particularly difficult set of cues for CI recipients to perceive, and a particularly noise-immune set of cues that aids NH listeners in speech perception in noise, this might provide important insight into the difficulty that these listeners have in noise. The following section provides an overview of the perception of prosody by CI recipients.

2.5 PROSODY PERCEPTION IN CI RECIPIENTS

The communicative functions of prosody can generally be divided into linguistic and non-linguistic functions (see Table 2.2). Table 2.3 provides an overview of a number of studies that have reported on the ability of CI users to perceive a number of prosodic patterns with specific linguistic functions, while Table 2.4 summarises reports on the perception of emotional prosody (a non-linguistic function of prosody). To enable fair comparisons between results from different studies, scores in both tables were corrected for guessing by using the equation of Boothroyd (1988),

$$S_c = (S_u - S_g) / (100 - S_g) \text{ x } 100, \qquad (2.1)$$

where S_c is the corrected score percentage, S_u is the uncorrected score percentage, and S_g is the percentage score expected from guessing (e.g. 50% if using a 2AFC test paradigm).

Table 2.3: Overview of reported findings on CI users' recognition of linguistic prosody. AFC = alternatives forced-choice; SPAC = Speech Pattern Contrast battery; MAC = Minimal Auditory Capabilities battery. All implants in these reports were unilateral.

Prosodic function	Language	Listener details	Processing strategy	Test paradigm	Results (corrected for guessing)	Reference
Sentence stress/ marking accented word in a sentence	German	CI (n = 12): aged 38-75 NH (n = 12): aged 34-68	HiRes-P (1) HiRes-S (1) ACE (3) CIS+ (6) FSP (1)	3AFC	CI: 59.5 (+/- 21.8)% NH 97.0 (+/- 4.6)% (significant difference between NH and CI)	(Meister *et al.*, 2009)
	English	CI (n = 6): aged 47-73	MPEAK	2AFC	SPAC: 82% MAC: 66%	(Richardson, Busby, Blamey and Clark, 1998)
	Hebrew	CI (n = 23): aged 17-65	ACE (19) CIS+ (2) CIS (1) HiRes (1)	3AFC	63.61 (+/- 22.82) %	(Most *et al.*, 2012)
	Hebrew	CI (n = 10): aged 8-15	ACE	3AFC	72.82 (+/- 15.8) %	(Most and Peled, 2007)
	English	CI (n = 16): aged 26-85	F0-F2 (9) F0-F1-F2 (7)	2AFC	SPAC 48%	(Waltzman and Hochberg, 1990)
Syllable stress/ accent	Hebrew	CI (n = 10): aged 8-15	ACE	2AFC	20.83 (+/- 27.56) %	(Most and Peled, 2007)
	Hebrew	CI (n = 23): aged 17-65	ACE (19) CIS+ (2) CIS (1) HiRes (1)	2AFC	54.67 (+/- 24.01) %	(Most *et al.*, 2012)

Table 2.3 (continued)

Prosodic function	**Language**	**Listener details**	**Processing strategy**	**Test paradigm**	**Results (corrected for guessing)**	**Reference**
Question/ statement distinction	English	CI (n = 16): aged 26-85	F0-F2 (9) F0-F1-F2 (7)	2AFC	SPAC: 38%	(Waltzman and Hochberg, 1990)
	English	CI (n = 9): aged 45-75	MPS (5) SAS (2) CIS (1)	2AFC	Male speaker: 38.6% Female speaker: 35.8%	(Green, Faulkner, Rosen and Macherey, 2005)
	Hebrew	CI (n = 10): aged 8-15	ACE	2AFC	42.5 (+/- 27.55) %	(Most and Peled, 2007)
	Hebrew	CI (n = 23): aged 17-65	ACE (19) CIS+ (2) CIS (1) HiRes (1)	2AFC	62 (+/-20.7) %	(Most *et al.*, 2012)
	German	CI (n = 12): aged 38-75 NH (n = 12): aged 34-68	HiRes-P (1) HiRes-S (1) ACE (3) CIS+ (6) FSP (1)	2AFC	CI: 64 +/- 10.7% NH: 98 +/- 2.0% (significant difference between CI and NH)	(Meister *et al.*, 2009)
	English	CI (n = 26) aged 7-20 NH (n = 17) aged 6-20	SPEAK (15) ACE (11)	2AFC	CI: 40.26 (+/- 14.5)% NH: 94 (+/- 4)% (significant difference between CI and NH)	(Peng, Tomblin and Turner, 2008)
	English	CI (n = 6): aged 47-73	MPEAK	2AFC 2AFC	SPAC: 76% MAC: 72%	(Richardson *et al.*, 1998)
Enhance speech recognition in noise	German	CI (n = 18): aged 19-81 NH (n = 13): aged 18-73	Not reported	Open set	SRT in noise advantage (normal vs. inverted F0): CI = 1.2 dB; NH: 2.1 dB	(Meister, Landwehr, Pyschny and Grugel, 2011)

Table 2.4: Overview of reported findings on CI users' recognition of emotional prosody. AFC = alternatives forced-choice. All implants in these reports were unilateral, except in the case of Cullington and Zeng (2011), where all users had bilateral implants.

Language	**Listener details**	**SP strategy**	**Speech material, speaker & test paradigm**	**Emotions**	**Results (corrected for guessing)**		**Reference**
					NH	**CI**	
English	CI (n = 17) NH (n = 18) Ages not reported	Not reported	Semantically neutral sentence 1 female speaker 4AFC	Angry		30.67	(House, 1994)
				Happy		18.67	
				Sad		29.33	
				Neutral		21.33	
				(total)	**97.33**	**25.33**	
Not reported	CI (n = 20): aged 33-79	SPEAK	Semantically neutral sentence & number 2 actors (1 female) 4AFC	Hot anger	84.00	61.33	(Pereira, 2000)
				Cold anger	65.33	4.00	
				Happy	77.33	17.33	
				Sad	93.33	41.33	
				Neutral	70.67	49.33	
				(total)	**78.67**	**34.67**	
				Amplitude normalised:			
				Hot anger	78.67	8.00	
				Cold anger	72.00	18.67	
				Happy	81.33	14.67	
				Sad	94.67	10.67	
				Neutral	70.67	33.33	
				(total)	**80.00**	**17.33**	
English	CI (n = 18): aged 7-13 NH (n = 18): aged 7-13	Not reported	Semantically neutral sentence Unspecified speaker 4AFC	Angry	86.67	51.11	(Hopyan-Misakyan, Gordon, Dennis and Papsin, 2009)
				Happy	68.89	31.11	
				Sad	77.77	33.33	
				Fearful	53.33	15.56	
				(total)	**71.67**	**33.89**	

Table 2.4 (continued). * Results for NH listeners in Cullington & Zeng (2011) were estimated from Fig. 5.

Language	Listener details	SP strategy	Speech material, Speaker & test paradigm	Emotions	Results (corrected for guessing)		Reference
					NH	CI	
English	CI (n = 8): aged 41-73 NH (n = 8): aged 22-40	ACE (3) SPEAK (5)	10 semantically neutral sentences (3 questions) 2 actors (1 female) 5AFC	Angry		29.75	(Luo, Fu and Galvin III, 2007)
				Happy		5.50	
				Sad		56.63	
				Neutral		57.00	
				Anxious		6.63	
				(total)	**87.25**	**31.13**	
				Amplitude normalised:			
				Angry		13.25	
				Happy		3.88	
				Sad		43.00	
				Neutral		44.88	
				Anxious		3.88	
				(total)	**83.88**	**21.75**	
Not reported	CI (n = 20): aged 10-17 (10 implanted before 6, 10 implanted after 6)	Not reported	1 semantically neutral sentence 1 professional actor 6AFC	Angry	61.70	40.94	(Most and Aviner, 2009)
				Happy	61.82	18.01	
				Sad	61.70	29.11	
				Fearful	38.06	3.96	
				Disgust	54.50	17.95	
				Surprise	11.04	3.00	
				(total)	**50.65**	**18.80**	
English	CI (n = 13): aged 38-75 NH (n = 27): aged 60-69	SPEAK (7) ACE (7) CIS (7) Hi-Res (3) MPS (2)	1 semantically neutral sentence 1 male speaker 6AFC	Angry			(Cullington and Zeng, 2011)
				Happy			
				Sad			
				Neutral			
				Disinterested			
				Surprised			
				(total)	**±83***	**62.69**	

Table 2.4 (continued)

Language	**Listener details**	**SP strategy**	**Speech material, Speaker & test paradigm**	**Emotions**	**Results (corrected for guessing)**		**Reference**
					NH	**CI**	
Hebrew	CI (n = 25): aged 15-67	Not reported	Nonsense utterance (3 2-syllable words) 1 female actress 4AFC	Angry		40.44	(Most *et al.*, 2012)
				Happy		22.66	
				Sad		52.00	
				Fearful		20.88	
				(total)		33.55	
Japanese	CI (n = 18): aged 5-13	ACE (12) SPEAK (4) SAS (2)	4 semantically neutral sentences, 1 female speaker 3AFC	***Amplitude normalised:***			(Nakata, Trehub and Kanda, 2012)
				Angry	100.00	8.50	
				Happy	100.00	41.50	
				Sad	100.00	46.00	
				(total)	100.00	32.00	

The data represented in Tables 2.3 and 2.4 clearly indicate that CI users had considerable difficulty with the recognition of most of the prosodic functions represented in these data. Although many of the studies on linguistic prosody did not include a group of NH control listeners, those that did reported recognition scores in the NH group between 90 and 100% for sentence accent and question/statement distinctions, which was significantly better than the performance of CI users in those studies (Meister *et al.*, 2009; Peng *et al.*, 2008). Studies on emotional prosody perception (Table 2.4) also showed substantial differences between NH and CI performance on the recognition of vocal emotion.

The difficulty that CI users experience in perceiving certain prosodic features may be examined in the light of the acoustic correlates of these features. The acoustic characteristics of stress or emphasis on a single-word level (syllable stress) or in a sentence (sentence accent or stress) are reported to be an increase in voice pitch, greater duration and greater intensity (Borden *et al.*, 2007; Cruttenden, 1997). The acoustic cues that differentiate questions from statements are reported to be a rising intonation pattern (or at least the use of higher pitch somewhere in the utterance) for questions (Borden *et al.*, 2007; Cruttenden, 1997; Ponelis, 1979; Thorsen, 1980), and a higher speech rate (Van Heuven and Van Zanten, 2005). The advantage of a natural intonation contour for speech intelligibility in noise (Laures and Bunton, 2003; Meister *et al.*, 2011) is also closely related to movements in voice pitch. The acoustic correlates of emotional prosody are related to changes in average F0 (which roughly corresponds to perceived voice pitch), F0 range, variability and contour, speech rate and high-frequency energy (Banse and Scherer, 1996). Looking at these acoustic correlates, the perception of voice pitch, which largely depends on perception of the speaker's voice F0 (Borden *et al.*, 2007), seems to play an important role in accurate perception of all of these prosodic patterns. Unfortunately, CI users seem to have great difficulty with pitch perception. In NH listeners, voice pitch is perceived through both the spectral resolution of low-frequency harmonics ("place cues", referring to stimulation of a specific area of the basilar membrane), and the resolution of the temporal fine structure of the input signal (temporal cues) (Green *et al.*, 2005; Kong, Stickney and Zeng, 2005). However, CI recipients using most current speech processors have limited access to both place and temporal voice pitch cues, owing to poor spectral resolution (as a result of a limited number of effective frequency channels) and a lack of temporal fine structure information in the signal provided by many processors (Brown and Bacon, 2010; Kong *et al.*, 2005; Qazi *et al.*, 2013; Shannon, Cruz and Galvin III, 2011). Although some of the older processing strategies (the F0-F2 and F0-F1-F2 strategies) explicitly encoded F0 cues (Loizou, 1999), many current processors convey F0 only through temporal modulation, and temporal fine structure cues are discarded (Stickney, Assmann, Chang and Zeng, 2007). Some CI manufacturers have recently introduced processing strategies that are designed to convey some temporal fine structure cues (Med-El's FSP or fine structure processing strategy, and Advanced Bionics' HiRes 120 strategy), but research to determine ways in which to deliver fine structure cues in such a manner that they can be fully utilised by CI users continues (Wilson and Dorman, 2008). While some results have demonstrated small improvements in speech recognition in noise when changing from continuous interleaved sampling+ (CIS+) to FSP (Lorens, Zgoda, Obrycka and Skarzynski, 2010), others showed no significant difference between the two strategies for this task (Qi, Krenmayr, Zhang, Dong, Chen, Schatzer, Zierhofer, Liu and Han, 2012). The difficulties with regard to pitch encoding and perception therefore remain a problem for many CI users and may provide at least a partial explanation for the problems that CI users have with prosody perception.

Besides voice F0, however, other acoustic cues such as speech rate, duration and intensity also support the perception of some prosodic features. These cues could also assist CI listeners with prosody perception. CI users' perception of duration cues (length of utterances and speech rate) are expected to be close to that of NH listeners, as their temporal resolution appears to be near normal according to psychophysical measures (Garadat and Pfingst, 2011; Moore and Glasberg, 1988). According to Rogers, Healy and Montgomery (2006), CI recipients are reportedly less sensitive than NH listeners to changes in intensity of speech stimuli (requiring differences of 3.1 dB to detect a change, whereas NH listeners can detect differences as small as 1.3 dB on average). However, there is large variability among CI users regarding this ability, and some CI listeners in the Rogers *et al.* (2006) study showed sensitivity within the normal range. It appears therefore that CI listeners might be relying heavily on intensity and durational cues to support the limited amount of success they are able to attain with prosody perception. Their reliance on intensity differences, at least for perception of emotional prosody, is supported by some of the findings shown in Table 2.4. The studies conducted by Pereira (2000) and Luo, Fu and Galvin (2007) both included a listening condition in their experiments where they normalised the amplitude of the speech materials, effectively eliminating the intensity cues to emotion. In the first study (Pereira, 2000), amplitude normalisation did not have a significant effect on the performance of NH listeners, while CI listeners' performance was much poorer (reduced from 35% to 17% if corrected for guessing, significance not reported) for amplitude normalised speech. Luo *et al.* (2007) reported that amplitude normalisation significantly reduced emotion perception in both NH and CI listeners, but the difference was also much larger in the CI group (9% versus 3% reduction in NH group, according to corrected scores). These findings suggest that CI recipients rely more heavily on amplitude (intensity) cues to perceive emotional prosody than NH listeners do, perhaps because they have limited access to the voice pitch cues that help to differentiate different emotions. The redundancy in the signal that enables NH listeners to achieve successful perception is not available to CI users to the same degree.

One of the limitations of the studies listed in Tables 2.3 and 2.4 is that almost none of these studies reported on the ability of their CI participants to perceive segmental information (vowels and consonants). Although it is clear from the data in the table that CI listeners had difficulty with prosody perception, other studies have also reported that CI recipients perform significantly worse than NH listeners on vowel and consonant perception (see e.g. Munson, Donaldson, Allen, Collison and Nelson, 2003; Stacey, Raine, O'Donoghue, Tapper, Twomey and Summerfield, 2010). Without a direct comparison of phoneme and prosody perception in the same group of listeners and in similar experimental conditions, however, it is not possible to say which set of features is more problematic for these listeners. The only study listed in Table 2.3 that also reported vowel perception in listeners used a 5AFC test paradigm on some of the listeners, and a 9AFC test paradigm on others (Green *et al.*, 2005) while testing prosody perception in a 2AFC paradigm. In that study, however, the goal was not to compare prosody and phoneme perception directly, but rather to document effects of an experimental processing strategy on prosody perception while monitoring its effects on vowel perception. The use of different test paradigms to assess perception of these two different speech features (prosody and vowels) makes it very difficult to compare perception fairly. It follows from this that listening experiments where CI recipients' perception of phonemes and prosody are directly compared in identical test paradigms could make a valuable contribution to the existing body of knowledge on prosody perception in CI users.

A second limitation of the studies listed in Tables 2.3 and 2.4 is that none of these reports on prosody perception in CI included noisy listening conditions, with the exception of Meister *et al*. (2011). This study, however, examined the effects of a natural versus inverted F0 contour in NH and CI listeners, and did not investigate the effects of background noise on the perception of specific prosodic functions (e.g. question/statement distinctions or vocal emotion). Although it is reasonable to assume that prosody perception will deteriorate with the addition of interfering noise, experimental work is required to investigate the effects of noise on prosody recognition and compare this to perception of other speech features.

The present study aimed to fill some of the research gaps left by the existing studies on prosody perception in NH and CI listeners through a series of listening experiments, as specified in the aims of the study (Chapter 1). The language in which all of these listening experiments were conducted is Afrikaans. The following section provides some background information on the test language.

2.6 BACKGROUND ON AFRIKAANS

Afrikaans is a major West Germanic language native to South Africa and similar to Dutch (Gooskens, 2007). In fact, Afrikaans originated from 17th century Dutch (Botha, 1996). In 1652, a group of Dutch officials arrived at the Cape (the south-western tip of Africa) with the aim to establish a refreshment station for an important trade route of their company (the Dutch East India Company) (Ponelis, 1993). Between 1658 and 1700 a large number of slaves arrived in the Cape, mostly from India, Madagascar and Indonesia (Davids, Ferreira, Links and Prinsloo, 1997). Around 1700, 220 people fleeing from religious persecution in France arrived in the Cape and settled in the Dutch community (Stoops, 1995). There were also a large number of Portuguese-speaking slaves in the community, and the settlement was surrounded by local nomadic Khoi people (Ponelis, 1993). A wide variety of languages was therefore represented in the small Cape community. However, since Dutch was the main language of the company that established the settlement in the Cape and the first school was founded in 1658 with the purpose of teaching Dutch to the slaves, Dutch was the transactional language and was spoken as a second language by most people in the community (Ponelis, 1993). The form of Dutch that resulted from its use by non-native speakers in the Cape, and its contact with the many other languages represented there at the time, evolved into the language that is now called Afrikaans, which was recognised as an official language in 1925 (Gooskens, 2007; Ponelis, 1993). Besides influences from Khoi, Portuguese, Malaysian and limited influence from German and French, English has also had a strong influence on Afrikaans on many levels, including vocabulary and stress patterns (Davids *et al.*, 1997; Donaldson, 1991; Stoops, 1995).

According to the latest census conducted in South Africa (in 2011), 13.5% of the country's population of 51 770 560 people use Afrikaans as their first language, which means that Afrikaans is the native language of close to seven million people in South Africa, making it the third most common first language after isiZulu and isiXhosa (Statistics South Africa, 2011). In two of South Africa's provinces (Northern and Western Cape), Afrikaans is the most common native language, and in three other provinces (Free State, Eastern Cape and North West) it is the second most spoken language (Statistics South Africa, 2012). Afrikaans is standardised on the basis of its eastern variety, the variety of Afrikaans spoken in Gauteng, where the present study was conducted (Ponelis, 1993).

There are very few published reports on the acoustic characteristics of phonemes and prosody in Afrikaans. A few reports provide data on the formant frequencies of Afrikaans vowels (Botha, 1996; Pretorius, Hanekom, Van Wieringen and Wouters, 2006; Taylor and Uys, 1988; Van der Merwe, Groenewald, Van Aardt, Tesner and

Grimbeek, 1993). As far as prosody is concerned, Afrikaans is reported to have a similar stress pattern to that of English, owing to its common Germanic background (Donaldson, 1991). A study on bi-syllabic compounds in Afrikaans has shown that the initial syllable is the default position for syllable stress, which is characteristic of Germanic languages (Wissing, 2007). In that study, vowels in stressed syllables were found to be longer in duration, and have higher F0 and greater intensity than in unstressed syllables. Besides these findings on syllable stress, there appears to be no published data on acoustic correlates of prosodic features in Afrikaans. For this reason, the test materials developed for each of the listening experiments in the present study were acoustically analysed and details of acoustic parameters were reported.

2.7 CONCLUSION

This chapter provided an overview of existing literature relevant to the research questions posed in Chapter 1. In summary, speech recognition in noise is particularly problematic for CI listeners, while NH listeners are able to perform surprisingly well on this task. In existing research on the speech cues that underlie successful speech recognition in noise there is a shortage of data on prosody perception in noise. Prosody fulfils many important communicative functions, and some findings in existing literature suggest that prosody might be particularly noise-immune. However, existing reports indicate that CI listeners have a great deal of difficulty with the perception of a variety of prosodic patterns. Listening experiments are needed to explore prosody perception in noise in NH listeners and CI recipients and to compare this to their perception of other speech features in noise. The subsequent chapters provide details on the background, methods, results and implications of the findings of the three sets of listening experiments conducted to answer the research questions formulated in Chapter 1.

CHAPTER 3 SENTENCE-LEVEL WORD RECOGNITION VERSUS PROSODY RECOGNITION IN NH LISTENERS IN NOISE

Parts of this chapter were published as an extended abstract in the Journal of Hearing Science (Van Zyl and Hanekom, 2011)

3.1 CHAPTER OBJECTIVES

This chapter describes the background, methods and results of the first set of listening experiments conducted in this study. This set of experiments was aimed at answering the first research question posed in Chapter 1, i.e. whether NH listeners are better at perceiving prosody on a sentence level than at recognising words in a sentence in background noise. The first two sub-aims specified in Chapter 1 were addressed in this chapter, namely i) the development and acoustic analyses of suitable pre-recorded speech materials for the experiment and ii) testing the perception of prosody on sentence level in noise in NH listeners and comparing this to their perception of words in a sentence at the same noise levels. In the first part of this chapter (section 3.2), a theoretical background to the speech material and experimental work reported here is provided from existing literature. Section 3.3 describes the development and analyses of the test materials by providing details on methods and results and a discussion of the findings. This is followed by section 3.4, which focuses on the listening experiment that compared word and prosody perception in noise, and also includes methods, results and discussion sections.

3.2 BACKGROUND

As discussed in Chapters 1 and 2, understanding which speech features contribute to NH listeners' perception of speech in noise could guide future efforts in improving speech recognition of CI users in noise. To date, however, prosody perception in noise has not been directly compared to the perception of other speech features in noise, and the listening experiment described in this chapter (section 3.4) aimed to address this gap in existing research. The listening experiment aimed to compare the recognition of the prosodic contrast described in section 3.3 of this chapter in background noise to the recognition of words in sentences in NH listeners. Word recognition is frequently used as a measure of speech perception success in audiology (Mendel and Danhauer, 1997), and words in a sentence provide a life-like measure with built-in redundancy to measure how well a listener has perceived the semantic content of an utterance. The experiment described in this chapter involved only NH listeners. It was considered prudent to first explore the possibility that prosody might be more noise immune than other speech features in NH listeners before involving CI recipients. If the hypothesis regarding prosody's noise immunity (see section 1.2, Chapter 1) was contradicted by the findings of this experiment, the possibility that poor prosody recognition has a substantial influence on CI speech recognition in noise would not have been strong enough to warrant further investigation. In addition, the first experiment provided a valuable opportunity to explore a suitable methodology for answering the research questions of the study. Determining what the limitations of this experiment were enabled improved design of the second and third experiments where CI users were involved.

The speech materials developed for the assessment of prosody in the listening experiment were sentences that used prosodic cues to indicate a specific attitude of the speaker and upcoming content of the sentence. The type of sentences developed was based on anecdotal evidence, which suggests that listeners are able to detect when a speaker grants permission or gives approval with hesitation or reluctance. This occurs in situations where, for example, a speaker would give permission for something (e.g. "You may borrow the book"), but reveals some sort of hesitation or reservation in their permission through prosodic cues. Permission given in the form of a sentence may contain the word "but" followed by a condition (e.g. "but not the pen"), but even before or without this condition being uttered, the listener may become aware of the speaker's reluctant attitude, communicated through prosodic cues in the first part of the utterance.

The anecdotal evidence for prosodic cues which communicate speakers' reluctance and their intention to add a condition to their permission or approval is supported by findings in existing literature, which suggest that listeners can use prosodic cues to make predictions about upcoming content, or the upcoming end of an utterance. Experimental work has shown that listeners are able to distinguish "finished" sentences from "unfinished" sentences based on prosodic cues (Berkovits, 1984). In spontaneous conversations, speakers and listeners use prosodic cues to facilitate smooth transitions between speaker turns. Speakers may use a specific accent, falling pitch, or a reduction in loudness to indicate the end of their speaking turn and yield the floor to their conversational partner (Wells and Macfarlane, 1998; Zuraidah and Knowles, 2006), or use a specific type of intonation contour to retain their turn in the conversation (Caspers, 1998). Some research has suggested that listeners may even be able to predict the length of the remaining part of a sentence after hearing the first few words in the sentence, based purely on prosodic cues (Grosjean, 1983; Grosjean and Hirt, 1996). Anecdotally, it seems possible to judge when a speaker is nearing the end of his or her turn in a conversation, and the findings reported by Grosjean might be a reflection of this ability. Furthermore, recent work by Swerts and Hirschberg (2010) has demonstrated that listeners are able to predict whether a speaker is about to convey good or bad news, according to acoustic and prosodic features in the utterance. Speakers can also vary the pitch accent pattern in a sentence according to upcoming information, especially if a word in the first part (or clause) of the sentence is about to be contrasted by a word in the second part of the sentence (Swerts, 2007).

In this study, sentence materials that communicate reluctant or conditional permission or agreement were developed and validated in a group of NH listeners. Compound sentences were used, containing a message of permission, approval or agreement in the first clause, and a condition to the permission, approval or agreement in the second clause. The motivation for using this type of utterance as a prosodic pattern for the prosody recognition task was twofold. Firstly, it provided an example of a prosodic pattern that conveys both attitudinal information about the speaker, and linguistic information in terms of accent or stress (acoustic analyses revealed a forward-looking contrastive accent on the noun). Secondly, it provided examples of utterances where prosody marked a change in the speaker’s attitude without a change in word order (as is the case with question/statement contrasts on sentence level in Afrikaans).

The materials developed to measure prosody recognition were validated in a group of NH listeners in quiet, and subsequently analysed to determine the acoustic characteristics that marked conditional versus unconditional prosody. In order to compare prosody recognition to word recognition on sentence level in noise (the listening experiment described in section 3.3), a second set of sentence materials was recorded from the same speakers from whom the prosody recognition materials had been recorded. This second set of sentences was based on previously validated material recorded from a different speaker (Theunissen, Swanepoel and Hanekom, 2011). The newly recorded version thereof was validated by having a qualified speech therapist approve its quality and intelligibility.

3.3 DEVELOPMENT AND ACOUSTIC ANALYSES OF TEST MATERIALS

3.3.1. Recording and editing of speech material: methods and results

The prosody recognition material consisted of a set of 12 sentences formulated for the purposes of this study, each denoting some form of permission (e.g. “You may borrow the car”), agreement (“I agree with his statement”), or approval of something (e.g. “He likes the house”). In seven of the sentences the noun was the final word in the sentence; in the other five sentences the noun was second or third to last. For each sentence, three separate versions were formulated (see Appendix A). In the first version, the sentence remained as it was, and denoted unconditional permission, agreement or approval. Before recording, speakers were instructed to read each of the sentences to themselves first, to reduce the effect of “read speech”. In a second version of the sentence, the statement (the main clause) was followed by a second clause that contained a condition pertaining to the noun in the sentence. This condition was introduced with the word “but” (e.g. “You may buy the tickets, but not the wine”). The third version of the sentence also included a condition starting with the word “but”, although this time the condition concerned the verb in the first clause of the sentence (e.g. “You may buy the tickets, but only if they are not too expensive”). To assist speakers in producing appropriate intonation, the researcher prompted the production of each of the sentences with conditional permission by asking a question (e.g. “Do you like the jam?”, to which the speaker would respond “I like the bread, but not the jam”; or “Would you like some more bread?”, to which the speaker would respond “I like the bread, but I have had enough”).

Recording of these sentences showed that the two conditional versions (one concerning the noun and the other concerning the verb) yielded entirely different prosodic realisations, with emphasis on different parts of the main clause. For example, if the condition concerned the noun in the main clause, the emphasis in the main clause was on the noun (e.g. "You may buy the *tickets,* but not the wine"), whereas a condition concerning the verb in the sentence resulted in emphasis on the verb (e.g. "You may *buy* the tickets, but only if they are not too expensive"), or the adverb (e.g. "You *may* buy the tickets, but only if they are not too expensive"). For the listening experiment, it was decided to use only the sentences that were recorded containing a condition related to the noun in the sentence as conditional utterances, since the versions that contained conditions related to the verbs had too much variability in their prosody. Only the sentences used in the listening experiments were subjected to acoustic analyses (see Table 3.1 for the results of these analyses).

A second set of sentences that could be used to assess word recognition was recorded from the same two speakers who produced the prosody materials. Sentences were selected from a collection of phonemically matched lists that were previously developed to assess speech recognition in noise and that had been shown to be of equivalent difficulty in noise (Theunissen *et al.*, 2011). These sentences were re-recorded for the purposes of the listening experiment described in section 3.3 so that the same speakers were used for the assessment of prosody recognition and word recognition. This was important to ensure that any differences that were found between recognition scores of the two types of material (the prosody materials and the phonemically matched lists) were not due to speaker differences. Different sentence lists were recorded for each speaker and each SNR that would be tested. The test forms showing these lists are included in Appendix B.

Speech materials were recorded digitally in a sound-proof booth, using an M-Audio Fast Track Pro external sound card (sampled at 44.1 kHz with 24-bit resolution). A Sennheiser ME62 microphone was placed on a microphone stand, 20 cm from the speaker's mouth. Two speakers were recorded (one male, one female, both native speakers of Afrikaans). After completion of the recordings, waveforms were edited using *Praat* software (Boersma and Weenink, 2010). Utterances conveying conditional permission were edited by removing the reduced coordinate clause that started with the word "but". Consequently, the two versions of each sentence (conditional and unconditional) were identical in content, differing only in terms of prosody. As the word "but" was preceded by a major prosodic boundary and therefore a short pause, the second clause could be removed without leaving any audible traces of the removed clause. Figure 3.1 shows an example of a sentence recorded with the coordinate clause included, illustrating the short pause between the part of the sentence that was retained (the first part), and the part that was removed (the second part).

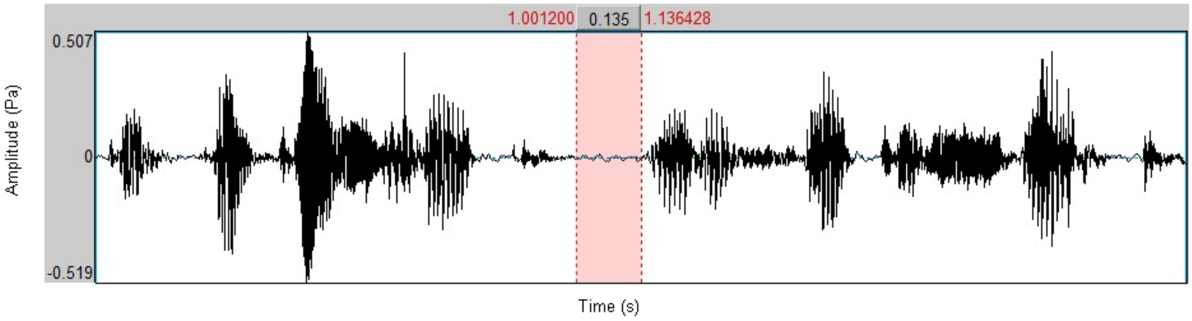

Figure 3.1: Example of a recorded sentence ("*Die boek is goed, maar die fliek is swak*", translated as, "The book is good, but the movie is bad") before editing. The highlighted section shows the pause between the first and second clause, which allowed removal of the second part of the sentence without leaving any audible traces of the removed clause.

Following the removal of the last part of each sentence, silences preceding and following the sentence were removed, leaving silences of 100 ms before and after the utterance, and the intensity (root-mean-square or rms value) of each sentence was re-scaled to 70 dB sound pressure level (SPL). This was done by importing each utterance recorded in .wav format into *Praat*, and using the "modify" function to scale the recording to the desired intensity. Recordings of the phonemically matched sentences intended for word recognition testing were edited in a similar fashion to the prosody materials, by removing unwanted silences before and after each sentence and re-scaling intensities to 70 dB SPL. Re-scaling intensities preserved relative intensity changes within the utterance, while ensuring that each item in the collection had the same intensity to ensure equal SNRs in the noise experiment. Edited materials were saved to hard disc in .wav format.

3.3.2 Validation of prosody recognition materials: methods and results

The prosody materials were validated in a group (n = 12, six male and six female) of NH listeners, who are all native speakers of Afrikaans (aged 18 – 26 years). The pure tone thresholds of all listeners were ≤ 20 dB HL at octave frequencies from 125 to 8000 Hertz (Hz). Listeners provided informed consent and were paid for their participation at the standard rate of the research group. For the listening experiment, listeners were seated in a sound-proof booth with the test administrator. Speech materials were presented using *Praat* as an interface, through an M-Audio EX66 Reference Monitor (-3dB bandwidth from 37 Hz to 22 kHz, with flat frequency response in between that allows maximum variation of +/- 1 dB). The recorded materials were presented at a comfortable loudness (measured as 70 dB SPL at ear level). Prior to testing, listeners were informed of the nature of the test items, as well as the methods used to record these items. They were encouraged to make swift judgements, following their “first instincts” rather than spend excessive time considering each decision. The presentation of the items in each playlist was controlled by the examiner, who presented the next item once the listeners had completed the discrimination task and recorded their response on a test form. Test duration was approximately 40 minutes per listener. Listeners were presented with one sentence at a time, and had to select (in a 2AFC paradigm) whether the speaker was giving conditional or unconditional permission. Sentence materials recorded from each speaker received a percentage score calculated according to the number of listeners who were able to identify the intended prosodic pattern correctly. For both male and female speakers, average recognition scores across 24 utterances (12 conditional and 12 unconditional) were 96.5% (standard deviation for male speaker 9.8%, and for female speaker 8.5%), with all utterances correctly classified by more than 65% of the listeners. These recognition scores were deemed sufficient to consider the recorded materials valid for use in the listening experiment.

3.3.3 Acoustic analyses of prosody materials: methods and results

To explore the acoustic cues contained in the prosodic pattern, aspects of fundamental voice frequency (F0), duration, intensity and voice quality in each utterance of the prosody materials were analysed using *Praat*. Each of the parameters analysed as described in this section was documented for the unconditional and conditional prosodic versions separately to enable comparison of the two prosodic conditions. In addition, the data for each speaker were documented separately, so that statistical comparisons could be made between the two conditions as produced by the same speaker. Before statistical comparisons were made, the normality of the distribution of the data was evaluated using the Kolmogorov-Smirnov test to determine whether parametric or non-parametric procedures should be used. The data were compared to look for significant differences using paired sample t-tests (for normal distributions) or the Wilcoxon signed-rank test for data with non-normal distributions.

3.3.3.1 F0 cues

F0 was extracted from the utterances with *Praat*'s autocorrelation algorithm, using assumed F0 ranges of 100 to 500 Hz for the female speaker and 65 to 300 Hz for the male speaker. This means that *Praat* identifies F0 values within the frequency band that lies between the low and high values of the assumed F0 range. The average, standard deviation and range between minimum and maximum F0 for the whole sentence were documented, as these variables are known to play an important role in prosody (see Chapter 2 for details). In addition, an approximation of the intonation contours of the sentences was determined by plotting the average F0 of each word in the sentence. Although this method could not accurately reflect all the small F0 changes in the sentences (see Figure 3.2 for an example), it provided a means to quantify “pitch accent” patterns (emphasis on particular words produced by an increase in F0) in the sentence.

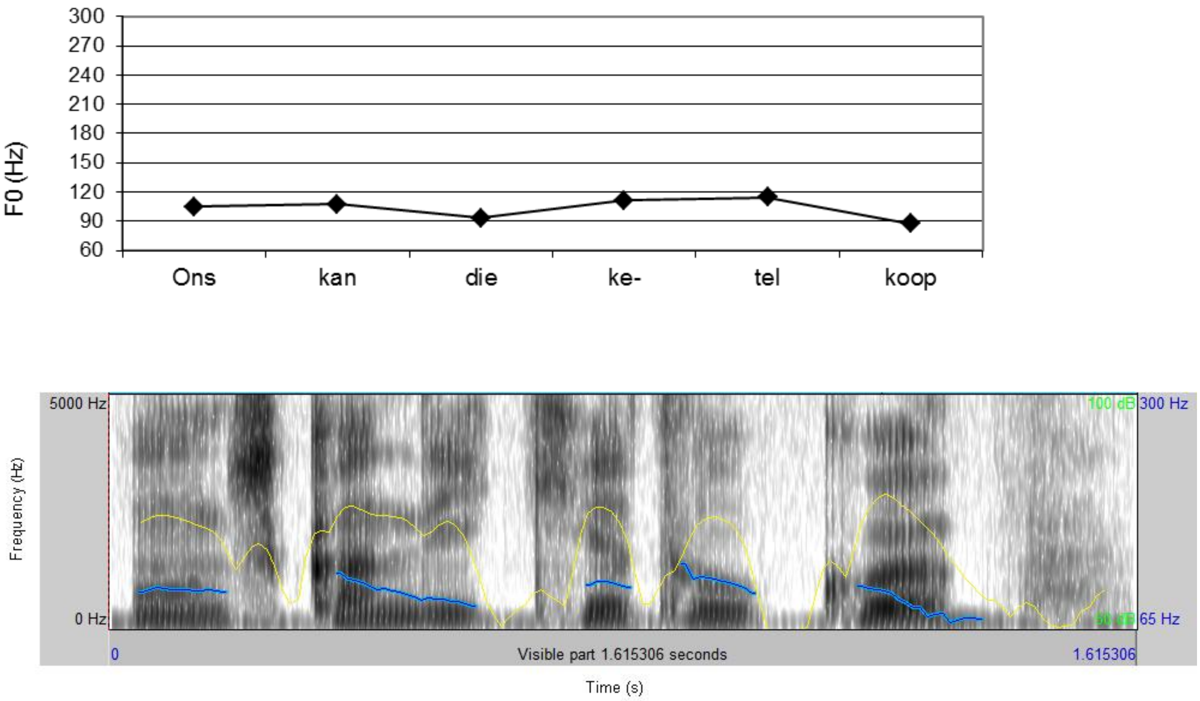

Figure 3.2: F0 patterns across the sentence "*Ons kan die ketel koop*" (We can buy the kettle*).* The top panel shows the pitch accent pattern plotted according to the mean F0 of each word in the sentence, while the bottom panel shows the F0 contours as plotted in *Praat* (indicated by the blue lines; yellow lines indicate amplitude variation). As an example of the difference between the two methods, the bottom panel shows how F0 was falling on the last word ("koop"), while in the top panel, this word's F0 is only depicted by a single value.

The calculated pitch accent patterns indicated that the male speaker especially used pitch accent to emphasise the noun in the conditional prosodic versions, and that the increase in F0 on these nouns was frequently preceded by a decline in F0 on the syllable immediately preceding the noun. Therefore, although the F0 on the accented noun, in some cases, did not differ greatly from the average F0 of the sentence, it still had a clear pitch accent due to the difference between the stressed syllable and the preceding unstressed syllable. To quantify this effect, termed "pitch accent difference" for the purposes of this study, the difference between the first (or only, if it was a monosyllabic word) syllable of the noun and the preceding unstressed monosyllabic word was calculated for each sentence in each prosodic condition. In all the multisyllabic nouns used in the present study, word stress naturally fell on the first syllable.

F0 cues analysed for the sentence materials are shown in Table 3.1. The average, standard deviation and range of F0, as well as the noun F0 and pitch accent difference, were higher in the conditional than unconditional versions for both speakers. With the exception of average F0 in the male speaker, all of these differences were statistically significant (*p* < 0.05 or smaller) according to Wilcoxon signed-rank tests.

Table 3.1: Mean values of acoustic parameters across utterances (n = 12) for unconditional and conditional prosody; standard deviations in brackets. Significant differences (*p* < 0.05 or smaller) indicated in bold.

	Female speaker				**Male speaker**			
	Unconditional		Conditional		Unconditional		Conditional	
Average F0 (Hz)	238.01		**272.50**		99.64		102.41	
Standard deviation F0 (Hz)	27.40		**51.67**		13.11		**19.06**	
F0 range (max - min) (Hz)	113.16	(28.06)	**186.89**	(45.32)	61.69	(16.02)	**78.65**	(17.05)
Noun F0 (Hz)	222.89	(18.41)	**239.52**	(18.92)	96.12	(12.91)	**115.54**	(14.31)
Pitch accent difference (Hz)	-7.01	(28.94)	**10.76**	(37.72)	1.32	(12.86)	**27.82**	(21.91)
Speech rate (syllables/second)	4.33	(0.57)	**4.64**	(0.44)	**3.89**	(0.47)	3.71	(0.42)
Final word duration (s)	0.40	(0.09)	0.41	(0.11)	0.50	(0.10)	**0.55**	(0.13)
Noun duration (s)	0.36	(0.12)	**0.40**	(0.14)	0.45	(0.16)	**0.51**	(0.19)
Final word intensity (dB)	62.87	(4.30)	63.55	(7.52)	69.30	(0.99)	70.30	(1.93)
Noun intensity (dB)	67.39	(3.31)	**70.05**	(2.56)	70.51	(1.62)	**72.47**	(1.32)
% Modal voice final vowel	52.92	(42.78)	27.73	(40.14)	69.42	(44.48)	78.29	(33.93)
Harmonics-to-noise ratio final vowel (dB)	8.14	(3.60)	6.58	(5.30)	5.34	(1.50)	8.18	(2.35)
Periodicity-to-noise ratio final vowel (dB)	7.18	(4.53)	7.88	(5.40)	4.33	(2.71)	7.74	(2.78)
TOTAL SIGNIFICANT DIFFERENCES:	8/13				8/13			

F0 movements at the end of the sentences were analysed to see whether the intonation contour at the end of the main clause provided an indication of the finality or non-finality of the final word in the clause. The position of sentence accent in the present sentence materials varied between the two prosodic conditions, and also within the collection of sentences. This variability could affect the realisation of the intonation contour, and had to be controlled for in the analyses. A number of the final words in the conditional versions of the sentences carried sentence accent, although some of these words were multi-syllabic, and carried stress on the first syllable, with one or two post-tonic syllables. To reduce the confounding effects of sentence accent on the analysis of the terminal intonation contour, the F0 contour of the very last syllable (or word, if the last word was monosyllabic) was analysed, but only in those sentences (n = 8) where these syllables did not carry the main sentence accent in either of the two prosodic conditions. The F0 estimates of the final words and syllables were obtained using *Praat* and plotted as a function of time. The resultant contours suggested that no single value (such as difference, average, or slope) could adequately describe the intonation contours of the final syllables. This was because many of the plotted intonation contours had one or more changes in direction somewhere during the course of the syllable (e.g. first falling and then rising, or vice versa), and a single value would fail to account for such intra-syllabic changes (see Figure 3.3 for an example).

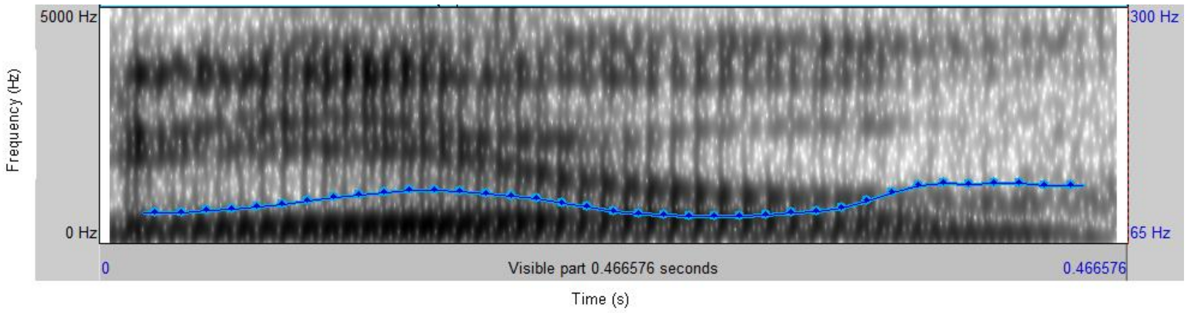

Figure 3.3: Intonation contour of the Afrikaans word "*reël*" ("rule" in English) as produced by the male speaker in the conditional version of sentence 6. F0 is depicted by the blue line and shows how the intonation started with a small rise, then fell and rose again at the end.

As an alternative method, each intonation contour was coded according to its shape as rising, falling, flat, or a combination of these (e.g. rising-falling, referring to a contour that rises and subsequently falls). The results of this method are reported in Table 3.2, which also shows the number of occurrences of each intonation contour type in each of the prosodic types.

The results in Table 3.2 show that the male speaker used rising contours more frequently in the conditional as opposed to unconditional versions (six rising contours for conditional compared to one rising contour for unconditional). For the female speaker, this pattern was not as clear, with only one rising contour in unconditional versions and only two rising contours in conditional utterances.

Table 3.2: Intonation contours of final word/syllable in eight of the recorded sentences. F = falling; R = rising; F, R refers to contours that fall and then rise. "No pitch" denotes samples where no F0 and therefore no pitch contour could be established owing to the presence of non-modal voice such as a whisper or creaky voice. Contours ending in a rise are depicted in bold.

Sentence number	Female speaker	Male speaker
	Unconditional	
1	**F, R**	Flat
2	F	F
3	*No pitch*	**R**
4	F	F
6	Flat	*No pitch*
7	*No pitch*	F
8	F	*No pitch*
12	F	F
Rising contours:	1	1
	Conditional	
1	F	Flat
2	*No pitch*	Flat
3	*No pitch*	**R**
4	**R**	**R**
6	F	**R, F, R**
7	F	**F, R**
8	*No pitch*	**R**
12	**R**	**F, R**
Rising contours:	2	6

3.3.3.2 Duration and rate cues

The durational parameters of the sentence materials were explored by documenting the speech rate (calculated as the number of syllables per second), as well as the duration of the noun and the final word in the main clause. Results are included in Table 3.1. Although the speech rate (syllables per second) differed significantly between prosodic conditions for both speakers, the difference did not show the same pattern for the two speakers. The female speaker used a higher speech rate in the conditional utterances, while the male speaker used a higher rate for unconditional versions. The duration of the final word was longer in the conditional version for both of the speakers, but this was significant only for the male speaker. Noun duration was significantly longer in both speakers.

3.3.3.3 Intensity cues

The overall intensity (across the frequency spectrum) of the final word and the noun in each sentence was measured in decibels. The intensity of the noun was significantly higher in the conditional version for both speakers (see Table 3.1). Final word intensity was higher in the conditional version of both speakers, but the difference between final word intensity of unconditional and conditional versions was not significant.

3.3.3.4 Voice quality cues

The voice quality of the final word in each sentence was investigated to determine whether the laryngealisation described in the literature occurred on these words, and whether it occurred more frequently in sentences where the final word truly was the last word that the speaker uttered, than in sentences where the "final word" was actually followed by a reduced coordinate clause in the unedited version. Other authors have mentioned the presence of laryngealisation at utterance endings (Kreiman, 1982; Lehiste, 1979), but instead of quantifying or describing the voice quality of these laryngealisations, the authors merely stated that laryngealisation was either present or absent. Upon inspection of the present recordings, however, it was often quite difficult to make such a distinction. The decrease of vocal effort towards the end of a sentence often resulted in disturbances in voice quality on the final word, although many of these disturbances could not be classified as "laryngealisations" or "creaky voice" by definition, which stipulates that this is a low-frequency periodic vibration, below the speaker's modal pitch register (Batliner, Steidl and Nöth, 2007; Laver, 1980). Quantifying voice quality is problematic, complicated by both the inconsistent labelling of different phonation types in the literature, and the technical difficulties in obtaining reliable measures of voice quality (Gobl and Ní Chasaide, 2003). Three methods of quantification were applied and compared in the present study. Firstly, the spectrogram of the syllable under scrutiny was inspected and any obvious disturbances in voice quality (such as an audible glottal fry, visible as low-frequency periodic vibration on the spectrogram, or whisper, visible as aperiodic noise) were identified and the duration of these disturbances was measured. The duration of modal voice was then documented as a percentage of the total duration of the syllable. Phonation was considered to be modal if it was audible, if the intensity curve indicated increased energy, as usually seen in voiced sounds, if there was clear periodicity in the time signal and if the estimated F0 fit the intonation contour reasonably well (this excluded disturbances such as octave jumps). This method made it possible to determine the proportion of each syllable that was uttered in non-modal voice (if any), and the frequency of occurrence of non-modal voice could thus be compared between the two prosodic conditions. Figure 3.4 shows examples of spectrograms (as shown in *Praat*'s sound editor window), which illustrate differences in voice quality between two recorded versions of the same word recorded from the female speaker. The spectrograms show the section of the Afrikaans word “stelling” (/s t æ l ə ŋ /) that is supposed to be voiced (/æ l ə ŋ /).

In the spectrogram shown in the top panel of Figure 3.4, the entire voiced part of the utterance was produced with modal voice for which F0 could be identified (depicted by the blue line in the figure). In the bottom panel, the spectrogram shows that the middle part of the utterance was produced with non-modal voice, as reflected by a lack of periodicity in the signal, and the lack of identifiable F0. Within *Praat*'s sound editor window, which displayed the spectrogram, it was possible to determine the duration of the part of the utterance produced with non-modal voice, as well as the duration of the part of the word that was supposed to be voiced, and consequently the percentage of the voiced section produced with non-modal voice could be calculated (in this case 59.38%).

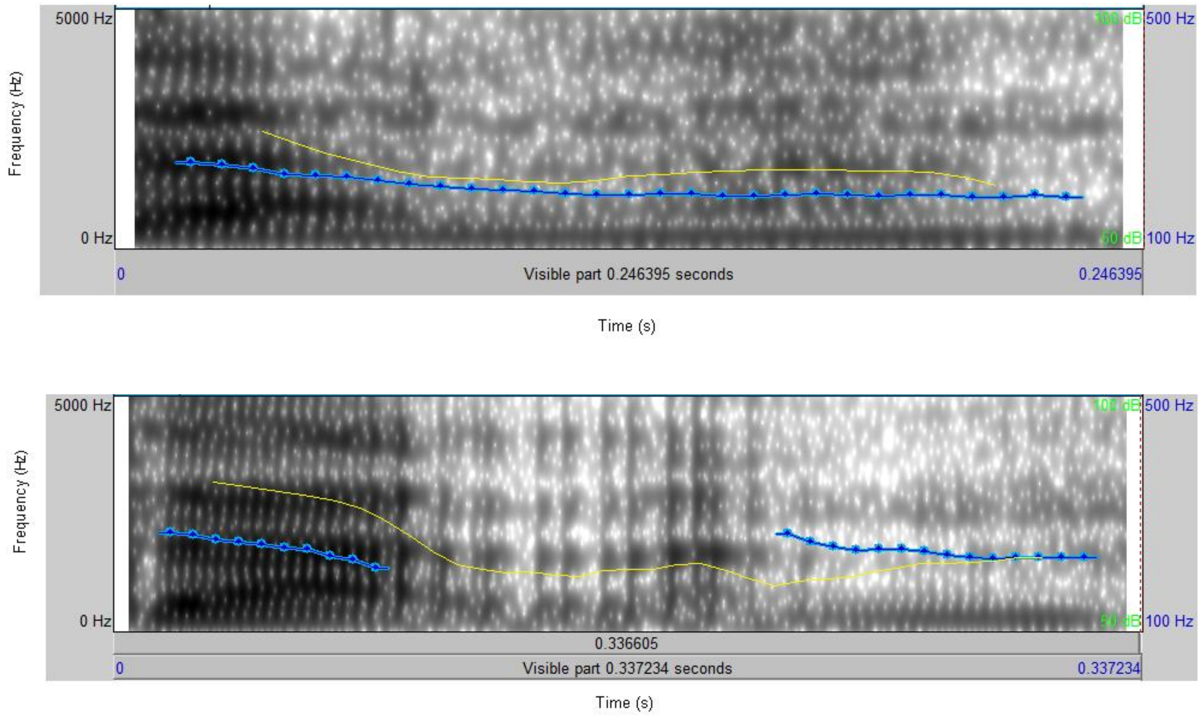

Figure 3.4: Spectrograms of the voiced part of the word "*stelling*" (/æ l ə ŋ /) as recorded from the female speaker. The blue lines indicate voice F0, while the yellow line shows intensity levels. The top panel shows the utterance excised from the unconditional version of the sentence, while the bottom panel shows the utterance excised from the conditional version (where the condition pertained to the noun).

The second method of voice quality analysis followed that of Cheang and Pell (2008), who measured the harmonics-to-noise ratio (HNR) as an indication of voice quality. The HNR was obtained using a function in *Praat* that uses cross-correlation analysis to detect acoustic periodicity and reflects the voice quality in areas where F0 has been determined. This measure has been validated as a measure of hoarseness, a pathological voice quality (Yumoto, Gould and Baer, 1982), but was also found useful to indicate voice quality differences between sincere and sarcastic utterances (Cheang and Pell, 2008). The third measure of voice quality was the periodicity-to-noise ratio. This measure indicates the periodicity-to-noise ratio in the entire syllable, irrespective of whether F0 could be estimated (as opposed to the HNR, which only measures voice quality of the parts of the utterance where F0 could be estimated). This third measure was used, as many of the utterances with irregular or non-modal voicing had sections where no F0 could be identified. Because of the confounding effects of sentence accent on end-of-sentence cues, only eight of the sentences were used for these analyses. In the remaining four sentences, the noun was a monosyllabic word that occurred at the end of the main clause. By implication, some of these words carried pitch accent and were therefore not produced with a possible decrease in voice quality that might characterise the last word in an utterance. In five of the sentences that were analysed, the final word was monosyllabic and free from sentence accent. The remaining three ended on a multi-syllabic word with the accent on the first syllable. The voice quality of the final, unstressed syllable or the entire word in the case of monosyllabic, unaccented final words was analysed.

Results of the voice quality analysis of eight of the final vowels in the sentence materials (those that did not carry sentence accent in the conditional versions) are shown in Table 3.1. Statistical analyses of these results revealed no significant differences between prosodic conditions for either of the speakers relating to any of the voice quality measures used.

3.3.4 Discussion

The findings of the validation experiment clearly indicated that NH listeners were able to distinguish conditional prosody from unconditional prosody at a level considerably better than chance for materials recorded from both speakers. Although the listening experiment offered listeners only two options to choose from, listeners were consistently able to assign utterances to the correct prosodic option and did not merely make a same/different discrimination. These results indicate that NH listeners are able to hear the difference between conditional and unconditional permission or agreement in the sentence materials used.

It was expected that the acoustic cues for this specific type of prosody would correspond to some degree to cues for emphasis and cues for finality or continuation of an utterance, although the result of the combination of cues could not be predicted beforehand. Cues for emphasis of the noun were expected to be stronger in the conditional version, indicating an upcoming contrast (Swerts, 2007), and could include such changes as an increase in F0, intensity and duration of the noun. Results of the acoustic analyses indicated that emphasis of the noun was indeed a prominent cue to conditional prosody, with both speakers producing the noun at a significantly greater intensity, average F0 and longer duration. This finding suggests that both speakers used a forward-looking contrastive accent on the noun, which refers to emphasis on the noun that indicates that the content of the upcoming clause would contain some reference to the noun of the main clause (e.g. emphasis on the word "dog" in contrast to "cat" in the sentence "You may have the *dog,* but not the *cat")*.

Cues for sentence ending or upcoming continuation that were analysed included speech rate and the intonation contour, duration, intensity and voice quality of the final word or final unaccented syllable of the main clause. Finality cues were not applied in the same manner by the two speakers. While the female speaker used a higher speech rate in the conditional utterances (which could indicate that there was more to come in the utterance), the male speaker did not. Also, while the male speaker substantially increased the use of rising contours for conditional versions, the female speaker did not. Another potential cue of finality was the duration of the final word in the main clause, which was expected to be longer if the sentence ended after this word (in other words, it was expected to be longer in the unconditional version). While the male speaker produced the final words of the conditional versions with increased duration, the female speaker's data showed no significant difference in final word duration between the two prosodic conditions. This effect could have been confounded by the interaction between emphasis and continuation cues, as many of the final words in the conditional versions were accented, which may have increased their duration in the male speaker. The intensity and voice quality of the final words in the unconditional versions were expected to be lower than in the conditional versions, as a cue of sentence ending (Kreiman, 1982; Local and Kelly, 1986). However, neither the intensity nor the voice quality of the final words differed significantly between the two versions for either of the two speakers. This finding may be specific to the prosody investigated here, but since no published studies on sentence end cues in Afrikaans report on this issue, it is not certain whether this cue is ever applied consistently at sentence or utterance boundaries. Therefore, while the male speaker produced more rising intonation contours in the conditional versions to mark continuation, the female speaker increased speech rate in the conditional versions, possibly because of the greater number of words in these versions as they were recorded (Grosjean, 1983). Since the prosody of both speakers' recordings were perceived with equal accuracy by listeners, this finding seems to indicate a possible cue-trading relationship with regard to finality or continuation cues in the speech materials used.

Other acoustic cues in the sentence materials that could not be exclusively categorised as either emphasis or continuation cues were the average, standard deviation and range of F0 across each sentence. Both speakers showed a greater amount of variation in F0 in the conditional versions, as indicated by larger standard deviations and range. This might be due to an increased amount of emphasis in the conditional versions (produced as pitch accent on the noun), or, in the case of the male speaker, a rising intonation pattern (a continuation cue).

3.4 LISTENING EXPERIMENT: PROSODY AND WORD RECOGNITION IN NOISE

3.4.1 Method

3.4.1.1 Speech materials

To compare the perception of prosody at sentence level with the recognition of words in a sentence, the recorded speech materials described in section 3.2 of this chapter were combined with SWN generated in a commercial software package for mathematics. The noise had a spectral envelope matching the average power spectral density of the entire set of sentences recorded from each speaker. This was achieved by determining the average spectrum for each speaker across all utterances, and using the envelope of this spectrum as a filter to shape the spectrum of white noise accordingly. The resulting noise was added to the recorded speech at SNRs of -2, -5 and -8 dB. The sentences that portrayed conditional versus unconditional prosody were distributed across the three test conditions (SNR-2, SNR-5 and SNR-8) so that each listener heard only one version (either conditional or unconditional) of each sentence at each SNR. To measure word recognition, a different group of phonemically matched sentence lists was used at each SNR to avoid familiarisation with the content (see Appendix B for test lists used at each SNR). Word recognition was tested using three lists of ten sentences each at each SNR.

3.4.1.2 Participants

Ten listeners participated in this experiment. All participants were young adults (ages 19-25 years), students at a tertiary education institution, native speakers of Afrikaans (the test language), and had normal hearing (pure tone thresholds ≤ 20 dB HL at 250, 500, 1000, 2000, 4000, and 8000 Hz). Informed consent was obtained from each listener prior to testing, and listeners were rewarded at the standard hourly fee specified by the research group.

3.4.1.3 Test procedure

Participants were seated in a sound-proof booth with the examiner (a qualified audiologist) for the duration of each experiment. Test materials were presented via the external sound card of a personal computer, through an M-Audio EX66 Reference Monitor. Listeners were seated approximately one metre from the loudspeaker, facing it squarely. Materials were presented at 65 dB SPL as measured at the ear level of the participant. This intensity level was selected as it was considered to be a comfortable listening level by the NH participants. The presentation of the test items was controlled by the administrator, using *Praat* software as an interface.

The first five participants started with the recognition of the phonemically matched sentences, with the female speaker. Testing commenced with 20 practice sentences, followed by three lists of ten sentences each. This was followed by the phonemically matched sentences as read by the male speaker (again 20 practice sentences followed by 30 test sentences). The sentences were presented one by one, and listeners were required to repeat whatever part of the sentence they were able to hear. The test administrator compared the listener's response to a written version of the sentence that was printed on a test form, and indicated on the test form the number of words in the sentence that were repeated correctly. After all the sentences for both speakers had been presented, the female speaker's version of the conditional/unconditional sentences were played, starting with three practice items, followed by the male speaker's version. No feedback or training was given after practice or test items. Each listener heard only one version of each sentence from each speaker. Listeners were required to classify each utterance they heard as either conditional or unconditional, and to indicate this on a printed test form. The second half of the group of listeners (listeners six to ten) completed the same test items but in reverse order, starting with the male speaker's version of the conditional/unconditional sentences. This was done to counterbalance any possible practice or learning effects that might have taken place during the course of the test session.

Every listener was first tested in the SNR-2 condition. This was followed by a two-week waiting period to minimise practice and memory effects, and then by the same sequence of tests in the SNR-5 condition. After another two-week wait, listeners were finally assessed in the SNR-8 condition.

3.4.2 Results

Prosody recognition performance was calculated as the percentage of sentences for which the prosodic version (conditional/unconditional) was identified correctly. Word recognition scores were calculated as the percentage of words repeated correctly from the three phonemically matched lists. Because the prosody task was a closed set (2AFC) task while word recognition was an open set task, prosody recognition scores were corrected for guessing using Boothroyd's equation (1988) as specified in Equation (2.1) of Chapter 2. Data from one of the listeners (listener number 2, a female listener) were excluded from the analyses, as this listener was a clear outlier on all the tests, performing considerably worse than all other listeners. This listener adhered to the selection criteria for the study, but may have misunderstood some of the instructions of the listening tasks. Results are depicted in Figures 3.5 and 3.6, the former showing results from the two speakers separately, and the latter showing results averaged across the two speakers.

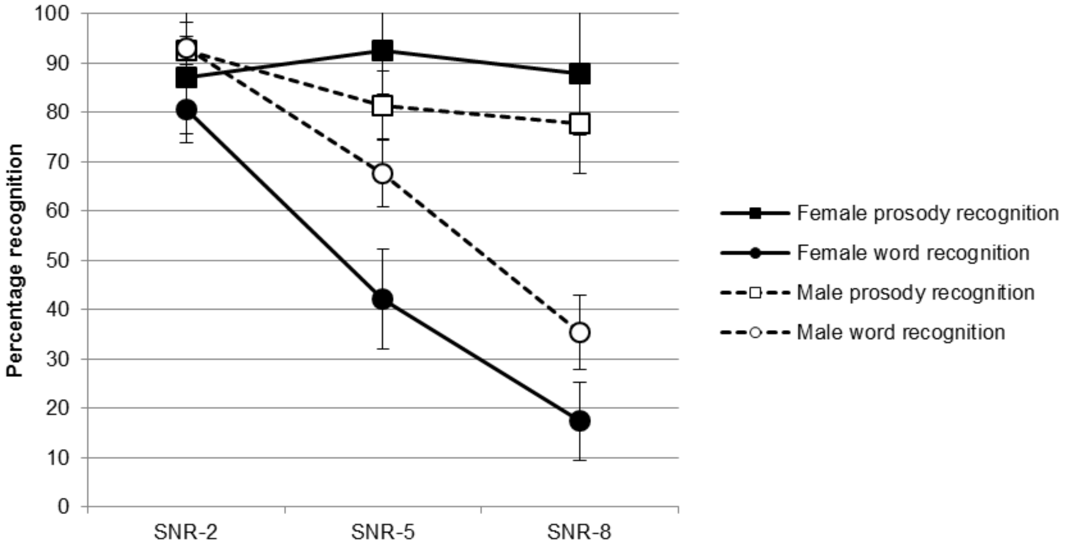

Figure 3.5: Prosody and word recognition of NH listeners (n = 9) in noise for male and female speaker separately. Error bars indicate one standard deviation from the mean.

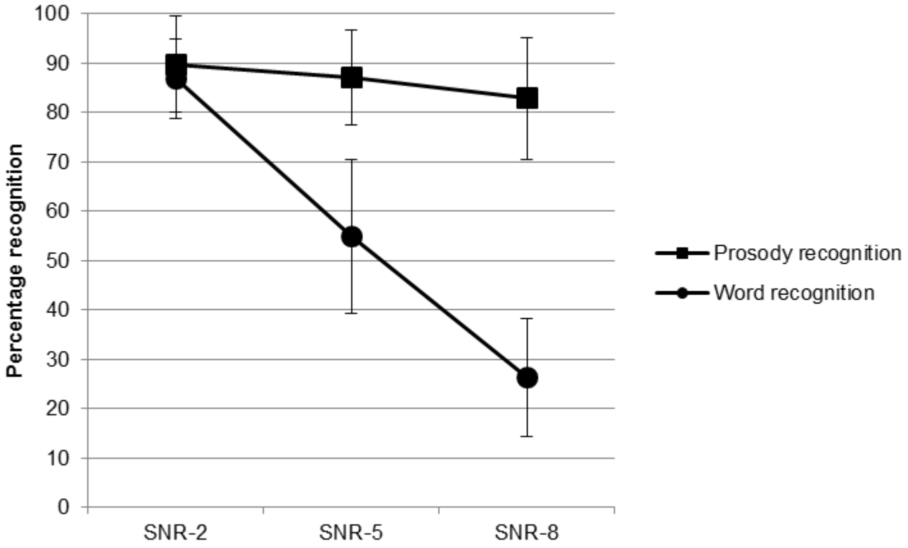

Figure 3.6: Prosody recognition and word recognition across speakers (n = 2) and listeners (n = 9). Error bars indicate one standard deviation from the mean.

Average prosody and word recognition scores across listeners were compared at the easiest and most difficult SNRs (SNR-2 dB and SNR-8 dB) using Wilcoxon signed-rank pairwise comparisons. A Bonferroni correction for the number of pairwise comparisons was applied, and all effects are reported at a 0.008 level of significance (0.05/6), according to asymptotic one-sided significance values. For the female speaker, word and prosody recognition did not differ significantly at SNR-2, T = 24.00, p = 0.43. The same was true for the average score across the two speakers, T = 24.00, *p* = 0.43, but for the male speaker, prosody recognition was significantly poorer than word recognition at SNR-2 dB, *T* = 44.00, *p* = 0.006, *r* = 0.60. However, at SNR-8, prosody recognition was significantly better than word recognition for both speakers. For the female speaker, average (corrected) prosody recognition was 58.54% better than word recognition at this SNR, *T* = 0, *p* = 0.004, *r* = -0.63. Recordings from the male speaker yielded a smaller difference of 20.19% between word and prosody recognition, but this difference was still significant, *T* = 0, *p* = 0.0075, *r* = -0.57. The average prosody score across speakers was 39.37% better than the average word recognition score at SNR-8, also a significant difference, *T* = 0, *p* = 0.004, *r* = -0.63.

The slope with which recognition of words and prosody deteriorated was compared by fitting a linear curve to the three data points (SNR-2, -5 and -8 dB) for each listener using a least squares estimate. An average deterioration slope could then be calculated for each task for the two speakers separately and averaged together. Differences in deterioration were statistically compared using Wilcoxon signed-ranks pairwise comparisons. A Bonferroni correction for the number of pairwise comparisons was applied, and all effects are reported at a 0.017 level of significance (0.05/3), according to asymptotic one-sided significance values. Table 3.3 shows the average slopes with standard deviations and statistical results across listeners.

Table 3.3: Slope of recognition deterioration for the two listening tasks (word and prosody recognition) across listeners (n = 9). *T* indicates the test statistic of the Wilcoxon signed-rank test, *p* denotes one-sided asymptotic significance, *r* indicates effect sizes, calculated as $z/\sqrt{n}$

		Average slope (%/dB)	**Standard deviation**	***T***	***p***	***r***
Female speaker	Word recognition	-10.55	1.61	45.00	0.004	0.63
	Prosody recognition	0.30	2.61			
Male speaker	Word recognition	-9.63	1.37	44.00	0.006	0.60
	Prosody recognition	-4.94	3.87			
Both speakers	Word recognition	-10.09	1.10	45.00	0.004	0.63
	Prosody recognition	-2.32	2.88			

The results in Table 3.3 show that for each of the speakers separately, as well as for the two speakers averaged together, word recognition deteriorated with a significantly steeper slope than prosody recognition as SNR deteriorated. In the case of the female speaker, prosody recognition showed a slight increase in recognition at poorer SNRs. This result might have been due to a slight practice effect that occurred, since the easier SNRs were tested first.

3.4.3 Discussion

Comparisons between word and prosody recognition scores indicated that at the easiest SNR tested (SNR-2 dB), the two tasks were either of equivalent difficulty, as indicated by a lack of significant differences between scores (for the female speaker and for the two speakers averaged together), or word recognition was significantly easier than prosody recognition, as found with results from the male speaker. However, at the most difficult SNR tested (SNR-8 dB), prosody recognition was significantly easier than word recognition. The slope at which prosody recognition deteriorated was also compared to the slope of deterioration of word recognition and results indicated that word recognition deteriorated with a significantly steeper slope with deterioration in SNR than prosody recognition.

The findings of this experiment suggest that, in NH listeners, the recognition of the prosodic contrast investigated here is more immune to the effects of background noise than the recognition of words in a sentence. Based on the results of the acoustic analyses on the prosody test materials used in the experiments, some of the acoustic cues that supported prosody recognition were changes in F0 (average and range), as well as a combination of acoustic cues that indicated emphasis of the noun in conditional versions (increased duration, intensity, and F0). Word recognition, on the other hand, relied on accurate recognition of phonemes (vowels and consonants), but was probably also supported by syntactic and semantic clues provided by the context of the sentence. Historically it has been estimated that in English, the structure of the language determines about 50% of the utterance, i.e. the redundancy of the language is around 50% (Shannon, 1948). Boothroyd and Nittrouer (1988) have also attempted to characterise the effect of context on word intelligibility. They reported that syntax and semantics result in a 170% increase in intelligibility of words from being presented in isolation to being presented in meaningful four-word sentences.

However, the effects of context on intelligibility depend on a number of factors, such as the type of sentence (e.g. stereotypical or fixed expressions versus meaningful sentences) (Lieberman, 1963), the predictability of the words in the sentence (Kalikow *et al.*, 1977; Leventhal, 1973), and the frequency of occurrence of words in the test language (Boothroyd and Nittrouer, 1988; Leventhal, 1973), as well as the intelligibility of the individual words in isolation (Boothroyd and Nittrouer, 1988; Grant and Seitz, 2000). Unfortunately, word frequencies for the test language are not well documented, so the frequency of occurrence of the words in the sentences was not known. The predictability and intelligibility of the individual words were also not known, so it was not possible to determine the exact contribution of sentence context to intelligibility in the present work.

According to a number of reports, the natural intonation contour (movements of voice F0 across the sentence) also supports word recognition in noise (Binns and Culling, 2007; Laures and Bunton, 2003; Laures and Weismer, 1999). Both the prosody and word recognition tasks therefore had some degree of built-in redundancy that may have supported recognition in noise. The results of the listening experiment suggest, however, that the acoustic cues that supported prosody recognition were more noise-immune than those needed for accurate word recognition, despite the support of sentence context in the word recognition task, and although F0 may have influenced both word and prosody perception. A possible explanation for this might be that the acoustic cues to the prosodic contrast (temporal, intensity and pitch cues) were spread out across the entire utterance and therefore perhaps more redundant than the cues required for word recognition, which entailed accurate perception of phonemes, a much shorter unit of speech.

This is in agreement with the findings of Mattys (2004), who demonstrated that syllable stress, a prosodic cue to word boundaries, was more resilient to background noise than co-articulation (a sub-segmental cue). Smith *et al.* (1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989)(Smith *et al.*, 1989) also showed that NH listeners could successfully perceive word boundaries and stress rhythm (the rhythm of stressed and unstressed syllables in an utterance) at SNR levels where segmental information (phonemes) was no longer discernible. The prosodic differences used by listeners to complete the prosody recognition task in the present experiment were also signified at least in part by stress differences (on the noun), which suggests, together with existing evidence, that syllable or word stress is an acoustically redundant speech cue with a high degree of noise immunity. This may be because stress or emphasis is realised through a combination of intensity, duration and F0 changes (Cruttenden, 1997; Fry, 1958). It remains to be seen whether other patterns of prosodic differences display the same level of noise immunity as the one investigated in this experiment.

A limitation of the present listening experiment was that the test paradigm of the two tasks (word and prosody perception) differed, with prosody recognition being measured in a closed set paradigm (2AFC) and word recognition in an open set paradigm. However, it could be argued that the word recognition task benefited from the redundancy inherent to meaningful sentences, where semantic and syntactic clues could aid perception, and successful word recognition did not depend on accurate perception of each phoneme in the word. Since it was not possible to characterise the exact contribution of the semantic and syntactic clues on word recognition (because of a lack of data on word frequencies, sentence predictability and word intelligibility in isolation), subsequent experiments (described in chapters 4 and 5) were designed so that prosody recognition and the recognition of segmental speech features (vowels and consonants) were tested in identical test paradigms. In the present experiment, an attempt was made to minimise the effects of the test paradigm difference by correcting prosody recognition scores for guessing.

3.5 CONCLUSIONS

The following can be concluded from the development and acoustic analyses of the prosody materials used in this experiment.

- NH listeners are able to distinguish conditional from unconditional permission, agreement or approval based on prosodic cues with a high degree of accuracy.
- Acoustic analyses showed that conditional prosody as realised in the prosody materials developed here differed from unconditional prosody on the basis of a number of F0 cues in both speakers.
- Both speakers produced increased F0, duration and intensity to emphasise the noun in the conditional utterances.
- Both speakers produced a type of continuation cue to indicate the upcoming clause in conditional utterances, but while the male speaker produced rising intonation as a continuation cue, the female speaker used an increase in speech rate to indicate the increased length of conditional utterances.

The following conclusions can be drawn from the results of the listening experiment.

- In the presence of SWN, at a poor SNR (-8 dB), NH listeners perform significantly better on the discrimination between conditional versus unconditional prosody than on the recognition of words in a sentence.
- In NH listeners, the recognition of words in a sentence deteriorates significantly faster as SNR deteriorates than the recognition of conditional/unconditional prosody on sentence level in SWN.
- In light of the acoustic analyses, it appears that the acoustic cues that marked conditional versus unconditional prosody, namely changes in voice F0 in general, and emphasis on the noun marked by increased intensity, F0 and duration, were more immune to background noise than the acoustic cues required to perceive words in a sentence.
- Although the test paradigm for the prosody recognition task was easier than the word recognition task in this experiment, prosody recognition scores were corrected for guessing and results suggest that the cues required to perform prosody recognition were more immune to the interference of SWN than the cues needed to recognise words in a sentence.

Further investigation is needed to answer a number of questions stemming from this listening experiment, namely i) whether the noise immunity of prosody found in this experiment can also be found in other prosodic contrasts (e.g. word-level prosody, linguistic prosody such as question/statement prosody, emotional prosody); ii) whether the noise immunity of prosody is higher than that of phonemes (vowels and consonants) when compared in identical test paradigms; and iii) whether CI recipients also perform better on prosody recognition than word or phoneme recognition in noise. As discussed in Chapter 2, CI recipients have considerable difficulty with prosody recognition, but also exhibit phoneme recognition abilities that are poorer than those of NH listeners. A direct comparison between these two abilities in CI users should therefore make a valuable contribution to existing knowledge. The following chapter describes the second set of listening experiments of the present study, which was aimed at addressing some of the research questions that remained unanswered after the first listening experiment.

CHAPTER 4 PERCEPTION OF VOWELS AND PROSODY BY CI RECIPIENTS IN NOISE

Parts of this chapter were published in the Journal of the Acoustical Society of America (Van Zyl and Hanekom, 2013b), while other sections were published in the Journal of Communication Disorders (Van Zyl and Hanekom, 2013a).

4.1 CHAPTER OBJECTIVES

This chapter describes the second listening experiment of the present study. These experiments were conducted to address the second and third research questions formulated in Chapter 1, namely whether NH listeners (question 2) and CI listeners (question 3) are better at perceiving prosody on a single-word level than at recognising vowels in single words in background noise. Suitable test materials were developed and acoustically analysed. A listening experiment was conducted to compare the perception of prosody with vowel perception in NH listeners and in CI listeners. Both vowel and prosody perception were tested in quiet and in an adaptive noise procedure, using a 2AFC test paradigm.

4.2 DEVELOPMENT AND ACOUSTIC ANALYSES OF TEST MATERIALS

4.2.1 Background

Results from the first set of listening experiments of this study (described in Chapter 3) suggested that NH listeners perform better on the recognition of prosody than word recognition on sentence level in SWN, and that prosody might therefore be more immune to the effects of noise than the cues needed for word recognition. However, to achieve the main aim of the present study (i.e. to compare the relative noise immunity of prosody and segmental speech information in NH and CI listeners), a number of remaining questions had to be addressed through further listening experiments. Firstly, it was necessary to determine whether the noise immunity of conditional prosody found in the first experiment can also be found in other prosodic contrasts. For the second listening experiment as described in this chapter, a new set of speech materials was therefore developed to measure the recognition of a linguistic function of prosody (a question/statement contrast) and an attitudinal function of prosody (a certain/hesitant attitude difference). Secondly, it was necessary to determine whether the noise immunity found in prosody on sentence level in the first experiment could also be observed on a single-word level. This was important to determine, since prosody may be particularly important in single-word utterances to differentiate speaker intent when contextual clues (i.e. semantic clues provided by additional words in the sentence, or word order clues) are not available. This can be seen, for example, in the case of the word “okay”, which is frequently used as a single-word utterance to fulfil a wide variety of functions (Gaines, 2011). Prosody plays an important role in differentiating the meaning of this word in different contexts (Gravano, Hirschberg and Benuš, 2012). Single-word utterances can also be used as either a statement or a question, with no inversion of word order to help the listener distinguish between the two possibilities, and prosody in this case is the only acoustic cue that can aid the listener in differentiating these (Chatterjee and Peng, 2008). A further motivation for using single words as test materials was that prosodic cues such as sentence stress and rhythm were eliminated from the segmental recognition task, thereby ensuring that prosody and segmental recognition could be tested separately. Thirdly, a limitation of the first set of listening experiments was that prosody recognition was tested in a closed set (2AFC) test paradigm, while word recognition was evaluated in an open set paradigm. The second listening experiment described in this chapter was designed to test prosody and segmental speech cue perception in identical test paradigms. The perception of prosodic cues is frequently evaluated in a 2AFC test paradigm, often because the very nature of these contrasts in everyday speech involves a choice between two alternatives. Examples of this include question/statement distinctions (e.g. Chatterjee and Peng, 2008; Most *et al.*, 2012), the discrimination of attitude as sarcastic or sincere (e.g. Cullington and Zeng, 2011), identification of phrase boundaries (Marslen-Wilson, Tyler, Warren, Grenier and Lee, 1992), and the resolution of sentence ambiguity based on prosodic cues (Price *et al.*, 1991). In contrast, phoneme recognition tasks often involve a larger set of alternatives for listeners to choose from. This means that the two tasks (phoneme and prosody perception) cannot be fairly compared, because the difficulty of the test paradigm is not the same. Because many prosody perception tasks call for the use of a 2AFC paradigm, the second listening experiment adopted this paradigm and cast the segmental cue recognition task into the same paradigm to provide a fair comparison between the two task types. The fourth question that stemmed from the work described in Chapter 3 was whether CI recipients would also perform better on prosody recognition than word or phoneme recognition in noise. Finally, the listening experiment described in Chapter 3 compared prosody recognition to the recognition of meaningful words in context-rich sentences. Word recognition in that task was supported by the perception of phonemes (vowels and consonants), as well as a number of other clues such as syntactic, semantic, and intonation clues. Since the main aim of the present work was to compare the relative noise immunity of prosody and segmental information, the second and third listening experiments were designed to compare the perception of specific prosodic cues to the perception of phonemes (vowels or consonants) without additional semantic or syntactic clues.

In order to address these questions that stemmed from the results of the first experiment, two types of speech material were developed for the second listening experiment. The first type was speech material aimed at measuring prosody perception on a single-word level, which included two different prosodic contrasts. The first prosodic contrast was a question/statement difference (a well-established linguistic function of prosody). In the test language of the present study, as in many other Germanic languages, the difference between questions and statements on sentence level is frequently indicated by an inversion of word order (Cruttenden, 1997; Ponelis, 1979). In single-word utterances, with no inversion of word order to indicate the difference between a question or statement, prosody is the only cue a listener can use to differentiate between the two types of utterance (Chatterjee and Peng, 2008). The acoustic cues of question prosody are reported to be a rising intonation pattern, or at least the use of higher pitch somewhere in the utterance (Borden *et al.*, 2007; Cruttenden, 1997; Meiring and Retief, 1991; Thorsen, 1980), and a higher speech rate (Van Heuven and Van Zanten, 2005). For the listening experiments described in this chapter, a single-word utterance (the word “coffee”, spelled “koffie” in Afrikaans) was recorded as either a statement or a question, with only prosodic cues differentiating the two utterance types.

The second prosodic contrast used in the present listening experiment denoted an attitudinal difference (certain versus hesitant) and was not such a well-documented prosodic difference. Therefore, the materials for this contrast were initially recorded from a larger number of speakers (n = 8), validated in a group of NH listeners (n = 12) in quiet, and thoroughly analysed (acoustically and statistically) to establish that this difference was recognisable to NH listeners, and to determine the acoustic differences between the prosody of the two attitudes. The word used as a vehicle for this prosodic expression was the word "okay", which is originally an English word, but is commonly used by Afrikaans speakers to convey the same meaning, and has been taken up into the Afrikaans lexicon with an altered spelling ("oukei") (Du Plessis, 2005). The reasons for using this prosodic contrast in the listening experiment were firstly that it provided an opportunity to investigate the perception of attitudinal prosody, which supplemented the use of linguistic and emotional prosody in the other experiments, and secondly, it represented a realistic use of prosody on a single-word level where semantic and syntactic clues do not communicate the speaker's attitude, using a word that is frequently used as a single-word utterance by speakers of the test language. In addition, after acoustic analyses of this prosodic contrast had been completed, it was discovered that durational cues played an important role in distinguishing the two attitudes (Van Zyl and Hanekom, 2013b). In light of the fact that question/statement contrasts are strongly related to intonation perception, which in turn is related to F0 perception, a particularly difficult task for CI users (Brown and Bacon, 2010; Chatterjee and Peng, 2008; Cullington and Zeng, 2011), it was considered useful to include a prosody perception task that relied more on the perception of duration differences, which appears to be an easier task for CI listeners (Moore and Glasberg, 1988).

The acoustic characteristics of the prosody materials that were examined included intensity, voice F0, durational and voice quality variables. Intensity can play a role in indicating emphasis on a particular syllable (Fry, 1955; Lieberman, 1960; Morton and Jassem, 1965). The height of the average F0 of each syllable could also indicate emphasis (Fry, 1958; Morton and Jassem, 1965), while F0 range across the entire utterance is often associated with the acoustic differences between different prosodic expressions (Breitenstein, Van Lancker and Daum, 2001; Hammerschmidt and Jürgens, 2007; Murray and Arnott, 1993). Duration or speech rate is frequently mentioned as an acoustic correlate of some forms of prosody expression (Fujie, Ejiri, Kikuchi and Kobayashi, 2006; Murray and Arnott, 1993; Williams and Stevens, 1972).

For the certain/hesitant contrast, voice quality was also analysed, as voice quality differences have previously been reported in attitudinal prosody (specifically sarcasm) (Cheang and Pell, 2008).

The second type of speech material developed for the listening experiment described in this chapter was intended to measure vowel recognition as an indication of segmental feature recognition. Vowels were used for this experiment as they have been reported to carry more information about sentence intelligibility than consonants, according to a study by Kewley-Port *et al.* (2007). In that study, recorded sentence materials were altered so that either vowels or consonants were replaced with speech-shaped noise, and intelligibility (sentence recognition) was measured in young NH and elderly hearing-impaired listeners. Across the two listener groups, sentences containing only vowels were significantly more intelligible than sentences containing only consonants (by a ratio of approximately 2:1). To conduct a fair comparison between prosody and vowel recognition, vowel recognition had to be measured in a 2AFC test paradigm. For this purpose, a number of vowel pairs had to be selected for the vowel recognition task. It was not considered practical to include all the vowels of the test language in the vowel recognition task, as testing each vowel against every other vowel in the 2AFC test paradigm would have resulted in 105 distinct vowel discrimination tasks. Rather than attempting this, three vowel pairs were carefully selected to represent specific acoustic differences. To select a suitable set of vowel pairs, a complete collection of 15 Afrikaans vowels was initially recorded and analysed, and three vowel pairs were selected according to the results of the acoustic analyses, on the basis that each vowel pair had specific acoustic differences and similarities. The acoustic characteristics of the vowels that were analysed were F1 and F2 frequencies and vowel duration. The importance of F1 and F2 frequencies for vowel discrimination has long been established (Assmann, Nearey and Hogan, 1982; Klatt, 1982; Miller, 1989; Nearey, 1989; Peterson and Barney, 1952). A number of studies also support the role of duration in vowel discrimination, especially with vowels lying close together in the F1-F2 vowel space (Ainsworth, 1972; Hillenbrand, Getty, Clark and Wheeler, 1995; Tartter, Hellman and Chute, 1992). Vowel pairs were selected so that each pair would represent a specific acoustic difference (F1, F2 or duration). The results of the acoustic analyses below provide further details on the choice of vowel pairs.

4.2.2 Recording and validation of speech material: methods and results

Digital recording of the speech materials was conducted in a double-walled sound booth, using an M-Audio Fast Track Pro external sound card and a Sennheiser ME62 microphone placed on a microphone stand 20 cm from the speaker's mouth. Recorded waveforms were edited using *Praat* software by removing unwanted silences (leaving silences of 100 ms before and after the utterance) and re-scaling the intensity of each utterance to 70 dB SPL before saving the material to hard disc in .wav format. Re-scaling intensities preserved relative intensity changes and cues within utterances, while eliminating any accidental intensity differences between utterances which might have occurred during recording and ensured accurate SNRs in the noise experiment.

For the prosody recognition task, test materials were developed that used the same word to express both versions of each contrast to ensure that the contrasts were purely prosodic and not related to the content of the utterance. For the question/statement contrast, the word "coffee" ("koffie" in Afrikaans, with pronunciation very similar to English) was used. Four speakers participated in the recordings (two male). Speakers had normal hearing and speech, and were native speakers of Afrikaans aged between 21 and 28 years. Untrained speakers were used, as the aim was to record speech materials that represented the speech of typical speakers, not trained actors. Fifteen interrogative (question) and 15 declarative (statement) versions were recorded from each speaker. The interrogative versions of the utterance were elicited by asking speakers to produce the word "koffie" in a manner as if asking someone if they would like a cup of coffee. The declarative version of the word was elicited by asking the talker a question (such as "what would you like to drink?"), and instructing them to produce the word "koffie" each time as a response to the question. The recorded materials were validated in a sample of NH listeners (n = 4) in quiet to ensure that recognition accuracy was ≥95% for each speaker's recordings across listeners.

The certain/hesitant contrast was represented by the word "okay", with half of the utterances produced with certainty, and the other half with hesitation or reluctance. The same four speakers used for the question/statement recordings participated in these recordings. Initially, however, four additional speakers (two male) were also recorded, in order to investigate the acoustic characteristics of the certain/hesitant contrast in a larger sample of speakers and to establish its validity as a recognisable prosodic contrast in NH listeners. Fifty repetitions of the word "okay" were recorded from each speaker, 25 of which conveyed unreserved (certain) permission, and 25 conveying reluctant (hesitant) permission. To elicit these utterances, a scenario was described to the speaker where someone would request to visit them on one of two different days. Speakers were informed that Friday would suit them in this scenario, whereas Monday would be inconvenient. Each elicited utterance was preceded by a question from the examiner (e.g. "Can I come on Monday?"), and speakers had to respond using only the word "okay", keeping in mind whether the requested time would be convenient or not. The same scenario was used to elicit all utterances, and this merits some explanation. Noting that "okay" performs a variety of communicative functions (Gaines, 2011), using the same scenario across elicitations ensured that the utterance was used to fulfil the same function in all instances. Also, different scenarios could potentially induce a variety of emotions in the speakers, which had to be avoided. Certain and hesitant elicitations were alternated to reduce task repetitiveness. Speakers were encouraged to produce each utterance as an authentic response to the examiner's question. Acoustic analyses of the recorded materials showed a high degree of variability within each speaker's collection of utterances, affirming that speakers were producing authentic responses rather than a rote repetition of the same utterance. Recorded speech materials were validated in 12 NH listeners (university students aged 19 to 29 years) using a 2AFC test paradigm. No prior training was given to listeners and no feedback was given during testing. This was to ensure that listeners would respond to the stimuli with everyday listening experiences as their only frame of reference. Different speakers' recordings were presented in counterbalanced order across listeners. Table 4.1 below shows the results of this validation procedure. The speakers whose recordings were used in the listening experiment in noise (described in section 4.3 of this chapter) are indicated in the table as FS2, FS3, MS1, and MS2.

Table 4.1: Results (percentage correct recognition) from the validation procedure for the certain/hesitant prosodic contrast obtained from NH listeners in quiet. Female speakers are denoted FS1-4; male speakers are M1-4; Ave. denotes average; SD denotes standard deviation.

	FS1	**FS2**	**FS3**	**FS4**	**MS1**	**MS2**	**MS3**	**MS4**	**Ave.**
Listener1	88.00	90.00	96.00	56.00	94.00	94.00	94.00	88.00	87.50
Listener2	92.00	90.00	92.00	80.00	92.00	96.00	96.00	96.00	91.75
Listener3	96.00	84.00	94.00	82.00	94.00	98.00	98.00	90.00	92.00
Listener4	92.00	78.00	90.00	74.00	96.00	98.00	100.00	94.00	90.25
Listener5	94.00	82.00	94.00	76.00	92.00	96.00	86.00	92.00	89.00
Listener6	92.00	98.00	84.00	66.00	90.00	82.00	84.00	94.00	86.25
Listener7	84.00	80.00	90.00	64.00	94.00	100.00	88.00	92.00	86.50
Listener8	92.00	78.00	92.00	82.00	94.00	92.00	94.00	90.00	89.25
Listener9	90.00	82.00	92.00	66.00	94.00	96.00	94.00	94.00	88.50
Listener10	92.00	82.00	92.00	64.00	90.00	96.00	82.00	82.00	85.00
Listener11	94.00	90.00	100.00	80.00	98.00	98.00	86.00	100.00	93.25
Listener12	78.00	76.00	82.00	76.00	92.00	96.00	90.00	86.00	84.50
Ave.	**90.33**	**84.17**	**91.50**	**72.17**	**93.33**	**95.17**	**91.00**	**91.50**	**88.65**
SD	*4.96*	*6.52*	*4.83*	*8.63*	*2.31*	*4.63*	*5.82*	*4.76*	*8.77*
Ave. (selected samples)		**95.00**	**94.56**		**99.44**	**96.67**			**96.42**
SD (selected samples)		*6.42*	*8.38*		*2.11*	*5.18*			*6.20*

Average scores for individual speakers across listeners varied between 72.17% and 95.17%. Using the results from the validation procedure, utterances that were correctly classified by at least 10 out of 12 listeners (i.e. significantly above chance, p < 0.05) were selected for the acoustic analyses. Recognition results for the samples that were selected for the listening experiments (n = 120) are indicated in Table 4.1 for the four speakers whose recordings were used in the noise experiment. Scores obtained from these four speakers' selected samples (30 utterances from each speaker, 15 certain and 15 hesitant) varied between 94.56% and 99.44% across listeners.

For the vowel discrimination tasks, a complete set of 15 Afrikaans vowels was recorded from each speaker in a /pVOWELt/ format and analysed in order to enable the selection of a representative subset of vowel pairs for the listening experiment, in light of the acoustic characteristics of all the vowels. The first and last consonants

(/p/ and /t/) were selected as they are both voiceless plosives, which enabled accurate isolation of the vowel segment for analysis. For each /pVOWELt/ combination containing a different vowel, 15 versions were recorded from each speaker. This enabled acoustic analysis of the vowels using average values across several utterances of the same vowel, thereby including in the analyses and listening experiments the natural variations that occur when a speaker repeats the same utterance (Peterson and Barney, 1952). Because of the ease of the vowel recognition task, and since a quiet condition with NH listeners was included in the listening experiment, the recorded vowel materials were not subject to validation in NH listeners. However, multiple repetitions of each vowel (n = 36) were recorded from each speaker, and a qualified speech and language therapist selected from these repetitions 15 versions of each vowel that were all considered to be good samples of the target vowel.

4.2.3 Acoustic analyses: methods and results

Because the certain/hesitant prosodic contrast is not a well-documented prosodic pattern in existing literature, a more detailed acoustic analysis was conducted on the recordings of this contrast from all eight speakers initially recorded. These methods and results are reported first. This is followed by a report on the methods and results of the acoustic analyses on the question/statement contrast and a summary of the certain/hesitant analyses on the four speakers used in the listening experiment to enable a quick comparison between the acoustic characteristics of the two types of prosody. Finally, methods and results of the acoustic analyses on the vowel materials are reported.

4.2.3.1 Acoustic analyses of initial recordings of certain/hesitant prosody

Acoustic characteristics of the certain/hesitant materials were investigated using *Praat* by examining aspects of voice F0, duration, intensity and voice quality in each utterance. As a number of the distributions deviated significantly from a standard normal distribution, the non-parametric Mann-Whitney test was used to determine whether differences between the two conditions were significant ($p < 0.05$ or smaller). Average F0 and F0 range across the utterance were extracted, with assumed

F0 ranges of 100 to 500 Hz for female speakers and 65 to 300 Hz for male speakers. The duration of the first syllable was measured from the onset up to the end of the silence preceding the plosive noise of the /k/, and the duration of the second syllable from the beginning of the release noise of /k/ to the end of phonation. The overall intensity (across the frequency spectrum) and voice quality of the voiced parts of the first and second syllables were determined separately. Voice quality was analysed through extraction of the HNR with cross-correlation analysis in *Praat*. The HNR reflects the degree of periodicity in the utterance and consequently voice quality in areas where a valid F0 has been determined. The results of these acoustic analyses are reported in Table 4.2. Note that only utterances that were correctly identified by ten or more out of 12 NH listeners in the validation procedure were included in the analyses, thus the differences in number of utterances analysed for each speaker.

Table 4.2: Mean values of acoustic parameters for certain (C) and hesitant (H) prosody. Female speakers are FS1-FS4, male speakers MS1-MS4. For significant differences ($p < 0.05$ or smaller) the greater of the two values is indicated in bold-face. S1 = 1^{st} syllable; S2 = 2^{nd} syllable.

Prosody:	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**
Speaker:	**FS1**		**FS2**		**FS3**		**FS4**	
Number of utterances (n)	22	20	19	15	20	23	12	10
Average pitch (Hz)	**246.44**	214.23	**208.23**	193.61	**271.37**	225.10	260.78	**288.53**
Pitch range (Hz)	127.31	**143.89**	94.20	**126.79**	**198.30**	97.85	194.22	209.44
Duration S1 (s)	0.19	**0.28**	0.10	**0.14**	0.15	**0.19**	0.16	0.17
% increased duration	43.05		37.01		24.25		8.41	
Duration S2 (s)	0.27	**0.40**	0.23	**0.32**	0.27	**0.53**	0.27	**0.38**
% increased duration	49.91		38.66		98.54		39.62	
Duration aspiration noise (s)	0.07	**0.20**	0.04	**0.09**	0.00	**0.03**	0.05	**0.11**
Total duration (s)	0.53	**0.88**	0.38	**0.54**	0.42	**0.75**	0.48	**0.66**
Intensity S1 (dB)	**74.30**	72.10	**71.74**	68.42	**72.76**	69.36	**74.36**	66.04
Intensity S2 (dB)	71.60	**72.57**	72.84	**73.49**	**72.72**	72.25	71.52	**72.98**
Intensity difference (S2-S1) (dB)	**-2.70**	0.47	1.10	**5.07**	-0.04	**2.89**	-2.84	**6.94**
Harmonics-to-noise ratio S1 (dB)	**15.81**	13.68	10.32	**13.85**	12.31	**13.99**	12.34	11.72
Harmonics-to-noise ratio S2 (dB)	**17.12**	13.84	14.69	14.98	15.70	**19.88**	12.72	**15.72**
Total significant differences:	*11/11*		*10/11*		*11/11*		*8/11*	
Speaker:	**MS1**		**MS2**		**MS3**		**MS4**	
Number of utterances (n)	21	21	25	23	25	18	25	19
Average pitch (Hz)	118.81	109.34	**125.22**	99.57	**143.41**	133.55	**136.73**	117.77
Pitch range (Hz)	77.37	80.82	55.41	**64.47**	61.57	74.85	61.18	60.13
Duration S1 (s)	0.19	**0.28**	0.16	**0.28**	0.19	**0.40**	0.16	**0.18**
% increased duration	47.20		72.06		113.26		12.51	
Duration S2 (s)	0.27	**0.48**	0.14	**0.28**	0.21	**0.27**	0.21	**0.36**
% increased duration	78.18		93.17		28.88		73.82	
Duration aspiration noise (s)	0.01	**0.05**	0.00	**0.06**	0.00	0.00	0.00	**0.03**
Total duration (s)	0.47	**0.82**	0.30	**0.62**	0.40	**0.67**	0.37	**0.56**
Intensity S1 (dB)	68.18	68.51	**72.29**	68.08	71.57	70.57	70.38	68.68
Intensity S2 (dB)	**73.31**	72.86	73.82	**74.73**	73.44	73.49	**73.29**	72.40
Intensity difference (S2-S1) (dB)	5.12	4.36	1.52	**6.65**	1.87	2.92	2.91	3.73
Harmonics-to-noise ratio S1 (dB)	7.76	6.74	6.87	6.44	8.76	**15.70**	7.05	7.13
Harmonics-to-noise ratio S2 (dB)	10.48	**12.57**	10.33	10.22	13.64	13.23	8.96	**11.39**
Total significant differences:	*6/11*		*9/11*		*5/11*		*7/11*	

The effect sizes of the differences between the two prosodic types (certain and hesitant) were calculated according to the Mann-Whitney test z-score and the total number of observations on which z is based (Field, 2009). Effect sizes are reported in Table 4.3.

Table 4.3: Effect sizes of differences between certain and hesitant versions for each speaker (female speakers FS1-FS4; male speakers MS1-MS4). Effect sizes representing differences that were statistically significant (p<0.05 or smaller) are depicted in bold-face. S1 = 1st syllable; S2 = 2nd syllable.

Speaker:	**FS1**	**FS2**	**FS3**	**FS4**	**MS1**	**MS2**	**MS3**	**MS4**
Average pitch (Hz)	**-0.67**	**-0.60**	**-0.73**	**0.52**	-0.26	**-0.78**	**-0.45**	**-0.72**
Pitch range (Hz)	**-0.31**	**-0.43**	**-0.60**	-0.15	-0.03	**-0.28**	-0.24	-0.01
Duration S1 (s)	**-0.84**	**-0.76**	**-0.43**	-0.30	**-0.83**	**-0.86**	**-0.84**	**-0.61**
Duration S2 (s)	**-0.86**	**-0.84**	**-0.86**	**-0.84**	**-0.86**	**-0.86**	**-0.78**	**-0.85**
Duration aspiration noise (s)	**-0.70**	**-0.64**	**-0.38**	**-0.65**	**-0.63**	**-0.68**	0.00	**-0.61**
Total duration (s)	**-0.84**	**-0.76**	**-0.86**	**-0.84**	**-0.86**	**-0.86**	**-0.85**	**-0.85**
Intensity S1 (dB)	**-0.56**	**-0.60**	**-0.65**	**-0.84**	-0.08	**-0.73**	-0.25	-0.26
Intensity S2 (dB)	**-0.35**	**-0.33**	**-0.42**	**-0.67**	**-0.50**	**-0.48**	-0.02	**-0.42**
Harmonics-to-noise ratio S1 (dB)	**-0.32**	**0.56**	**0.35**	-0.06	-0.17	-0.12	**0.70**	-0.04
Harmonics-to-noise ratio S2 (dB)	**-0.52**	-0.20	**-0.49**	**-0.67**	**-0.25**	-0.39	-0.14	**-0.57**
Average effect size:	*-0.60*	*-0.46*	*-0.51*	*-0.45*	*-0.45*	*-0.60*	*-0.29*	*-0.49*

According to the results reported in Table 4.2, average F0 across the utterance was significantly higher in the certain version for six of the eight speakers. Speaker FS4 produced a higher F0 average in the hesitant version, while speaker MS1 showed no significant difference between the F0 averages of certain and hesitant versions. The F0 range (difference between maximum and minimum across the utterance) differed significantly between prosodic conditions for four speakers, three of which used a significantly greater range for reluctant prosody (FS1, FS2, MS2), whereas one speaker (FS3) produced a greater F0 range in the baseline condition.

All eight speakers used significantly longer total word and second syllable duration for hesitant utterances. The first syllable had a significantly greater duration in the hesitant versions of seven speakers. Except for MS3, the durational increase of the second syllable was greater than that of the first. MS3 used a longer first syllable, sometimes preceded by glottal fry or nasalisation of the vowel, as a prominent cue of hesitant prosody. All the other speakers also produced an audible aspiration noise at the end of most utterances, and this noise was significantly longer in the hesitant versions of these speakers. Table 4.2 indicates the percentage of duration increase for each of the two syllables in the hesitant versions.

The intensity of the first syllable was significantly greater in the certain version for five speakers, while the second syllable's intensity was significantly greater in the hesitant versions of six speakers. HNRs showed that voice quality differed significantly in one or both syllables for seven speakers, with FS1 having a higher HNR in both syllables for utterances expressing certainty, FS3 producing higher HNR in both syllables for hesitant utterances, and FS2, FS4, MS1, MS3 and MS4 producing higher HNRs for hesitant prosody on either the first or the second syllable.

An average intonation contour of the final syllable was determined for each speaker in both conditions. This required the elimination of duration differences between utterances without affecting their spectral characteristics, which was accomplished using phase vocoding methods (Ellis, 2002). The certain and hesitant intonation curves for each speaker were then compared using Zhao's *Z*-statistic for comparing trend curves (Zhao, 2011). The intonation contours of the final syllable, as averaged over all the sampled utterances for each speaker, are shown in Fig. 4.1a (female speakers) and Fig. 4.1b (male speakers). Certain and hesitant curves of each speaker were compared using a *Z*-statistic (Zhao, 2011), with resulting *p*-values (Table 4.4) showing that five speakers produced intonation curves that differed significantly between the two versions. Table 4.4 also shows which half (first or last) of the utterances differed significantly.

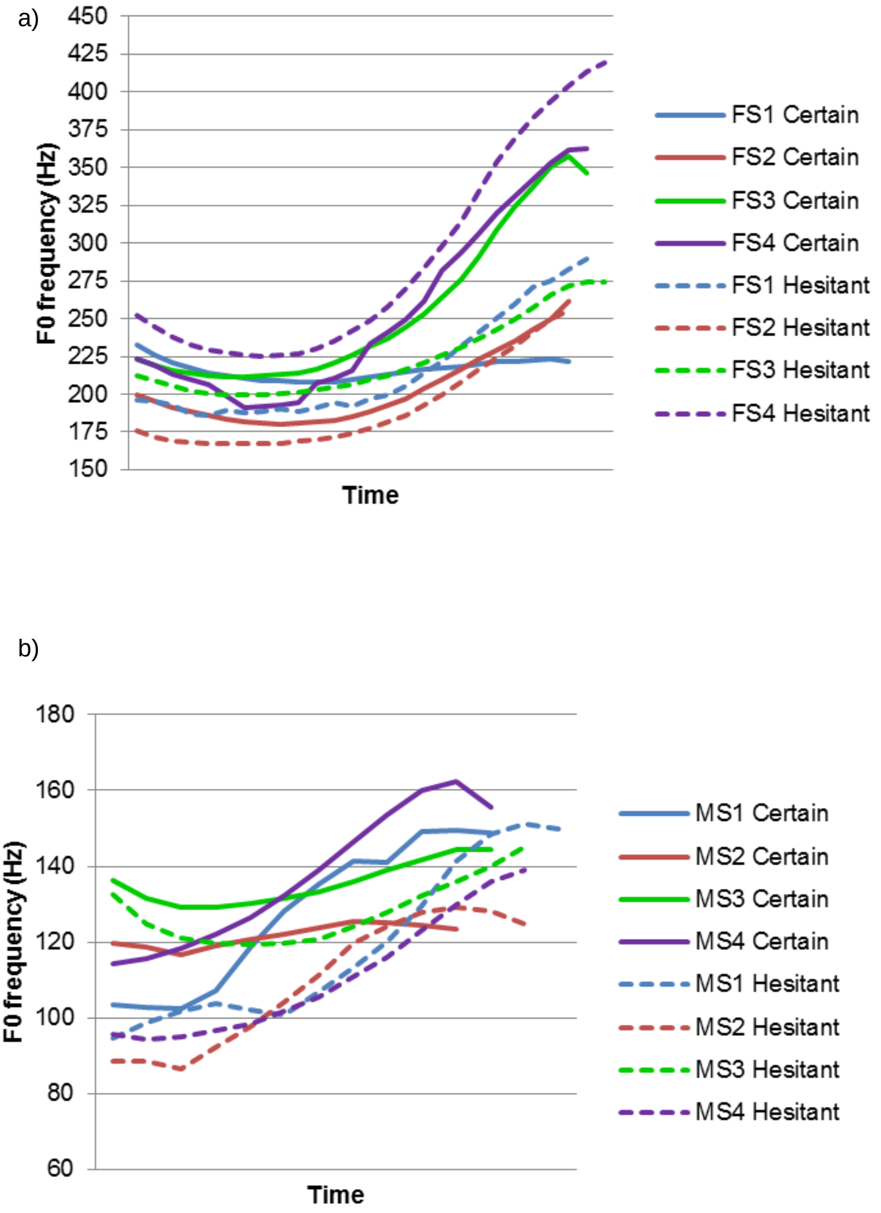

Figure 4.1: Average intonation contours of final syllables of certain/hesitant utterances as produced by female speakers numbered FS1 to FS4 (panel a), and male speakers numbered MS1 to MS4 (panel b), showing certain and hesitant utterances separately.

Table 4.4: Results of the *Z*-statistic comparing the two utterance types (certain and hesitant) of each speaker (female speakers FS1-FS4; male speakers MS1-MS4). Significant differences ($p < 0.05$) are depicted in bold-face.

	Whole curve (*p*-value)	**1st half (*p*-value)**	**2nd half (*p*-value)**
FS1	0.430	**0.010**	**0.010**
FS2	**0.005**	**<0.001**	0.213
FS3	**<0.001**	0.142	**<0.001**
FS4	**0.004**	**0.016**	**0.037**
MS1	0.052	0.156	0.134
MS2	**0.011**	**<0.001**	0.397
MS3	0.056	0.153	0.119
MS4	**<0.001**	**<0.001**	**<0.001**

Logistic regression analyses were carried out to explore the relative importance of the different cues in predicting to which category (certain or hesitant) an utterance belonged. Different models were tested with predictors selected from the cues in Table 4.2. All validated utterances were included in the analyses. Utterances from male and female speakers were analysed separately, as especially F0 parameters differed substantially between genders. Nagelkerke's R^2 was used as indicator for the variance accounted for. For both genders, all models that could account for more than 90% of the variance in the data set included duration as a predictor. Conversely, all models excluding duration as a predictor accounted for at most 66% of the variance. This confirms observations from Table 4.2 regarding duration being the most consistent cue. However, models that included only duration as a predictor did not fully explain the data (male speakers, deviance = 47.9, degrees of freedom = 175, Nagelkerke's R^2 = 0.895; female speakers, deviance = 44.6, degrees of freedom = 139, Nagelkerke's R^2 = 0.876). Adding other predictors improved the models and evidence of cue trading relationships was observed. For example, for female speakers, models that included duration and either the intensity of both syllables (deviance = 24.4, degrees of freedom = 137, Nagelkerke's R^2 = 0.937) or average F0 and F0 range (deviance = 22.0, degrees of freedom = 136, Nagelkerke's R^2 = 0.944) did not differ significantly (p = 0.124).

4.2.3.2 Acoustic analyses of question/statement and final collection of certain/hesitant prosody materials

Acoustic analyses of the question/statement prosody materials examined the intensity of each syllable, the average F0 of each syllable, the range of the voice F0 across both syllables, and the duration of the entire utterance. Intensity was measured in dB SPL, and reflected the rms value of the intensity of each syllable separately. Intensity analysis included only the vowels, since all the consonants were voiceless and therefore did not reflect voice intensity. The intensity of the two syllables were analysed separately, as the characteristic of interest was the relative difference in intensity of the two syllables as an indication of syllable stress. The average and range of voice F0 were determined using *Praat*, and were expressed in Hz. The same acoustic cues were investigated in the final selection of certain/hesitant utterances (30 utterances, of which 15 were hesitant, from each of the four speakers used for the question/statement prosody recordings) and are summarised with the results from the question/statement analyses in Table 4.5. Note that in this final analysis, speakers were numbered differently than in the first analysis of the certain/hesitant recordings with eight speakers. In Table 4.5, speakers MS1, MS2, and FS2 represent the same speakers as in Table 4.2, but speaker FS1 in Table 4.5 corresponds to speaker FS3 in Table 4.2.

The data in Table 4.5 indicate that the question-versus-statement utterances differed in both F0 and intensity characteristics, with differences that exceed the difference limens (DLs) reported for NH, and in some cases, for CI listeners. The average duration difference between question and statement utterances was, however, smaller than DLs reported for both NH and CI listeners (Moore and Glasberg, 1988; Small and Campbell, 1962). F0 and intensity cues therefore seemed to be the most prominent cues for this contrast. Certain/hesitant utterances differed in intensity, F0, and duration for most speakers. The intensity differences produced by speaker MS1, however, were below DLs reported in existing literature, even for NH listeners, and the difference in F0 of the second syllable was only above the NH DL, and not above the DL reported for CI users (Rogers, Healy and Montgomery, 2006). According to the results reported in section 4.2.3.1, cue trading relationships existed between other cues, but the most consistent cue was a duration difference between certain and hesitant utterances.

Table 4.5: Results of acoustic analyses on prosody materials. Values indicate means; standard deviations are shown in brackets. S1 and S2 denote first and second syllables respectively. * indicates values greater than difference limens (DLs) for NH, ** indicates DLs larger than NH and CI DLs. Intensity and frequency DLs are from Rogers *et al.* (2006), and duration DLs from Small and Campbell (1962)[3].

Statement/Question contrast

			FS1		FS2		MS1		MS2	
Intensity	S1	Statement	77.11	(0.6)	75.75	(1.0)	76.53	(0.4)	77.43	(0.4)
(dB SPL)		Question	73.84	(0.7)	74.68	(1.3)	71.74	(1.6)	75.14	(1.1)
		Difference	*-3.27***		*-1.07*		*-4.79***		*-2.29**	
	S2	Statement	67.70	(2.7)	68.28	(3.0)	67.95	(1.7)	59.77	(2.9)
		Question	72.60	(0.8)	71.75	(1.6)	73.61	(0.6)	70.76	(1.6)
		Difference	*4.90***		*3.47***		*5.66***		*10.99***	
Mean F0	S1	Statement	207.93	(8.4)	168.42	(7.2)	96.97	(7.4)	103.7	(8.1)
(Hz)		Question	220.87	(5.5)	173.28	(6.0)	104.00	(6.6)	87.22	(11.5)
		Difference	*12.94**		*4.87**		*7.02**		*-16.49**	
	S2	Statement	188.19	(9.8)	163.87	(7.8)	78.58	(7.5)	71.16	(4.7)
		Question	381.65	(9.2)	232.91	(12.2)	151.23	(6.8)	121.4	(22.1)
		Difference	*193.46***		*69.04***		*72.65***		*50.23***	
F0 range (Hz)		Statement	54.42	(10.2)	47.21	(19.6)	33.31	(14.2)	40.80	(30.8)
		Question	257.61	(17.7)	105.82	(20.8)	63.74	(10.4)	57.06	(17.9)
		Difference	*203.19*		*58.61*		*30.42*		*16.25*	
Duration	(s)	Statement	0.39	(0.02)	0.36	(0.02)	0.40	(0.02)	0.27	(0.02)
		Question	0.45	(0.02)	0.35	(0.02)	0.42	(0.02)	0.34	(0.02)
		Difference	*0.06*		*-0.01*		*0.02*		*0.06*	

Certain/Hesitant contrast

			FS1		FS2		MS1		MS2	
Intensity	S1	Certain	72.52	(1.5)	72.04	(2.2)	68.68	(6.8)	72.81	(1.9)
(dB SPL)		Hesitant	69.12	(3.7)	68.42	(1.7)	68.85	(2.3)	67.96	(2.1)
		Difference	*-3.40***		*-3.62***		*0.17*		*-4.85***	
	S2	Certain	72.74	(0.6)	72.86	(0.8)	73.08	(1.6)	73.64	(1.3)
		Hesitant	72.33	(0.4)	73.49	(0.5)	72.81	(0.2)	74.71	(0.5)
		Difference	*-0.41*		*0.63*		*-0.27*		*1.07*	
Mean F0	S1	Certain	279.65	(75.7)	209.56	(19.6)	97.43	(4.0)	133.9	(13.9)
(Hz)		Hesitant	215.21	(25.2)	164.09	(6.8)	97.64	(4.9)	83.71	(16.4)
		Difference	*-64.44***		*-45.47***		*0.21*		*-50.21***	
	S2	Certain	263.63	(16.4)	209.41	(10.5)	119.48	(27.6)	121.5	(11.7)
		Hesitant	225.27	(13.6)	200.17	(13.6)	113.65	(18.8)	111.7	(6.7)
		Difference	*-38.36***		*-9.24**		*-5.83**		*-9.81**	
F0 range (Hz)		Certain	193.54	(80.4)	84.54	(46.0)	70.51	(32.2)	53.67	(26.0)
		Hesitant	93.73	(41.7)	126.79	(22.9)	73.57	(17.8)	65.13	(10.1)
		Difference	*-99.81*		*42.25*		*3.06*		*11.46*	
Duration	(s)	Certain	0.42	(0.03)	0.37	(0.05)	0.47	(0.05)	0.30	(0.04)
		Hesitant	0.65	(0.1)	0.54	(0.07)	0.83	(0.1)	0.59	(0.1)
		Difference	*0.23***		*0.17***		*0.36***		*0.29***	

[3] The duration DLs reported here are for NH listeners (Small and Campbell, 1962); no report of such DLs measured in CI users could be found in existing literature. However, evidence from existing literature indicates that the temporal resolution of CI users is close to that of NH listeners (Moore and Glasberg, 1988).

The average intonation contours of the final syllable of the question/statement recordings were also determined for each speaker by eliminating duration differences using phase vocoding methods (Ellis, 2002). These contours are depicted in Fig. 4.2.

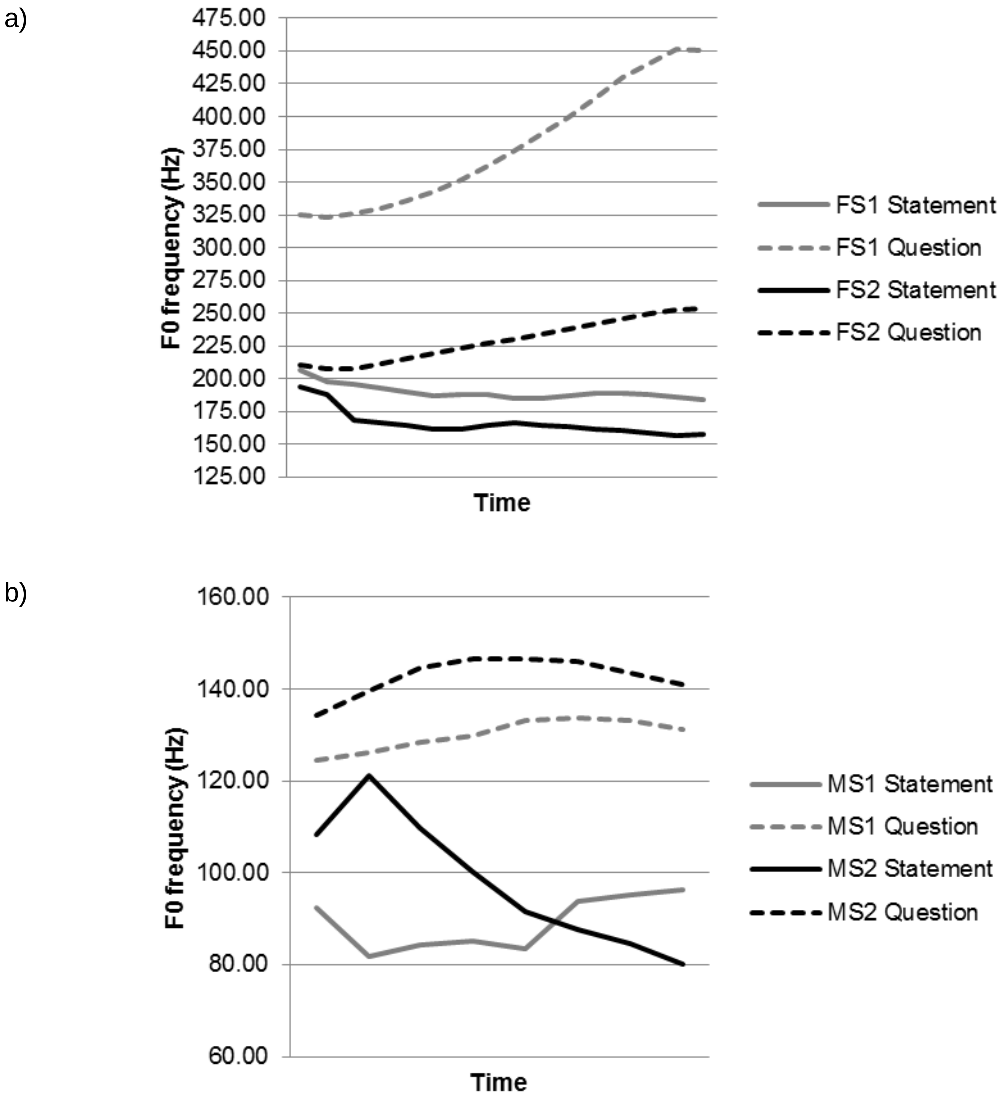

Figure 4.2: Average intonation contours of final syllables of question/statement recordings as produced by female speakers FS1 and FS2 (panel a), and male speakers MS1 and MS2 (panel b), showing question and statement utterances separately.

The intonation contours depicted in Fig. 4.2 show that both female speakers used a falling or flat intonation contour for statements, and a rising contour for questions. MS1 produced intonation contours with a slight rise for both question and statement utterances, while MS2 produced a falling contour for statements. Interrogatives produced by MS2 showed a slight rise followed by a fall in intonation.

4.2.3.3 Acoustic analyses of vowel materials

The vowels contained in the /pVOWELt/ utterances were analysed to determine the F1 and F2 frequencies and duration of the vowel. The start and end times of the vowels had to be determined first, since the extraction of the formant frequencies depended on the accurate definition of the vowel segments. To ensure accurate analysis, each vowel's beginning and end times were identified manually with the help of a Matlab graphic user interface (GUI) developed specifically for this purpose. The selection was made based on visual inspection of the waveform (zoomed to achieve a high resolution) and auditory inspection of waveform segments around the beginning and ending of the utterance. Formant frequencies were subsequently extracted from one time frame spanning the middle 80% of the vowel using the formant estimation algorithm of *Praat* that is based on linear predictive coding.

The results of the acoustic analyses of the recorded vowels are shown in Table 4.6, which shows duration analysis results, and Fig. 4.3, which shows the vowels on an F1-F2 vowel plane.

Table 4.6: Average duration of vowels selected for the listening experiment

	FS1	**FS2**	**MS1**	**MS2**	**Average**
Average duration /pɔt/ (s)	0.13	0.13	0.11	0.10	**0.12**
Average duration /pɛt/ (s)	0.13	0.13	0.10	0.09	**0.11**
Difference (/pɛt/ - /pɔt/) (s)	*0.00*	*0.00*	*-0.01*	*-0.01*	***0.00***
Average duration /pat/ (s)	0.12	0.14	0.13	0.10	**0.12**
Average duration /put/ (s)	0.09	0.10	0.09	0.07	**0.09**
Difference (/put/ - /pat/) (s)	*-0.03*	*-0.04*	*-0.04*	*-0.03*	***-0.04***
Average duration /pɛt/ (s)	0.13	0.13	0.10	0.09	**0.11**
Average duration /pɛ:t/ (s)	0.27	0.26	0.32	0.22	**0.27**
Difference (/pɛ:t/ - /pɛt/) (s)	*0.14*	*0.13*	*0.22*	*0.13*	***0.16***

The average difference in duration between the recorded /pɔt/ and /pɛt/ utterances ranged between 0 and 0.01 seconds for the different speakers, and duration differences for /pat/ and /put/ ranged 0.03 and 0.04 seconds across speakers. The average duration differences for /pɛt/ and /pɛː t/ were between 0.13 and 0.22 seconds for the different speakers.

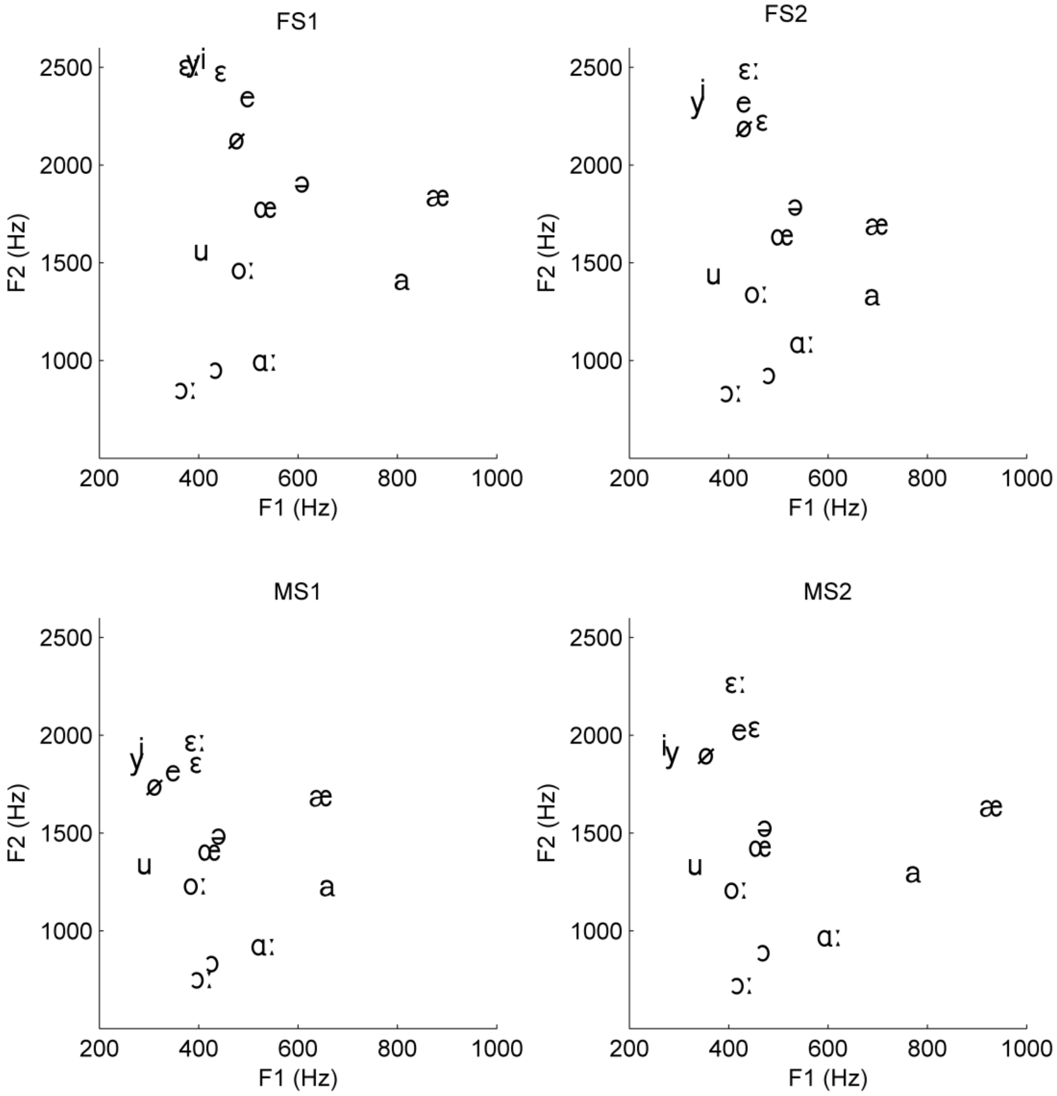

Figure 4.3: F1-F2 vowelspace of the 15 Afrikaans vowels recorded from four speakers (FS1 and FS2 are female speakers, MS1 and MS2 are the male speakers). Values indicate average frequencies calculated from 15 distinct utterances of each vowel.

These results were used to select vowel pairs that could be used for the listening experiment in a 2AFC test paradigm based on their acoustic characteristics. Since only a limited number of vowel pairs could be used, the selection had to represent specific acoustic differences and similarities. A number of acoustic characteristics of the vowel contribute in combination to vowel identification accuracy. These include the availability of formant frequency information, vowel duration, formant movement over time, and distance between vowels (or dispersion of vowels) in the F1-F2 vowel space (Neel, 2008). In the present study, vowels were selected to differ in the three most prominent steady state cues reported in literature (F1, F2 and duration, as discussed in section 4.2.1). The first vowel pair (/pɔt/ and /pɛt/) was selected to differ primarily in terms of their average F2 frequencies, and having similar average F1 frequencies and durations for all four speakers. Despite the relatively large difference in F2, this vowel pair posed a difficult task to listeners owing to the highly similar F1 frequencies, a cue that has been shown to be particularly important for vowel recognition in noise (Parikh and Loizou, 2005). Swanepoel *et al.* (2012) have also shown that while F2 is more important than F1 in quiet and low noise conditions, listeners increase reliance on F1 as noise levels increase. The second vowel pair (/pat/ and /put/) differed primarily in F1 frequency, while having similar F2 frequencies and durations. As only either F1 or F2 differed within a vowel pair, the F2-F1 difference between the two vowels of a vowel pair was relatively large. Although this difference may have had an influence on the degree of difficulty in vowel comparisons, the work of Neel (2008) suggests that distinctiveness of vowels based on formant frequencies, duration and formant movement over time may more strongly influence vowel identification than dispersion in vowel space. The third pair (/pɛt/ and /pɛ: t/) differed mainly in duration, while being closely spaced in the F1-F2 plane. This pair was selected in order to examine the noise immunity of duration as a cue to vowel identity in cases where formants are very similar and could not be used as cues to distinguish these vowels. This was important to consider, particularly since CI users are reported to have relatively good temporal resolution (close to that of NH listeners) (Moore and Glasberg, 1988), and durational cues were particularly prominent in the certain/hesitant prosodic contrast (Van Zyl and Hanekom, 2013b).

4.2.4 Discussion

Perceptual validation confirmed that NH listeners were able to discriminate accurately between certain/hesitant and question/statement prosody in the recorded materials of all the speakers, despite the inter-speaker differences in acoustic cues. Acoustic analyses of the certain/hesitant materials revealed that the cue for hesitant prosody that was used with greatest consistency across speakers was an increase in duration. The importance of duration as a cue was confirmed by the effect sizes of the differences between prosodic versions and the amount of variance that this cue accounted for. Increased duration was also reported by Fujie *et al*. (2006) as an important cue of a negative response attitude, in addition to a smaller F0 range (which was not found to be a consistent cue in the present study), but the consistency of the cues across speakers was not reported. Other cues were used less consistently and the logistic regression analysis pointed to cue-trading relationships.

Observations regarding average F0 show some agreement with findings on other types of non-linguistic prosody such as sarcasm, where a reduction in F0 has been shown to be the most consistent prosodic cue (Cheang and Pell, 2008), and emotional prosody, where F0 changes constitute an essential acoustic cue (Williams and Stevens, 1972). Word-final intonation has been reported to be important in the interpretation of the word “okay” in isolation (Gravano *et al.*, 2012). Some of the speakers in the present study used the intonation contour to differentiate certain and hesitant attitudes, but different speakers applied intonation differently. Speakers FS1, MS2 and MS3 (as numbered in the original eight-speaker recordings) produced falling intonation contours in utterances conveying certainty and rising contours in hesitant utterances, corresponding to findings regarding uncertainty in factual answers (Brennan and Williams, 1995), while the other speakers produced some form of rising pitch for both utterance types. Statistical comparison of the intonation curves showed that comparing the entire curve of utterances expressing certainty with the entire curve of hesitant utterances may be useful in cases such as those of speakers FS3 and MS4, where the two curves did not have any interaction, but may produce less informative results in cases such as that of FS1, where the two curves clearly differed in shape (one rising and one falling or flat). Speakers may use intensity as a cue to their attitude, but again the manner in which they apply this cue varies across speakers. Previous studies on cues for uncertainty in responses to factual questions did not report findings on intensity differences or values (Brennan and Williams, 1995; Krahmer and Swerts, 2005). Voice quality cues did not show consistent patterns across speakers, and effect sizes were small in comparison to most of the other investigated parameters. Higher HNRs observed in the hesitant versions of six of the speakers are in contrast to findings reported in a study on sarcasm, where a negative attitude corresponded to a lower HNR (Cheang and Pell, 2008).

Analyses of the question/statement contrasts showed that all four speakers used a higher intensity and a higher average F0 of the second syllable in interrogative (question) utterances than in statements. The differences between the average values for these variables across the 15 versions of each utterance type (question or statement) exceeded the DLs of both NH and CI listeners reported in the literature for speech stimuli (Rogers *et al.*, 2006). All four speakers also used a greater F0 range in the interrogative utterances (measured across the whole utterance). This difference is also reflected in the fact that the difference between the average F0 of the first and second syllables was greater in the interrogative than the declarative versions. The F0 of the first syllables in statements was higher than that of the first syllables in questions, while the F0 of the second syllables in questions was considerably higher than that of the second syllables in statements. This indicates that speakers used a contrast between the F0 of the first and second syllables to mark a rising intonation for interrogative utterances, even lowering the F0 of the first syllable to make the higher F0 in the second syllable more prominent. Although the use of rising intonation in questions has been questioned in a previous report (Geluykens, 1988), the acoustic analyses of materials recorded for this experiment strongly suggest that speakers used a rising intonation to mark interrogative prosody on these single-word materials. Speech rate has also been reported to mark question/statement contrasts on sentence level (Van Heuven and Van Zanten, 2005), but in the single-word materials recorded in the present work, durational differences between questions and statements were below DLs reported for NH listeners (Small and Campbell, 1962), and speech rate therefore does not appear to play a role in marking question/statement differences in these recordings.

The acoustic analyses of the vowel materials provided information on the acoustic characteristics that were necessary to select suitable vowel pairs for the listening experiments. As can be seen in Fig. 4.2, a number of other possible vowel pair selections would have constituted an easier listening task, owing to differences in more than one important acoustic characteristic (e.g. /a/ versus /ɛ, i, y, ɔ/). Also, vowel pairs that may have constituted a more difficult listening task appear (e.g. /y/ versus /i/, or /ə/ versus /œ/), but because of the low frequency of occurrence of /y/ and /œ/ in Afrikaans (Van Heerden, 1999), plus the fact that these vowels are often reduced to /i/ and /ə/ in conversational speech, these vowel pairs were not used. The vowel pairs selected for the listening experiment were representative of specific differences observed within the complete collection of vowels, and were balanced in terms of their difficulty level.

The following section reports on the background, methods, results and discussion of the listening experiments that were conducted using the prosody and vowel materials described in section 4.2.

4.3 LISTENING EXPERIMENTS

4.3.1. Background

The speech materials described in section 4.2 were used to conduct listening experiments on a group of CI recipients and a control group of NH listeners, as described in this section. The aim of this second listening experiment was to compare the perception of prosody and vowels on single-word level in background noise (SWN), in both NH and CI listeners. As discussed in Chapter 2, CI recipients have considerable difficulty with speech perception in noise. The signal received by CI users contains a reduced set of speech cues compared to the cues available to NH listeners, as some of the cues required for redundancy are absent (Xu, Thompson and Pfingst, 2005). Spectral information, for example, is degraded in CIs (Chatterjee and Peng, 2008), with CI users having a limited number of spectral channels available when compared to NH listeners (Friesen, Shannon, Baskent and Wang, 2001). As a result, CI recipients reportedly have difficulty with the recognition of some prosodic cues, especially those features closely related to F0. Voice F0 plays an important role in many important prosodic functions, such as conveying normal intonation patterns, which helps with speech recognition in noise (Laures and Bunton, 2003), marking the differences between questions and statements (Grant and Walden, 1996; Lakshminarayanan, Ben Shalom, Van Wassenhowe, Orbelo, Houde and Poeppel, 2003), conveying the emotion or attitude of a speaker (Breitenstein *et al.*, 2001; Cheang and Pell, 2008; Dmitrieva, Gel'man, Zaitseva and Orlov, 2008; Murray and Arnott, 1993), and marking accented words in a sentence (Breen, Fedorenko, Wagner and Gibson, 2010; Pell, 2001). CI recipients derive less benefit than NH listeners from natural intonation patterns in noise (Meister *et al.*, 2011), and perform significantly worse than NH listeners on question/statement distinctions and sentence accent perception (Meister *et al.*, 2009). Also, CI recipients perform poorly in the recognition of vocal emotions (Hopyan-Misakyan *et al.*, 2009; Luo *et al.*, 2007).

However, it is not only with prosodic cues that CI recipients have difficulty. Vowels, which have been shown to be a particularly important segmental feature in speech recognition (Kewley-Port *et al.,* 2007), also pose a challenge to these listeners. Many CI recipients are unable to attain 100% recognition of vowels even in quiet listening conditions. Munson *et al.* (2003), for instance, reported that better-performing CI users in their study scored 86.6 % (± 5.8%) on vowel recognition while worse-performing listeners scored only 53.7% (± 16%) for vowels that were recognised with 95% accuracy by NH listeners (Hillenbrand *et al.*, 1995). A more recent study reported average vowel recognition accuracy of 45% in CI recipients (Stacey *et al.*, 2010). Introducing background noise makes vowel recognition even harder for these listeners (Xu and Zheng, 2007), who require significantly more favourable SNRs than NH listeners to attain 50% recognition (Goldsworthy, Delhorne, Braida and Reed, 2013).

From the studies mentioned it is clear that generally CI recipients experience difficulty with the recognition of both prosody and vowels. However, most existing reports do not directly compare perception of the two types of speech features. A direct comparison between prosody and vowel perception could provide deeper insight into the difficulty that CI listeners experience with speech perception in noise by showing which speech features are worst affected by noise. Given the reported redundancy and noise robustness of prosodic cues as discussed in Chapter 2 and illustrated by the findings reported in Chapter 3, it is possible that NH listeners use these cues to augment speech perception in noise when segmental information such as vowels is degraded. It is not clear from existing data whether the speech features and cues that are most immune to noise effects for NH listeners also remain useful to CI listeners in noise. Even direct comparisons between vowel and prosody perception in quiet are rare. One study that compared vowel and prosody perception is that of Luo, Fu, Wu and Hsu (2009), who investigated the perception of Mandarin Chinese tones and vowels in CI users using their clinically assigned speech processors. Four vowels were each produced with four different tones (which correspond to changes in voice F0). Listeners responded in a 16-alternative forced-choice paradigm, and results were analysed to determine the number of correctly identified syllables, tones and vowels. Findings indicated that CI users performed better on vowel recognition than tone recognition, but were still able to score above 60% on average on tone recognition in quiet. This finding agrees with the findings of Wei, Cao, and Zeng (2004), who also found an average tone recognition score of above 60% for the CI users in their study. However, it is still unclear whether the F0 cues that are available to CI listeners in quiet remain available in background noise (Brown and Bacon, 2010), and how the perception of other prosodic cues compare to vowel recognition in CI listeners.

Therefore, the listening experiment described below was conducted to explore how well CI recipients perceive prosodic cues in background noise, and how the perception of prosody by CI recipients compares to their perception of important segmental information (specifically vowels) in quiet and in noise. A control group of NH listeners was also included, to provide a baseline against which to compare CI listeners' performance, and to compare the relative noise immunity of the different speech features in the two listener groups. The hypothesis was that perception of prosody would be better than vowel perception in noise in both NH and CI listeners. Although existing literature reports that CI listeners have difficulty with prosody perception related to changes in voice F0, it was hypothesised that durational and intensity cues in combination with available F0 cues would present enough redundancy in prosodic cues to provide an advantage over vowel cues.

4.3.2 Methods

4.3.2.1 Listeners

Ten CI recipients (aged 21-70) participated in the study. All participants used Cochlear devices, and years of implant use ranged from five to 19 years. Nine participants had unilateral CIs. One recipient (S19) used a hearing aid in the non-implanted ear, and one (S15) had bilateral implants. She (S15) was requested to switch off the processor on the ear that she considered weakest, while the hearing aid user was asked to switch the hearing aid off during testing, so that all recipients were evaluated with only one implant. All CI recipients were tested with their processors set to the program and settings that they used most frequently. A control group of listeners matched to the CI group in gender and age also participated in the study. All control subjects had normal hearing (pure tone thresholds ≤ 20 dB HL at octave frequencies from 250 to 8000 Hz). All participants (NH and CI) were native speakers of Afrikaans. Ethics clearance was obtained from the relevant ethics committee at the institution where the research was conducted, and participants provided informed consent prior to testing. Table 4.7 provides information on the CI recipients who participated in the study.

Table 4.7: Details of CI recipients who participated in the listening experiments. Speech recognition scores reflect the percentage of words in pre-recorded sentences that were identified correctly. Speech recognition data for CI7 were not available. C and CA refer to the Contour electrode and Contour Advance electrodes respectively. Details of S15's second implant and processor are not included, as this processor was switched off during testing.

Subject number	Gender	Age	Processor	Implant	Strategy	Post-/Pre-lingual deafness	No of years implanted	Ear(s) implanted	Speech recognition %
S15	F	23	Freedom	CI22M	SPEAK	Post	19	Left	96
S24	F	21	Freedom	CI24RE (CA)	ACE	Post	5	Right	100
S22	M	41	CP810	CI24RE (CA)	ACE	Post	5	Right	100
S23	M	21	ESPrit 3G	CI22M	SPEAK	Pre	15	Right	87
S28	F	58	Freedom	CI24RE (CA)	ACE	Post	5	Right	100
S26	M	22	CP810	CI24R (C)	ACE	Pre	9	Left	96
S27	F	70	Freedom	CI24RE (CA)	ACE	Post	5	Left	-
S14	M	30	CP810	CI24R (C)	ACE	Post	9	Left	92
S5	F	44	Freedom	CI24M	SPEAK	Post	12	Right	92
S19	F	43	CP810	CI24RE (CA)	ACE	Post	6	Right	75

4.3.2.2 Procedures

Listeners were seated in a double-walled sound booth with the test administrator. Speech materials were presented through an M-Audio EX66 Reference Monitor. All test materials were presented in a single-interval 2AFC paradigm, through a GUI developed in Matlab showing the two alternatives on the screen. Participants had to click on a start button, and subsequently had to click on the alternative they heard to prompt the presentation of the next item.

Each listener had to complete five listening tasks for each of the four speakers, in two listening conditions. The five listening tasks included two prosody discrimination tasks (question/statement and certain/hesitant) and three vowel discrimination tasks (/pɔt/ and /pɛt/; /pat/ and /put/; /pɛt/ and /pɛ:t/). The two listening conditions were quiet and an adaptive noise condition, using an SWN specific to each speaker. In quiet, a total of 36 stimuli were presented in each task. The first six items in each task were practice items, purposefully selected to include three items of each of the two alternatives. The remaining 30 stimuli were presented in random order, and performance was scored as the percentage of correct responses. No feedback was given to listeners on the correctness of their responses.

To test recognition in noise, an adaptive procedure was used to prevent floor and ceiling effects that could occur when using a fixed noise condition, especially in the CI population where there is great inter-individual variability. The SNR was changed adaptively via a transformed two-down, one-up staircase procedure, where equilibrium occurs at an SNR corresponding to the 71% correct point on the psychometric function (Levitt, 1971). The 71% correct point is an attractive equilibrium point in a 2AFC paradigm where the chance level is 50%, since it is approximately halfway between guessing and perfect (100%) recognition (Hartmann, 1998). To minimise practice or learning effects, each listener first completed all five listening tasks in both quiet and noise with recordings from an additional female speaker to ensure familiarity with the tasks and procedures. Furthermore, the order of the different tasks and speakers was counterbalanced across listeners. The total testing time was around seven hours per listener. One adaptive procedure took approximately three minutes to complete, and had to be repeated four to six times for each task to ensure accurate determination of recognition thresholds. This resulted in 3 min x average 5 repetitions x 5 tasks (two prosody and three vowel tasks) x 4 speakers = 6 hours, excluding the time required for training and testing recognition in quiet.

The initial step size of the SNR adjustment (until the first reversal) was 2 dB, and the subsequent step size (following the listener's first error) was 1 dB. A pilot experiment was conducted to determine the test procedure that would result in a minimum amount of variance without extending testing time unnecessarily. Ten reversals in the adaptive procedure and four to six repetitions of the procedure resulted in the smallest attainable standard deviation (± 2 dB). The adaptive procedure was therefore terminated after ten reversals, of which the last six reversal points were used to calculate the 71% point. The procedure was repeated a minimum of four times for each task, and if any of the four results differed more than 4 dB (allowing for ± 2 dB deviation from the mean), two additional repetitions were carried out. No feedback was provided on the correctness of individual items, but at the end of each completed test, listeners were informed about their performance rate, which helped to keep listeners motivated.

Speech and noise were combined adaptively to attain the desired SNR, in such a way that the combined stimulus had an intensity of 60 dB SPL, i.e., both the speech and noise levels were adapted after each response so that the desired SNR was obtained, while maintaining the stimulus level at 60 dB SPL. The stimulus level was measured with a sound level meter at the approximate location of the listeners' ears. Speech was always presented above the threshold of a particular listener. Across CI listeners, the SNR varied from -15 dB to 10 dB during the adaptive procedure, so that noise and speech levels varied between 58.6 and 47.6 dB SPL, and 43.6 and 57.6 dB SPL respectively. The 60 dB level was selected, as this is considered to be the average level at which most conversational speech occurs (Firszt, Holden, Skinner, Tobey, Peterson, Gaggl, Runge-Samuelson and Wackym, 2004; Pearsons, Bennett and Fidell, 1977).

4.3.3 Results

Results for the listening experiments are depicted in Fig. 4.4(a) and 4.4(b), which show the average scores for each of the five listening tasks (question/statement, certain/hesitant, pɛt/pɔt, pat/put, and pɛt/pɛ:t discrimination) across all four speakers for NH and CI listeners separately. Data from individual CI listeners are included in Appendix C.

The results depicted in Figure 4.4(a) show that on all tasks in quiet, CI listeners obtained a poorer average score than NH listeners, and a larger variance. The difference between the two listener groups across tasks was analysed using Mann-Whitney's U (owing to the small sample size) and indicated that CI listeners performed significantly worse than NH listeners (U = 4.0, z = -3.48, p < 0.001). Pairwise comparisons between listener groups on the question/statement and certain/hesitant task indicated that CI listeners performed significantly worse than NH listeners on the question/statement task (U = 9.0, z = -3.1, p < 0.001), but not on the certain/hesitant task (U = 31.0, z = -1.44, p = 0.165). The CI listeners' results for the vowel discrimination tasks were compared to NH listeners' results using a one-sample t-test, since all NH listeners scored 100% for all the vowel tasks. Scores did not differ significantly from 100% for any of the three vowel tasks.

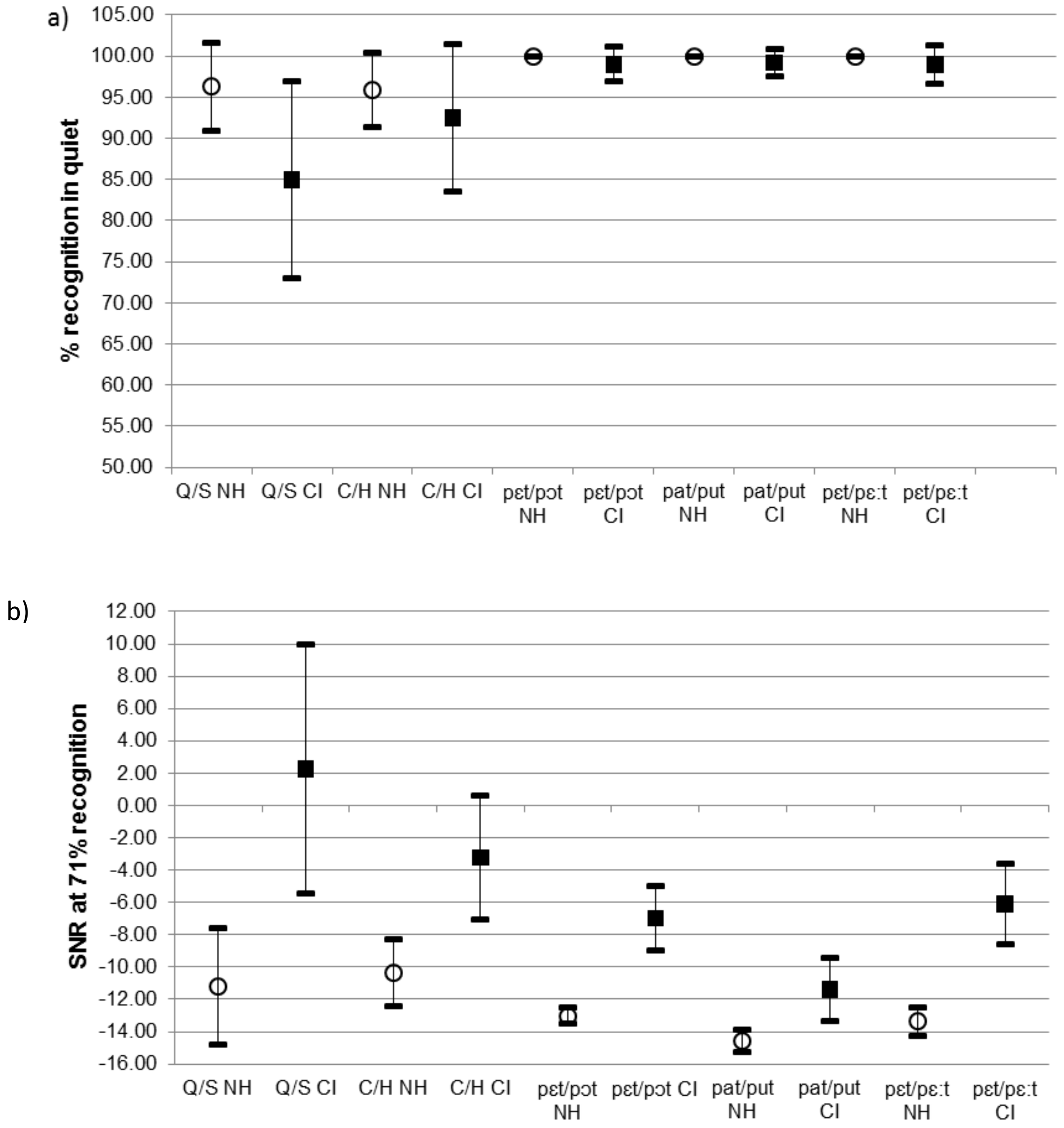

Figure 4.4: (a) Percentage recognition scores obtained in quiet. (b) SNR levels at which 71% recognition was obtained for each task type and listener group. Q/S denotes question/statement discrimination, C/H denotes certain/hesitant discrimination, NH denotes NH listeners and CI denotes CI recipients. Error bars indicate one standard deviation from the mean.

Both listener groups performed best in the vowel recognition tasks. Friedman's analysis of variance (ANOVA) revealed significant differences (p < 0.05) between the five tasks for both groups of listeners, and *post hoc* Wilcoxon signed-rank tests revealed that performance on the two prosodic tasks did not differ significantly in either listener group; neither did performance on the three vowel tasks. In the NH group, the two prosody tasks both differed significantly from each of the three vowel tasks. In the CI group, the question/statement task differed significantly from each of the vowel tasks, while the certain/hesitant task differed only from the pat/put vowel task. A Bonferroni correction was applied to correct for the number of pairwise comparisons (significance reported at a level of $p = 0.005$).

Figure 4.4(b) (listening in noise) shows a similar pattern to that of Figure 4.4a (listening in quiet), with NH listeners performing better than CI listeners on all tasks (as demonstrated by a lower SNR at which 71% correct is achieved) and displaying smaller variance across listeners. A mixed design ANOVA was performed on the data measured in noise, with within-subject factors defined as task (three levels, i.e. question/statement discrimination, certain/hesitant discrimination, and vowel discrimination) and speaker (four levels, namely FS1, FS2, MS1 and MS2). Listener group (NH or CI) was the only between-subject variable. Between-subject effects measured found a significant overall effect of listener group, F (1, 18) = 62.03, p < 0.001. Within-subject measures showed a significant overall effect of task, F (4, 72) = 62.46, $p < 0.001$, as well as significant interaction between task and listener group (NH or CI), F (4, 72) = 23.74, $p < 0.001$. *Post hoc* pairwise comparisons with Bonferroni corrections indicated that across all speakers and listeners, each of the five listening tasks differed significantly from the other four tasks ($p < 0.001$). Friedman’s ANOVAs were conducted on the average results across speakers for the two listener groups separately, and indicated that there were significant differences between the five listening tasks in both groups ($p < 0.001$). Wilcoxon pairwise comparisons in the NH group indicated that none of the vowel tasks differed significantly from each other, and the two prosody tasks also did not differ significantly. The question/statement task differed significantly from the pat/put vowel task, and the certain/hesitant task differed from all three vowel tasks ($p < 0.005$). In the CI group, the two prosody tasks also did not differ significantly from each other; neither did the three vowel tasks. In this group, the question/statement task was significantly more difficult than all three of the vowel tasks, and the certain/hesitant task was significantly more difficult than the pat/put task, but not more difficult than the other two vowel tasks. Differences between CI and NH performance on the different tasks are reflected in the differences between the SNR required for CI listeners to obtain 71%, and those required by NH listeners to achieve the same level of accuracy with each task (averaged across speakers). Table 4.8

documents the average SNR improvement required for CI listeners to enable them to perform at the same level as NH listeners for each of the tasks. The average values show that performance of NH listeners varied by 4.24 dB between the easiest (pat/put) and most difficult (certain/hesitant) task, while CI listeners showed a variation of 13.65 dB between best (pat/put) and worst (question/statement) performance. The question/statement task yielded the biggest difference between NH and CI listeners.

Table 4.8: Differences between SNRs required by each listener group to obtain 71% recognition for each listening task (averaged across speakers).

	Question/	**Certain/**			
	statement	**hesitant**	**pƐt/pƆt**	**pat/put**	**pƐt/pƐ:t**
NH average SNR	-11.21	-10.34	-13.02	-14.58	-13.37
CI average SNR	2.27	-3.21	-7.00	-11.38	-6.09
Difference	**13.48**	**7.12**	**6.03**	**3.20**	**7.28**

The effects of different speakers on discrimination performance in noise are shown in Figures 4.5(a) and 4.5(b) for NH and CI listeners, respectively.

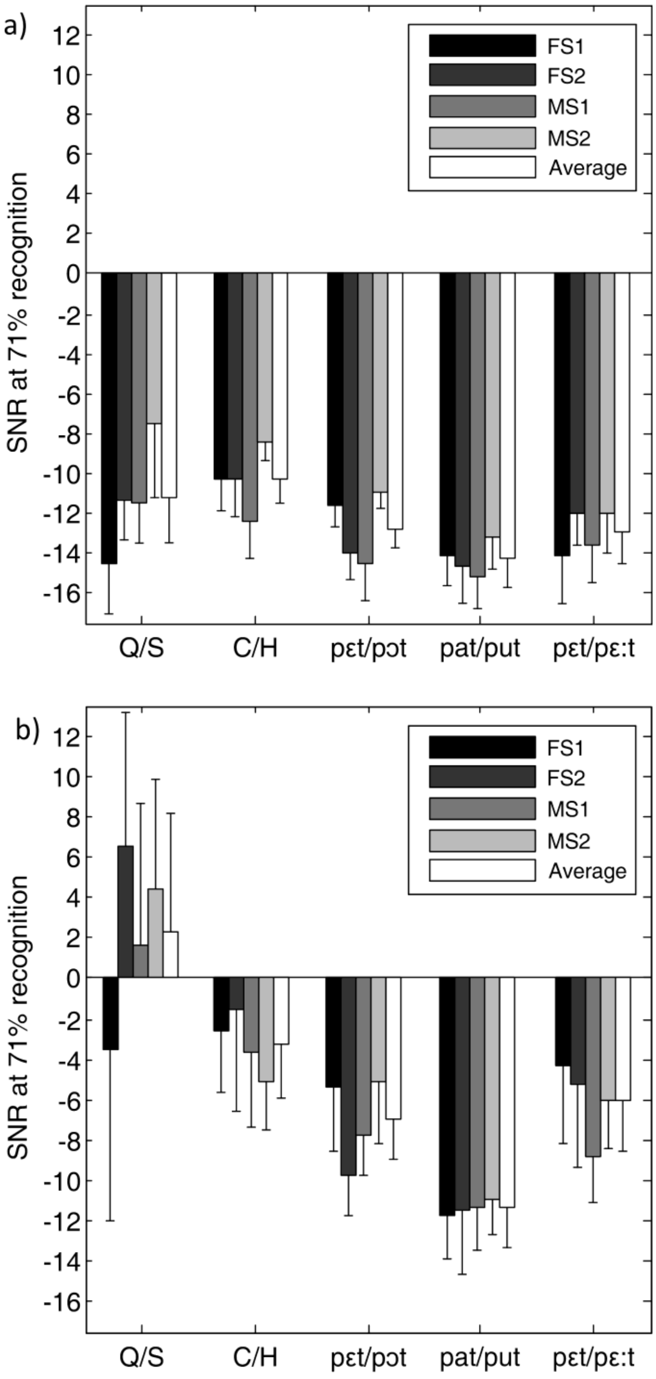

Figure 4.5: (a) Average SNR at 71% recognition attained by NH listeners (n = 10). (b) Average SNR at 71% recognition attained by CI recipients (n = 10). Error bars indicate one standard deviation from the mean. Q/S denotes question/statement recognition and C/H denotes certain/hesitant recognition. Female speakers are FS1 and FS2; male speakers are MS1 and MS2. Significant differences (p <0.05) were found between speakers on all tasks in the NH group, and on all but one of the tasks (pat/put) in the CI group.

The overall effect of speaker was found to be significant using a mixed design ANOVA, F (3, 54) = 15.18, p < 0.001. *Post hoc* pairwise comparisons using Bonferroni corrections showed that across all five tasks and both listener groups, results obtained with speaker FS1 differed significantly from those obtained from MS2, while outcomes from MS1 differed significantly from FS2 and MS2 (all significant at a level of $p < 0.008$). There was significant interaction between speaker and listener group, F (3, 54) = 3.44, $p < 0.05$. This interaction can be seen, for example, in the question/statement discrimination task, where speaker MS2 elicited the poorest performance from NH listeners, while speaker FS2 elicited the poorest performance from CI listeners. Mauchly's test of sphericity indicated that the speaker by task interaction violated the assumption of sphericity, and Greenhouse-Geisser estimates of sphericity were therefore used to correct the degrees of freedom for this interaction ($\varepsilon = 0.46$). The speaker by task interaction was significant at the level of $p < 0.001$, F (5.53, 99.51) = 13.80. Speaker by task interaction was particularly salient for the question/statement task, where speaker FS1 elicited the best recognition performance for both NH and CI listeners. Significant three-way interaction of speaker by task by listener group was found, F (5.53, 99.51) = 4.78, $p < 0.001$. This can be seen, for example, in results found with FS2, who yielded the poorest performance for CI recipients in both prosody tasks, but not in the vowel tasks, while the same speaker yielded a performance close to the average across all speakers on all five tasks from NH listeners. Friedman's ANOVAs were used to compare results from each speaker for each task and listener group separately. Results indicated significant differences between speakers for each of the five tasks in the NH group, and for all but one of the tasks (pat/put discrimination) in the CI group.

Correlations between performance in quiet and performance in noise were analysed and compared between listener groups. Figure 4.6 shows the linear regression lines for the two prosody tasks, with results grouped across speakers but separated for the two listener groups. Vowel tasks were not included because of the ceiling effect and lack of variance in results obtained in quiet. Spearman's rho was used to determine the strength and significance of the correlations, as the data were not normally distributed in all instances.

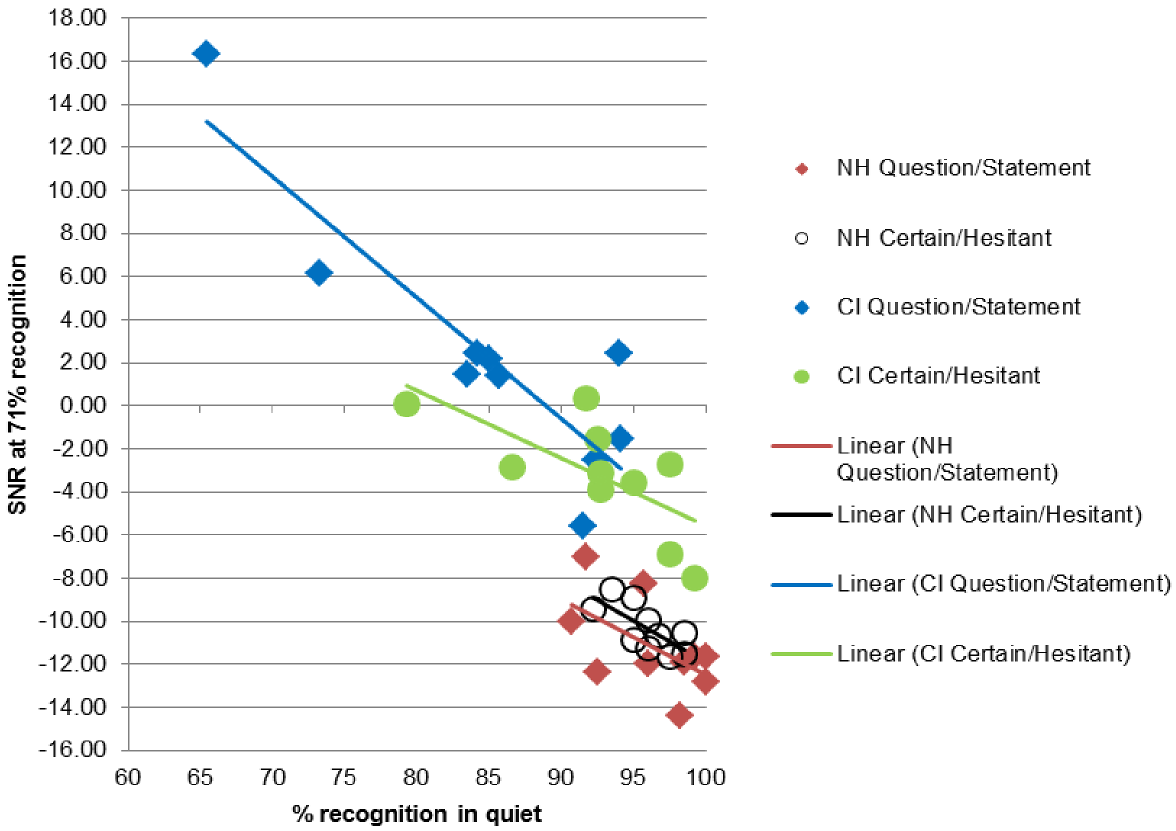

Figure 4.6: Recognition in noise as a function of recognition in quiet. NH indicates NH listeners, CI indicates CI recipients. Results reflect the average scores obtained across four speakers.

In the NH group, performance in noise was significantly related to performance in quiet for the certain/hesitant discrimination task, $r_s = -0.67$, p (one-tailed) < 0.05, but not for the question/statement discrimination task, $r_s = -0.44$, $p = 0.10$. In the CI listener group, question/statement discrimination results in noise were significantly related to results obtained in quiet, $r_s = -0.70$, p (one-tailed) < 0.05, as were certain/hesitant discrimination results, $r_s = -0.74$, p (one-tailed) < 0.01.

4.3.4 Discussion

4.3.4.1 Prosody versus vowel perception in quiet

Results from the quiet listening condition showed that NH listeners performed significantly better on all three vowel tasks than on prosody discrimination tasks. However, in the CI group, while the question/statement task was significantly more difficult than the vowel tasks, the difficulty of the certain/hesitant task did not differ significantly from two of the vowel tasks. These were specifically those of which the primary underlying cues were F2 and duration. This observation suggests that durational differences (underlying the certain/hesitant contrast) available to CI listeners in quiet are more salient than changes in voice F0 and intensity (required for accurate perception of question/statement differences). Results from the question/statement task are comparable to existing reports on CI performance in listening tasks that also involve F0 perception. Results in quiet agree with the report by Luo *et al*. (2009) on tone and vowel perception in Mandarin-speaking CI users (n = 8), which demonstrated that CI users scored better on vowel recognition (90%) than on tone recognition (63%) in quiet. A second similarity with these data was the difference between NH and CI performance that was smaller for vowel recognition (non-significant in the present study) than for tasks that involved tone or intonation recognition (the question/statement discrimination task yielded a significant difference between listener groups in the present work). The difference between NH and CI performance on the certain/hesitant task in the present study was not significant in quiet, supporting the suggestion that the underlying cues of this task were more readily available to CI users than the cues underlying the question/statement distinction.

4.3.4.2 Prosody versus vowel perception in noise

In SWN, NH listeners showed the poorest performance for the certain/hesitant contrast, while CI listeners performed worst on the question/statement task. Although the differences between prosody tasks were not significant in either group, differences between these two tasks and the three vowel tasks showed some interesting effects. In NH listeners, the certain/hesitant task was significantly more difficult than all three vowel tasks, while question/statement distinction was only more difficult than the pat/put task (the easiest vowel task for both groups). In quiet, the question/statement task differed significantly from all three vowel tasks, and it appears therefore that the added noise reduced the performance differences between question/statement and vowel recognition somewhat, resulting in more similar performance levels in noise. On a temporal level, both the F0 contour cues marking question/statement contrasts and the formant cues underlying vowel discrimination are at least partly supported by temporal fine structure cues (Rosen, 1992; Smith, Delgutte and Oxenham, 2002), which would have been severely affected by the SWN (see section 5.6.2.5 for more detail). On a spectral level, vowel identity is supported by formant frequencies (Parikh and Loizou, 2005), which are peaks in the spectral shape at specific frequencies. F0, on the other hand, is represented by harmonics occurring at integer multiples of F0 across the spectrum, and can therefore be perceived even if F0 itself is masked (Oxenham, 2013). The SWN used in the experiment, being a broad-band noise, would have had an adverse effect on spectral cues across the speech spectrum (including harmonics and formants), and would therefore have reduced both F0 and formant frequency perception. It is possible that a narrowband noise, or a noise with amplitude modulations (such as multi-talker babble) could have masked or distorted the spectral peaks required to perceive the formants (Parikh and Loizou, 2005), while some periodicity cues (Rosen, 1992) or harmonics could have remained accessible to listeners in the troughs of the modulated signal, enabling F0 perception (see section 6.5.3).

In the CI group, the difficulty of the two prosodic tasks showed differences to the NH group, with the question/statement task being significantly more difficult than all three vowel tasks, while the certain/hesitant task was only more difficult than the easiest vowel task (pat/put). These results confirm the difficulty that CI listeners had with the question/statement task, as demonstrated by results from the quiet listening condition. However, their perception of a prosodic contrast that was heavily dependent on duration differences did not differ significantly from two of the vowel discrimination tasks – one that depended especially on perception of duration, and another which depended more on F2 perception. This finding suggests that both F2 and durational cues on a vowel level showed similar resistance to noise than the durational cue on a prosody level for these listeners. It seems therefore that CI listeners do have access to some prosodic cues even in background noise, but these cues are not more immune to noise effects than vowel cues. In fact, the prosodic cues required to make the question/statement distinction (F0 and intensity) were severely affected by noise in this listener group.

The outcome that prosody perception tasks yielded poorer performance in noise than vowel tasks was somewhat unexpected, given evidence in the literature that some prosodic cues are quite redundant and immune to noise effects (Dmitrieva *et al.*, 2008; Grant and Walden, 1996; Lakshminarayanan *et al.*, 2003; Smith *et al.*, 1989) and the findings reported in Chapter 3 of the present work. A possible explanation for this finding is that most of the studies demonstrating the redundancy of prosodic cues used longer utterances than the single words used in the current experiment. It is conceivable that prosodic cues are less redundant and noise-resistant on a single-word level than in longer utterances such as phrases or sentences. Among others, the stress rhythm and word boundary cues reported in Smith *et al.* (1989) will not be available in single-word prosody. The second listening experiment of this study therefore contributes to researchers' understanding of the relative robustness of prosody on a single-word level, and suggests that the resilience of prosody that has previously been reported seems to function mainly on the level of longer utterances such as phrases or sentences.

4.3.4.3 Relative performance of CI and NH listeners in noise

Differences between CI users and NH listeners' performances in noise, expressed as the difference in SNR required to obtain 71% recognition (Table 4.8), showed the smallest difference in the pat/put discrimination task (around 3 dB). This task relied heavily on F1 perception, and the small difference in performance between CI and NH listeners suggests that the noise immunity of F1 cues was similar in the two groups. The difference between the two groups was similar for the certain/hesitant prosody task (which was strongly connected to duration perception) and two of the vowel tasks, one relying mostly on F2 perception and the other on duration perception (6-7 dB). The largest difference in performance in noise between NH and CI listeners (13.48 dB) was found in the question/statement task. Acoustic analyses showed that the question/statement contrast was marked by large differences in average F0 of the second syllable and in F0 range for most of the speakers, as well as some intensity differences. F0 perception is a known problem area for CI recipients even in quiet (Rogers *et al.*, 2006). Although NH listeners also performed significantly worse on this task than on vowel discrimination tasks, the large difference between CI and NH performance on this task in noise demonstrates that the addition of background noise can highlight differences between listener groups and different tasks that cannot be detected when testing in quiet, agreeing with the finding of Luo *et al*. (2009) that the introduction of a competing talker resulted in a larger difference between NH and CI performance. Furthermore, the average F0 differences between questions and statements was above DLs reported for CI users in the literature (Rogers *et al*., 2006) for all four speakers, but the findings of the present work suggest that the F0 differences successfully perceived by these listeners in quiet are more vulnerable to the effects of noise than the duration differences that marked the certain/hesitant contrast. This finding is an important first step in answering the question raised by Brown and Bacon (2010) about whether F0 cues perceived by CI listeners in quiet remain available in noise.

4.3.4.4 Relationship between performance in quiet and performance in noise

Although correlations do exist between discrimination abilities in quiet and in noise for CI listeners, these should be interpreted with care. Specifically, individual listeners' performance in noise cannot necessarily be predicted from their performance in quiet. A good example of this is CI listener S28, who attained the second highest score in question/statement discrimination in quiet (94%), but was in the bottom half of performers in this task in noise. This finding indicates the importance of directly evaluating the perception of important speech features in noise, instead of assuming a similar pattern of recognition behaviour in noise as was observed in quiet.

4.3.4.5 Speaker-dependent differences in performance

Different speakers yielded largely different results in both listener groups, especially with the prosody perception tasks (as indicated by larger standard deviations than for the vowel perception tasks). In light of the acoustic analyses, this may be because different speakers used different acoustic cues to indicate the specific prosodic version (e.g. question or statement), and the size of acoustic differences between the two prosodic types varied greatly between speakers. For example, MS2 yielded the poorest performance on the question/statement task for NH listeners, and was also the speaker with the smallest differences in average and range of F0. FS1 yielded the best question/statement discrimination scores in both listener groups, and was also the speaker with the largest differences in F0 average and range between the two versions. This was the only speaker for whom CI recipients were able to obtain 71% discrimination at a negative SNR, suggesting that the large F0 differences produced by this speaker was more immune to the effects of noise than the smaller F0 differences produced by the other speakers, and that this cue remained useful to CI recipients even in strong background noise. As with the results reported by Meister *et al.* (2009), even large F0 differences (as produced by FS1) yielded only a small difference in performance in quiet but could result in a distinct advantage in noise, especially for CI users.

The certain/hesitant discrimination task showed smaller differences between scores obtained with the different speakers in both listener groups. MS1 yielded the best performance on this task from NH listeners and also had the largest duration difference between certain and hesitant utterances (0.35s on average), but MS2 produced duration differences that were very similar (0.31s) and yet elicited the poorest performance from NH listeners. In quiet, however, listeners performed slightly better with recordings from MS2 (99.1 +/- 1.45%) than with those from MS1 (96.1 +/- 3.84%). This suggests that the performance difference in noise was probably not due to duration cues. Regression analyses in which the acoustic cues in the certain/hesitant contrast were considered have shown a degree of cue trading involving duration, frequency and intensity cues (Van Zyl and Hanekom, 2013b). It is possible that differently weighted acoustic cues supported the prosodic contrast in each of these two speakers, and that the cue set supporting perception of the contrast in speaker MS1 was more immune to noise than that of MS2. This underscores the importance of measuring perception of specific cues in noise, since performance in noise cannot necessarily be predicted from performance in quiet.

4.3.4.6 Comments on the experimental design

Interestingly, results for the question/statement distinction obtained using single words in the present listening experiment (NH listeners 96 ± 5.35%, CI listeners 85 ± 11.99%) were similar to the results of Meister *et al.* (2009), who measured question/statement discrimination using sentence materials (NH listeners 99 ± 2.0%, CI listeners 82 ± 10.7%). This supports the use of single-word prosody tasks in the listening experiments reported in this chapter.

For practical reasons, the vowel tasks in this listening experiment did not include all the vowels of the test language; testing each vowel against every other vowel in the 2AFC test paradigm (required to allow a fair comparison between prosody and vowel recognition) would have resulted in 105 distinct vowel discrimination tasks. Rather than attempting this, three vowel pairs were carefully selected to represent specific acoustic differences. This provided insight into the perception of specific acoustic cues by the two listener groups and allowed comparison between duration perception on a vowel level and duration perception on a prosody level. It is conceivable that some other vowel pair selections may result in a different outcome in terms of the noise immunity of specific vowels, but the findings regarding the availability of different cues are expected to remain the same. The listening experiment described in Chapter 5 was designed to test the perception of a more comprehensive collection of vowels in order to explore whether the findings reported here can be generalised to a larger vowel collection.

4.4 CONCLUSIONS

The following can be concluded from the development and acoustic analyses of the prosody materials used in this experiment.

- Prosodic cues can differentiate certain (unreserved) and hesitant permission on the level of a single word.
- The most consistent prosodic cue for distinguishing between certain and hesitant single-word utterances was found to be duration, while cue trading between other cues was observed.
- The cues that communicate a certain/hesitant attitude on a single-word level are different from those that communicate emotion or that differentiate questions from statements.

- Previous work on acoustic characteristics of the word "okay" reported that word-final intonation, intensity, duration, mean F0 and voice quality all serve to differentiate different functions of the word (Gravano *et al.*, 2012), while the acoustic analyses presented in this chapter identified how these cues are applied to communicate the speaker's attitude, and the consistency with which these cues are applied by different speakers.

- Changes in voice F0, specifically a rising intonation, and intensity differences (greater intensity of the final syllable in interrogative utterances) were found to be consistent cues of interrogative (question) prosody in the materials analysed.

The following conclusions can be drawn from the results of the listening experiment.

- Although it was expected that prosodic cues may aid listeners in noise, so that performance on prosody recognition would decline less than performance on vowel recognition tasks in noise, the opposite was found for both NH and CI listeners. This may be because prosodic cues on a single-word level contain less redundancy and are therefore less noise-resistant than those contained in longer utterances.

- The two prosody tasks yielded similar performance in both listener groups in both quiet and noise. However, while NH listeners performed worst on the certain/hesitant distinction in noise, CI listeners performed worst on the question/statement task, suggesting that these listeners may not have received the F0 and intensity cues used by speakers to mark the question/statement distinction, while the certain/hesitant task contained more duration and intensity cues that were better preserved to these listeners in noise.

- The two types of tasks (prosody recognition and vowel recognition) yielded significantly different performance, and these differences were amplified by adding interfering noise. Evaluations of different CI speech processors (Balkany *et al.*, 2007), preprocessing strategies (e.g. Gifford and Revit, 2010) and processor settings (Davidson, Skinner, Holstad, Fears, Richter, Matusofsky, Brenner, Holden, Birath, Kettel and Scollie, 2009) typically include only phoneme and/or sentence recognition tests. Differences between vowel perception and prosody perception found in the listening experiment reported here suggest that such assessments should also include tasks specifically aimed at evaluating prosody perception, especially if the effects of background noise on speech perception need to be determined.

- It should be noted that individuals' performance on a particular task in noise is not necessarily predictable from their performance in quiet.
- The present experiment used an adaptive SNR procedure to measure speech recognition in noise, in order to prevent floor and ceiling effects, especially in the CI population. A limitation of this method is the extended testing time that results from the need of repeated threshold measurements. For this reason, the third listening experiment (reported in Chapter 5) used fixed SNRs to measure both prosody and phoneme recognition.

In conclusion, CI users performed better on vowel recognition than prosody recognition in both quiet and an adaptive noise paradigm, but some prosodic cues remained more useful to these listeners than others in noise. This finding differed from the results obtained using sentence-length utterances (reported in Chapter 3), where prosody recognition was retained better in noise than word recognition. This may be due to differences in the test paradigm; in Chapter 3, word recognition was tested in an open set paradigm, while prosody recognition was measured using a closed set, 2AFC paradigm. It is also possible that the difference in results occurred because the second experiment used single-word utterances while the first experiment used sentences, which may have caused a difference in the amount of redundancy of prosodic cues. To address the possible causes of the differences in results, the listening experiment described in Chapter 5 was designed to measure recognition of prosody and phonemes in sentence-length utterances, but this time using identical test paradigms.

CHAPTER 5 PERCEPTION OF PHONEMES AND EMOTIONAL PROSODY BY CI RECIPIENTS IN NOISE

Parts of this chapter have been submitted to the Journal of Communication Disorders.

5.1 CHAPTER OBJECTIVES

This chapter describes the third listening experiment of the present study. This experiment was conducted to address the fourth and fifth research questions formulated in Chapter 1, namely whether the slope of deterioration of prosody recognition with increasing levels of background noise is shallower than the slope of segmental feature (phoneme) recognition in NH listeners (question 4) and CI users (question 5). This was achieved by testing the perception of emotional prosody in sentences in noise at specific (fixed) SNR levels in NH listeners and in CI listeners, as well as vowel and consonant perception at the same SNRs. Statistical comparisons could subsequently be made between the deterioration slopes of recognition of the two feature types (prosody and segmental), as well as between recognition scores at specific SNRs and between the two listener groups.

5.2 BACKGROUND

The listening experiments described in Chapters 3 and 4 yielded conflicting results. The results from the first listening experiment, described in Chapter 3, suggested that the recognition of prosody at sentence level is more resistant to the effects of background noise than the recognition of individual words in a sentence, when measured in NH listeners. In contrast, results from the second listening experiment (described in Chapter 4) showed that recognition of prosody on single-word level elicited poorer performance than vowel recognition in noise in NH listeners. There are two possibilities for the difference in findings. Firstly, Experiment 1 compared prosody in a 2AFC paradigm to word recognition in an open-set paradigm (thereby possibly advantaging the prosody task), while Experiment 2 compared prosody and vowel recognition in identical 2AFC test paradigms. Secondly, Experiment 1 assessed prosody on a sentence level, while Experiment 2 used single-word materials. Therefore, the differences in results may have occurred either because of the differences in test paradigms of the first experiment, or because the noise immunity of prosody exists only on a sentence level, and was not exhibited by the single-word materials of the second experiment. To determine which of these two possibilities was more likely, a third listening experiment was designed using identical test paradigms and sentence-level speech materials.

5.2.1 Addressing limitations of previous experiments

The following limitations of the methodology of the previous two experiments were addressed in this third experiment. In the first experiment, prosody recognition was evaluated in a 2AFC test paradigm, while word recognition was tested in an open set paradigm. Since the difference in test paradigms may have affected the outcome of the experiment, the second experiment was designed to test both prosody and vowel recognition in a 2AFC paradigm. The value of using a 2AFC paradigm is that many prosodic contrasts in everyday speech occur naturally as a 2AFC phenomenon, e.g. questions versus statements, ambiguous phrases that are clarified by phrase boundaries (a prosodic feature), and certain versus hesitant permission or approval. However, a 2AFC paradigm constitutes a fairly easy listening task and it is possible for listeners to use a "yes/no" strategy – if they are able to recognise only one of the two possible options, anything that does not sound like this option can easily be classified as the alternative, even without clear perception of the stimulus. In light of these limitations of the first two experiments, the third experiment was designed in such a way that prosody and phoneme perception could be tested in identical test paradigms while giving listeners more than two alternatives to choose from.

The main aim of the present study was to compare the relative noise immunity of prosody and segmental speech information in NH and CI listeners. In the first experiment, prosody was compared to word recognition in a sentence context. Although word recognition in such a context relies on the recognition of segmental information (phonemes), it is also supported by other clues, such as syntax, semantics, and even the natural intonation contour (Laures and Bunton, 2003; Meister *et al.*, 2011). It is therefore difficult to deduce from the findings of the first experiment how segmental features specifically are affected by noise. Only vowels were included in the second experiment to represent segmental feature recognition.

This was due to the nature of the experiment, which used an adaptive procedure and required a lengthy testing time (more than six hours per listener), and because vowels have been reported to have a greater impact on speech intelligibility than consonants (Kewley-Port, Burkle and Lee, 2007). To provide a more comprehensive comparison between prosody and segmental cue perception, the third experiment included both vowels and consonants in the evaluation of segmental cue perception.

The first experiment used fixed SNRs, which provides an opportunity to compare the relative slope of deterioration in prosody and segmental cue perception. However, the difference in test paradigms between the two cue types meant that no direct comparison could be made in terms of the rate of deterioration. An adaptive procedure was used when measuring perception in noise in the second experiment. This was done to avoid floor and ceiling effects, especially in light of the relatively easy 2AFC task, and because both NH and CI listeners were tested. The results of the adaptive task, however, did not provide data on the slope of perception deterioration with deteriorating SNRs. The third experiment again used fixed SNRs to enable a comparison of the slopes with which prosody and segmental cue perception deteriorate. This time, however, the two tasks were cast in identical test paradigms (4AFC). The possibility of floor and ceiling effects was reduced by offering four alternatives instead of two, and also by presenting stimuli at different SNRs to the NH listeners than to the CI users.

5.2.2 Types of speech material

To enable testing in a 4AFC paradigm, emotional prosody was selected as a vehicle for testing prosody perception. Emotional prosody was considered suitable for the present work firstly because it has been established that NH listeners are able to differentiate emotions based on prosodic cues alone (Pell, Jaywant, Monetta and Kotz, 2011) and secondly because emotional prosody enabled the use of a test paradigm with more than two possible alternatives (as opposed to a number of other prosodic contrasts, such as question/statement differences). Four emotions were selected for use in the present study, namely anger, happiness, sadness and fear. These four emotions were selected on the basis that the ability of both NH listeners and CI users to distinguish these emotions in quiet has been demonstrated in existing literature (Juslin and Laukka, 2003; Most *et al.*, 2012).

To evaluate segmental (phoneme) perception, vowel and consonant recognition was also evaluated in a 4AFC paradigm to ensure that the two types of features (phonemes and prosody) were tested in identical test paradigms, enabling direct comparison of the noise immunity of the two types of features. To facilitate equivalence of the listening tasks further, both prosody and phoneme perception were tested using sentence materials.

Existing literature indicates that CI recipients have difficulty with both phoneme and prosody perception in quiet. Munson *et al*. (2003) report that better-performing listeners in their study scored 86.6 % (± 5.8%) on vowel recognition for vowels that are recognised with 95% accuracy by NH listeners (Hillenbrand *et al.*, 1995) and 70.4% (± 7.8%) on consonant recognition, as measured with consonants identified with > 90% accuracy by NH listeners. Poorer-performing listeners scored only 53.7% (± 16%) for vowel recognition and 40% (± 12.8%) for consonant recognition. A more recent study reported average vowel recognition accuracy in CI recipients of 45% and consonant recognition scores of 40% (Stacey *et al.*, 2010). Although there are large discrepancies between the phoneme recognition scores reported by these studies, it is clear that CI recipients are unable to attain the same level of accuracy with this task as NH listeners. CI listeners have even greater difficulty with phoneme perception in background noise, requiring significantly more favourable SNRs than NH listeners to attain 50% recognition, especially for consonants (Goldsworthy *et al.*, 2013). Prosody perception also poses a difficult task for CI recipients even in quiet, as indicated by the data summarised in Table 2.4 (see Chapter 2).

To date, however, no published work has directly compared phoneme and prosody perception in noise in either NH or CI listeners, so the relative noise immunity of the two speech feature types remains unknown. The listening experiment described in this chapter endeavoured to address this gap in the research by directly comparing phoneme and prosody perception in identical test paradigms.

5.3 DEVELOPMENT, VALIDATION AND ACOUSTIC ANALYSES OF SPEECH MATERIALS: METHODS AND RESULTS

The phonemes selected for the vowel and consonant recognition tasks were phonemes with a proportional representation of ≥ 1% in the speech sample collected in a study on phoneme occurrence in Afrikaans (Van Heerden, 1999). The selected vowels (n = 10) were /ə, a, i, Ɛ, ɔ, ɑ, e, æ, o, u/ and consonants included (n = 15) were /t, n, s, r, k, l, d, x, f, m, v, p, ɦ, b, j/. Note that the /r/ phoneme in Afrikaans is pronounced as a voiced alveolar trill, which involves a vibration of the tongue tip against the alveolar ridge (as in the Spanish /r/), and the / ɦ/ phoneme is produced in the same place and manner as the English /h/, but is voiced, rather than voiceless. Vowels and consonants were all embedded in meaningful consonant-vowel-consonant (CVC) contexts, with the initial consonant being the target phoneme for consonant testing. Appendix D shows phonetic transcriptions of the target CVC utterances, along with the alternatives offered to the listeners in the GUI. All CVC utterances were contained within the same sentence that did not provide any semantic clues as to the identity of the test word ("he would have said [CVC] now"). Within this sentence in Afrikaans ("Hy sou toe [CVC] gesê het"), an equal number of syllables (n = 3) followed and preceded the target utterance, and the word directly preceding the target word ended on a vowel (/u/), so that target consonants were effectively contained in a vowel-consonant-vowel (VCV) context. The use of sentence materials for vowel and consonant recognition tasks ensured that these abilities were tested at typical speech rates, and in utterances that are more representative of daily communication than the isolated CVC or VCV utterances often used in research or evaluations. It also ensured greater task equivalence with the prosody recognition task, which also used sentence-length utterances of seven syllables each.

To measure the recognition of emotional prosody, a set of jabberwocky sentences (Pannekamp *et al.*, 2005; Silva-Pereyra *et al.*, 2007) was developed. These are sentences in which content words (nouns, verbs, adjectives and adverbs) are replaced with words that consist of phonemes and phoneme combinations that occur in the test language, but that do not have any meaning, while function words (e.g. “the”, “a”, “in”, “is” etc.) are preserved (Silva-Pereyra *et al.*, 2007). Jabberwocky was used to prevent the semantic content of the sentence from biasing listeners towards any particular emotion, while preserving the natural rhythm, phonetic structure and intonation patterns of the test language. Sixteen jabberwocky sentences, each seven syllables long, were created (see Appendix E) and each sentence was recorded with each of the four emotions used in this study (happiness, anger, fear, and sadness).

All speech materials were recorded from two speakers, one male (aged 29 years) and one female (aged 24), both native speakers of the test language who graduated from the same tertiary education institution where the research was conducted. Both had acting experience and training. Recordings of test materials were conducted in a double-walled sound booth, using an M-Audio Fast Track Pro external sound card and a Sennheiser ME62 microphone placed on a microphone stand 30 cm from the speaker’s mouth. Three utterances (repetitions) of all sentence material (jabberwocky for the emotional prosody perception task, and sentences for the phoneme recognition task) were recorded.

Emotional prosody materials were elicited from the speakers using affective story recall (Turnbull, Evans and Owen, 2005), where speakers were asked to recall a specific event in their own lives when they experienced one of the target emotions, and to describe the event to the examiner. The utterances of each of the 16 jabberwocky sentences were recorded, with the speaker aiming to express the target emotion in each of the sentences. After recordings had been completed for one of the target emotions, the speaker was given a break, and a number of neutral utterances (the vowel or consonant materials) were recorded before proceeding to the next emotion.

Recorded waveforms were edited using *Praat* to remove unwanted silences, leaving silences of 100 ms before and after the utterance. For the vowel and consonant materials, the mean intensity (rms level) of each utterance was re-scaled so that all the target CVC words had the same intensity. Intensities of the jabberwocky utterances were left unedited, as intensity is an important cue to emotional prosody perception (Juslin and Laukka, 2003). The male speaker produced the jabberwocky utterances with a range of average sentence intensity of 19 dB, while the female speaker produced sentences with a range of 16 dB in average intensity. For the validation procedure, all three recorded versions of each jabberwocky utterance depicting each emotion were retained. For phoneme recognition, two out of three recorded versions of each phoneme in each CVC context were selected, as a negligible amount of variation in pronunciation was observed between the recorded versions.

All the recorded materials were validated in 12 adults (six female), aged between 18 and 25 years with normal hearing sensitivity (pure tone thresholds ≤ 15 dB HL at 250, 500, 1000, 2000, 4000 and 8000 Hz). Both phoneme and prosody recognition tasks were conducted in the same 4AFC test paradigm followed for the main experiment. A pilot experiment showed that performance on the prosody recognition task tended to improve towards the end of the test, probably as the listener became more familiar with the speaker's manner of expressing the different emotions. For this reason, each listener completed the entire prosody recognition task twice for both speakers. The first presentation served as familiarisation, and data were only collected from the second presentation. No feedback was given either on the correctness of individual items or on their overall performance, and listeners were not trained to listen for specific acoustic cues, to ensure that listener responses were spontaneous and not trained.

Average phoneme recognition across listeners and speakers in the validation procedure was 99.72% (standard deviation = 0.17%). All vowel and consonant utterances were perceived correctly by at least 11 out of 12 listeners (only nine out of 300 utterances were incorrectly identified, and then only by one listener, and only once out of the two recorded versions of the same phoneme). One recorded version of each phoneme in each of its three CVC contexts was therefore retained for the main experiment. Emotional prosody recognition yielded lower accuracy than phoneme recognition; the overall recognition rate for four emotions across two speakers and all versions of all utterances was 87.75% (87.42% for the female speaker, and 88.08% for the male speaker), with a standard deviation of 4.48% across listeners. A recognition accuracy score was calculated for each version of each recorded utterance by calculating the percentage of listeners (out of 12) that correctly identified the emotional prosody of that utterance. Utterances depicting anger yielded the highest score for both speakers (average of 97.02%), followed by sadness (92.45%), happiness (89.33%) and fear (71.72%). For the main experiment, only utterances for which the emotion was correctly identified by ≥ nine out of 12 listeners (75%, or three times chance) were selected. If more than one recorded version of the same jabberwocky sentence depicting the same emotion scored ≥ 75%, only the highest scoring version (or one of the highest scoring versions) was included. Sentences for which any of the emotions did not yield one recorded version scoring above 75% for a particular speaker were excluded from the collection for all four emotions to ensure equal sample sizes of each emotion. This resulted in a final collection of 14 distinct jabberwocky sentences (each depicting all four of the target emotions) for the female speaker, and 13 for the male speaker, with an overall prosody recognition rate of 96.19% (95.29% for the female speaker, and 97.10% for the male speaker).

Acoustic analyses were conducted on the validated materials using *Praat* to obtain data on a number of its basic acoustic characteristics, which could aid interpretation of the listening experiment's results. For the phoneme test materials, which were produced with neutral intonation, speech rate was determined (expressed in syllables per second), along with the average F0 and the range of F0 (maximum F0 minus minimum F0), expressed in Hz. Average intensity was not reported, since this was normalised across these utterances. For the emotional prosody (jabberwocky) utterances, speech rate, average F0, F0 range, and average intensity (in dB SPL) were determined. The results of these analyses are shown in Table 5.1.

Table 5.1: Acoustic characteristics of recorded sentence materials. Average values are reported with standard deviations in brackets.

Emotion	**Intensity (dB SPL)**		**Speech rate (syllables/s)**		**F0 average (Hz)**		**F0 range (max - min) (Hz)**	
	Female	Male	Female	Male	Female	Male	Female	Male
Neutral	70	70	4.69	5.20	201.34	103.14	92.96	52.32
			(0.23)	(0.36)	(6.29)	(8.88)	(15.20)	(21.62)
Fear	67.29	61.99	4.95	5.38	192.96	160.34	98.38	110.65
	(3.36)	(2.12)	(0.39)	(0.28)	(14.32)	(11.28)	(21.22)	(19.30)
Happiness	73.69	64.93	5.24	4.97	201.23	170.18	100.92	156.91
	(1.55)	(2.06)	(0.45)	(0.55)	(14.57)	(28.02)	(31.97)	(51.15)
Sadness	64.53	55.64	4.44	4.01	200.11	118.80	83.49	78.93
	(1.51)	(2.35)	(0.20)	(0.62)	(7.87)	(8.14)	(18.16)	(17.55)
Anger	70.60	66.81	5.11	4.08	190.94	147.18	106.72	153.01
	(1.84)	(1.33)	(0.54)	(0.64)	(20.46)	(13.13)	(36.67)	(42.27)

The table shows that fear and sadness tended to have lower intensities, while happiness and anger were produced with higher intensities by both speakers. Both speakers produced sadness at the slowest speech rate, but while the male speaker produced fear at the highest speech rate, the female speaker produced happiness at the highest rate. The female speaker did not show large differences between average F0 for the different emotions (the highest average was 201.23 Hz for happiness, and the lowest was 190.94 Hz for anger, a difference of around 10 Hz), while the male speaker produced a low average F0 (118.80 Hz) for sadness, and high F0 averages for fear (160.34 Hz) and happiness (170.18 Hz). For the four emotions, both speakers used the smallest F0 range for sadness (although the male speaker produced an even smaller range for neutral utterances) and the largest F0 ranges for happiness and anger.

5.4 PILOT LISTENING EXPERIMENT

To determine suitable SNRs for testing both NH and CI listeners, a pilot experiment was conducted with three NH listeners and one CI listener. The first NH listener was presented with a sample of the jabberwocky sentences from both speakers and vowels and consonants recorded from the female speaker at 16 different SNR levels, ranging from -16 to 14 dB, with 2 dB intervals. Results were examined to look for floor and ceiling effects. Because of the 4AFC test paradigm, chance performance (floor performance) is at 25%. The results are shown in Figure 5.1 below. Results are shown only up to 0 dB SNR, as a plateau was reached at higher SNRs.

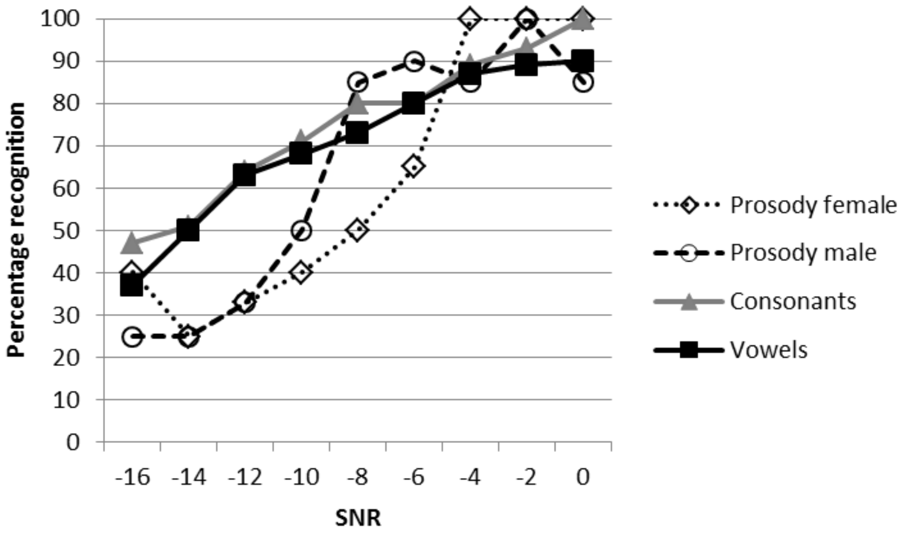

Figure 5.1: Results of first NH listener in pilot experiment

During testing of the first NH listener, it was noted that the wide range of intensities at which stimuli were presented was slightly distracting to the listener. This wide range occurred because of both the range of intensities of the recorded materials and the range of SNRs (a range of 30 dB), since the intensity of the combined speech and noise varied with both of these variables. Therefore, the second NH listener was tested with an altered set of stimuli. To test this listener, each jabberwocky utterance was combined with noise at all the SNRs that would be tested (SNRs of 2, -2, -6, -10, and -14 dB), and then rescaled back to the original intensity of the utterance. This way, the intensity of the utterance was preserved as a cue to the emotion expressed, but the range of output intensities was reduced. However, this listener reported using the intensity of the background noise as a cue to the emotion, and selecting "anger" whenever she heard that the noise was loud. Therefore, the intensities of the combined speech and noise stimuli were not rescaled in the main experiment. The distraction experienced by the first listener was expected to be reduced in subsequent testing owing to the use of a smaller range of SNRs. Testing with the second listener also revealed a slight practice effect, as the listener had not been familiarised with the materials beforehand. A third listener was therefore tested, to ensure that testing at the selected SNRs could be conducted without problems following adequate familiarisation, and to determine the duration of the test when assessing the listener with four different SNRs (as would be done in the main listening experiment).

Subsequently, one CI user was assessed with the vowels, consonants and jabberwocky recorded from both speakers and presented in quiet and at one SNR (0 dB SNR). The results of this test are shown in Table 5.2 below.

Table 5.2: Results of CI user in pilot study in quiet and at 0 dB SNR

	Vowels	**Consonants**	**Prosody**
		Quiet	
Female speaker	80	80	43
Male speaker	73	76	54
		0 dB SNR	
Female speaker	43	56	40
Male speaker	67	36	38

The results obtained from this CI user showed that recognition of some of the features (the male speaker's consonants and prosody, especially) was close to chance performance (25%) and 0 dB SNR therefore constituted a relatively difficult listening condition for this listener. However, since the aim of the main experiment was to investigate the effects of noise on the recognition of particular speech features, it was important for the listening task in the main listening experiment to be difficult enough to demonstrate deterioration between the highest and lowest SNR clearly. Allowing for the possibility that some CI users might perform better than the listener in the pilot experiment, 0 dB SNR was selected as the second lowest SNR in the main experiment. Although the range of SNRs for the NH listeners ranged across 15 dB, the range for CI users was 3 dB smaller in the experimental design. This range was selected because the intensities of the combined speech and noise materials varied with SNR, and CI users have a dynamic range that is much smaller than that of NH listeners (Loizou, Dorman, Poroy and Spahr, 2000).

5.5 MAIN LISTENING EXPERIMENT

5.5.1 Methods

5.5.1.1 Participants

Fourteen listeners participated in the main listening experiment. Seven listeners were CI recipients (four female), with ages ranging from 24 to 71. Details of these listeners are reported in Table 5.3. Three of the CI users had bilateral implants (S15, S22 and S5); two of them had less than 12 months' experience with the second implant. These three listeners were asked to remove the processor from the ear with the more recent implant and to complete the listening tasks using only one implant to ensure equivalence across the group. Details of these second implants are excluded from Table 5.3, since the processors were switched off during testing. A control group of NH listeners (n = 7), each matching the gender and age (within three years) of one of the CI recipients also participated in this experiment.

Table 5.3: Details of CI recipients participating in main listening experiment. Speech recognition scores reflect word recognition measured in meaningful sentences.

Subject number	Gender	Age	Processor	Implant	Strategy	Post-/Pre-lingual deafness	No of years implanted	Ear(s) implanted	Speech recognition %
S15	F	24	Freedom	CI22M	SPEAK	Post	20	Left	96
S22	M	42	CP810	CI24RE (CA)	ACE	Post	6	Right	100
S28	F	60	Freedom	CI24RE (CA)	ACE	Post	6	Right	100
S26	M	24	CP810	CI24R (C)	ACE	Pre	10	Left	96
S27	F	71	Freedom	CI24RE (CA)	ACE	Post	6	Left	100
S14	M	33	CP810	CI24R (C)	ACE	Post	11	Left	92
S5	F	45	Freedom	CI24M	SPEAK	Post	13	Right	92

5.5.1.2 Procedure

To ensure equivalence between the recognition tasks for segmental and prosodic features, both were tested in a closed-set, single-interval, 4AFC paradigm. Tests were conducted using computer-based GUIs. Phoneme perception was assessed with the use of 10 distinct vowels and 15 distinct consonants. In each listening condition (quiet and each of the four SNRs), each phoneme was presented three times, every time in a different CVC context, with three other phonemes offered as alternatives in the GUI (see Appendix D). Listeners were presented with an utterance containing the target word and had to select which word they thought they heard from the four options offered in the GUI. In the vowel recognition task, the target word and the three alternatives all had the same initial and final consonants, with only the vowel differing. The vowel alternatives were selected so that each time a target vowel's recognition was tested, three different alternatives were offered. As a result, each of the nine other vowels in the total collection was offered as an alternative in one of the three trials for each target vowel. For the consonant task, all four options had the same vowel and final consonant, with only the initial consonant differing between options. The consonants presented as alternatives were selected on the basis of their distinctive features (voicing, place of articulation, and manner of articulation). In each of the three trials of each consonant, the other three alternatives were selected so that one of them differed only in terms of one distinctive feature, a second differed in terms of two features, and the third differed in terms of three distinctive features. Appendix F provides details on the distinctive features of the consonants used. All the words used in the forced-choice tests were meaningful words to prevent any bias that listeners might have towards meaningful words from influencing results.

Emotional prosody was assessed using 14 distinct jabberwocky sentences presented by the female speaker, and 13 sentences presented by the female speaker. In each of the five listening conditions (quiet and the four different SNRs), each sentence was presented four times, each time depicting a different emotion (anger, fear, happiness, or sadness). Recorded utterances were presented one at a time in random order, and listeners had to select from a 4AFC GUI which of the four emotions they perceived following the presentation of each utterance.

Testing was conducted in a double-walled sound booth with speech materials presented through an M-Audio EX66 Reference Monitor. Each participant listened to all the test materials in quiet and in SWN (specific to each speaker). Four different SNRs were used in the noise condition, based on pilot experiments. For NH listeners, SNRs of 0 dB (with speech and noise at equal intensities), -5 dB, -10 dB, and -15 dB were used. For CI recipients, SNR levels of 8 dB, 4 dB, 0 dB and -4 dB were used. The intensities of all the phoneme recognition materials were re-scaled to the same average intensity in *Praat* after having combined the signal and noise. During testing, all phoneme recognition materials were presented at an average intensity of 65 dB SPL as measured at the level of the listener's ear. To obtain the desired SNR for the emotional prosody materials, noise levels were calculated for each utterance individually according to its intensity, since the intensities of these utterances were not normalised so as to preserve the intensity cue to the emotion expressed. Also, no adjustments were made to the overall intensity of the prosody materials after combining the signal with SWN at the desired SNR. To prevent listeners from using the intensity of the combined signal and noise as a cue to emotion (instead of listening to the speech material only), utterances with different SNRs were all presented in a mixed block in random order. As the individual utterances' intensities varied, the resultant speech-and-noise combinations also varied according to both the intensity of the utterance and the SNR of the test item, with test items presented at a poorer SNR having a higher overall intensity. This resulted in a great variation in intensity levels of the individual test items in the emotional prosody task, which made it unlikely that listeners would rely on the intensity of the noise to determine the emotion expressed. Presentation levels for the prosody materials varied from 50 to 75 dB SPL (average intensity of the playback of the utterance), depending on the emotion expressed by the speaker. Prior to testing the CI listeners, free-field thresholds were measured for each listener using SWN to ensure that the utterances with the lowest intensity were at least 5 dB above their audibility threshold.

Prior to testing, listeners were familiarised with the test materials. Samples of each of the recorded phonemes from each speaker were played to the listeners, in the same sentence context that would be used during the phoneme test. Before starting the emotional prosody task, the examiner explained to the listener what jabberwocky was, and read the jabberwocky sentences to them using neutral intonation. Subsequently, all the validated jabberwocky sentences recorded from each speaker were played to the listener, one emotion at a time (i.e. all the utterances depicting fear, followed by all the utterances depicting happiness and so on). This was done to compensate for the fact that the speakers were not known to the listeners, which proved to introduce learning effects, as discovered during pilot testing and the validation procedure. The order of the different listening tasks (phonemes and emotional prosody) and recordings from the two speakers were counterbalanced across listeners to reduce any possible practice effects.

5.5.2 Results

The results for the quiet listening condition are depicted in Figure 5.2, with results for the two speakers combined. Results from the two listener groups were compared using the non-parametric Mann-Whitney's U (owing to the small sample size). CI users performed significantly worse than NH listeners (using one-tailed exact significance) on vowel recognition (U = 48.50, z = 3.196, p < 0.001, r = 0.85), consonant recognition (U = 44.00, z = 2.55, p < 0.01, r = 0.68) and prosody recognition (U = 49.00, z = 3.13, p < 0.001, r = 0.84). In view of the poor performance of CI users on the prosody recognition task, a one-sample t-test was used to determine if their performance was significantly above chance (25%). It was found that CI users' prosody recognition was significantly above chance (p < 0.001). A non-parametric (Kruskal-Wallis) ANOVA was performed to compare the performance on the different tasks within each listener group. For the NH listeners, the ANOVA indicated significant differences between tasks (H(2) = 12.37, p < 0.05), and Mann-Whitney tests were used to conduct *post hoc* pairwise comparisons. Effects are reported at a 0.0167 level of significance to correct for the number of comparisons (Bonferroni correction). Pairwise comparisons revealed that while vowel and consonant recognition did not differ significantly from each other (U = 1.86, z = 0.61, p = 0.545), prosody recognition was significantly poorer than both vowel recognition (U = 10.14, z = 3.30, p < 0.001, r = 0.88) and consonant recognition (U = 8.29, z = 2.70, p < 0.01, r = 0.72). The same pattern of findings was observed in the CI listener group, with the ANOVA indicating a significant difference between the three tasks (H(2) = 10.87, p < 0.05). Pairwise comparisons indicated that in this group, as with the NH group, vowel and consonant recognition did not differ significantly (*U* = -1.00, *z* = -0.30, *p* = 0.763), while prosody recognition was significantly poorer than vowel recognition (*U* = 8.93, *z* = 2.69, *p* < 0.01, *r* = 0.72) and consonant recognition (*U* = 9.93, *z* = 3.00, *p* < 0.01, *r* = 0.80). Since prosody recognition in the validation procedure was only slightly poorer than phoneme recognition (96.19% prosody recognition, compared to 99.72% phoneme recognition), the significant difference in the main listening experiment was unexpected. However, since the comparison between phoneme and prosody recognition in the noise condition would not be based on absolute values of recognition scores alone, but also on the slope of deterioration, this difference was not considered a problem.

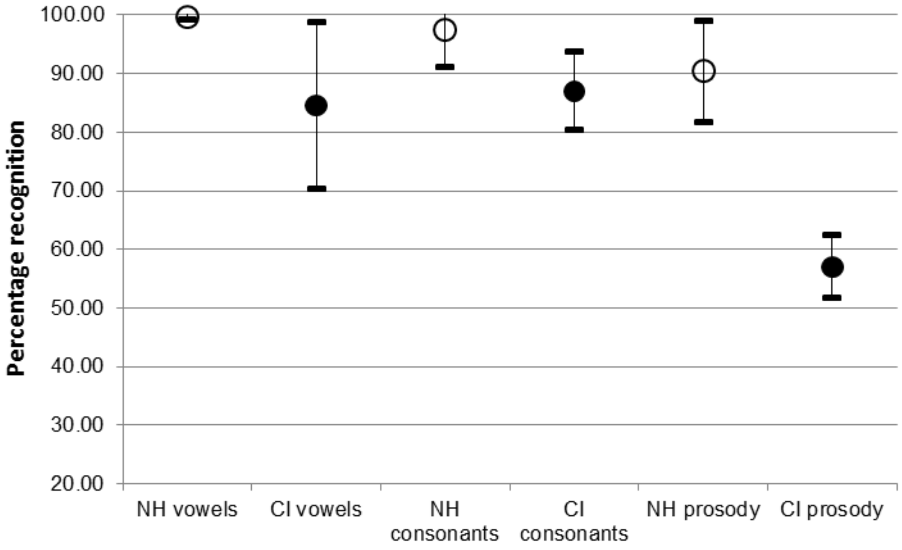

Figure 5.2: Average percentage correct recognition in quiet across listeners in each group. Error bars indicate one standard deviation from the mean. Open circles = NH listeners; filled circles = CI recipients.

Results obtained in quiet from each of the different emotions separately are depicted in Figures 5.3 (NH listeners) and 5.4 (CI users).

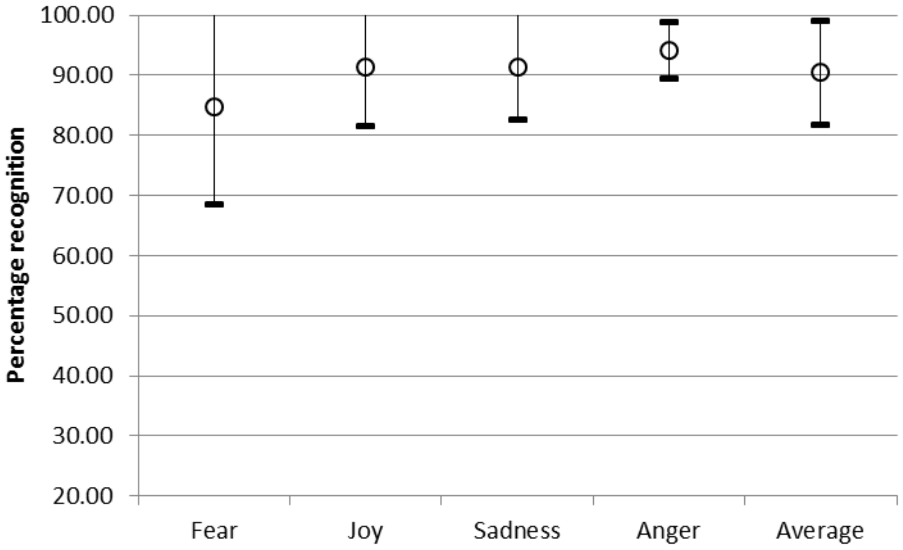

Figure 5.3: NH listeners' recognition of the different emotions in quiet, with results of the two speakers averaged together

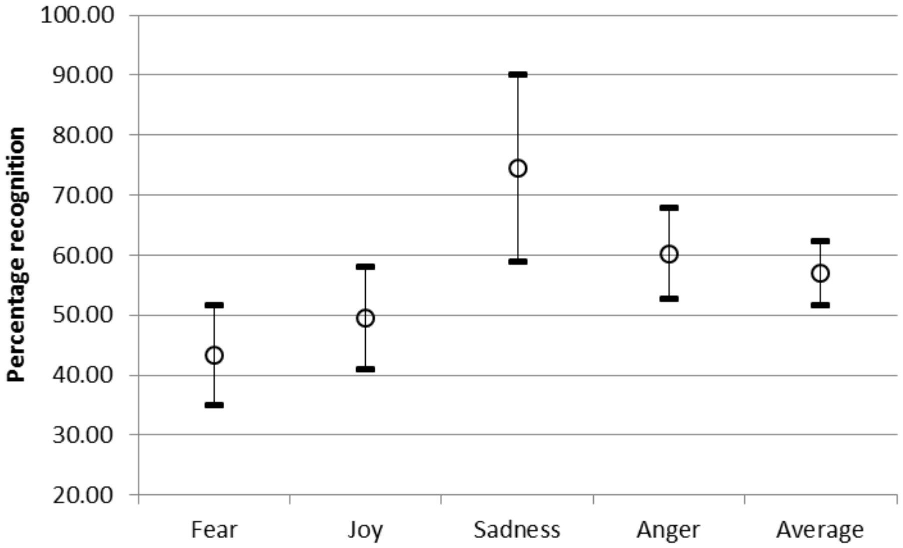

Figure 5.4: CI users' recognition of the different emotions in quiet, with results from the two speakers averaged together

The results shown in Figures 5.3 and 5.4 were analysed using a Friedman's ANOVA to explore the differences in scores between the different emotions in each listener group. In the NH listener group, the ANOVA indicated that the recognition of the different emotions did not differ significantly ($\chi^2(3) = 6.09$, $p = 0.11$). In the CI recipient group, there was a significant difference between the emotions ($\chi^2(3) = 11.78$, $p < 0.01$). This finding was further explored using *post hoc* Wilcoxon pairwise comparisons, which revealed that the recognition of fear and sadness differed significantly in this listener group (z = -3.00, p < 0.008, which is equivalent to a 0.05 level of significance when correcting for the number of comparisons).

A confusion matrix depicting CI listeners' confusions between the different emotions in quiet is shown in Table 5.4 below (NH listeners' recognition scores in quiet were too high to warrant a confusion matrix). These results show that happiness and anger were mutually confused. Fear showed the poorest recognition rate and was most frequently confused with sadness in the case of the female speaker, and happiness in the case of the male speaker. Sadness in turn yielded the best recognition rate and was most frequently confused with fear and never with anger.

Table 5.4: Confusion matrix showing results from CI users (n = 7) in quiet listening condition. Values represent percentages, with values in bold indicating correct recognition, while values in italics indicate the emotion with which the target emotion was most frequently confused.

	Fear	**Happiness**	**Sadness**	**Anger**
		Female speaker		
Fear	**36**	27	*29*	8
Happiness	11	**41**	3	*45*
Sadness	*15*	6	**79**	0
Anger	16	*28*	3	**53**
		Male speaker		
Fear	**49**	*29*	15	7
Happiness	15	**59**	3	*23*
Sadness	*26*	0	**74**	0
Anger	3	*27*	3	**67**
		Both speakers		
Fear	**42**	*28*	23	7
Happiness	13	**49**	3	*35*
Sadness	*20*	3	**77**	0
Anger	10	*27*	3	**60**

The results from the noisy listening condition are shown in Figures 5.5 and 5.6 (results for NH and CI listeners respectively). To determine whether the introduction of background noise had a greater effect on CI users compared to NH listeners for the different tasks, differences between recognition in quiet and recognition at 0 dB SNR (the only SNR level tested in both groups) were calculated for each listener and each task, and statistically compared between groups using Mann-Whitney's U for independent samples. Results indicated that the difference between vowel recognition in quiet and at 0 dB SNR was significantly greater in CI listeners (U = 44.50, z = 2.57, p < 0.01, r = 0.69), and the same was found for consonant recognition (U = 49.00, z = 3.14, p < 0.001, r = 0.84). However, the difference between prosody recognition in quiet and at 0 dB SNR did not differ significantly between listener groups (U = 26.00, z = 0.192, p = 0.902).

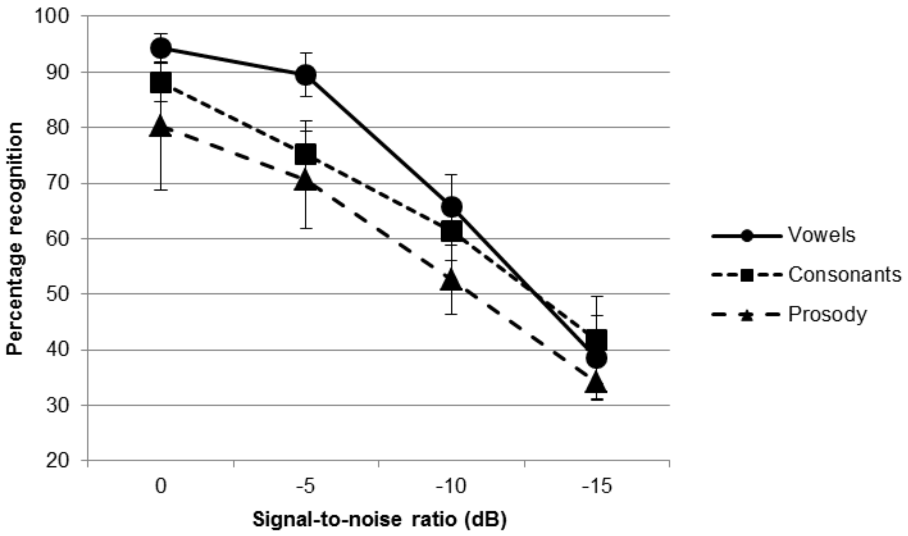

Figure 5.5: Average percentage recognition of vowels, consonants and emotional prosody at different SNRs in NH listeners. Error bars indicate one standard deviation from the mean.

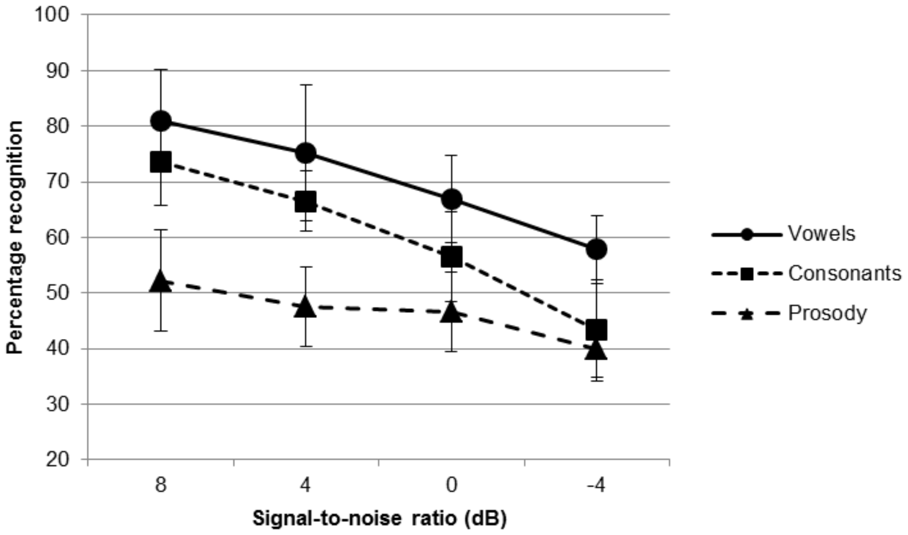

Figure 5.6: Average percentage recognition of vowels, consonants and emotional prosody at different SNRs in CI recipients. Error bars indicate one standard deviation from the mean.

Using the results obtained at the different SNRs, the slope of recognition deterioration with decreasing SNR (expressed as percentage per dB) was determined for each listener and each task separately, using a least squares estimate to fit a linear slope to the data points. The distribution of the average slope could then be compared between listener groups and listening tasks. Table 5.5 shows the average deterioration of recognition with standard deviations for each task and listener group.

Table 5.5: Average percentage deterioration in recognition per dB of decreasing SNR for each of the three listening tasks. Standard deviations are shown in brackets.

	NH listeners	**CI recipients**
Consonants	-3.06 (0.61)	-2.52 (0.62)
Vowels	-3.82 (0.43)	-1.94 (0.79)
Prosody	-3.12 (0.69)	-0.95 (0.58)

Statistical comparisons revealed that the slope of consonant recognition deterioration did not differ significantly between NH and CI listeners (U = 14.00, z = -1.34, p = 0.21), while vowel recognition deteriorated with a significantly steeper slope in NH listeners than CI listeners (U = 0, z = 0, p < 0.001), as did prosody recognition (U = 0, z = 0, p < 0.001). ANOVA comparisons of the three listening tasks within each listener group showed significant differences between the three tasks for both NH listeners ($H(2)$ = 6.91, p < 0.05) and CI listeners ($H(2)$ = 10.92, p < 0.01). *Post hoc* pairwise comparisons (measured at a 0.0167 level of significance to correct for the number of comparisons) showed that in NH listeners, deterioration of prosody recognition did not differ significantly from either vowel recognition (U = -7.00, z = -2.11, p = 0.04) or consonant recognition (U = 1.00, z = 0.302, p = 0.76), but vowel and consonant recognition deteriorated with significantly different slopes, with vowels showing a steeper deterioration slope (U = -8.00, z = -2.31, p < 0.0167, r = -0.62). In the CI recipient group, the deterioration slope of vowels and consonants did not differ significantly (U = 3.43, z = 1.04, p = 0.30), and neither did the deterioration slope of vowels and prosody (U = -7.29, z = -2.20, p = 0.03). Consonant recognition deteriorated with a significantly steeper slope than prosody recognition (U = -10.71, z = -3.24, p < 0.001, r = -0.87).

In light of the noticeably flat slope (small overall deterioration) of the prosody recognition in CI listeners (with a difference of only 12.34% between recognition at the best and worst SNRs), recognition scores at the best and worst SNR for this task in this group were compared using a Wilcoxon signed rank test, which showed that recognition at -4 dB SNR was significantly poorer than recognition at 8 dB SNR (z = -2.37, p < 0.05, r = -0.63). In addition, a one-sample t-test was conducted and results showed that the slope was significantly non-zero. To determine whether the differences found between prosody and phoneme recognition in quiet persisted under adverse listening conditions, recognition scores at the lowest SNRs were compared between tasks for each listener group using Friedman's non-parametric ANOVA for related samples and *post hoc* Wilcoxon pairwise comparisons. Results showed that in NH listeners at the lowest SNR (-15 dB), recognition scores on the three listening tasks (vowels, consonants and prosody) did not differ significantly (χ^2 (2) = 4.57, p = 0.102), while in CI recipients at -4 dB SNR there was a significant difference between the three tasks (χ^2 (2) = 11.14, p < 0.01). *Post hoc* pairwise comparisons in this listener group showed that prosody and consonant recognition did not differ significantly at this SNR (z = 0.80, p = 0.423), but vowel recognition was significantly better than both consonant recognition (z = 2.41, p < 0.0167) and prosody recognition (z = 3.21, p < 0.001).

Deterioration of the recognition of each of the different emotions separately is depicted in Figures 5.7 (NH listeners) and 5.8 (CI users).

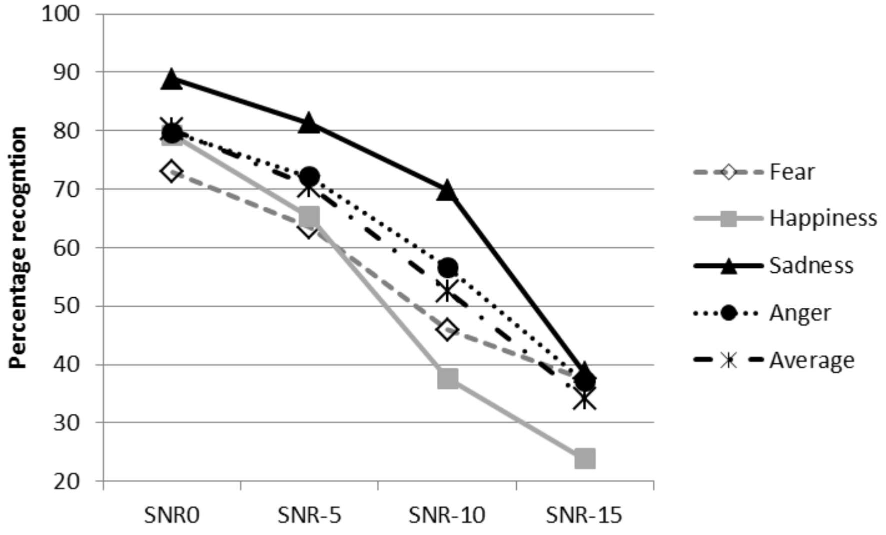

Figure 5.7: Deterioration of emotion recognition with decreasing SNR for each of the different emotions as measured in NH listeners

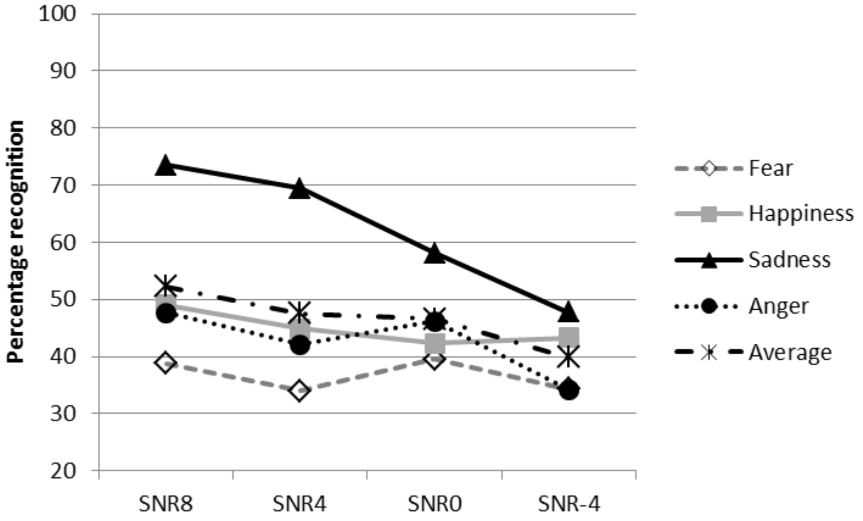

Figure 5.8: Deterioration of emotion recognition with decreasing SNR for each of the different emotions as measured in CI recipients

The slope with which the recognition of each emotion deteriorated with increasing noise was determined for each listener by fitting a linear slope using a least squares estimate. An average slope could then be determined for each emotion in each listener group, and recognition deterioration of the different emotions could be statistically compared using Friedman's ANOVA and *post hoc* Wilcoxon pairwise comparisons. The average slopes of each emotion's deterioration (with standard deviations) are shown in Table 5.6.

Table 5.6: Average percentage deterioration in recognition per dB of decreasing SNR for each of the four emotions. Standard deviations are shown in brackets.

	NH listeners	CI recipients
Fear	-2.48 (1.40)	-0.19 (1.09)
Happiness	-3.88 (1.01)	-0.50 (1.04)
Sadness	-3.25 (0.93)	-2.21 (1.61)
Anger	-2.86 (1.46)	-0.90 (1.79)

Results from the Friedman's ANOVA indicated that there were significant differences within the collection of slopes obtained from the NH listeners ($\chi^2(3) = 8.66$, $p = 0.03$) and *post hoc* pairwise comparisons revealed that the deterioration slope of happiness recognition was significantly steeper than that of fear ($z = 2.69$, $p < 0.007$). The slopes obtained from CI users did not show any significant differences ($\chi^2(3) = 5.57$, $p = 0.13$).

Confusion matrices for the different emotions at the lowest SNR tested in each group are shown below in Tables 5.7 (NH listeners) and 5.8 (CI users).

Table 5.7: Confusion matrix showing results from NH listeners (n = 7) at the lowest SNR measured (-15 dB SNR). Values represent percentages, with values in bold indicating correct recognition, while values in italics indicate the emotion with which the target emotion was most frequently confused.

	Fear	Happiness	Sadness	Anger
		Female speaker		
Fear	**39**	13	22	*26*
Happiness	27	**22**	12	*39*
Sadness	*36*	9	**40**	15
Anger	*28*	27	10	**36**
		Male speaker		
Fear	**36**	22	10	*32*
Happiness	22	**25**	10	*43*
Sadness	*38*	16	**37**	8
Anger	22	*33*	7	**38**
		Both speakers		
Fear	**38**	17	16	*29*
Happiness	24	**24**	11	*41*
Sadness	*37*	13	**39**	12
Anger	25	*30*	8	**37**

Table 5.8: Confusion matrix showing results from CI users (n = 7) at the lowest SNR measured (SNR-4). Values represent percentages, with values in bold indicating correct recognition, while values in italics indicate the emotion with which the target emotion was most frequently confused.

	Fear	Happiness	Sadness	Anger
		Female speaker		
Fear	**36**	*33*	18	13
Happiness	21	**46**	6	*27*
Sadness	*31*	17	**44**	8
Anger	30	*40*	6	**24**
		Male speaker		
Fear	**33**	*33*	14	20
Happiness	18	**41**	5	*36*
Sadness	*36*	11	**52**	1
Anger	15	*35*	5	**44**
		Both speakers		
Fear	**34**	*33*	16	16
Happiness	20	**43**	6	*31*
Sadness	*33*	14	**48**	5
Anger	23	*38*	6	**34**

The confusion matrix of the CI users shows that in noise, as in quiet, sadness was the easiest emotion to recognise (averaged across speakers), although the difference in recognition scores between sadness and the other emotions was much smaller in noise. Another similarity with the data measured in quiet was that anger and happiness were mutually confused. However, in noise, anger yielded the poorest performance in the case of the female speaker, while fear was the most difficult to recognise in the male speaker's case. NH listeners also showed mutual confusion between anger and happiness, although in the case of the female speaker anger was even more frequently mistaken for fear. These listeners performed most poorly with the recognition of happiness in both speakers' cases, with the recognition rates of the other emotions very close together. In both listener groups, sadness was most often mistaken for fear. However, while NH listeners most frequently confused fear with anger, CI users most frequently mistook fear for happiness. The results depicted in the confusion matrices are discussed in light of the acoustic analyses in section 5.6.2.

For the consonant recognition task, the alternatives offered in the GUI were selected according to the number of distinctive features on which they differed from the stimulus (see section 5.5.1.2 and Appendices D and F). Because of this method, the confusion patterns of listeners were examined to see if there was a bias towards particular consonants in the recognition task. Figure 5.9 shows the distribution of responses across the different options offered in the GUI. The data indicate that listeners selected the option with the smallest amount of difference (in terms of distinctive features) from the stimulus more frequently than options with a greater degree of difference, but this effect was not particularly strong.

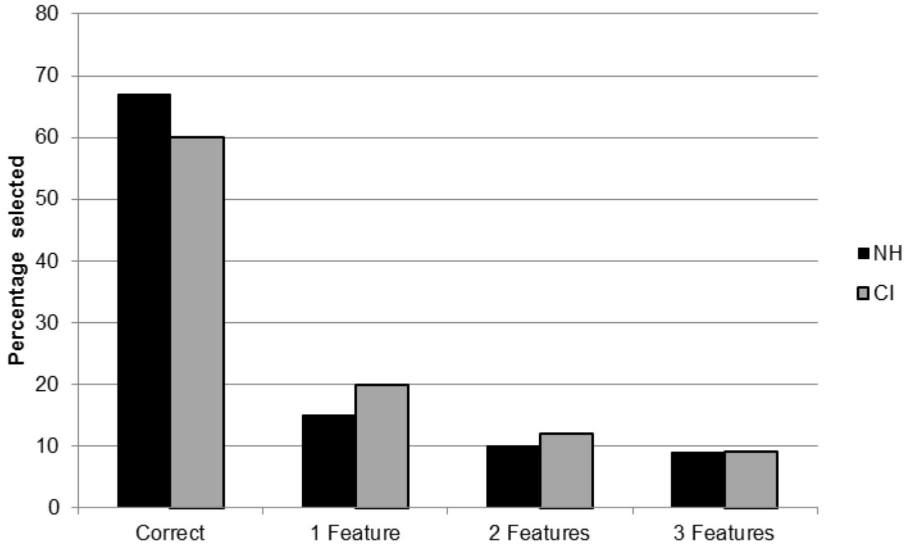

Figure 5.9: Frequencies of selections of different consonant options offered in the GUI. "1 Feature" indicates the option that differed from the correct option on the basis of only one distinctive feature (manner, place, or voice); 2 Features indicate an option that differed with two distinctive features, etc.

The error patterns of listeners on consonant recognition were further analysed according to the individual phonemes. Figure 5.10 below illustrates recognition scores for the individual consonants. Consonants on the x-axis are arranged according to manner of articulation, with plosives to the left of the axis, followed by fricatives, semi-vowels and nasals.

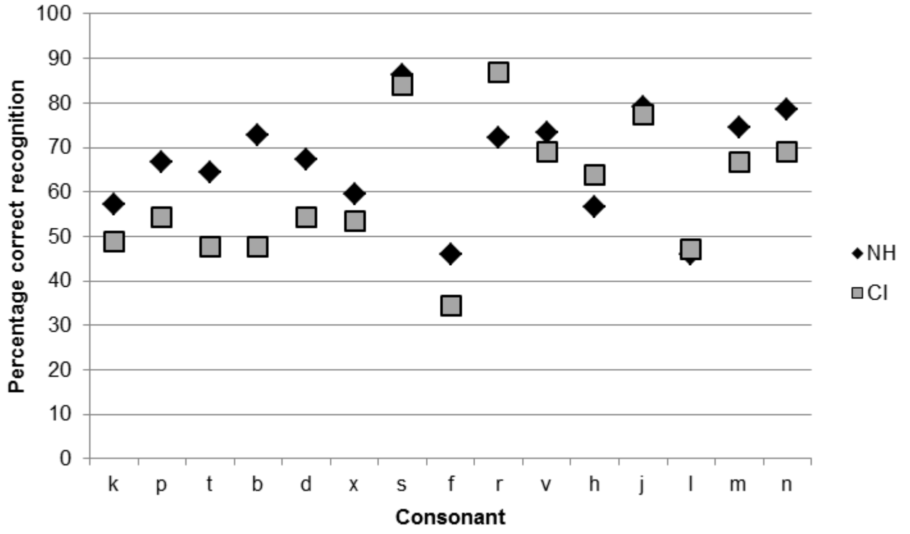

Figure 5.10: Percentage correct recognition in noise of each consonant (averaged across speakers and SNRs). SNRs for NH listeners were 0, -5, -10 and -15 dB; SNRs for CI listeners were 8, 4, 0 and -4 dB.

The data in Figure 5.10 suggest that the difference in NH and CI performance was especially noticeable for plosives (/k,p,t,b,d/).

5.6 DISCUSSION

5.6.1 Acoustic analyses

Acoustic analyses of the emotional prosody materials showed that utterances portraying sadness were characterised by a low intensity, slow speech rate and small F0 range in both speakers, which is in agreement with reports in the literature from other languages (Banse and Scherer, 1996; Juslin and Laukka, 2003). A low average F0 is also commonly found in sad utterances, according to these reports. In the present work, the male speaker used a low F0 to convey sadness, but the female speaker used an average F0 similar to the other emotions (this speaker did not produce obvious differences in F0 average between different emotions). Anger is reported to be associated with an increase in average F0, F0 range, intensity, and speech rate (Banse & Scherer, 1996). Angry utterances recorded from both speakers in this study showed a high intensity, similar to that of happy utterances. The female speaker produced both happy and angry utterances at a high speech rate, while the male speaker produced anger at a slow rate. F0 average values for anger were not particularly high for either speaker. The male speaker produced anger and happiness with similarly large F0 ranges. The female speaker, as mentioned before, did not produce obviously different F0 ranges for the different emotions, but nevertheless used the largest F0 range for anger. Happiness or joy reportedly presents with similar acoustic characteristics as anger (high intensity, speech rate, F0 average and large F0 range) (Banse and Scherer, 1996; Juslin and Laukka, 2003). In the present study, anger and happiness showed some similarities – high intensity and a relatively large F0 range. The female speaker produced happiness at the highest speech rate of all the emotions, followed by anger, while the male speaker used the highest rate for fear, followed by happiness. The male speaker also produced happy utterances at the highest average F0 of all the emotions, while the female speaker used an average F0 similar to neutrality and sadness to depict happiness.

According to Banse and Scherer (1996), fearful utterances are frequently associated with high arousal levels, resulting in increases in intensity, speech rate, F0 average and F0 range. In the present work, however, fear was produced with a low to average intensity by both speakers, and an F0 range similar to that of neutral utterances. The female speaker also produced fear at a low to average speech rate and average F0. The male speaker, however, used a high average F0 to produce fear, as well as a high speech rate. During the listening experiments, listeners remarked that fear and sadness tended to sound the same (fear was found to be the most difficult of all the emotions to recognise), while happiness and anger sounded similar. In summary, all of the acoustic characteristics that were analysed in this study showed some differences between the different emotions that were expressed. The following section discusses the data from the listening experiment and explores how specific acoustic cues may have been related to specific confusion patterns.

5.6.2 Listening experiment

5.6.2.1 Quiet listening condition: NH and CI performance

CI recipients performed significantly worse than NH listeners on all three tasks (vowels, consonants and prosody) in quiet, and both listener groups performed significantly worse on prosody than on phoneme recognition. Despite the fact that CI listeners were able to perceive prosody at an accuracy level (57%) significantly above chance (25%), it was clear that they had difficulty with this task. Although NH listeners also performed significantly better in phoneme recognition than prosody recognition, the difference between the performance of CI recipients and NH listeners on prosody recognition (33%) was much larger than the difference in phoneme recognition performance (14%). This suggests that CI users did not have the same degree of access to the cues supporting prosody perception as to those supporting phoneme perception.

The small number of CI users that participated in this experiment did not allow for reliable statistical conclusions regarding the effects of non-auditory factors (such as age, duration of deafness and implant experience) on the recognition of emotional experience. Nonetheless, the following table was compiled with data on a number of non-auditory features of the CI participants, along with their performance on emotional prosody in quiet to see if any apparent trends emerged.

Table 5.9: Non-auditory features and prosody recognition in quiet (CI users)

Listener	Pre-/postlingual deafness	Age	CI use	Speech recognition (% in quiet)	Prosody recognition (% in quiet)
S15	Post	24	20	96	63
S22	Post	42	6	100	62
S5	Post	45	13	92	59
S26	Pre	24	10	96	57
S14	Post	33	11	92	56
S27	Post	71	6	100	54
S28	Post	60	6	100	48

The data in the table were arranged according to the percentage of emotional prosody recognition obtained in quiet (from best to worst). Two possible trends seem to emerge from the data. Firstly, with the exception of S22, it appears that longer implant experience resulted in better prosody recognition. In fact, S15, who attained the best prosody recognition score, only had four years of normal hearing before contracting meningitis and receiving her implant. This observation seems counterintuitive, as one would have expected that listeners who have had longer exposure to prosodic cues by means of normal hearing would have been better at recognising these cues. The second possible trend is the age of the listeners. Notably, S27 and S28 who were the two oldest participants showed the poorest prosody recognition scores, despite both having had late onset hearing losses and therefore many years of exposure to prosody through normal hearing. Existing literature reports that a listener's age affects the way in which they process auditory stimuli, e.g. Vongpaisal and Pichora-Fuller (2007) showed that F0 difference limens deteriorate with age, even in NH listeners.

5.6.2.2 Performance in noise: NH versus CI listeners

Phoneme perception

Results from measurements at 0 dB SNR indicated that CI recipients' recognition of vowels and consonants was significantly more affected by noise than NH listeners' perception of these features at the same level, which corresponded to existing reports on sentence and phoneme recognition in noise (Goldsworthy *et al.*, 2013; Qazi *et al.*, 2013). CI users in the present study required approximately 10 dB better SNR conditions to achieve the same levels of phoneme recognition as NH listeners. Goldsworthy *et al*. (2013) reported a smaller SNR difference between CI users and NH listeners for vowel recognition (6.3 dB), and a larger difference for consonant recognition (14.7 dB). The difference between the present researcher's observations and those of Goldsworthy *et al*. (2013) may be due to methodological differences (different test paradigms, different number of speakers, feedback given during testing, or test language difference).

Although CI users required an 11 dB improvement in SNR to obtain the same level of consonant recognition as NH listeners, the slope of consonant recognition deterioration did not differ significantly between NH and CI listeners. This indicates that although CI listeners performed more poorly on consonant recognition in noise than NH listeners, increasing levels of noise had a similar effect on consonant recognition in the two groups.

Prosody recognition

In contrast to phoneme perception, the difference between prosody recognition in quiet and at 0 dB SNR did not differ significantly between listener groups. This may be in part because the recognition of prosody by CI recipients was already relatively poor in quiet, but since their prosody recognition remained well above chance even at -4 dB SNR (40%), this observation is not ascribed to a floor effect. The shallow slope of prosody recognition deterioration in noise in this group showed that this ability was relatively unaffected by increasing noise levels, despite their relatively poor recognition of prosody in quiet. This result suggests that the acoustic cues that CI listeners relied on to attain the limited degree of prosody recognition they achieved in quiet were more immune to the effects of background noise when compared to the cues that NH listeners relied on for prosody recognition.

5.6.2.3 Noise immunity of prosody and phonemes in NH listeners

Comparisons between the deterioration slopes of prosody and phonemes in noise in NH listeners showed that prosody recognition did not deteriorate with a significantly different slope than phoneme recognition. This finding was in contrast to suggestions of prosody's noise immunity in existing literature (as discussed in Chapter 2, section 2.4) and to the findings of Chapter 3. The similarity in deterioration slopes also seems counterintuitive – in everyday experience it appears to be easier to recognise a speaker's emotion than to discern individual phonemes in difficult listening conditions. This experience, however, was not supported by the auditory-only data measured in this study, and the discrepancy suggests that listeners might use other cues (such as visual or contextual cues) to support emotion recognition in real-life situations (Paulmann and Pell, 2011).

Vowel recognition deteriorated significantly faster than consonant recognition in NH listeners, particularly at SNR levels poorer than -5 dB. Fig. 5.5 shows that vowel recognition was better than consonant recognition at 0 dB and – 5 dB SNR. This may be because vowels are relatively high in intensity when compared to consonants (Orr, Montgomery, Healy and Dubno, 2010). However, at lower SNRs (-10 and -15 dB) the advantage of vowels over consonants was eliminated by the noise.

5.6.2.4 Noise immunity of prosody and phonemes in CI listeners

In the CI listener group deterioration of prosody and vowel recognition did not differ significantly. Despite the similar deterioration slopes of vowel and prosody recognition, however, vowel recognition was still significantly better than prosody recognition at the lowest SNR, because of the poor prosody recognition by CI users even at the best SNR. This observation agrees with the findings reported in Chapter 4, which showed that CI users' vowel recognition was significantly better than prosody recognition in noise, although the present experiment's methods allowed comparisons between vowel and prosody recognition at specific SNRs, which the method followed in Chapter 4's experiment did not. Consonant recognition, on the other hand, deteriorated faster than prosody perception. Consequently, although prosody recognition was significantly poorer than consonant recognition in quiet in this population, this difference was eliminated in noise. Consonant and vowel recognition deterioration did not differ significantly in CI users, but consonant recognition remained poorer than vowel recognition across SNRs.

In both listener groups, vowel recognition was significantly better than prosody recognition in quiet, and deteriorated with a similar slope. At the higher SNRs tested (0 dB, -4 dB and -5 dB), vowel recognition was better than prosody and consonant recognition in both listener groups. It appears therefore that the cues underlying vowel recognition are relatively resistant to the effects of noise in low noise levels for both listener groups. If a hierarchy of noise immunity were compiled based on the present findings, vowels could be considered most noise resistant in both groups, followed by prosody, with consonants showing the greatest vulnerability to background noise. These differences are especially noticeable at low noise levels (from 8 dB SNR to -5 dB SNR).

Some caution is warranted in the interpretation of the slopes. The type of speech material used in a speech recognition test could influence the slope, depending on the redundancy thereof. E.g. some studies measuring sentence recognition in noise report a deterioration slope as steep as 17%/dB (Van Wieringen and Wouters, 2008), while studies measuring word recognition in noise generally report more gradual slopes, around 3-4%/dB (Beattie, Barr and Roup, 1997; Shi and Zaki, 2014). However, this might depend on the test methods and materials, as other studies using sentence materials (specifically, a number of versions of the Hearing In Noise Test in different languages) report slopes of around 10%/dB (Soli and Wong, 2008), and the Words-in-Noise test which uses monosyllabic words yield slopes of around 8-9%/dB (Wilson, Carnell and Cleghorn, 2007). Nonetheless, a gradual slope could mean that listeners had a lower performance plateau (Shi and Zaki, 2014), and might not necessarily imply a high degree of noise immunity of the test materials. Notably, the gradual slope of prosody recognition in the CI users in the present work may have been due in part to the low performance plateau in this task. However, the slopes of vowel recognition and prosody recognition did not differ significantly in this group, despite vowel recognition having a significantly higher score than prosody recognition in quiet and at 8 dB SNR.

5.6.2.5 Acoustic cues to prosody perception: effects of noise and CI processing

The acoustic analyses of the emotional prosody materials showed that the distinctions between the four expressed emotions are marked by differences in F0, speech rate and intensity characteristics. Although a speech-weighted noise specific to each speaker was used, it is possible that the different underlying cues were differentially affected by the masking noise. Appendix G shows the temporal envelope waveforms as well as the spectrograms of a sample jabberwocky sentence depicting happiness (recorded from the female speaker), in quiet and at SNRs of 0 dB and -4 dB. In the following sections, the possible effects of noise on the different underlying cues are discussed with reference to the figures in Appendix G. In addition, the perception of the underlying cues by CI users is discussed in light of the present data and existing literature, since CI users might not have had equal access to the different cues even in quiet.

F0 cues

Voice pitch, which roughly correlates with voice F0, has long been known to be an important cue to emotional prosody (Lakshminarayanan *et al.*, 2003; Williams and Stevens, 1972), and the acoustic analyses of the emotional prosody materials used in the present study showed that both F0 average and range were used by the two speakers to convey different emotions. Temporal envelope or amplitude modulation information in an acoustic signal can convey pitch information to some degree (Burns and Viemeister, 1976), but temporal fine structure cues evoke a stronger percept of pitch (Smith *et al.*, 2002; Stickney *et al.*, 2007). The masking created by speech-weighted noise used in the present work likely eliminated most of the fine structure cues in the speech signals. However, adding noise to a speech signal can also affect temporal envelope cues by filling the troughs (Drullman, 1995). Figures G.2 and G.3 (Appendix G) illustrate how added speech-weighted noise masked the temporal fine structure of the sample sentence. The estimations of F0 and formant frequencies were affected by the noise, probably due to the lack of fine structure cues. Although the noise also affected temporal envelope cues by filling in the troughs, the shape of the temporal envelope and especially the peaks of the signal remain visible on the waveform.

The difficulty that CI users have with F0 perception owing to a lack of temporal fine structure cues, among other factors (as discussed in Chapter 2), are likely to have limited their ability to make use of F0 cues to perceive emotional prosody in the present work. Note that none of the participants in the present study used processing strategies that were designed to convey fine structure cues. Data from existing studies on prosody perception by CI recipients indicate that these listeners have particular difficulty with prosodic cues that are closely related to changes in F0, such as question/statement contrasts (Meister *et al.*, 2009; Peng *et al.*, 2008) and tonal contrasts (Luo, Fu, Wu and Hsu, 2009). It is therefore reasonable to assume that CI users in the present work had limited access to F0 as a cue to emotional prosody, and might have relied to a greater extent on other cues such as intensity and speech rate (discussed below) as cues to differentiate between emotions.

Duration and speech rate cues

Slow fluctuations (between 2 Hz and 50 Hz) in the overall amplitude of the speech signal can be described as envelope information and is related to acoustic features such as duration and intensity (Rosen, 1992). The amplitude onsets and offsets represented in the temporal envelope can also convey information about speech rate and rhythm (Rosen, 1992). As mentioned under the discussion of F0 cues, added noise can affect envelope cues by filling in the troughs (Drullman, 1995), but the amplitude peaks in the speech envelope that were not masked by the noise (see Figures G.2 and G.3 in Appendix G), might have provided some information about speech rate and/or word durations to listeners.

The temporal envelope of speech is explicitly encoded in CI speech processors, and studies have shown that CI users have access to this cue (Van Wieringen and Wouters, 1999). In contrast to F0 perception, CI listeners' ability to make use of certain cues in the time domain (measured with a task such as gap detection) is reported to be close to that of NH listeners (Moore and Glasberg, 1988; Sagi, Kaiser, Meyer and Svirsky, 2009). Speech rate and other durational cues might therefore have been more accessible to CI users than F0 cues. If CI users depended primarily on speech rate as a cue to emotion, a high degree of confusion between sadness and anger as produced by the male speaker could be expected, because of the similarity in speech rate of these two emotions (sadness = 4.01 syllables/second with a standard deviation of 0.62; anger = 4.08 syllables/second with a standard deviation of 0.64). However, CI users never mistook sadness for anger in quiet, and anger was only mistaken for sadness 3% of the time. In noise, there was also a very low degree of confusion between sadness and anger in CI listeners. It appears therefore that although speech rate may have been an important cue that was used in conjunction with other cues, CI users were not solely relying on speech rate to judge the expressed emotion. It could be said that speech rate contributed to the redundancy of the prosodic cues that could aid perception in noise.

Intensity cues

In the present work, intensities of the recorded emotional prosody materials were not normalised in order to preserve intensity as a cue to the emotion expressed. The limitation of this method is that the intensity of the added SWN also varied between utterances, and may have served as a cue to the emotion of the speaker. However, since the intensity of the combined speech and noise materials also varied with the specific SNR at which it was presented, and different SNRs were combined in a random order during testing, it is unlikely that listeners were able to rely on this cue consistently. The alternative – normalising the intensities of all the utterances and therefore ensuring equal intensity of the interfering noise at all SNRs – would have eliminated intensity as a cue to emotional prosody. Given the evidence that intensity is an important cue to emotional prosody, especially for CI users (Juslin and Laukka, 2003; Luo *et al.*, 2007), this method was not used, as it would not have accurately represented all the cues to emotional prosody that occurs in everyday life, and therefore would have eliminated the possibility of a fair comparison between phoneme and prosody perception.

Data from the confusion matrix for NH listeners at -15 dB SNR indicated that at this level, utterances expressing happiness were most frequently labelled as angry. The results of the acoustic analyses indicated that happiness and anger had similar intensity levels. At low SNR levels, the combined speech-and-noise stimuli had higher intensities than at high SNR levels. The confusion of happiness with anger at -15 dB SNR suggests that listeners were inclined to label the high-intensity stimuli as angry at this SNR level, and might indicate that intensity cues remained available to the listeners at low SNRs where other cues (such as F0 cues) may have been completely masked by noise.

The intensity resolution of CI users is reported to be poorer than that of NH listeners, and their dynamic range is much smaller (Loizou *et al.*, 2000), but while DLs for F0 in CI users are reported to be poorer than those of NH listeners by almost one order of magnitude, intensity DLs are only poorer by a factor of 2.4 dB (Rogers *et al.*, 2006). In the present study, happiness and anger had similar average intensities in both speakers' productions, and there was a high degree of confusion between these two emotions by CI users in both quiet and noise, as well as by NH listeners in high levels of noise. It should be added, however, that these two emotions also showed similar sized F0 ranges in both speakers' productions. Listening to the male speaker's productions in quiet, CI users confused sadness only with fear, which was closest in average intensity. This confusion pattern also occurred in noise in both listener groups. The female speaker also produced fear and sadness with similar intensities, and these two emotions were mutually confused in quiet by CI users. In noise, however, fear was more frequently mistaken for happiness, despite a 6.4 dB difference in average intensity between these emotions. These observations suggest that although CI listeners did not rely exclusively on intensity cues, intensity played an important role in their perception of emotional prosody. This is in line with the findings of Luo *et al*. (2007) who reported that amplitude normalisation significantly reduced emotional prosody recognition in CI users, indicating these listeners' dependence on intensity cues.

Complementary or redundant cues

The underlying acoustic cues in the speech materials of the present work occurred concurrently, as they would in natural speech, and were not isolated or manipulated to determine the effects of the individual cues on prosody recognition. Listeners therefore had access to multiple acoustic cues supporting the same speech features.

In quiet and in NH listeners, some of these features may have been redundant, i.e. one of the available cues may have been adequate to recognise the prosodic pattern, but in noise and in CI users, listeners may have had to rely on different cues that were better preserved. Peng, Chatterjee and Lu (2012) investigated NH and CI listeners' use of F0, intensity and duration (speech rate) cues for question/statement discrimination. They reported that with full-spectrum stimuli, NH listeners relied primarily on F0 cues, but when the stimuli were spectrally degraded, these listeners were less sensitive to F0 contour and relied more on intensity and duration cues. CI users in that study were specifically sensitive to differences in peak intensity ratio (an intensity cue), and used durational cues in a variable way. Another recent study (Morris, Magnusson, Faulkner, Jönsson and Juul, 2013) reported that the intensity difference limens of CI users was a stronger predictor of their performance in a number of prosody discrimination tasks (vowel length, word stress, and word boundary discrimination) than their duration or F0 difference limens. In conclusion, both intensity and duration appear to be fairly noise resistant and therefore relied upon to a greater degree than F0 cues in noise by NH listeners, while CI listeners appear to rely mostly on intensity cues.

5.7 CONCLUSIONS

The following can be concluded from the development and acoustic analyses of the prosody materials used in this experiment.

- The jabberwocky materials recorded for this experiment can be successfully used to assess the perception of emotional prosody in the absence of linguistic cues, as indicated by its perception by NH listeners in quiet. However, the utterances portraying fear yielded relatively low scores in NH listeners and should be used with caution.

- The acoustic features of the emotional prosody materials showed some similarities to those reported in existing literature, although the characteristics of utterances portraying fear were somewhat different to those reported in the literature.

- The speech materials developed for the assessment of phoneme perception in this experiment constitute a useful means by which to assess phoneme recognition in a sentence context, without semantic clues as to the identity of the phoneme. These phonemes were recognised with 100% accuracy by NH listeners in quiet.

The following conclusions can be drawn from the results of the listening experiment.

- In quiet, CI recipients were able to attain average phoneme recognition of around 85%, while prosody recognition was around 55%. Given that perception of emotional prosody is an important skill that is central to human social interactions (Paulmann, Pell and Kotz, 2008), this finding highlights the importance of continued efforts to improve the delivery of acoustic cues underlying prosody perception to CI users. Observations from the present experiment suggest that improved access to F0 should serve to improve perception of emotional prosody in these listeners.

- NH listeners also performed significantly worse on prosody than phoneme recognition in quiet, but the difference between the two tasks was much smaller (around 3%).

- Contrary to expectations, the deterioration slope of prosody recognition in NH listeners did not differ significantly from that of either vowel or consonant recognition. This finding suggests that the relative success that NH listeners have with speech recognition in noise is not due to the noise immunity of prosody in particular. In light of the findings of both this experiment and the one described in Chapter 4, vowels might be playing a more important role in ensuring noise immunity of speech as perceived by NH listeners.

- In CI listeners both vowel and prosody perception deteriorated more slowly with increasing noise levels than consonant recognition.

- It was also demonstrated that the phoneme perception of CI listeners is more susceptible to the effects of noise than that of NH listeners, which underscores the difficulty that these listeners have in noise.

- In CI recipients the limited amount of prosody perception they were able to attain in quiet remained largely intact with increasing noise. However, owing to the poor prosody perception of these listeners even in quiet, this ability remained significantly poorer than vowel recognition at the lowest SNR. Efforts to improve the delivery of prosodic cues to CI recipients should therefore continue.

CHAPTER 6 GENERAL DISCUSSION AND CONCLUSION

6.1 RESEARCH OVERVIEW

The objective of this study was to investigate the relative noise immunity of prosody and segmental speech information in both NH and CI listeners. Three listening experiments were conducted to achieve this objective. In the first experiment (Chapter 3), NH listeners' perception of prosody that occurs on sentence level was compared to word recognition in meaningful sentences. Sentence materials that contained prosody conveying either conditional or unconditional permission, agreement or approval were recorded from a male and female speaker and validated in NH listeners. Perception of the prosodic contrast on three different levels of SWN (-2, -5 and -8 dB SNR) was compared to perception of words in meaningful sentences produced by the same speakers and presented at the same SNRs. This experiment tested perception only in NH listeners to explore the suggestions in existing literature that some prosodic cues are more noise-immune than segmental speech information when perceived by NH listeners (Mattys, 2004; Mattys *et al.*, 2005; Smith *et al.*, 1989).

The second experiment (Chapter 4) was designed to test prosody and vowel recognition in identical (2AFC) test paradigms. Perception of two prosodic contrasts that occur on single-word level were compared to perception of vowels, also presented in single-word contexts. One of the prosodic contrasts was attitudinal (a certain/hesitant contrast), while the other was linguistic (a question/statement contrast). Vowel perception was tested using three vowel pairs selected from a complete set of vowels on the basis of a few important acoustic characteristics (frequency of F1 and F2, and duration). Both NH listeners and CI users participated in this experiment. Validated test materials recorded from four speakers (two male) were used in the experiments. Perception was tested in quiet as a baseline listening condition, and in an adaptive noise paradigm.

Because of the conflicting findings reflected in Chapters 3 and 4, the experiment reported in Chapter 5 was designed to compare prosody and phoneme perception on sentence level (as in Chapter 3), using identical test paradigms (as in Chapter 4). In this experiment, fixed SNRs were used instead of an adaptive paradigm to enable a simple direct comparison of the deterioration slope of phoneme and prosody recognition. Both NH listeners and CI users participated. Prosody perception was tested using nonsense sentences (jabberwocky) depicting four different emotions (fear, happiness, sadness, and anger), while phoneme perception was tested using sentences that did not provide semantic or syntactic clues to the identity of the phonemes. Both prosody and phoneme perception were tested in a 4AFC test paradigm.

6.2 SUMMARY OF RESULTS

The key results of the study are shown in Table 6.1, along with the research questions addressed in each experiment.

Table 6.1: Summary of research questions and results

RESEARCH QUESTION	RESULT
Experiment 1 (Chapter 3)	
1. Are NH listeners better at perceiving prosody on sentence level than at recognising words in a sentence in background noise?	• At the most difficult SNR tested (-8 dB SNR), prosody recognition scores, corrected for guessing, were significantly better than word recognition ($p < 0.05$). • Across the three SNRs and the two speakers, the deterioration slope of word recognition was significantly steeper ($p < 0.05$) than that of prosody recognition, suggesting a greater degree of noise immunity of the acoustic cues marking the conditional/unconditional prosodic contrast than the cues that supported word recognition.
Experiment 2 (Chapter 4)	
2. Are NH listeners better at perceiving prosody on single-word level than at recognising vowels in single words in background noise?	• Results from the adaptive noise condition indicated that vowel perception was significantly easier than prosody perception for both NH and CI listeners.
3. Are CI listeners better at perceiving prosody on single-word level than at recognising vowels in single words in background noise?	• The most difficult task for CI recipients in both quiet and noisy listening conditions was the question/statement distinction, which relied heavily on perception of F0 and intensity differences.
Experiment 3 (Chapter 5)	
4. Is the slope of deterioration of prosody recognition shallower with increasing levels of background noise than the slope of phoneme recognition in NH listeners? 5. Is the slope of deterioration of prosody recognition shallower with increasing levels of background noise than the slope of phoneme recognition in CI recipients?	• The results of the third listening experiment showed that in NH listeners the deterioration slope of emotional prosody perception with decreasing SNR did not differ significantly from the deterioration slope of either vowel or consonant perception and at the poorest SNR tested, prosody recognition did not differ significantly from either vowel or consonant recognition. • In contrast, prosody perception in the CI user group deteriorated with a significantly shallower slope than consonant perception in the third experiment. No difference was found between the deterioration slopes of prosody and vowels in this listener group. • CI users performed noticeably worse on emotional prosody recognition than NH listeners, even in quiet and at the best SNR tested. This result indicated that CI users had very limited access to the cues required for accurate perception of emotional prosody even under ideal listening conditions. However, it seemed that whichever cues they were relying on to attain some degree of prosody perception in quiet and low noise levels remained largely available to them even at the poorest SNR (higher noise levels) tested.

The first hypothesis of this study was that, in noise, CI listeners would perform worse on the recognition of prosody in comparison to the recognition of phonemes and words, based on existing literature demonstrating the difficulty that these listeners have with prosody perception (see Tables 2.3 and 2.4 for an overview). Results from

Chapter 4 indicated that prosody recognition was significantly poorer than vowel recognition in noise when tested with single-word stimuli. Findings from Chapter 5 also indicated that emotional prosody perception, along with consonant perception, were significantly poorer than vowel perception when tested with sentence materials. Even in quiet, CI users performed poorly on prosody recognition.

The second hypothesis of the study was that for NH listeners, the perception of prosody in noise would be better than the perception of phonemes or words in noise. While the results of the first experiment (Chapter 3) supported this hypothesis, results of the subsequent experiments did not. It was reported in Chapter 4 that prosody on single-word level was significantly more difficult to recognise than vowels, while results in Chapter 5 showed no significant difference between prosody and phoneme recognition at the most difficult SNR tested. It was also shown that the deterioration of prosody perception with decreasing SNR did not differ significantly from the deterioration of phoneme perception.

6.3 DISCUSSION

6.3.1 Normal-hearing listeners

The present study posed the hypothesis that prosody may be more noise-immune than segmental features and endeavoured to test this hypothesis by directly measuring and comparing recognition of prosody and segmental information in both NH and CI listeners. The results of the first experiment supported this hypothesis, by showing that NH listeners performed better on prosody recognition than word recognition as SNR decreased. However, the results of the second and third experiments did not confirm this finding. This difference between the findings reported in Chapter 3 and those of Chapters 4 and 5 may be due to methodological differences. While the listening experiment of Chapter 3 involved sentence materials, the experiment of Chapter 4 used single-word stimuli, and it was therefore speculated that the redundancy of prosody observed on sentence level might be absent on single-word level. However, the third listening experiment (Chapter 5) also used sentence-level materials, and did not corroborate the results of Chapter 3 either. Another possible explanation for the discrepancy between the first and following experiments may be a difference in test paradigms. In the first experiment, prosody and word recognition were tested using different test paradigms (closed set 2AFC and open set, respectively), while both the second and third experiments used identical test paradigms to test prosody and segmental perception. Open-set testing, where listeners are not offered a limited number of options to choose from, constitutes a more difficult task than closed-set testing, especially if there are only two alternatives in the closed set. Reducing the number of options offered in a closed set reduces the difficulty of the listening task. This is demonstrated, for example, in the study by Green *et al*. (2005), who used a smaller number of alternatives (five or nine) in a vowel recognition test for CI users compared to NH listeners (who were tested with 12 alternatives) because of the poor performance of the CI users. If increasing the number of options in a closed set test notably affects performance, it is to be expected that an open set without specific options (theoretically constituting an infinite number of options) will be more difficult than a closed set with limited options.

An additional difference between the first and following experiments was the type of prosody used in the test materials. Recognition of the prosodic pattern presented in the first experiment (Chapter 3) relied to some extent on the perception of stress (accent) on the noun in the sentence, while the prosody patterns used in the other two experiments did not specifically relate to perception of stress. Acoustically, stress or emphasis is marked by a combination of intensity, duration and F0 changes (Cruttenden, 1997; Fry, 1958), which might make it a particularly redundant prosodic feature. This suggestion is supported by previous reports such as the study by Smith *et al*. (1989), who demonstrated that NH listeners were able to perceive word boundary and rhythm cues (which rely on the perception of stressed and unstressed syllables) at an SNR where phonemes were not discernible. Other work on word boundary cues (Mattys, 2004; Mattys *et al.*, 2005) has also suggested that stress is more tolerant to signal degradation than acoustic-phonetic cues such as co-articulation cues. In contrast, both the second and third experiments (Chapters 4 and 5) showed that phoneme recognition was an easier task than prosody recognition for both NH and CI listeners in quiet, and prosodic cues did not show a greater degree of noise immunity than phonemes. These experiments tested prosody perception using patterns of linguistic, attitudinal and emotional prosody, which did not specifically involve perception of sentence stress. However, the perception of these prosodic patterns, especially perception of emotional prosody, was also supported by a combination of intensity, duration and F0 changes. Therefore, it seems unlikely that stress should be a more salient prosodic feature than, for example, emotional prosody.

Another possible explanation for the difference between the present findings and the suggestions of noise immunity of prosody in the literature may be the differences in methodology. The study by Smith *et al*. (1989) used only one SNR (-10 dB), and reported that segmental features could not be recognised at this SNR. It is possible that the listeners in their experiments might have been able to recognise some of the segmental features in the speech materials, especially if they had been offered a limited number of options, as with the prosodic cues. The study by Mattys (2004) compared co-articulation (a sub-segmental cue) and stress (a suprasegmental or prosodic cue) as cues to word boundaries. In the present work, however, prosody recognition was compared to recognition of segmental information (phonemes), which might be more noise-immune than a sub-segmental cue such as co-articulation. No studies could be found that compare the relative noise immunity of co-articulatory cues and phonemes directly, but findings on co-articulation seem to indicate that these cues are relatively vulnerable to noise effects, because of their dependence on accurate processing of fine-grained, low-level acoustic properties (Fernandes, Ventura and Kolinsky, 2007; Mattys, 2004; Mattys *et al.*, 2005).

The results reported in Chapters 4 and 5 indicate that, when directly comparing NH listeners' ability to explicitly recognise a specific prosodic pattern (a combination of prosodic cues that convey a specific meaning such as the speaker's emotional state) in noise to their ability to recognise phonemes in noise, perception of prosody is not particularly noise-immune. However, in addition to the important communicative functions that prosody fulfils by conveying explicit meaning through the combination of prosodic cues in a particular pattern (e.g. emotional prosody or question/statement contrast), prosody also plays a supporting role in speech recognition, especially in noise. Prosody has been shown to act as a structure that organises spoken language and that plays a role in the perception of word boundaries, syntactic and semantic structure (see Cutler *et al.*, 1997 for an overview of supporting research evidence). It appears that prosody also plays an important role in supporting speech perception in the presence of noise. The studies mentioned earlier (Mattys, 2004; Mattys *et al.*, 2005; Smith *et al.*, 1989) demonstrate that the prosodic cues to stress and word boundaries (which support sentence and word recognition) remain available to listeners at low SNRs. Other studies have shown that flattening or inverting the F0 contour of speech has a negative influence on speech intelligibility in noise (Binns and Culling, 2007; Laures and Bunton, 2003; Laures and Weismer, 1999; Watson and Schlauch, 2008). This effect is especially ascribed to the importance of F0 movement for highlighting specific content words and thereby directing the listener's attention towards them, as well as the presence of appropriate stress and speech rhythm that a natural F0 contour produces by accenting specific syllables, which helps listeners to identify words according to their stress patterns (Binns and Culling, 2007; Laures and Weismer, 1999). These studies did not directly compare segmental and prosody perception, but their results indicated the availability of F0 variation (a prosodic cue) in background noise, as well as its support of speech intelligibility. When one considers the results of the present study in light of these reports, it appears that it is not the explicit recognition of specific prosodic patterns (e.g. a question/statement intonation, or the expression of a specific emotion) that makes prosody an important cue for speech recognition in noise. Rather, perception of the variation of F0 *supports* the perception of meaningful speech.

Neuro-imaging studies on the processing of spoken language and prosody suggest a difference between explicit and implicit processing of emotional prosody, with explicit processing requiring a conscious judgement of the speaker's emotion and implicit processing involving perception of emotional prosody, without the need for this explicit judgement (Bach, Grandjean, Sander, Herdener, Strik and Seifritz, 2008). Results from functional magnetic resonance imaging indicate the involvement of different brain regions in implicit and explicit prosody tasks, supporting the theory that these are two separate processes (Bach *et al.*, 2008; Fruhholz, Ceravolo and Grandjean, 2012). The prosody perception tasks in the present work all involved explicit judgements of specific prosodic patterns. It is possible that implicit processing of prosody, including both emotional prosody and linguistic prosody (which supports word boundary cues, for example) may be a simpler task and that it is this kind of processing that supports speech recognition in noise even when the level of noise interferes with explicit judgements of specific prosodic patterns. This hypothesis is based on the view that the implicit use of prosody as supportive cues is analogous to a discrimination task – a simpler task than a recognition task, which is what explicit prosody recognition would require. Another way in which emotional prosody might enhance perception of speech may be through the recruitment of additional neuronal resources when a specific emotion such as anger is perceived (Grandjean, Bänziger and Scherer, 2006). This could mean, for example, that angry speech may draw listeners' attention and cause them to allocate additional processing resources, thereby improving their perception of the speech, and this might also apply in challenging listening conditions. In conclusion, the present findings, viewed in conjunction with existing reports, suggest that although explicit recognition of prosody might not be particularly noise-immune, implicit use of prosody as supportive cues to speech recognition may play an important role in speech recognition in noise.

6.3.2 Cochlear implant users

The listening experiments involving CI listeners emphasised the difficulty that these listeners have with the perception of prosody. Although the CI users performed worse on vowel and consonant recognition than NH listeners in quiet, as also shown by previous studies (Munson *et al.*, 2003; Stacey *et al.*, 2010), it was insightful to find that prosody perception was even poorer than phoneme perception in the CI group. This finding agreed with that of Luo *et al.* (2009), who reported that Mandarin-speaking CI users performed better on vowel recognition than tone recognition (which is related to perceiving changes in voice F0). However, the present work offered the first direct comparison between phoneme and prosody recognition in a non-tonal language, using prosody materials that involved more than simply an intonation contour difference. This comparison revealed that prosody recognition was significantly poorer than phoneme recognition in the CI user population, especially in quiet and low levels of noise. At the poorest SNR tested, prosody and consonant recognition did not differ significantly, but both of these were significantly poorer than vowel recognition. The difficulty that CI users had with prosody recognition may have been due to inadequate access to acoustic cues that underlie prosody, especially F0 cues, as discussed in the following paragraphs.

Accurate perception of voice pitch (roughly correlated to voice F0) underlies the perception of many important prosodic contrasts (Lakshminarayanan *et al.*, 2003). As discussed in Chapter 2, CI users have limited access to pitch cues. This is due to a variety of factors such as inaccurate encoding of low-frequency information (Kong *et al.*, 2005); a lack of temporal fine structure in the signal (Brown and Bacon, 2010; Kong *et al.*, 2005; Qazi *et al.*, 2013; Shannon *et al.*, 2011); and current spread, which makes it difficult for users to distinguish stimulations of electrodes that are close together which, in turn, results in a limited number of effective frequency channels (Brown and Bacon, 2010). The processing strategy used by most of the CI users in this study (ACE) does not explicitly encode F0, nor does it explicitly encode temporal fine structure. Existing reports have shown that CI users have difficulty with the perception of prosodic contrasts that rely on accurate F0 perception, such as question/statement contrasts (Meister *et al.*, 2009; Most *et al.*, 2012; Peng *et al.*, 2008) and emotional prosody perception (Luo *et al.*, 2007; Most *et al.*, 2012; Pereira, 2000). The findings of both Chapters 4 and 5 corresponded with these reports. Chapter 4 reported that CI recipients had the greatest degree of difficulty with the question/statement discrimination task (as compared to vowel perception and the certain/hesitant discrimination task), which relied to a large extent on F0 changes, while Chapter 5 showed the difficulty they experienced with emotional prosody perception, in which F0 cues also play an important role.

In contrast to their difficulty with pitch perception, CI users' perception of some temporal cues (such as those measured in gap detection tasks) is reported to be close to that of NH listeners (Moore and Glasberg, 1988; Sagi *et al.*, 2009). In the listening experiment reported in Chapter 4, it was found that CI users performed better with the certain/hesitant discrimination task than the question/statement discrimination. This may be because the former task depended more on durational cues (Van Zyl and

Hanekom, 2013b), while the latter relied more on pitch perception. The confusion patterns between different emotions reported in Chapter 5 also suggested that CI users might have relied on intensity and durational cues more than on voice pitch (F0) cues to recognise emotional prosody.

As with temporal cues, CI users' perception of intensity cues is close to that of NH listeners. Rogers *et al*. (2006) measured intensity DLs in CI users in free field using their everyday clinical processors. They reported that, on average, DLs were larger for CI users than NH listeners, but there was substantial variability within the CI group and some CI users were able to discern intensity changes within the same range as NH listeners. However, the sensitivity of some CI users to small intensity changes is to some degree offset by a dynamic range that is much smaller than that of NH listeners. This small dynamic range results in a small number of discriminable intensity steps (approximately 10 to 20 steps compared to 50 to 200 in NH listeners) (Kreft, Donaldson and Nelson, 2004; Rogers *et al.*, 2006). Nonetheless, the limited number of discriminable intensity steps CI users attain appears not to affect their performance in consonant and vowel recognition tasks (Loizou *et al.*, 2000). For sentence recognition, an inverse relationship exists between the number of intensity steps and spectral channels required, with as few as two discriminable intensity steps being needed when a SPEAK-type processing strategy is used and up to 16 steps being needed in the case of four frequency channels (Loizou *et al.*, 2000). Although this study by Loizou *et al*. (2000) investigating the effects of intensity discrimination on speech perception did not test prosody perception, their findings suggest that the intensity resolution provided by most CI processors is adequate to facilitate accurate perception of phonemes and sentences, which makes it plausible that this resolution is also sufficient to support prosody perception. Two previous studies have shown empirical support for this hypothesis. Both Pereira (2000) and Luo *et al*. (2007) have shown that amplitude normalisation (equalising the amplitude levels of recorded emotional prosody speech materials) has a significant impact on CI users' perception of emotional prosody, while the effect on the perception of NH listeners is much smaller. In the present work, intensity cues were preserved in the test materials of the third experiment (Chapter 5). The results from this experiment showed that the limited amount of emotional prosody perception CI users were able to attain in quiet remained relatively intact despite increasing noise levels, indicating that the cues they relied on to perceive prosody were fairly noise-immune. The confusion matrix results reported in Chapter 5 (section 5.5.2) showed mutual confusion between emotions with similar average intensities. Although intensity DLs were not directly measured in the CI participants of the present work, the findings suggested that CI users may have relied heavily on intensity cues to attain the limited amount of accuracy they did in the emotional prosody perception task, and these cues remained accessible even at the lowest SNR level measured.

The hypothesis discussed in section 6.3.1. was that prosodic cues might support the speech perception in noise of NH listeners, as suggested by a number of studies (e.g. Binns and Culling, 2007; Laures and Bunton, 2003). A study by Meister *et al.* (2011) has demonstrated that CI users derive less benefit than NH listeners from a naturally varying intonation contour to improve speech intelligibility when compared to an inverted contour. This may be because of a lack of the supporting structure that prosody, and specifically F0, provides for lexical segmentation. Spitzer *et al.* (2009) showed that although CI users are able to use syllabic stress cues for word segmentation in degraded speech, listeners with residual low-frequency acoustic hearing in the ear opposite the implant were able to derive more benefit from F0 cues for lexical segmentation. The findings of the present work, which showed that prosody perception, particularly perception of F0-related differences, was poorer than phoneme perception (especially vowel perception) in noise, suggest that their inability to use the natural F0 contour might be an important contributor to the difficulty that CI listeners have with speech recognition in noise.

The relative difficulty that CI users in the present work experienced with prosody recognition compared to vowel recognition raises the question why the cues involved in vowel perception (specifically formant frequencies) are more accessible to these listeners than the cues needed for voice pitch (F0) perception. Formant frequencies may be perceived based on place cues (tonotopic representation), as different formant frequencies result in stimulation of different electrodes (Svirsky, Silveira, Suarez, Neuburger, Lai and Simmons, 2001). The spectral information relevant to vowel discrimination is, however, presented to the incorrect place on the auditory nerve array of implant users, owing to the limited insertion depth of the electrode (Rosen, Faulkner and Wilkinson, 1999). This mismatch between the input acoustic frequency assigned to the electrodes and the tonotopic stimulation range of the cochlea can impair vowel recognition significantly, depending on the size of the spectral shift (Baskent and Shannon, 2007; Fu and Shannon, 1999). Dorman *et al.* (1997) showed that simulated electrode insertion depths of 22 and 23 mm resulted in poorer speech recognition (vowels, consonants, and sentences) than normal, while performance with simulated 25 mm insertion was closer to normal. However, there is evidence that CI users might adapt to the spectrally shifted speech over time, thus improving their recognition of segmental speech information (Fu and Shannon, 1999; Li, Galvin III and Fu, 2009; Rosen *et al.*, 1999). It has also been shown that NH listeners presented with envelope cues (such as those perceived by implant users) can identify consonants at a level above chance, and the envelope information can be used to divide consonants into four envelope feature groups, which in combination with visual clues can theoretically convey 95% of consonant information (Van Tassell, Soli, Kirby and Widin, 1987).

F0 perception, on the other hand, relies on temporal fine structure cues, which are inadequately represented in CIs (Kong *et al.*, 2005). A recent review by Oxenham (2013) suggested that place cues might also play an important role in perception of F0 in complex sounds, and this would require access to low-numbered harmonics. Oxenham, Bernstein, and Penagos (2004) have demonstrated that NH individuals listening to multiple low-frequency harmonics presented to high-frequency regions of the cochlea were unable to extract F0 from these stimuli, indicating the importance of correct tonotopic presentation in pitch perception. In the case of CI users, place coding of low-frequency harmonics is hampered by the limited insertion depth of the electrode array, which could also provide an explanation for the difficulty these listeners experience with F0 perception. This may be why providing CI users with access to low-frequency acoustic stimuli through the use of either a hybrid electric-acoustic implant (which preserves low-frequency acoustic hearing) or the use of a hearing aid on the contralateral ear tends to improve their F0 perception (Cullington and Zeng, 2011; Kong *et al.*, 2005). The results of the present work, which directly compared prosody and phoneme perception, suggest that the limitations of CIs that affect F0 perception in particular and prosody perception in general (lack of temporal fine structure and shallow electrode insertion depth), appear to be have a more prominent effect than the limitations that spectral shift imposes on phoneme perception.

6.4 IMPACT OF THE STUDY

6.4.1 Test materials

The test materials and methods developed for the present study provide valuable resources for further research comparing prosody and segmental feature perception. In future developments of new speech processors and pre-processing and processing strategies, these methods could be used as part of the experimental evaluations of these developments. The methods described in Chapter 4 provide a useful tool to assess and compare word-level prosody and vowel perception. Although these materials represent only a small subset of vowels, they can be used to indicate which vowel cues (e.g. F1, F2 or duration) are problematic for the listener. The methods of the third experiment (Chapter 5) are useful to assess sentence-level emotional prosody and directly compare this to phoneme perception. Continued efforts to improve CI users' access to F0 cues (see Brown and Bacon, 2010 for an overview), which play an important role in conveying many prosodic cues, could be evaluated with these methods to determine the relative effects of new processing strategies on segmental and prosodic information, as these methods compare segmental and prosody perception in identical test paradigms.

6.4.2 Noise immunity of prosody in NH listeners

The findings of the listening experiments contribute to researchers' understanding of speech perception in both NH listeners and CI users. Pertaining to NH listeners, the first experiment confirmed the initial hypothesis that prosody appears to be more noise-immune than the recognition of words in a sentences. However, the second and third experiments used a more rigorous approach in comparing prosody and segmental information through the use of identical test paradigms, and did not confirm this finding. The findings of this study as a whole therefore suggest that NH listeners' perception of specific prosodic patterns (such as question/statement contrasts or emotion) is not more immune to the effects of noise than their perception of consonants and particularly vowels. Since no other reports in existing literature directly compare prosody and segmental feature perception in different degrees of noise and identical test paradigms, this contributes to the current understanding of how NH listeners perceive speech in noise. It demonstrates that the relative success (compared to CI users) which NH listeners exhibit in recognising speech in noise is probably not primarily due to their perception of specific prosodic patterns. This does not mean, however, that some prosodic cues (such as a naturally varying intonation contour) do not play an important role in supporting speech recognition in noise, as demonstrated in previous research (e.g. Laures and Bunton, 2003) and discussed in section 6.3.1.

6.4.3 Prosody perception in CI users

The present study has confirmed previous reports that CI listeners have difficulty with the perception of prosody. However, by directly comparing perception of segmental and prosodic features in both quiet and noise, this study contributed to existing knowledge by demonstrating that CI users' perception of prosody is significantly poorer than their perception of segmental information (phonemes), especially in quiet and at low noise levels. In light of the fact that prosody fulfils a variety of important communicative functions in daily life (as summarised in Chapter 2), this finding indicates the importance of ongoing efforts to improve prosody perception in CI users. Results of the listening experiments suggested that of all the different cues underlying prosody perception (duration, intensity and pitch cues), F0 or pitch cues were particularly problematic. Improving the delivery of pitch cues to CI users could therefore be helpful in improving their perception of prosody. Attempts to improve F0 perception include techniques such as current steering, i.e. weighting current delivered to adjacent electrodes to create virtual frequency channels (e.g. Geurts and Wouters, 2004), amplitude modulations of electric pulses with an extracted F0 (Green *et al.*, 2005), and processing strategies designed to convey temporal fine structure cues (Qi *et al.*, 2012), but the success of these efforts has been limited thus far (Brown and Bacon, 2010; Wilson and Dorman, 2008). Results obtained using recordings from one of the female speakers in the second experiment (FS1), suggested that an exaggerated F0 contour might facilitate better pitch perception in CI users, especially in noise. This possibility should be investigated in future research to determine the effects and viability of delivering an exaggerated F0 contour to CI users. A different approach to improving F0 perception, as mentioned under 6.3.2, is the addition of acoustic hearing by means of a hearing aid on the contralateral ear (if there is residual hearing in that ear) or electric-acoustic (hybrid) stimulation, which uses a modified electrode that preserves low-frequency acoustic hearing in the implanted ear (Brown and Bacon, 2010; Cullington and Zeng, 2011; Kong *et al.*, 2005). Recent work has also indicated that simple music training of CI users might result in a notable improvement in prosody perception, as well as a small improvement in speech perception in noise, although this finding is based on preliminary results (Patel, 2014).

In both NH and CI listeners, the present work compared the noise immunity of different prosodic patterns assessed in the same test paradigms. In the second experiment, question/statement discrimination was compared to discrimination between a certain and hesitant attitude on single-word level. In both listener groups, no significant difference was found between the recognition of the two prosodic contrasts, although the differences between the prosody tasks and the vowel recognition tasks in that experiment indicated that while NH listeners performed slightly poorer on the certain/hesitant task, CI users had particular difficulty with the question/statement task. This difference could be due to the underlying acoustic cues; while certain/hesitant discrimination was supported by durational cues, question/statement discrimination relied more heavily on F0 perception, which is known to be problematic in CI users. The difficulty that CI users experience with this task was highlighted by the addition of noise, and the result suggested that durational cues were more noise-immune in this group than F0 cues, although the same was not true in NH listeners. This finding suggests that further exploration of the relative noise immunity of duration and F0 cues in CI users may be valuable, especially if ways to exploit durational cues to improve speech recognition in noise can be found. Recent work has shown promising results of auditory training using simple stimuli to improve speech recognition in noise in CI users (Oba, Fu and Galvin III, 2011).

Furthermore, there is evidence that behavioural training can improve cortical processing of temporal cues in cats, even after long periods of deafness (Vollmer and Beitel, 2011). In light of these findings, auditory training methods teaching CI users to make use of speech rate and utterance or phoneme duration cues may be a possible way to increase the exploitation of durational cues to improve speech recognition in noise.

6.4.4 Fixed versus adaptive SNR test protocol

The first and third experiments of this study used fixed SNRs, while the second experiment used an adaptive procedure, which altered the SNR according to the listener's performance. Since both of these methods were used here in the same study with similar test methods and listener groups, this provides an opportunity to compare the two methods, especially for the use of testing speech recognition in noise in CI users. The use of a fixed SNR was a suitable technique to enable a simple comparison of deterioration slopes, as was done in the third experiment. The limitation of using this method in a population of CI users is that it requires pilot testing to determine suitable SNR levels that would not result in floor and ceiling effects. An adaptive SNR technique avoids floor and ceiling effects – a useful attribute when testing the CI population, which may show a broad range of performance levels - but this method resulted in very long testing times owing to the need to repeat the procedure several times for each task and each listener. Despite the need for pilot testing, the use of fixed SNRs seems to be a more appropriate technique for comparing listening tasks in noise.

6.5 LIMITATIONS OF THE STUDY AND SUGGESTIONS FOR FUTURE RESEARCH

6.5.1 Test paradigm

The listening experiment described in Chapter 3 compared prosody recognition in a closed set (2AFC) test paradigm to word recognition in an open set paradigm. Although prosody recognition scores were corrected for guessing, this method may not have resulted in a fair comparison between the two tasks. To address this limitation, subsequent experiments described in Chapters 4 and 5 were designed to test the two speech feature types (prosody and segmental information) in identical test paradigms. It may be insightful to compare prosody and segmental information recognition using an open set paradigm (not offering listeners a limited set of options to choose from) for both feature types in future work, as such a paradigm is more representative of everyday listening experiences. However, the number of possible options in an open set prosody recognition task is not infinite, and neither is the number of possible options in a phoneme recognition task. Therefore, even an "open set" test paradigm, where listeners are not explicitly offered a limited number of options, could result in performance differences between prosody and phoneme recognition, since the number of possible options in an open set prosody recognition task is likely to be different from the number of possible options in a phoneme recognition task.

6.5.2 Vowel selection

In the second listening experiment, a small sub-set of vowels selected from a complete collection of vowels was used to evaluate segmental feature perception. These phonemes were selected on the basis of specific acoustic cues (F1, F2 and duration). Each vowel pair used in the 2AFC paradigm differed with regard to one of these important acoustic cues. The motivation behind selecting only a small number of vowel pairs was to limit testing time, since the adaptive procedure used in that experiment resulted in approximately seven hours of testing per listener. The limitation of this method is that the selection of vowel pairs may not have been representative of the complete collection of Afrikaans vowels, and findings on vowel recognition could therefore not necessarily be generalised to all vowels in the test language. To address this limitation, the third listening experiment used a larger collection of vowels (n = 10), which included all vowels with a proportional representation of ≥ 1% in the speech sample collected in a study on phoneme occurrence in Afrikaans (Van Heerden, 1999). The results of the third listening experiment corresponded well with those of the second experiment, indicating that the selection of a small sample of vowels for the second experiment did not have a major effect on its outcomes.

6.5.3 Background noise type

The present study used only SWN as background noise, as this was considered to represent a difficult listening condition and it could be ensured that all speakers' voices were equally masked through the use of speaker-specific noise. It remains to be seen, therefore, whether a different type of background noise, such as multi-talker babble, would result in the same findings reported here. Multi-talker babble produces both energetic masking, which means that parts of the signal is rendered inaudible by the interfering noise, and informational masking, which puts an additional cognitive load on listeners and competes for their attention (Cooke, Lecumberri and Barker, 2008). However, the amplitude modulations of multi-talker babble allow for some masking release, which could in some cases make the listening task easier by allowing listeners to "glimpse" the signal through gaps in the masker noise (Cooke, 2006; Jin and Liu, 2012). Multi-talker babble could also be considered more representative of the type of background noise that listeners are faced with daily, and listening experiments exploring the relative noise immunity of segmental and prosodic information in this type of noise could therefore be valuable. Results reported by Parikh and Loizou (2005) have shown that at low SNRs, multi-talker babble has a greater effect on phoneme intelligibility than SWN, while Laures and Bunton (2003) reported similar effects of multi-talker babble and SWN on speech intelligibility when the F0 contour is flattened. It is possible, therefore, that the advantage of phonemes over prosody found in the present work might be smaller or even absent if multi-talker babble is used as a masker.

6.5.4 Auditory-only cues

In the present work all stimuli, including the emotional utterances of the last experiment, were presented in an auditory-only condition, with no visual cues. Investigating the perception of speech cues in an auditory-only condition is an important endeavour, as listeners are sometimes faced with situations where visual cues are not available (e.g. on the telephone, or in the case of visually impaired listeners), and the addition of visual cues may mask the effects of specific auditory cues on perception. However, in many communication situations, auditory and visual cues are both available to listeners. It has long been known that visual cues aid in the perception of phonemes (Sumby and Pollack, 1954), and recent work has shown that the addition of visual cues helps to improve perception of emotional prosody (Paulmann and Pell, 2011). The relative contributions of auditory and visual cues to phoneme and prosody perception have not been investigated, and future work should address this research gap. If such a comparison revealed, for example, that phoneme perception gained more support from visual cues than prosody, it could indicate that improved access to prosody on an acoustic level should receive precedence over improved access to segmental information, as listeners rely more heavily on acoustic cues to perceive prosody.

6.5.5 Concurrent cues

The second and third listening experiments of this study have provided some indirect indications of which acoustic cues underlying prosody are particularly problematic for CI users. However, the use of natural speech that contains concurrent intensity, duration, and pitch cues to prosody makes it difficult to determine the individual contribution of each of these different underlying cues. Future work should further explore how CI recipients use different cues to perceive emotional prosody by systematically manipulating the underlying acoustic cues. The work by Luo *et al.* (2007) has shown that eliminating intensity cues has a greater effect on CI users' perception of emotional prosody than on NH listeners' perception. The present work also suggested that CI users might rely on intensity cues for perception of emotional prosody, and further indicated that these cues might be particularly noise-resistant, as CI users' emotional prosody perception remained relatively unaffected by increasing levels of noise. Future studies should explore this finding further and examine the possibility of exploiting intensity cues to improve prosody perception in CI recipients. In addition, by eliminating F0 cues by flattening F0 (see e.g. Laures and Bunton, 2003), and durational cues using phase vocoding techniques (Ellis, 2002), the relative contributions of these cues could be compared within and between listener groups (NH and CI listeners). Further insights into the relative importance of these underlying cues could help direct future efforts in improving prosody perception in CI users, and measuring the effects of attempts to improve this in a systematic manner.

6.6 FINAL CONCLUSIONS

The final conclusions drawn from the main findings of the study are as follows.

- Although prosody may be important for successful speech recognition in noise in NH listeners, cues needed for explicit recognition of specific prosodic patterns are not particularly noise-resistant.
- In NH listeners, the deterioration slope of prosody recognition did not differ significantly from that of vowel or consonant recognition.
- The study confirmed that CI users have difficulty with prosody perception in quiet, and demonstrated that their prosody recognition is significantly poorer than their phoneme perception.
- The difference between CI users' prosody and phoneme perception remains evident at low noise levels.
- The deterioration slope of emotional prosody recognition in CI users is significantly shallower than that of consonant recognition, but not of vowel recognition.
- The limited amount of emotional prosody recognition that CI users are able to attain remains relatively intact at poor SNRs, possibly because of their reliance on intensity and durational cues.
- Improving prosody perception, and particularly F0 perception, in CI users should improve their speech perception in quiet, and could indirectly benefit their speech recognition in noise.
- Further exploitation of durational and intensity cues might be valuable in improving CI users' speech perception in noise.

REFERENCES

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

Assmann, P. & Summerfield, Q. (2004). "The Perception of Speech Under Adverse Conditions," in *Speech Processing in the Auditory System*, S. Greenberg et al., eds., Springer-Verlag, New York, pp. 231-308.

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

Bach, D. R., Grandjean, D., Sander, D., Herdener, M., Strik, W. K. and Seifritz, E. (2008). The effect of appraisal level on processing of emotional prosody in meaningless speech, *NeuroImage* **42**(2): 919-927.

Balkany, T., Hodges, A., Menapace, C., Hazard, L., Driscoll, C., Gantz, B., Kelsall, D., Luxford, W., McMenomy, S., Neely, J. G., Peters, B., Pillsbury, H., Roberson, J., Schramm, D., Telian, S., Waltzman, S., Westerberg, B. and Payne, S. (2007). Nucleus Freedom North American clinical trial, *Otolaryngology - Head and Neck Surgery* **136**(5): 757-762.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psychology* **70**(3): 614-636.

Baskent, D. and Shannon, R. V. (2007). Combined effects of frequency compression-expansion and shift on speech recognition, *Ear and Hearing* **28**(3): 277-289.

Batliner, A., Steidl, S. and Nöth, E. (2007). Laryngealizations and emotions: How many Babushkas?, *Proceedings of International workshop on paralinguistic speech - between models and data, 2007, Saarbrücken, Germany*, pp. 17-22.

Beattie, R. C., Barr, T. and Roup, C. (1997). Normal and hearing-impaired word recognition scores for monosyllabic words in quiet and noise, *British Journal of Audiology* **31**(3): 153-164.

Berckmoes, C. and Vingerhoets, G. (2004). Neural foundations of emotional speech processing, *Current Directions in Psychological Science* **13**(5): 182-185.

Berkovits, R. (1984). A perceptual study of sentence-final intonation, *Language and Speech* **27**(4): 291-308.

Binns, C. and Culling, J. F. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech, *Journal of the Acoustical Society of America* **122**(3): 1765-1776.

Boersma, P. & Weenink, D. Praat: doing phonetics by computer. [5.1.32]. 2010. http://www.praat.org/.
Ref Type: Computer Program

Boothroyd, A. (1988). Perception of speech pattern contrasts from auditory presentation of voice fundamental frequency, *Ear and Hearing* **9**(6): 313-321.

Boothroyd, A. and Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition, *Journal of the Acoustical Society of America* **84**(1): 101-114.

Borden, G. J., Harris, K. S. and Raphael, L. J. (2007). *Speech science primer: Physiology, acoustics, and perception of speech*, 5th edn, Lippincott Williams & Wilkins, Philadelphia.

Botha, L. (1996). Towards modelling acoustic differences between L1 and L2 speech: The short vowels of Afrikaans and South-African English, *Institute of Phonetic Sciences Proceedings,* Vol. 20, University of Amsterdam, pp. 65-80.

Botinis, A., Granström, B. and Möbius, B. (2001). Developments and paradigms in intonation research, *Speech Communication* **33**(4): 263-296.

Bradley, J. S. (1986). Predictors of speech intelligibility in rooms, *Journal of the Acoustical Society of America* **80**(3): 837-845.

Breen, M., Fedorenko, E., Wagner, M. and Gibson, E. (2010). Acoustic correlates of information structure, *Language and Cognitive Processes* **25**(7): 1044-1098.

Breitenstein, C., Van Lancker, D. and Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample, *Cognition and Emotion* **15**(1): 57-79.

Brennan, S. E. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers, *Journal of Memory and Language* **34**(3): 383-398.

Broersma, M. and Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise, *Speech Communication* **52**(1-2): 980-995.

Brown, C. A. and Bacon, S. P. (2010). Fundamental frequency and speech intelligibility in background noise, *Hearing Research* **266**(1-2): 52-59.

Burns, E. M. and Viemeister, N. F. (1976). Nonspectral pitch, *Journal of the Acoustical Society of America* **60**(4): 863-869.

Caposecco, A., Hickson, L. and Pedley, K. (2012). Cochlear implant outcomes in adults and adolescents with early-onset hearing loss, *Ear and Hearing* **33**(2): 209-220.

Carlson, K. (2009). How prosody influences sentence comprehension, *Language and Linguistics Compass* **3**(5): 1188-1200.

Caspers, J. (1998). Who's next? The melodic marking of question vs. continuation in Dutch, *Language and Speech* **41**(3-4): 375-398.

Chatterjee, M. and Peng, S. C. (2008). Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition, *Hearing Research* **235**(1-2): 143-156.

Cheang, H. S. and Pell, M. (2008). The sound of sarcasm, *Speech Communication* **50**(5): 366-381.

Clark, G. (2003). *Cochlear Implants: fundamentals and applications*, Springer-Verlag New York, Inc., New York.

Cooke, M. (2006). A glimpsing model of speech perception in noise, *Journal of the Acoustical Society of America* **119**(3): 1562-1573.

Cooke, M., Lecumberri, M. L. G. and Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception, *Journal of the Acoustical Society of America* **123**(1): 414-427.

Cruttenden, A. (1997). *Intonation*, 2nd edn, Cambridge University Press, Cambridge.

Cullington, H. E. and Zeng, F. G. (2011). Comparison of bimodal and bilateral cochlear implant users on speech recognition with competing talker, music perception, affective prosody discrimination, and talker identification, *Ear and Hearing* **32**(1): 16-30.

Cutler, A., Dahan, D. and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review, *Language and Speech* **40**(2): 141-201.

Davids, A., Ferreira, J., Links, T. and Prinsloo, K. (1997). *Afrikaans in Afrika*, J.L. van Schaik Publishers, Pretoria.

Davidson, L. S., Skinner, M. W., Holstad, B. A., Fears, B. T., Richter, M. K., Matusofsky, M., Brenner, C., Holden, T., Birath, A., Kettel, J. L. and Scollie, S. (2009). The effect of instantaneous input dynamic range setting on the speech perception of children with the Nucleus 24 implant, *Ear and Hearing* **30**(3): 340-349.

De Pijper, J. R. and Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues, *Journal of the Acoustical Society of America* **96**(4): 2037-2047.

Dmitrieva, E. S., Gel'man, V. Y., Zaitseva, K. A. and Orlov, A. M. (2008). Dependence of the perception of emotional information of speech on the acoustic parameters of the stimulus in children of various ages, *Human Physiology* **34**(4): 149-153.

Donaldson, B. C. (1991). *The influence of English on Afrikaans*, Academica, Pretoria.

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

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility, *Journal of the Acoustical Society of America* **97**(1): 585-592.

Du Plessis, M. (2005). *Pharos Afrikaans-Engels / English-Afrikaans Dictionary*, Pharos Dictionaries, Cape Town.

Ellis, D. P. W. A Phase Vocoder in Matlab. 2002. http://labrosa.ee.columbia.edu/matlab/pvoc/.
Ref Type: Computer Program

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners, *Journal of the Acoustical Society of America* **116**(4 I): 2365-2373.

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.

Fernandes, T., Ventura, P. and Kolinsky, R. (2007). Statistical information and coarticulation as cues to word boundaries: A matter of signal quality, *Perception and Psychophysics* **69**(6): 856-864.

Field, A. (2009). *Discovering statistics using SPSS*, 3rd edn, SAGE Publications Ltd, London.

Firszt, J. B., Holden, L. K., Skinner, M. W., Tobey, E. A., Peterson, A., Gaggl, W., Runge-Samuelson, C. L. and Wackym, P. A. (2004). Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems, *Ear and Hearing* **25**(4): 375-387.

Flamme, G. A., Stephenson, M. R., Deiters, K., Tatro, A., VanGessel, D., Geda, K., Wyllys, K. and McGregor, K. (2012). Typical noise exposure in daily life, *International Journal of Audiology* **51**(S1): S3-S11.

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.

Fruhholz, S., Ceravolo, L. and Grandjean, D. (2012). Specific brain networks during explicit and implicit decoding of prosody, *Cerebral Cortex* **22**(5): 1107-1117.

Fry, D. B. (1958). Experiments in the perception of stress, *Language and Speech* **1**(2): 126-152.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress, *Journal of the Acoustical Society of America* **27**(4): 765-768.

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.

Fujie, S., Ejiri, Y., Kikuchi, H. and Kobayashi, T. (2006). Recognition of positive/negative attitude and its application to a spoken dialogue system, *Systems and Computers in Japan* **37**(12): 45-55.

Gaines, P. (2011). The multifunctionality of discourse operator okay: Evidence from a police interview, *Journal of Pragmatics* **43**(14): 3291-3315.

Garadat, S. H. and Pfingst, B. E. (2011). Relationship between gap detection thresholds and loudness in cochlear-implant users, *Hearing Research* **275**(1-2): 130-138.

Geluykens, R. (1988). On the myth of rising intonation in polar questions, *Journal of Pragmatics* **12**(4): 467-485.

Geurts, L. and Wouters, J. (2004). Better place-coding of the fundamental frequency in cochlear implants, *Journal of the Acoustical Society of America* **115**(2): 844-852.

Gifford, R. H., Dorman, M. F., Shallop, J. K. and Sydlowski, S. A. (2010). Evidence for the expansion of adult cochlear implant candidacy, *Ear and Hearing* **31**(2): 186-194.

Gifford, R. H. and Revit, L. J. (2010). Speech perception for adult cochlear implant recipients in a realistic background noise: Effectiveness of preprocessing strategies and external options for improving speech recognition in noise, *Journal of the American Academy of Audiology* **21**(7): 441-451.

Gobl, C. and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude, *Speech Communication* **40**(1-2): 189-212.

Goldsworthy, R. L., Delhorne, L. A., Braida, L. D. and Reed, C. M. (2013). Psychoacoustic and phoneme identification measures in cochlear-implant and normal-hearing listeners, *Trends in Amplification* **17**(1): 27-44.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages, *Journal of Multilingual and Multicultural Development* **28**(6): 445-467.

Grandjean, D., Bänziger, T., & Scherer, K. R. (2006). "Intonation as an interface between language and affect," in *Progress in Brain Research*, S. Anders et al., eds., Elsevier, pp. 235-247.

Grant, K. W. and Seitz, P. F. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context, *Journal of the Acoustical Society of America* **107**(2): 1000-1011.

Grant, K. W. and Walden, B. E. (1996). Spectral distribution of prosodic information, *Journal of Speech, Language, and Hearing Research* **39**(2): 228-238.

Gravano, A., Hirschberg, J. and Benuš, S. (2012). Affirmative cue words in task-oriented dialogue, *Computational Linguistics* **38**(1): 1-39.

Green, T., Faulkner, A., Rosen, S. and Macherey, O. (2005). Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification, *Journal of the Acoustical Society of America* **118**(1): 375-385.

Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language, *Linguistics* **21**(3): 501-529.

Grosjean, F. and Hirt, C. (1996). Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects, *Language and Cognitive Processes* **11**(1/2): 107-134.

Hammerschmidt, K. and Jürgens, U. (2007). Acoustical correlates of affective prosody, *Journal of Voice* **21**(5): 531-540.

Hartmann, W. M. (1998). *Signals, Sound, and Sensation*, Springer Science+Business Media, Inc., New York.

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): 3099-3111.

Hopyan-Misakyan, T. M., Gordon, K. A., Dennis, M. and Papsin, B. C. (2009). Recognition of affective speech prosody and facial affect in deaf children with unilateral right cochlear implants, *Child Neuropsychology* **15**(2): 136-146.

House, D. (1994). Perception and production of mood in speech by cochlear implant users, *ICSLP 94, 18 September 19940, Yokohama, Japan*, pp. 2051-2054.

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.

Jin, S. H. and Liu, C. (2012). English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners, *Journal of the Acoustical Society of America* **132**(5): EL391-EL397.

Juslin, P. N. and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code?, *Psychological Bulletin* **129**(5): 770-814.

Kalikow, D. N., Stevens, K. N. and Elliot, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability., *Journal of the Acoustical Society of America* **61**(5): 1337-1351.

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.

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

Kong, Y. Y., Stickney, G. S. and Zeng, F. G. (2005). Speech and melody recognition in binaurally combined acoustic and electric hearing, *Journal of the Acoustical Society of America* **117**(3 I): 1351-1361.

Krahmer, E. and Swerts, M. (2005). How children and adults produce and perceive uncertainty in audiovisual speech, *Language and Speech* **48**(1): 29-53.

Kreft, H. A., Donaldson, G. S. and Nelson, D. A. (2004). Effects of pulse rate and electrode array design on intensity discrimination in cochlear implant users, *Journal of the Acoustical Society of America* **116**(4): 2258-2268.

Kreiman, J. (1982). Perception of Sentence and Paragraph Boundaries in Natural Conversation, *Journal of Phonetics* **10**(2): 163-175.

Ladd, D. R. & Cutler, A. (1983). "Introduction. Models and Measurements in the Study of Prosody," in *Prosody, Models and Measurements*, A. Cutler & D. R. Ladd, eds., Springer-Verlag, Berlin, pp. 1-8.

Lakshminarayanan, K., Ben Shalom, D., Van Wassenhowe, V., Orbelo, D., Houde, J. and Poeppel, D. (2003). The effect of spectral manipulations on the identification of affective and linguistic prosody, *Brain and Language* **84**(2): 250-263.

Laures, J. S. and Bunton, K. (2003). Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions, *Journal of Communication Disorders* **36**(6): 449-464.

Laures, J. S. and Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level, *Journal of Speech, Language, and Hearing Research* **42**(5): 1148-1156.

Laver, J. (1980). *The phonetic description of voice quality*, Cambridge University Press, Cambridge.

Lehiste, I. (1970). *Suprasegmentals*, M.I.T. Press, Cambridge, MA.

Lehiste, I. (1979). "Perception of sentence and paragraph boundaries," in *Frontiers of Speech Communication Research*, B. Lindblom & S. Öhman, eds., Academic Press, New York, pp. 191-201.

Lehiste, I. (1976). "Suprasegmental features of speech," in *Contemporary Issues in Experimental Phonetics*, N. Lass, ed., Academic Press, New York, pp. 225-239.

Leventhal, G. (1973). Effect of sentence context on word perception, *Journal of Experimental Psychology* **101**(2): 318-323.

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

Li, T., Galvin III, J. J. and Fu, Q. J. (2009). Interactions between unsupervised learning and the degree of spectral mismatch on short-term perceptual adaptation to spectrally shifted speech, *Ear and Hearing* **30**(2): 238-249.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English, *Journal of the Acoustical Society of America* **32**(4): 451-454.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech, *Language and Speech* **6**(3): 172-187.

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

Local, J. and Kelly, J. (1986). Projection and 'silences': Notes on phonetic and conversational structure, *Human Studies* **9**(2-3): 185-204.

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

Loizou, P. C., Dorman, M., Poroy, O. and Spahr, T. (2000). Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution, *Journal of the Acoustical Society of America* **108**(5): 2377-2387.

Lorens, A., Zgoda, M., Obrycka, A. and Skarzynski, H. (2010). Fine Structure Processing improves speech perception as well as objective and subjective benefits in pediatric MED-EL COMBI 40+ users, *International Journal of Pediatric Otorhinolaryngology* **74**(12): 1372-1378.

Luo, X., Fu, Q. J. and Galvin III, J. J. (2007). Vocal emotion recognition by normal-hearing listeners and cochlear implant users, *Trends in Amplification* **11**(4): 301-315.

Luo, X., Fu, Q. J., Wu, H. P. and Hsu, C. J. (2009). Concurrent-vowel and tone recognition by Mandarin-speaking cochlear implant users, *Hearing Research* **256**(1-2): 75-84.

Marslen-Wilson, W. D., Tyler, L. K., Warren, P., Grenier, P. and Lee, C. S. (1992). Prosodic effects in minimal attachment, *Quarterly Journal of Experimental Psychology* **45 A**(1): 73-87.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation, *Journal of Experimental Psychology: Human Perception and Performance* **30**(2): 397-408.

Mattys, S. L., White, L. and Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework, *Journal of Experimental Psychology: General* **134**(4): 477-500.

Meiring, B. A. and Retief, R. (1991). *Funksionele Afrikaans*, J.L. van Schaik Pty (Ltd), Pretoria.

Meister, H., Landwehr, M., Pyschny, V. and Grugel, L. (2011). Use of intonation contours for speech recognition in noise by cochlear implant recipients, *Journal of the Acoustical Society of America* **129**(5): EL204-EL209.

Meister, H., Landwehr, M., Pyschny, V., Walger, M. and Wedel, H. V. (2009). The perception of prosody and speaker gender in normal-hearing listeners and cochlear implant recipients, *International Journal of Audiology* **48**(1): 38-48.

Mendel, L. L. & Danhauer, J. L. (1997). "Characteristics of sensitive speech perception tests," in *Audiologic evaluation and management and speech perception assessment*, L. L. Mendel & J. L. Danhauer, eds., Singular Publishing Group, Inc., San Diego, pp. 59-99.

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

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel, *Journal of the Association for Research in Otolaryngology* **85**(5): 2114-2134.

Millotte, S., Wales, R. and Christophe, A. (2007). Phrasal prosody disambiguates syntax, *Language and Cognitive Processes* **22**(6): 898-909.

Monrad-Krohn, G. H. (1947). Dysprosody or altered "melody of language", *Brain* **70**(4): 405-415.

Moore, B. C. J. and Glasberg, B. R. (1988). Gap detection with sinusoids and noise in normal, impaired, and electrically stimulated ears, *Journal of the Acoustical Society of America* **83**(3): 1093-1101.

Morris, D., Magnusson, L., Faulkner, A., Jönsson, R. and Juul, H. (2013). Identification of vowel length, word stress, and compound words and phrases by postlingually deafened cochlear implant listeners, *Journal of the American Academy of Audiology* **24**(9): 879-890.

Morton, J. and Jassem, W. (1965). Acoustic correlates of stress, *Language and Speech* **8**(3): 159-181.

Most, T. and Aviner, C. (2009). Auditory, visual, and auditory-visual perception of emotions by individuals with cochlear implants, hearing aids, and normal hearing, *Journal of Deaf Studies and Deaf Education* **14**(4): 449-464.

Most, T., Gaon-Sivan, G., Shpak, T. and Luntz, M. (2012). Contribution of a contralateral hearing aid to perception of consonant voicing, intonation, and emotional state in adult Cochlear Implantees, *Journal of Deaf Studies and Deaf Education* **17**(2): 244-258.

Most, T. and Peled, M. (2007). Perception of suprasegmental features of speech by children with cochlear implants and children with hearing aids, *Journal of Deaf Studies and Deaf Education* **12**(3): 350-361.

Moudon, A. V. (2009). Real noise from the urban environment: How ambient community noise affects health and what can be done about it, *American Journal of Preventive Medicine* **37**(2): 167-171.

Mozziconacci, S. J. L. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values, *User Modeling and User-Adapted Interaction* **11**(4): 297-326.

Munson, B., Donaldson, G. S., Allen, S. L., Collison, E. A. and Nelson, D. A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as function of overall speech perception ability, *Journal of the Acoustical Society of America* **113**(2): 925-935.

Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustical Society of America* **93**(2): 1097-1108.

Nakata, T., Trehub, S. E. and Kanda, Y. (2012). Effect of cochlear implants on children's perception and production of speech prosody, *Journal of the Acoustical Society of America* **131**(2): 1307-1314.

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

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy, *Journal of Speech, Language, and Hearing Research* **51**(3): 574-585.

Nygaard, L. C., Herold, D. S. and Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning, *Cognitive Science* **33**(1): 127-146.

Oba, S. I., Fu, Q. J. and Galvin III, J. J. (2011). Digit training in noise can improve cochlear implant users' speech understanding in noise, *Ear and Hearing* **32**(5): 573-581.

Orr, S. B., Montgomery, A. A., Healy, E. W. and Dubno, J. R. (2010). Effects of consonant-vowel intensity ratio on loudness of monosyllabic words, *Journal of the Acoustical Society of America* **128**(5): 3105-3113.

Oxenham, A. J. (2013). Revisiting place and temporal theories of pitch, *Acoustical Science and Technology* **34**(6): 388-396.

Oxenham, A. J., Bernstein, J. G. W. and Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception, *PNAS* **1001**(5): 1421-1425.

Pannekamp, A., Toepel, U., Alter, K., Hahne, A. and Friederici, A. D. (2005). Prosody-driven sentence processing: An event-related brain potential study, *Journal of Cognitive Neuroscience* **17**(3): 407-421.

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.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis, *Hearing Research* **308**: 98-108.

Paulmann, S. and Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli?, *Motivation and Emotion* **35**(2): 192-201.

Paulmann, S., Pell, M. D. and Kotz, S. A. (2008). How aging affects the recognition of emotional speech, *Brain and Language* **104**(262): 269.

Pearsons, K. S., Bennett, R. L., & Fidell, S. (1977). *Speech levels in various noise environments*, Bolt, Beranek and Newman Inc., Canoga Park, CA, EPA-600/1-77-025.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions, *Journal of the Acoustical Society of America* **109**(4): 1668-1680.

Pell, M. D., Jaywant, A., Monetta, L. and Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues, *Cognition and Emotion* **25**(5): 834-853.

Peng, S. C., Chatterjee, M. and Lu, N. (2012). Acoustic cue integration in speech intonation recognition with cochlear implants, *Trends in Amplification* **16**(2): 67-82.

Peng, S. C., Tomblin, J. B. and Turner, C. W. (2008). Production and perception of speech intonation in pediatric cochlear implant recipients and individuals with normal hearing, *Ear and Hearing* **29**(3): 336-351.

Pereira, C. (2000). "The perception of vocal affect by cochlear implantees," in *Cochlear Implants*, S. Waltzman & N. L. Cohen, eds., Thieme, New York, pp. 343-345.

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.

Phatak, S. A., Lovitt, A. and Allen, J. B. (2008). Consonant confusions in white noise, *Journal of the Acoustical Society of America* **124**(2): 1220-1233.

Ponelis, F. A. (1979). *Afrikaanse Sintaksis*, J.L. van Schaik Pty (Ltd), Pretoria.

Ponelis, F. A. (1993). *The development of Afrikaans*, Verlag, Frankfurt.

Pretorius, L. L., Hanekom, J. J., Van Wieringen, A. and Wouters, J. (2006). 'n Analitiese tegniek om die foneemherkenningsvermoë van Suid-Afrikaanse kogleêre inplantinggebruikers te bepaal, *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* **25**(4): 195-208.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. (1991). The use of prosody in syntactic disambiguation, *Journal of the Acoustical Society of America* **90**(6): 2956-2970.

Qazi, O., van Dijk, B., Moonen, M. and Wouters, J. (2013). Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility, *Hearing Research* **299**: 79-87.

Qi, B., Krenmayr, A., Zhang, N., Dong, R., Chen, X., Schatzer, R., Zierhofer, C., Liu, B. and Han, D. (2012). Effects of temporal fine structure stimulation on Mandarin speech recognition in cochlear implant users, *Acta Oto-Laryngologica* **132**(11): 1183-1191.

Richardson, L. M., Busby, P. A., Blamey, P. J. and Clark, G. M. (1998). Studies of prosody perception by cochlear implant patients, *Audiology* **37**(4): 231-245.

Rogers, C. F., Healy, E. W. and Montgomery, A. A. (2006). Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions, *Journal of the Acoustical Society of America* **119**(4): 2276-2287.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects, *Philosophical Transactions of the Royal Society B: Biological Sciences* **336**(1278): 367-373.

Rosen, S., Faulkner, A. and Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants, *Journal of the Acoustical Society of America* **106**(6): 3629-3636.

Sagi, E., Kaiser, A. R., Meyer, T. A. and Svirsky, M. A. (2009). The effect of temporal gap identification on speech perception by users of cochlear implants, *Journal of Speech, Language, and Hearing Research* **52**(2): 385-395.

Shannon, C. E. (1948). A mathematical theory of communication, *The Bell System Technical Journal* **27**(3): 379-423.

Shannon, R. V., Cruz, R. J. and Galvin III, J. J. (2011). Effect of stimulation rate on cochlear implant users' phoneme, word and sentence recognition in quiet and in noise, *Audiology and Neurotology* **16**(2): 113-123.

Shi, L. F. and Zaki, N. A. (2014). Psychometric function for NU-6 word recognition in noise: effects of first language and dominant language, *Ear and Hearing* **35**(2): 236-245.

Silva-Pereyra, J., Conboy, B. T., Klarman, L. and Kuhl, P. K. (2007). Grammatical processing without semantics? An event-related brain potential study of preschoolers using jabberwocky sentences, *Journal of Cognitive Neuroscience* **19**(6): 1050-1065.

Small, A. M. Jr. and Campbell, R. A. (1962). Temporal differential sensitivity for auditory stimuli, *The American Journal of Psychology* **75**(3): 401-410.

Smith, M. R., Cutler, A., Butterfield, S. and Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech, *Journal of Speech and Hearing Research* **32**(4): 912-920.

Smith, Z. M., Delgutte, B. and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception, *Nature* **416**(6876): 87-90.

Snedeker, J. and Trueswell, J. (2003). Using prosody to avoid ambiguity: effects of speaker awareness and referential context, *Journal of Memory and Language* **48**(1): 103-130.

Soli, S. D. and Wong, L. L. N. (2008). Assessment of speech intelligibility in noise with the Hearing in Noise Test, *International Journal of Audiology* **47**(6): 356-361.

Spitzer, S., Liss, J., Spahr, T., Dorman, M. and Lansford, K. (2009). The use of fundamental frequency for lexical segmentation in listeners with cochlear implants, *Journal of the Acoustical Society of America* **125**(6): EL236-EL241.

Stacey, P. C., Raine, C. H., O'Donoghue, G. M., Tapper, L., Twomey, T. and Summerfield, A. Q. (2010). Effectiveness of computer-based auditory training for adult users of cochlear implants, *International Journal of Audiology* **49**(5): 347-356.

Statistics South Africa. (2012). *The South Africa I know, the home I understand*, Statistics South Africa, Pretoria.

Statistics South Africa. (2011). *Census 2011 key results*, Statistics South Africa, Pretoria.

Stickney, G. S., Assmann, P. F., Chang, J. and Zeng, F. G. (2007). Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences, *Journal of the Acoustical Society of America* **122**(2): 1069-1078.

Stoops, Y. (1995). *Bobbejane of bavianen: Afrikaans versus Nederlands*, Coda, Mechelen.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America* **26**(2): 212-215.

Summerfield, Q., Sidwell, A. and Nelson, T. (1987). Auditory enhancement of changes in spectral amplitude, *Journal of the Acoustical Society of America* **81**(3): 700-708.

Svirsky, M. A., Silveira, A., Suarez, H., Neuburger, H., Lai, T. T. and Simmons, P. M. (2001). Auditory learning and adaptation after cochlear implantation: a preliminary study of discrimination and labeling of vowel sounds by cochlear implant users, *Acta Oto-Laryngologica* **121**(2): 262-265.

Swanepoel, R., Oosthuizen, D. J. J. and Hanekom, J. J. (2012). The relative importance of spectral cues for vowel recognition in severe noise, *Journal of the Acoustical Society of America* **132**(4): 2652-2662.

Swerts, M. (2007). Contrast and accent in Dutch and Romanian, *Journal of Phonetics* **35**(3): 380-397.

Swerts, M. and Hirschberg, J. (2010). Prosodic predictors of upcoming positive or negative content in spoken messages, *Journal of the Acoustical Society of America* **128**(3): 1337-1345.

Tartter, V. C., Hellman, S. A. and Chute, P. M. (1992). Vowel perception strategies of normal-hearing subjects and patients using Nucleus multichannel and 3M/House cochlear implants, *Journal of the Acoustical Society of America* **92**(3): 1269-1283.

Taylor, J. R. and Uys, J. Z. (1988). Notes on the Afrikaans vowel system, *Leuvense Bijdragen* **77**(2): 129-149.

Theunissen, M., Swanepoel, D. and Hanekom, J. J. (2011). The development of an Afrikaans test of sentence recognition thresholds in noise, *International Journal of Audiology* **50**(2): 77-85.

Thorsen, N. G. (1980). A study of perception of sentence intonation - evidence from Danish, *Journal of the Acoustical Society of America* **67**(3): 1014-1030.

Tomlinson, J. M. and Fox Tree, J. E. (2011). Listeners' comprehension of uptalk in spontaneous speech, *Cognition* **119**(1): 58-69.

Turnbull, O. H., Evans, C. E. Y. and Owen, V. (2005). Negative emotions and anosognosia, *Cortex* **41**(1): 67-75.

Van der Merwe, A., Groenewald, E., Van Aardt, D., Tesner, H. E. C. and Grimbeek, R. J. (1993). Die formantpatrone van Afrikaanse vokale soos geproduseer deur manlike sprekers, *South African Journal of Linguistics* **11**(2): 71-79.

Van Heerden, R. (1999). *Die voorkomsfrekwensie van die spraakklanke van Afrikaans met die oog op fonetiese balansering van oudiologie woordelyste*, B Communication Pathology dissertation, thesis, Department of Communication Pathology, University of Pretoria.

Van Heuven, V. J. and Van Zanten, E. (2005). Speech rate as a secondary prosodic characteristic of polarity questions in three languages, *Speech Communication* **47**(1-2): 87-99.

Van Tassell, D. J., Soli, S. D., Kirby, V. M. and Widin, G. P. (1987). Speech waveform envelope cues for consonant recognition, *Journal of the Acoustical Society of America* **82**(4): 1152-1161.

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

Van Wieringen, A. and Wouters, J. (2008). LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands, *International Journal of Audiology* **47**(6): 348-355.

Van Zyl, M. and Hanekom, J. J. (2011). Speech perception in noise: A comparison between sentence and prosody recognition, *Journal of Hearing Science* **1**(2): 54-56.

Van Zyl, M. and Hanekom, J. J. (2013b). When "okay" is not okay: Acoustic characteristics of single-word prosody conveying reluctance, *Journal of the Acoustical Society of America* **133**(1): EL13-EL19.

Van Zyl, M. and Hanekom, J. J. (2013a). Perception of vowels and prosody by cochlear implant recipients in noise, *Journal of Communication Disorders* **46**(5-6): 449-464.

Vollmer, M. and Beitel, R. E. (2011). Behavioral training restores temporal processing in auditory cortex of long-deaf cats, *Journal of Neurophysiology* **106**(5): 2423-2436.

Vongpaisal, T. and Pichora-Fuller, M. K. (2007). Effect of age on F0 difference limen and concurrent vowel identification, *Journal of Speech, Language, and Hearing Research* **50**(5): 1139-1156.

Waltzman, S. and Hochberg, I. (1990). Perception of speech pattern contrasts using a multichannel cochlear implant, *Ear and Hearing* **11**(1): 50-55.

Watson, D. and Gibson, E. (2005). Intonational phrasing and constituency in language production and comprehension, *Studia Linguistica* **59**(2-3): 279-300.

Watson, P. J. and Schlauch, R. S. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours, *American Journal of Speech-Language Pathology* **17**(4): 348-355.

Wei, C., Cao, K. and Zeng, F. G. (2004). Mandarin tone recognition in cochlear-implant subjects, *Hearing Research* **197**(1-2): 87-95.

Wells, B. and Macfarlane, S. (1998). Prosody as an interactional resource: turn-projection and overlap, *Language and Speech* **41**(3-4): 265-294.

Williams, C. E. and Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates, *Journal of the Acoustical Society of America* **52**(4): 1238-1250.

Wilson, B. S. and Dorman, M. F. (2008). Cochlear implants: Current designs and future possibilities, *Journal of Rehabilitation Research and Development* **45**(5): 695-730.

Wilson, R. H., Carnell, C. S. and Cleghorn, A. L. (2007). The Words-In-Noise (WIN) test with multitalker babble and speech-spectrum noise maskers, *Journal of the American Academy of Audiology* **18**(6): 522-529.

Wilson, R. H. and Cates, W. B. (2008). A comparison of two word-recognition tasks in multitalker babble: Speech Recognition in Noise Test (SPRINT) and Words-in-Noise test (WIN), *Journal of the American Academy of Audiology* **19**(7): 548-556.

Wissing, D. (2007). Basiese akoestiese korrelate van klemtoon in Afrikaans, *Southern African Linguistics and Applied Language Studies* **25**(3): 441-458.

Woods, D. L., Yund, E. W., Herron, T. J. and Ua Cruadhlaoich, M. A. I. (2010). Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise, *Journal of the Acoustical Society of America* **127**(3): 1609-1623.

World Health Organization. (2000). *Guidelines for community noise*, World Health Organization, Geneva.

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.

Yamada, Y. and Neville, H. J. (2007). An ERP study of syntactic processing in English and nonsense sentences, *Brain Research* **1130**(1): 167-180.

Yumoto, E., Gould, W. J. and Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness, *Journal of the Acoustical Society of America* **71**(6): 1544-1550.

Zhao, Z. (2011). Power of tests for comparing trend curves with application to national immunization survey (NIS), *Statistics in Medicine* **30**(5): 531-540.

Zuraidah, M. and Knowles, G. (2006). Prosody and turn-taking in Malay broadcast interviews, *Journal of Pragmatics* **38**(4): 490-512.

APPENDIX A UNCONDITIONAL AND CONDITIONAL SENTENCES FOR EXPERIMENT 1

Sentences recorded for the first listening experiment conveying unconditional or conditional permission, agreement, or approval. English translations of the original Afrikaans sentences are provided in italics.

Table A.1: Sentences recorded to convey unconditional or conditional permission, agreement or approval

1	Jy mag die hond kry.
	You may have the dog.
1a	Jy mag die hond kry, maar nie die kat nie.
	You may have the dog, but not the cat.
1b	Jy mag die hond kry, maar dit gaan jou kos.
	You may have the dog, but it is going to cost you.
2	Ons kan die ketel koop.
	We can buy the kettle.
2a	Ons kan die ketel koop, maar nie die koppies ook nie.
	We can buy the kettle, but not the cups.
2b	Ons kan die ketel koop, maar nie vandag nie.
	We can buy the kettle, but not today.
3	Jy kan die kaartjies koop.
	You can buy the tickets.
3a	Jy kan die kaartjies koop, maar nie die wyn ook nie.
	You can buy the tickets, but not the wine as well.
3b	Jy kan die kaartjies koop, maar net as dit nie te duur is nie.
	You can buy the tickets, but only if they are not too expensive.
4	Jy mag die boek vat.
	You may take the book.
4a	Jy mag die boek vat, maar nie die tas ook nie.
	You may take the book, but not the suitcase as well.
4b	Jy mag die boek vat, maar net as jy dit terugbring.
	You may take the book, but only if you will return it.
5	Ek stem saam met die reel.
	I agree with the rule.
5a	Ek stem saam met die reel, maar nie met die toepassing daarvan nie.
	I agree with the rule, but not with its application.
5b	Ek stem saam met die reel, maar ek hou nie baie daarvan nie.
	I agree with the rule, but I do not like it very much.

6 Ek stem saam met sy stelling.

I agree with his statement.

6a Ek stem saam met sy stelling, maar nie met sy redes nie.

I agree with his statement, but not with his reasons.

6b Ek stem saam met sy stelling, maar ek hou nie baie daarvan nie.

I agree with his statement, but I do not like it very much.

7 Ek glo ook aan oefening.

I also believe in exercise.

7a Ek glo ook aan oefening, maar nie aan diëte nie.

I also believe in exercise, but not in diets.

7b Ek glo ook aan oefening, maar ek doen dit nie graag nie.

I also believe in exercise, but I do not like to do it.

8 Ek glo ook aan spoke.

I also believe in ghosts.

8a Ek glo ook aan spoke, maar nie aan feëtjies nie.

I also believe in ghosts, but not in fairies.

8b Ek glo ook aan spoke, maar ek is nie bang vir hulle nie.

I also believe in ghosts, but I am not afraid of them.

9 Hy hou van die brood.

He likes the bread.

9a Hy hou van die brood, maar nie van die konfyt nie.

He likes the bread, but not the jam.

9b Hy hou van die brood, maar hy wil nie nog hê nie.

He likes the bread, but he does not want any more.

10 Ek hou van die huis.

I like the house.

10a Ek hou van die huis, maar nie van die tuin nie.

I like the house, but not the garden.

10b Ek hou van die huis, maar ek wil dit nie koop nie.

I like the house, but I do not want to buy it.

11 Ek hou van die man.

I like the man.

11a Ek hou van die man, maar nie van die vrou nie.

I like the man, but not the woman.

11b Ek hou van die man, maar ek dink nie hy is bekwaam nie.

I like the man, but I do not think he is competent.

12 Die boek is goed.

The book is good.

12a Die boek is goed, maar die fliek is swak.

The book is good, but the movie is bad.

12b Die boek is goed, maar ek sal dit nie weer lees nie.

The book is good, but I will not read it again.

APPENDIX B TEST FORMS FOR PHONEMICALLY MATCHED SENTENCES

SNR-2 Listener:

Female speaker

Practice lists:

	Lys 8	
2	Sy sny met ‘n mes.	
17	Die klein babatjie slaap.	
49	Die lemoen was nogal soet.	
51	Die gesin het ‘n huis gekoop.	
59	Die tafel het drie pote.	
136	Die meisie het verkoue gekry.	
140	‘n Meisie kom by die deur in.	
142	Piesangs is geel vrugte.	
195	Hulle steek 'n kers op.	
207	Sy kam haar pop se hare.	

	Lys 11	
11	Die seun het die speletjie geken.	
24	‘n Seuntjie hardloop in die pad af.	
27	Hulle staan op hulle knieë.	
65	Die skoonmaker gebruik ‘n besem.	
69	Hy het die brief gaan pos.	
94	Hy het sy geld laat val.	
111	Die bestuurder het verdwaal.	
183	Die kar se ratte maak 'n geraas.	
211	Sy lees 'n dik boek.	
218	Hy praat met sy mond vol kos.	

Test lists:

	Lys 1	
4	My pa sluit die voorhek.	
30	Die vuurhoutjies lê op die rak.	
36	Die hond gee ‘n kwaai grom.	
54	Sy skryf vir haar boetie 'n brief.	
73	Die hoender het eiers gelê.	
90	Die plant staan langs die deur.	
158	Daar was baie min mense.	
159	Die paleis het 'n pragtige tuin.	
171	Ons was gister biblioteek toe.	
208	Hy het kaas en melk gaan koop.	

	Lys 2	
10	Hy het sy vinger gesny.	
53	Hulle het gaan kaas koop.	
72	Die emmers is vol water.	
77	Die twee boere gesels lekker.	
82	Pa het by die hek betaal.	
86	Die seun het 'n rooi karretjie.	
92	Die vragmotor ry teen die bult op.	
98	Die gras word nou lank.	
110	Die polisieman soek ‘n hond.	
197	My pa plant 'n boom.	

	Lys 3	
41	Die vrou het haar huis opgeruim.	
43	Die vrugte lê op die grond.	
61	Hy luister na sy pa.	
102	Daar is oulike mense wat kom.	
105	Die dogters het tafel gedek.	
152	Iemand het die boek by my geleen.	
174	Sy plak 'n seël op die brief.	
175	Die weerlig slaan hard.	
189	Die bank is gister beroof.	
221	Sy het die stoof aan vergeet	
	TOTAL	

SNR-2 Listener:

Male speaker

Practice lists:

	Lys 15	
7	Die seun loop op sy hande.	
95	Hulle het al die eiers gebreek.	
112	Hulle het na die prent gestaar.	
157	Hy klim op tot bo.	
165	Hy kruip agter die bos weg.	
168	Hy trek 'n sirkel om die woord.	
187	Die brood is van graan gemaak.	
191	Die seuntjie vee die stoep.	
199	Ons moet vroeg in die bed klim.	
204	Die hasie sit in sy hok.	

	Lys 21	
31	Hulle hardloop verby die huis.	
47	My pa het die brood vergeet.	
75	Die polisieman ken die pad.	
79	Die voortuin lyk baie mooi.	
87	Hulle het ‘n uur lank gewag.	
104	Die klein babatjie is mooi.	
155	Sy skryf haar naam op die bord.	
160	Die visstok se katrol is stukkend.	
170	Daar is 'n geraamte in die kis.	
182	Die meisie het sproete op haar neus.	

Test lists:

	Lys 4	
23	Die ou man is bekommerd.	
26	Hy het sy boetie gekry.	
33	Hy het by die venster uitgeval.	
66	Sy het in die spieël gekyk.	
101	Die seuntjie hardloop skool toe.	
103	Daar groei blomme in die tuin.	
125	Hulle het die muur geverf.	
156	Die wolke gaan reën bring.	
167	Sy buk om haar tas op te tel.	
177	Die seuns is baie lui.	

	Lys 5	
39	Die lorrie ry in die straat af.	
40	Die slim dogtertjies lees boek.	
58	Sy het naby haar venster gestaan.	
64	Die kar het in 'n muur vasgejaag.	
107	Hy het sy oë toegemaak.	
121	Hulle hou van appelkooskonfyt.	
138	Die peperpot was leeg.	
139	Die hond het uit ‘n bak gedrink.	
153	Sy pak die mandjie vol kos.	
217	Ons moet oor die brug stap.	

	Lys 6	
8	Die roomys was pienk	
20	Die vrou het ‘n trui aangehad.	
42	Die hond het teruggekom.	
70	Die melk staan op die tafel.	
89	Die aarbeikonfyt was soet.	
133	Hy probeer die lepel bykom.	
137	Die twee kinders lag.	
145	My pa het druiwe gepluk.	
149	Ons hond is baie siek.	
186	Die dogtertjie wil 'n ponie hê.	
	TOTAL	

SNR-5 Listener:

Female speaker

Practice lists:

	Lys 1	
4	My pa sluit die voorhek.	
30	Die vuurhoutjies lê op die rak.	
36	Die hond gee ‘n kwaai grom.	
54	Sy skryf vir haar boetie 'n brief.	
73	Die hoender het eiers gelê.	
90	Die plant staan langs die deur.	
158	Daar was baie min mense.	
159	Die paleis het 'n pragtige tuin.	
171	Ons was gister biblioteek toe.	
208	Hy het kaas en melk gaan koop.	

	Lys 2	
10	Hy het sy vinger gesny.	
53	Hulle het gaan kaas koop.	
72	Die emmers is vol water.	
77	Die twee boere gesels lekker.	
82	Pa het by die hek betaal.	
86	Die seun het 'n rooi karretjie.	
92	Die vragmotor ry teen die bult op.	
98	Die gras word nou lank.	
110	Die polisieman soek ‘n hond.	
197	My pa plant 'n boom.	

Test lists:

	Lys 7	
12	Kersfees is in die somer.	
14	Die muis hardloop na sy gat toe.	
37	Iemand het die geld gevat.	
50	Die nuwe pad is op die kaart.	
91	Die seun het swart hare.	
116	Hy drink uit sy beker.	
151	Hy het laat by die huis gekom.	
176	Die konstabel groet vriendelik.	
181	Die polisieman is gewapen.	
215	Sy vurk het op die vloer geval.	

	Lys 10	
18	Die hond het met 'n stok gespeel.	
29	Die kind gryp die speelding.	
34	Die park is naby die pad.	
62	Hulle is weg met vakansie.	
80	Hy het sy hoed verloor.	
93	Die ou vrou was by die huis.	
109	Sy betaal vir die brood.	
114	Die kar ry baie vinnig.	
219	Sy pa het 'n bok gaan skiet.	
222	Die mot vlieg al om die lig.	
	TOTAL	

	Lys 9	
1	Die hanswors het ‘n snaakse gesig.	
13	Die polisie het die kar gejaag.	
16	Daar is ‘n hoop hout onder die boom.	
55	Die speler het die bal verloor.	
71	Die grond was te hard.	
124	Die wolke bring reën.	
126	Die handdoek het op die vloer geval.	
184	Hulle het die vakansie gaan ski.	
188	Daar is 'n swerm bye by die nes.	
203	Hy koop 'n lamp vir sy bedkassie.	

SNR-5 Listener:

Male speaker

Practice lists:

	Lys 4	
23	Die ou man is bekommerd.	
26	Hy het sy boetie gekry.	
33	Hy het by die venster uitgeval.	
66	Sy het in die spieël gekyk.	
101	Die seuntjie hardloop skool toe.	
103	Daar groei blomme in die tuin.	
125	Hulle het die muur geverf.	
156	Die wolke gaan reën bring.	
167	Sy buk om haar tas op te tel.	
177	Die seuns is baie lui.	

	Lys 5	
39	Die lorrie ry in die straat af.	
40	Die slim dogtertjies lees boek.	
58	Sy het naby haar venster gestaan.	
64	Die kar het in 'n muur vasgejaag.	
107	Hy het sy oë toegemaak.	
121	Hulle hou van appelkooskonfyt.	
138	Die peperpot was leeg.	
139	Die hond het uit ‘n bak gedrink.	
153	Sy pak die mandjie vol kos.	
217	Ons moet oor die brug stap.	

Test lists:

	Lys 12	
3	Die huis het nege kamers.	
9	Die leer staan by die deur.	
60	Die vyf mans werk hard.	
67	Hulle het die paadjie gevolg.	
81	Die krane is bokant die wasbak.	
119	Sy bel haar dogter.	
127	Die hond eet 'n stuk vleis.	
130	Suiker is baie soet.	
169	Die bal het hom teen die kop getref.	
196	Sy spring oor die muurtjie.	

	Lys 13	
19	Hulle sê ‘n klomp lawwe goed.	
32	Die trein het ontspoor.	
76	Die seuntjie klim in die bed.	
84	Die wedstryd is verby.	
88	Die groot hond is gevaarlik.	
128	Die reën val op die dak.	
162	Sy was gister by die haarkapper.	
164	Sy het haar elmboog gestamp.	
179	Ek en my pa speel skaak.	
180	Die vrou dra baie juwele.	

	Lys 14	
6	Die sak sleep op die grond.	
46	Die bal het gehop.	
78	Ma het blomme gepluk.	
115	Die verwer het 'n kwas gebruik.	
131	Die skoonmaker vee die vloer.	
178	Die hond vee sy snoet aan my af.	
185	Die vrou is deftig aangetrek.	
190	Sy streel haar pop se hare.	
192	Die hondjie se pels blink mooi.	
193	Hy is uitgeput na die wedstryd.	
	TOTAL	

SNR-8 Listener:

Female speaker

Practice lists:

	Lys 7	
12	Kersfees is in die somer.	
14	Die muis hardloop na sy gat toe.	
37	Iemand het die geld gevat.	
50	Die nuwe pad is op die kaart.	
91	Die seun het swart hare.	
116	Hy drink uit sy beker.	
151	Hy het laat by die huis gekom.	
176	Die konstabel groet vriendelik.	
181	Die polisieman is gewapen.	
215	Sy vurk het op die vloer geval.	

	Lys 9	
1	Die hanswors het 'n snaakse gesig.	
13	Die polisie het die kar gejaag.	
16	Daar is 'n hoop hout onder die boom.	
55	Die speler het die bal verloor.	
71	Die grond was te hard.	
124	Die wolke bring reën.	
126	Die handdoek het op die vloer geval.	
184	Hulle het die vakansie gaan ski.	
188	Daar is 'n swerm bye by die nes.	
203	Hy koop 'n lamp vir sy bedkassie.	

Test lists:

	Lys 16	
25	Die huis het 'n mooi tuin.	
96	Sy help haar maatjie.	
99	Die vuur was baie warm.	
129	Die gesin eet graag vis.	
141	Die pad loop teen die bult op.	
147	Hulle het geld verloor.	
148	Sy skep dit met 'n lepel.	
161	Hy blaas die stof van sy kas af.	
201	Die stoel staan in die hoek.	
206	Die vrou kom by die winkel uit.	

	Lys 18	
38	My pa kom huis toe.	
44	Die bus het vroeg gery.	
45	Hulle het twee leë bottels.	
74	Die bestuurder wag op die hoek.	
85	Sy dra 'n klomp inkopiesakkies.	
117	Hulle het aan die venster geklop.	
118	Die skêr is nogal skerp.	
122	Sy ma het die venster toegemaak.	
172	Die blaartjie dryf in die stroom af.	
214	My boonste knoop het afgeval.	
	TOTAL	

	Lys 17	
22	Die vrou het haar man gehelp.	
35	Die kok het uie gesny.	
63	Die trein beweeg vinnig.	
97	Die bordjie wys die pad aan.	
113	Die oond se deur was oop.	
194	Sy pluk 'n rooi roos.	
198	n By het my sussie gesteek.	
202	Die hond jaag die kat.	
205	Hy maak die boot met 'n tou vas.	
216	Hulle speel buite met die bal.	

SNR-8 Listener:

Male speaker

Practice lists:

	Lys 12	
3	Die huis het nege kamers.	
9	Die leer staan by die deur.	
60	Die vyf mans werk hard.	
67	Hulle het die paadjie gevolg.	
81	Die krane is bokant die wasbak.	
119	Sy bel haar dogter.	
127	Die hond eet 'n stuk vleis.	
130	Suiker is baie soet.	
169	Die bal het hom teen die kop getref.	
196	Sy spring oor die muurtjie.	

	Lys 13	
19	Hulle sê ‘n klomp lawwe goed.	
32	Die trein het ontspoor.	
76	Die seuntjie klim in die bed.	
84	Die wedstryd is verby.	
88	Die groot hond is gevaarlik.	
128	Die reën val op die dak.	
162	Sy was gister by die haarkapper.	
164	Sy het haar elmboog gestamp.	
179	Ek en my pa speel skaak.	
180	Die vrou dra baie juwele.	

Test lists:

	Lys 19	
5	Hulle kyk na die horlosie.	
15	Die vrou maak 'n speelding.	
52	Die beker staan op die rak.	
57	Die boek vertel ‘n storie.	
120	Die hond het die kat gejaag.	
123	Hy speel buite saam met sy maatjie.	
135	Die vadoek is nogal nat.	
144	Hy het sy sussie bang gemaak.	
209	Ons eet pap en wors vanaand.	
220	Die mes is vol botter.	

	Lys 20	
48	Die meisie het ‘n inkleurboek.	
100	Hy suig nog sy duim.	
106	Hulle het oor die gras geloop.	
132	Die badwater was warm.	
134	Hy het sy rekening betaal.	
150	Hy het 'n fiets geleen.	
163	Die kinders groet die juffrou.	
200	Ons soek die pad op die kaart.	
210	Die bome se blare val af.	
321	Die kamer word nou koud.	

	Lys 22	
28	Die meisie het haar pop verloor.	
56	Die meisies luister musiek.	
68	Die hond spring op die stoel.	
83	Ons het gaan brood koop.	
108	Hulle het die ambulans gebel.	
146	Die ketel het vinnig gekook.	
154	Daar is 'n mier op sy voet.	
166	Hulle gaan na die wedstryd kyk.	
212	Die vlag wapper in die wind.	
213	Hy het sy been gebreek.	
	TOTAL	

APPENDIX C INDIVIDUAL DATA FOR CI LISTENERS (EXPERIMENT 2)

The following graphs illustrate the data for individual CI recipients on the listening tasks of Experiment 2 (Chapter 4).

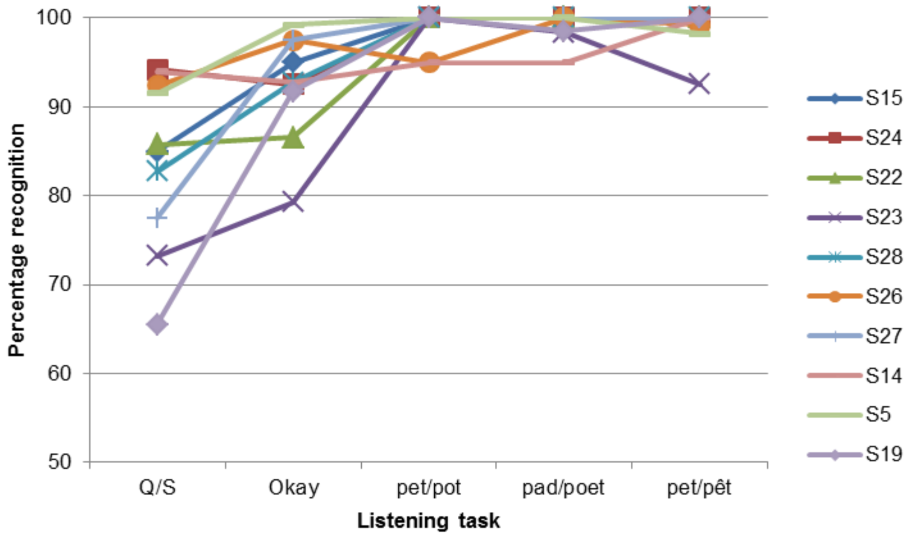

Figure C.1: Recognition scores of individual CI users on the different listening tasks as measured in quiet

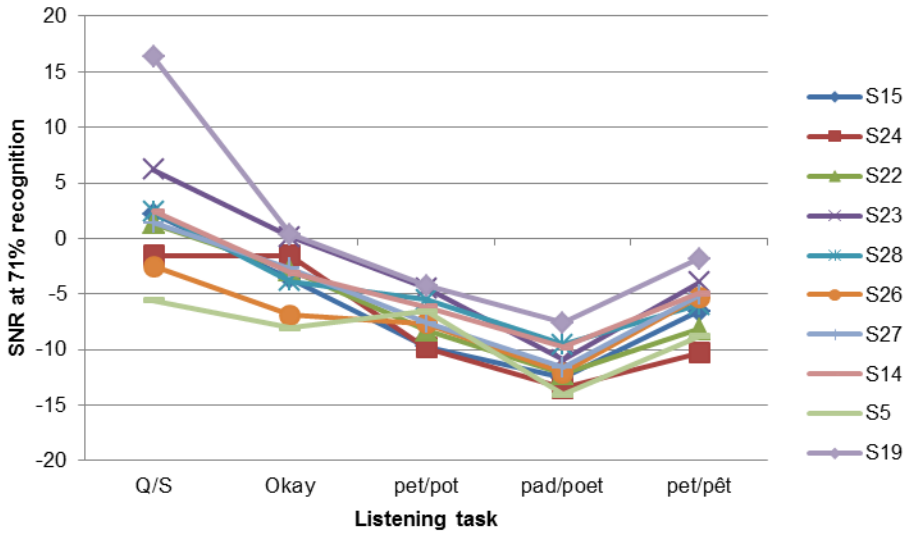

Figure C.2: SNR levels at 71% recognition as measured from individual CI users on the different listening tasks in an adaptive noise procedure

APPENDIX D PHONETIC TRANSCRIPTION OF CVC STIMULI AND GUI ALTERNATIVES

Table D.1: CVC-combinations used for vowel testing

Target phoneme	**Target word**	Alternative 1	Alternative 2	Alternative 3
ə	**sət**	sat	sut	sɔt
	bət	bɛt	bet	bot
	vəx	væx	wɑx	vix
a	**pat**	pɔt	pɛt	pot
	vax	vɑx	vex	væx
	bat	but	bət	bit
i	**kis**	kɑs	kəs	kus
	vil	væl	val	vɔl
	bit	bɛt	bot	bet
ɛ	**mɛs**	mas	mɔs	mus
	rɛt	rɑt	rət	rit
	bɛt	bet	bot	bɑt
ɔ	**rɔk**	ræk	rɑk	rak
	lɔs	lɛs	les	lis
	bɔt	bət	bot	but
ɑ	**tɑl**	tal	təl	tæl
	fɑx	fɔx	fox	fux
	lɑs	lɛs	les	lis
e	**mes**	mɛs	mɔs	mus
	let	lat	lit	lot
	vex	vəx	væx	vɑx
æ	**ræk**	rɔk	rok	rɑk
	væx	vax	vix	vex
	ɦæk	ɦuk	ɦək	ɦɔk
o	**rok**	rɑk	rak	ræk
	pot	pɔt	pət	pɛt
	bot	bit	bet	but
u	**sun**	sen	sən	sin
	rus	ros	rɑs	rɛs
	buk	bæk	bɔk	bak

Table D.2: CVC-combinations used for consonant testing

Target phoneme	**Target word**	Alternative 1	Alternative 2	Alternative 3
	tak	pak	bak	ɦak
t	**tɑk**	sɑk	rɑk	vɑk
	tɔl	dɔl	fɔl	mɔl
	duk	buk	kuk	ɦuk
d	**del**	rel	mel	xel
	dak	tak	sak	fak
	bak	dak	tak	fak
b	**bɑn**	mɑn	lɑn	xɑn
	bɛs	pɛs	rɛs	sɛs
	pək	tək	dək	ɦək
p	**pul**	kul	ful	vul
	pak	bak	mak	rak
	kɛn	pɛn	dɛn	ɦɛn
k	**kas**	tas	jas	las
	kɔm	xɔm	sɔm	ɦɔm
	sak	rak	ɦak	bak
s	**sɔp**	tɔp	kɔp	mɔp
	sɑx	fɑx	lɑx	jɑx
	xor	for	ɦor	bor
x	**xɔm**	sɔm	kɔm	dɔm
	xas	kas	jas	las
	fak	sak	rak	bak
f	**fel**	xel	kel	del
	fɛt	vɛt	pɛt	mɛt
	vas	ras	xas	tas
v	**vɑr**	ɦɑr	dɑr	pɑr
	vɔl	fɔl	mɔl	kɔl
	mɛs	nɛs	rɛs	sɛs
m	**mɔs**	bɔs	pɔs	kɔs
	mal	val	dal	fal
	nat	mat	vat	fat
n	**nɛs**	rɛs	bɛs	pɛs
	nis	lis	sis	kis
	ɦir	vir	dir	tir
ɦ	**ɦɑs**	rɑs	bɑs	kɑs
	ɦak	sak	dak	pak
	jas	las	das	pas
j	**jɑx**	mɑx	vɑx	fɑx
	jol	ɦol	kol	sol
	rɑs	ɦɑs	xɑs	kɑs
r	**rɛt**	nɛt	bɛt	pɛt
	rɔt	sɔt	bɔt	mɔt
	lɔk	jɔk	ɦɔk	kɔk
l	**lu:r**	ru:r	bu:r	fu:r
	lat	nat	sat	xat

APPENDIX E JABBERWOCKY SENTENCES FOR EMOTIONAL PROSODY TESTING

Table E.1: Jabberwocky sentences

1 Hy het die krawe geliep.
2 Sy gen die mole waksel.
3 Hy wou 'n bligter gol riep.
4 Sy wippel die rane foop.
5 Sy sal lieke of hag plo.
6 Hy is naster met skalpe.
7 Sy dif 'n jabbel met gik.
8 Hy wil donkel by die klaf.
9 Sy moet die sloegte bewap.
10 Sy het 'n nefte rakel.
11 Hy kal die troke soewe.
12 Hy was tiffel oor die nos.
13 Sy sou lare gekim het.
14 Hy drabel die giewe mef.
15 Sy het die fille gewom
16 Hy wou kalle of roog blo.

All sixteen sentences were recorded from both speakers. Following the validation procedure, sentences 1 and 6 were excluded from the female speaker's collection, and sentences 1, 2 and 15 were excluded from the male speaker's collection, due to poor scores obtained in quiet.

APPENDIX F DISTINCTIVE FEATURES OF CONSONANTS

Table F.1: Classification of consonants according to distinctive features

Place of articulation				
Bilabial	Labiodental	Alveolar	Velar	Glottal
p, b, m	f, v	t, d, n, l, s, r	k, x, j*	ɦ

Manner of articulation			
Stop	Fricative	Semi-vowel	Nasal
p, t, k, b, d	f, x, s, r**, ɦ, v	l, j	m, n

Voicing	
Voiced	Voiceless
b, d, l, n, m, r, ɦ, j, v	p, t, s, f, k, x

* /j/ is produced mid-palatal, but since it is the only Afrikaans consonant with this place of articulation, it was grouped with the velar consonants (the closest place of articulation to mid-palatal).

** /r/ is a voiced alveolar trill, but since it is the only Afrikaans consonant with this manner of articulation, it was grouped with the fricatives (the closest manner of articulation to a trill)

Table F.2: Differences between each target consonant and all other consonants according to distinctive features

	Difference with other consonants						
Target consonant	Place only	Manner only	Voice only	Place & voice	Place & manner	Manner & voice	Place, manner & voice
t	p,k	s	d	b	f,x	r,l,n	m,j, ɦ,v
d	b	n,l,r	t	p,k	v,m,j, ɦ	s	f,x
b	d	m	p	t,k	n,l,r,j, ɦ,v	-	s,x,f
p	t,k		b	d	s,x,f	m	n,l,r,j, ɦ,v
k	p,t	x	g*	b,d	s,f	j	n,l,r, ɦ,v
s	x,f	t	r	v, ɦ	p,k	d,l,n,	b,m,j
x	f,s	k	-	v, ɦ,r	p,k	j	b,m,d,n,l
f	s,x	-	v	r, ɦ	p,t,k	-	b,m,d,n,l,j
v	r,h	-	f	s,x	b,d,l,m,n,j	-	p,t,k
m	n	b	-	-	d,l,r, ɦ,j,v	p	t,s,f,k,x
n	m	d,l,r	-	-	b, ɦ,j,v	t,s	p,f,k,x
ɦ	v,r	-	-	f,x,s	b,d,l,m,n,j	-	p,t,k
j	l	-	-	-	b,d,n,m,r, ɦ,v	k,x	p,t,s,f
r	v,h	d,n,l	s	f,x	b,j,m	t	p,k
l	j	d,n,r	-	-	b,m, ɦ,v	t,s	p,f,k,x

APPENDIX G NOISE EFFECTS ON PROSODY

The following figures illustrate waveforms and spectrograms of a jabberwocky sentence (“*Hy is naster met skalpe*”) depicting happiness, as recorded from the female speaker. Blue lines on the spectrograms indicate estimated voice pitch (F0), while red lines indicate estimated formant frequencies.

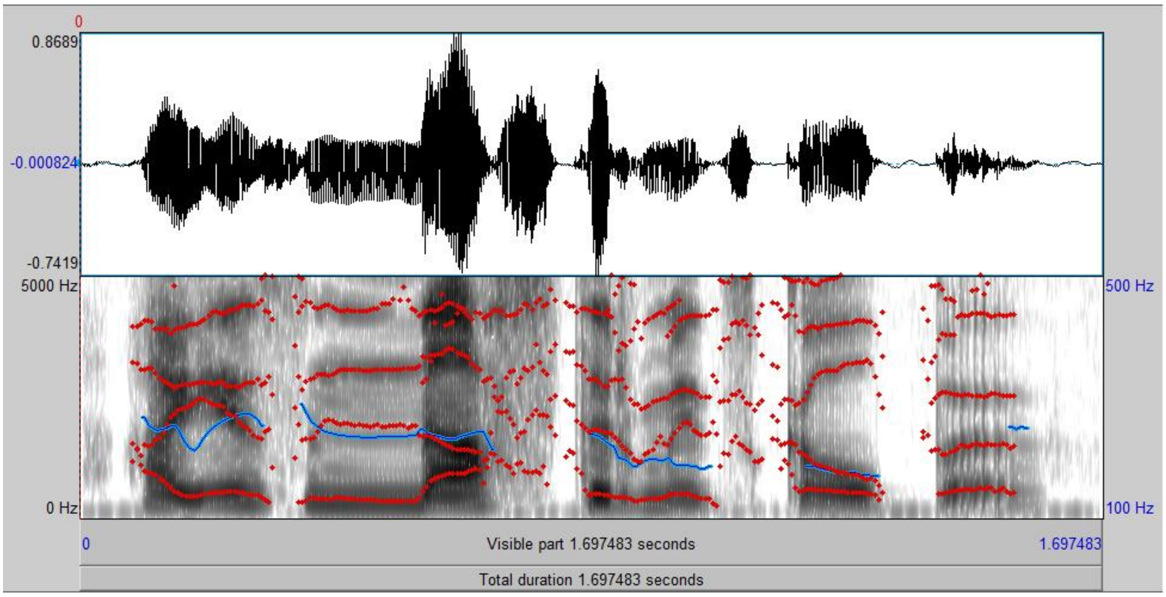

Figure G.1: Original waveform and spectrogram of the sentence as recorded in quiet.

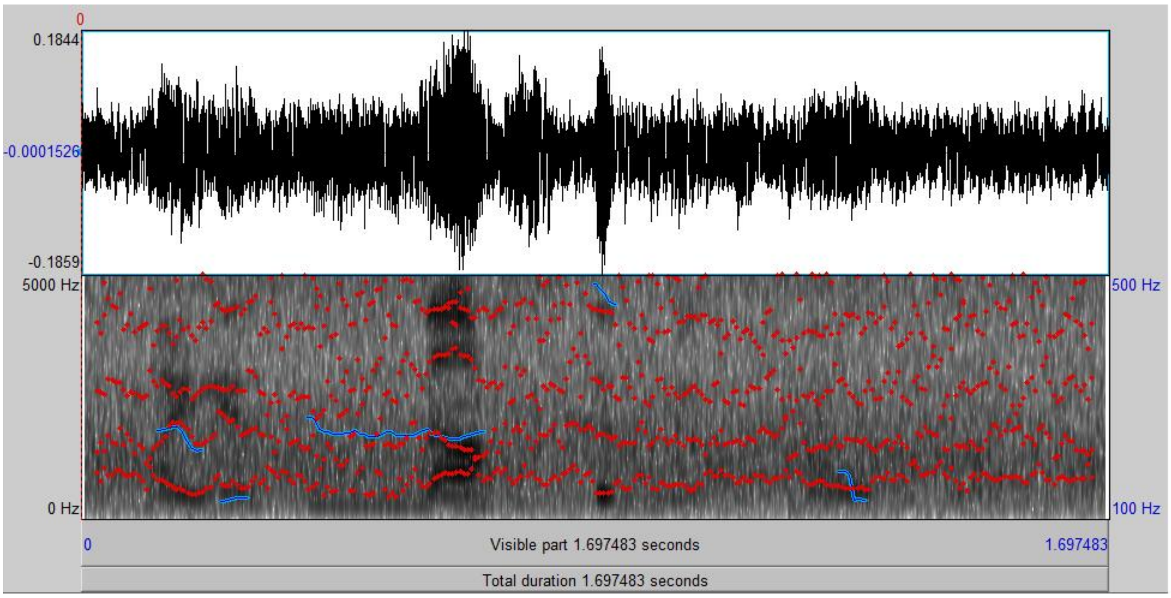

Figure G.2: Waveform and spectrogram of the sentence with added speech-weighted noise (at -0 dB SNR)

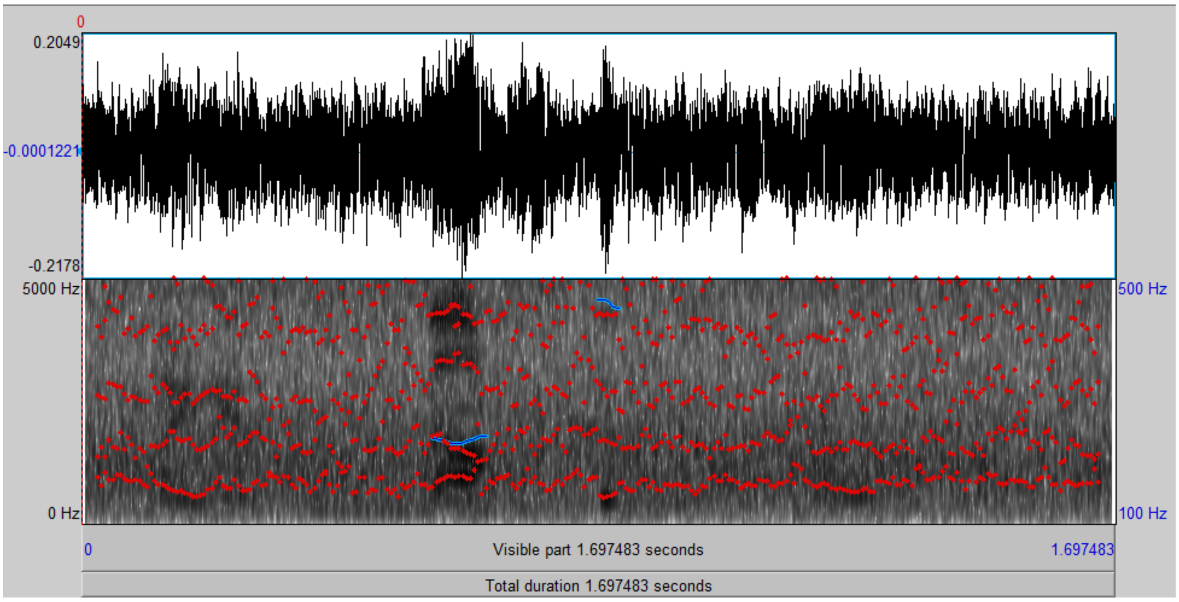
0
0.2049
-0.0001221
-0.2178
5000 Hz
0 Hz
500 Hz
100 Hz
0
Visible part 1.697483 seconds
1.697483
Total duration 1.697483 seconds

Figure G.3: Waveform and spectrogram of the sentence with added speech-weighted noise (at -0 dB SNR)