UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Obstruent acoustic landmark enhancement for cochlear implants

By

Cherith Mercedes Webb

Thesis for the degree of Doctor of Philosophy

February 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Doctor of Philosophy

OBSTRUENT ACOUSTIC LANDMARK ENHANCEMENT FOR COCHLEAR IMPLANTS

By Cherith Mercedes Webb

Cochlear implant users are typically able to achieve high levels of speech recognition in quiet but struggle to follow speech even in moderate levels of background noise. It may be possible to improve cochlear implant users' speech perception scores in noise, by making more efficient use of the limited bandwidth available for transmitting important speech information, rather than increasing the amount of information transmitted. Acoustic landmarks are locations in the speech signal which are rich in information and allow a listener to identify a particular speech sound as a vowel, sonorant consonant or obstruent consonant; it is around these regions that important speech cues tend to be concentrated. Obstruent consonants are signalled by sudden amplitude and spectral changes and the onset/offset of a period of noise. It has been shown that the auditory system is particularly responsive to rapid spectral changes, manifested as increased firing rates of auditory nerve fibres, particularly at onsets of signals. Cochlear implant users commonly confuse speech sounds with rapidly changing spectral patterns, possibly due to the poor transmission of obstruent landmark information.

The aim of the present work was to develop an obstruent landmark enhancement strategy which could be integrated into current cochlear implant processing. The first stage of this process required the identification of obstruent landmarks from the noise-mixed speech stimuli. An existing automatic landmark detection algorithm did not achieve the high levels of accuracy required for use in the present study and so a set of hand-generated labels were created, and used to guide the proposed obstruent landmark enhancement strategy. A series of cochlear implant simulation experiments were conducted to help evaluate the strategy and guide further developments. Results from the simulation studies suggest that the proposed method of obstruent landmark enhancement does not help to improve speech recognition in noise for normal hearing listeners listening to a cochlear implant simulation. It is likely that the strategy outlined in this thesis did not help to improve the saliency of obstruent landmark events as the enhancement was applied to the noise as well as the target speech signal, making it difficult for listeners to resolve the boosted landmark information. However, the results also highlight the limitations of using cochlear implant simulations to evaluate the strategy and so the findings are not necessarily a predictor of actual cochlear implant user performance.

List of Contents

List of Figures

List of Tables

Declaration of Authorship

I, **CHERITH MERCEDES WEBB** declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

OBSTRUENT LANDMARK ENHANCEMENT FOR COCHLEAR IMPLANTS

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. None of this work has been published before submission

Signed:

Date: 18/05/15

Acknowledgements

Firstly I would like to thank my supervisor, Carl Verschuur, for allowing me the opportunity to follow my passion and expand my thinking surrounding the area of speech perception. His guidance and encouragement has helped shape me into the researcher I am today.

I must also thank my co-supervisor, Anna Barney, for stepping in at the last minute and for giving me valuable feedback and suggestions at various stages of my work. Thanks also goes to the rest of the staff in the Hearing and Balance Centre for sharing their knowledge and inspiring me to enter the fascinating world of audiological research – I have learnt so much.

A special mention goes to Amit Juneja, who, despite the great distance and the new arrivals gave up his spare time to help develop the automatic landmark detectors used in this thesis. A massive thank you also goes to Falk-Martin Hoffmann- you have been my signal processing and MATLAB guru, and I really cannot thank you enough for the help you have given me over the last few years.

Thank you to Action on Hearing Loss and Rosetrees Trust for your financial support and continued interest in my research.

To my friends and family- this thesis would not have been completed if it were not for your continual love and support and the occasional motivational hug, picture, text and phone call.

And finally to my fiancé, Damien, quite simply- I would not have made it without you. I look forward to becoming one half of "The Drs Campbell-Bell".

List of Abbreviations

ACE- Advanced Combination Encoder
ADRO- Automatic Dynamic Range Optimisation
AGC- Automatic gain control
AL- acoustic landmark
ALD- acoustic landmark detection/detector
ANF- auditory nerve fibre
ANOVA- analysis of variance
ASR- automatic speech recognition
BKB- Bamford-Kowel-Bench
BM- basilar membrane
C- level- comfort level
C-V- consonant-vowel
CF- characteristic frequency
CI- cochlear implant
CIS- continuous interleaved sampling
CNC- consonant-nucleus-consonant
COG- centre of gravity
CSE- cochlea -scaled entropy
CUNY- City University of New York
dB- decibel
DR- dynamic range
EAS- electro-acoustic stimulation
EBS- event-based system
EE- envelope enhancement
EECIS- envelope enhancement continuous interleaved sampling
***f*0**- fundamental frequency
F1/2/3- first/second/ third formant
FFT- fast fourier transform
GMM-Gaussian Mixture Model
GUI-graphical user interface
HI- hearing impaired
HINT- hearing in noise test
HL-hearing level
HMM- Hidden Markov Model
HTK- Hidden Markov Model toolkit
Hz- Hertz
IBM- Ideal Binary Mask
ICA- independent component analysis
IDR- input dynamic range
IHR- Institute for Hearing Research
ILD- interaural level difference
ITD- interaural time difference
JND- just noticeable difference
LAFF- Lexical Access from Features
LBP- landmark boost package
LLI- language learning-impaired
LND- landmark detection toolkit
MANOVA- multivariate analysis of variance
MDT- modulation detection threshold
MFCC- mel frequency cepstral coefficient
MI- Massachusetts Institute of Technology
NH- normal hearing
NICE- National Institute for Clinical Excellence
NMT- Nucleus Matlab Toolbox
n-*of*-m- number-of-maxima
PCA- principle component analysis
PER- potential enhancement regions
pps- pulses per second
PSTH- post-stimulus time histogram
QSS- quasi-steady-state
RAU- rationalised arcsine transform
RMS- root-mean-square
SINFA- sequential information transfer
SMSP- spectral maxima sound processor
SNR- signal-to-noise ratio
SPEAK- Spectral Peak
SPL- sound pressure level
SRT- speech reception threshold
SS- speech-shaped
SVM- seport vector machine
T-level- threshold level
TESM- Transient Emphasis Spectral Maxima
T-F- time-frequency
TTN- two-talker noise
V-C- vowel-consonant
VCV- vowel-consonant-vowel
VOT- voice onset time

Chapter 1- Introduction

1.1 Background

Speech is an efficient and robust information carrier and is the primary means of communication amongst humans, and although humans are able to produce and perceive speech with very little effort, the processes involved in the transmission and understanding of speech are extremely complex. Although much of the anatomy and physiology of the speech chain is now understood, the processes involved in the perception and organisation of speech in the brain are still being explored. Understanding these processes is further complicated when considering listening in adverse conditions, and/or hearing and speech processing disorders, where communication can, at times, be severely impaired. Understanding which elements of the speech signal are important for good speech reception in different listening situations, and knowing the best method for optimising their transmission, has pronounced implications for the development of technology in areas such as telecommunications, automatic speech recognition (ASR), speech synthesis and hearing prostheses. In particular, a better understanding of processing and perception of the speech signal by a human listener may help in the development of new cochlear implant (CI) systems/strategies which bring CI user speech recognition more in line with that of a normal hearing (NH) listener.

A CI is an auditory prosthesis that can provide profoundly deaf patients with a sensation of hearing by bypassing the auditory processing of the outer, middle and inner ear and directly stimulating auditory nerve cells. There are currently around 11,000 CI users in the United Kingdom (www.earfoundation.org.uk) and approximately 324,000 users worldwide (www.nidcd.nih.gov). Due to the nature of electrical stimulation, CIs are limited in their ability to reproduce the coding of speech frequencies and intensities, resulting in a bottleneck in the flow of information from the implant to the auditory system. Poor spectral resolution and a severely limited dynamic range are two of the main contributing factors to this so-called electro-neural bottleneck.

Despite these limitations, advancements in CI technology have seen marked improvements in speech recognition for CI users, with average scores of 70-80 percent correct for sentences in quiet (Stickney et al., 2004). However, even the best performing CI users do not perform as well in background noise as NH listeners, requiring much higher signal-to-noise ratios (SNRs) to obtain comparable speech recognition scores in noise to NH listeners (Dorman et al.,

1998a; 1998b) and even to NH listeners listening to vocoded speech (Friesen et al., 2001). To understand at least 50 percent of speech in noise, CI users typically need an SNR which is 10-15 dB higher than for NH listeners when listening in speech-shaped (SS) noise and as much as 30 dB in fluctuating maskers (Nelson et al., 2003; Zeng et al., 2005). A large degree of variability in speech perception scores is also often observed between CI users (Fu and Nogaki, 2004). Implant signal processing and device characteristics impose a performance ceiling (Skinner et al., 2002a; Dorman et al., 2002) and methods for optimising implants for individual differences are still poorly understood (Wilson, 2004).

1.2 Enhancing acoustic landmarks for cochlear implant users

Methods for improving the intelligibility of speech in the presence of noise can often be categorised as either noise reduction strategies or strategies which deliberately manipulate or distort the speech signal. Noise reduction techniques include methods such as spectral subtraction, adaptive filtering, adaptive noise cancellation and the use of dual/multi microphones. However, these methods often only lead to improvements in speech quality and provide little or no benefit to speech intelligibility (Moore, 2003a; Levitt, 2001; Wang et al., 2009). A number of studies have investigated the use of speech enhancement techniques which attempt to manipulate parts of the speech signal believed to be perceptually important, in the hope of increasing their saliency (Vandali, 2001; Geurts and Wouters, 1999; Koning and Wouters, 2012). It is possible that CI users may gain significant improvements in speech perception with their current implant devices if the speech processing strategy focused more on transmitting perceptually important elements of the speech signal.

Evidence from neurophysiological studies has shown that the auditory system is particularly responsive to rapid spectral and amplitude changes (Delgutte, 1997), and it has been suggested that this is because these regions of change contain the most important information for speech perception (Stilp and Kleunder, 2010). These regions of change are often referred to as acoustic landmarks (ALs) and allow a listener to distinguish between vowel, glide and consonant segments, and it is around these regions where cues important for speech sound identity tend to be concentrated. This thesis makes a distinction between the high-frequency landmarks associated with the obstruent consonants (sudden onsets and offsets of a segment of aperiodic noise) and the low-frequency landmarks associated with vowels and sonorants (slowly varying, periodic segments). Hazan and Simpson (1998; 2000) showed that enhancement of speech cues related to ALs can help to improve speech perception in noise

for NH listeners. A series of studies were conducted by Li and Loizou (2008a; 2009; 2010) which explored the importance of obstruent ALs for speech perception in noise for both NH listeners and CI users. Their focus on obstruent landmarks was motivated by the fact that CI users most commonly misperceive sounds with rapidly changing spectral patterns, such as the obstruent consonants (Munson and Nelson, 2005).

Findings from these studies suggest that speech perception in noise for both NH listeners and CI users can be improved if they are given clear access to obstruent ALs. They also highlight that obstruent landmark information may be poorly transmitted with current speech processing strategies, particularly due to envelope compression which reduces vowel-consonant ratios. Li and Loizou (2010) investigated the use of selective compression, which was more compressive during vowel segments and less compressive during obstruent consonants, as a method for increasing vowel-consonant ratios, with the aim of making times of obstruent landmarks more distinct. Although benefits in speech perception were seen for CI users with selective compression it is not clear whether this method specifically enhanced the transmission of obstruent landmarks.

1.3 Contribution to knowledge

The aim of the present study is to investigate the importance of ALs for CI users and explore methods for improving their transmission with current CI processing. In particular, the study focuses on obstruent landmarks, as the obstruent consonants are affected more by noise corruption than vowels and sonorants, and they are more commonly misperceived by CI users. The improvement of obstruent landmark transmission with CI processing requires both the development of an enhancement strategy and a method for automatically detecting these landmarks from the signal. The development of ASR systems which incorporate ALs into their front-end processing has shown that it is possible to automatically detect ALs with quite a high degree of accuracy (see Juneja and Espy-Wilson, 2004; Juneja and Espy-Wilson, 2008). This thesis explores whether an existing automatic landmark detection algorithm can be adopted for CI processing and used to accurately identify and label obstruent landmarks from within the speech signal (both in quiet and in noise). Hand-generated landmark labels have also been created for this study, the criteria for which were developed by the author based on evidence from the literature and through spectrographic observations.

Alongside this, a new method for enhancing obstruent landmarks has also been developed. The obstruent landmark enhancement strategy applies a boost to the speech signal at times

where obstruent landmarks are identified (using either automatically detected landmark labels or manually generated labels). The temporal and spectral characteristics of the boost function applied to the signal depend on the landmark type, for example, plosive or fricative.

Speech recognition scores for NH listeners listening to simulated CI speech in noise were recorded in order to assess the effect of the landmark enhancement strategy on speech perception and to guide optimisation of the boost functions and labelling. In addition, evidence from spectrographic observations and knowledge about neural representations of sudden onset signals (relating to obstruent consonants) led to the development of a new, clearer definition of 'obstruent landmarks'; this also helped guide improvements to the landmark enhancement strategy. Ultimately, the experiments outlined in this thesis looked to determine whether obstruent landmark enhancement could be beneficial for CI users when listening in noise. Although not investigated in this thesis, other applications for a landmark enhancement technique have already been demonstrated by Hazan and Simpson (1998; 2000), and also have the potential to be adapted for other amplification devices, such as hearing aids.

1.4 Research Questions

This thesis intends to answer the following questions (main questions highlighted in bold):

1. What effect does current CI processing have on the transmission of ALs?
2. **What is the most appropriate method for enhancing obstruent ALs with CI processing?**
3. **Can AL enhancement help to improve speech recognition in noise for CI-processed speech?**
4. Can obstruent ALs be accurately detected (automatically) in quiet and in noise?

1.5 Aims and objectives

1.5.1 Aims

This thesis has the following aims:

1. Develop a clear definition of acoustic landmarks
2. Investigate the role of ALs in speech perception, particularly in noise
3. Understand the effect current CI processing may have on AL transmission

4. Develop and evaluate a method of automatically detecting obstruent ALs that can be used with CI processing

5. Develop and evaluate a method of enhancing obstruent ALs

The first three aims will be achieved by critically evaluating relevant literature.

1.5.2 Objectives

This thesis plans to meet the following objectives (main objectives highlighted in bold):

- **To develop a method of automatically detecting obstruent ALs which is compatible with CI processing**
- **To develop an appropriate method to emphasise obstruent landmarks**
- **Compare speech recognition scores for NH listeners (listening to a CI simulation) and, if appropriate, CI users with and without landmark enhancement**
- Compare the transmission of ALs with the new AL enhancement algorithm and current speech processing strategies

1.6 Thesis outline

The present chapter has introduced the topic of the thesis and summarised its main aims and objectives.

Chapter 2 covers the basics of CI systems and electrical stimulation, and reviews approaches to speech processing with CIs. This chapter also discusses the limitations of CI processing with regards to speech perception and looks at new methods for improving speech intelligibility in noise.

Chapter 3 introduces ALs and discusses the importance of spectral and amplitude change for speech perception in quiet and in noise, drawing on evidence from neural responses to speech, and models of speech perception that incorporate spectral change into the initial stages of speech processing.

Chapter 4 explores the evidence that CI users may benefit from better access to ALs and discusses methods for improving the transmission of landmark information with CI

processing. This includes the development and evaluation of channel-specific landmark detectors which could be used to guide channel selection in a CI speech processing strategy.

Chapter 5 outlines the development of a strategy to selectively enhance obstruent landmarks; first by using automatically detected landmark labels and then by manually created landmark labels. This chapter also reports results from experiments comparing speech perception scores of NH listeners listening to CI simulated speech, with and without landmark enhancement, at the various stages of development.

Chapter 6 discusses the results of these experiments in more detail and considers their wider implications. In addition, this chapter summarises the outcomes of this thesis and considers the next steps to be taken in continuing the investigation into whether obstruent landmark enhancement can be used to improve speech perception for CI users.

Chapter 2- Speech perception by cochlear implant users

2.1 Statement of the problem: speech perception in noise

Speech perception abilities of CI users have significantly improved since the introduction of the first single channel devices in the 1970s. However, despite these advances, CI users typically require an SNR which is at least 10-15 dB higher than for NH listeners, when listening to speech in background noise, to obtain a speech recognition score of 50 percent correct (Nelson et al., 2003; Zeng et al., 2005). The following chapter gives an overview of speech processing in CIs and explores possible reasons why CI users perform worse in background noise than both NH listeners and NH listeners listening to vocoded speech. This chapter also discusses possible methods for improving speech perception in noise for CI users.

2.2 Overview of cochlear implants

The loss and/or absence of hair and nerve cells in the cochlea disrupts the transmission of information about incoming sound to the auditory nerve, resulting in higher auditory thresholds. A CI bypasses the auditory processing of the outer, middle and inner ear, directly stimulating spiral ganglion cells connected to the auditory nerve. In the United Kingdom a patient may be considered for cochlear implantation if they have a bilateral, severe-profound hearing loss (thresholds >90 dB HL at 2-4 kHz) and are not receiving adequate benefit from acoustic hearing aids; defined as a Banford-Kowel-Bench (BKB) (Bench et al., 1979) sentence test score of <50 percent at 70 dB SPL (NICE, 2009).

Figure 2.1 shows the external and internal components of a CI. The role of the external unit is to detect, code and transmit the incoming speech signal to the internal unit; it is comprised of a microphone, speech processor and transmitter coil. The internal receiver decodes the signal and converts it into electrical currents which are delivered to the electrode array. The electrode array is inserted through the round window of the cochlea into the scala tympani, as close to the spiral ganglion as possible (Wilson, 2004); allowing for direct stimulation of surviving neurons in the auditory nerve. Figure 2.2 shows the position of the electrode array through a cross-section of the cochlea.

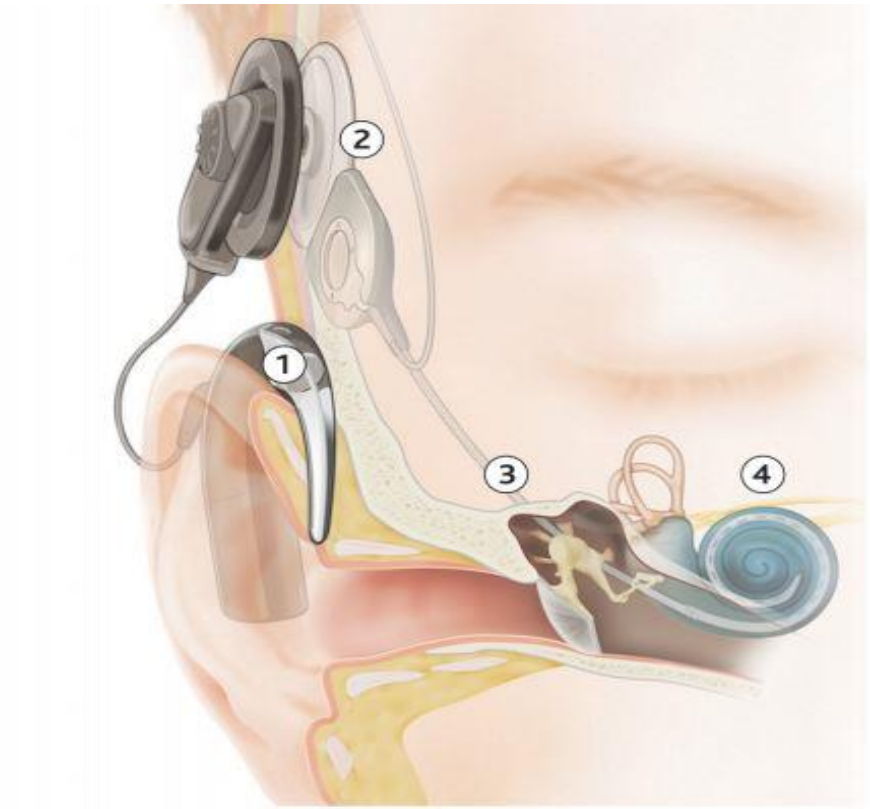


Figure 2. 1 The internal and external components of a cochlear implant. (1) speech processor, (2) external headpiece and internal coil, (3) electrode array and (4) auditory nerve. Courtesy of Cochlear Limited (www.cochlear.co.uk).

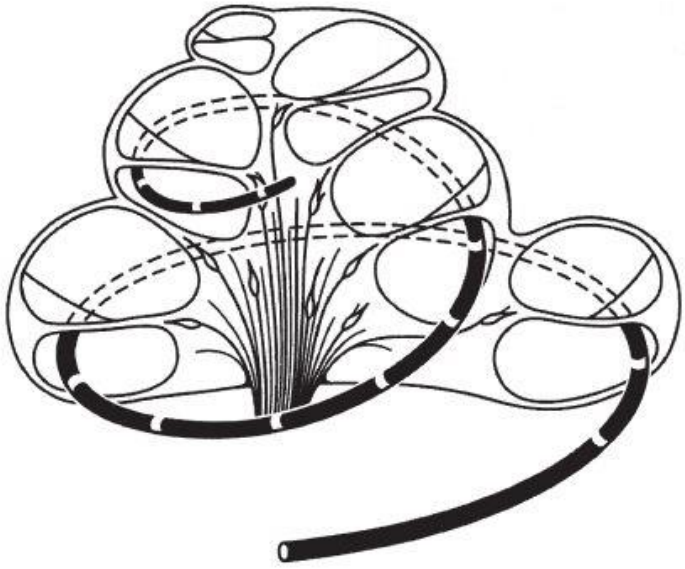


Figure 2. 2 Cross-section through an implanted cochlea, demonstrating the position of the electrode array in the scala tympani. The white squares on the electrode array represent the electrode contacts. From Zeng (2004b: 10). Reproduced with permission.

A cochlear implant attempts to replicate the tonotopic organisation of the cochlea (Moore, 2003b), stimulating basal electrodes to represent high-frequency sounds and apical electrodes to represent low-frequency sounds. Electrode arrays typically comprise of between 12 and 22 active electrodes; this number varies between manufacturers and devices.

Stimulation of the electrodes is in the form of charge-balanced biphasic pulses (an equal amount of charge is passed in both directions). The amplitude of the speech signal is coded for by the amount of electrical charge delivered to each electrode pair and is controlled by the current amplitude and pulse width or duration (see figure 2.3); the greater the amplitude and/or width of the pulse, the more electrical charge is delivered and the louder the perceived sound. High stimulation rates (number of pulses per second) can be used to represent finer temporal detail in the resulting stimulation pattern, however, there is a trade-off between temporal resolution and spectral resolution; the greater the number of channels selected in each frame, the lower the overall stimulation rate.

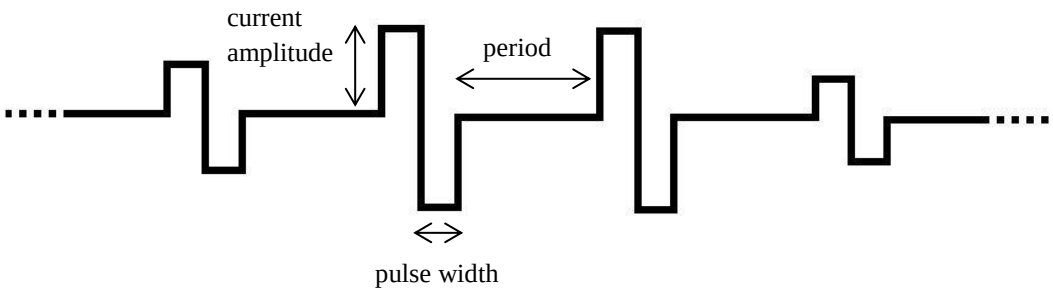


Figure 2. 3 Example of a section of a pulse train for one stimulation channel

2.3 Speech processing strategies

Preferably, speech signals should be coded in such a way as to preserve as many of the acoustic cues required for good speech perception as possible. As the information bandwidth of a CI is limited, the speech processing strategy plays an essential role in maximising the transmission of important speech cues. Speech coding strategies are therefore instrumental to a CI user's ability to perceive and communicate via speech. Figure 2.4 illustrates the stages of electric hearing and gives an overview of the processing that is performed by the CI's speech processor. Although many current speech processing strategies share similar front-end processing, they differ in their method of channel selection and, therefore, their resulting

pattern of stimulation. Present strategies are generally categorised as either fixed-channel or channel-picking strategies and these are explored in the following sections. Specifics are given for Cochlear™ devices (Cochlear Corporation, www.cochlear.org).

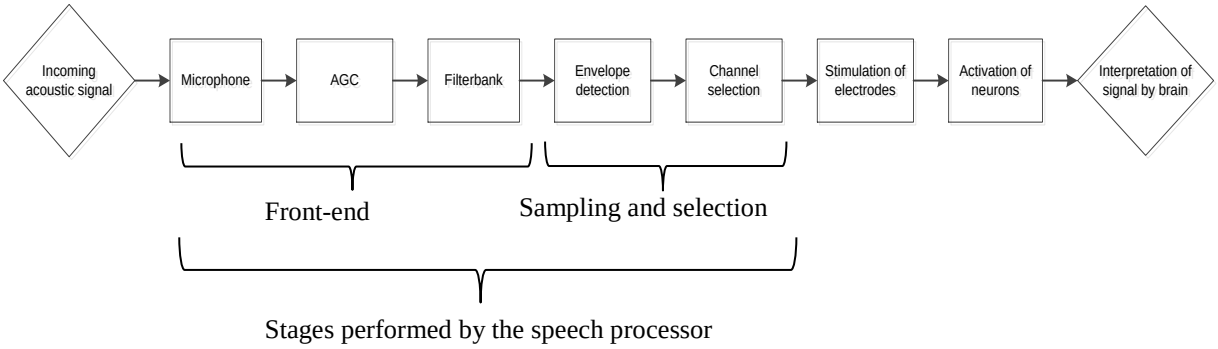


Figure 2. 4 Overview of the stages of electrical hearing

2.3.1 Fixed-channel strategies

For fixed-channel strategies the number and position of stimulation sites is fixed, with each channel stimulated during each frame. The Continuous Interleaved Sampling (CIS) strategy is a fixed-channel strategy which utilises a relatively small number of input channels (Dorman et al., 2002), ranging from four to twelve out of a possible 22. By stimulating only a small number of channels in each frame, CIS is able to utilise high stimulation rates (typically between 600-1800 pps) and thus better represent rapid temporal changes in speech; however, this comes at cost of reducing the representation of finer spectral changes. The bandwidth of each channel is determined by the total number of channels selected, and current level is in direct proportion to the energy in the corresponding frequency band. CIS does not make any assumptions about how speech is produced or perceived and therefore no specific speech features are extracted or represented in the resulting electrical output.

2.3.2 Channel-picking strategies

Channel-picking strategies, often referred to as 'number of maxima' (n-*of*-m) strategies, select channels corresponding to the spectral maxima of the incoming signal (frequency bands containing greater than a certain defined amplitude) (Allum, 1996). Spectral Peak (SPEAK) and the Advanced Combination Encoder (ACE) are both n-*of*-m strategies employed in the Nucleus 24 device.

SPEAK uses a total of 20 bandwidths with up to ten of the highest amplitudes selected for any one stimulation frame. As '*n'* can vary for each stimulation frame, SPEAK is considered as an adaptive n-*of*-m strategy. The exact number of maxima selected depends on the spectral composition and intensity of the signal. The number of maxima delivered per stimulation frame affects the overall stimulation rate; fewer channels result in a higher stimulation rate. On average, SPEAK stimulates each channel at a rate of 250 pulses per second (pps), providing an overall stimulation rate of up to 2500 pps (Skinner at al., 2002b).

For the ACE strategy, *n* is fixed and maxima are selected based on the intensity and frequency distribution within the signal. As channel selection is not fixed, as with CIS, the stimulated channels can vary across the array and are related to the spectral input. The ACE strategy uses higher stimulation rates than SPEAK (600-1800 pps), which results in better representation of temporal information. However, there is a trade-off between the stimulation rate and number of maxima that can be chosen as the overall stimulation rate is limited to 14, 400 Hz (Skinner et al., 2002b); the greater the number of selected maxima, the lower the stimulation rate for each channel.

2.3.3 Comparison of fixed-channel and channel-picking strategies

Both fixed-channel and channel-picking strategies can provide high levels of speech understanding. Skinner et al. (2002a) conducted a comparison study which looked at speech perception scores and preference with ACE, CIS and SPEAK for 12 newly implanted users. Each of the participants used the three strategies in three, two week blocks. The participants were split into three groups, with each group using the strategies in a different order. The participants did not know which order they had used until the end of the study. At the end of each two week block, the authors tested the speech perception abilities of each implant user for each of the three strategies. The authors found that half of the participants scored significantly higher on sentence testing with the ACE strategy. Overall group mean scores for the sentence recognition task showed that ACE gave statistically significantly higher scores than for CIS and SPEAK, and seven of the 12 participants also showed an overall preference for ACE at the end of the study. Participants completed the sentence test in noise, using a SNR which had been individually determined at the start of the six week trial. Sentence testing in noise was used to minimise ceiling effects (this ranged from no noise right down to +3 dB for better performing listeners). However, despite using a SNR of +10, two subjects consistently achieved scores close to or at ceiling level. Results from this study suggest that

CI users should be able to choose from a range of processing strategies in order to maximise speech perception and that this might be of more importance to poorer performing users.

Despite similarity in speech recognition scores with fixed-channel and channel-picking strategies, CI users often show a preference for channel-picking strategies over fixed-channel strategies (Dorman et al., 2002; Kiefer et al., 2001; Skinner et al., 2002a; 2002b). Nevertheless, current strategies are not able to yield comparable speech recognition scores with NH listeners (except in the cases of some star performers). As it stands, neither channel-picking nor fixed-channel strategies appear to provide CI users with the important cues they require to follow speech in background noise.

2.4 Acoustic simulations of cochlear implant processing

CI simulations are often used as a method for predicting the pattern of CI user performance when assessing the effects of specific aspects of CI processing on speech perception; for example, channel number, stimulation rate, stimulation mode and electrode insertion depth. They are also often used to help guide development of new processing strategies. Conducting simulation studies (also referred to as vocoder studies) allows researchers to review results in the absence of confounding variables caused by patient differences, including duration of deafness prior to implantation, degree and pattern of nerve survival, the position of the electrode array within the cochlea (e.g. insertion depth) and the proximity of the electrodes to the surviving nerves. It can be difficult to recruit a homogeneous group of CI users (although it is attempted, see Verschuur, 2009), and this can make it difficult to delineate differences in results between users in an experiment (Chen and Loizou, 2011:1).

The literature reviewed in this thesis discusses studies which have used CI user and/or CI simulation experiments. In chapter 5 a series of experiments are presented which used NH listeners, listening to CI simulated speech, to evaluate the proposed obstruent landmark enhancement strategy. It is therefore important to understand how changing the parameters of vocoder processing can effect speech perception, and whether the different approaches are comparable. This may affect whether conclusions made in a particular study using a CI simulation (or from the experiments conducted as a part of this thesis) are applicable to CI users. This section will therefore explore the different parameters of vocoder processing.

2.4.1. Basics of vocoder processing

Studies have shown good agreement between vocoder simulations and actual CI performance (Dorman et al., 1997; Dorman and Loizou, 1998; Fu et al., 1998; Friesen et al., 2001; Verschuur, 2009). Vocoders were first developed by Dudley (1939) and use a series of band-pass filters to extract envelope modulations of the speech signal, which are then used to modulate a carrier signal. Figure 2.5 shows how these filters can be used to estimate the overall shape of the output spectrum. A large number of channels are required to accurately represent formant peaks in the output spectrum of a channel vocoder (Holmes and Holmes, 2001), however, as few as four may be sufficient to provide enough detail for speech communication (Shannon et al., 1995). The smoothing of formant peaks through vocoder processing is demonstrated in figure 2.5, whereby the amplitude of the peaks in the output spectrum, particularly in the higher frequencies, does not match that of the input (or desired) spectrum. This figure also demonstrates how more channels are used to represent the lower frequencies and that channel width broadens with increasing frequency, meaning that a single high channel represents information across a wider frequency range.

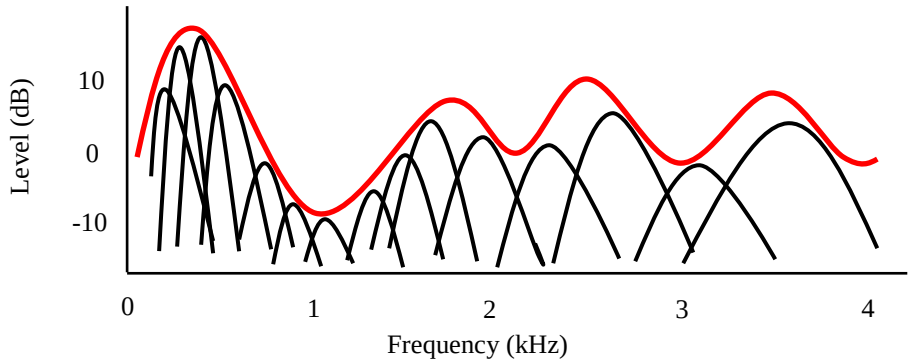


Figure 2. 5 The contribution of individual channels (black lines) in a 15-channel vocoder to the output spectrum. The red line represents the desired spectrum shape. After Holmes and Holmes (2001: 53).

Acoustic simulations of cochlear implant stimulation use processing similar to that of the front-end of a real CI device (pre-emphasis, filterbank, channel selection). The primary difference is that an acoustic signal is produced (using a carrier signal modulated by the envelope signal) rather than a series of electrical impulses. Nonetheless, the resulting information delivered by the simulation to the listener is restricted in a similar way to that delivered by a CI. In simulations, the signal is first filtered into a number of frequency bands

(n), then the resulting envelopes from the frequency bands are used to modulate a carrier tone; this may be either a sine-wave which matches the centre frequency of the filter, or narrow-band noise whose bandwidth equals that of the analysis band. The resulting signals are then rectified and smoothed. The resynthesis stage of the vocoder simulation typically uses a set of band-pass filters, with cut-off frequencies which match those of the analysis filterbank, to filter the noise-band carrier or determine the frequency of the sine carrier (equal to the centre frequency of the analysis filter). The filtered noise-band carrier or sine carrier for each channel is amplitude modulated (mimicking the electrical impulse) and summed with the amplitude modulated signals for the other channels; this is represented in figures 2.6 and 2.7.

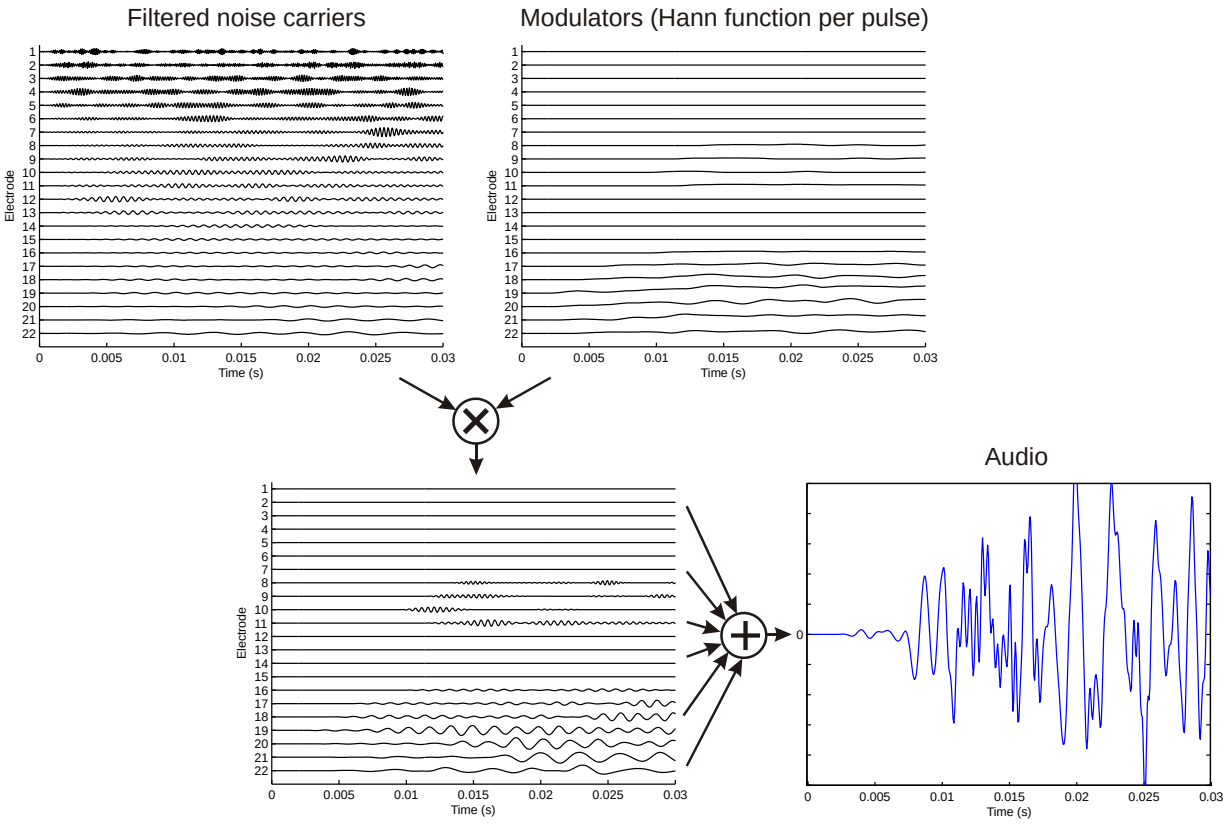


Figure 2. 6 Noise-band vocoder (courtesy of Swanson and Mauch, Cochlear Ltd., 2006)

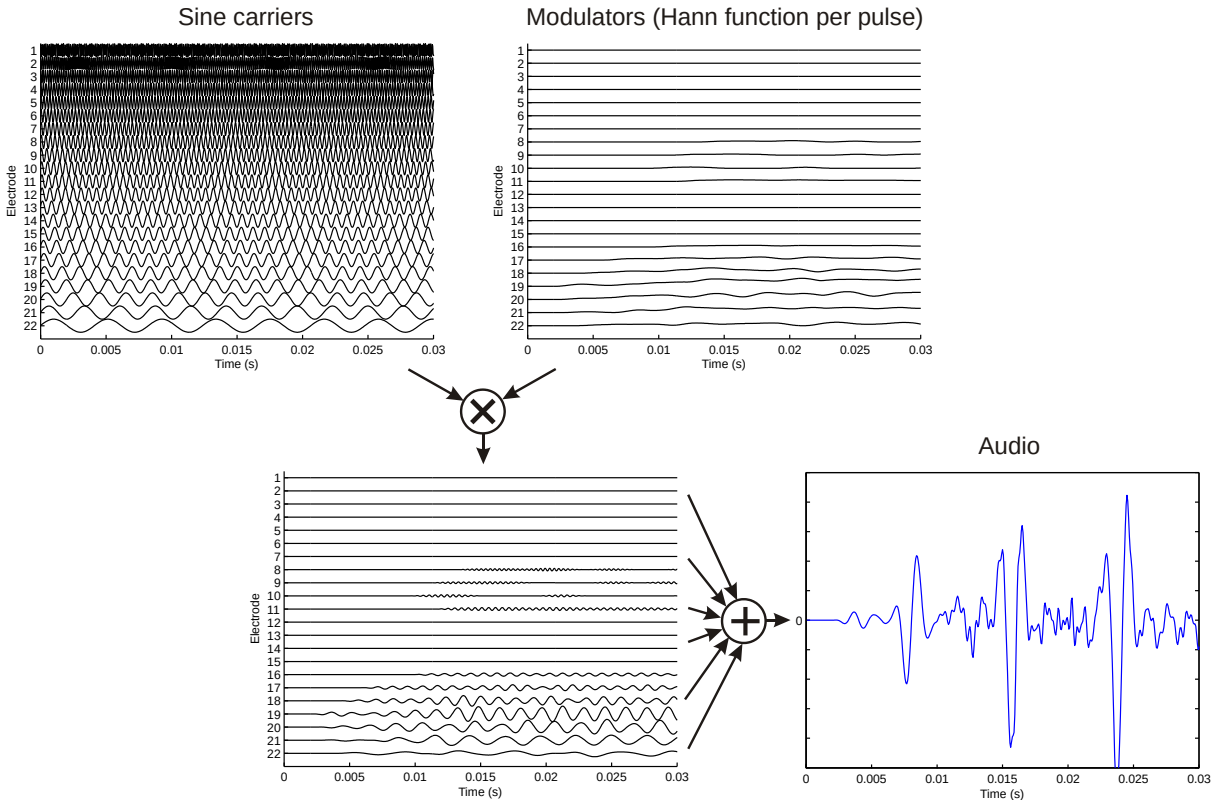


Figure 2. 7 Sine vocoder (Courtesy of Swanson and Mauch, Cochlear Ltd., 2006)

2.4.2 Carrier type

There is no agreed upon vocoder carrier type that best represents cochlear implant performance, and researchers often do not provide a justification for the carrier type used in their experiments. Shannon et al. (1995) showed that high levels of speech recognition could be achieved by NH listeners when listening through a four channel noise-band vocoder. Since this finding there have been a number of studies which have looked at the effect of vocoder carrier on speech perception scores. Dorman et al. (1997) compared speech intelligibility scores in quiet, with noise-band and sine vocoders, for eight NH listeners with two tonine channels. The authors looked at a range of speech materials including three vowel recognition tests, consonant identification and sentence recognition. Differences in scores with the noise-band and sine vocoders were small and often non-significant. They also found that five channels were sufficient to produce 100 percent recognition scores for sentences but for more difficult speech material, such as multi-talker vowels, eight channels were required to reach performance maximum.

Fu et al. (2004) explored the contribution of temporal and spectral information in the identification of talker gender (in quiet) for six NH listeners listening to CI simulated speech and for 11 CI users. NH listeners were tested using sine vocoded vowel stimuli to simulate four, eight, 16 and 32 channels. The decision to use a sine vocoder was made based on pilot data which found that NH listeners scored at chance level in the four channel condition with a noise-band vocoder. Results from the comparison study found the CI users' gender discrimination to be in line with scores obtained for NH listeners listening with four toeight channels (sine vocoder).

Gonzalez and Oliver (2005) looked at the effects of noise-band versus sine vocoders on a listener's ability to discriminate gender and identity of different speakers. Fifteen NH (native Spanish) listeners were asked to determine whether the Spanish sentence ('How old is your cousin from Barcelona?') was produced by a male or female speaker when vocoded using three-16 channels. The sentence was recorded by 40 different speakers, half of whom were male and the other half female, and subjects were presented with 10 sentences per condition. On average, the sine vocoder yielded results (proportion of correct responses) near ceiling for all channel conditions, whilst the noise-band vocoder ranged from around 65 percent correct for responses with three channels to 90 percent correct with 16 channels. Results with the sine vocoder are in line with those found by Fu et al. (2004) for eight-16 channels; however, they are higher than those obtained for CI users, especially for the poorer performing participants. Results with the noise-band vocoder were also above chance for all conditions. The difference between the two studies may be a result of the speech material used; with the sentence task having a much longer duration over which to resolve gender cues than for short duration vowel stimuli. The authors also compared their findings with those of an earlier study by Chinchilla and Fu (2003, as referenced in Fu et al., 2004) who compared gender discrimination and vowel recognition for CI users and NH listeners listening to sine and noise-band vocoded speech. As with Dorman et al. (1997), the authors found that recognition scores with the different vocoder types were very similar, however, the sine vocoder yielded higher gender discrimination scores.

Whitmal et al. (2007) conducted a study to compare the speech intelligibility with noise-band and sine vocoders both in quiet and in noise. They recorded sentence recognition scores for 12 NH native English speakers listening to speech through a six-channel vocoder in quiet and at 3, 8, 13 and 18 dB SNRs for SS noise and two-talker noise (TTN). The participants were tested using both the sine and noise-band vocoders, with half of the participants listening in

the SS noise condition first and the other half listening in the TTN first. Results showed that the sentences were significantly more intelligible with the sine vocoder in all conditions. On average, participants scored 13 percent higher with the sine vocoder in quiet and 20 percent higher when listening in noise. However, results did uncover a significant learning effect with scores for the second carrier being higher than those from the first.

In a second experiment, Whitmal et al. (2007) looked at the effect of carrier type on the intelligibility of consonants when listening in SS noise at the same SNRs used for the sentence test. Floor effects were observed at +3 dB SNR for both the sine and noise-band vocoders (around 23 percent correct) and scores with the two carrier types did not differ significantly in quiet. At all other SNRs, (8, 13, 18 and 23 dB), performance was significantly better with the sine vocoder. In a further experiment, the authors looked at consonant confusions for the two carrier types in both quiet and at 8 dB SNR (SS noise). For voicing, scores were near 100 percent correct in quiet for both carriers (99 percent for the sine vocoder and 93 percent for the noise-band vocoder). In noise, the sine vocoder gave better voicing results (91percent) than the noise-band vocoder (86 percent). As the noise-band vocoder caused more unvoiced-voiced conversions, the fact that there were slightly more voiced consonants in the test material likely gave the sine vocoder an advantage.

Results showed similar overall intelligibility for manner-of-articulation (see section 3.1) for both sine and noise-band vocoders in both quiet and noise (94 percent for sine and 92 percent for noise-band in quiet and 79 percent and 75 percent in noise, respectively). However, the pattern of errors was different for each vocoder type; the sine vocoder was better for identification of stops, nasals and semivowels and the noise-band vocoder gave better recognition of fricatives. Scores for place-of-articulation (see section 3.1) were lower than for manner-of-articulation (for both carrier types and both in quiet and noise). Once again, in quiet, scores for place were similar for the sine and noise-band vocoders (both 70 percent) but again error patterns were different; sine vocoders were better for the recognition of labial (/p, b, m, w, f and v/) and velar consonants (e.g. sonorants and stops with a spectral emphasis in the low frequencies) whereas the noise-band vocoder was better for alveolar consonants (i.e. fricatives and stops with high-frequency emphasis). In noise, results for place with the sine vocoder were slightly higher than those for the noise-band vocoder (53 percent and 61 percent respectively). The overall pattern of results (voicing and manner scores higher than place scores) matches that shown by Miller and Nicely (1955) for natural speech and also Dorman et al. (1997) for vocoded speech.

These results contradict results from earlier studies which showed no advantage of sine over noise-band vocoders. The apparent advantage of sine vocoders observed by Whitmal et al. (2007), Fu et al. (2004) and Gonzalez and Oliver (2005) may be due to the presence of spectral sidebands which lead to harmonic-like spectral components that can be resolved by the listener (Whitmal et al., 2007; Gonzalez and Oliver, 2005). These spectral sidebands can give listeners additional information relating to pitch periodicity, and therefore better $f0$ information, which aids speech perception (Souza and Rosen, 2009). This is not the case for noise-band vocoders and therefore may make noise-band vocoders more realistic; however, envelope fluctuations of the noise-band carrier can interfere with envelope cues and mask important speech envelope modulations (Dau et al., 1999). This can impair a listeners' ability to extract useful speech envelope information and does not occur in actual CI processing. Studies by Dau et al. (1999) and Kohlrausch et al. (2000) suggest that amplitude modulation detection is better with sine vocoders and that results are more in line with actual CI user modulation detection than for noise-vocoded speech (Shannon, 1992; Zeng, 2004a).

Dorman et al. (1997: 2404) recorded an observation from many of their CI patients that when a single channel is stimulated that it sounds like 'beep tones' rather than bands of noise. This led the authors to suggest that sine vocoders may provide a more realistic simulation of listening with a cochlear implant. The main consideration of whether to use sine or noise-band carriers in simulation studies is therefore not which carrier type yields the highest scores, but rather which is more in line with actual CI user performance. Many simulation studies show a lower inter-subject variability than that observed for actual CI users and this may be due to the nature of acoustically simulating performance (i.e. a healthy cochlea with intact nerves).

Souza and Rosen (2009) demonstrated the complex interaction between envelope cut-off frequency, carrier type and number of bands. They found that for lower envelope cut-off frequencies (30 Hz) performance was best with the noise-band vocoder in consonant, vowel and sentence recognition tasks, and that subjects scored higher with the sine vocoder for higher cut-off frequencies (300 Hz). This would appear to support the theory that sine vocoders can provide some information on $f0$, and possibly even intonation. Souza and Rosen (2009) also found that differences in scores for the noise-band and sine vocoders were only significant above three channels which may mean there is no particular benefit of using a sine vocoder when simulating only a limited number of channels. Importantly, this study

highlighted that performance with each vocoder type may be highly dependent on the parameters used.

2.4.3 Channel number/interaction

A common trend amongst CI simulation studies is that NH listeners listening to CI simulations show improving speech perception scores as channel number is increased (Shannon et al., 1995; Dorman et al., 1998; Fu et al., 1998; Loizou et al., 1999). Although, CIs can have up to 20-22 channels available for stimulation, Fu and Nogaki (2005) found that CI users' speech recognition in noise is more comparable with that of NH listeners listening to four spectrally smeared channels. They also reported that actual CI user performance was more in line with the results obtained with noise-band vocoders with broadly overlapping channel filters than for channel filters which were steeply sloping. Simulating channel interaction can be achieved by changing the order of the resynthesis filters of the vocoder; a lower order simulates a greater degree of current spread. A slightly cruder method is simply to use fewer channels, thus limiting the amount of spectral information represented in the final acoustic output. This method has been adopted in CI vocoder studies (for example Whitmal et al., 2007), whereby the number of channels is selected to produce a performance that is more in line with actual CI user performance.

Oxenham and Kreft (2014) measured sentence recognition in different masker types at a range of SNRs for 12 CI users and four NH listeners using a 16 channel sine vocoder. Typically, the NH listeners scored higher than the CI users, with scores at ceiling performance at 5 dB SNR, compared to an average score of 50 percent at the same SNR for the CI users, and even at 20 dB SNR CI users' scores were below 90 percent. However, when the authors modelled spectral smearing in the vocoder simulation, performance for NH listeners was then found to be in line with that of actual CI users at a range of SNRs and for different noise types. It is possible that scores for NH listeners may have been lower with a noise-band vocoder without smearing, as noise stimulates a wider region of the basilar membrane than a pure tone and could possibly be a better model of electrical stimulation (this is discussed in more detail in section 2.5.3).

Due to the use of different masker types, it is not possible to directly compare sentence recognition scores in noise from this study with those obtained with the sine vocoder by Whitmal et al. (2007). Whitmal et al. (2007) tested participants using six spectrally independent channels to try and mimic CI user performance whereas Oxenham and Kreft

(2014) initially chose to use 16 channels because the majority of their CI users had 16 channels of stimulation, however, as this did not accurately represent their performance the authors introduced spectral smearing to model channel interaction. Despite these methodological differences, there does appear to be some agreement in performance levels between the studies for NH listeners listening to the sine vocoded speech in noise (at least for lower SNRs). This could suggest that using fewer spectral channels, rather than modelling channel interaction could be used in vocoder studies to predict CI user performance.

Oxenham and Kreft (2014) argued that their findings were more in line with those found by Dorman et al. (1997), who demonstrated that eight channels were needed for participants to reach asymptote performance; however, this was for vowel stimuli rather than sentences. In fact, Dorman et al. (1997) found that for sentence material (both with sine and noise-band vocoders) five channels were sufficient to reach asymptote performance; this does not seem too dissimilar from results reported by Shannon et al. (1995). The differences in channel number needed to reach asymptote performance between the studies may have resulted from the sentence material used. Shannon et al. (1995) and Dorman et al. (1997) did use different vocoder types, however, it has already been demonstrated that this does not necessarily have an effect on sentence recognition tasks.

Fu et al. (2004) explored the contribution of temporal and spectral information in the identification of vowels (in quiet) for six NH listeners listening to CI simulated speech and for 11 CI users. NH listeners were tested using a sine vocoder with four, eight, 16 and 32 channels. It was found that with eight channels, NH participants were able to score around 80 percent correct for vowel recognition and reached ceiling performance with 16 channels. Results for the CI users were widely distributed between around 55-85 percent correct, putting them in line with scores with four-eight channels for the NH listener group. These results would suggest that the high levels of speech recognition achieved with eight channels in previous studies might be more representative of results for better performing CI users.

Verschuur (2009) looked at the role of channel interaction in modelling consonant recognition scores in noise for CI users. It was found that spectral smearing plays only a small role in simulating consonant recognition performance for CI users (better performing users) and that loss of important information in the signal processing stage is a larger contributor. Nonetheless, it does appear that channel interaction does affect transmission of consonant place-of-articulation cues (explained in section 3.1). For the poorer performing CI

users tested, even the largest degree of channel overlap used in the vocoder study did not accurately predict their performance. The findings could not determine whether a greater level of channel overlap would more accurately model poorer CI user performance or whether there are other factors, not included in the acoustic simulation that resulted in the difference between NH listeners listening to acoustic simulation and the CI user group.

Results from these studies suggest that selecting the correct number of channels to accurately simulate CI performance is difficult to determine. Performance is closely linked to methodological factors such as speech material (Loizou et al., 1999), whether speech is presented in quiet or in noise, masker type and whether or not channel interaction is included in the simulation. Verschuur (2009) recommends that acoustic simulations of CIs should be careful to match the exact processing parameters used by the group of CI users with which their results are being compared with. However, findings from the aforementioned studies would seem to suggest that CI user performance can be simulated, with a reasonable degree of accuracy, using four-eight channels. Nonetheless, it is still not clear whether this more closely resembles performance for better performing CI users and further investigation is needed to explore this.

2.4.4 Experience of listening to vocoded speech

There is evidence to suggest that NH subjects need extensive training to reach a plateau in their performance with vocoded speech. Dorman et al. (1997) replicated the earlier study by Shannon et al. (1995) whereby NH listeners listened to the vowel stimuli with a four-channel noise-band vocoder. Dorman et al. (1997) found that participants scored 76 percent correct whereas Shannon et al. (1995) reported a mean score of 95 percent correct. Scores for consonant and sentence recognition were comparable. Dorman et al. (1997) suggested that one contributing factor to the difference in vowel scores was the amount of training given. Shannon et al. (1995) gave their participants between eight and 10 hours of practice before testing, whereas Dorman et al. (1997: 2407) gave their participants ‘less practice’. This is of interest when considering training given in other vocoder studies, for example, Whitmal et al. (2007) found TTN to be a more effective masker than SS noise when listening to sentences presented in noise. However, prior to testing, participants were given training in SS noise only, meaning the TTN was unfamiliar.

2.5 Information loss with cochlear implant processing

The vocoder studies discussed in section 2.4 have shown that NH listeners do not require access to the full speech signal to obtain high levels of speech recognition, even in noise. However, CI users often do not achieve the same levels of performance. This is mainly the result of the electro-neural interface, which dramatically reduces the amount of information that can be transmitted by the implant, creating an information bottleneck. However, information loss with CIs is also the result of:

- Loss of spatial cues from transformations by the pinna
- Loss of the cochlear amplifier $\longrightarrow$ poorer spectral resolution
- Loss of cochlear compression $\longrightarrow$ reduced dynamic range
- Loss of stochastic firing of the auditory nerve fibres $\longrightarrow$ reduced dynamic range
- Possible reduction in the ability of the brain to interpret sound

Figure 2.8 illustrates the different transformations that take place when sound enters a normal, healthy ear and the information loss that occurs at different stages of cochlear implant processing. The following sections explore how this bottleneck in the flow of information from the implant to the brain impacts on a CI user's ability to understand speech in noise.

2.5.1 Loss of binaural processing

An advantage is observed for listening with two ears, especially in multi-talker backgrounds where competing talkers tend to be separated spatially; their voices are coming from different locations around the listener. It has been found that if a noise source is coming from a spatially separate location from the target speech then speech intelligibility improves. This is known as spatial release from masking (Loizou, 2007). Due to this spatial separation, the signals from these different sources will reach each ear of the listener at different times and at different amplitudes; a result of the head shadow effect (Middlebrooks and Green, 1991). The cues to localising a sound source are interaural level differences (ILDs) and interaural time differences (ITDs). Low- frequency ITD and high-frequency ILD cues are used to localise the azimuth of a sound source (Lorenzi et al., 1999). NH listeners are able to make use of ITDs to segregate the different talkers' auditory streams. The head shadow also allows the ear, contralateral to the noise source, to receive a better SNR, therefore making it easier to understand the words of the target speaker. This apparent binaural advantage diminishes as the number of noise sources increases and become more widely distributed around the head (Bregman, 1994).

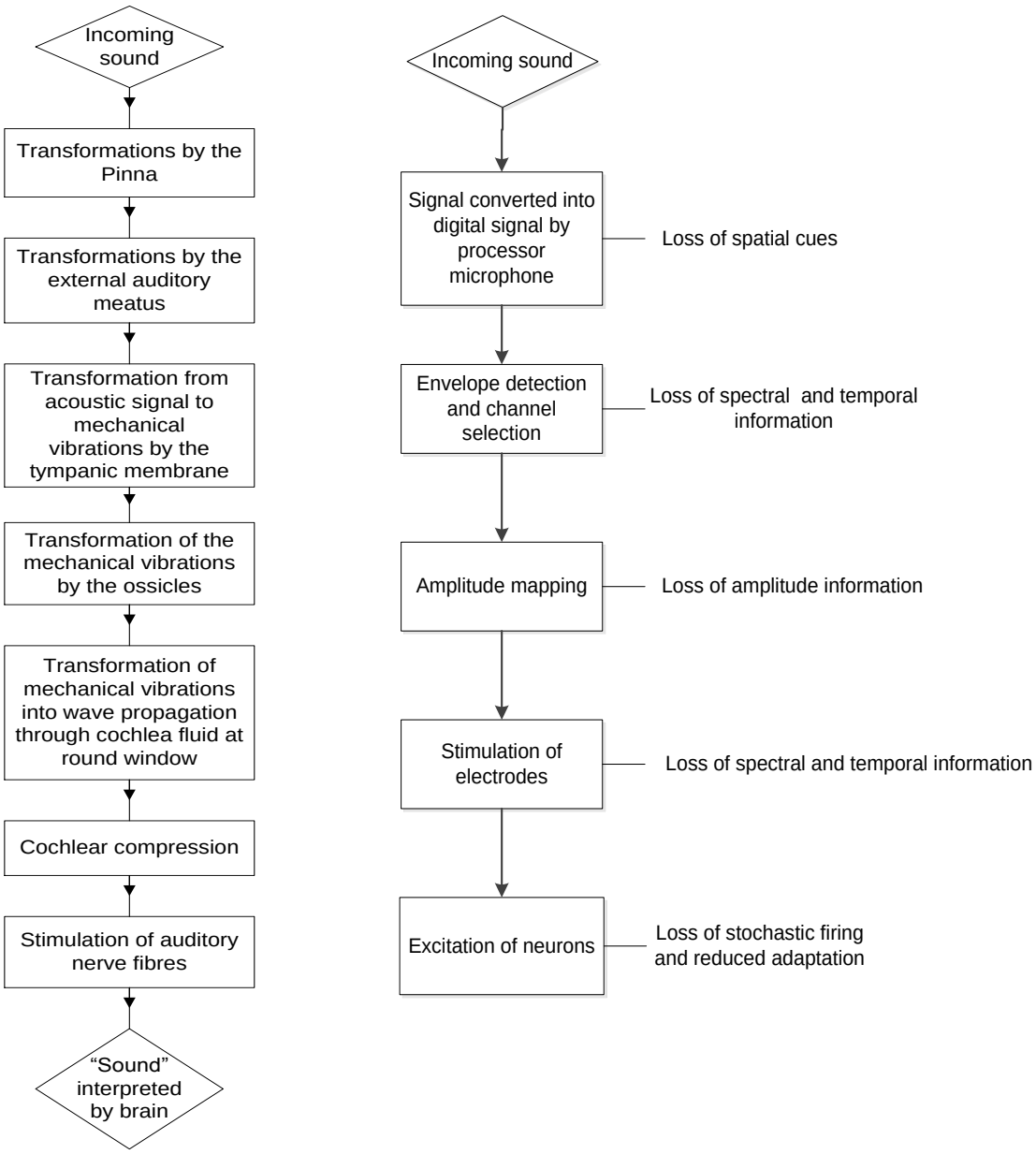


Figure 2. 8 The transformations of acoustic signals by the outer, middle and inner ear (left panel) and the stages of information loss in cochlear implant processing (right panel).

CI users do not benefit from the acoustic transformations which occur in the NH pathway, such as pinna cues required for localisation. A number of studies have explored the ability of bilaterally implanted CI users to make use of ITDs and ILDs when listening in noise (see van Hoesel et al., 1993; van Hoesel and Clark, 1997). These studies have shown that bilaterally implanted CI users are able to lateralise sounds using ILDs but are limited in their ability to make use of ITDs. This is likely the result of a mismatch between the two implanted electrode arrays, which means that the same electrode on both sides can elicit different pitch

and loudness percepts. It appears that for CI users to make use of ITDs it is necessary to reduce the mismatch by stimulating regions of the cochleae which are within a critical band of each other (Neutzel and Hafter, 1981).

There is still some debate as to whether CI users with bilateral implants actually obtain better speech recognition scores in noise than those who have a single implant (Loizou et al., 2009). However, as results from some studies suggest that having two implants is more beneficial than one (Tyler et al., 2002; Müller et al., 2002; Buss et al., 2008) many children are now implanted bilaterally. Ultimately though, many adult patients in the United Kingdom are still only implanted unilaterally, and therefore if not aided on the contralateral ear, will have little or no access to binaural cues such as ITDs and ILDs.

2.5.2 Reduced dynamic range

Normal hearing listeners are able to process acoustic inputs over a range of around of 120 dB (Bacon, 2004). This large dynamic range (DR) allows NH listeners to easily deal with the 30-60 dB range over which amplitude variations in speech occur (Boothroyd et al., 1994). Auditory nerve cells only have a DR of 30-40 dB; therefore one role of the cochlea is to compress the wide range of input levels into the much narrower DR of the nerve cells. When presented with a sound of lower intensity, the outer hair cells of a healthy cochlea act in a non-linear fashion to help increase the DR of hearing. When they are damaged, this function is lost and the DR is reduced to that of the inner hair cells/ auditory nerve cells. As cochlea function diminishes, a person's DR decreases and this results in a much smaller range over which sounds go from being just audible, to comfortable, too loud and then very loud; this is known as a person's loudness growth function (see figure 2.9).

The DR of a CI user may be as small as 5-15 dB (Fu and Shannon, 1999a; 1999b; Hong et al. 2003) but can range up to 20-30 dB in the best cases (Loizou et al., 2000a). Minimum and maximum tolerable electrical stimulation levels are measured for each patient and are referred to as electrical threshold levels (T-levels) and electrical comfortable listening levels (C-levels). T and C-levels are used to determine a person's electrical dynamic range for each electrode pair and stimulation must therefore fall between these two values. T and C- levels are influenced by a number of factors, including type of electrode used, distance between the electrode and the nerve, the electrode-tissue interface and the degree and pattern of nerve survival (Zeng, 2004b). As these factors vary between patients they can result in significant differences in electrode thresholds between CI users.

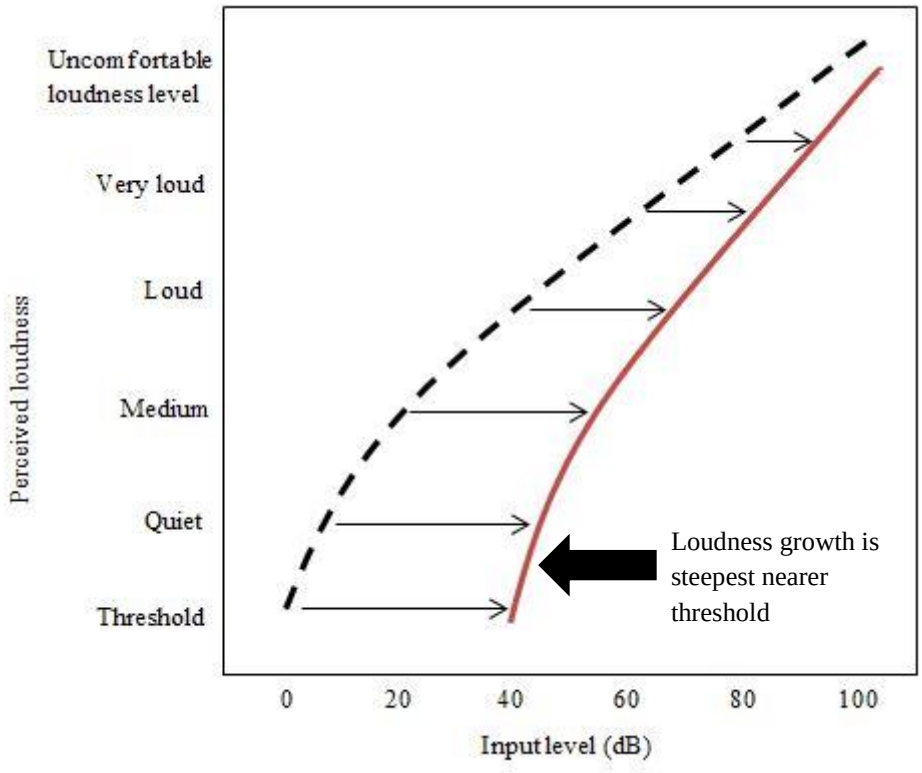


Figure 2. 9 Examples of loudness growth functions for a normal hearing listener (dashed black line) and a hearing impaired listener (solid red line). The difference between the two loudness growth functions reduces as input level increases, creating a much steeper curve for the hearing impaired listener.

Loizou et al. (2000a) tested 10 NH listeners listening to speech (consonants, vowels and sentences) through a six-channel sine vocoder to investigate the effects of (output) DR on speech perception. The authors compressed the amplitude envelopes to create conditions with six, 12, 18 and 24 dB amplitude ranges. The authors noted that the exact mapping function used in the experiment does not match that used in an actual CI device. Analysis of channel output levels highlighted that peak-to-trough amplitude differences are progressively reduced as DR is reduced. This ultimately means that compression results in the reduction of spectral contrast among channel amplitudes (Loizou et al., 2000a). In summary, sentence recognition was more robust to the effects of a reduced DR than transmission of place of articulation information and vowel recognition. This is possibly due to the loss of spectral peak and trough information, and therefore spectral contrast, introduced with compression. The authors posited that their results suggest CI users are forced to make use of across-channel level difference to detect variance in signal frequency. This argument is based on the fact that CI users have access to only a small number of spectral channels.

In an earlier study, Loizou et al. (1999) investigated the use of relative differences in across-channel amplitudes to code spectral information. They restricted channel amplitudes into a number of finite steps (two, four, eight and 16) and measured speech intelligibility for NH listeners listening to six and 16 channel sine vocoded sentences. The authors hypothesised that when spectral resolution is high, the number of amplitude steps would become less important for coding spectral information. When listening with just six channels, eight discriminable amplitude steps were needed to achieve high levels of speech recognition. For the 16 channel condition, just two discriminable steps were sufficient to provide high levels of speech recognition (>90 percent). This is in line with results obtained by Drullman et al. (1995) who found that speech intelligibility remained largely unaffected by poor amplitude resolution when spectral resolution was high. When spectral resolution is poor (fewer discriminable channels), speech perception scores are significantly reduced when the number of steps is small. Results from Loizou et al. (1999; 2000a) would indicate that patients with smaller DRs are likely to perform less well on speech recognition tasks than those with larger DRs and performance is likely to be even worse when the input signal level is low.

Effects of input dynamic range

Typically, modern amplification devices try to compensate for this reduction in a listener's DR by applying less gain for louder input levels than quieter input levels that fall on or below a person's threshold (compression). As the DR varies as a function of frequency, different levels of compression are applied to each analysis frequency band. For cochlear implants, compression is implemented by the automatic gain control (AGC) of the front-end processing, prior to channel selection and at the mapping stage. The input dynamic range (IDR) of the CI processor plays an important role in ensuring that sounds are mapped comfortably within a patient's electrical output DR. The AGC therefore compresses the incoming signal to reduce peaks so that they fall within a user's dynamic range, whilst trying to minimise envelope distortions. The knee-point of the AGC (point at which compression is applied) is determined by the sensitivity control of the microphone (this is set by the clinician). The sensitivity control varies the gain applied to the incoming signal before it is analysed and determines the minimum input level (in SPL) that will result in stimulation.

Figure 2.10 depicts the minimum acoustic energy required to cause stimulation for different sensitivity settings of the Sprint processor of the Nucleus 24 implant; when the sensitivity is set higher a lower input level will result in stimulation and vice versa. If a peak in the signal

is greater than the maximum comfort level of the implant user then the resulting stimulation will be at the maximum comfort level, and never above this. Typically, clinicians will set the sensitivity control to 8, with the minimum audible level falling 30 dB below the knee-point (for Nucleus 24 device). IDR differs between implant devices, though typically IDRs of 30-60 dB are used (for example MED-El devices use an IDR of 60 dB) as this should be sufficient to cover the DR of speech (Zeng et al., 2002).

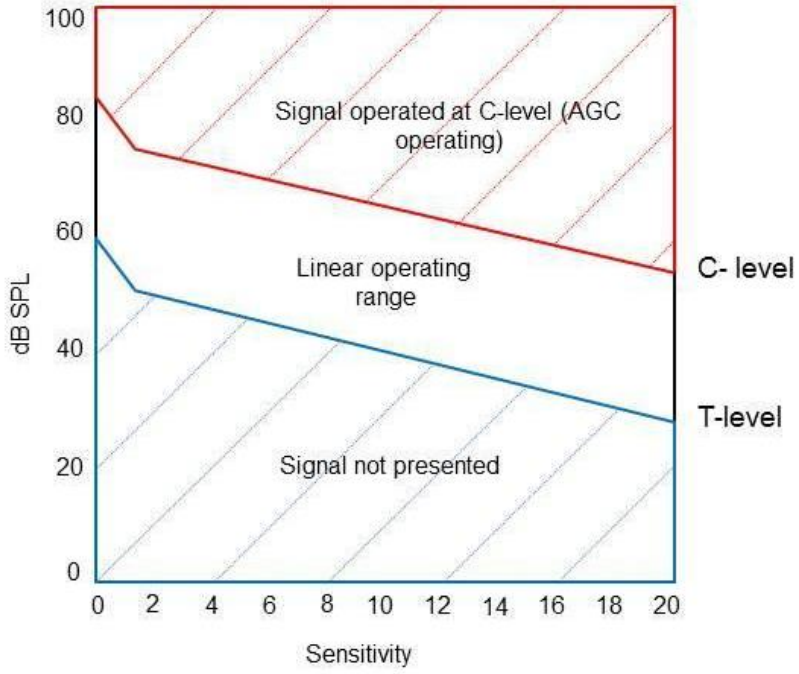


Figure 2. 10 Acoustic energy required which results in stimulation with a Nucleus 24 Sprint processor. After the Nucleus Technical Reference Manual (1999).

Fu and Shannon (1999a) measured the effects of reducing input DR on phoneme recognition in quiet with CI users by using peak and centre clipping. They tested three CI users with a four-channel CIS device and found that an input range as low as 30 dB still provided the participants adequate information about the speech signal to correctly identify the phonemes. Input ranges below 30 dB resulted in poorer phoneme identification and peak clipping had a detrimental effect on the identification of vowels. Cosendai and Pelizzone (2001) also investigated the effects of peak and centre clipping on the intelligibility of vowel and consonant sounds for three CI users and similarly found that vowel sounds were degraded by the processing. Results from their study also suggest that a larger input DR of around 45 dB

is required for optimum speech perception. It is possible that the recommended input DRs from these studies are applicable only to the processing strategies used by the authors (four-channel and five-channel high-rate CIS strategies respectively) and to the speech stimuli used, i.e. consonant and vowel stimuli. Subject numbers for both studies were also very small and may not be representative of a wider range of CI user performance.

Spahr et al. (2007) investigated the role of IDR in the differences in speech perception sometimes observed between users of different implant devices. They observed that for narrow IDRs, such as that used in Cochlear devices, patients may benefit from being able to adjust their sensitivity control for different listening environments; a high sensitivity setting can result in greater gain being applied to background noise, making it difficult for listeners to follow speech. On the other hand, if the sensitivity is set too low in quiet, then listeners may struggle to pick up soft sounds, such as high-frequency consonants. Improved speech perception scores were observed for low input levels when IDR was increased for users of the 3G Cochlear implant processor, however, this had an adverse effect on medium speech input levels (also reported by James et al., 2003).

Overall, the authors found that using a narrow IDR (30 dB) and a low-medium sensitivity control yielded the highest performance for speech recognition in noise. Further support to using a narrow IDR is given by Zeng et al. (2002) who found that wider IDRs impairs consonant recognition for CI users. It is important to note however, that Spahr et al. (2007) documented a wide variation in the scores of their participants and subjects did not necessarily complete every condition; therefore it might be difficult to generalise this statement to all CI users.

The AGC typically uses a fast “attack time”, meaning that when a high-level peak is detected in the signal the instantaneous gain is quickly reduced to ensure the resulting electrical pulse falls within the implant user’s comfort levels (McDermott et al., 2002). The gain is then slowly increased once the peak has passed (release time). Moore (2003b) posits that fast-acting compression in CIs may result in poorer performance in noisy backgrounds because it reduces important amplitude modulation patterns. The AGC applies compression prior to the filterbank and this can lead to ‘spurious amplitude fluctuations, which are correlated across different frequency bands’ (Moore, 2003b: 247), and which may be fused by the listener, making it difficult for them to distinguish between the masker and the target speech signal. Khing et al. (2013) propose the use of a “dual loop” system whereby slow-acting

compression is implemented in the front-end processing, compensating for slower variations in the speech signal (McDermott et al., 2002), but with an additional fast-acting compressor that accounts for faster variations and sudden, intense sounds.

Studies have shown that IDR affects speech perception, especially in noise, however it is not clear whether the effects of IDR are correlated with a person's electrical DR. Spahr et al. (2007) used only the best performing implant patients so as to avoid floor effects in some of their test conditions, and they did not report the subjects' electrical DRs (as given by their T and C-levels) so it cannot be concluded that their "star" performers also had relatively large DRs. It is difficult to draw any conclusions on the interaction between IDR and electrical DR from the studies mentioned above as they have typically investigated IDR between implant devices, although, as previously stated processing strategy and device may have little effect on overall speech recognition scores. Nonetheless, further investigation is required into the interaction between IDR and electrical DR for speech recognition.

In addition, although it is possible to measure an implant user's DR (through T and C- levels), this does not necessarily indicate how many intensity steps they can discriminate within that range, given by the just-noticeable difference (JND) for changes in intensity. This may be of particular importance for CI users, who typically rely more on amplitude cues in speech as a result of their reduced spectral resolution (Cosendai and Pelizzone, 2001; James et al., 2002). Loizou et al. (1999) posit that amplitude cues are essential for accessing important formant frequency information for implant users. Normal hearing listeners can typically discriminate between 50 and 200 intensity steps within their 120 dB DR. On average, CI users can discriminate about 10-20 steps (differences in current amplitude), however it could range anywhere from 7-45 steps (Nelson et al., 1996).

Effects of the amplitude compression function

The amplitude compression function of a CI speech processor is applied during the mapping stage and is the result of three factors; the output DR (T and C-levels measured by the clinician), the shape of the function and the input DR. The ultimate aim of the compression function is to restore normal loudness growth (as seen in figure 2.10) for CI users (Khing et al., 2013). Fu and Shannon (1998) argue that important envelope cues are lost if amplitude mapping does not maintain normal loudness relations. The shape of the compression function is usually logarithmic (Cosendai and Pelizzone, 2001), in an attempt to model the logarithmic compression of the cochlea which is lost with implantation.

Fu and Shannon (1998) explored the effect of amplitude mapping on vowel and consonant recognition for CI users. The authors explored the use of a power-law transformation with varying exponents rather than the traditional logarithmic mapping function. They tested three implant users with the following power function exponents (P) in the mapping stage: 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5 and 0.75 (demonstrated in figure 2.11). Performance in the consonant task was generally best for P= 0.2 and this was also true for the transmission of place, manner and voicing information; with place information being most affected. Varying the power-law mapping function appeared to have very little effect on vowel recognition however. In comparison, the logarithmic mapping function traditionally used in implants is similar to a power law function where P= 0.25 (figure 2.11), suggesting that the amplitude mapping method used does not have a significant impact on speech recognition scores.

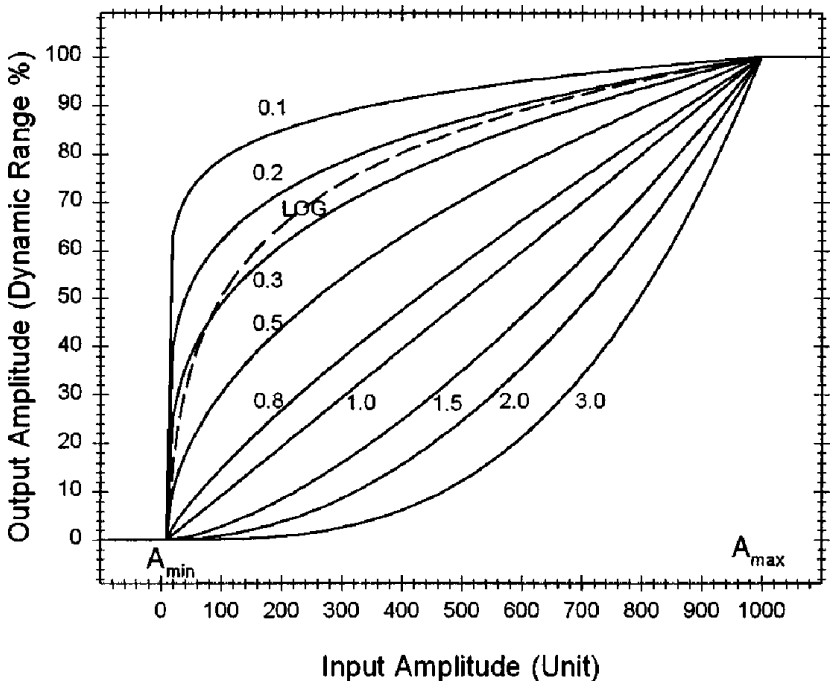


Figure 2. 11 Examples of the power-law amplitude mapping functions used in the study by Fu and Shannon (1998: 2572). P=0.1 would be considered a strong compressive function whereas P=0.5 would be a weak compressive function. Reproduced with permission, copyright Acoustic Society of America.

However, only three participants were tested, so the data will not reflect a wide range of implant users. Also their electrical DRs were not measured, so it is not possible to explore whether patients with narrower DRs are likely to be more affected by changes in amplitude mapping than those with wider DRs. The participants were also tested using a four-channel CIS strategy when they were all daily users of the SPEAK strategy. The authors report that the participants had previous experience with the CIS strategy from taking part in other experiments but this may still have had an effect on their overall scores. In addition, participants were given no training with the different mapping functions so it is possible that they performed best in the condition which most closely mimicked their day-to-day strategy. However, Loizou et al. (2000b) found a similar pattern of results when they tested consonant recognition for six implant users with a six-channel CIS and varied the exponent of a power-law mapping function between -0.1 and 0.6 (linear mapping function).

Fu and Shannon (1999b) conducted a similar experiment to explore the effect of amplitude mapping on speech recognition in noise. Once again, a power-law mapping function that was with an exponent similar to that of a logarithmic mapping function produced the best recognition scores for consonant stimuli both in quiet and in noise. However, weakly compressive mapping functions resulted in only a slight deterioration in scores in noise and a strongly compressive functions cause quite dramatic reductions in scores. Zeng and Galvin (1999) on the other hand did not find any significant effect of compression on speech perception scores in noise. Results from these studies therefore suggest that although the exact method used for acoustic-to-electric amplitude mapping does not necessarily impact speech recognition, the amplitude mapping function should be optimised to help restore normal loudness growth for CI users and maintain critical envelope cues. However, these conclusions can only be made for vowel and consonant stimuli and not for sentence length material. The effect of compression (specifically acoustic-to electric mapping) on speech perception is explored further in section 4.3.1.

Dynamic range and temporal information

Shannon et al. (1995) and Turner et al. (1995) have demonstrated the importance of temporal envelope cues for HI individuals and CI users in perceiving speech. Temporal information is represented by the stimulation rate of CI stimulation; the higher the pulse rate, the more temporal information is conveyed. However, as has already been discussed, there is often a trade-off between temporal and spectral resolution- the greater the number of stimulation

channels, the lower the overall pulse rate for each channel. Temporal information in speech can be split into three categories relating to envelope, periodicity and fine structure (Rosen, 1992). Envelope information is carried in modulations between 2-50 Hz and primarily includes segmental cues to manner of articulation. Periodicity information, between 50-500 Hz, contains strong segmental information relating to voicing and also some information about manner of articulation. Voicing and manner of articulation are linked to some degree as, for example, nasal consonants are always voiced and aperiodicity is a cue for fricatives. This means that some information relating to manner of articulation can be derived from information about voicing. Place of articulation cues are given by fine structure information (>500 Hz) although some segmental cues can also be derived.

CI users typically have good access to envelope and periodicity information (Shannon 1992; Shannon et al., 1995; Zeng et al., 2002). Loizou et al. (2000a) found manner and voicing cues to be fairly robust to the effects of a reduced DR. However, the fastest temporal fluctuations CI users can typically detect are in the region of 300- 500 Hz (Shannon et al., 2004), though some may be able to detect fluctuations as high as 1000 Hz. Detection of these faster fluctuations are critical for speech perception as they give place of articulation cues. Higher stimulation rates have been found to increase a CI user's DR (Galvin and Fu, 2005; Green et al., 2012) however, there is little evidence to suggest that it provides any improvement in speech perception scores (Loizou et al., 2000b; Vandali et al., 2000; Holden et al., 2002 and Friesen et al., 2005). It has been suggested by Green et al. (2012) that this could be the result of increased channel interaction (explored in section 2.5.3).

Galvin and Fu (2005) found that a higher stimulation rate of 2000 pps (as might be used with ACE) produced larger DRs for CI users than a lower stimulation rate of 250 pps (as for SPEAK). This is believed to be because of lower thresholds achieved with higher stimulation rates. The authors wanted to investigate whether CI users were in fact able to make use of speech cues provided by higher stimulation rates, and so measured modulation detection thresholds (MDTs) for six CI users at the different stimulation rates (*ibid.*). Despite the larger DRs measured with the 2000 pps rate, MDTs were significantly higher than for the 250 pps rate. This would seem to suggest that a larger DR does not necessarily provide CI users with better amplitude resolution- i.e. a larger DR does not equate to better envelope processing. Galvin and Fu therefore concluded that due to the limiting factor of electrical DR, increasing stimulation rates for CI users may not actually be of any benefit to their speech recognition. Fu and Shannon (2000), Holden et al. (2002) and Vandali (2000) have all found that

relatively low stimulation rates can provide sufficient temporal envelope information for CI users.

Current CI processing is able to represent envelope and periodicity information but is limited in its ability to represent fine structure cues (though this is being explored). Differences in the abilities of CI users to access different temporal information in speech may account, in part, for differences in performance.

2.5.3 Loss of spectral information

The basilar membrane (BM) of a healthy cochlea is very sharply tuned; with different sections corresponding to a characteristic frequency (i.e. giving maximum vibration for a particular frequency). This sharp tuning is partly the result of the mechanical properties of the basilar membrane whereby width increases and stiffness decreases from base to apex; meaning the base responds best to the high frequencies and the apex the low frequencies. The outer hair cells also help to sharpen frequency tuning in the cochlea by increasing the movement of the BM at particular regions through an active process as a consequence of changing their shape, length and stiffness (Moore, 2003b); this is known as the "cochlear amplifier" (Gelfand, 2004). Inner hair cells then detect the amplified vibrations of the basilar membrane and generate action potentials in their associated auditory nerve fibres. The cochlea has in the region of 3000 inner hair cells, each with 10-20 connected auditory nerve fibres (Zeng, 2004c) and as a result the human ear is able to process frequencies from 20-20000 Hz.

In contrast, CIs are limited to around 20-22 stimulating electrodes that can be inserted into the cochlea. The speech signal must therefore be filtered into a smaller number of filterbands than coded for by a healthy cochlea. The filterbank stage of a CI constrains the amount of temporal and spectral information that is transmitted and as explored in section 2.5.2 the AGC reduces spectral contrast within and across segments. Spectral resolution is further degraded by interaction of stimulated channels and the non-specific activation of neurons. Loss of spectral information in CIs is therefore a result of both CI processing and the electro-neural interface, and is considered to be a large contributory factor to CI users' susceptibility to background noise (Fu et al., 1998). Henry et al. (2005) demonstrated that the limited spectral resolution capabilities of CIs is, in part, linked to the reduction in spectral peaks in the signal and found that reduced spectral peak detection was at least moderately correlated with vowel and consonant recognition scores for CI users.

Shannon et al. (1995) and Loizou et al. (1999) have demonstrated that NH listeners are able to understand 90 percent of speech in quiet with just four or five spectral channels. Results from Fu et al. (2004) suggest that CI speech recognition in quiet is in line with NH listeners listening with four-eight spectral channels. Dorman and Loizou (1998) compared the vowel and consonant recognition scores of seven CI users, using a six-channel CIS strategy, with that of 10 NH listeners, listening to sine-vocoded speech, as recorded by Dorman et al. (1997). They sought to investigate whether it is possible for CI users to extract sufficient information from the speech signal to reach similar performance levels as NH listeners listening to vocoded speech with the same number of channels.

For both synthetic and multi-talker vowels, a six-channel vocoder gave scores of around 80 percent correct and five out of the seven CI users tested in the 1998 study were able to obtain similar scores for synthetic stimuli and four out of seven for the multi-talker vowels. For consonant recognition only place-of-articulation percentage transmission was reported and compared because NH listeners reached ceiling performance for voicing and manner. Again, five out of the seven CI users reached similar performance levels as the NH listeners (around 85 percent). These results suggest that at least some CI users are able to extract similar amounts of information about the speech signal as NH listeners listening to spectrally degraded speech. However, the authors do note that the sample size was only small, and that these findings might not be generalizable to a larger, random sample. The authors did explore DR as a differentiating factor between the better and poorer performing CI users in the study and they found that, on average, the poorer performers had a DR (average across electrodes) of 10-14 dB, whereas three out of the four top performers had DRs >19 dB. As one of the four top performers had a DR more in line with the poorer performing implant users (12 dB) it suggests that DR is not the only factor that determines whether a CI user is a good or poor performer.

The loss of spectral information has a more pronounced effect on speech recognition in noise than in quiet (Fu et al., 1998). Dorman et al. (1998a) looked at speech recognition in noise for NH listeners listening to vocoded speech as a function of the number of channels. They found that in quiet, only four channels were required to obtain 90 percent speech understanding accuracy, but for noisy situations, the number of channels required to reach maximum performance increased to 12 channels for +2dB SNR and 20 channels for -2dB SNR. Based on their findings, Dorman et al. (1998a) recommended that CI users should have at least 12 active channels to help maximise speech recognition in noise.

In a further study, Dorman et al. (2002) explored the optimum channel number for speech perception in quiet and in noise for fixed-channel and channel-picking strategies. They tested NH listeners listening to vocoded vowels, consonants and sentences. In order to reach a maximum performance of 70-80 percent correct for each stimulus type the authors had to use different SNR levels; -2 dB for vowels, +4 dB for consonants and 0 dB for sentences. These levels would indicate that consonants are more affected by the addition of background noise than vowels. Table 2.1 summarises the number of channels required for each strategy to reach maximum performance in quiet and noise for the different recognition tasks.

	Quiet		Noise	
	Fixed-channel	Channel-picking	Fixed-channel	Channel-picking
Vowels	8	3-of-20	10	10-of-20
Consonants	6	3-of-20	6	9-of-20
Sentences	6	6-of-20	10	9-of-20

Table 2. 1 Number of channels required to reach maximum performance for the different speech stimuli, in quiet and noise, for fixed-channel and channel-picking strategies as per Dorman, et al. (2002).

These results indicate that in both quiet and noise, fewer output channels are required for channel-picking strategies than for fixed-channel strategies in order to reach maximum performance. Nonetheless, similar maximum performance scores can be achieved with both strategies, as long as channel number is optimised for each patient. Results from the studies by Dorman et al. (1998a; 2002) suggest to maximise speech perception. CI users require a greater number of channels when listening in noise

However, it has been shown that better performing CI users' speech recognition scores in noise do not improve beyond six-eight channels, and poorer performing users typically do not improve beyond three-four channels (Friesen et al., 2001). This means they are unable to make use of 20 independent spectral channels. In addition, recommendations from Dorman et al. (1998a; 2002) were based on performance levels at SNRs in which CI users may typically score at chance level. As a trend was seen that more channels were required to reach maximum performance as SNR decreased, it could be assumed that similar performance may be achieved with fewer channels at higher SNR levels. Yet, at more favourable SNR levels

such as +16 dB, even better performing users only score in the region of 50-60 percent correct for sentences in noise (Nelson et al., 2003). This suggests that CI users would require many more than the currently available 20-22 channels to reach a similar performance at +2 dB SNR, as that observed in the study by Dorman et al. (1998a). Results from Dorman et al. (1998a; 2002) would suggest that if the spectral resolution of CIs could be improved, then speech recognition scores in noise would also improve. Nonetheless, "simple" reduction of the number of channels available for stimulation does not account fully for poorer results seen in quiet and in noise for CI users.

Effects of channel interaction

Despite the limited number of spectral channels available for stimulation with CIs, vocoder studies (as discussed in section 2.4) suggest that this alone is not enough to drastically reduce speech perception scores. Several studies have compared actual CI user performance with vocoder simulation and found that NH listeners listening to CI simulated speech obtain higher recognition scores in noise with increasing channel number, (see for example Friesen et al., 2001; Xu and Zheng, 2007) whereas CI users often reach a plateau at or below eight channels. A CI is not able to replicate the frequency specificity of the cochlea due to the size of the implanted electrodes and spread of current from each stimulated electrode; CIs are not able to stimulate auditory nerves cells as specifically as in a healthy, functioning cochlea. Auditory neurons need sufficient time to recover after being stimulated otherwise the response to subsequent stimulations will be reduced and thresholds will be increased; this is known as forward masking (Abbas and Miller, 2004). Channel interaction, the overlapping of electrode currents (represented in figure 2.12), can reduce effective tonotopic selectivity and limit spectral resolution through the smearing of amplitude cues across segments. Chatterjee and Shannon (1998) found that channel interaction and the non-specific excitation of auditory neurons can significantly increase forward masking.

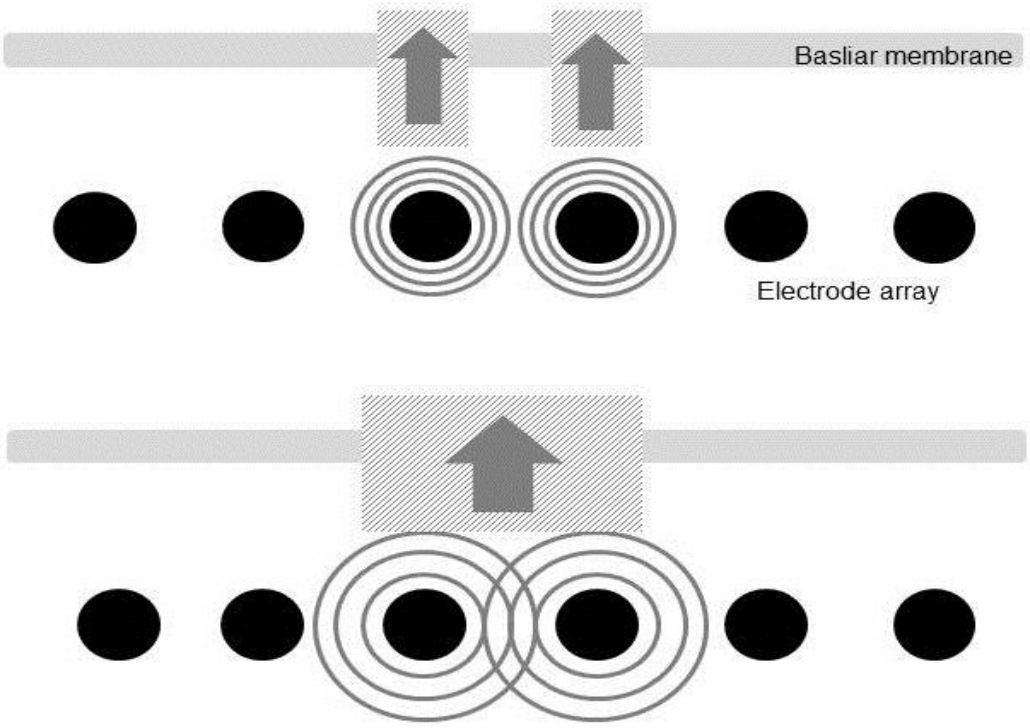


Figure 2. 12 Schematic representation of channel interaction. In the top panel, two stimulated electrodes create narrow, independent fields of excitation along the basilar membrane. In the bottom panel, the current fields generated by the electrodes are broader and overlap. Rather than stimulating two independent sites, they stimulate one, broader site along the basilar membrane.

During electrical stimulation current flows between an active and a paired "reference" electrode (typically referred to as a channel). The location of the reference electrode, in relation to the active electrode, can be altered to create different modes of stimulation. The most common modes of stimulation are monopolar and bipolar. For monopolar stimulation the reference electrode is extracochlear (outside the cochlea), with the active electrode inside the cochlea. For bipolar stimulation, both the active and reference electrodes are intracochlear.

Monopolar stimulation creates the lowest T and C-levels, meaning less energy is needed to produce auditory sensations. This allows for higher rates of stimulation and prolongs battery life. For bipolar stimulation, the reference electrode may be located adjacent (in an apical direction) to the active electrode, or with one or two electrodes between the two. Moving the reference electrode further from the active electrode reduces the number of possible independent channels by stimulating a more diffuse region of spiral ganglion cells (Hanekom

and Shannon, 1998). However, higher stimulation levels are required when the electrode pairs are more closely spaced and this can also result in a broader field of excitation.

Studies have also shown that introducing spectral smearing in the vocoder simulation results in performance closer to that of real implants users. Boothroyd et al. (1996) found that for NH listeners, speech perception in noise is made more difficult by the introduction of spectral smearing. Fu and Nogaki (2004) simulated the effects of channel interaction on speech in noise for six NH listeners listening to vocoded speech, with various degrees of spectral smearing. Spectral smearing was achieved by changing the slope of the carrier band filters of the vocoder and gave two conditions: (1) simulating a CI with little channel interaction and (2) simulating a CI with a significant degree of channel interaction. As spectral smearing was used to simulate channel interaction with electrical stimulation, the exact method did not match that used by Boothroyd et al. (1996). Speech reception thresholds (SRTs) for sentences were measured without processing, and for both levels of smearing, using a four, eight and 16- channel vocoder. The different conditions were then tested in steady-state noise and six levels of gated noise. These were then compared to SRT results for the same speech material and noise types for 10 CI users.

The average SRT (defined as the 50 percent correct level) for the CI group in steady-state noise was +10 dB and did not vary much between the other noise conditions. There was more variation amongst the conditions for the NH group (without processing), with lowest SRTs measured in the steady-state noise (around -4 dB). Relative to the unprocessed speech condition, NH listeners performed significantly worse in all vocoder conditions. In summary, the results indicate that the amount of masking release was highly dependent on the degree of spectral resolution (i.e. number of spectral channels). However, even when spectral resolution was relatively high, no significant release from masking was observed when smearing was introduced. The mean performance of the CI users was comparable to that of NH listeners listening to four spectrally smeared channels.

The CI users included in the study were “good performers” only and were able to reach around 100 percent correct for sentences when listening in quiet. The device and speech processing strategy used varied between participants, with the majority using either SPEAK or CIS. Individual scores for each noise condition showed a large degree of variation both between groups (split into the different strategies and processors) and within the groups. A better comparison could be made if a larger number of CI users had been tested and matched

in terms of implant device, processing strategy and number of active channels. However, the authors reported no clear differences in the trends in performance between the different CI devices/strategies and so suggested that perhaps these variables do not have much relation to the underlying cause of the implant users' poor performance when listening in noise. As the number of channels for each participant ranged from 8-22, but performance did not necessarily improve with increasing channel number, this would indicate that these channels were not spectrally independent and that the users were experiencing reasonable levels of channel interaction. One participant who had the Clarion device and uses the HiRes strategy achieved SRTs that matched more closely to that of a NH listener listening to eight spectrally smeared channels. This CI user had a relatively high number of stimulation channels (sixteen) and stimulation rate (4640 pps). The overriding conclusion of the study was that both reduced spectral resolution (number of independent channels) and a loss of the fine spectro-temporal cues caused by channel interaction are the main causes for poor speech intelligibility of CI users in noise.

Chaterjee and Shannon (1998) have shown that channel interaction is highly individually variable. The excitation of specific neural fields by the electrode array can be affected by a number of factors, including the survival pattern of spiral ganglion cell populations, the proximity of these neurons to the electrodes and the configuration of coupled electrodes (Hanekom and Shannon, 1998; Wilson, 2004). Spiral ganglion cell survival differs between patients, and between different aetiologies (Leake and Rebscher, 2004). Moreover, the survival of spiral ganglion cells tends to deteriorate over time and can be worse for the basal end of the cochlea than the apical half. As the placement of electrodes moves further from the target neurons, broader excitation fields may be produced, increasing the chance of overlap between different channels of stimulation. However, for those who have only a limited number of surviving neural sites along the cochlea, a broader stimulation pattern may be more beneficial. This variety in nerve survival between CI users may mean vocoder simulations that try and model channel interaction may not be an accurate predictor of performance as they are stimulating a healthy cochlea with good spread of nerve populations.

To help improve the selectivity of excitation fields, and therefore reduce the number of overlapping stimulation sites, electrodes should be placed as close to the inner wall of the scala tympani as possible. This may also help to reduce thresholds and therefore increase a person's DR (Shepherd et al., 1993). However, controlling fields of excitation along the basilar membrane is quite a complicated task, and the approach taken may need to be tailored

to the individual depending on their neural survival pattern. Fu and Shannon (1999c; 1999d) demonstrated that electrode location, spacing and frequency allocation can affect speech recognition scores, and that although implant users may be able to acclimatise to the shifted tonotopic pattern created by their implant, it is likely that this is only a partial accommodation.

Effects of front-end processing

Section 2.5.2 discussed how smearing of amplitude envelope cues introduced by the AGC can reduce spectral resolution, but information loss also occurs in the first stages of processing. To maximise the limited bandwidth supplied by the electrode array, implant processors employ an anti-aliasing filter and a sampling frequency of 16 kHz. This means that only frequencies up to 8 kHz are represented in the digitized signal, however, this should represent the main frequencies in speech. The microphone of the speech processor contains a pre-emphasis filter which aims to improve the transmission of relatively weak consonant sounds by attenuating frequencies below 1.2 kHz at 6 dB/octave. However, to help give the best representation of formant information, decomposition of the signal by the filterbank gives more emphasis to the low frequency channels, with low frequency bands being spaced linearly and high frequency bands spaced logarithmically (Laneau et al., 2006). This means that more channels are given over to representing the low frequencies (see figure 2.13), resulting in much narrower bandwidths than for the higher frequencies.

Most CI processors utilise a simple linear filterbank to mimic frequency decomposition and the tonotopy of the human cochlea (Kim et al., 2009); for example, the Nucleus 24 implant uses a 128-point FFT (Fast Fourier Transform) filterbank. For the standard 16 kHz sampling rate, a 128-point FFT filterbank would have a window length of 8 ms, giving a minimum channel bandwidth of 125 Hz. Improved spectral resolution could be achieved if a longer analysis window was used; however, this would come at the cost of temporal resolution.

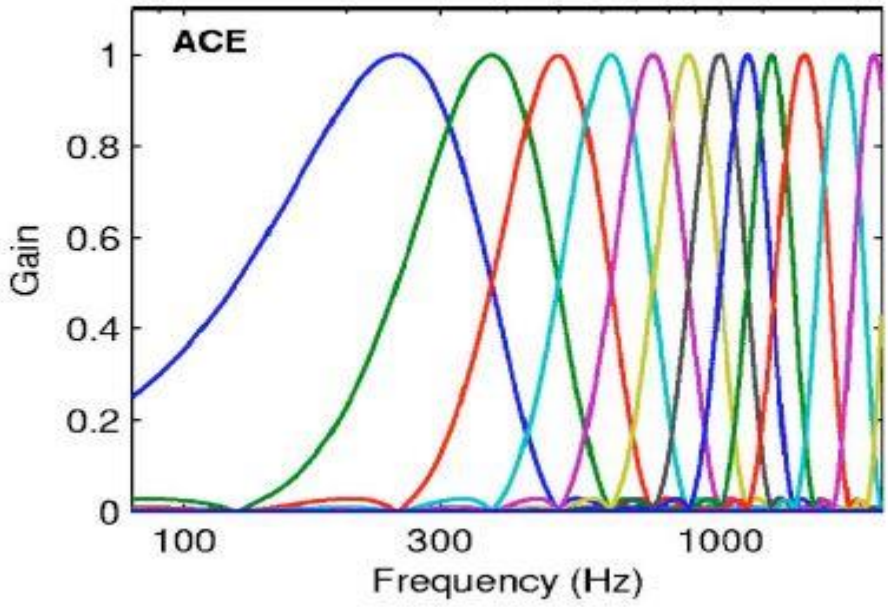


Figure 2. 13 FFT filterbank settings for the ACE strategy in the Nucleus 24 implant processor. Laneau et al., (2006: 495). Reproduced with permission, copyright Acoustic Society of America.

Relative contribution of speech processing and electro-neural interface to reduced spectral resolution

Nelson et al. (2003) explored speech recognition for NH listeners, NH listeners listening with a four-channel noise-band vocoder and CI users in both modulated and SS maskers. In particular, they were interested in CI users' abilities to make use of temporal fluctuations in the masker to glimpse information about the target speech signal; known as release from masking. The authors found that NH listeners obtained significant release from masking in modulated noise, whereas CI users and NH listeners listening to a CI simulation are much more affected by background noise, even at more favourable SNRs. They tested nine implant users and on average they demonstrated very good speech recognition in quiet (80 percent correct), but for SNR levels typical of many environmental situations, for example +8 dB (modulated noise), key word identification dropped by about 50 percent, with similar results obtained for the simulation group. In comparison, the NH listeners (without vocoder processing) achieved similar levels of recognition in SS at 0 dB SNR and for many of the modulated noise conditions at -8 dB SNRs. However, at +16 dB SNR, the CI users were only achieving recognition scores of around 60 percent; a score being reached by NH listeners (without vocoder processing) at -16 dB SNR in modulated noise. Nelson et al. (*ibid.*)

concluded that background noise (modulated or steady) disrupts the ideal amplitude envelope cues coded by the implant processor, even at low levels.

Of interest is that results in the simulation experiment were worse than for the implant user group, even in quiet (average of 50 percent correct compared with an average of around 80 percent correct for CI users). This is much lower than observed in other comparison studies and suggests that the simulation was not a good model of CI performance, or at least the "good" performers recruited for this study. The lower than expected scores for the NH group could be the result of lack of experience with the vocoder processing as the authors do not comment whether they received any training prior to the experiment. It is not possible to ascertain whether the lack of release from masking observed in the implant group and simulation group was a result of reduced spectral resolution in certain frequency regions. In other words, was it because of their inability to detect certain spectral information pertaining to consonants and vowels? This is not easy to determine with vowel and consonant recognition tasks, as the processing applied (addition of noise) will affect each stimulus differently due to their short duration. The authors compared results for different speech processors (Nucleus-22, Clarion 1.2 and Clarion HiFocus) and strategies (SPEAK and CIS) and although there was variation in individual scores, all groups showed a similar pattern of performance.

The authors suggest that the general lack of release from masking observed in the study was the result of the electro-neural interface rather than specific processing parameters, including speech coding strategy, AGC, filterbank etc. However, absolute performance may be influenced by these parameters as results appear to indicate that the Nucleus-22 users had a lower overall performance in the gated noise than the two Clarion device groups. It is not possible though, to determine whether this is the effect of the processing strategy or other processing parameters of the front-end of the device. It is also not evident whether the Nucleus-22 users scored more poorly in noise than the Clarion 1.2 and HiFocus users, as the authors only report scores averaged across all three groups. As demonstrated in section 2.5.2, IDR does have an effect on speech recognition in noise and that this may be more of a problem for Nucleus devices which typically have a smaller IDR. As participants were allowed to set their sensitivity control to allow for a comfortable level this may account for the difference between implant groups. The authors suggest that the ability to detect information in the dips was not the reason for the lack of release from masking observed for implant users, but it may account for differences between groups.

In contrast to these findings, Munson et al. (2003) suggest that CI processing is responsible for the general reduction in transmission of consonant cues, whereas channel number/interaction may account for individual differences. They tested vowel and consonant recognition scores, in quiet, with 30 users of the Nucleus-22 Clarion device. They split the participants into best and poorer performers based on averaged phoneme recognition scores (combined for vowels and consonants). Vowel recognition was better for both groups than consonant recognition, but scores for the poorer performing users were more widely distributed than for the better performers. Results from information transmission analysis showed that although less information was transmitted overall for the poorer performing group (voicing, place and manner), a similar pattern of results was seen across the groups. Consonant confusion matrices showed that both poorer-performing and better-performing users had the most difficulty distinguishing the obstruent consonants. These results suggest that CI processing is responsible for the general reduction in transmission of spectral cues and that channel interaction may account for individual differences in performance.

Existing data for 26 participants (13 from each group) for word identification was also available, and this showed that on average, better performing users, as defined by the vowel and consonant task, scored significantly higher for word recognition than the poorer performers. The two implant devices were found to be equally distributed across the two performance level groups, suggesting this was not a factor in performance. The authors were also able to rule out duration of deafness, age at implantation and duration of implant use as significant between the two groups because of the large number of participants recruited. One observation is that nine of the 13 subjects who had progressive losses prior to implantation were in the better performers group and the top five performers all had progressive losses. Although aetiology was reported for each participant it is not clear whether this was correlated with perception scores. This could be interesting to explore as it may affect the physiology of the cochlea and therefore the pattern of nerve survival and/or placement of electrodes. Nonetheless, these results are in agreement with Verschuur (2009), who found that, for better performing CI users, channel interaction played a significantly smaller role in explaining the reduction in the transmission of consonant information than did the potential loss of spectral information from CI signal processing.

2.6 Improving speech in noise for cochlear implants

Vocoder studies (as discussed in sections 2.4 and 2.5) have demonstrated that NH listeners do not require more than a few channels of information to understand speech, even in noise. However, due to CI processing and the overlap of individual perceptual channels, CI users are not able to resolve as much spectral detail from the speech signal. The reduction in the number of spectral cues perceptually available and smearing of amplitude cues by front-end processing means that CI users are not able to reach the high levels of speech perception in noise observed in NH listeners listening to acoustic simulations of CI processed speech. There are currently two main approaches to improving the transfer of information through the electro-neural bottleneck:

1. **Increase channel capacity** by increasing the number of discriminable channels.
2. **Increase channel efficiency** by making better use of the capacity already available.

Improving channel capacity focuses on increasing the number of discriminable channels available and reducing the problem of channel interaction. This will involve providing more focused stimulation of the auditory nerve (Stafford et al., 2014) and include techniques such as current steering to produce "virtual channels" (Koch et al., 2007; Landsberger and Srinivasan, 2009; Landsberger et al., 2012; Srinivasan et al., 2012), use of tripolar stimulation (Berenstein et al., 2008), stimulation of neurons via infrared lasers rather than electrical pulses (Littlefield et al., 2010; Raiguru et al., 2010; Albert et al., 2012) and improving surgical techniques and implant design to allow electrodes to be implanted closer to the neurons they are stimulating (Shepherd et al., 1993; Middlebrooks and Snyder, 2007; 2010). Although promising, many of these techniques are still only in the very early stages of development and it is not yet clear whether increasing channel number, and by extension frequency selectivity, will in fact improve the representation of specific speech cues important for good speech recognition in noise.

Therefore, rather than trying to increase the amount of information transmitted by a cochlear implant, some researchers believe focus should be placed on minimising the transmission of the redundant parts of the speech signal, so as to maximise encoding resources. Smith and Lewicki (2006; Lewicki, 2002) state that the goal of sensory coding is to form an efficient code which reduces the bandwidth of the source, and for this reason it is proposed that CI devices should maximise the transmission of the essential information from the speech signal. By ensuring the most important elements of speech are preserved it should be possible to use

the limited bandwidth of CIs more effectively. Increasing channel efficiency may be achieved by two methods; improving bandwidth for the transmission of speech cues by employing noise reduction and/or the direct manipulation of the speech signal in attempt to maximise the transmission of specific elements of the speech signal. These two methods are discussed in the following sections.

2.6.1 Noise reduction

A number of speech enhancement algorithms have been developed which aim to reduce noise interference of speech signals for CI users. Traditional Wiener filters and spectral subtraction techniques (Yang and Fu, 2005) attempt to improve the SNR of a signal by estimating the spectrum of the noise signal and subtracting it from the spectrum of the target speech signal. Although these methods have been shown to successfully improve the SNR of signals, this has not always translated into improvements in speech intelligibility (Moore, 2003a; Levitt, 2001; Wang et al., 2009). Improvements in speech recognition in noise for CI users have, however, been observed for multiple-mic approaches (Loizou, 2006). Implant devices employing multiple-mics with two or more microphone ports are able to make use of directivity to improve the SNR of the incoming signal, and are able to significantly improve speech recognition in noise for some implant users (see Van Hoesel and Clark, 1995; Wouters and Berghe, 2001). Nonetheless, benefits seen with different noise reduction techniques appear to vary between implant users.

It has been argued that CI research should move away from basic modelling of cochlea filtering and move towards strategies that focus on extracting the information that the brain will find useful (McAlpine, 2011). The processing strategies discussed below can be considered as approaches to improving the SNR of speech signal through implant processing; however they are of particular interest because they focus primarily on the specific transmission of "essential" elements of the target signal (i.e. speech) by considering neural representations of speech and not just removing noise from the signal.

Ideal Binary Mask or Channel specific SNR

Hu and Loizou (2008) highlight that strategies which base channel selection on amplitude, such as ACE, may be limited in their ability to provide high levels of speech recognition in noise because of the likelihood that they will select channels dominated by the masker. Selecting and stimulating masker-dominated channels may make it difficult for CI users to segregate the masker from the target signal due to informational masking (for modulated

maskers, such as speech) and energetic masking. For channels where the masker is dominant, it may be impossible to resolve the target signal; louder sounds make quieter sounds inaudible. Therefore, Hu and Loizou (2008) proposed an n-of-m strategy based on channel-specific SNR as a way to reduce the transmission of masker-dominated segments of speech and improve speech intelligibility in noise.

The channel-specific SNR strategy proposed by Hu and Loizou is motivated by the principle of Ideal Binary Mask (IBM, see Wang et al., 2008; 2009). IBM is an approach used in computational auditory scene analysis whereby a matrix is constructed for noise-mixed speech where 1 indicates a time-frequency (T-F) section in which the SNR exceeds a certain threshold and 0 indicates otherwise (Wang et al., 2008). The values can then be used to discard T-F segments which are dominated by noise (SNR is below threshold) and therefore allow listeners to focus on regions of the signal where the target speech signal is relatively strong. IBM is able to generate glimpses in the speech signal, this is particularly important for the low-mid frequencies (up to 3 kHz) which include information relating to F1 and F2 (Li and Loizou, 2007). Wang et al. (2009) found that IBM can be used to help improve hearing impaired (HI) listeners' speech recognition in noise and bring it to a level comparable with NH listeners.

Unlike the ACE strategy, channel specific SNR is an n-of-m strategy where *n* is not fixed but instead changes depending on the SNR in each cycle, only selecting channels with SNRs ≥ 0 dB and discard channels with low SNRs (<0 dB). Hu and Loizou (2008) measured speech recognition scores in noise and in quiet for six implant patients, using a CIS processing strategy. They measured sentences recognition in babble and SS noise for both the CIS and the channel-specific SNR strategy. Results for the channel-specific SNR conditions showed that the CI users obtained speech recognition scores comparable with those obtained in quiet for the original CIS strategy and significantly higher scores compared with the CIS strategy in both types of noise. It is not clear as to why the authors did not compare the IBM strategy with ACE. They report that a pilot study with one patient showed a similar pattern of results and therefore speculated that IBM will perform significantly better in noise than ACE. Ideally this should be explored in a more quantitative way as a comparison with another channel-picking strategy may help to further demonstrate the relative advantages of a strategy based on IBM.

As *n* is not fixed for IBM the number of channels selected in each frame can range from zero to 16. The authors found that for sentences presented in babble noise, the most frequently selected number of channels was zero (between 17-21 percent). This was attributed to the fact that speech sounds which are low in energy, such as fricatives and stops, occur frequently in speech but are easily masked by background noise, leading to a high incidence of channels with an SNR <0 dB. Although the authors suggest that selecting no channels during a heavily masked stop burst (for example) may be beneficial for speech intelligibility, this approach may be no better at improving the transmission of these weak but perceptually important speech cues than current speech processing strategies.

The channel-specific SNR strategy used in the study by Hu and Loizou was 100 percent accurate because the authors pre-processed their stimuli to have a pre-defined SNR value and therefore the SNR did not have to be estimated. However, for real-world listening environments, the SNR would not already be known and would therefore have to be calculated for each channel from the mixture envelopes. Using an SNR estimation algorithm, Hu and Loizou (2008) looked at the effect of wrongly retaining/discarding channels and found that listeners can tolerate an error rate of up to 25 percent before performance was compromised.

Li and Loizou (2008b) investigated the required accuracy to estimate the binary mask without compromising speech intelligibility. They tested seven NH listeners listening to IEEE sentences which had been corrupted with SS noise, two-talker masker or 20-talker babble at a -5 dB SNR. The T-F units were computed (as per Hu and Loizou, 2008) and then a fixed percentage of the IBM values were flipped (i.e. changed 0 to 1 or 1 to 0) in each 20-ms analysis frame to introduce errors. This percentage of errors was varied from as little as 5 percent to as much as 40 percent. Participants were tested for each masker type, for each level of error and for unprocessed speech (IBM strategy not applied). The same pattern of results was observed for all masker types, with participants scoring at ceiling for error rates at or below 10 percent. Scores began to drop considerably as error rate increased, with an overall score of 30-40 percent correct for a 40 percent error rate. The authors noted that if the overall aim of noise reduction is to restore normal speech perception scores (as obtained in quiet) then the error rate of IBM must be kept at no greater than 10 percent. However, when scores are compared with performance in noise without any noise reduction then higher recognition scores can be achieved with IBM even when the error rate is as high as 30 percent.

Li and Loizou (2008b) also looked at the effects of error type on speech recognition scores with increasing error rates. They measured sentence recognition scores in which either only type I errors (an original 0 was changed to a 1) or type II errors (an original 1 was changed to a 0) were introduced. Error rate was varied between 20 and 90 percent. Speech recognition performance remained high with type II errors up to 60 percent error rate, whereas type I errors produced recognition scores significantly lower than for type II errors at all error rates until 95 percent. Results were likely poorer for type I errors because more masking noise would have been introduced as a result of changing a 0 to a 1. Nonetheless, it would seem that both types of error need only to be lower than 85 percent to achieve scores better than with unprocessed speech.

Li and Loizou (2008c) explored the potential benefit of the IBM strategy when spectral resolution is reduced, as is the case for CI users. The authors tested 14 NH listeners listening to sine-vocoded IEEE sentences filtered into 6, 12, 16, 24 and 32 channels and corrupted by SS noise and multi-talker babble at -5, 0 and 5 dB SNRs for unprocessed and IBM processing conditions. As channel number increased, performance increased for all conditions, with a plateau in performance being reached with 16 channels for IBM at -5 and 0 dB SNR in babble and 12 channels at 5 dB SNR. Although the benefits of IBM were restricted with decreasing spectral resolution, scores obtained with IBM were significantly higher than for the baseline at all conditions, but particularly in low SNR levels.

In a further study, when the frequency location to which IBM was applied was restricted (by altering frequency cut-off) participants were able to obtain reasonably high levels of speech recognition for cut-off frequencies between 1 and 2 kHz suggesting the importance of F1 and F2 to speech recognition. Performance increased as cut-off frequency was increased, but asymptote performance was reached around 3000 Hz. This suggests that access to high frequency speech cues is also important for improving speech recognition in noise. However, the authors only tested this for a 32 channel vocoder which is not representative of CI users' spectral resolution; this might be because they were limited in the number of sentences available for testing. It would be therefore be good to investigate this effect for smaller channel numbers. Nonetheless, it does appear that IBM should, in theory, be beneficial for CI users when listening in background noise.

Hazrati and Loizou (2013) investigated the potential benefits of the IBM strategy for CI users who traditionally use the ACE strategy. They found that CI users could reach very high levels

of speech recognition (around 80 percent correct) at SNR levels as low as -5 dB SNR, whereas the same users' scores were at floor level with the ACE strategy for this SNR. Even at higher SNR levels (5 dB) users' scores improved in the region of 40 percent with the IBM. However, the real-world benefit of IBM is likely to be much smaller than this as a priori knowledge about the true SNR values was available to compute the IBM values. It may be important that CI users do not receive the full benefit of the IBM strategy, as although background noise is detrimental to speech perception, it is important in giving listeners and awareness of their environment and therefore it is not recommended that noise is completely removed from the signal. Further studies should consider replicating the earlier experiment of Li and Loizou (2008b) and explore the effect of different error rates on the potential benefit of IBM for CI users as this has not been investigated for CI processing.

Another factor that also needs investigation is the time delay that will be introduced as a result of the processing required to calculate T-F segments, as this will affect its feasibility as a useable, real-time strategy. Further to this, the effect of IBM processing on speech recognition in quiet needs to be investigated. Hu and Loizou (2008) proposed an n-of-m strategy where n can range from 0 to 16, meaning that in quiet situations, all channels 16 channels are likely to be stimulated. Based on literature explored in section 2.5.3, the likely result of stimulating all 16 channels is the creation of a large degree of channel interaction which would subsequently worsen speech recognition. The IBM strategy therefore needs further development to optimise the maximum value for n and allow it to only be applied to noisy speech. Nevertheless, the potential high levels of speech recognition that can be achieved in even relatively low SNRs with IBM suggests that increasing speech recognition in noise for CI users does not require an increase in channel number and argues for approaches to channel selection that minimises the redundant elements of the signal.

Sparse Coding

Information Theory states that only a few neurons of a sensory system fire at the same time. Information is therefore represented by a relatively small number of simultaneously active neurons (Olshausen and Field, 2004). Barlow (1959) suggests that neurons higher up the processing stage become less active than those lower down and that this helps to create greater specificity in neural representations of sensory input. As such, sparse codes are often considered as being energy efficient.

Researchers are currently exploring implementing sparse coding principles into the design of a new speech processing strategy. Due to the nature of electrical stimulation, CI users often have neural firing patterns which are synchronised with the electrical pattern of the CI, determined by the speech processor. This means that CI users cannot make use of sparse coding in the auditory system and may be one explanation as to why CI users do not necessarily achieve the high levels of speech recognition that are observed for NH listeners listening to vocoded speech. It has therefore been suggested that to make the electrically stimulated auditory neuron firing patterns of CI users more sparse, they must receive a sparse representation of the speech spectrum, delivered by the speech processing strategy.

Li and Lutman (2008) developed the SPARSE algorithm, which aims to create a more sparse representation of the speech signal for CI users with the idea that it may help improve speech recognition scores in noise. Following the detection of the envelope signals by ACE, the SPARSE algorithm transforms the signal, using principle component analysis (PCA) to help "denoise" the signal. The resulting signal is then subjected to independent component analysis (ICA) which detects the important elements of the signal so that redundant elements (e.g. noise) can be further reduced. The result of such processing leads to a more sparse representation of the signal on which channel selection is then based (as with ACE, this is based on envelope amplitudes).

The authors measured speech perception scores with both ACE and the SPARSE strategy, for NH listeners listening to a CI simulation and for CI users. VCV stimuli were presented in quiet and in babble noise at +15, +10 and +5 dB SNRs. Results indicated that for NH listeners listening to the CI simulation, speech recognition scores were only significantly different for the two strategies in the +5 dB SNR condition, with SPARSE giving better results. The lack of difference between the two strategies at higher SNR levels was likely due a ceiling effect as subjects regularly scored higher than 70 percent, even for the noisy conditions. The improvement seen with the SPARSE algorithm at 5 dB SNR may be due to its ability to reduce the noise in the signal, allowing listeners to make better use of the information relating to the speech signal. It should be noted that an improvement may not have been seen at the higher SNRs because although the participants are listening to a CI simulation, they are still listening and processing the signal with a healthy auditory system. This may mean that their auditory systems are creating their own sparse representations of the processed signals, perhaps further reducing the information in the signal.

For the CI users, the authors noted a wide variation in scores, even in quiet. In addition to this, it was noted that participants who had better scores in quiet tended to show little or no improvement with the SPARSE strategy. However, participants who generally scored quite low in quiet showed an improvement with the SPARSE strategy and that this improvement was larger than that seen for NH listeners. Li and Lutman (2008) therefore suggest that sparse coding may be particularly beneficial for CI users who are typically considered to be "poor performers". Three potential advantages of sparse coding are presented by the authors, these include (1) the reduction in potential channel interaction due to sparse representation of the signal, (2) the algorithm selects the essential elements of the signal meaning the limited DR of CIs can be used more effectively, and (3) it causes the neurons in the auditory system to act more sparsely. It is possible that greater improvements may be seen with the SPARSE strategy if participants are given more training or take-home experience; this may be particularly true for better performing users who have reached a plateau with their current strategy.

SPARSE has been further developed by Hu et al. (2011; 2012). Hu et al. (2011) evaluated the SPARSE strategy with six NH listeners listening to vocoded speech mimicking the ACE and SPARSE strategies. The participants listened to BKB sentences presented in babble noise at 0, 5 and 10 dB SNRs with both strategies and percent correct scores were calculated. A significant improvement in speech recognition was observed for the SPARSE strategy over ACE at 0 dB SNR, however, no such benefit was observed for higher SNRs and this is likely the result of participants reaching ceiling performance for both strategies in these conditions. Although results from this study suggest some benefit of SPARSE for CI users, this may be limited by the SNR used; therefore sentence recognition in noise with SPARSE should be evaluated in the future using CI users listening to speech presented in SNRs in which they typically have difficulty.

Using sparse representations of the speech signal to capture information important for speech perception may help to deal with the potential "overload" of spectral-temporal detail of the speech spectrum, which in itself is highly influenced by a large range in variability in the production of speech (within and between talkers). Sparse coding not only looks to limit redundant information in the speech signal but also attempts to replicate normal auditory processing for CI users, in terms of activating neurons in a sparse way. However, it is unclear from the literature what exactly is deemed as important or essential information. Indeed it may be inferred that anything that is speech is considered as important and that anything

which is noise is considered redundant. If so, sparse coding strategies should be thought more of as noise reduction strategies rather than strategies which aim to explicitly improve the transmission of certain (important) elements of the speech signal.

In addition, the SPARSE algorithm proposed by Li and Lutman (2008) still assumes that the essential information from the signal can still be represented by the channels containing the highest amplitudes and, as discussed earlier, channel selection based on amplitude may not be the best approach. It is also important that performance of the SPARSE strategy should be measured for single-talker backgrounds as it is likely that the algorithm will find it harder to separate the two signals as both will likely be identified as speech. Sparse coding is likely to work better for continuous and more modulated maskers and indeed has so far been evaluated using babble noise only.

2.6.2 Improved representation of specific speech elements

These techniques try to improve the transmission/representation of specific elements of the speech signal which are believed to be important for speech perception. This may include finding new ways to increase information transmission, for example, combined electric and acoustic stimulation (EAS). EAS allows patients with residual low-frequency hearing to use a hearing aid in addition to their CI (either on contralateral ear or on implanted ear). Studies have shown that use of a hearing aid can help improve CI users' access to low-frequency speech cues such as $f0$, which can help to segregate signals from competing talkers (Turner et al., 2004; Kong et al., 2009). It may also involve direct manipulation of the speech signal by applying channel-specific gain to enhance specific frequency regions/events in the signal. Channel-specific gain is applied prior to channel selection and can therefore influence channel selection for n-of-m strategies such as SPEAK and ACE. Channel gain has no effect on channel selection for the CIS strategy as all activated channels are stimulated, however, it will affect stimulation levels. The following sections explore examples of channel-specific gain that can be applied to current implant processing strategies.

Adaptive Dynamic Range Optimisation

Adaptive Dynamic Range Optimisation (ADRO) was developed as a strategy for reducing distortions to the natural peaks and troughs in speech introduced by compression. The aim of ADRO was to optimise the input DR of the signal in each analysis frequency band (of the FFT) so that the output (electrical pulse) produces a percept that is comfortable and audible to the listener (Blamey, 2005). Blamey (2005) argues that CI processors should select the region

of the input DR that will result in the most important information in the incoming signal being transmitted, and at level which the user can actually access. Studebaker and Sherbecoe (2002) state that the most important information in the signal falls within the upper region of its DR. As ADRO does not convey all incoming information, only the important elements, it is able to utilise the limited input DR of a CI more effectively, maintaining natural variations in intensity over time, preventing distortion of the signal.

ADRO uses a series of fuzzy logic rules to control the gain in each channel, meaning that sounds are only made softer or louder if they fall above or below the audibility target. If input signals fall within the range which is both audible and comfortable (T and C- levels) then the gain remains unchanged (linear). This processing can be, and is, applied within current speech coding strategies; for example in the Nucleus 24 device. As ADRO has been designed to improve access to softer input signals it was hoped that it could help to improve speech perception in the following situations:

- when listening to talkers who are quietly spoken
- when listening to conversation between two talkers
- when listening to conversation in a small group of people
- when listening across a quiet room

James et al. (2002) compared speech recognition scores for nine Nucleus 24 users, currently using the Sprint processor, with and without ADRO processing. Seven out of the nine participants used ACE as their day-to-day strategy and two users used SPEAK. Each participant was given at least one week's take home experience with ADRO and was asked to use it in as many different listening situations as possible. For closed-set spondees, scores with and without ADRO processing were both high at ~97 percent (no significant difference observed between scores) at a test level of 60 dB SPL. At 50 dB SPL, participants scored around 9 percent higher with ADRO than without it, however this was not significant. At 40 dB SPL a significant difference was observed with average scores of 35.6 percent correct with ADRO and 14.9 percent correct for the user's normal strategy. Again, for CUNY sentences presented in quiet, no improvement was seen with ADRO at the higher input level (70 dB SPL); however, scores were significantly higher with ADRO at both 50 and 60 dB SPL. The authors also measured speech perception scores in noise; this was to test their hypothesis that ADRO would have no adverse effects on speech when presented in noise.

Participants were tested in +15 and +10 dB SNRs at an input signal level of 70 dB SPL and once again, no significant differences were found between scores with and without ADRO, at either SNR. The authors asked participants to fill out a listening quality questionnaire, covering a range of different listening situations. In general, the subjects either demonstrated a preference for ADRO in most conditions (59 percent) or had no preference at all (31 percent), with only a small number of participants showing a strong preference for their normal strategy over ADRO. The authors therefore concluded that ADRO can be used to improve both the audibility and sound quality of low level speech for cochlear implant users and it has subsequently been implemented in implant processors as part of the ACE and SPEAK processing.

More recently, Ali et al. (2014) compared speech perception scores in noise for CI users with and without ADRO processing. These patients already use ADRO in their everyday MAP. Overall, the authors found no benefit of ADRO, but as in the earlier study by James et al. (2002) it did not negatively impact on scores either. However, these results are only applicable to the moderate listening level of 65 dB SPL that was used and so it is not clear from this study whether ADRO can be used to improve speech recognition for softer input levels. The authors also noted a large variation in scores between users, with some participants performing better with ADRO, some showing no difference and some performing worse and this may have led to the overall non-significant finding. ADRO may therefore improve speech perception for some users, possibly those with narrower DRs.

Transient Emphasis Spectral Maxima

Transient Emphasis Spectral Maxima (TESM) (Vandali, 2001) is based on the knowledge that low-level speech cues that are also short in duration, such as the burst noise of stop consonants (e.g. /p/ and /t/) and formant transitions, may be poorly coded for in current processing strategies as they fall close to or below the minimum input level coded for by the implant system (Holden et al., 2005). In particular, these short duration cues can be used to distinguish place of articulation for consonants, which is typically more poorly transmitted than manner of articulation or voicing with CIs (Dorman et al., 1990). TESM therefore seeks to specifically emphasise short-duration acoustic cues by applying a multiband AGC transient emphasis algorithm to the filterbank outputs of the coding strategy, before channel selection. Additional gain is applied to any filterbank band for which a rapid rise in the slow-varying envelope signal is measured (Vandali, 2001) and even greater gain is applied where a rapid

rise is followed by a rapid fall (associated with a noise burst or formant transition). This gives the channels containing these cues a greater chance of being selected.

Vandali (2001) measured speech recognition scores for eight CI users with the TESM strategy and the spectral maxima sound processor (SMSP) strategy (a precursor to SPEAK). Participants were given take home experience with both strategies and had weekly training sessions with material similar to that used for the final evaluation. The author monitored each participants perceptual scores at the training sessions to give an indication of when they had adjusted fully to each strategy (a plateau in scores was observed). Testing for each strategy included a four week evaluation, whereby speech recognition results for consonant-nucleus-consonant (CNC) words in quiet, and sentences in noise, were recorded on a weekly basis. Results were analysed individually for each of the eight participants and showed that the TESM strategy gave significantly higher scores for the sentence tests in noise for four out of the eight subjects, for the other four no significant difference was observed in scores between the two strategies. For CNC words, six out of the eight subjects scored significantly higher with the TESM strategy for at least one of the word test categories.

Confusion matrices were constructed for the consonant scores, showing an increase primarily in the transmission of place of articulation information, particularly for consonants in the final position but also for manner of articulation. However, the author gives no indication that these results are statistically significant. It was suggested that the increase in consonant scores was due to an improvement in the perception of nasals, fricatives and stops. The author concluded that the improvement in speech recognition scores, seen for four out of the eight participants, was the due to the better contrast provided by the TESM strategy between the onset envelopes of the speech signal and the background noise (Vandali, 2001).

Holden et al. (2005) later compared the TESM strategy with ACE, where patients' MAPs had been optimized for soft sounds (using ADRO). They tested eight CI users who used ACE as their everyday strategy, using a take-home strategy through the use of a research processor for vowel, consonant, word and sentence recognition. They also collected preference data from each of the participants via way of a questionnaire, asking them decide whether they preferred listening with TESM or ACE for a number of situations, including conversation in quiet, conversation in small groups, watching TV and listening to music. During the acclimatisation period, subjects were to blind to which strategy they were using at any one time. For the speech recognition tasks, participants were tested in two SNRs; a better SNR

which produced average scores of 75 percent correct and a poorer SNR, giving scores of 50 percent correct. These SNR levels were determined for each participant prior to testing and were used in attempt to replicate realistic listening situations.

Results showed no significant differences in percent scores obtained with TESM and ACE for any of the speech material tested at either SNR, and in fact scores were almost identical for the two strategies. Information transmission analysis for the consonant recognition task did show an improvement in the transmission of manner and voicing features with TESM, however this was for lower input levels only (55 dB SPL), and only showed an improvement of around five percent. The authors were not able to replicate the findings of Vandali (2001) that perception of nasals, fricatives and stops is improved with TESM for moderate input levels (65-70 dB SPL). Results from the questionnaire indicated that initially, TESM sounded louder than ACE and patients were able to perceive the significant time delay (30 ms) introduced as the results of the TESM processing, however they were able to acclimatise to this after several weeks experience with the strategy.

However, patients did report that they could detect soft environmental sounds better with TESM and half of the participants reported that they preferred TESM to ACE for half of the listening situations listed on the questionnaire. In addition, all but one participant said they would prefer use TESM for certain situations if it was available as an option on their current processor. The authors suggest that TESM may be more beneficial for some patients, particularly those with a very limited DR and therefore concluded that although TESM and ACE may yield similar speech recognition scores TESM may be more successful in transmitting low-level short-duration speech cues for some CI recipients and should therefore be available for clinicians to choose from. Ultimately though, current methods of enhancing low-level speech cues with speech specific gain have not yet realised the higher levels of speech recognition, in quiet and in noise that they hoped to achieve.

2.7 Summary

This chapter has explored how CIs are limited in their ability to transmit speech signals, particularly in noise. This is largely attributed to current CI processing and the electro-neural interface which results in reduced spectral resolution and a limited DR. However, information loss with CIs is not just due to poor cochlea pre-processing but is also the reflection of the loss of neural processing. Speech enhancement techniques that have tried to better replicate neural representation of speech signals, such as IBM and SPARSE, have shown promise in

improving speech perception in noise for CI users. An overarching theme of these strategies is that they do not attempt to increase the amount of information transmitted by the implant, but rather look to minimise the redundant elements of the speech signal; therefore only transmitting the important or "useful" information. Other speech enhancement techniques have looked at overcoming the effects of reduced spectral resolution and compression (and subsequently the smearing of important spectral and amplitude changes) by specifically enhancing certain regions of the signal to try and make them more salient. However, so far results with such strategies have shown only limited benefit, and are highly independently variable between CI users. There is a possibility that these strategies have more potential for poorer performing users, although this has not yet been established in the current literature.

Although ACE conveys more spectral information than fixed-channel strategies, such as CIS, basing channel selection on amplitude may not be of benefit in noise, particularly for situations where the noise is dominant. SPARSE and IBM have shown the potential of modifying n-*of*-m strategies to select "*n*" in a more sophisticated way; based on selecting the important information in the signal. However, neither of these strategies has properly defined what they consider to be the "important" or "essential" elements of the speech signal. Both strategies aim to minimise the redundant elements of the signal but their approaches seem to attribute this redundancy to the noise portion of the signal, meaning that anything relating to the target speech signal is deemed essential and therefore retained. This may in fact be a valid approach towards maximising speech intelligibility in noise, as it does not sacrifice the transmission of some cues in order to improve the transmission of others. This means CI users should have access to a broad range of speech cues with which to resolve the target speech cues. On the other hand, these approaches do not make any attempt to improve the transmission of speech cues which are typically blurred or lost with CI processing, yet these may be particularly important for poorer performing implant users.

The following chapters explore how CI processing, especially the sampling and selection stage, may be modified to improve the transmission of specific speech elements considered important for speech perception, without having to increase channel number. In particular, chapter 3 draws on evidence from neurophysiological studies that suggest the auditory system is particularly responsive to sudden changes in the signal, and explores the effect of current CI processing on the ability of CI listeners to utilise these perceptually important cues, especially in noise.

Chapter 3- The role of spectral change in speech perception: an introduction to acoustic landmarks

Speech recognition is incredibly robust to signal degradation, with listeners able to achieve high levels of speech perception with only a small number of spectral channels. This is because speech is very rich in information, giving it high redundancy and allowing listeners to make use of only small glimpses of signal to understand the overall message. The extraordinary redundancy of speech has contributed to the high levels of speech perception achievable (in quiet) with a cochlear implant. However, CI users are not able to make full use of this redundancy when listening in noise; firstly, because the input signal is already severely restricted and secondly, they cannot make use of the same mechanisms by which NH listeners are able to follow speech in the presence of additive noise (e.g. binaural processing and glimpsing).

The previous chapter explored how CI users' speech perception in noise might be improved if speech processing strategies focused more on selecting and transmitting the "important" or perceptually salient elements of speech. Li and Lutman (2010) advocated the concept that "less may be more" and incorporated principles of sparse coding into the widely used ACE speech processing strategy. Both the SPARSE and IBM strategies have highlighted the importance of considering how the auditory system processes sound, beyond simply trying to mimic cochlea mechanics. However, despite arguing that future CI processing techniques should concentrate on preserving the perceptually important elements of the speech signal, neither strategy has clearly defined what these are and have instead focused more on minimising the amount of background noise transmitted.

Exploring information in speech, this chapter starts by considering its production and the resultant acoustic cues. It then argues for the importance of preserving or improving the transmission of rapid spectral changes in the speech signal, drawing on evidence from studies which have investigated the processing of sensory inputs by biological systems.

3.1 Acoustic cues for speech sounds

On a linguistic level, speech may be broken down into segmental and suprasegmental information. Segmental information refers to those elements which help a listener to identify and understand *what* is being said. For example, they include the acoustic cues which allow

listeners to identify phonemes, syllables, words and sentences. Suprasegmental information is related more to cues which provide information on the identity of the speaker (including age and gender), as well as prosody, stress and intonation. Combined, these two elements allow a listener to determine not only what a person has said but the meaning behind the message as well. Chapter 2 also showed how information in the speech signal can be considered in terms of its temporal, spectral and amplitude characteristics; these can be visualised in the form of a spectrogram which displays the relative amplitude of different frequencies across time (an example of which can be found in figure 3.1). These variations in frequency and amplitude relate to information which is instrumental to identifying the different sounds in speech (Drullman et al., 1994). For example, changes in formant frequencies across the utterance are clearly visible in figure 3.1; represented by the dark horizontal bands.

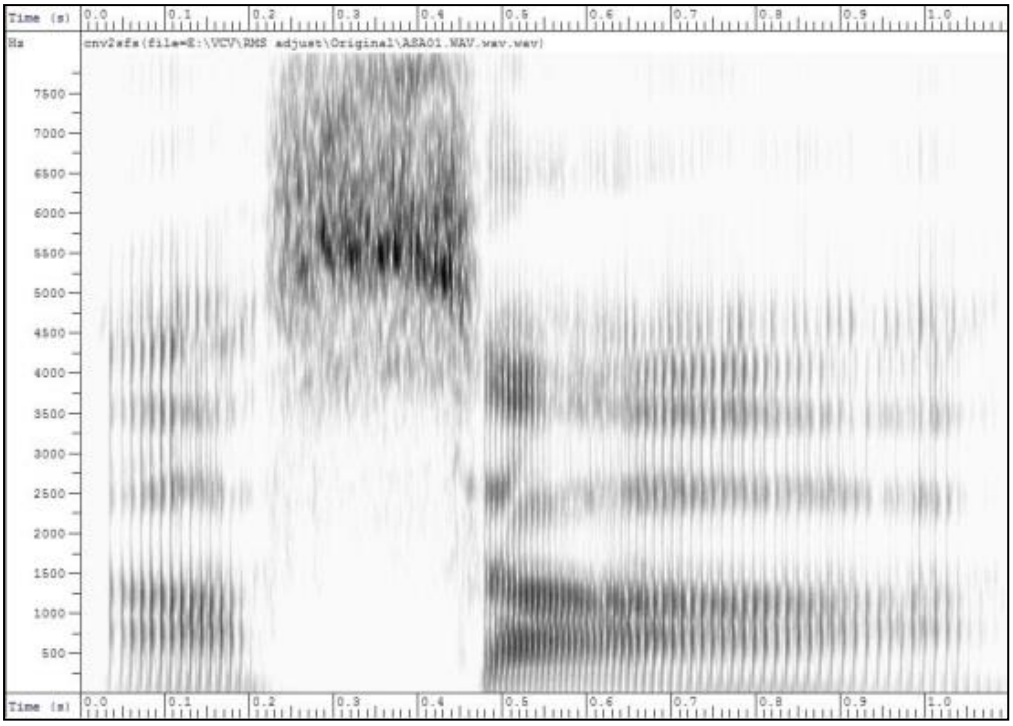


Figure 3. 1 Broadband spectrogram for the nonsense syllable /afa/. Frequency is displayed along the vertical axis, time along the horizontal axis and amplitude is represented by relative darkness.

Speech is often described as being made up of a series of individual sounds or phonemes. Each phoneme has a distinct spectrum that is a result of a sound source being modified by a variable filter (Fant, 1964). In speech production there are two main sources of energy: (1) a

stream of air from the lungs during exhalation and (2) forming a partial or complete constriction at some point along the vocal tract and as the air passes through the constriction it becomes turbulent. This creates an episode of aperiodic noise which can vary in duration depending on how abruptly the constriction is released; an example of this can be seen in the spectrogram of figure 3.1 for the sound /ʃ/ ("sh"), between 200 and 450 ms.

During voiced speech the vocal folds vibrate periodically as a result of the opening and closing of the glottis. The rate at which the vocal folds vibrate is known as *f*0 (fundamental frequency). *f*0 varies continually during speech and differs among speakers, with the most noticeable differences occurring between male, female and preadolescent speakers; average *f*0s of 120 Hz, 225 Hz and 265 Hz respectively (Fry and Blumstein, 1988). When a speaker changes the rate of their vocal fold vibration (and therefore *f*0), this is heard as a change in the pitch of their voice. Voiced speech sounds include vowels and some consonants, for example /d/ and /g/. When constrictions are formed in the absence of vocal fold vibration the speech sounds produced are voiceless. However, it is possible to form these constrictions and blockages whilst the vocal folds vibrate simultaneously; creating sounds such as /z/. Figure 3.2 shows the supralaryngeal vocal tract and the articulators that can be used to modify its overall shape.

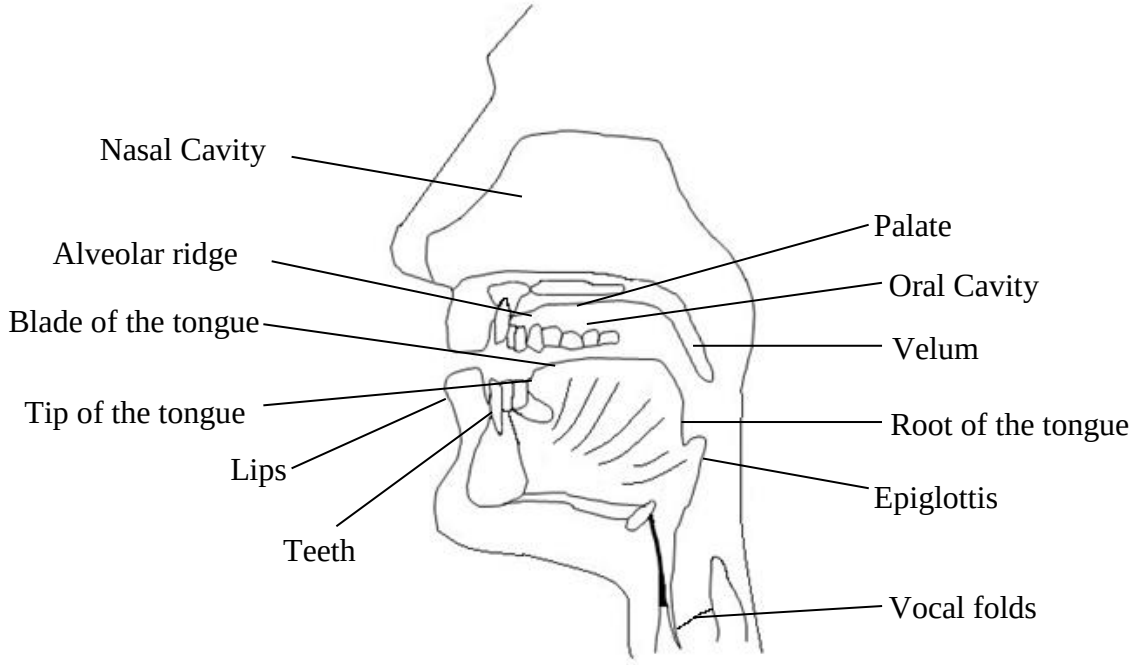


Figure 3. 2 The articulators of the human vocal tract

The overall shape of the vocal tract acts to dampen the transmission of certain frequencies (Greenberg and Ainsworth, 2004), resulting in peaks and troughs of energy in the output spectrum (Ashby and Maidment, 2005). Formants are regions of energy maxima, with the lowest formant (F1) having the highest amplitude. Formant structure can be varied by changing the height and location of the tongue and rounding of the lips during production (Pickett, 1999). Formants of different speech sounds are only affected minimally by changes in *f*0, therefore it is the shape of the filter that determines which sound is produced. Figure 3.3 demonstrates the effect of changing the resonant characteristics of the vocal tract on the source spectrum.

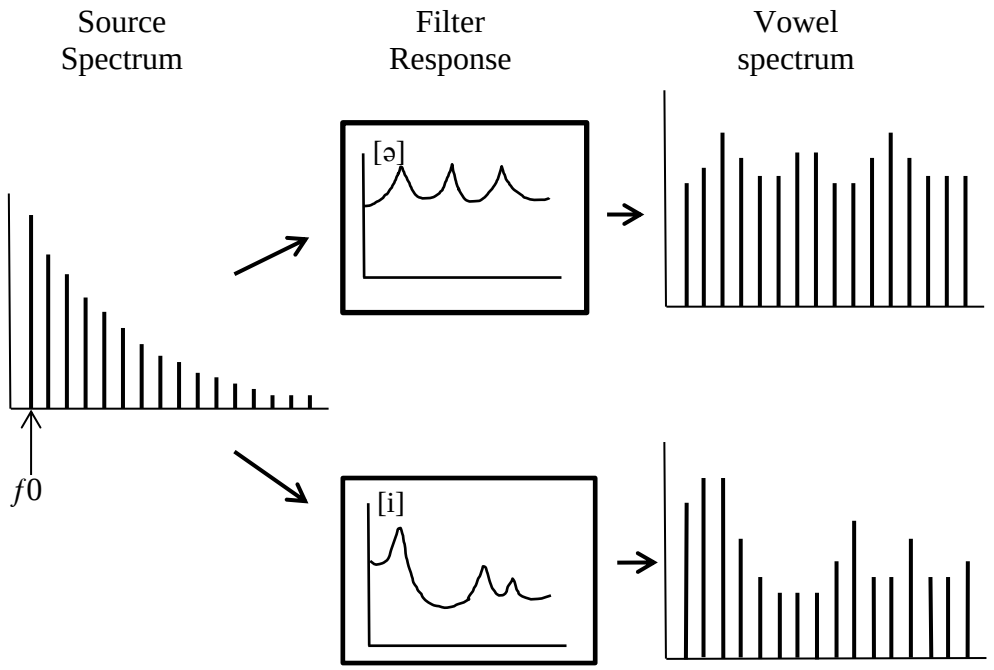


Figure 3. 3 The production of the vowels [ə] and [I] in accordance with the source-filter theory. The filter response of the vocal tract is different for each vowel, modifying the source spectrum (input) to give distinctive peaks and troughs in the final vowel spectrum (output). The frequencies of the resonance peaks in the output are not affected by the f0 of the source.

Consonants can be sub-divided into different classes based on their acoustic properties, which are a result of their production. These properties include voicing (voiced or voiceless), manner-of-articulation (how the constriction is formed) and place-of-articulation (where the constriction is formed). Tables 3.1 and 3.2 show the different manner and place categories for some consonant sounds in English.

Manner classification	Speech sounds
Stop	p, t, k, b, d, g
Fricative	v, z, ʒ, s, ʃ, f, θ, h
Nasal	m, n, ŋ
Approximant	r, l, w, j

Table 3. 1 Manner classifications for some English consonants

Place classification	Speech sounds
Bilabial	p, b, m
Labiodental	v, f
Dental	θ
Alveolar	d, t, s, z, n
Palato-alveolar	ʒ, ʃ
Velar	ŋ, k, g
Laryngeal	h

Table 3. 2 Place classifications for some English consonants

The acoustic properties of the speech signal that help a listener distinguish between different classes of sounds are known as acoustic cues. An overview of some of the acoustic cues for distinguishing between different classes of speech sounds are outlined in the sections below.

3.1.1 Vowels

There are typically many fewer vowels than consonants in the English language. Vowels are always voiced and are high in intensity. Formant frequencies (especially F1, F2 and F3) are very important for vowel identification (Denes and Pinson, 1993). Figure 3.4 shows the distribution of vowels in a two-dimensional space (known as the vowel space) in terms of their relative F1 and F2 frequencies. This is dependent on how open or closed the oral cavity is during phonation and the position of the tongue (front, middle or back). For example the vowel /i/ or “ee” is produced with the mouth relatively closed and the tongue near the front of the mouth, whereas the vowel /ɑ/ is produced with the mouth open and the tongue towards the back. The schwa sound, /ə/, is a very neutral vowel and is produced with the mouth about half way open and with no distinct tongue position (neither front nor back); it is found directly in the middle of the vowel space in figure 3.4.

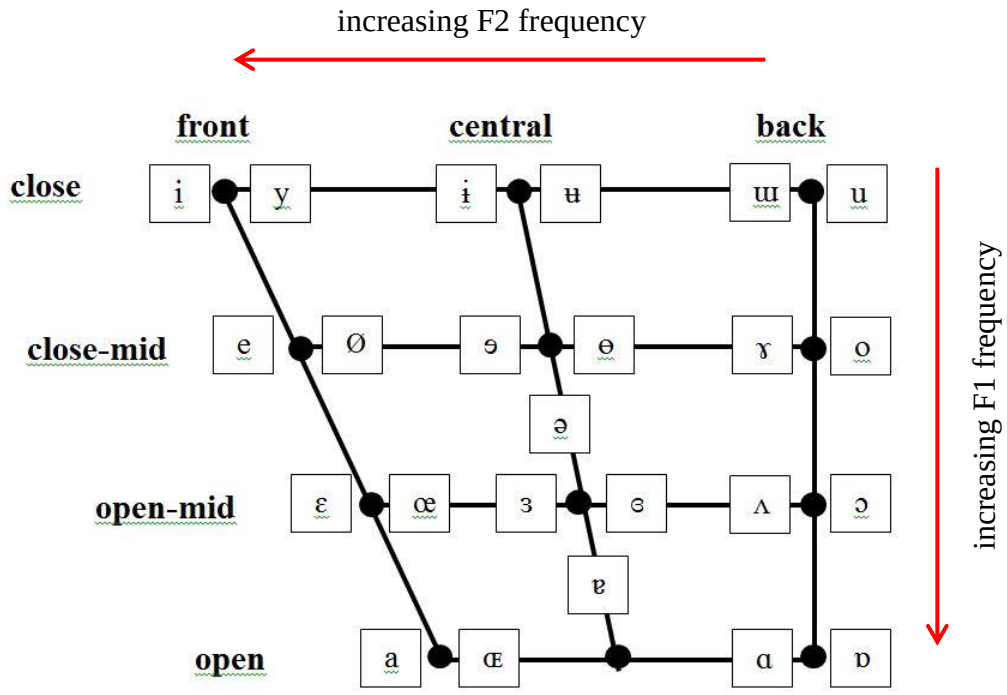


Figure 3. 4 The "vowel space"

3.1.2 Sonorants

Sonorant sounds can be split into approximants, which can be sub-divided into liquids (/l/ and /r/) and glides (/w/ and /j/), and nasals (/m/ and /n/). Sonorants have a similar formant pattern to that of vowels but are traditionally classified as consonants because they occur at the onsets and offsets of syllables, and their formant patterns do not reach a steady-state like vowels. They are also lower in amplitude than vowels (Wright, 1997). Approximants are produced with a partial constriction of the vocal tract and are characterised by formant transitions that are >100ms in duration. Formant transitions into and out of an approximants are an important cue to place of articulation, especially transitions associated with F2 and F3 (Reetz and Jongman, 2008). During the production of a nasal sound, the nasal cavity is open but the oral cavity remains closed (due to lowering of the velum). The nasal cavity acts as a larger, longer resonator, resulting in a lower resonant frequency and as such nasals are typically characterised by a low-frequency resonance (often referred to as the nasal murmur) at around 200-500 Hz. The higher frequency formants are heavily attenuated, resulting in anti-formants, and these act as a manner cue. The nasal sound /m/ tends to be lower in frequency and shorter in duration that /n/ (Denes and Pinson, 1993).

3.1.3 Plosives

Also referred to as stops, plosives are produced by a complete occlusion of the vocal tract, resulting in a brief period of silence in the spectrum followed by a short burst of noise (<100 ms); both these characteristics act as cues to manner of articulation). The duration of this period of silence acts as a cue to help distinguish between plosive sounds (/p, b, t, k, d, g/). Voice onset time (VOT), the time between the release of the burst and the onset of vocal fold vibration, is also an important cue for voiced plosives (/d, g, b/) (Benki, 2001; Raphael, 2005). Voiced plosives also tend to have a burst which is higher in intensity then unvoiced plosives (Borden et al., 2007). Place cues are derived from transitions of the lower formants as well as spectral tilt and frequency of the burst (Cooper et al., 1952; Stevens and Blumstein, 1978).

3.1.4 Fricatives

Sometimes referred to as continuants due to their prolonged period of aperiodic noise, cues to manner-of-articulation for fricatives are determined by the duration of the noise following the constriction (frication). They are produced by passing a continuous flow of air through a narrow constriction along the focal tract and can be distinguished from plosives by their relatively long rise time. Intensity, shape of spectra and F2/F3 transitions all act as cues to place of articulation (Reetz and Jongman, 2008). Cues tend to lie above 1 kHz and go spread to as high as 8-10 kHz for sounds such as /s/ (Wright, 1997). Voicing cues come from the presence/absence of low frequency energy in the spectrum. For unvoiced, weaker fricatives, F2 transitions are particularly important (Denes and Pinson, 1993).

The above sections have outlined the different types of information found within the speech signal and some the acoustic cues that can be utilised by listeners to identify and distinguish between different speech sounds. Many of these cues relate to changes in frequency and amplitude information over time (e.g. onsets and offsets and formant transitions) and as such this chapter will further explore and argue for the importance of spectral change in the perception of speech.

3.2 The relative importance of vowels and consonants in speech perception

Research has yet to determine whether vowels or consonants convey the most information in speech. One difficulty is that phonemes are not produced in isolation during running speech. In fact, cues for different speech sounds overlap due to the nature of the movement of the articulators from one sound to the next; this is known as coarticulation. It is therefore extremely difficult to split the speech signal into distinct vowel and consonant segments.

Cole et al. (1996) investigated the effect of removing information about vowels and consonants on speech perception for NH listeners. Using sentences from the TIMIT speech corpus (Lamel et al., 1986) they grouped sounds into consonants, vowels and what they referred to as "weaksons"; consisting of liquids, glides and nasals. This gave 20 consonants, 20 vowels and 12 weaksons. The TIMIT speech corpus has time-aligned phonetic transcriptions and allocated segment boundaries which assign half the formant transitions to vowels and half to consonants. In the first experiment, the authors replaced either the consonants or vowels with noise leaving the other sounds unaltered (the weaksons were left unaltered in both the consonant only and vowel only conditions). The authors selected 60 sentences which contained a fairly balanced number and duration of consonant and vowel segments. They tested speech perception scores for 35 participants listening to clean speech (original sentence), and to sentences which had either the consonants or vowels replaced by noise. Participants were able to listen to the sentences up to 5 times before a final response was taken (written down as best they could). They were scored both in terms of the number of correctly identified words and number of complete sentences correctly identified. The results showed that when only vowels were available almost twice as many words were recognised than when only consonants were available (around 47 percent correct with consonants only and 85 percent with vowels only). Sentence recognition differed by up to 40 percent, with recognition scores of around only 11 percent in the consonant only condition.

As the weaksons were available in both consonant, and vowel only conditions, the authors wanted to investigate whether leaving the weaksons unaltered was more beneficial for vowel or consonant recognition. The authors did not reuse any of the same participants for the second experiment but numbers were significantly smaller (13 participants). The authors used a similar noise replacement paradigm as in experiment one, but this time created three conditions; consonants only, vowels only and weaksons only. Participants were able to identify more words in the vowel only condition (56.5 percent) than in the consonant only (14.4 percent) and the weaksons only conditions. Weaksons produced a very low average score of 3.1 percent correct. In addition to this, participants were only able to correctly identify full sentences (i.e. all words in a sentence) in the vowel only condition. Interestingly, scores for the vowel only and consonant only conditions were significantly lower in the second experiment when weakson segments were removed, even though word identification with the weaksons was only 3.1 percent.

The authors considered the effects of coarticulation and questioned whether formant information in the vowel only portions was enough to provide information on neighbouring consonants. As speech cues are found across segment boundaries and are shared between the vowels and consonants it is possible that consonant boundaries are affected more than vowel boundaries when the other is replaced. In a final experiment, Cole et al. (1996) looked at the effect of increasing and decreasing vowel and consonant segments by expanding and reducing the vowel-consonant (V-C) boundaries by 10 msecs. They hypothesised that expanding boundaries (by 10 msecs) for consonant only segments would provide more information about neighbouring vowel segments and therefore improve recognition in this condition. They tested 14 new participants using the method from experiment one. The authors reported that results from this experiment were ambiguous as a similar effect was observed for both vowel only and consonant only conditions, however the effect was significant for both conditions.

Based on their results the authors argued that vowels are more important for speech perception because they contain more coarticulatory information at the V-C boundaries. However, results from these noise replacement studies need to be viewed with caution as they do not consider the phenomenon of "phonemic restoration"; whereby when a portion of the speech signal is removed and replaced with noise (particularly segments with a noise-like spectrum, e.g. fricatives), listeners are able to perceive the sentence as being intact and correctly identify the utterance, as long as the surrounding context remains intact (Warren and Sherman, 1974). Cole et al. (1996) make no reference to this phenomenon in the discussion of their results, which could, in part, explain why recognition was much higher for the vowel only condition; participants may have been able to partially resolve the consonant information.

Kewley-Port et al., (2007) looked to extend the work of Cole et al. (1996) to investigate the role of vowels versus consonants on speech intelligibility for elderly HI listeners with a typical loss. A typical age-related hearing loss is a bilateral symmetrical sloping high-frequency sensorineural loss, resulting in a reduction in the ability of these listeners to perceive information relating to high-frequency consonant information. This is often described by patients as a loss of the clarity of speech.

The authors used a similar experimental method to Cole et al. (1996) but classified weaksons as consonants (as phonetically this is how they are typically classified). This resulted in 32

consonants versus 20 vowels. Their first experiment looked to replicate the results of Cole et al. (1996) and tested a group of young, NH listeners in clean, vowel only and consonant only sentences (again from the TIMIT corpus). Results showed almost 100 percent correct word recognition for clean sentences, 74 percent for vowel only and 34 percent for consonant only sentences. These results are very similar to those obtained by Cole et al. (1996). Kewley-Port et al. (2007) argued that vowels may be more important for speech perception because they are more intense and longer in duration that consonants as well as containing a considerable amount of information important for neighbouring consonant identity (from formant transitions). Although a similar noise replacement paradigm was used in the experiment by Kewley-Port et al. (2007) as to that by Cole et al. (1996), they do at least consider the effect phonemic restoration and how it may be more important for consonant than vowel identification.

Kewley-Port et al. (2007) went on to test a group of elderly HI listeners and a new group of NH listeners (as a baseline for comparison) in the same task, but at a louder listening level (95 dB SPL, compared with the 70 dB SPL of the first experiment). This level was chosen so as to ensure that all segments were within the HI listeners' audible range, without being uncomfortably loud. The 16 recruited HI participants all had a long-standing, bilateral, symmetrical, moderate sensorineural hearing loss. The authors state that they tried to ensure participants were only included if their hearing loss would give some impairment for vowels at 2000 Hz (considered to be a minimum of 35 dB HL) and good audibility for consonants between 2000 and 4000 Hz. Although no justification for these criteria was given, it can be assumed that it was chosen to ensure some impairment of F1 and F2 information for vowels. However, formant transitions have also been shown to be important for the identification of consonants.

Results showed that the HI listeners and NH listeners scored comparably in the clean sentence condition. For the noise-replaced sentences, as was predicted, the NH group performed significantly better than the HI group. Both groups scored better for the vowel only condition than for the consonant only condition, with the HI group scoring almost double in the vowel only condition; 40 percent words correctly identified versus 20 percent for consonant only. The distribution of scores from the HI group was not reported, however, one might predict greater variation in scores for the HI group than the NH group.

Figure 3.5 shows the audiometric data for each of the 16 HI listeners in the study. Below this, figure 3.6 demonstrates the "speech banana"; the distribution of English speech sounds across level (dB HL) and frequency. From figure 3.5 it can be seen that there is a large variation in the audiometric configuration of the participants, this is to be expected as it is extremely difficult to recruit a homogenous group of HI subjects. Although the authors give some justification as to the hearing threshold criteria they chose, figure 3.6 shows that this covers only a very narrow region of the speech signal. This region may cover information pertaining to formant transitions but considering the speech banana in figure 3.6, there is also a lot of information found below this region, especially relating to vowels and sonorants. With a test level of 95 dB SPL and lower threshold of hearing in the low frequencies it could be expected that vowel and sonorant sounds would be more audible than consonants.

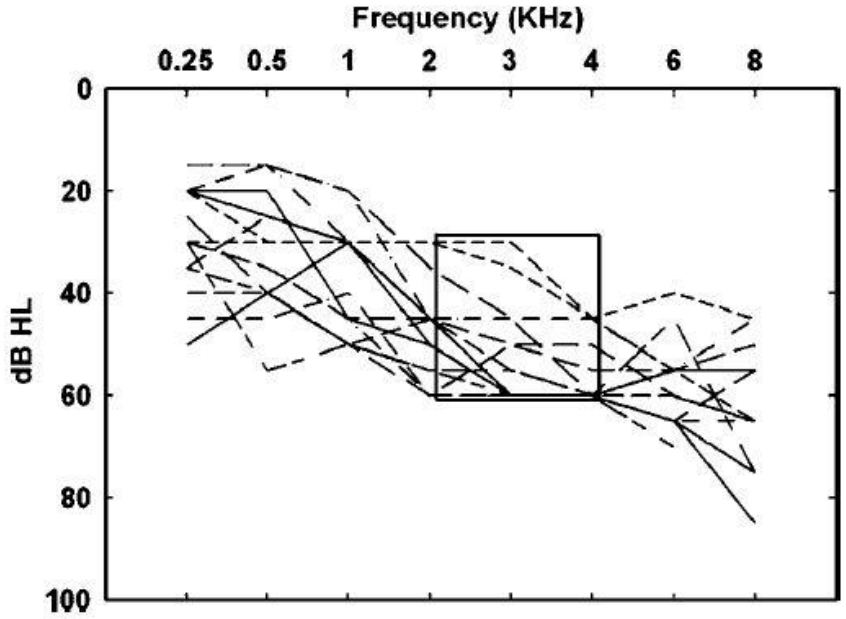


Figure 3. 5 Audiometric data for the HI participants in the Kewley-Port et al. study. The box indicates the hearing threshold criteria required (Kewley-Port et al., 2007: 2). Reproduced with permission, copyright Acoustic Society of America.

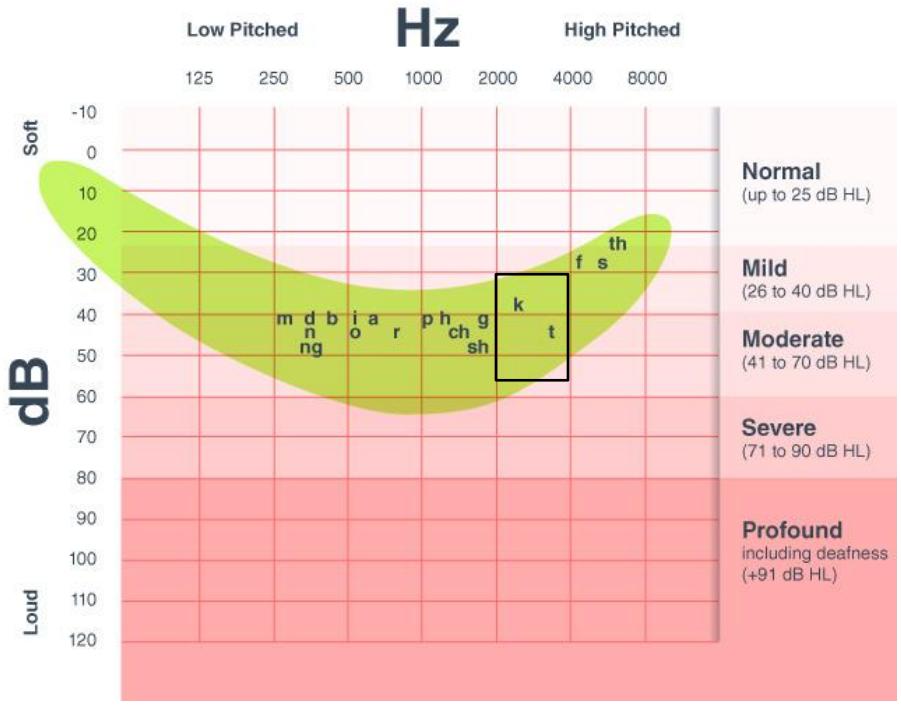


Figure 3. 6The Speech Banana. This demonstrates the distribution of the main speech sounds in English in terms of their frequency and intensity when spoken at a normal conversational level. The black box indicates the hearing threshold criteria as shown in figure 3.5. From http://www.hearinglikeme.com/facts/what-hearing-loss/now-hear. Audiogram image courtesy of HearingLikeMe.com and Phonak.

The overall conclusion that even HI listeners rely more on vowel information for good speech perception seems counter-intuitive, as with a typical high frequency hearing loss, vowel information is going to be more audible than consonant information. The authors argue that their findings would appear to support that amplification for patients with these types of losses should ensure good transmission of low frequency, vowel information. However, in reality, patients prefer less amplification in the low-mid frequencies than in the higher frequencies; as is demonstrated in the gain differences between the NAL-NL1 and NL2 fitting formulas (Keidser et al., 2011). The authors should therefore have extended their study to look at the relative importance of vowels versus consonants for aided HI subjects.

Owren and Cardillo (2006) investigated the relative importance of vowel and consonant identification for speaker and meaning discrimination. Their experiment looked to test the hypothesis that participants (listening to individual words) would be better able to

discriminate talker identity from vowel only information, and meaning discrimination would be better with consonant only information. The most commonly used vowels in speech are always voiced and therefore have a robust harmonic structure and hold important information about *f*0; this means they are important for providing supra-segmental information, including cues which help to discriminate between talkers.

The authors produced words with vowel only and consonant only segments. Segments were replaced with silence but it is not clear from their method whether this was done manually, by visual inspection of the spectrogram for each word, or by a computer programme. Obvious formant transitions were also removed between the segments which resulted in vowel segments consisting only of their steady-state centres. However, the authors were not able to completely remove information about adjacent consonants from vowel segments.

One hundred and twenty eight words were recorded from eight male and eight female speakers, all of a similar age (five year spread). The words were presented in pairs and had either the same/different meaning, or were produced by either the same or different speakers. The authors tested 40 participants in total, with half completing the speaker identity task, and half completing the meaning task. In both conditions participants were tested listening to intact, vowel only and consonant only words. The overall findings did not show any effect of listener sex and found that when listening to intact speech, participants were better at identifying the meaning (same versus different word) than talker identity. For the consonant only and vowel only conditions, identity was more difficult that meaning, however, listeners were better at discriminating between talkers in the vowel only condition and performed better in the meaning task with consonant only information.

In their method, the authors did state that almost twice as many of the words began with consonants than began with vowels (82 and 46 respectively). They also differed in the number of syllables (48 were two syllables long, 44 were three syllables long, 23 were four syllables, and three were five syllables). Further analysis in the meaning condition showed that results with the consonant-only words were not affected by whether the word started with a vowel or a consonant. However, in the meaning task, listeners seemed to score highest with vowel only words starting with a vowel and lowest when a vowel only word started with a consonant. Unfortunately, the authors did not do any further analysis to look at the effect of syllable duration; for example, does talker discrimination improve with increasing syllable duration?

Unlike in previous studies, Owren and Cardillo (2006) suggest that vowels and consonants play different roles in speech identification, with vowels providing more information about talker identity and consonants providing more information about word meaning. Although fundamentally a different task, results from this study could be seen to contradict those of Cole et al. (1996) and Kewley-Port et al. (2007) who argue strongly for vowel superiority in speech recognition. In order to form a better comparison between these studies, it would be benefitial to extend the work of Owren and Cardillo to look at sentence length material. It may also be useful to compare results for noise replaced and silence replaced segments, as both experimental paradigms have their limitations. In particular, noise replacement studies must deal with the issue of phonemic restoration for consonant sounds, and a better understanding of the effects of forward and backward masking on the intelligibility of consonant and vowel segments is also required.

The aforementioned studies have so far highlighted the possible importance of information held at vowel-consonant (V-C) boundaries. To investigate this further Fogerty and Kewley-Port (2009) looked to examine the effect of varying acoustic information at these V-C boundaries and determine their relative importance to vowel and consonant recognition. Their first experiment involved shifting the pre-determined V-C boundaries of the TIMIT sentences. Segmentation boundaries of the TIMIT sentences have been defined by highly salient and abrupt acoustic changes (to mark phoneme boundaries for stops) and formant transitions have been split in half during slow periods of change for liquids (Fogerty and Kewley-Port, 2009: 849).

Twenty normal hearing listeners were tested using the same TIMIT sentences as used in the Kewley-Port et al. (2007) study. As in previous studies, consonant and vowel only conditions were created, and removed segments were replaced with noise. V-C boundaries were manipulated for vowel and consonant segments only and not between consonant and consonant or vowel and vowel sections. The V-C boundaries were shifted so as to increase the amount of vowel information included in the consonant only condition, and to decrease formant transition information in the vowel only condition (vowel centre information only). Although only a short familiarisation task was given, once again the authors were able to replicate similar findings of vowel superiority as in Cole et al. (1996) and Kewley-Port et al. (2007).

Results from this experiment showed that speech recognition in the consonant only condition improved as more transitional information was added (i.e. formant trajectories to and from surrounding vowels), however, the same benefit was not observed for the vowel only condition. These results would seem to suggest that transitional information is less important for good vowel recognition than for consonant recognition and given that these transitions are generally short in duration, this may be used to argue for the importance of abrupt spectral changes in the perception of consonant speech sounds. However, these results are in conflict with those of Strange et al. (1983) who demonstrated that it is possible for listeners to identify vowels from dynamic information found at V-C boundaries. Indeed, in an earlier study, Strange et al. (1976) found that vowels are more accurately identified when in a V-C context rather than isolated, suggesting that they depend on coarticulatory information for correct identification.

Fogerty and Kewley-Port (2009) also looked at the overall duration of consonants and vowels across the TIMIT sentences used in the experiment. They found that consonants made up 55 percent of the total sentence duration, yet there were almost twice as many consonants compared with vowels in each sentence. Consonants often appeared in strings (74 percent of all consonants) and were of a shorter duration than vowels. In contrast only 12 percent of vowels appeared in strings. It is therefore surprising that, although much of the sentence is comprised of consonant information, according to several of the studies presented, vowels give more information on intelligibility.

Ultimately, these studies highlight that acoustic cues for vowels and consonants overlap at their boundaries, meaning that segmenting them is difficult. Although considering the speech signal in terms of vowels and consonants and the amount of information they carry may be convenient, it may also be irrelevant, as the overwhelming conclusion from these studies is that the important information is found at V-C boundaries. Continuing along this line of thought, subsequent sections in this chapter explore the role of information at V-C boundaries further, and more specifically, information held by rapid spectral changes in these regions; this forms an argument for a theory of speech perception that moves away from the concept of discrete segmentation.

3.3 Sensitivity to change in the perception of speech

Dynamic spectral changes, such as those found at V-C boundaries, may be the key to identifying important regions of the speech signal and the following sections in this chapter

explore the evidence for this hypothesis further. Furui (1986) also gives evidence for the importance of spectral transitions in speech perception, particularly for vowel identification and argues that the steady state portions of speech provide little information.

3.3.1 Neurophysiological responses to spectral change

Investigations into the neural representations of speech have shown that the auditory system is particularly responsive to change; this is manifested by increased firing rates of auditory nerve fibres (ANFs) at the onsets of syllables and during transient events such as bursts (Delgutte, 1980; Delgutte and Kiang, 1984; Palmer and Shamma, 2004). Delgutte (1997) explored how recordings from single auditory neurons in response to speech can enable us to follow the neural representations of the signal through each stage of auditory processing. It has been found that rapid changes in amplitude and spectrum (related to onsets and offsets) are clearly represented and readily observed in the discharge patterns of ANFs. Delgutte (1997) illustrates this by showing how rapid amplitude and spectral changes which are evident in the example broadband spectrogram for the sentence 'Joe took father's green shoe bench out' (figure 3.7) are also apparent in the corresponding neurogram (figure 3.8). The neurogram uses post-stimulus time histograms (PSTH) to display the average discharge rate of a small number of ANFs following the onset of the stimulus.

These moments of rapid amplitude change in different spectral regions are indicated by the arrows on the neurogram and spectrogram. The filled arrows mark events where fibres with low characteristic frequencies (CF) show an abrupt increase in their discharge rate and the spectrogram shows a similar abrupt increase in energy in the low-frequencies. These events can be linked to times of transition from an obstruent to a sonorant segment, and the onset of voicing of stop consonants (Delgutte, 2007: 6). Similarly, the empty or open arrows indicate events where high CF fibres are observed as having a rise in discharge rate which is mirrored by a rapid increase in energy in the high frequency region of the spectrogram (Delgutte, 2007: 6). This happens at times of sonorant to obstruent transitions and at the onset of the release burst for stop consonants. The ovals on both the neurogram and spectrogram show the movement of F2.

The neurogram in figure 3.8 also shows that, following the prominent peaks there is a gradual decay in the instantaneous discharge rate of the neurons, where responses to stimuli following the peak are depressed; this is known as *adaptation*. Delgutte (1997: 7) highlights the four main roles of adaptation in encoding the speech signal in the auditory nerve:

1. Peaks in the discharge rate point to spectro-temporal regions which are rich in phonetic information
2. Increase the temporal precision with which onsets are represented
3. Enhancement of spectral contrast between successive speech segments
4. Encoding of the phonetic contrasts based on characteristics of the amplitude envelope

ANFs appear to have a robust response to acoustic events relating to rapid spectral changes and enhance them, making them more prominent. Indeed Delgutte (1997) asserts that 'acoustic characteristics of speech that are phonetically the most important are prominently and robustly encoded in neural responses'.

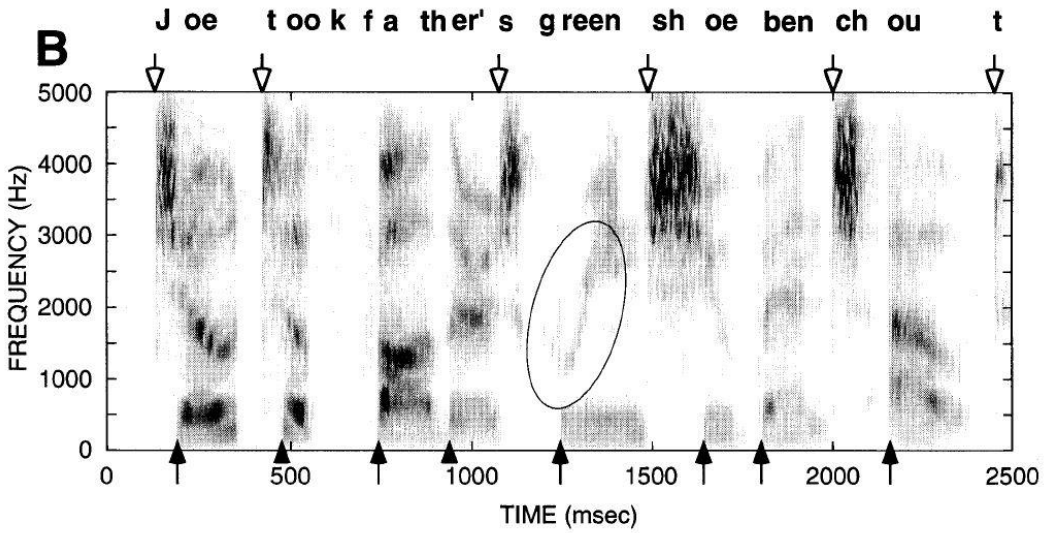


Figure 3. 7 Broadband spectrogram for the sentence "Joe took father's green shoe bench out". From Delgutte (1997). Reproduced with permission, copyright Wiley-Blackwell.

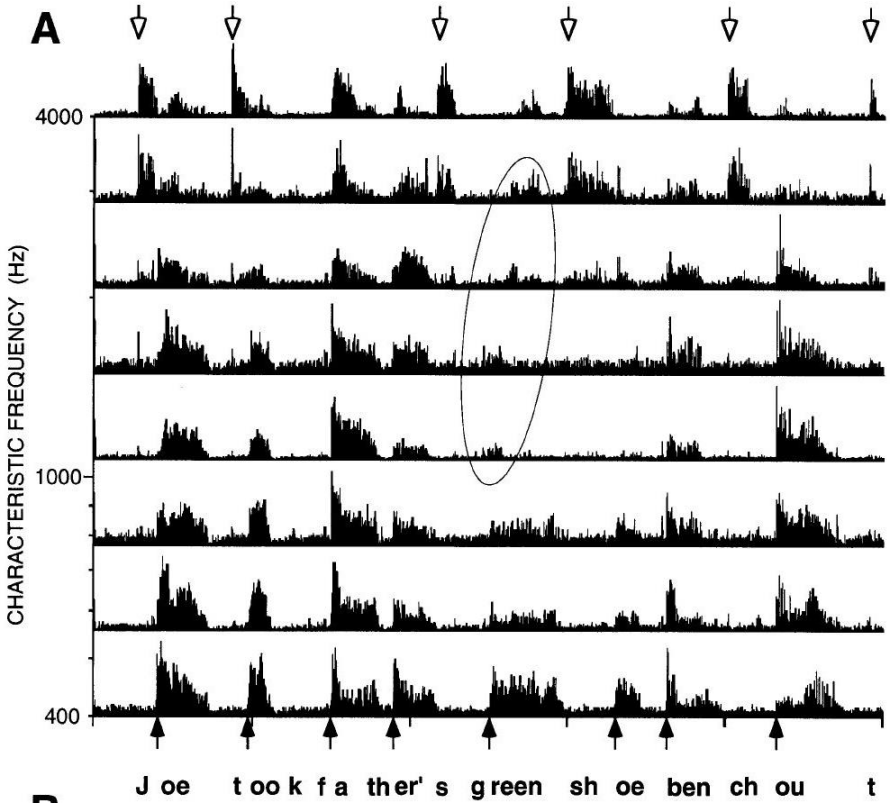


Figure 3. 8 Neurogram for the same utterance "Joe took father's green shoe bench out", as produced by a female talker. The traces represent the average post-stimulus-time histogram for 2-7 ANFs. From Delgutte (1997). Reproduced with permission, copyright Wiley-Blackwell.

Christovich et al. (1982) suggested a model of auditory processing which is devised of two systems working in parallel. The first of these is a tonic system which gives a constant spectral representation of the stimulus and the second is a phasic system. The phasic system is closely linked to adaptation in ANFs as its role is to detect the acoustic transients (i.e. onsets and offsets) in different frequency bands. The phasic system is described as having two major functions: provide temporal cues which help with identifying phonetic segments, and to give a set of "pointers" at times which are of particular importance using these to guide sampling of the output of the tonic system. Cells at central locations show sensitivity for a particular amplitude modulation frequency, for example, the inferior colliculus (midbrain) responds best to modulations between 10-300 Hz and cortical neurons to 3-100 Hz. These cells will not phase-lock to the pitch period but should respond well to the major changes in amplitude and spectrum which contain information relating to manner-of-articulation.

Morris et al. (1991: 1442) used Mutual Information (the measure of two random variables' mutual independence) to measure the distribution of phonetic information in relation to Onset and Offset events in the signal (measured from spectrograms). They found that information relating to place of articulation for plosives is concentrated equally at On and Off positions, but that information for vowel identification is found close to the point of maximum energy for the corresponding vowel. This compliments data from ANF discharge pattern studies, and supports the notion that onset and offset events are potentially important for efficient data reduction and the organisation of speech processing at higher levels.

It is important to note that these rapid changes in spectrum and amplitude are not the only events in the speech signal that ANFs respond to, as although such "pointers" identify regions rich in information, cues for vowel identity, and also place of articulation for consonants, rely more on detailed spectral features related to the fine structure of the signal. Miller and Sachs (1983) demonstrated that formants, which are important acoustic cues, are also well represented in the discharge patterns of neurons. Nonetheless, these results show that, even at more central locations, gross changes in amplitude and spectrum are well represented, suggesting that spectral change plays an important perceptual role in the recognition of speech.

3.3.2 Cochlea scaled entropy

It is known that perceptual systems respond primarily to change (Kluender et al., 2003) and this has been demonstrated in the neural discharge patterns of ANFs (Delgutte, 1997). Shannon's Information Theory (1948) posits that there is no new information when events do not change or they are predictable. Biological systems respond to change because this is where important information is held. Stilp and Kleunder (2010) therefore suggest that we move away from linguistic assumptions and the traditional view of speech separated into vowels and consonants and instead base importance of segments on a measure of how much the signal changes as a function of time. As biological systems respond to relative change, rather than absolute values (i.e. for amplitude or frequency) this may help to account for how the auditory system is able to deal with the issue of coarticulation (Kleunder et al., 2003).

Despite arguing that speech perception should not be thought of in terms of individual phonemes (vowels and consonants), Stilp and Kleunder (2010) do still maintain that some parts of speech are more information bearing than others. They developed a method for measuring the potential information in a speech signal, based on the relative (un)predictability of successive spectral slices; this is known as cochlea-scaled entropy (CSE). For this measure, more CSE (relating to

more uncertainty/unpredictability) is equivalent to more information. In their first experiment, Stilp and Kleunder (2010) replicated the results of Cole et al. (1996), Kewley-Port et al. (2007) and Fogerty and Kewley-Port (2009) using a noise replacement paradigm, indicating vowel superiority in speech perception. The authors then went on to conduct a similar experiment, whereby vowel and consonant distinctions were ignored and instead segments of either high CSE or low CSE were replaced with noise. Sentences which retained high CSE provided higher speech intelligibility (Stilp and Kleunder, 2010).

Interestingly, in the high CSE replacement condition, the authors report that more vowels than consonants were replaced. In fact, fricatives and stops were replaced least often in the high CSE condition. This is a surprising finding as the nature of the production of these speech sounds (constriction of the vocal tract and sudden release of energy) would suggest that the resultant rapid spectral changes would hold more uncertainty and therefore, more information (and greater CSE). The authors suggested that vowels and other sonorant sounds may have greater CSE because more weighting was given to the low frequency prominences (e.g. F1 and F2). However, considering the theoretical background behind the measure of CSE, it seems extremely unlikely that entire vowels or consonant sounds were replaced in this experiment. It is not clear from the reported results whether V-C boundaries were replaced more in high or low CSE conditions.

Nonetheless this study would seem to suggest that CSE is able to predict intelligibility of speech signals; with poorer listener performance in conditions where high CSE segments were replaced by noise. This study appears to suggest the importance of spectral change in the perception of speech, as change signals new and important information within the speech signal.

3.3.3 Lexical Access from Features model of speech perception

This chapter has so far argued for the importance of rapid amplitude change occurring in different spectral regions, often found at CV boundaries, for good speech recognition. Stevens (2005) presents a theory of speech perception that incorporates these important spectral events which is outlined in the following section.

Chomsky and Halle (1968) theorised that words can be described in terms of sequences of discrete segments, with each segment consisting of bundles of binary distinctive features. Distinctive features act as “phonological markers” that are used to make phonemic distinctions (Diehl and Lindblom, 2004). They are considered in a binary nature as a listener is able to assign a + or – value based on the presence or absence of a particular property. It could be argued that distinctive features are building blocks of the traditional “phoneme”,

with a change in the + or - value of a given feature resulting in a different word being uttered or heard. This change in the feature value is the consequence of a change in the place-of-articulation, manner-of-articulation or voicing characteristics of a speech sound. Stevens (2002) built upon the work of Chomsky and Halle (1968) to develop a more complete model of speech perception, based upon distinctive feature theory; this is known as the Lexical Access from Features (LAFF) model of speech perception. Figure 3.9 demonstrates how the distinctive feature bundles of an utterance can be used to match those of words stored in the lexicon.

Stevens (2002; 2005) splits distinctive features into those which are considered to be "articulator-free" and those which are "articulator-bound". Articulator-free features are not constrained to any particular articulator but instead refer to the type of constriction formed and the resulting acoustic consequences (closely related to manner-of-articulation); classifying segments as vowels, consonants or glides. Stevens (2002) asserts that the presence of a vowel, consonant or glide segment within an utterance is signalled by "acoustic landmarks"; specific acoustic events around which the acoustic cues required to determine feature values are concentrated (Slifka et al., 2004). Landmarks may therefore be considered as acoustic events which signal the presence of new linguistic information in the speech signal and listeners can exploit these events to do further processing of the signal, typically within a few tens of milliseconds of the landmark event. This links closely with the concept of CSE, whereby only regions of change are considered important and signal new information in the speech signal.

Acoustic landmarks are moments of rapid amplitude changes occurring in a particular frequency region which can be used to distinguish between vowel, sonorant and consonant segments. Vowel, consonants and glides are signalled by different landmarks: vowels by a peak in low-frequency amplitude in the region of F1, consonants by an abrupt spectral discontinuity (rapid amplitude change) at the onset and release of the constriction, and glides by a minimum in low-frequency amplitude (in the absence of any spectral discontinuity). The arrows on the spectrograms in figure 3.10 show the position of vowel and consonant landmarks (a) and glide landmarks (b).

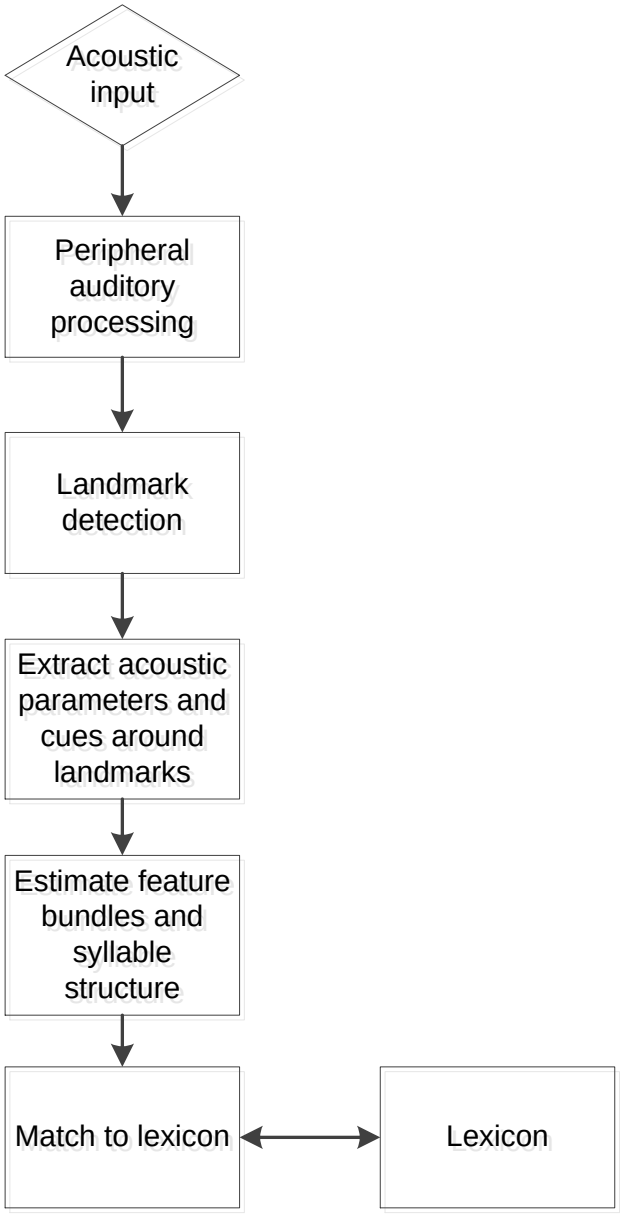


Figure 3. 9 Schematic of the processes involved in the LAFF model of speech perception. After Stevens (2005).

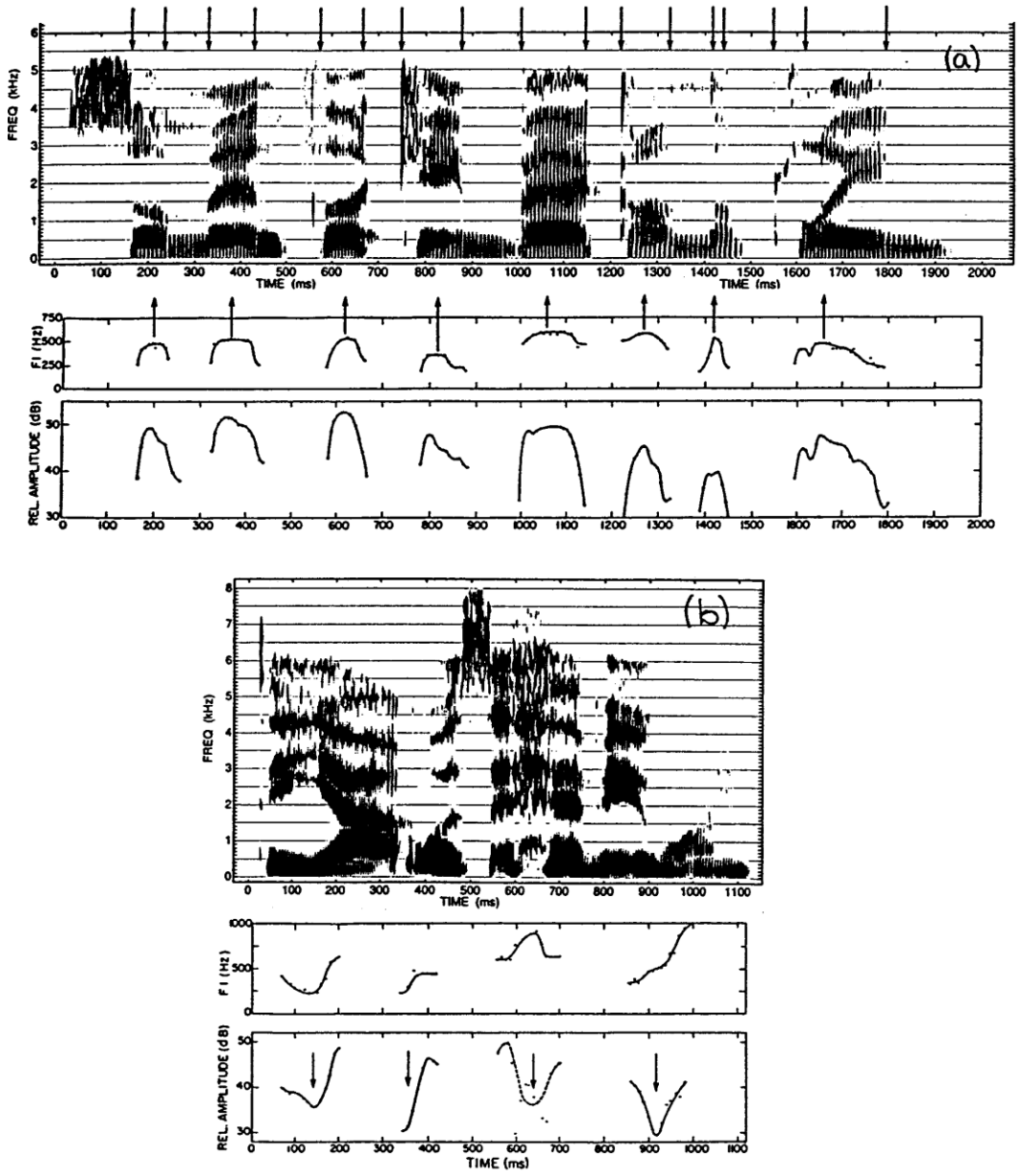


Figure 3. 10 (a) is spectrogram for the sentence "Samantha came back on the plane", produced by a male speaker. Immediately below is a plot of F1 frequency versus time, during vocalic regions. Below this is a plot of the amplitude of the F1 prominence, during vocalic regions. The arrows at the top of the spectrogram indicate the presence of a consonantal landmark, whereas the plots below indicate the presence of a vowel landmark. (b) is a spectrogram for the sentence "The yacht was a heavy one", produced by a female speaker. The plot immediately below shows the F1 frequency in the vicinity of the glides with the amplitude of the F1 prominence in the plot below this. The arrows in (b) identify glide landmarks in regions of amplitude minima. From Stevens (2002). Reproduced with permission, copyright Acoustic Society of America.

Articulator-bound features specify which of the articulators are used to create the vowel, consonant or glide segment (Stevens, 2002; 2005), and are closely related to place-of-articulation and voicing. They also specify the shape and position of the active articulators. Movement of the articulators produces distinctive acoustic properties that can be used to distinguish between speech sounds. Figure 3.11 shows how the articulators can be classified into three groups (Stevens 2002; 2005): those that affect the configuration of the oral cavity (blue boxes), those that control the shape of the vocal tract in the laryngeal and pharyngeal regions (red boxes), and finally, those which describe how stiff or slack the vocal folds are (green box).

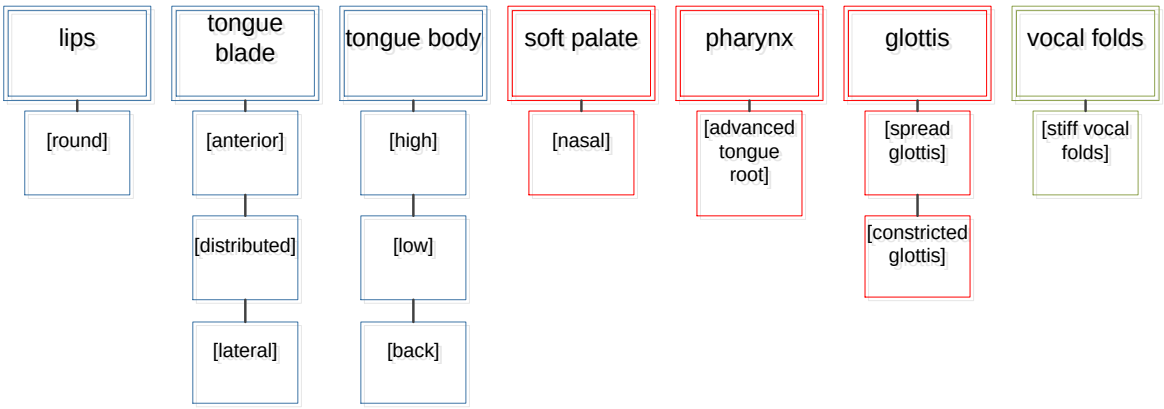


Figure 3. 11The three groups of articulators and the phonemic distinctions that they make.

The features listed in the blue boxes of figure 3.11 relate to the active (or primary) articulator used to produce consonant segments. These are the articulators that form the constriction and resulting spectral discontinuity which create consonantal landmarks. The feature [stiff vocal folds] relates to whether a segment is voiced or not. As it does not specify which articulator is forming the constriction it is therefore considered a secondary articulator. The values for the distinctive features of some consonants can be found in table 3.3 and help to demonstrate more clearly how a feature bundle is formed.

The major distinguishing features for vowels are those that specify the positions of the tongue-body ([high], [low] and [back]), affecting the formant pattern in the region of the vowel landmark. Although typically glides have fewer distinctive features than vowels and consonants, they can use any of the features involving the first six articulators (figure 3.11).

The features arising from manipulation of the tongue-body are important for distinguishing the sounds /w/ and /j/, whilst [spread glottis] is the only feature used to specify /h/. Articulator-bound features for some English vowels and glides are given in table 3.4.

	b	d	g	p	f	s	z	ʃ	m	l
continuant	-	-	-	-	+	+	+	+	-	-
sonorant	-	-	-	-					+	+
strident					-	+	+	+		
lips	+			+	+				+	
tongue blade		+				+	+	+		+
tongue body			+							
round	-			-	-				-	
anterior		+				+	+	-		+
lateral									-	+
high		+								
low		-								
back		+								
nasal									+	-
stiff vocal folds	-	-	-	+	+	+	-	+		

Table 3. 3 Articulator-free and articulator-bound features for some consonants in English. [continuant], [sonorant] and [strident] are examples of articulator-free features as they are used to further classify consonants as stops, fricatives and sonorants.

	i	ɛ	æ	a	ʌ	u	w	j	h
high	+	-	-	-	-	+	+	+	
low	-	-	+	+	-	-	-	-	
back	-	-	-	+	+	+	+	+	
round				-	-	+	+		
tense	+	-			-	+	+	+	
spread glottis							-	-	+

Table 3. 4 Articulator-bound features for some vowels and glides in English.

The process of speech perception may then require a local analysis of the speech signal in which ALs are first located in order to identify vowel, consonant and glide segments. Further processing then takes place around these landmark times to extract the cues required for feature identification, thus ALs can be considered as providing a "road map" for further acoustic analysis of the speech signal (Slifka et al., 2004). In a model incorporating landmark

detection as an initial stage, if landmarks are not correctly identified then errors may be made in further stages of processing, leading to inaccurate word identification; this highlights the potential importance of ALs for good speech perception.

Stevens (2002; 2005) attempts to deliver a comprehensive method for describing the speech signal, however, his theory does not clearly explain how the auditory system is able to use the "cues" located around landmarks to perform the binary classifications required to determine the sequence of features for each word. Nor is it evident how the brain is able to determine which feature values belong to which word; the LAFF model does not successfully solve the issue of segmentation. Hawkins (1999) asserts that the LAFF model does not yet adequately explain the role of suprasegmental information (intonation, prosody etc.) which is known to play an important role in the understanding of the speech utterance.

A further criticism of models of speech perception based on distinctive feature theory (such as LAFF) is that they require more empirical evidence to support their claims. The main assumptions of distinctive feature theory can be summarised as:

(1) There is a universal set of features (i.e. common to all languages)

(2) This set of universal features is relatively small (20 or so)

(3) The acoustic speech signal contains invariant properties

(4) There is a direct relationship between the invariant properties of the signal and the features

These assumptions have been made primarily on the basis of evidence obtained from the physical limitations of the motor and auditory systems involved. This is not necessarily an unreasonable link to make, as clearly they must play some role in the perception of speech. However, a better understanding between the production mechanism and the resulting acoustic output is really required. Feature theory, if nothing else does appear as a useful tool for classifying and distinguishing different sounds within languages. Whether it plays a role in the perception of speech is still to be determined.

The term feature is referring to something abstract- just as the phoneme has been used in the past to help phoneticians and speech researchers better understand the construction of speech (and language). This has biased our view on the perception of speech, as it is always tempting to describe things in these terms and assume that this is how the auditory system and higher

functions involved perform classifications. It is possible to break the speech signal down into its component parts of frequency, amplitude and phase and these are the variables that the auditory system deals with. Humans are able to manipulate their production mechanism (vocal tract) generating changes in the properties of the resulting waveforms, in order to create distinctions and convey different meanings. Fant (1973) asserts that the perception of speech based on detection of distinctive features does not rely on the strict reconstruction of the articulation/production of the utterance.

This thesis will not attempt to directly determine the feasibility of such a model based on distinctive features; indeed some speech researchers do not necessarily agree that words are stored and analysed in the form of a discrete set of segments (see for example Port, 2010). However, the LAFF model does attempt to account for variation in the production of speech by suggesting that the perception of speech sounds within an utterance does not rely on the specific acoustic patterns of the waveform. In particular, landmarks appear to be fairly robust to the issue of variability as they do not rely on specific spectral characteristics but rather gross-spectral changes. As such, there is no "pure" example of how a particular consonant or vowel landmark should look, rather, listeners need only be able to identify the fact that a spectral peak or valley has occurred in the low frequency portion of the signal or that an abrupt change in amplitude has occurred in the mid-high frequency region of the signal. This strongly ties in with evidence from the aforementioned neurophysiological studies on the importance of rapid spectral and amplitude changes for speech perception.

3.4 The role of acoustic landmarks when listening in noise

Although studies discussed in earlier sections of this chapter have argued that important information in speech may found around regions of spectral change, or "acoustic landmarks", the role of these landmark events has not been explored for speech perception in noise. Li and Loizou (2008a) looked to determine the contribution of obstruent consonants (stops, fricatives and affricates), and their corresponding ALs, to speech recognition in noise. The authors chose to focus on obstruent consonants because important cues for place-of-articulation, such as spectral tilt and burst frequency of stop consonants (Stevens and Blumstein, 1978), are significantly altered by the addition of noise (Parikh and Loizou, 2005). Li and Loizou (2008a) measured percent correct scores for 13 NH subjects listening to sentences and VCV stimuli corrupted by 20-talker babble noise at -5 and 0 dB SNRs. It was observed that when noise-corrupted obstruent consonants were removed and replaced

with clean versions of the same consonant (with sonorant segments left corrupted), speech recognition scores significantly improved. This was true for both SNRs but the effect was larger in the more difficult listening situation (-5 dB SNR); this may be because participants were reaching ceiling performance in the clean consonant condition at 0 dB SNR.

Li and Loizou's (2008a) results suggest that providing access to clean obstruent consonants can help to improve speech intelligibility. However, the authors noted that during sections containing the clean obstruent spectra, subjects had access to the full spectrum and therefore a number of different cues, including F1 and F2 transitions, voicing information and also better access to ALs. Interestingly, even when the consonant segments were completely removed from the signal and replaced by silence, subjects were still able to score above chance level in the VCV task. The authors attributed this to the fact that listeners use cues from vowel formant transitions to gain information about consonant identity. This supports the conclusion from section 3.2 that information held at V-C boundaries is important for the perception of consonants.

A second study was conducted to determine which temporal and/or spectral cues were contributing most to the improvement seen in experiment one (Li and Loizou, 2008a). This was achieved by systematically varying the amount of spectral and temporal information present in the clean obstruent spectra. Only partial clean spectral information was available, spanning a range of 0-F_c Hz, where F_c was equal to 500, 1000, 2000 or 3000Hz (frequencies higher than F_c were left corrupted by noise). The aim of using these frequency ranges was to assess the relative contributions of low and high-frequency landmarks to the improvement seen in speech intelligibility during the first experiment. However, the terms 'high-' and 'low-frequency landmarks' are not clearly defined within this study, though one can assume that high-frequency landmarks (>2000 Hz) are those that signal fricatives (characterised by high-frequency turbulent noise), with stops being signalled by discontinuities in the lower frequency ranges (regions incorporating F1 and F2 information). This definition of landmarks differs from that originally outlined by Stevens (2002).

Results showed that significant improvements in intelligibility can be achieved, even when subjects do not have access to the full spectrum of the clean obstruent consonants (F_c=12.5 kHz). At -5dB SNR significant improvements were seen for all values of F_c. The largest improvements were seen for F_c= 3000Hz, with scores for sentances improving from 24 percent correct in the baseline condition (F_c= 0Hz), to 51 percent and from 67 percent correct

to 80 percent correct for VCVs. Consonant confusion matrices were constructed and subjected to information transmission analysis, focusing on the features of voicing, place and manner. For lower values of F_c (e.g. 500Hz), the improvement in speech intelligibility was found to be largely due to better transmission of voicing information, whereas manner and place cues were better transmitted with high values of F_c (e.g. 3000Hz).

The authors concluded that 'better and more reliable access to voicing leads in turn to better access to low-frequency landmarks' (Li and Loizou, 2008a: 3954). However, it is unclear exactly what the authors mean by low-frequency landmarks for obstruent consonants, as typically, low-frequency landmarks are associated with vowel and sonorant segments (amplitude maxima and minima around F1). Furthermore, the onset of voicing can create discontinuities in the signal, however they are not generally considered to play an important role in underlying contrastive segments (Slifka et al., 2004). Indeed, in the model proposed by Stevens (2002) it is the identification of ALs that should drive further acoustical analysis of the signal to determine voicing and manner cues, and not the other way around. Also of interest is that there was little or no improvement in the transmission of manner features when subjects had access to the clean obstruent consonants. If an improvement in speech perception scores was, in part, due to better access to reliable acoustic landmark information, it would be expected that an increase in the transmission of manner features would be observed, as landmarks are believed to provide initial information on the manner of articulation.

It is difficult to assess the contribution of ALs from the experiments outlined in Li and Loizou (2008a), as although providing subjects with clean obstruent consonants (or only specific frequency regions) will undoubtedly have given them better access to clear acoustic landmark information, subjects also had access to other important cues such as $f0$, F1 and F2 information. These cues are useful in determining the identity of a vowel or consonant, but it is the landmarks that first signal that a particular segment is present. However, landmarks act only as signposts to regions rich in information, so it is not possible to completely disassociate spectral changes associated with ALs from their surrounding acoustic cues (e.g. formant transitions).

It is possible that the results from these experiments underestimate the importance of ALs when listening at low SNR levels. At SNR levels of 0 and -5 dB, NH listeners should not have great difficulty in following the noise corrupted speech. It is therefore possible that the

importance of having access to clear acoustic landmark information may be more pronounced at lower SNRs, as listeners become more reliant on vowel-consonant boundaries than on other spectral cues. The 20-talker babble noise used to corrupt the speech will have only limited spectral dips in which the participants would have been able to glimpse the target signal. This noise type is more comparable with a steady-state noise, such as speech-shaped noise, and is therefore not representative of a real-world listening situation. It may be of interest to investigate the role of landmarks when listening in more fluctuating maskers as better access to ALs during the dips in fluctuating maskers may help contribute to release from masking. This is because better access to clear ALs helps to facilitate segmentation and the identification of word boundaries and allows listeners to fuse the glimpsed portions of the signal across time. It is therefore the combination of the favourable SNR and the favourable timing of dips in the masker that lead to masking release (at times when ALs are present). This hypothesis is supported by findings from an earlier study (Li and Loizou, 2007) in which NH listeners obtained high speech recognition scores in noise when they were able to glimpse information in the F1/F2 frequency region (0-3 kHz).

Parikh and Loizou (2005) looked at the influence of noise on different vowel and consonant cues. They identified three important cues to place of articulation for stop consonants; spectrum, spectral change from burst to voicing onset and formant transitions. The aim of the investigation was to ascertain which speech features are robust in the presence of background noise and therefore which cues listeners may use for speech sound identification in noise. They measured the effect of SS noise and four-talker babble (two male, two female) on the burst frequency (largest spectral magnitude of the burst spectrum) and spectral tilt (positive, negative or diffuse) of the stop consonants /b, d, g, p, t, k/. The authors found that both burst frequency and spectral slope were significantly affected by noise, suggesting that they are not robust cues to place of articulation when listening in noise. The authors then measured stop-consonant identification scores for nine NH listeners in the same noise conditions. Results showed that identification of the stop consonants remained high, between 80-90 percent, even in the presence of background noise. As the authors had already shown that burst frequency and spectral tilt were not robust cues to place of articulation when listening in noise, they suggested that listeners must use other cues to aid identification of stop consonants and highlighted that spectral change may in fact play an important role.

3.5 Chapter summary

It is not yet clear exactly how the brain decodes the speech signal, however, there is reasonable evidence to suggest that ALs are important for identifying regions around which further analysis of the signal should be concentrated. Even if further processing around landmark times is not to uncover so called "feature values", important cues which help to determine voicing, place and manner-of-articulation do appear to be concentrated within the vicinity of these landmarks. These include cues such as formant transitions (important for both vowel and consonant recognition), amplitude of frication noise and the spectral region in which it occurs (important for fricative place-of-articulation), and the presence or absence of vocal fold vibration near a landmark (which indicates whether or not a speech sound is voice). Timing between onset and offset landmarks of consonants can be used to determine duration of frication noise and help distinguish between stops, fricatives and affricates.

Chapter 2 argued for an approach to speech processing in CIs which focuses on making more efficient use of current channel capacity by transmitting only the important elements of the speech signal. The following chapters will look to build on the evidence that the detection of spectral change is essential for the perception of speech, and investigate how present CI processing may limit CI users' access to these important landmark events in the speech signal. The term acoustic landmark will be used from this point on to describe important acoustic events in speech that signal sudden spectral and amplitude change.

In this thesis, the term "high-frequency" or "obstruent" landmarks will be used in relation to the obstruent consonants (plosives and fricatives) and are related to the rapid onsets and offsets of a segment of aperiodic noise (i.e. burst release or frication) within the signal. "Low-frequency" landmarks will be used to describe the slowly varying, periodic segments of vowels and sonorants.

Chapter 4- Landmark transmission with cochlear implants

The previous chapter argued for the importance of spectral change and so called "acoustic landmarks" in good speech perception. This chapter explores the transmission of these landmarks with current CI processing and asserts that improved speech perception in noise for CI users may be realised if they are given better access to obstruent acoustic landmarks. Further to this, it also considers different methods for enhancing landmarks for CI users and introduces the concept of automatic detection of landmarks to guide potential enhancement strategies.

4.1 The contribution of obstruent acoustic landmarks to speech perception in noise by CIs

Studies by Li and Loizou (2007, 2008a) suggest that providing NH listeners with access to reliable AL information (for obstruent consonants) can significantly improve speech intelligibility in noise. Following on from their earlier experiments, Li and Loizou (2009) wanted to determine whether improved speech perception in noise could be seen for CI users if attempts are made to improve the transmission of obstruent landmarks through the CI device.

To test this hypothesis, Li and Loizou (2009) tested NH subjects listening to CI simulated sentences which had been degraded by noise. As in previous experiments they constructed sentences in which the obstruent consonants were left clean and only the sonorants were masked by noise. This was done at -5, 0 and +5 dB SNRs, for steady-state and two-talker competing speech maskers (female talker masker). The authors used a fixed-channel noise-band vocoder strategy and tested in three different channel conditions; six, 12 and 22. Participants were given some training listening to vocoded sentences for the different channel conditions but it is not clear from the procedure whether they received any training in noise (for either masker type). However, participants listened to two lists per condition and averaging percent correct scores across the two lists may have helped balance out any learning effects for a particular condition.

Results showed that for the unprocessed conditions, where obstruent consonants were left corrupted by noise, that there were no significant differences between scores for the different

masker types for six and 12 channels; the authors attributed this to floor effects at -5 and 0 dB SNR conditions. However, scores with the steady-state masker were significantly higher than for the two-talker masker in the 22 channel condition. When they had access to the non-degraded obstruent consonants, listeners showed substantial improvement in all channel number conditions, at all SNRs and for both masker types. For six and 12 channels, results were significantly better with the fluctuating masker than for the steady masker at all SNR levels, but for the 22 channel condition, the fluctuating masker only gave greater improvement at -5 dB SNR. The authors argued that the improvement in speech recognition scores for the clean obstruent consonant conditions was the result of improved access to clear acoustic landmark information, and therefore better access to segmentation cues.

Using NH subjects in this experiment allowed the authors to eliminate any confounding factors associated with CI users, for example, implant device, electrode insertion depth, number of active electrodes and stimulation rates. However, the authors tested using channel numbers that are not representative of the spectral resolution capabilities of actual CI users. As shown in chapter 2, many CI users are not able to make use of more than 4-6 spectrally independent channels, and therefore results in the 12 and 22 channel conditions may over-inflate the potential benefits of improved access to obstruent consonant information for CI users when listening in noise. Indeed, the authors reported a smaller improvement with clean obstruent consonant information for six channels at -5 and 0 dB SNR than for 12 and 22 channels.

Although percentage improvement was similar for all channel conditions at 5 dB SNR (between 30-40 percentage points), total percent correct scores were still only between 50-60 in the six channel condition, whereas participants scored closer to that obtained in the no noise condition when listening with either the 12 or 22 channel vocoder (between 80-90 percent correct). These results appear to highlight the importance of good spectral resolution for listening to speech in noise as generally, greater improvement was seen with increasing channel number. Further to this, the study highlights the difficulty CI users face when listening in noise, as scores in the clean consonant six-channel condition at 0 dB SNR are almost half (in terms of percent correct) of those obtained by NH participants in an equivalent condition in the earlier study by Li and Loizou (2008a).

Li and Loizou (2010) went on to conduct a similar experiment using CI users. They tested seven CI patients, using a similar test paradigm as in their previous study (Li and Loizou,

2009) but due to the floor effects observed at -5 and 0 dB SNR the authors decided to test the implant users at 5 and 10 dB SNRs. As with the NH subjects listening to CI simulated speech, access to the clean obstruent consonants offered a significant improvement when listening in noise (at both levels of SNR); however, the improvement was greater at 5 dB SNR (20-30 percentage points) than at 10 dB SNR (10-15 percentage points). Interestingly, the CI users' scores in unprocessed noisy condition at +5 dB SNR were more in line with those obtained in the 12 channel condition (at the same SNR) than in the six channel condition from the 2009 study.

The authors concluded that the dependency on SNR level suggests that CI users may make use of different sets of acoustic cues at different levels of spectrally degraded speech and that the importance of ALs is most prominent at lower SNRs (where there is more confusion between the target and the masker). As in the previous studies, the authors identified that subjects had access to the full spectrum of the clean obstruents and therefore more information than just ALs. As a result, a second experiment was conducted to determine how much of the benefit shown was due to access to clear ALs. This was done by applying selective envelope compression to only the low-frequency channels (<1 kHz) of the corrupted stimuli. A logarithmic mapping function was applied during sonorant segments (as was used already implemented in the participants' speech processors) and a weakly-compressive function during obstruent consonant segments. The aim of the selective envelope compression was to improve vowel-consonant boundaries, which should help CI users to identify times of obstruent ALs whilst leaving information relating to consonant identity corrupted.

Once again, participants were tested at +5 and +10 dB SNRs, in SS and two-talker noise and with and without processing but unlike in earlier experiments, the processing left both sonorant and obstruent segments corrupted by noise. Results from this experiment showed that selective compression significantly improved speech intelligibility (relative to the baseline condition); however no release from masking was observed at either SNR. The improvement in scores with selective compression at +5 dB SNR was less than for when the full, clean, obstruent segment was available to the listener in the previous experiment (10-15 percentage points and 20-30 percentage points respectively). However, the amount of improvement seen at +10 dB SNR was similar to that obtained with the full obstruent consonant spectra in the earlier experiment (10-15 dB percentage points). This is in line with the degree of improvement seen in their 2008a study when only information up to 1000 Hz

was left uncorrupted in the obstruent consonant segments of the sentence. Li and Loizou (2008a) attributed the improvement seen to an increase in the transmission of voicing information, and so they concluded that the improvement seen with selective compression of obstruent segments (<1 kHz) is also due to the increase in the transmission of voicing information.

Chen and Loizou (2010) investigated the role of obstruent ALs in the improvement of speech perception observed in patients using combined electric-acoustic stimulation. Typically, patients who use EAS have reasonable thresholds up to 750 Hz (20-60 dB HL) and are therefore able to process information at these frequencies through the normal acoustic pathway. The authors conducted a simulation study using a similar test paradigm as in Li and Loizou (2008a; 2009; 2010) whereby noisy obstruent consonants were replaced with clean versions of the same consonants. Participants were seven NH listeners and were tested in six different conditions:

1. Eight band sine-vocoded speech (V)
2. Speech which has been low pass filtered (<600 Hz) to simulate acoustic hearing (LP)
3. A combination of the first two conditions (using upper five channels of the vocoder only) to simulate EAS (LP+V)
4. EAS condition whereby the vocoder portion of obstruent consonants was left clean (LP+Vc)
5. EAS condition whereby the acoustic portion of the obstruent consonants was left clean (LPc+V)
6. EAS condition whereby both vocoder and acoustic portions of the obstruent consonants were left clean LPc+Vc)

Percent correct scores for sentence recognition showed an improvement when listeners had access to both clean acoustic and vocoded portions of the obstruent consonants, compared to the same EAS condition where both low-frequency acoustic information and high frequency vocoded information were left corrupted. This improvement ranged from 15 percentage points in the 5 dB SNR condition to 30 percentage points at 0 and -5 dB SNRs; this once again follows the trend of previous experiments whereby improvement was greater for more challenging SNRs. What is surprising is that although an improvement of 15 percentage points was seen in the positive SNR condition, it was not found to be significant. Overall the results showed a trend of participants scoring highest in conditions where the high-frequency vocoded portions of the obstruent consonants were left uncorrupted by noise (LPc+Vc >

LP+Vc > LPc+V > LP+V > V > LP). Floor effects were seen at -5 dB SNR for the V, LP, LP+V and LPc+V conditions.

The results from this experiment would appear to suggest that information held in both the low-frequency acoustic portion of the clean obstruent segments, and the clean, high-frequency vocoded obstruent segments, can help to improve speech perception in noise for EAS users. However, as percent correct scores were higher in the LP+Vc condition than in the LPc+V condition it is reasonable to suggest that the majority of the benefit seen for the LPc+Vc condition was due to access to clean obstruent information in the high-frequencies, rather than clean low-frequency acoustic information.

To summarise, the work completed by Loizou and colleagues demonstrates that when obstruent consonant information is restricted participants do not perform as well as when they have access to the full (clean) spectrum. It has therefore been suggested that CI users need access to both ALs, and the surrounding cues to consonant identity, which are easily masked by noise. In particular, CI users may benefit from improved access to information in the high frequencies (or at least > 600 Hz). The contribution of vowels and other sonorant sounds to speech perception in noise by CI users was not assessed in any of the studies discussed above but the authors argued that selective compression should increase contrast at V-C boundaries, and this could be used to improve identification of both the sonorant and consonant segments (as important cues for both segments are found at these boundaries). The rapidly changing F1 and F2 frequencies/ trajectories, found at V-C boundaries, are considered important cues to the identity of many speech sounds; however, due to limited spectral resolution they are not reliably transmitted with current CI processing (Munson and Nelson, 2005).

The presence of background noise degrades important cues for obstruent consonants and, along with reduced frequency resolution, smears important spectral contrasts and landmark information (Li and Loizou 2009; Leek and Summers, 1996). Loizou and Poroy (2001) identified a number of contributing factors to reduced spectral contrast for CI users, including compression (linked to reduced dynamic range), the steepness of the compression function, the setting of a user's sensitivity control and the addition of background noise. The 2009 study by Li and Loizou showed that if the SNR is relatively high (e.g. ≥ 0 dB), and spectral resolution is reasonable, then subjects are less reliant on ALs, as they have access to a number of other cues (e.g. formant transitions). They therefore concluded that increasing the number of channels gives access to a greater number of cues. However, as previously

discussed, CI users are only able to make use of a small number of spectrally independent channels and therefore are more likely to rely on the detection of landmarks when listening in noise. As improving spectral resolution for CI users is extremely difficult, a method for improving access to landmark information (i.e. amplitude changes in specific frequency regions) by adapting current CI processing is required.

4.2 Automatic detection of landmarks

So far, classification of landmarks has been described in terms of manual labelling using the waveform/spectrograms of the clean stimuli. However, to improve the transmission of landmarks with CI processing it is likely that landmarks will need to be automatically detected from within the speech signal, and in noise, to help guide further processing. The following sections look at automatic landmark detection algorithms that have been developed for use in automatic speech recognition (ASR) systems.

4.2.1 Automatic speech recognition

Traditional ASR frameworks employ Artificial Neural Networks and/or Hidden Markov Modelling (HMM). These are purely statistical approaches to speech recognition and are based on the assumption that successive speech frames are conditionally independent and uncorrelated. This means they require training for specific environments if they are to work optimally (Bitar and Espy-Wilson, 1995; Liu, 1996), which can require a very large data set. Such statistical methods often perform well for speech in quiet and for well-articulated (such as read) speech, but often fall short for recognition tasks in noise. This has motivated researchers to return to more knowledge-based approaches, such as those based on feature detection, as it allows the possibility to build in knowledge about speech into the speech recogniser. These approaches are designed to model parts of the human perception process; such systems focus more specifically on processing speech rather than processing signals in general (Liu, 1996). ASR systems built around specific speech-based knowledge can be a useful scientific tool in helping to determine the strengths and limitations of a particular model of speech perception, in particular, how good the model is with coping with different sources of variation in the speech signal.

Since the 1980's, researchers have been developing ASR systems that are based on feature models, including earlier iterations of Stevens' (2002) LAFF model. Notable early work includes that of the development of the SUMMIT system (Bitar and Espy-Wilson 1995; Bitar, 1997; Liu, 1996; Zue et al., 1989; 1990), The systems proposed in these papers all

incorporate a landmark detector in their front-end processing. Similar to the argument put forward by Stilp and Kluender (2010), Smith (1995) suggested that ASR systems may benefit from front-end segmentation processing techniques based on detecting onset events in the speech signal as this more closely represents the biological auditory system.

ASR systems that incorporate landmark detection (sometimes referred to as event-based systems (EBS)) aim to extract the acoustic correlates of linguistic features by first identifying important events within the speech signal which point to times 'when a feature is most acoustically or perceptually prominent' (Jansen and Niyogi, 2008: 1743). Therefore, although there may be little phonetic information contained within a landmark segment of the speech signal (Hasegawa-Johnson et al., 2005), they point to regions rich in information relating to manner of articulation (Hasegawa-Johnson, 2001), and further processing around these events can uncover place feature values.

4.2.2 Early landmark detection algorithms

The SUMMIT speech recognition system (Zue et al., 1989; 1990) was developed at the Massachusetts Institute of Technology (MIT) and makes use of acoustic-phonetic knowledge to perform classifications. It uses a similar framework as Stevens' (2002) LAFF model and relies on an initial stage of landmark detection to guide further acoustic-phonetic feature extraction (Zue et al., 1990: 49). Figure 4.1 demonstrates the three main stages of the SUMMIT system; clear similarities can be seen with the stages of processing involved in the LAFF model of speech perception as outlined in figure 3.9.

The first stage of the SUMMIT system is outlined in the blue box in figure 4.1 and involves creating an acoustic-phonetic transcription of the incoming speech utterance and extracting features for phoneme recognition. The speech signal is first transformed into a representation based on an auditory model (Zue et al., 1990) that helps in the identification of onset and offset events connected with landmarks. The output of this model is then used to segment the signal and phonemes are described in terms of a series of acoustic events. Pattern classification algorithms are then implemented to perform phonetic classifications. SUMMIT attempts to take into account variations in pronunciation in the second stage (outline in the red box) with as many of the alternatives entered into the lexicon of the speech recogniser. The output of the stage outlined in the blue box is matched to representations in the lexicon and a "best match" approach is used. The final stage of the SUMMIT system (not explored in

the 1989 or 1990 paper) would be to utilise higher-level linguistic knowledge to *understand* what is being said in a particular speech utterance.

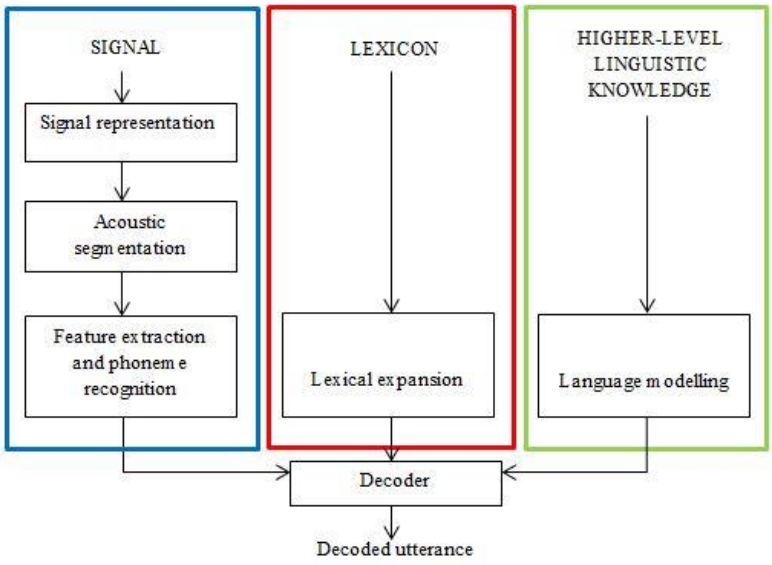


Figure 4. 1 The three stages of the SUMMIT system. After Zue et al. (1989).

Liu (1996) also proposed a speech recognition system that closely resembles Stevens' (2002) LAFF model of speech perception (see figure 4.2). Liu (1996: 3419) argues that landmarks act as *foci*, and it is around the times of landmarks when further processing takes place (and not between). This is because information relating to speech sound identity tends to be concentrated around these regions. This differs from more traditional frame-based processing and segmentation where further processing takes place across an entire frame (typically 5-30 ms long) or segment (corresponding to a phone). Essentially, landmark-based processing should be more efficient than other, more traditional, methods as it can "target the level of effort where needed" (Juneja and Espy-Wilson, 2008: 1154).

Liu categorised landmarks into four distinct groups: (1) abrupt-consonantal (marks closure and release of constriction), (2) abrupt (abrupt changes in the sound caused by glottal or velopharyngeal activity but without accompanying articulator movement), (3) nonabrupt (as for glides) and (4) vocalic (for vowels). Liu notes that landmarks which are acoustically abrupt are most numerous within the speech signal, equating to approximately 68 percent of the total number of landmarks in speech, and therefore decided to focus on detecting these abrupt landmarks. A landmark detection algorithm that gives an output of a series of

landmarks specified by time was developed. Initial processing involved the detection of peaks in the waveform which represent times of abrupt spectral change. This was achieved by first dividing a generated broadband spectrogram in to six frequency bands (0.0-0.4, 0.8-1.5, 1.2-2.0, 2.0-3.5, 3.5-5.0 and 5.0-8.0 kHz).

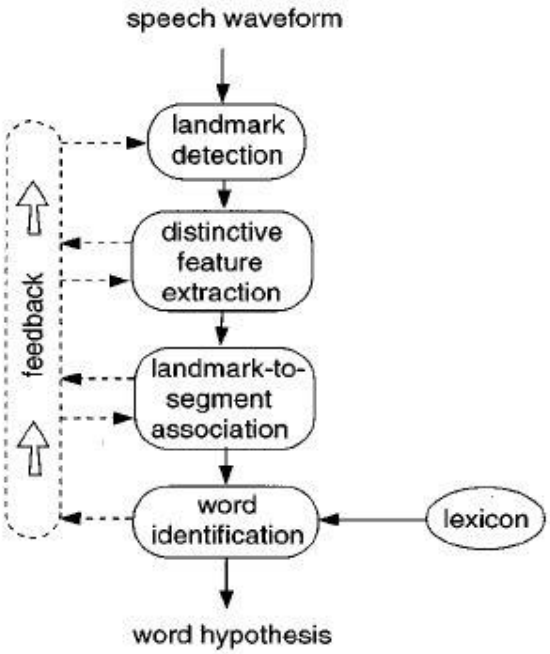


Figure 4. 2 Landmark-based speech recognition system as proposed by Liu (1996:3418). Reproduced with permission, copyright Acoustic Society of America.

For each of the bands, an energy waveform was constructed and the derivative of the energy was computed. Each band was designed to monitor different types of spectral change, for example, band 1 monitored the presence or absence of glottal vibration, which is why the band did not extend above 400 Hz. Following the detection of the peaks, a landmark type-specific processing stage was applied to find three landmark types: glottis (time vocal folds start/stop vibrating), sonorant (closure or release of a nasal and/or /l/) and burst (affricate or aspirated stop burst-onset and offsets). Interestingly, Liu classifies the sonorant consonants (/n/, /m/ and /l/) as having abrupt landmarks. Abrupt spectral changes are commonly observed in the higher frequency regions of the spectrum and although the nasal consonants are often grouped with the stop consonants, they could also be considered as having landmarks which are found in the more slowly varying low-frequency spectral components of

the speech signal, relating to a minimum in F1 energy (as with other sonorant consonants such as glides). This highlights that Liu's definition of landmarks differs slightly from that of Stevens (2002) and from the definition used in this thesis (as outlined in section 3.5).

The error rate of the landmark detection algorithm was calculated as the sum of the deletion (missed landmarks), substitution (landmark detected but of the wrong type) and insertion (false landmark) rates. Deletion and insertion rates were considered more serious than the substitution of a landmark as not identifying or incorrectly inserting a landmark could misguide further processing for acoustic cues and produce errors in lexical representation, and therefore, lexical access. The overall error rate for the three detectors was found to be 16%, with the sonorant detector contributing most to errors. Liu suggested that the poorer performance of the sonorant landmark detector may have been due to the large variation in phonetic context in which these speech sounds can be found. However, it is also possible that it was inappropriate to design a sonorant landmark detector based solely on abrupt changes in the signal (corresponding to the closure and release of the sonorant segment) as nasal sounds are also cued for by dominance in the low frequencies, particularly in the region of F1 and F2. Nonetheless, Liu (1996) demonstrated that it was possible to detect landmarks from within the speech signal, with relatively high accuracy, and this technique is still used by researchers today to automatically label landmarks in speech (see Shattuck-Hufnagel and Veilleux, 2007).

Bitar (1997) focused on the acoustic modelling and analysis of speech based on phonetic features (voicing, place and manner) as the basis for a speaker-independent speech recognition system. He developed a series of rule-based algorithms that were able to extract the acoustic properties of phonetic features, based on evidence from acoustic phonetics and spectrographic analysis. Bitar (1997: 7) argues that phonetic features (voicing, place and manner) are the only candidate that can satisfy the three requirements for the choice of minimal speech units used for speech recognition by machine (not phonemes or syllables etc.). He defined these as (1) their ability to uniquely describe lexical items which are usually words, (2) their allowance for the generalizability of different phonological processes that occur during fluent speech and (3) that from these units, it is possible to develop reliable acoustic models that sufficiently describe their acoustic manifestations from a limited amount of data.

Earlier work by Bitar and Espy-Wilson (1995) on a feature-based representation of speech aimed to target only the linguistic information contained in the speech signal. They assert that

focusing on the phonetic content of the speech signal could help to reduce model complexity and therefore lesson the demand on training data. The following section explores how Bitar's (1997) landmark-based system has been developed and improved to create modern day frameworks based on phonetic features.

4.2.3 Current landmark-based ASR systems

A large body of work has been carried out at the University of Maryland which follows closely on from the work of Bitar (1997), including that of Juneja and Espy-Wilson (2003; 2004), Juneja (2004), Salomon et al. (2004) and Espy-Wilson et al. (2007). Juneja and Espy-Wilson (2003; 2004; 2008) developed a landmark-based system which uses a probabilistic framework, as previous rule-based systems such as those described by Liu (1996) and Bitar (1997), were unable to handle pronunciation variability. The development of a probabilistic framework was considered important if the landmark-detector system is to be scaled to achieve large-vocabulary recognition tasks which integrate higher level information.

The authors developed a set of Acoustic Parameters (APs) to extract information from the speech signal depending on the type of landmark detected. APs are "the measures performed on the speech waveform and its time-frequency transformation(s) in order to seek evidence for the acoustic properties of phonetic features" (Bitar, 1997: 23). Traditional HMM frameworks which utilise Mel Frequency Cepstral Coefficients (MFCCs) differ markedly from this as they assume that speech frames are independent from one another (known not to be the case) and as such, all frames of the signal are analysed and all MFCCs looked at. MFCCs are amplitudes derived from a cepstral (spectrum of a spectrum) representation of the input signal based on the mel-scale (based on pitch comparisons). Landmark-based systems are therefore considered to be much more efficient as processing only takes places around the times of landmarks, and different resolutions (window lengths) can be applied depending on the type of landmark detected. This has been motivated by the fact that speech sounds, and their acoustic cues, vary in duration; for example a stop burst is much shorter than a vowel (Juneja and Espy-Wilson, 2004). The APs chosen were based on those proposed by Bitar (1997) and can be found in table 4.1.

Phonetic Feature	APs
Silence	(1) $E[0,F3\text{-}1000]$, (2) $E[F3,f_s/2]$, (3) ratio of spectral peak in [0,400 Hz] to the spectral peak in $[400,f_s/2]$, (4) Energy onset (Bitar, 1997) (5) Energy offset (Bitar, 1997)
sonorant	(1) $E[0,F3\text{-}1000]$, (2) $E[F3,f_s/2]$, (3) Ratio of $E[0,F3\text{-}1000]$ to $E[F3\text{-}1000,f_s/2]$, (4) $E[100,400]$
syllabic	(1) $E[640,2800]$ (2) $E[2000,3000]$ (3) Energy peak in [0,900 Hz] (4) Location in Hz of peak in [0,900 Hz]
continuant	(1) Energy onset (Bitar, 1997), (2) Energy offset (Bitar, 1997), (3) $E[0,F3\text{-}1000]$, (4) $E[F3\text{-}1000,f_s/2]$

Table 4. 1 APs used in broad class segmentation by Juneja and Espy-Wilson (2008: 1156). Reproduced with permission, copyright Acoustic Society of America.

As with the landmark-based system outlined by Juneja and Espy-Wilson (2008), other ASR systems such as segment-based (Glass et al., 1996) and syllable-based systems, carry out multiple segmentations for further analysis and use probabilistic frameworks. However, these systems do not selectively use knowledge-based APs to detect landmarks and other articulatory features. It is also important to note that some HMMs have been designed to include knowledge-based APs but they do not have an initial stage of segmenting or detecting perceptually and acoustically important events in speech (Juneja and Espy-Wilson (2004; 2008). In earlier work (Juneja and Espy-Wilson, 2004), the authors associate landmarks with five broad classes: vowel (V), fricative (Fr), sonorant consonant (SC), stop burst (ST) and silence (SILEN). The probabilistic framework can be used to organise these broad classes into a feature hierarchy and is shown in figure 4.3. In the hierarchy, the most important feature classifications are found at the top; these features are less context dependent and are therefore more robust to noise (Espy-Wilson et al., 2007). Manner features are therefore found at the top of the hierarchy and place features at the lower nodes, with the lowest nodes displaying phonemes. Hierarchies such as this can be useful in demonstrating why place features are more difficult to detect in noise than manner features (Miller and Nicely, 1955); they require more spectral information and therefore further processing (Hasegawa-Johnson, 2001).

At each node of the hierarchy a Support Vector Machine (SVM) classifier is applied, using only those APs which are relevant for the feature at that node (these are automatically extracted from the signal). The authors point out however that they do not rule out the use of neural networks or Gaussian mixture models instead of SVMs. Following this, probability decisions are obtained for each SVM and are combined with class dependant duration

probability densities to give one or more segmentations into the broad classes (V, Fr, SC, ST and SILEN). This locates landmarks corresponding to each of the broad class segments.

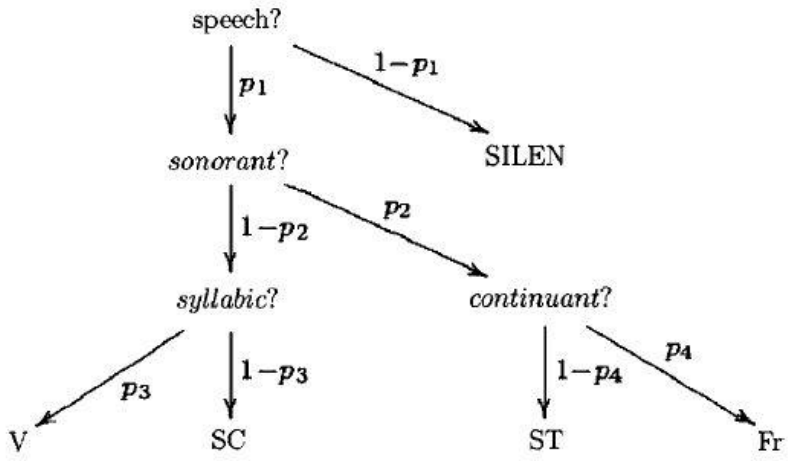


Figure 4. 3 Probabilistic feature hierarchy as proposed by Juneja and Espy-Wilson (2008: 1156). Reproduced with permission, copyright Acoustic Society of America.

Unlike Liu's (1996) system, EBS are designed to extract both abrupt and non-abrupt landmarks closely resembling the landmark classifications of Stevens (2002). Table 4.2 outlines the landmarks detected by EBS and their corresponding broad classes. The performance of the EBS framework was demonstrated using broad-class recognition and landmark detection tasks (using the TIMIT database) and a vocabulary-constrained broad-class recognition task (using isolated digits from the TIDIGITS database). For both tasks, only information relating to manner was sought.

Evaluation included a four-way comparison to determine the effectiveness of a probabilistic landmark framework against a statistical HMM system and knowledge-based APs versus MFCCs. Therefore, the following systems were compared:

1. EBS using APs
2. HMM using MFCCs
3. EBS using MFCCs and
4. HMM using APs

Broad class	Landmark(s)
Vowel (V)	Syllabic peak (P) Vowel onset point (VOP)
Stop (ST)	Burst (B)
Sonorant consonant (SC)	Syllabic dip (D) SC onset (Son) SC offset (Soff)
Fricative (Fr)	Fricative onset (Fon) Fricative offset (Foff)

Table 4. 2 Landmarks and their corresponding broad classes as detected by EBS.

Results for the landmark detection task showed that unconstrained segmentation using the TIMIT data yielded an accuracy of 79.5 percent (for EBS using APs). When detection of landmarks was constrained using broad class pronunciation models, correct segmentation was obtained for about 68.7 percent of the words (Juneja and Espy-Wilson, 2008: 1167). This was found to be comparable with HMM systems using APs (67.8 percent) and for HMM models using MFCCs (63.8 percent). Overall, the accuracy of broad class recognition was found to improve for constrained segmentation (84.2 percent), compared to unconstrained segmentation (74.3 percent) and once again this was similar to the results obtained for the HMM system using APs.

Another system that has also been developed to incorporate landmark detection to guide further processing is that of Jansen and Niyogi (2008). They propose a system which is based specifically on the detection of distinctive features (as in Stevens, 2002). Similar to previous systems based on landmark detection, Jansen and Niyogi outline a set of feature detectors (based on manner-of-articulation) and order these in a hierarchical way (as in Juneja and Espy-Wilson, 2004; 2008). Although rooted in the distinctive feature theory, the authors assert that traditional views which organise distinctive features into a set of bundles (as with Chomsky and Halle, 1968; Stevens, 2002) suggest that the features are synchronised in time. Jansen and Niyogi (2008) developed a feature hierarchy that considers the specific time at which feature detectors "fire" and this is represented in figure 4.4 by the horizontal time axes

at each node of the hierarchy, with the arrows indicating times when each feature detector fires. The times when feature detectors fire can be closely linked to the concept of landmarks, as they do so around perceptually and acoustically important events in the speech signal. This leads to a sparse, point process representation of the speech signal which can be likened to the spike train patterns of specific neurons in the auditory cortex (Jansen and Niyogi, 2008: 1739). This spike-based representation of speech is what differentiates this approach from other frame-based representations used in most other ASR systems, including those already discussed.

. The authors make close links between their hierarchical structure and the sonority profile, suggesting that the first most important distinction to make in speech is that between sonorant and obstruent sounds. The sonority distinction is therefore found at the top of the hierarchy outlined in figure 4.4.

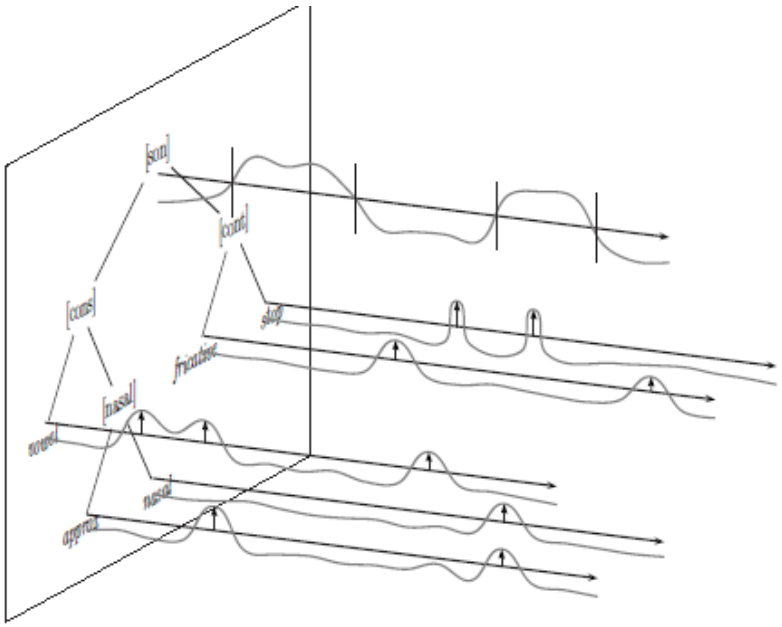


Figure 4. 4 Feature hierarchy timing tier as proposed by Jansen and Niyogi (2008). Reproduced with permission, copyright Acoustic Society of America.

Broad class feature detectors were built that could separate a + feature value from a – feature value for vowels, approximants, nasals, fricatives, stops and silence. A sonorant segmenter was applied to the speech signal in the first instance to separate sonorants from obstruents. If the segmenter was perfect, it would only be necessary for each feature detector to distinguish

between other features in the same broad class, however, this is not the case and so a "one-versus-all" classifier (using MFCCs and SVMs) was required. This means that each feature detector had to distinguish between all other features, and not just those in the same broad class. Results from this landmark-based speech recognition framework were again found to be competitive with equivalent HMM-based systems. However, the authors suggest that to reduce possible insertion and deletion errors it may more optimal to use APs because the stop detector, which gave superior performance compared to all other detectors, was the only detector to make use of knowledge-based APs.

This section has discussed several ASR systems which incorporate landmark detection, and it has been argued that an initial landmark detection stage can be used to help guide more efficient methods of extracting information important for speech sound identity. Current ASR systems do not perform as well at recognising speech as human listeners and it has been suggested by several researchers (Zue et al., 1989; Liu, 1997; Bitar, 1997) that ASR should take more of an acoustic-phonetic approach, building in information about the speech signal. Although performances with the aforementioned landmark-based ASR techniques have produced recognition rates comparable with other ASR systems using HMMs, they have not yet managed to surpass them. This research lends some weight to speech perception models based on distinctive features and demonstrates the importance of ALs in helping to identify regions of the speech signal which are rich in information.

These studies clearly demonstrate that it is possible to automatically detect landmarks from within the speech signal, however, reliability of such landmark-based systems is usually only presented for speech recognition in quiet and future work on determining the reliability of these frameworks in noise would be extremely beneficial for exploring the importance of landmarks when listening in degraded conditions. In particular, it is not clear from these studies how frameworks based on an initial stage of landmark detection would perform in different types of noise, particularly noise consisting of competing talkers (i.e. single and two-talker maskers). These types of maskers are likely to introduce conflicting landmark cues and therefore recognition systems will need to build in methods for parsing the different streams. This is of particular importance for CI landmark enhancement strategies based around automatic detection of landmarks as they would want to avoid transmitting information from the wrong speech stream. Developing landmark detectors that are robust in noise will also be important for work outlined later in this thesis.

4.3 Methods for enhancing landmarks

The literature has shown that ALs are specific events in the speech signal that may be of great perceptual importance, in particular they help to signal the presence of obstruent and sonorant sections and point to regions in the signal where other important cues can be found. Studies by Li and Loizou (2009; 2010) have shown that the transmission of ALs with CI processing may be poor and therefore methods should be developed to improve the representation of these important events in the resulting electrical pattern delivered by a CI device. It is possible that landmark detection algorithms, such as those developed for ASR systems, may be useful for guiding some form of landmark emphasis in CI speech processors. The following sections explore three methods which may be used to enhance landmarks for CIs.

4.3.1 Compression

Li and Loizou (2009; 2010) have suggested using automatic landmark detection in conjunction with a dynamically changing compression function which applies less compression during obstruent consonant segments and more during sonorant sound segments. This form of compression was initially proposed by Zeng et al. (2002), as compression in CIs appears to have more of a detrimental effect on the perception of consonants (as discussed in chapter 2).

Li and Loizou (2010: 1268) proposed a method for automatically identifying obstruent consonants based on a two-class Bayesian classifier which splits the (corrupted) speech signal into sonorant and obstruent segments. The authors split the corrupted speech signal into 16 channels, using the same frequency spacing as implemented in the participants' implant processors, and calculated band-normalised energy values for each channel. These 16 band energy values served as features for the detection framework and fed into a Gaussian Mixture Model (GMM) which was trained using a selection of 180 IEEE sentences and used to classify each 10 ms frame of the sentence as either sonorant or obstruent (*ibid.*:1269). They tested the accuracy of the Bayesian classifier by measuring detection and false alarm rates for 540 IEEE sentences (not already used in the training of the GMM) corrupted by SS noise in either +5 or +10 dB SNR. The sonorant/obstruent classifications as provided by the Bayesian classifier were then compared with manually generated segmentations for the same stimuli. In general, the classifier performed well in both levels of noise, with detection rates (correct identification of obstruents) of around 93 percent and false alarm rates (incorrect classification of sonorants as obstruents) of <11 percent.

This automatic detection of segments was then used to guide a selective compression function as used in their earlier experiment (Li and Loizou, 2010). As per previous studies, CI users listened to sentences with and without selective compression (processed offline and presented via a research processor) and corrupted by SS noise at +5 and +10 dB SNRs. Results showed that selective compression of sonorant and obstruent segments, as guided by the Bayesian classifier, provided significant improvements to speech perception in noise (compared with the standard log compression function used in CI processing). Scores with the automatically detected segments and manually annotated segments (as in the previous experiment) were almost identical. Although this classification method does not specifically detect landmark events from within the speech signal, the method of selective envelope compression guided by a GMM classifier, as proposed by Li and Loizou (2010), could be used to enhance landmarks at syllable and V-C boundaries and improve obstruent consonant identification. This type of classifier should be fairly easy to implement into current CI processors. Nonetheless, further research is required to determine the accuracy of the GMM classifier in other types of interfering noise (e.g. single talker and two-talker backgrounds) and at different levels of SNR.

4.3.2 Channel selection

It is possible that landmark detection may be used to help guide channel selection in an n-of-m strategy, a simple model for which is outlined in figure 4.5. Such a strategy would aim to provide CI users with better access to the acoustic cues necessary for speech perception by selecting channels to be stimulated based on the probability of them containing an acoustic landmark (rather than on amplitude). In the initial stages of this thesis, a landmark detection algorithm based on the method proposed by Juneja and Espy-Wilson (2008) was developed to detect landmarks within different frequency bands, corresponding to the filterbank of a CI device (e.g. 22 channels of a standard Nucleus device).

The algorithm detected landmarks for vowels, fricatives, stops, sonorants as well as also detecting periods of silence. Therefore, unlike many of the other methods of landmark/spectral change enhancement discussed in this thesis, this method does not specifically focus on the obstruent consonants. The outputs of these different landmark detectors for a single channel for the stimulus /ada/ (in quiet) can be seen in figure 4.6. A broadband spectrogram for the same stimulus is shown in figure 4.7 for comparison. Figure 4.6 demonstrates that appropriate peaks occur at the times of silence between the vowel portion of the stimulus (blue arrow) and the onset of the burst noise of the stop (green arrow).

Results are shown for channel 13, a low-mid frequency channel (~1.5 kHz), therefore more prominent peaks may be expected for the stop detector in more high-frequency channels as obstruent landmarks are characterised by abrupt spectral changes in the mid-high frequencies. The stimulus had a period of silence inserted at the beginning and at the end of the segment (as seen in the spectrogram in figure 4.7) and this is demonstrated by the very high probability values shown for the silence detector at the beginning and end of the probability outputs.

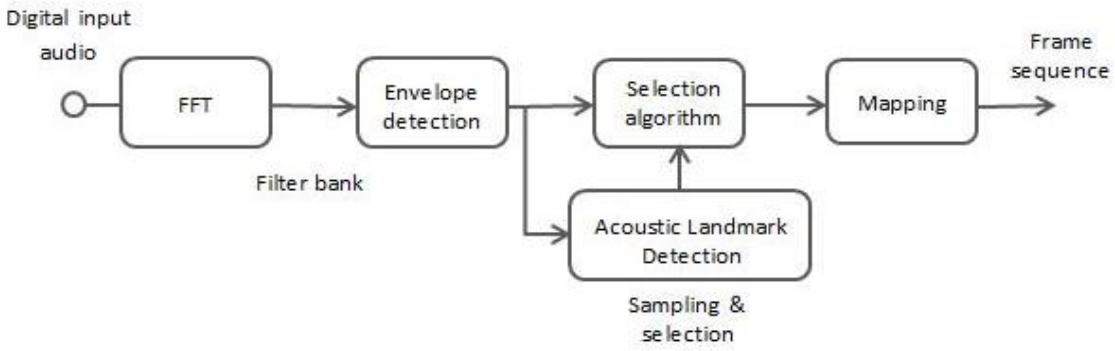


Figure 4. 5 Conceptual landmark-based n-of-m speech processing strategy.

The vowel landmark detector "fires" during times when a vowel is also present in the broadband spectrogram and is "dormant" during the portion of the signal corresponding to the stop closure. The probability values for the vowel detector are quite low at times when it is firing, however this would be expected for a mid-frequency channel as probability values are expected to be much higher for channels closer to the frequency of F1. The fricative detector also appears to be firing at nearly all times during the signal, except for during periods of silence.

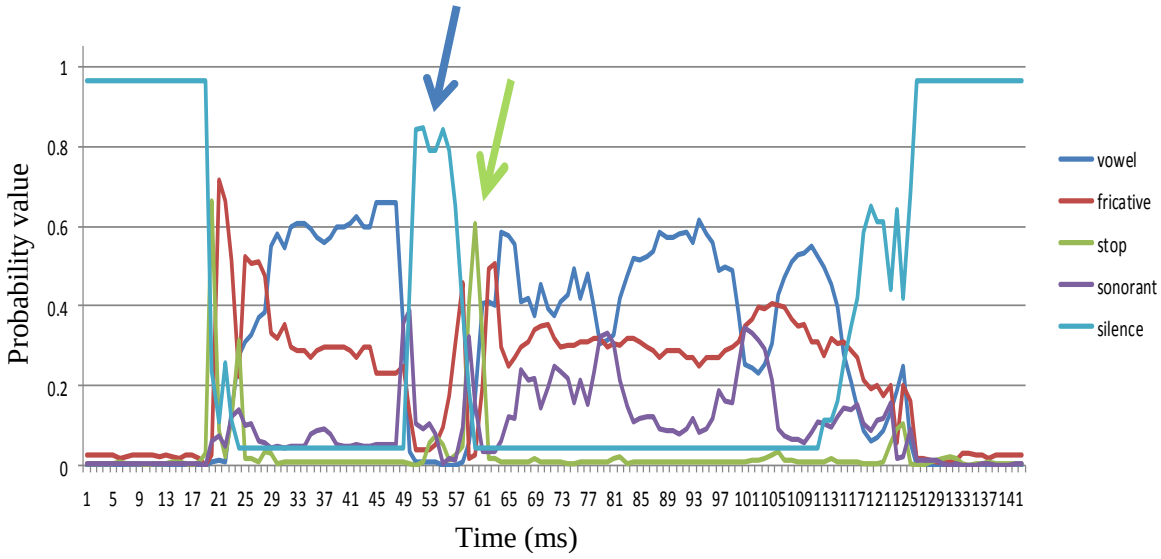


Figure 4. 6 Probability ouput from each of the landmark detectors (vowel, fricative, stop, sonorant and silence) for channel 13 of 22 for the stimulus /ada/, in quiet, at 10 ms intervals. The blue arrow indicates correct identificaion of a period of silence by the silence detector and the green arrow identifies a correct detection of a stop landmark by the stop detector.

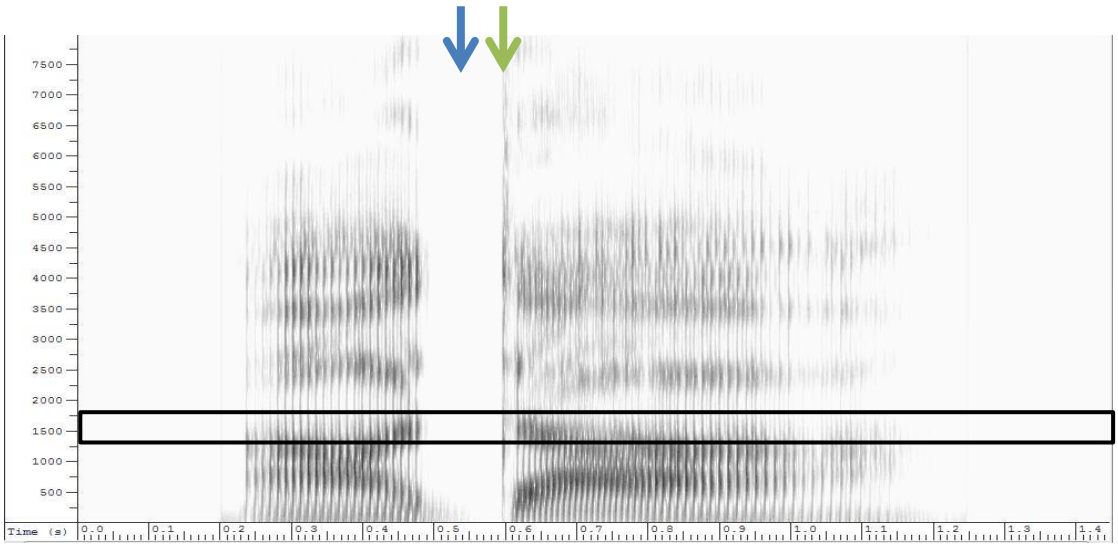


Figure 4. 7 Broadband spectrogram for the stimulus /ada/, in quiet. The black box highlights the frequency region which is included in channel 13. The blue arrow points to the region of silence during consonant closure and the green arrow indicates the time of the burst noise following consonant release.

Although initial results in quiet were quite encouraging, particularly for the stop, vowel and silence detectors, a different pattern emerges when noise was added to the speech signal. Results for the same channel and same stimulus but in +10 dB SNR (eight-talker babble) are presented in figure 4.8. Looking at the landmark probabilities in this condition it can be seen that although a peak is still evident for the stop detector, it is markedly reduced (low probability). This may reduce the likelihood of this landmark being coded accurately. The fricative detector also appears quite "noisy", despite the fact that there are no fricative landmarks present in this signal. This may be the result of the landmark detector mistaking the background noise as a fricative segment.

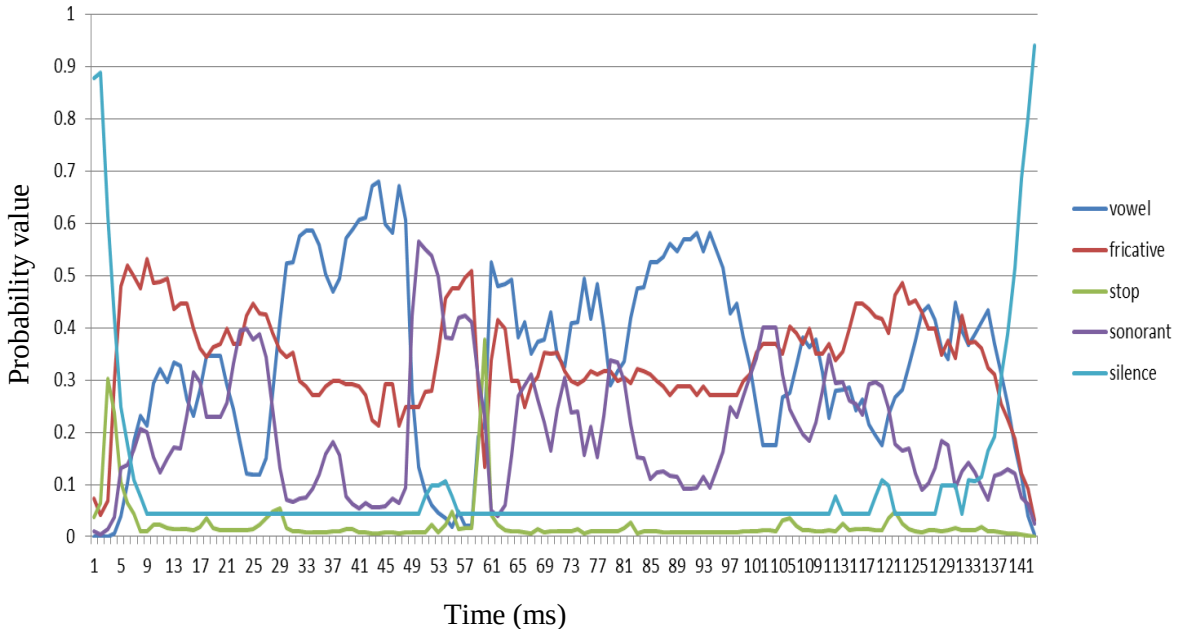


Figure 4. 8 Probability ouput from each of the landmark detectors (vowel, fricative, stop, sonorant and silence) for channel 13 of 22 for the stimulus /ada/, at +10 dB SNR, at 10 ms intervals.

Although results from only one channel and for one speech sound have been presented, results from the initial analysis of the channel-specific landmark detectors showed that their performance is markedly reduced even by moderate levels of noise. As discussed in chapter 3, if landmarks are identified incorrectly this could be detrimental to further processing of the speech signal. It is possible that the channel specific landmark detectors were not accurate enough because they require information from more than a single channel.

As already outlined, sonorant landmarks (vowels, nasals and glides) are found in the low-frequency regions, and obstruent landmarks (stops and fricatives) in the high frequency

regions, therefore it may be more appropriate to constrain sonorant and vowel landmark detectors to low frequency channels and the stop and fricative landmark detectors to the high frequency channels. Channel specific landmark detectors are instrumental in the design of an ALD-based n-of-m strategy; however developing accurate channel specific landmark detectors currently requires further research. It was therefore decided, for the purposes of this study, to focus on using an existing n-of-m strategy and instead develop a method of pre-emphasis that would improve the transmission of landmarks.

4.3.3 Selective enhancement

Pre-emphasis of the speech signal takes place prior to the envelope detection stage of CI processing. The aim of pre-emphasis is to make a particular feature of the speech signal more salient; often by increasing the gain in corresponding channels. This should result in these channels having an increased chance of being selected for stimulation and therefore hopefully increase the transmission of that particular feature. Some enhancement techniques will apply enhancement in blanket form, for example, applying selective compression across all segments as proposed by Li and Loizou (2010). Selective enhancement techniques apply enhancement to specific regions of the speech signal, such as was used in the TESM strategy proposed by Vandali (2001). The following sections explore current methods for enhancing spectral change within the speech signal and considers their potential for use in CI signal processing.

Specific cue enhancement

Typically, speech enhancement is constrained to working with an already noise corrupted signal; however, telecommunications is in a unique position whereby it is possible to modify the speech signal *before* it is corrupted. This is because the problem of noise is mainly at the end of the receiver, i.e. the person on the end of the phone may be listening in background noise. Modifying the signal before degradation occurs can help to make relatively weak but important speech cues more robust and therefore more salient. Therefore, this technique may be used to investigate which elements of speech are particularly important for listening in noise.

When talking in noise, speakers often modify the production of their speech to help make the signal more robust and salient. Some of these techniques include increasing the length of consonant segments and consonant-vowel intensity ratios. Hazan and Simpson (1998) posit that these techniques could be modelled in speech enhancement to aid speech recognition. As

these modifications are made by talkers to improve the salience of specific cues, it could be argued that these cues are particularly important for good speech perception in noise. However, Hazan and Simpson (1998: 213) go on to argue for a more perceptually motivated approach to speech enhancement, instead of trying to mimic specific speaker modifications. This would involve analysing the signal in terms of its phonetic components and attempting to improve the saliency of the information bearing regions by increasing their intensity or duration.

The authors decided to focus on enhancing regions which are associated with acoustic landmarks and therefore rich in important acoustic cues. They tested NH listeners listening to sentences and VCVs in noise (SS noise), where the noise was added after enhancement had been applied to the speech signal. The overall aim of the enhancement experiment was to improve the saliency of the important "landmarks" by amplifying them relative to surrounding vowel amplitude levels. The authors manually labelled the speech stimuli for (a) vowel onset and offsets and (b) consonantal occlusion/constrictions, categorised as burst transients, burst aspirations, frication or nasality portions (*ibid.*: 214). Five different manipulations of the stimuli were investigated and these included:

1. Unaltered or "natural" condition
2. Boosting of bursts, frication and nasals (increasing amplitude)
3. Boosting bursts, frication and nasals as well as surrounding formant transitions
4. Filtering and then boosting of the signal at times of bursts and frication to enhance the most perceptually important spectral regions. Nasals were boosted as before but not filtered.
5. Same as for condition four but with vowel onsets/offsets also boosted

Thirteen NH subjects listened to five repetitions of 12 consonants in three separate vowel environments (/a, i, u/) at SNRs of 0 and -5 dB and for each enhancement condition. Results showed that for both SNRs, participants scored significantly higher (percent correct) in all enhanced conditions when compared with the natural condition. However, this improvement was only small in the 0 dB SNR condition, with the highest mean increase being 6 percent for the final enhancement condition (filtering and boosting of consonants + boosting of vowel onsets/offsets). Only the final enhancement condition showed significantly higher scores than

for the other enhancement conditions, and this was only in the -5 dB SNR condition and once again the improvement seen was quite small. Scores in the /u/ context were significantly poorer than in the /a/ and /i/ conditions, however, scores for the different enhancement conditions were not explored in terms of vowel environment. Information transmission analysis showed that enhancement significantly improved the perception of place and manner-of-articulation, especially in the /u/ vowel environment. It also particularly helped with the discrimination of voiceless stops and fricatives.

The effects of coarticulation and context variability are not readily observable with VCV stimuli so the authors felt it important to test the different levels of enhancement with sentence length material. As highest scores from the VCV task were recorded with the final enhancement condition (having the most manipulations), only one enhancement condition was used (but with some minor alterations to the filtering and boosts). In total, 32 participants were tested but split into two groups, with 16 participants tested at 0 dB SNR and 16 at +5 dB SNR. The participants listened to 50 semantically unpredictable sentences which were divided into 25 per condition (natural and enhanced). Twenty-four of the participants also repeated a similar VCV task as per experiment one, but at an SNR of 0 dB only.

Participants gave written responses to each sentence and were scored in terms of number of key words correctly identified (four key words per sentence). Results showed no significant effect of enhancement in the 0 dB SNR condition (74.6 percent correct for natural and 74.1 percent correct for enhanced sentences) but a small, yet significant, improvement with enhancement in the 5 dB SNR condition. However, this improvement was in the region of only 2 percent (scores of 92.4 percent correct for natural and 94.6 percent correct for enhanced sentences) and it does not seem likely that this would translate to any significant real-world benefit. These scores are also close to ceiling level, which would limit any potential of the enhancement strategy. It was also found that in the +5 dB SNR condition, presentation order had a significant effect on scores; with a greater improvement seen between natural and enhanced conditions when listeners heard the natural condition first. More encouraging was that the VCV task was able to replicate very similar results to those found in experiment one (improvement of five percent with enhanced condition).

The authors thought it possible that the limited benefit seen with enhancement may have been restricted as a result of the enhancement strategy causing affricates and approximants to sound unnatural. As such, Hazan and Simpson (1998) decided to conduct a further

experiment in which only plosives, fricatives and nasals were manipulated. In this final experiment, both frication and plosive burst filtering was dropped from the enhancement strategy. Sixteen new participants were tested using the same test protocol as experiment two but at 0 dB SNR only. A significant effect of enhancement was observed, but once again this was only small; in the region of four percent. The authors looked to assess the test-retest reliability by comparing the scores for the natural conditions across both sentence test experiments. They found that the scores did not differ significantly from each other, however, the difference between them (around three percent) is very similar to the difference between the natural and enhanced conditions of experiment three (four percent), which was considered significant.

When the authors looked at scores for individual sentences and individual listeners, they found great variability in the effect of enhancement. It is possible that the sentence material used by Hazan and Simpson (1998) was not phonetically balanced and therefore, considerably varying levels of enhancement may be applied to different sentences. The method also does not explicitly state that subjects were given any familiarisation with the sentence material or conditions. Although order effects were not observed in experiment three, it is possible that participants may benefit more from the enhancement if they are given training. Finally, annotation of the waveform is highly subjective and becomes much more difficult for more fluid speech than for VCV stimuli, and any errors may have led to inappropriate enhancement of the signal in some regions, potentially creating artefacts that confuse the listener. Conversely, if important landmarks are missed during the annotation stage then they will not then they will not benefit from enhancement.

Hazan and Simpson (2000) wanted to investigate the effects of talker (particularly male vs. female) on speech intelligibility with "cue-enhancement". They also wanted to test the "robustness" of cue-enhancement using casually spoken speech which more closely resembles a real world listening environment. The authors recorded two iterations for each of 12 consonant sounds (comprised of plosives, nasals and fricatives) in three separate vowel environments (/a, i, u/) by four speakers with no phonetic training (two male, two female). Three of the four speakers had a south-eastern British accent and one of the male speakers had a north-eastern British accent.

As in their earlier experiment (Hazan and Simpson, 1998), landmark regions were hand-annotated using the waveform for each stimulus.

The following annotations were used:

1. Vowels
2. Nasal onsets and offsets
3. Fricative onsets and offsets
4. Burst transients
5. Aspiration

A similar enhancement strategy was employed with bursts, frication and nasals boosted by 6 dB, aspiration by 12 dB (between the onsets and offsets) and vowel transitions ramped from 2-4 dB across a 5 ms period to help reduce discontinuities in the waveform and blend successive enhanced segments. Fourteen NH listeners were tested at a SNR of 0 dB SNR and listened to three repetitions of the two recordings for each consonant, in each vowel environment and in both the natural and enhanced condition.

Percent correct scores showed a mean overall intelligibility (for all four talkers) of 73.8 percent in the natural condition and 82.9 percent in the enhanced condition. This improvement of around nine percent is greater than was seen in their earlier study. Analysis of the results revealed a significant effect of condition, speaker and vowel environment, and also a significant interaction between speaker and condition. Post-hoc analysis found that results for each speaker differed significantly from all others, with the best scores seen for the two female speakers. Scores with enhancement were significantly higher than for the natural condition for all speakers; this effect was higher for the male speakers and is likely because the female speakers were already more intelligible. As per their 1998 study, participants scored significantly worse in the /u/ vowel environment compared with /a/ and /i/.

Consonant confusion matrices showed that identification of plosives benefited the most from enhancement. Once again, nasals did not benefit from the enhancement. This may be because the authors chose to enhance the onsets and offsets of nasal portions, as they are often classified as stops; however they can also be classified as sonorants, which are signalled by low-frequency minima. Transmission of place-of-articulation, manner-of-articulation and voicing information followed a similar pattern to that of overall intelligibility scores.

Although the participants were given ten minutes of familiarisation with natural and enhanced tokens similar to those used in the actual experiment is not clear whether these were recorded from the same speakers. One of the male speakers was noted as having a

distinct accent from the other three speakers and was indeed the least intelligible speaker. Therefore, results from this speaker may have been skewed depending on the familiarity of the listeners with that talker's particular accent.

A larger improvement was seen in the 2000 experiment compared with the1998 study; however, the authors did not look at the effect of cue enhancement on sentence length material in the 2000 study. Regardless of the fact that the speakers were not given phonetic training, the isolated VCV tokens were likely produced more clearly than similar consonant and vowel sounds within a sentence spoken in a more conversational manner. It would also be interesting to examine the differences in annotations of sentence material to compare the natural salience of particular landmarks between speakers. Although the authors argue that cue enhancement has shown benefit for speech perception in noise, this has yet to be replicated at sentence level, for different SNRs and with different noise types.

Ortega et al. (2000) followed on from previous experiments by Hazan and Simpson (1998; 2000) and looked to move away from hand annotated waveforms and instead develop a method of automatically detecting landmarks or "potential enhancement regions" (PERs). PERs were estimated using a broad-class HMM classifier (as described in Huckvale, 1997) and this included six context-free HMMs representing silence, vocalic, fricative and nasal regions as well as stop-gaps and stop-aspirations. Phonetically annotated, read speech material from the SCRIBE corpus was used to train and then determine the accuracy of the models (using material not used in the training). Rates of misses and false alarms for the detection of burst, fricative, nasal, vowel onset and offset PERs typically ranged between 10-25 percent, with an overall class labelling accuracy of 71 percent (Ortega et al., 200: 45). It is important to note at this stage that the proposed method of automatic detection of PERs was developed for use prior to corruption of the speech signal by noise (i.e. clean spectra).

The authors had to make some changes to the enhancement levels used in the study by Hazan and Simpson (2000) to take into account the most common errors made when estimating PERs. Aspiration, frication and plosive bursts were required to be covered by a single label and a uniform level of 9 dB enhancement applied to all three. Enhancement at vowel onsets and offsets also had to be reduced to just 1 dB at 10 ms slices. In their first experiment, the authors compared intelligibility scores between unenhanced (natural), manually enhanced (as per Hazan and Simpson, 2000) and automatically enhanced speech for word and sentences tests. No significant improvement was observed with enhancement (manually or

automatically labelled) compared with intelligibility of the natural speech and in some cases a negative effect was observed. Again, there was no significant difference between intelligibility scores for the manually labelled and automatically labelled speech, suggesting that potential errors introduced by the automatic detection of PERs do not have a significant impact on intelligibility. It is not clear from this paper whether the speech material were presented in the presence of interfering noise or in quiet.

A second experiment explored the possibility that the first experiment had failed to realise any improvement with enhancement because the gain applied to aspiration, nasal and frication segments had been increased in relation to that use in Hazan and Simpson (2000); and potentially increased consonant confusion by introducing conflicting cues. Participants listened to VCV stimuli (as used by Hazan and Simpson, *ibid*) in SS noise at 0 dB SNR which were either natural, manually enhanced with a 9 dB boost or manually enhanced with a 6 dB boost. For this task an improvement was seen in speech recognition scores when the speech was enhanced (both levels of enhancement) but this improvement was only half that observed in the study by Hazan and Simpson (2000) and there was no difference observed in scores for the different levels of enhancement.

In their final experiment, the authors compared speech intelligibility scores for manually labelled speech, manually labelled speech using the PER labels, automatically labelled speech and unenhanced (natural) speech when listening in SS noise at -5 dB SNR. Once again no significant differences were observed for enhancement speech compared with the natural speech. These disappointing results may have been due to limitations in the study design, including the fact that listeners in the final experiment were split into four groups and heard only one condition each. The authors noted a large variability in scores amongst listeners and provided a much shorter training period prior to testing in the final experiment which may have reduced the effects of enhancement observed in earlier experiments.

Results from these experiments potentially suggest some improvement could be seen with enhancement based on PERs but further investigation is needed into the best method for enhancing each feature. Further studies should investigate individual contribution of features to determine which contributes most to the improvement seen in earlier experiments and then consider how to optimise the other features to further improve performance. Automatic detection of PERs needs to be more sensitive and reliable for use in CIs, particularly because these techniques have been applied to clean speech and with noise added after enhancement.

Further work would also be required to ascertain if there is a similar pattern of results when this method is applied to noise-corrupted speech.

Enhancement of transients in speech

Chapter 3 explored how the auditory system responds well to sudden changes in the speech signal and that these changes correlated well with the concept of acoustic landmarks introduced by Stevens (2002). Yoo et al. (2005; 2007) developed a method for enhancing these transient components of the signal related to transitions between vowels and consonants. This differs from the approach outlined in the studies above as it splits the speech signal into only two categories: quasi-steady-state (QSS) components (linked to vowels and consonant hubs) and transient components. The QSS component is identified and then removed to leave the transient component, which is then amplified and recombined with the original signal.

Yoo et al. (2007) used the Modified Rhyme Test to evaluate the intelligibility of the enhanced speech (using a gain factor of 12) compared with natural speech. They tested eleven NH subjects at six SNR levels (-25, -20, -15, -10, -5 and 0) in SS noise (added after enhancement). Results showed an improvement in speech recognition scores when listening to the enhanced speech in the more difficult SNRs (-25, -20 and -15 in particular). Although no statistical analysis of the results was presented in this study, the improvement from around 20 percent correct to 55 percent correct at -25 dB SNR and 45 percent correct to 70 percent correct at -20 dB SNR would suggest a significant difference between scores. Scores at -5 and 0 dB SNRs were similar for the enhanced and natural speech and were near ceiling level. This may help to explain why results from Hazan and Simpson (2000) and Ortega et al. (2000) showed little to no benefit with their enhancement technique. The results from this experiment are encouraging. The authors also demonstrated in another experiment that although the transient component only comprises about 2 percent of the energy in speech (compared with 14 percent held by the QSS component), when played in isolation it is more intelligible than the QSS component. This would suggest that the loudest parts of speech are not necessarily the most important for speech perception, and could be used to argue for an n-of-m CI speech processing strategy that does not base channel selection on bands with the most energy.

As in the studies by Hazan and colleagues, the transient enhancement strategy has been developed for use in the telecommunications industry and therefore, noise in the experiments by Yoo et al. (2005; 2007) was added after enhancement. So far studies have shown improvement at low SNRs only, and benefit in positive SNRs needs to be demonstrated for

CI users. Performance of the enhancement strategy also needs to be evaluated for different types of noise and for sentence length material. This method possibly could not be directly implemented in CI processing as the authors high-pass filtered the speech above 700 Hz, essentially removing any information about $f0$. This is possible for NH listeners as they are able to make use of the harmonic structure in the higher frequencies to uncover information about $f0$, and indeed the authors showed that the high-pass filtered speech was almost as intelligible as the original speech. However, this may not be suitable for CI users who already receive a significantly restricted signal and do not reliably perceive important formant transitions <1kHz. In addition to this, the authors also noted a large time-delay was introduced with the processing, approximately 40 times the duration of the original speech signal; this would introduce an unacceptable delay for the CI users.

Speech envelope enhancement

The previous chapter highlighted the importance of adaptation by auditory nerve fibres for enhancing regions of sudden change in the speech signal, such as occurs at the onsets of speech sounds (and are closely linked with the concept of landmarks). This adaptation effect is not the same for nerve fibres stimulated by electrical hearing; there is a larger time constant and much smaller amplitude (Geurts and Wouters, 1999: 2477). This is the result of the CI bypassing the inner hair cells and stimulating the auditory nerve fibres directly. Geurts and Wouters (1999) tried to incorporate adaptation into CI processing by enhancing peaks in the speech envelope. This method was based on the CIS speech processing strategy and differed only in the way the envelope was extracted; the strategy was therefore labelled "EECIS". For EECIS, envelopes for each channel are run through two low-pass filters to give the slow envelope (from the filter with lower cut-off frequency) and fast envelope. The slow envelope is amplified and then subtracted from the fast envelope, leaving only the peak signal. This peak signal is amplified and then added to the standard speech envelope (as per CIS) to give the enhanced signal. The amplification of the peak signal is determined for each patient individually to take into account their DRs and resultant T and C-levels (see Geurts and Wouters, 1999 for procedure).

The authors compared intelligibility scores for eight CI users who used an eight channel CIS strategy with their normal CIS processing and for EECIS. The participants listened to stop consonant (/p, t, k, d, b, g/) VCV stimuli in the vowel contexts /a, i, u/ as well as monosyllabic CVC words. The authors chose to evaluate the effects of EECIS on the stop

consonants as adaptation in the healthy auditory system is greatest for these sounds (Delgutte, 1997).

EECIS showed a small, but significant, improvement in the /a/ context only (nine percentage points). Scores with CIS and EECIS in the other vowel contexts did not differ significantly from each other. Feature analysis applied to scores in the /aCa/ condition showed that the small improvement seen with EECIS was primarily due to improved transmission of voicing and place-of-articulation information. Results from the CVC word task found that scores with CIS and EECIS did not differ significantly for the correct identification of initial and final consonants. However, EECIS did appear to significantly improve the identification of medial vowels, though again this effect was only small (seven percent). The authors concluded that there does appear to be some benefit to speech perception with the selective enhancement of certain elements of the speech signal (relating to sudden changes in amplitude), however, this is only small and appears to be constrained to certain contexts. It is also not possible to say from these results that the EECIS strategy is replicating adaptation in a healthy auditory system, as this would require recording the response patterns from auditory nerves with and without the peak enhancement. A suggested improvement to the strategy, proposed by the authors, was that the amplification applied to the peak signal should perhaps be channel dependent, rather than using a uniform gain across all channels.

Koning and Wouters (2012) continued investigations with onset envelope enhancement for CI users, this time considering its benefits for listening in noise. The authors used an envelope enhancement strategy (EE) that was based on the EECIS strategy, however, the peak signal for EE was computed before the envelope compression stage (this was done after compression for EECIS) making it more sensitive to onsets in speech which are very low in energy.

The authors conducted three experiments to help evaluate the EE strategy compared with CIS when listening in noise corrupted speech. Each experiment used 10 NH participants listening to speech processed through an eight channel noise-band vocoder. Experiment one compared speech recognition scores in SS noise with the CIS and EE strategies; the peak signal for the EE strategy was enhanced based on the clean speech signal and then added with the vocoded noise signal. Percent correct scores were measured for participants listening to LIST sentences (Dutch/Flemish sentences) produced by a female speaker in five SNR conditions ranging from + 4 to -4 dB in 2dB intervals. Significant benefits of (ideal) onset envelope

enhancement were observed at -4, -2 and 0 SNRs with increases in scores (compared with CIS) of 29, 30 and 17 percent respectively. The authors also report a significant improvement of 19 percentage points with EE compared to CIS at +2 dB SNR, however, their graph displaying percent correct scores for the two strategies (figure 3, *ibid.*: 2575) would suggest an improvement of around half this magnitude. Scores at +2 dB SNR were not significantly different for the two strategies and this would appear to be in line with previous studies which have shown that selective enhancement of the signal shows greatest benefit for more negative SNRs.

Experiment two looked to assess performance of EE and CIS in SS noise with processing more akin to that performed by an implant in a real-world situation. The peak signal of the EE strategy was therefore calculated from the noise-mixed signal which had been subjected to a Wiener filter (in an attempt to reduce the noise in the signal) and then added back to the noisy signal. The Wiener filter was either given *a priori* knowledge of the noise and speech signals or no *a priori* knowledge (to represent a more realistic real-world implementation) thus creating two conditions for the EE strategy. Once again 10 NH listeners were tested listening to noise-band vocoded LIST sentences for CIS and the two EE strategies using a similar test paradigm as in experiment one. Percent correct scores were significantly better for the EE strategy with *a priori* knowledge built into the Wiener filter than for CIS at -4, -2, 0 and +2 dB SNR and EE with no *a priori* knowledge at -4, -2 and 0 dB SNRs, whilst scores for CIS and EE with no *a priori* knowledge did not differ significantly. Scores for the EE strategy with *a priori* knowledge were very similar to those obtained with EE in experiment one.

The final experiment was designed to evaluate the ability of the EE strategy to cope with more realistic masker types, i.e. two-talker condition where the target speech was produced by a female and the interferer signal was produced by a male speaker. The authors created three conditions for the EE strategy (*ibid.*: 2574):

1. Onsets from the clean target and clean interferer signals were used separately as input envelopes of peak extraction [EE(T)+EE(I)]

2. Noisy mixture used as input to peak extraction [EE(T+I)]

3. Only onsets of the target signal enhanced [EE(T)+I]

For this experiment, Koning and Wouters (2012) computed SRTs for 50 percent correct word identification for the three EE strategies and CIS. Results showed no significant differences in the measured SRTs for CIS, EE(T)+(I) and EE(T+I) with SRTs of 1.8, 1.7 and 1.9 dB respectively. However, the SRT of -0.8 dB measured with the EE(T)+I condition was significantly different from the SRTs measured in all other conditions. Results from these experiments suggest that EE could be used to improve speech recognition scores in noise for CI users. However, an improvement was only seen with the EE strategy under ideal conditions, i.e. when estimation of the peak signal was done in the absence of noise. The authors posit that the improvement seen in these experiments is the result of improved contrast between the speech and noise signals and at phoneme boundaries.

It is very difficult to compare the results of Koning and Wouters (2012) with the other speech enhancement strategies discussed in this thesis (Vandali, 2000; Hazan and Simpson, 1998; 2000; Yoo et al., 2007) because none of the aforementioned studies have tested their strategies using NH participants listening to a CI simulation. However, as in previous studies, enhancement only showed benefit at very low SNRs, with participants reaching ceiling level in the 4 dB SNR condition. van Wieringen and Wouters (2008) calculated 50 percent correct SRTs for the LIST sentences in SS noise for both NH listeners and CI users and found that performance for the CI users ranged from +0.5 dB to +15 dB. In the study presented by Koning and Wouters (2012), calculated SRTs for the CIS simulation conditions in experiments one and two equated to -0.3 dB and -0.7 dB respectively. This could mean that results from these experiments could only be compared with the best performing CI users suggesting the vocoder simulation used is not a good predictor of performance for "poor performing" CI users, and it is these patients who may benefit most from enhancement strategies. It is also not possible to directly compare results from this study with those of Geurts and Wouters (1999) as they used different speech material and only tested CI users in quiet. It might therefore be useful to conduct a similar experiment with the improved EE strategy using CI users listening to a range of speech material in noise.

Koning and Wouters (2012) suggest that a potential benefit of the EE strategy is that it has low complexity and should be fairly easy to implement in a real-time device, without introducing a significant time-delay due to increased processing of the signal. However, results from their study suggest that improvements in speech perception with EE cannot be realised with current noise-reduction techniques implemented in CI devices.

4.4 Gaps in knowledge

Spectral and amplitude changes have been shown to be of great importance to the perception of speech for NH listeners, both in quiet and in noise. However, very little research has been done to determine the contribution of these rapid spectral and amplitude changes and their corresponding ALs to speech perception in noise by CI users. Transmission of landmark information with current CI processing is likely to be poor due to reduced spectral contrast as a result of poor spectral resolution, and a severely reduced dynamic range. Several possible methods for improving the saliency of landmarks, particularly in noise, have been proposed, however many of these techniques have not been designed with CI processing in mind. Furthermore, potential benefits with such strategies seem limited in real-world listening conditions, particularly at SNRs and masker types in which CI users are known to struggle. The literature has therefore raised the following questions:

1. Can landmark enhancement be used to improve speech perception in noise for CI users?
2. What is the best way to improve landmark transmission for CI users?
3. Can benefits of landmark enhancement be realised in positive SNRs?
4. Can ALD algorithms developed for ASR be successfully adapted for use in CIs?
5. Can ALs be accurately detected in noise?
6. Should landmark enhancement focus on one specific set of landmarks (e.g. sonorants vs. obstruent consonants)

The current study will explore how landmarks can be selectively enhanced within the context of CI processing by first detecting the landmarks from within the signal (using ALD or hand-generated labels) and then applying a boost to the channels associated with the landmarks. This type of processing does not require channel-specific landmark detectors as proposed in section 4.3.2 but instead uses the entire signal (similar to detectors outlined in section 4.2.3) to provide a set of confidence values and landmark labels.

The study will focus on obstruent landmarks only for four main reasons:

1. Obstruent consonants are more easily corrupted by noise than sonorant sounds (Parikh and Loizou, 2005), section 3.4.

2. Obstruent consonants are most commonly confused by CI users during listening tests (Munson and Nelson, 2003), section 4.1.
3. Acoustically abrupt landmarks account for 68 percent of all landmarks in speech (according to Liu, 1996), section 4.2.2.
4. Obstruent landmarks (especially stops and fricatives) are more robustly detected by algorithms than nasals or glides (Amit Juneja, personal communication).

Chapter 5- Development of a landmark enhancement strategy

5.1 Introduction

The literature supports the idea that speech perception in noise for CI users can be improved by increasing the transmission of perceptually important information in the speech signal, rather than simply developing implant systems that convey *more* information. Chapter three argued that new cochlear implant processing strategies should focus on the perseveration of rapid spectral and amplitude changes in the speech signal, namely those relating to acoustic landmarks which point to regions rich in information. Chapter four then explored how these landmarks are conveyed, if at all, with current CI processing and introduced the concept of using an automatic landmark detection algorithm to guide a landmark enhancement strategy, with a particular focus on the obstruent landmarks. The following chapter presents a series of experiments that were conducted as part of the development and evaluation of such an obstruent landmark enhancement strategy. An in depth discussion of the overall findings from the experiments is given in chapter 6.

5.1.1 Main aim and hypothesis

Aim: develop a method for identifying and emphasising obstruent acoustic landmarks in noise so as to make them more salient and explore the effect on speech perception scores in noise for NH listeners listening to a CI simulation.

Hypothesis: boosting the speech signal at times of obstruent landmarks will make channels containing information relating to important spectral changes more likely to be selected with an n-of-m strategy (such as ACE), therefore increasing their transmission, and subsequently improving speech perception in noise.

5.2 General method

The following sections outline methods shared between the experiments outlined in this chapter and help to provide justification for certain methodological decisions. Where the methodology for a particular experiment differs from the general method, this will be specified in the introduction for the experiment. All experiments were approved by the

University of Southampton Faculty of Engineering and the Environment Ethics Committee (see Appendix 9).

5.2.1 Speech material

Sentence testing

Open-set speech recognition scores were measured using the BKB sentence lists. The BKB sentences were chosen as they are currently used as a measure of speech recognition for cochlear implant patients in the United Kingdom. The BKB sentence material consists of 21 lists of pre-recorded short open-set sentences, approximately two seconds in duration. The sentences have basic syntactic structure and each list is phonetically balanced. Each list is comprised of 16 sentences contains with three or four key words per sentence, with a total of 50 key words per list. The BKB sentences were originally developed to test the speech recognition abilities of hearing-impaired children and this is reflected in the vocabulary used. Nonetheless, the sentences have high face-validity as this form of speech material is more in line with real-world speech than that of monosyllabic words or even nonsense syllables. Hazan and Simpson (1998: 213) highlighted the importance of testing the effect of cue-enhancement techniques with sentence-length material as sentences contain a greater amount coarticulation and context variability than nonsense syllables.

Although it can be hard to standardise difficulty across all sentences and lists, Thornton and Raffin (1978) propose that one list per test condition should be sufficient to account for the random variation within the test material; meaning any significant differences in scores for different conditions should be the result of the parameters being investigated. Equivalent sentences tests have been developed for both non-English speaking countries and English-speaking countries; for example, the BKB sentences were adapted to create the HINT sentences which are commonly used in America. This should mean that results from this study should be comparable with those which have used the HINT sentences.

Practice lists were comprised of the Institute for Hearing Research (IHR) sentence lists (McLeod and Summerfield, 1990) and are equivalent to the BKB sentence lists. The IHR sentences are comprised of 18 lists of 15 sentences with three key words per sentence and have the same syntactic structure as the BKB sentences. They have also been recorded using the same male speaker as for the BKB sentences. The IHR sentences were chosen for the practice lists so as to remove the need to repeat sentence material during the test and therefore

helped to minimise learning effects. The present study used the male speaker recordings for both the BKB and IHR sentences as there are only male recordings available for the IHR lists.

The BKB and IHR sentence lists can be found in Appendix 1 and 2 respectively. Key words in each sentence have been underlined.

"*Children like strawberries"* is an example of a sentence taken from the BKB sentences.

"*He hid his money"* is an example of a sentence taken from the IHR sentences.

Consonant recognition task

For some experiments, participants also listened to VCV stimuli. VCV tasks allow researchers to look at the effect of signal manipulation on a CI user's ability to resolve spectral and temporal information (Loizou, 1998) and consequently the perception of specific speech cues (relating to voicing, place and manner) by computing consonant confusion matrices (Miller and Nicely, 1955). Such stimuli are closed-set, meaning that subjects are encouraged to guess when they are not sure of which is the correct response. The consonant material consisted of 20 isolated /VCV/ syllables: /p, t, k, b, d, g, f, θ, s, ʃ, v, z, tʃ, dʒ, m, n, l, r, j, w/. Although the sonorant consonants /m, n, r, l, w, y/ were not enhanced in any of the experiments (except for experiment I), they were included in the list of available response. Three recordings of each of the selected consonants were used (Hazan and Simpson, 1998).

For the first experiment, participants were tested in a single vowel context, /aCa/, produced by a male speaker with a standard southern British accent. However, section 3.2 highlighted the importance of contextual information held at vowel-consonant boundaries, therefore, in the final experiment participants were tested in three vowel contexts: /aCa/, /iCi/ and /uCu/. This larger inventory was chosen so as to get a better understanding of the effects of specific processing on particular speech contrasts and whether the processing applied is affected by vowel environment (something which is not possible to assess with open-set sentences). These experiments used stimuli produced by a female speaker as there were more recordings available than for the male speaker.

5.2.2 Noise

The majority of studies investigating CI speech perception in noise have focused primarily on the use of speech and speech-like maskers. This is because competing speech signals pose the greatest difficulty for CI users and for the noise reduction and speech enhancement

algorithms used in CI processors. These maskers can range from speech maskers, such as a single competing talker to speech-like maskers such as SS noise (filtered white noise) and gated noise, which have similar spectral and/or temporal characteristics as the target speech signal. The decision of which masker type/s to use will depend on the specific question being asked. Competing speech signals provide both "energetic" and "informational" masking of the target signal. Energetic masking occurs at the periphery of the auditory system and arises from spectral overlap between the target and masker signal; rendering one or both signals inaudible to the listener (Brungart, 2001). Informational masking occurs at a higher more central level of the auditory processing pathway and is the result of the listener being unable to delineate the masker and target signal due to linguistic similarities (Durlach et al., 2003).

As the number of contributing speech sources increases, the similarity of the masker and target decreases and as the SNR increases above 0 dB, the contribution of informational masking decreases and the main masking component comes from energetic masking (Freyman et al., 2004). For multi-talker maskers (or babble), such as eight or more talkers, the ambiguity between the target and masker signal is lost, as the streams of the competing talkers are indistinguishable, and the resulting spectrum of the masker signal is almost completely continuous. For more continuous maskers such as babble and SS noise, the cognitive load is reduced and masking occurs primarily at the periphery. The highest degree of informational masking will occur in a two-competing talker scenario (Freyman et al., 2004).

As the present study is concerned with the issue of signals competing for limited bandwidth with current CI processing, it was decided to use maskers providing primarily energetic masking. SS noise (white noise filtered to have the same long-term average spectrum as the BKB sentences, as spoken by a male speaker) was chosen as it has no linguistic content, is continuous, and has little to no uncertainty in the signal; this means that masking occurs mainly due to energy from the target and masker signals occurring in the same critical band. An eight-talker babble noise (four male, four female) was also chosen because it is more representative of a real world environment e.g. a busy restaurant. The small amplitude and spectral fluctuations, as well as some linguistic content, will introduce a small element of informational masking. The babble noise was only used in experiments I and II of this study as no significant differences were found between the two masker types. The decision to continue with the SS noise was made so that results from further experiments could more easily be compared with the studies conducted by Li and Loizou (as outlined in Chapter 4).

For stimuli mixed with the babble noise, a section of the babble noise was randomly selected and mixed with the sentence (at the appropriate SNR). For both babble and SS noise, the noise added started 500 ms before the beginning of each sentence and VCV stimulus and stopped 500 ms after the end of the sentence/VCV stimulus to avoid masking overshoot effects (Gelfand, 2004), For all experiments, noise was always added before the enhancement stage to give a more realistic estimation of any potential improvement.
Although fixed procedures are more sensitive to ceiling and floor effects (Lutman, 1997), fixed SNR levels were used in all experiments rather than an adaptive procedure. When using adaptive procedures, results with different processing conditions are compared in terms of SRTs, the SNR required to achieve a particular score (e.g. percent correct). Results are therefore discussed in terms of "SNR improvement" rather than an overall increase/decrease in percent correct scores. However, due to the nature of the processing applied during this study (enhancement applied after the addition of noise) it would have been complicated and less time-efficient to have used an adaptive approach and therefore fixed SNRs were chosen. Using fixed SNR levels also allowed for the enhancement strategy to be tested at SNR levels in which CI users commonly struggle. The exact SNR levels were chosen based on results from pilot studies.

5.2.3 Cochlear implant simulation and the Nucleus MATLAB Toolbox

Section 2.4 has shown that neither vocoder type (sine or noise-band) has been proven to most accurately predict/resemble actual CI performance however noise stimulates a broader section of the basilar membrane than a tone and therefore may represent stimulation with an actual implant more closely. The present author wanted overall performance in the listening tasks to be sufficiently low at more positive SNRs (e.g. +10 dB) so as to better represent conditions in which CI users have difficulty (realistic) and it was felt this could be better achieved using a noise-band vocoder (reduce ceiling effects).

It is possible to use a research processor with CI users, allowing the researchers to manipulate the speech processing strategy without having to make any changes to the patient's implant actual processor (see Li and Lutman, 2008 and Verschuur, 2009). However, implant users typically require a period of acclimatisation when changes are made to their processing strategy, necessitating the development of a take home strategy (see for example Vandali, 2001) or multiple training sessions with the new strategy; neither of which would have been practical in the time-frame of this project. Ultimately, the simulation experiments conducted

as part of the present study serve as pilot studies in the development of the optimum processing parameters of the landmark enhancement strategy and as such it would not have been practical to have used actual implant users during this period. If a beneficial effect of the processing was found then the next stage would be to trial it with actual implant users.

Stimuli for all experiments were processed to simulate listening with a CI using the Nucleus MATLAB toolbox (Swanson and Mauch, 2008). Developed by Cochlear Ltd, the NMT is a research tool which models the processing of the Nucleus 24 implant to create acoustic simulations of what it is like to listen with a cochlear implant. It models front end, filterbank and sampling and selection for the ACE, CIS and SPEAK processing strategies. All stimuli recordings were down sampled to a rate of 16 kHz to be in line with the typical sampling rate used by the Nucleus 24 implant and ACE processing strategy. Stimuli were then mixed with the appropriate noise (eight-talker babble or SS noise) at the required SNRs and then landmark enhancement applied (to be explained further in following sections).

The NMT was then employed to create the CI acoustic simulation for each enhanced noise-corrupted sentence. Firstly the enhanced sentence was pre-emphasised (low pass below 1200 Hz with 6 dB/octave roll-off) to simulate the response of the processor microphone. The signal was then band-pass filtered using a 128 point FFT filterbank, resulting in 62 useful bins, spaced linearly and each with a width of 125 Hz. These bins were then summed and weighted to give 20 frequency bands. The envelope of the signal in each frequency band was extracted by full-wave rectification and low-pass filtering (4^{th} order Butterworth filter) with a 400 Hz cutoff frequency. The channel amplitudes were then estimated from the RMS energy of the envelopes. For the first three experiments, the 12 channels with the highest amplitude were selected to modulate the band-pass filtered white noise carrier using a resynthesis filterbank with cut-off frequencies identical to those used by the analysis filter. The result was an acoustical representation of a 12-*of*-22 ACE strategy. For the final two experiment (for reasons which will be discussed later), a 3-*of*-22 ACE strategy was used. Figure 5.1 shows a block diagram of the ACE strategy as implemented by the NMT, and used in this study.

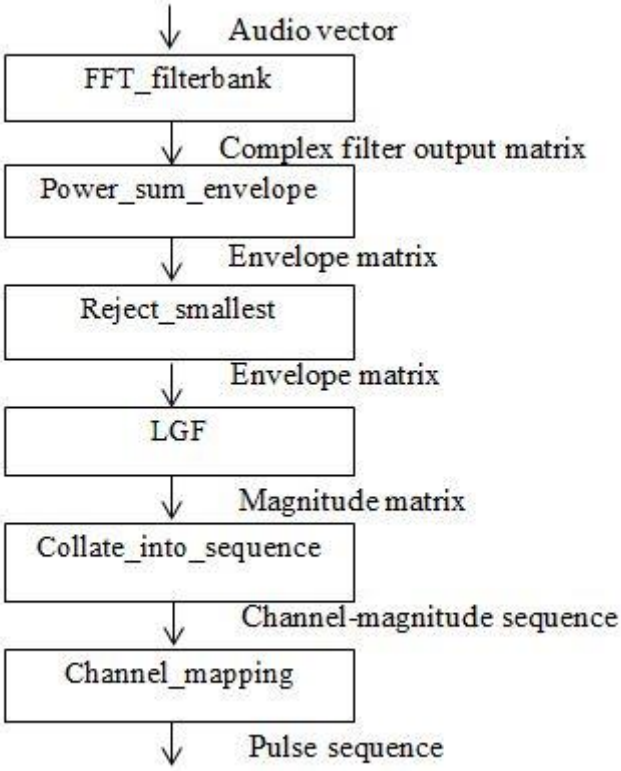


Figure 5. 1 Nucleus Matlab Toolbox implementation of the ACE processing strategy. After Swanson (2008).

5.2.4 Apparatus and calibration

For all experiments subjects were tested in a quiet, sound-attenuated booth. Stimuli were generated offline and presented using a Dell Latitude E4310 laptop, routed through a Creek Audio OBH-21SE headphone amplifier and presented monaurally through HDA 200 Sennheiser circumaural headphones. Monaural presentation was chosen as the majority of adult implant users in the UK are implanted unilaterally.

For all experiments, subjects were tested at a level of 65 dB (A), to simulate listening at a comfortable conversational level. To ensure that the level of each stimulus presentation was at 65dB (A), speech-shaped noise with the same long-term RMS level as that of the BKB sentences (averaged across the lists) was played through the headphones and the sound pressure level (dB (A)) measured using an artificial ear with a flat plate coupler. Volume levels within the test system were altered as necessary to ensure that the level of the test stimuli was at 65dB (A), as measured in the coupler. Identical volume settings and transducers were used throughout testing and daily checks were performed by the tester to ensure the apparatus was working correctly and that the stimuli were free from distortion.

5.2.5 Presentation and scoring

For both open-set speech recognition and VCV tests, stimulus presentation and scoring was done using Praat (version 5.3.32), a speech analysis and software testing package (Boersma and Weenink, 2011). For speech testing, a PRAAT script file (see Appendix 3) played each sentence and following the participant's response, allowed the tester to select the number of key words correctly identified (cross-referenced with the BKB/IHR transcript) and after the selection had been made, the next sentence was then presented. During test conditions, responses were scored loosely, meaning that tense and plural errors were not scored as incorrect responses. No feedback was given regarding the correct content of the sentences. The participant was not able to see the screen during testing.

During testing in the VCV task, the PRAAT script provided participants with a graphical interface (shown in figure 5.2) which provided a display of response alternatives. Participants were asked to select the consonant they heard by clicking on the corresponding symbol using the mouse. Once the participant clicked on the symbol the next VCV stimulus was presented. Again, no feedback was given regarding whether or not the participant had selected the correct consonant for each presentation. The VCV PRAAT script (Appendix 4) also randomised the presentation order of the stimuli.

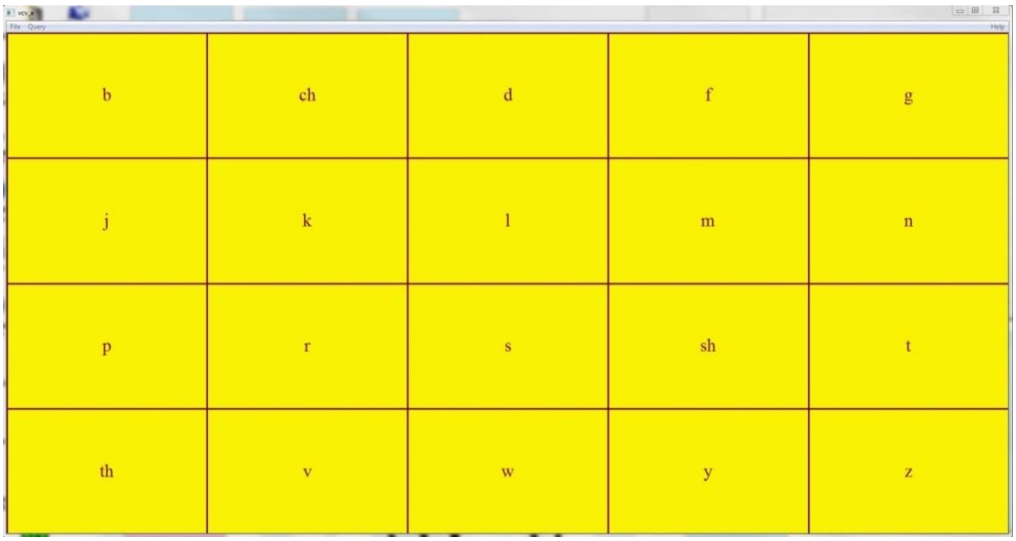


Figure 5. 2 Screenshot of the PRAAT GUI used for VCV tests- 'j' was used to represent the consonant /dʒ/ and 'y' was used to represent the consonant /j/.

5.2.6 Analysis

Where appropriate, consonant confusion matrices were computed (Miller and Nicely, 1955) for the VCV tasks. The responses for each condition, as recorded by the PRAAT software, were imported into Excel and ordered alphabetically by the stimulus label (e.g. b b b ch ch ch

d d d….) and saved as a .txt file. The consonant confusions were computed using the SCORE computer programme which has been developed by the Department of Phonetics and Linguistics at University College London (www.phon.ucl.ac.uk/resource/software.html). SCORE utilises a feature matrix to categorise the errors made according to a set of pre-determined phonological features. Errors can be categorised according to the features voicing, place and manner, however, in addition to these three categories, the features nasal (experiment I only), fricative and plosive were also used in this study. The fricative and plosive features were included in the matrix to determine whether the landmark enhancement applied to the stimuli provided better distinction between these two manner categories. The feature matrix used in the experiments is given in appendix 7.

Consonant confusions can be reported in terms of percentage correct for each feature; however, the different features have different chance scores based on the number of sub-categories for that feature (this may differ between languages and between studies). Reporting consonant confusions in this way makes it difficult to confidently compare the perception of specific features. Computing the percentage information transmitted is considered a more robust comparison of perception across features and this is achieved via a method known as sequential information transfer analysis (SINFA). SINFA analysis is a statistical method proposed by Miller and Nicely (1955) and further developed by Wang and Bilger (1973), and is based on principles similar to multiple regression; using recursion to partial out the independent contributions of the phonological features defined in the feature matrix. These recursions give a series of "iterations", with the first iteration giving the unconditional estimated information transmitted for each of the defined features; therefore adjusting for the different chance levels. Where indicated, percentage information transmission was therefore used to analyse feature transmission for the VCV tasks undertaken in this study. Single-iteration SINFA analysis was achieved using the FIX programme (also developed by the Department of Phonetics and Linguistics at University College London).

For all experiments percentage scores were converted using RAU arcsine transform before statistical analyses (Studebacker, 1985). In order to determine whether parametric statistical analyses could be used for each of the five experiments, the Shapiro-Wilk test of normality was conducted for the RAU score for each independent variable and interpreted in conjunction the corresponding histograms to determine whether the data were normally distributed. The assumption of normal distribution was not broken for any variable in any experiment, except for

the feature "nasal", in experiment I, which had a negatively skewed distribution. Mauchly's test of sphericity was also assessed for each variable and indicated that variances were not significantly different.

The consonant confusion tasks conducted in experiment I and V resulted in a large number of dependent variables following the calculation of percentage information transmission for each of the selected features. Therefore the appropriate statistical technique for analysing was considered to be a multivariate analysis of variance (MANOVA). Although the feature "nasal" violated the assumption of normal distribution, Howell (2012) has shown the F test used in the MANOVA technique is robust to the problem of a skewed distribution. As only one variable violated the assumption of normal distribution it was concluded that the MANOVA would be appropriate.

As there was only one dependent variable for the sentence recognition tasks (the transformed percent correct scores) in each experiment and because the same participants were used in each condition, a repeated-measures analysis of variance (ANOVA) was used to compare means for each condition. Sidak's *post-hoc* analyses were performed to identify any significant relationships between conditions, whilst controlling for type I errors as the result of making multiple comparisons.

5.2.7 Participants

All participants were recruited from students at the University of Southampton, primarily from within the Faculty of Engineering and the Environment. Due to the nature of speech testing, all participants were required to have normal hearing and be native English speakers. The following inclusion criteria were used for recruiting participants across all experiments:

- Aged between 18-35 years
- Normal audiometric thresholds (assessed prior to testing via a hearing screen at 0.5, 1, 2 and 4 kHz)
- Good ear health- no current or recent ear infections, ear canal obstructions (including occluding wax). Assessed via otoscopic examination.
- No recent noise exposure (within the past 24 hours)
- Free from troublesome tinnitus
- No reported difficulties listening to speech in noise

Where possible, participants with no (or very little) experience with the sentence material were recruited. When it was not possible to only recruit participants with no experience of the BKB and IHR sentences this was noted so that results between the experienced and non-experienced listeners could be compared if necessary. For the final experiment, five of the original nine participants were retested but care was taken not to reuse any of the same sentence lists during testing. An example of the participant questionnaire and consent form are given in Appendix 5 and 6 respectively.

5.3 Experiment I- Enhancement of high-frequency fast modulations in speech

5.3.1 Introduction

It is possible that obstruent landmarks may be enhanced by a method which does not rely on the specific detection these events. Improving the transmission of obstruent landmarks has been highlighted as the main focus of the current study as obstruent consonants are more commonly confused by CI users, particularly in the presence of background noise. Obstruent landmarks are characterised by rapid spectral and amplitude changes in the high frequency regions of the speech signal, therefore modulation enhancement may be an appropriate method by which to target, and boost these rapid changes in amplitude.

Nagarajan et al. (1998) theorised that children who are language learning-impaired (LLI's) may be limited in their ability to recognise and process elements of the speech signal which are both short in duration and that change rapidly, for example, formant transitions that may help to distinguish between two stop consonants. The authors therefore developed an algorithm that helps to amplify the faster phonetic elements in speech in the modulation region of 3-30 Hz. The aim behind the algorithm was to enhance the rapid changes in speech in order to help improve the segmentation of successive phonetic elements. Initial results with the algorithm showed that, with training, the LLI children were better able to identify rapidly successive sounds and fast consonant-vowel stimuli.

Li and Loizou (2010) have suggested that the transmission and therefore perception of obstruent landmarks may be improved with increased spectral contrast. ALs are closely linked to changes in manner of articulation (Stevens, 2002) and therefore it is possible that enhancement of modulation frequencies relating to envelope information (2-50 Hz) may help to improve spectral contrast between successive segments in speech and therefore improve

the transmission of landmarks for CI users. Modulation enhancement should be applied only in the high frequencies as obstruent landmarks are found within the high frequency regions of speech. Formant transitions (F1 and F2) which are important for vowel identity and act as place of articulation cues for consonants are also characterised by fairly rapid changes in the spectrum (though not as fast as obstruent landmarks). As the focus of the study was on obstruent landmarks, it was considered that modulation enhancement should be applied to frequencies above the range of F2. It is also not clear what degree of enhancement will be required to improve the saliency of obstruent landmarks.

This first experiment was therefore conducted to explore the use of modulation enhancement to boost rapid spectral changes in speech related to obstruent ALs, when listening to CI simulated speech in noise. As modulation enhancement is being explored as a method which targets but does not specifically select obstruent landmarks for enhancement, processing was applied to the entire speech signal; incorporating all vowel and consonant segments.

5.3.2 Method

Fourteen native English speaking NH listeners participated in this experiment (seven male, seven female). Subjects were paid for their participation and their ages ranged from 18-29 years. Three recordings of each of the 20 VCV stimuli (male recordings) were used in the context /aCa/. The speech tokens were corrupted with eight-talker babble and SS noise at +5 and 0 dB SNRs and processed to simulate a 12-*of*-20 ACE speech processing strategy.

For both types of noise and both levels of SNR, each of the subjects was tested in four conditions, plus an additional condition which measured their speech recognition scores in quiet (un-corrupted with no additional processing). For the first condition, the noise-corrupted stimuli were left unaltered (control condition). For the other three conditions, fast amplitude changes in the region of 3-100 Hz, in the high frequencies (2.5-8 kHz) were boosted by 5, 10 or 20 dB. These conditions will be referred to as Mod5, Mod10 and Mod20, respectively. These parameters were chosen to attempt to enhance the fast spectral and amplitude changes associated with manner of articulation that occur for obstruent landmarks. A high-pass cut-off of 2.5 kHz was chosen to minimise the chances of enhancing transitions relating to the lower formants (F1 and F2), which are important cues for place of articulation and for vowel identification. Processing was done in Praat using the deepen band modulation function which is based on the algorithm presented in Nagarajan et al. (1998).

5.3.3 Results

Mean total percent correct scores for all conditions are shown for SS noise and babble in figure 5.3. A three-way repeated measures ANOVA was performed with three factors (SNR, noise types and level of enhancement) and two levels for SNR (5 and 10 dB) and noise type (babble and SS noise) and four levels for level of enhancement (none, 5, 10 and 20 dB). Analysis revealed a significant effect of SNR ($F(1,209)$=32.35, p= 0.00), no significant effect of noise type or level of enhancement (p>0.05) and no significant two or three-way interactions.

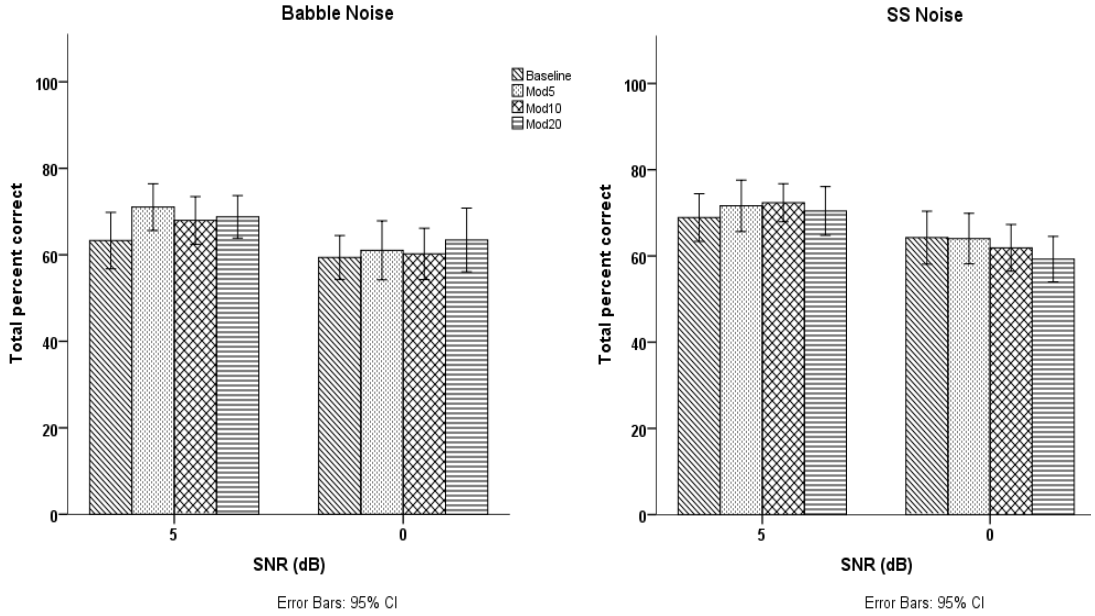


Figure 5. 3 Mean percent correct scores for SS and babble noise for all conditions.

Consonant confusion matrices were computed for each of the conditions (excluding the quiet condition) and subjected to information transmission analysis for the features voicing, place, manner, nasal, plosive and fricative.

A MANOVA was performed with six dependent variables (percentage information transmission for the features voicing, place, manner, nasal, plosive and fricative), factors were as before. Analysis showed a significant effect of SNR for all features (p<0.05), a significant interaction between noise type and enhancement for several features (p<0.05) but no significant three-way interactions. A two-way repeated measures ANOVA for SS noise showed a significant effect of SNR for all features (p<0.05), no significant effect of

enhancement for any feature (p>0.05) and no significant two-way interactions. However, a two-way repeated measures ANOVA for babble noise showed a significant effect of SNR for all features except for the fricative feature ($F(1, 153)$= 2.95, p= 0.089), a significant interaction effect of level of enhancement for the features nasal, plosive and manner (p<0.05) and no significant interactions between SNR and level of enhancement.

Sidak's *post-hoc* test revealed that nasal transmission significantly increased in the Mod5 condition, compared with the baseline condition (p= 0.001) and the Mod10 condition (p= 0.026) but not the Mod20 condition. Scores in the Mod10 and Mod20 conditions did not differ significantly from one another or the baseline condition. Plosive transmission was significantly higher for all three enhancement levels compared with the baseline condition (p< 0.05), however, percentage information transmission for the three enhancement levels did differ significantly from one another. Manner transmission significantly increased (p<0.05) from baseline in the Mod20 condition only (p= 0.018) but percentage information transmission did not differ significantly between the three levels of enhancement. However, the improvement seen in the Mod5 condition was only just above significant level at p= 0.063. Figure 5.4 shows the mean percentage information transmission for each of the six features in babble noise, for both SNRs and for each level of enhancement.

5.3.4 Discussion

The improvement in consonant identification in babble noise may be due to listeners being able to make use of temporal dips in the masker to glimpse information relating to these features. Boosting the faster modulations in the signal may have improved the SNR at the times of these dips as the signal should have been boosted whilst leaving the noise unchanged. Unlike fluctuating maskers such as babble noise, SS noise does not have temporal dips and this is possibly why no improvement was seen for any feature with this type of noise. Greater improvement might be seen for more fluctuating maskers, such as two or four-talker babble; however, eight-talker babble may be more representative of real-world listening situations.

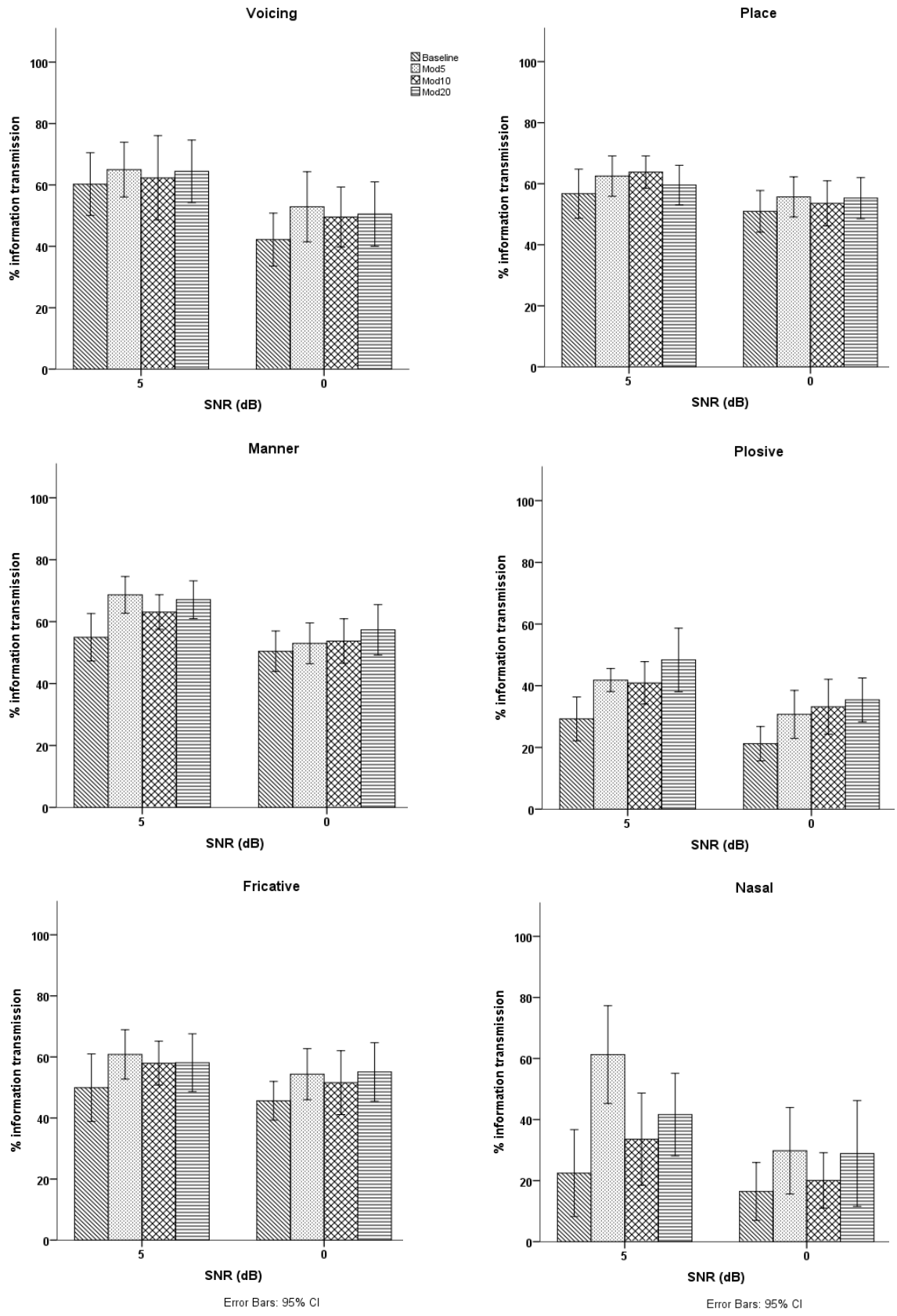


Figure 5. 4 Percentage information transmission for the features voicing, place, manner, fricative, plosive and nasal in babble noise at +5 and 0 dB SNRs for the baseline condition and three levels of enhancement.

An improvement in the transmission of manner information is an encouraging result as ALs are linked to manner of articulation cues in speech. The improvement in the transmission of plosive information may be the result of enhanced spectral information relating to the burst portion of the consonant which helps with place of articulation distinctions. Slightly more unexpected was the improvement in nasal information transmission. Although nasals are sometimes classified as stop consonants, enhancement in the high frequencies should potentially be of little benefit for nasal identification as they are characterised by predominantly low-frequency information. However, it is possible that some of the higher formants (e.g. F3 and F4) may have been enhanced and these higher formant transitions can be useful for classification.

Initial results with the TESM strategy (Vandali, 2000), also showed an improvement in the transmission of nasals and stops. However, unlike the TESM strategy, no improvement was seen in the transmission of fricatives with modulation enhancement. Further to this, an improvement was not seen for either place or voicing features in the current experiment, suggesting that even though listeners may have had better access to landmarks, it is possible that they were still unable to resolve the spectral information that could help them distinguish between different consonant sounds. Nonetheless, it is important to note that enhancement, of any level, did not significantly reduce the transmission of any feature for either noise type.

For features where information transmission did improve, no effect of level of enhancement was observed. However, for the features nasal and manner the 10 dB enhancement condition did not significantly improve transmission compared with the baseline condition even though an improvement was seen for both 5 and 20 dB enhancement. It is not clear why the 10 dB enhancement condition did not always provide a significant improvement but these results would appear to suggest that a 5 dB enhancement is sufficient to give an improvement for some features and that a greater boost (i.e. 20 dB) is neither more beneficial nor detrimental to the improvement seen.

It is possible that no improvement for overall scores was seen due to the “blanket-style” processing that was applied to each VCV (modulation enhancement during both vowel and consonant segments). Li and Loizou (2009) discussed how landmarks can help listeners to identify syllable boundaries but that the important vowel-consonant ratios required for this may be blurred by CI processing due to their limited DR. The current experiment aimed to boost ALs and improve spectral contrast between segments, however, the high frequency

regions of the vowel spectra were also boosted, and therefore this type of processing probably did not help to enhance boundaries.

Figure 5.5 shows a broadband spectrogram for the stimulus a/t/a in +5dB SNR babble noise and the corresponding electrodograms for the unprocessed condition and 20 dB enhancement. Electrodograms can be used to visualise the information transmitted with CI processing by showing the activation pattern of electrodes. Electrode number is given on the y-axis with low-frequency (apical electrodes) at the bottom and high-frequency (basal electrodes) at the top. Time is shown along the x-axis and amplitude is represented by the height of the vertical bar, indicating channel activation in a given analysis frame.

Point (A) in the spectrogram demonstrates that energy during the burst portion of the /t/ is concentrated in the high frequencies (particularly around 5 kHz). At the same point in the electrodogram for the unprocessed condition there is very little activation in the high frequency channels with more dominance in the low-frequency channels; potentially making it harder to resolve not only the acoustic landmark at the onset of the burst but also other important spectral information which may help with determining place of articulation cues. Looking at the same point in the electrodogram corresponding to the 20 dB enhancement condition, activation of the high frequency channels has not only significantly increased but has also been extended into channels which were not activated in the unprocessed condition (e.g. channel four). This also means that activation of the low frequency channels has also decreased. It is possible that this increased activation in the high frequency channels is the reason why information relating to the feature plosive was improved with modulation enhancement.

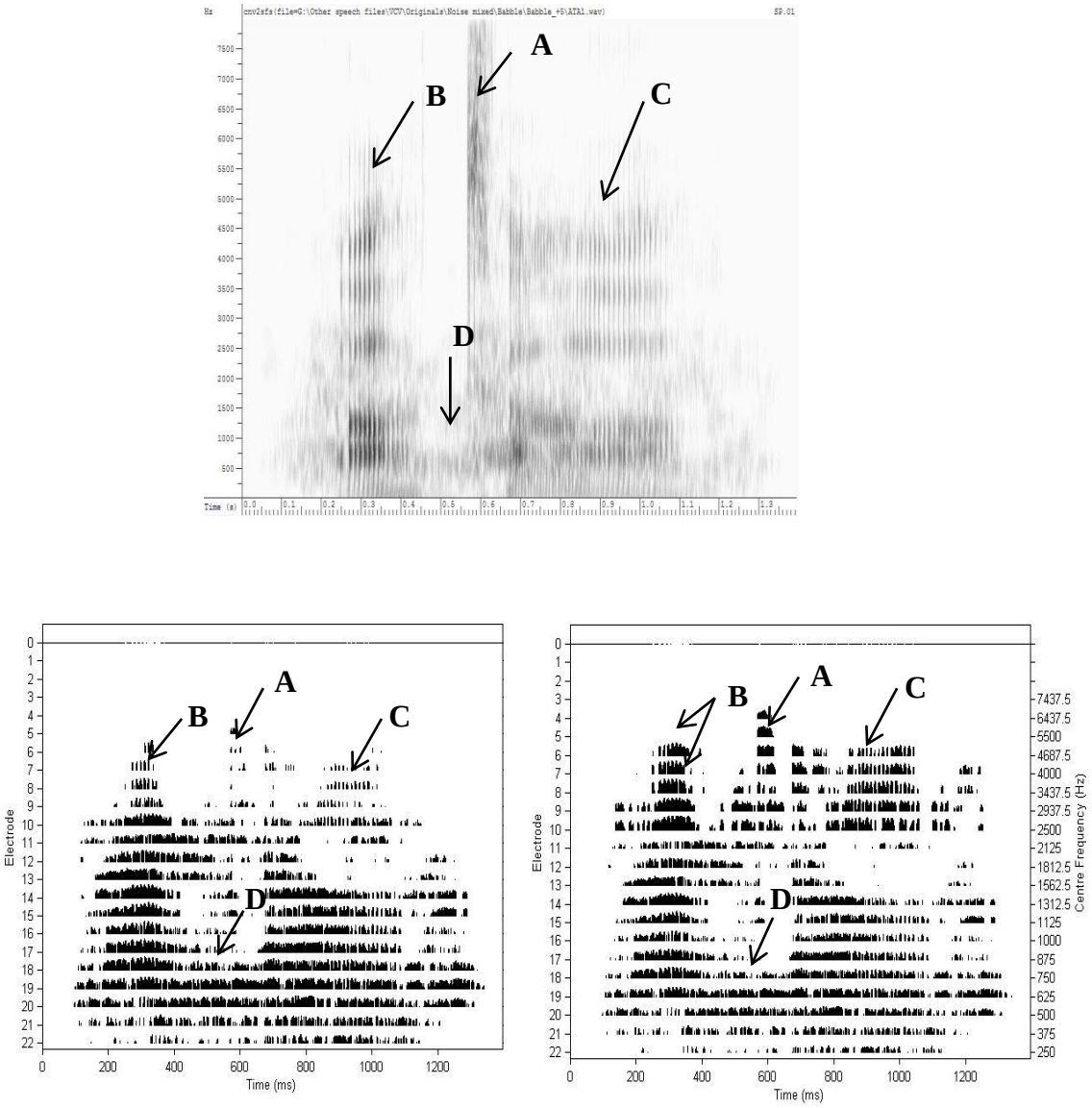


Figure 5. 5 **Broadband spectrogram for the stimulus /ATA/ in +5 dB SNR babble noise (top panel). Electrodograms for the same stimulus for the unprocessed condition (bottom left panel) and for the Mod20 condition (bottom right panel) are also shown. (A) Indicates the time of the burst, (B) and (C) indicate energy in the high frequency region of the vowel segments and (D) indicates the contribution of noise to the energy seen in the low frequency region of the "silence" during the stop.**

The points labelled (B) and (C) indicate the energy in the high frequencies during the vowel segments (/a/). The spectrogram shows there is little energy in the high frequencies above F5 (around 4.5 kHz). In the electrodogram for the unprocessed condition there is very little activation of channels above the frequency of about F3 (around 2.5 kHz). Typically this would not be expected to affect the perception of the vowel as the first three formants are regarded as being most important for vowel identity. However, in the electrodogram corresponding to the 20 dB enhancement condition shows that emphasis has been taken away

from the lower frequency channels (and therefore the lower formants) and that there is greater activation in channels corresponding to F4 and F5. Although the study was not concerned with the effect of modulation enhancement on the perception of vowels, the increased activation of high frequency channels during the vowel segments will mean that, even though the burst of the /t/ has been enhanced, the vowel-consonant ratio will be reduced. Ultimately, by also enhancing fast spectral changes during vowel segments, relating to the higher formants, spectral contrast is not improved and this is potentially the reason why greater improvement was not observed during this study.

Point D indicates the contribution of noise in the lower frequencies during the period of silence prior to the consonant release. Due to the increased activation of high frequency channels for the 20 dB enhancement condition, there is less emphasis on the low frequency channels during the stop closure. However, there does appear to be more activation in the high frequencies for the 20 dB enhancement condition, prior to the consonant release; potentially introducing conflicting burst cues.

Manner of articulation information has been linked to modulation rates to 2-50 Hz however in the current experiment modulations up to 100 Hz were enhanced, to include some (limited) voicing information which has been shown to help manner of articulation distinctions. It is possible that the modulation range selected was not optimal for enhancing landmarks and therefore further research should look to determine which modulation frequencies would be most beneficial to boost. As the processing is applied in a generic way to the speech signal (and not specifically around the time of landmarks), further research may also help us to determine which modulation rates allow for better targeting of landmarks.

Another explanation as to why little improvement was seen may be that some subjects had very little experience with CI simulations and generally tended to struggle more than those with some experience. Indeed, scores (even in quiet) tended to vary greatly between participants. One possibility is that due to their inexperience with the simulation, subjects may have not been able to exploit the subtle changes in the enhanced conditions. Although participants were given time to familiarise themselves with the simulated stimuli, they were given examples of the quiet, unprocessed syllables only. Future studies should provide more extensive training, particularly in noise, to ensure that participants are familiar with the sound of CI simulated speech. This is important, particularly for experiments which may test actual CI users because although participants should be those who have fully acclimatised to

listening with their implant, they may require time to get used to listening to speech material with a slightly different sound quality. It therefore seems reasonable that subjects listening to CI simulated speech should also be given adequate time to familiarise themselves with the quality of the signal in future experiments and that this may need to be monitored in some way.

5.3.5 Conclusions

Modulation enhancement can be used to improve the transmission of some features, particularly plosives and nasals. Overall, the results indicate that a more selective approach to landmark enhancement may be more beneficial; applying boost only at times when obstruent landmarks occur. This may help to improve V-C ratios, and therefore listeners' abilities to utilise these important acoustic events. An approach that applies enhancement specifically to obstruent landmarks (e.g. using ALD) may therefore prove more effective at increasing spectral contrast and also the perception of ALs. Results from this experiment can be compared with the performance of specific landmark enhancement to help better understand whether specific enhancement, using ALD, is more beneficial for improving speech recognition scores in noise for CI users.

5.4 Experiment II- Specific landmark enhancement using Automatic Landmark Detectors

5.4.1 Introduction

Experiment I showed some, albeit limited, benefit of enhancing ALs for speech perception in noise for NH listeners listening to CI simulated speech. However, results suggested that a more selective approach to landmark enhancement, such as that explored by Hazan and Simpson (1998; 2000), may yield better outcomes. For use in real-time CI processing, a method which selectively enhances obstruent landmarks needs to incorporate a process by which these events are automatically detected from within the incoming signal. A number of ALD algorithms were discussed in Chapter 4 and the following experiment outlines the development of a method to automatically detect and enhance obstruent landmarks by adapting the ALD algorithm proposed by Juneja and Espy-Wilson (2008). Further to this, the development of a method for enhancing the detected obstruent landmarks is also discussed and results from a preliminary experiment using this method are presented.

5.4.2 Development of the landmark detection toolkit and boost package

Landmark detection toolkit

A landmark detection toolkit (LND) was developed for this study by Amit Juneja and is based on previous work for ASR (Juneja and Espy-Wilson, 2008). The development of the LND is outlined below (the details of which were provided by Amit Juneja, (personal communication).

The landmark detection toolkit applies four binary classifiers in each analysis frame for the features sonorant, syllabic, plosive and silence. The posterior probabilities from these classifiers are then combined in a probabilistic phonetic hierarchy to obtain frame probabilities of vowels, stop bursts, fricatives, sonorant consonants and silence and a dynamic search algorithm is then applied to obtain the landmark sequences. In the original implementation of the landmark detection toolkit (Juneja and Espy-Wilson, 2008), knowledge-based APs were used by each classifier meaning that the parameters could be different for each classifier. The newly developed toolkit can use any set of acoustic parameters and contains default models for mel-frequency filterbank parameters computed from the Hidden Markov Model Toolkit (HTK) at 16 kHz (to be in line with CI processing). Mel-frequency filterbank parameters have been used as they closely resemble filterbank settings used in a Cochlear Nucleus implant processor and are commonly used in CI processors (Loizou, 1998), with narrower bandwidths used in the low frequencies and wider bandwidths in the high frequencies. Ideally, the filterbank parameters of the LND toolkit would match those of the analysis filterbank of the CI device in which it was being implemented; however, this was not possible for this experiment due to copyright issues.

In an attempt to improve the noise-robustness of the landmark detection system for use in cochlear implant processing, cepstral mean subtraction and variance normalization were applied to the mel-frequency filterbank energies. To apply the cepstral mean subtraction, the filterbank energies for each channel were observed across the whole utterance and the mean and variance for each channel was computed. Following this, the mean was subtracted from each channel and the resulting values were divided by the standard deviation for that channel. The SVM models that the LND toolkit requires were then re-trained using the TIMIT database and the filterbank coefficients computed using the HTK.

The LND toolkit was delivered as an executable file and the output included a .bcout file for each stimulus. This file can be loaded as a transcription alongside the spectrogram for a

stimulus to show the locations and classifications of the detected landmarks. The labels generated by the LND toolkit included V (vowel), Sc (sonorant), ST (stop), Fr (fricative) and SIL (silence).

Landmark boost package

The landmark boost package (LBP) was developed in MATLAB (The Mathworks, Inc.) in conjunction with Falk-Martin Hoffmann (University of Southampton). The LBP was designed to take the resulting landmark labels from the LND toolkit and apply a boost to the noise-corrupted speech signal in the corresponding analysis frames. Although landmark labels were generated for sonorants, vowels and periods of silence, the present study focused on the perception and transmission of the obstruent consonants and so only stop and fricative labels were used and enhanced by the LBP; the other three labels were simply ignored. The development of the LBP allowed for selective enhancement of obstruent consonants, relative to sonorant sections within the signal.

The main stages of signal processing applied by the LBP for this experiment are outlined below and shown in figure 5.6.

1. Each BKB recording was mixed with a random section of the noise stimulus (either babble or SS). The noise-mixed files were produced in such a way that the noise signal was ramped over a duration of 0.5 seconds before the speech signal started and after the speech signal had ended. To ensure that both the speech and noise signal were of the same duration, the speech signal was extended by zero padding at the beginning and end of the stimulus.

2. Landmark labels were generated for the noise-mixed stimuli using the LND toolkit (generation of the .bcout file).

3. Each .bcout file was read by LBP and a boost applied (in dB) to the signal in the corresponding time frames, linearly between 2500-8000 Hz (similar to the parameters used in experiment I). The time development characteristics of plosives and fricatives (as discussed in section 3.1) were considered when defining the duration of the boost, with the boost applied at plosive landmark labels being relatively short in duration (6 ms) compared with that applied at fricative labels (12 ms).

4. The NMT was then implemented to create a 12-*of*-22 vocoder simulation of the boosted, noise-mixed recording.

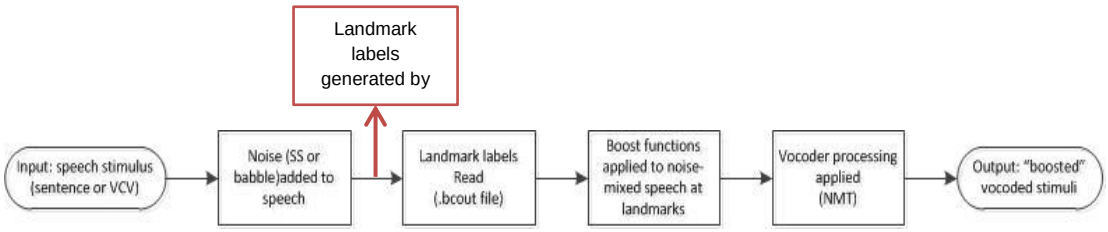


Figure 5. 6 Processing stages of the landmark boost package.

5.4.3 Research questions

1. Can enhancing landmarks using the proposed method help to improve the transmission of these events with CI processing and subsequently improve speech recognition in noise for normal hearing listeners listening through a cochlear implant simulation?

2. If an improvement is observed, how is it affected by the noise condition e.g. noise type (babble vs. speech-shaped) and SNR?

3. What level of enhancement (in dB) gives the greatest improvement (if any) and again, is this affected by the noise condition?

5.4.4 Method

Fourteen native English speaking NH listeners, aged 18-29 years participated in this experiment. Participants were not paid for their participation.

Results from a small pilot study of six NH listeners, revealed that ceiling effects were reached using an SNR of +10 dB for both babble and SS noise. BKB sentences were therefore corrupted with babble and SS noise at +5 and 0 dB SNRs. Each participant completed the experiment for both types of noise, at both SNR levels and for three processing conditions: a baseline condition (no enhancement, standard ACE processing), +5 dB boost and +10 dB boost applied at the times of detected obstruent landmarks.

To avoid bias towards a particular list and to cancel out any effects from a particular list within a condition, no two participants were presented with the same list for the same

condition (lists 1-20 were used for this experiment). Test order was randomised for each participant and test ear was also alternated between subjects to avoid any possible ear effects. Participants completed either all babble noise conditions first followed by all SS noise conditions, or vice versa. Prior to testing, participants were given time to listen to three practice lists to help familiarise themselves with the CI simulated stimuli. The first practice list simulated listening in quiet, the second listening at +10 dB SNR in babble noise and the third in +10 dB SNR in SS noise. Speech material for the practice lists consisted of the first three lists of the IHR sentences. Participants completed testing in one two hour session, allowing time to complete the practice lists and have regular breaks.

5.4.5 Results

Mean total percent correct scores for all conditions are shown for SS noise and babble in figure 5.7. A three-way repeated measures ANOVA was performed with three factors (SNR, noise type and level of enhancement) with two levels for SNR (0 and 5 dB) and noise type (babble and SS noise) and three levels of enhancement (baseline, 5 and 10 dB). Results indicated a significant effect of SNR ($F(1, 153)$=247.3, p=.00), boost ($F(2, 152)$= 40.91, p= .00) and noise type ($F(1, 153)$=71.77, p=.00). There was a non-significant interaction between SNR and level of enhancement ($F(2, 152)$=0.21, p= 0.81) and between SNR and noise type ($F(1, 153)$=0.007, p=0.93). However, a significant interaction was observed between noise type and level of enhancement ($F(2, 152)$=7.93, p=.00). There was a non-significant three-way interaction between SNR, level of enhancement and noise type ($F(2, 152)$=2.42, p=0.09). *Post-hoc* Sidak's test showed that scores in the baseline condition were significantly higher than for both enhancement conditions and that scores in the 5 dB boost condition were significantly higher than for the 10 dB boost condition (all at p=.00).

As the interaction between level of enhancement and noise type was found to be significant a two-way repeated measures ANOVA was conducted for each noise type. There were two factors (level of enhancement and SNR), with three levels for enhancement and two levels for SNR (as before). Results for SNR and level of enhancement were significant for both babble and SS noise (all at p=.00) and neither noise type showed a significant interaction between SNR and level of enhancement). Sidak's *post-hoc* test revealed that scores in the baseline condition did not differ significantly from scores in the 5 dB boost condition in babble noise, and that both the baseline condition and the 5 dB boost condition yielded significantly higher scores than for the 10 dB boost condition. For SS noise, scores in the baseline condition were

significantly higher than for either of the boosted conditions and scores were significantly worse in the 10 dB boost condition compared with the 5 dB boost condition.

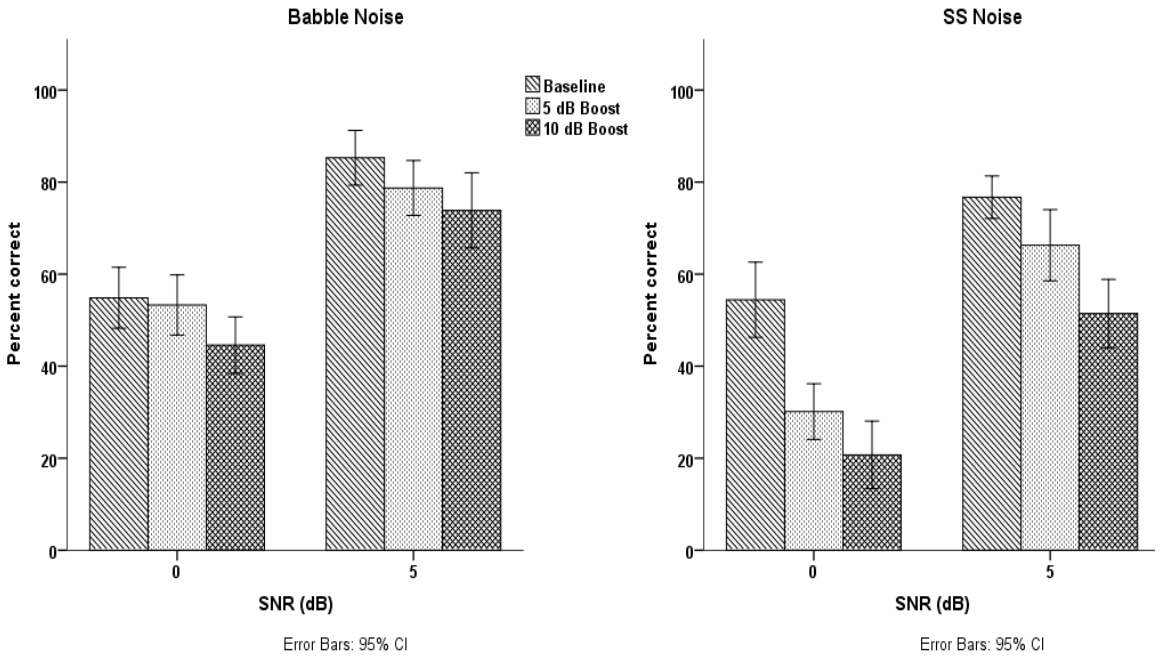


Figure 5. 7 Mean percent correct scores in babble and in SS noise for each SNR and all levels of enhancement.

5.4.6 Discussion

Results from this experiment indicate that the proposed method of obstruent landmark enhancement was not beneficial for improving speech recognition in either type of noise. The landmark enhancement strategy was in fact found to be detrimental to performance, with the highest level of boost (10 dB) resulting in the poorest performance. This effect was found to be more pronounced for SS noise than for babble noise, with an average decrease in percent correct scores of 33 percentage points in SS noise at 0 dB SNR (10 dB boost) and 26 percentage points at +5 dB SNR. Whereas, for the same processing conditions in babble noise the decrease in performance was only 10 and 11 percentage points respectively.

The observed reduction in performance with the enhanced stimuli is possibly a result of the LND toolkit and its poor accuracy in noise. On visual inspection of several of the .bcout files outputted by the detector, compared with their respective spectrograms, it became apparent that the fricative and plosive detectors were not particularly accurate in the SNRs used in this experiment. Figure 5.8 gives an example of a SS noise-corrupted sentence at 0 dB SNR

(panel a), the same sentence at 20 dB SNR (panel b) and the corresponding landmark label transcriptions as given by the LND toolkit. The red circles highlight the incorrect identification of fricative landmarks and the green arrows in the bottom panel (c) suggest times at which an obstruent landmark should have been identified (based on the clean signal). The figure demonstrates that even at fairly high SNR levels, fricative landmarks are incorrectly identified at several points throughout the utterance and both plosive and fricative landmarks have been missed on other occasions. Of note is that the vowel landmark detector appears to be fairly robust, even at a fairly low SNR. These results are reminiscent of those found with the channel-specific landmark detectors (as proposed in chapter 4).

The introduction of incorrect fricative labels would have led to these sections of the noise signal being boosted, increasing the chances of the corresponding channels being selected in the vocoder processing stage. This may have introduced masking effects as well as signal artefacts. The poorer scores observed with enhanced speech in SS noise compared with babble noise may be the result of the characteristics of the noise signals used. The aperiodic nature of the SS noise spectrum possibly closely resembles that of the spectrum of fricative noise. SS noise is a continuous signal whereas babble noise has natural peaks and troughs, meaning that its energy is not constant across the whole utterance. This may have resulted in more insertions of fricative landmarks for SS noise, occurring at the onset of any noise dominant segments in the high frequencies, i.e. between sonorant segments. Another explanation for the differences in scores observed in the different types of noise is that the landmark classifiers were trained using the same eight-talker babble noise as was used to corrupt the sentences for the experiment. This may have resulted in a higher degree of accuracy in landmark identification for babble noise than for SS noise and therefore, fewer insertions of incorrect landmark labels.

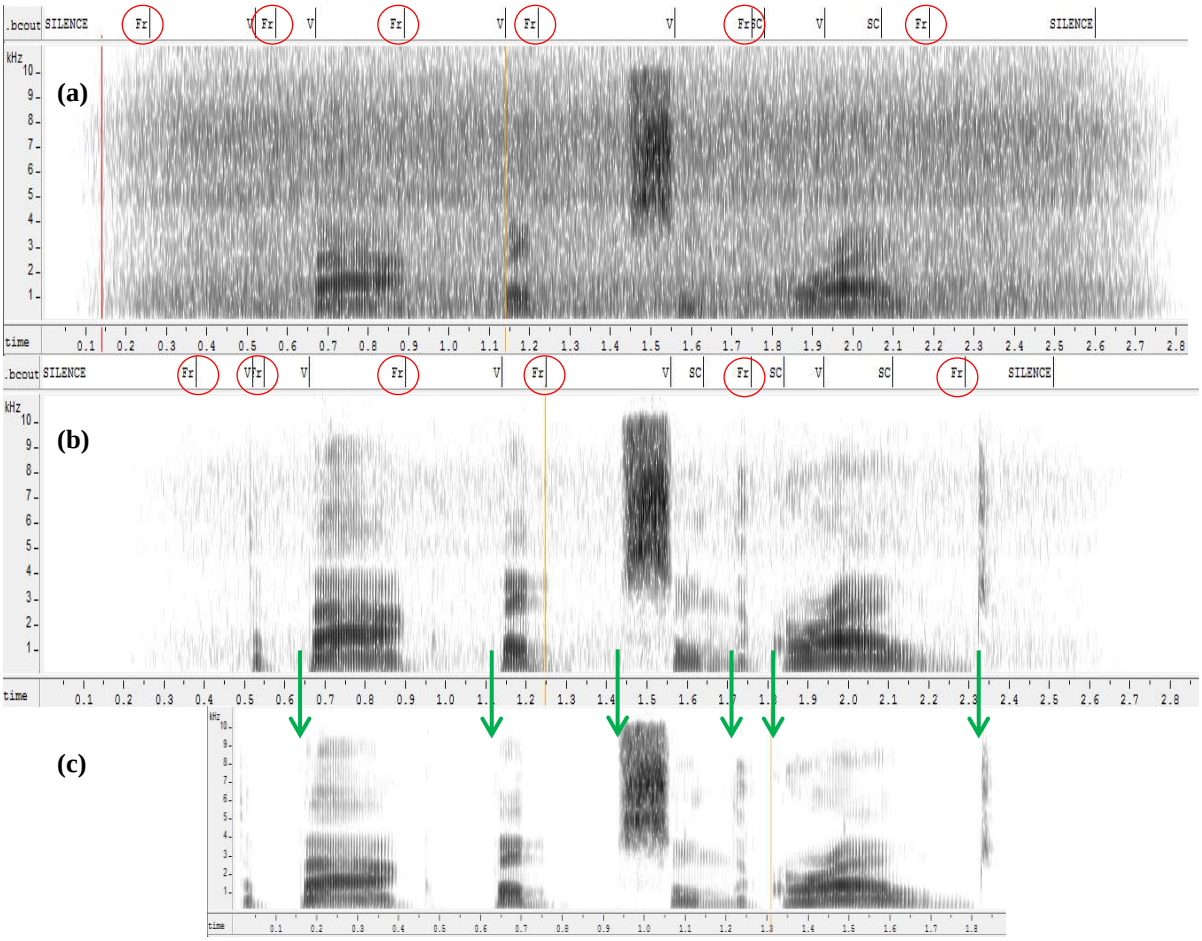


Figure 5. 8 Broadband spectrograms for the sentence "The bag bumps on the ground" in SS noise at 0 dB SNR (panel a) and 20 dB SNR (panel b). The corresponding landmark transcriptions as generated by the LND toolkit can be found above each spectrogram. The bottom panel (c) shows the same sentence but in quiet- note the noise-corrupted sentences had a period of silence added at the beginning and end of the sentence to allow for ramping of the noise signal, therefore the spectrogram for the quiet sentence has been time-aligned for ease of comparison.

Although during the development of the LND toolkit the classifiers were modified and trained to be more noise-robust, the level of accuracy was not available prior to the experiment, although 'it was observed that the new models were more noise-robust than the original landmark detector' following the testing of the newly trained landmark models (Juneja, 2012, personal communication). The classifiers had also been trained using sentences from the TIMIT database and not the BKB sentences; however, it is likely that this would mostly result in the LND toolkit missing landmark events.

Based on these findings, it was therefore concluded that further work using the LND toolkit would require the improvement and documentation of the accuracy of the landmark detectors, for both types of noise, as calculated for the BKB sentences. For the next experiment it was

suggested that the landmark labels, as generated by the LND toolkit for the present study, should be visually inspected and corrected by hand. These new labels should therefore represent "perfect" landmark detection and would allow the author to determine the potential of the proposed LBP using a similar test paradigm as used in the present study. Following further developments to improve the accuracy of the classifiers in noise, these results could then be compared with scores obtained with the LBP utilising the LND toolkit. It would also be possible to manipulate the hand-corrected labels to introduce certain types of errors, including missed landmarks, substitutions and insertions, and quantify their effect on speech perception scores. Although this next stage of development would not allow for real-time application of the LBP in a CI device, it will provide a clearer estimate of whether optimised enhancement may be beneficial for speech recognition in noise.

5.5 Experiment III- Generation of hand labels and further development of boost functions

5.5.1 Introduction

The initial aims of this experiment were to:

1. Determine the accuracy of the improved LND toolkit for different levels of noise and for different noise types. To be measured in terms of misses, substitutions and insertions.

2. Check and, where necessary, amend the labels generated by the automatic landmark detectors.

3. Assess whether "perfect" landmark detection with applied enhancement could help to improve speech perception in noise (when listening to CI simulated speech).

4. Explore the effects of deleted, substituted and inserted landmark labels on speech perception scores with NH listeners listening to CI simulated speech.

In order to generate the hand-annotated landmark labels for the BKB (and IHR) sentence lists, the .bcout files as generated by the LND toolkit (in quiet) were loaded alongside their corresponding spectrogram and obstruent landmark labels (for plosives and fricatives) were altered, removed and created as necessary. However, during this process of spectrographic analysis it was considered that there may be merit in distinguishing between plosives which were very short in duration (almost instantaneous, for example, /d/ in the word "dog") and

plosives of longer duration, i.e. with a period of aspiration after the initial burst lasting longer than 20 ms (for example, /k/ in the word “king”). The landmarks occurring at the beginning of these different plosives will have different time developments in terms of their rise and fall times, as well as their overall duration and therefore may benefit from individual processing to help listeners distinguish between them. As the sudden spectral and amplitude changes that occur after the cessation of aspiration and fricative noise are also important cues for speech perception, an offset label was also generated. The new set of proposed landmark labels are described in the following section along with the further developments which were made to the boost parameters.

Unfortunately, at this stage, further development of the LND toolkit with Amit Juneja was not able to continue within the time-frame of this thesis. The author was also not permitted to continue the development of the toolkit with another researcher and therefore, all subsequent experiments using the LBP are based on manually generated landmark labels only.

5.5.2 Hand labelling of landmarks

The hand generated labels were as follows:

- **P** – a plosive of short duration (sudden onset and offset)
- **Pn** - a plosive of longer duration, defined as having a considerable period of aspiration following the initial burst (sudden onset and offset)
- **Fr**- fricative, long segment of aperiodic noise (onset and offset longer not as abrupt)
- **Off** - offset landmark following the cessation of longer duration plosives and fricatives.

Broadband spectrograms for each BKB and IHR sentence were visually inspected and annotated in Wavesurfer (Beskow and Sjölander, 2004). Label transcriptions for the sentences were saved as .bcout files so that they could be read by the LBP. An example of a labelled spectrogram can be seen in figure 5.9. Unlike for experiment II, the labels were generated from the clean spectrogram, rather than from the noise-corrupted stimulus. As a result, the landmark labels had to be shifted by 500 ms by the LBP prior to the boost being applied as the length of the stimuli were altered by the introduction of a period of silence at the beginning and end of the sentences to allow for ramping of the noise signal.

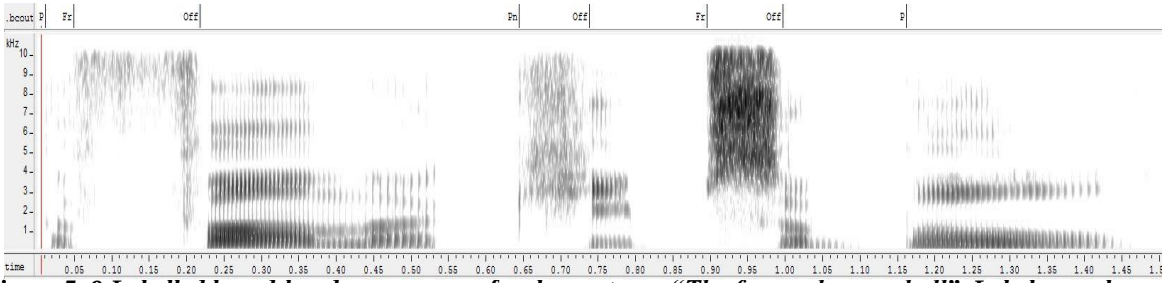


Figure 5. 9 Labelled broad-band spectrogram for the sentence "The farmer keeps a bull". Labels are shown, time-aligned in the pane above the spectrogram.

5.5.3 Further development of boost functions

Time-Variant Gain Stage

Figure 5.10 shows the time development of the gain applied at each of the four landmark labels. The gain development over time is defined through the number of frames it takes to rise to full gain (N_R), fall back to zero gain (N_D) and the overall width (N_O) in frames (frame length of 12 ms with 6 ms frameshift). How these parameters differ for the four landmark labels is demonstrated table 5.I. Gain development for short duration plosives (**P**) was rapid, with a fast rise and decay time and only of short duration. Gain for longer duration plosives (**Pn**) also had a rapid rise time but a slightly longer decay time and a longer overall width duration. Rise time for fricative (**Fr**) landmarks was slightly longer than for **P** and **Pn** landmarks, with a longer overall boost duration (width) and longer decay time. Rise time for the offset landmark (**Off**) was the same as for **Fr** landmarks but overall width and decay time were longer. As can be seen in figure 5.10, the gain applied at **Off** landmarks is actually negative. These have been chosen to suit the characteristics of the different landmarks (as outlined in section 5.5.2).

Landmark	Rise time (N_R)	Decay time (N_D)	Overall width (N_O)
P	1	1	4
Pn	1	2	5
Fr	2	3	8
Off	2	4	10

Table 5. 1 Parameters defining the gain development over time for the four types of landmarks (in frames).

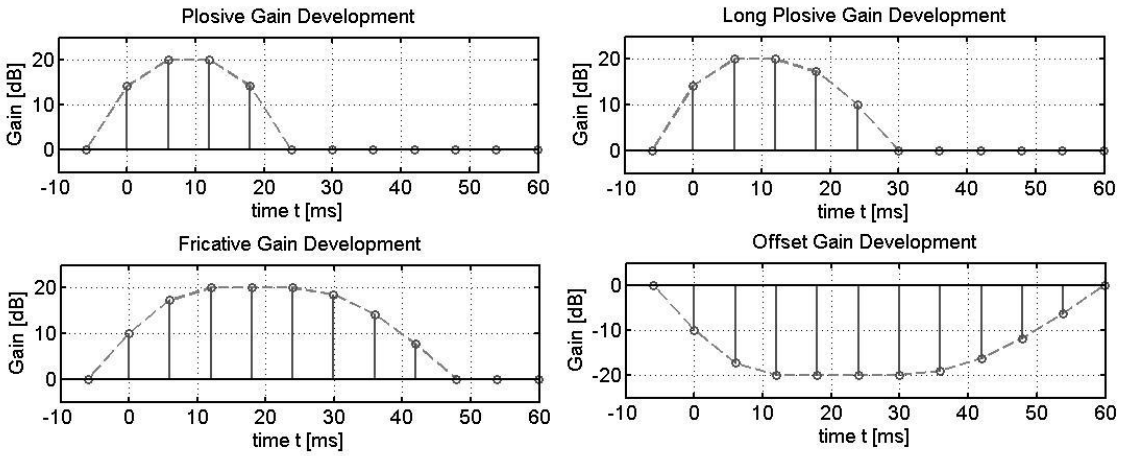


Figure 5. 10 Time development for the four landmark gain functions (in dB), with a maximum of 20dB or -20dB respectively. The stem mark at t = 0 represents the gain for the frame corresponding to the occurrence of a landmark.

Frequency Content-based Gain Stage

Further to changes to the time-development of the gain applied at different landmarks, it was also considered that it may be more beneficial to focus the enhancement on specific regions considered important for speech sound identification. This is similar to the approach of Hazan and Simpson (1998) who filtered frication to include only a certain region and examined the burst spectrum of plosives to find the area of maximum energy. This does, in part, contradict the approach taken in experiments I and II, which attempted to enhance only the sudden spectral and amplitude changes relating to ALs, and not those relating to formant transitions or other important speech cues. However, if information relating to these cues is not transmitted, or is impaired, then even though better transmission of obstruent landmark information might help listeners identify word/segment boundaries, it will not necessarily aid with speech sound identification/discrimination. Spectral peak and relative amplitude are important cues to fricative place of articulation and the frequency and intensity of the release burst are important cues to voicing and place of articulation cues for plosives (Reetz and Jongman, 2008). Therefore, in order to optimise the results of the processing, the gain was also subjected to content-based weighting.

Centre of gravity (COG), or Centroid, of the noisy speech spectrum was computed, and was defined for the frame with index i by

$$\mathrm{COG}_i = \frac{\sum_{k=L_{low}}^{L_{high}} kX_{k,i}}{\sum_{k=L_{low}}^{L_{high}} X_{k,i}},$$

for the discreet signal. Where X is the noisy speech spectrum and L_{high} and L_{low} are the bins corresponding to the lower and upper edge frequencies, individually chosen for the four landmarks, and are given in Table 5.2. The COG was therefore calculated within a limited frequency range and the f_{low} and f_{high} for each landmark type was chosen to capture the important energy in the spectrum, across frequencies, which occur at the onset of each event. This was based on both spectrographic analysis and evidence from section 3.1.

Landmark	Lower edge frequency f_{low} (Hz)	Upper edge frequency f_{high} (Hz)
P	2500	8000
Pn	2500	8000
Fr	1000	8000
Off	200	8000

Table 5. 2 Frequency ranges analysed to control frequency dependent gain processing for the different types of landmarks.

The highest gain is applied to the COG bin and decays in a sinusoidal slope towards both the lower and higher edge frequency. While gain falls back to 0 dB towards the lower edge, it only decreases by 20 percent towards the upper edge. The remaining part above the upper edge decays in a log-shaped slope towards a 0dB gain at the Nyquist frequency bin. The resulting gain development across the entire frequency range is shown in figure 11. To keep the computation complexity low, the COG is only computed for the first frame of a detected landmark. The following frames are subject to the same relative gain development. Further details about the COG calculation can be found in the MATLAB script file for the final version of the LBP in Appendix 8.

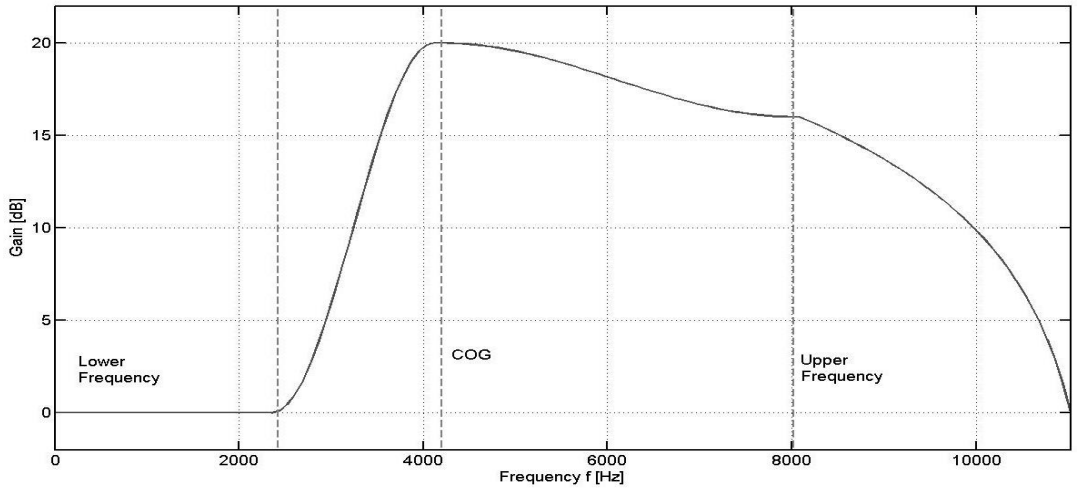


Figure 5. 11 Gain development across all frequencies for an example plosive landmark, with the COG at f= 4.2 kHz and a maximum gain of 20dB.

5.5.4 Method

Fifteen NH listeners, aged 18-29, participated in this experiment, four of whom had some previous experience of listening to vocoded stimuli. Participants received no financial incentive for taking part in the study. As per experiment II, percent correct scores for BKB sentences were measured in terms of number of key words correctly identified. The sentences were corrupted with babble and SS noise at +3 and 0 dB SNRs. For both types of noise and both SNRs, each of the subjects was tested for three levels of obstruent landmark enhancement; no enhancement (baseline condition), +5 and +10 dB. As before, the enhanced sentences were processed to simulate a 12-*of*-22 ACE strategy. Each participant completed the experiment for both types of noise, at both SNR levels and for three processing conditions: a baseline condition (no enhancement, standard ACE processing), +5 dB boost and +10 dB boost applied at the times of labelled obstruent landmarks. These conditions mirror those used in the previous experiment, however, the +5 dB SNR condition was changed to +3 dB SNR to avoid ceiling effects (as determined during pilot testing). Test methodology and training were identical to that used in experiment II.

5.5.5 Results

Mean total percent correct scores for all conditions are shown for SS noise and babble in figure 5.12. A three-way repeated measures ANOVA was performed with three factors (SNR, noise type and level of enhancement) with two levels for SNR (0 and +3 dB) and noise type

(babble and SS noise) and three levels of enhancement (baseline, 5 and 10 dB). A significant effect of SNR was observed (*F(1, 164)*=113.03, *p*=.00), with scores worse in the 0 dB SNR condition for both noise types and for all levels of enhancement. No significant differences were observed in scores between noise types or for level of enhancement and there were no significant interactions.

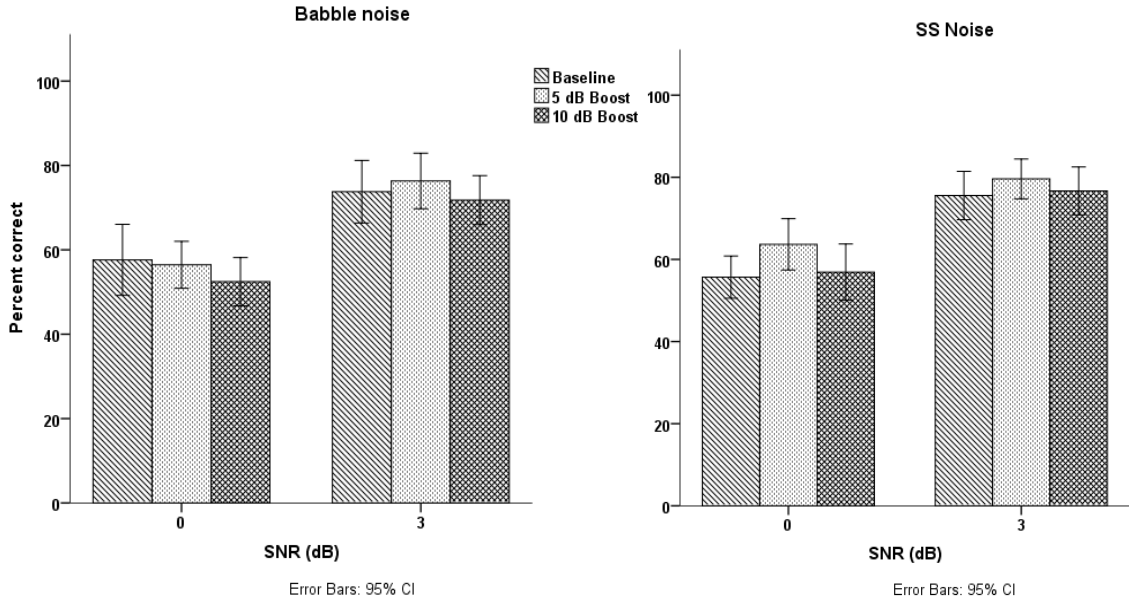


Figure 5. 12 Mean percent correct scores in babble and in SS noise for each SNR and all levels of enhancement.

As previously stated, four of the participants had some previous experience of listening to vocoded speech. To determine whether level of experience had an effect on the scores observed for this experiment, the RAU scores were pooled across noise types (as this was not shown to have a significant effect on scores) and a three-way repeated measures ANOVA was performed with three factors (level of experience, SNR and level of enhancement) with two levels for SNR (0 and +3 dB) and level of experience (some experience and no experience) and three levels of enhancement (baseline, 5 and 10 dB). As before, a significant effect of SNR was observed (*F(1, 164)*= 157.3, *p*= .00) and a significant difference was also observed between the scores for the experienced and non-experienced group, with the experienced participants scoring higher for all conditions than the non-experienced group (see figure 5.13). Results from the ANOVA, however, did not reveal any significant interactions between the factors. This suggests that level of experience did not affect the ability of a participant to make use of the enhanced obstruent landmark information.

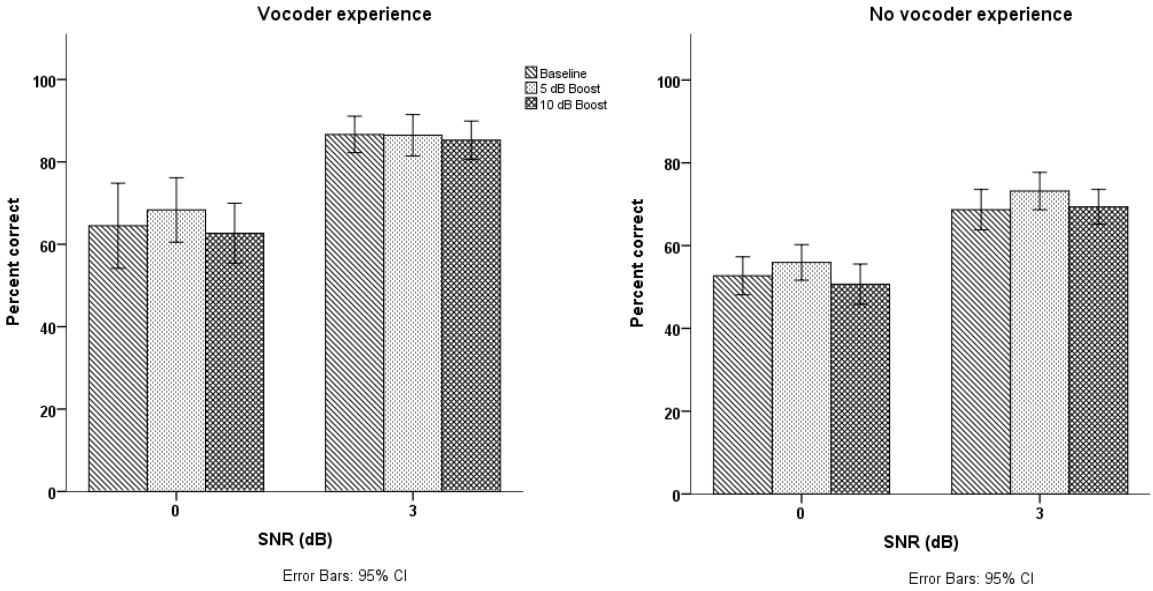


Figure 5. 13 Pooled mean percent correct scores for experienced and non-experienced participants for each SNR and all levels of enhancement.

5.5.6 Summary

Overall, results from the present experiment do not indicate any benefit of obstruent landmark enhancement when listening to vocoded speech in the presence of a competing noise source. The initial aims of the experiment were to compare speech recognition results obtained with the improved ALD toolkit with those which represented "perfect" landmark detection (by manually labelling each sentence) and explore the effects of inserting, deleting and omitting landmark labels. However, circumstances were such that further development of the LND toolkit could not continue and so comparisons between the two strategies could not be made.

It was considered whether the gain applied at **P** labels (short-duration plosives, <20 ms) was too long, with enhancement being applied beyond the end of the burst noise. It is also possible that the lower edge frequency for the COG calculation may not be low enough, resulting in conflicting burst cues. The lower edge frequency for the **Off** label may also have caused the reduction in the transmission of information relating to important formant transitions into adjacent vowel segments. A further period of spectrographic analysis was

therefore proposed to reconsider the labelling of some segments as well as the time and frequency developments of the four gain functions.

5.6 Experiment IV- Further developments to hand labelling and LBP

5.6.1 Introduction

Considering the findings from experiment III, further changes were made to the time development gain functions for the four landmark labels. In order to more clearly distinguish between very short duration plosives and those with at least some duration of aspiration following the release burst, the gain applied for labels designated **P** was made to be much shorter in duration and almost instantaneous. In order to achieve this, the frame length was changed to 3 ms with a frameshift of 1.5 ms). Based on further spectrographic analysis across all sentence material, it was determined that the duration of the boost applied at the label **Pn** should not be greater than 20 ms so as to incorporate landmarks which had originally been classified as **P** but were now classified as **Pn**. This change in gain development would ensure that enhancement was not applied beyond the aspiration of the identified segment and would allow for the correct placement of an **Off** label. The lower edge frequency used for both plosive labels was also reduced to 1000 Hz. The duration of the negative gain applied at **Off** labels was considerably reduced to minimise effects of interactions between closely occurring onset (**Fr** and **Pn**) and **Off** landmarks. The lower edge frequency for **Off** landmarks was also increased to 3000 Hz so as not to impair the transmission of important formant transitions (in the region of F1 and F2). The new gain development parameters for the four landmark labels can be found in tables 1.3 and 1.4 and in figure 5.14.

Landmark	Rise time (N_R)	Decay time (N_D)	Overall width (N_O)
P	0	0	1
Pn	3	6	14
Fr	6	9	24
Off	4	5	14

Table 5. 3 Parameters defining the new gain development over time for the four types of landmarks (in frames).

Landmark	Lower edge frequency f_{low} (Hz)	Upper edge frequency f_{high} (Hz)
P	1000	8000
Pn	1000	8000
Fr	1000	8000
Off	3000	8000

Table 5. 4 New frequency ranges analysed to control frequency dependent gain processing for the different types of landmarks.

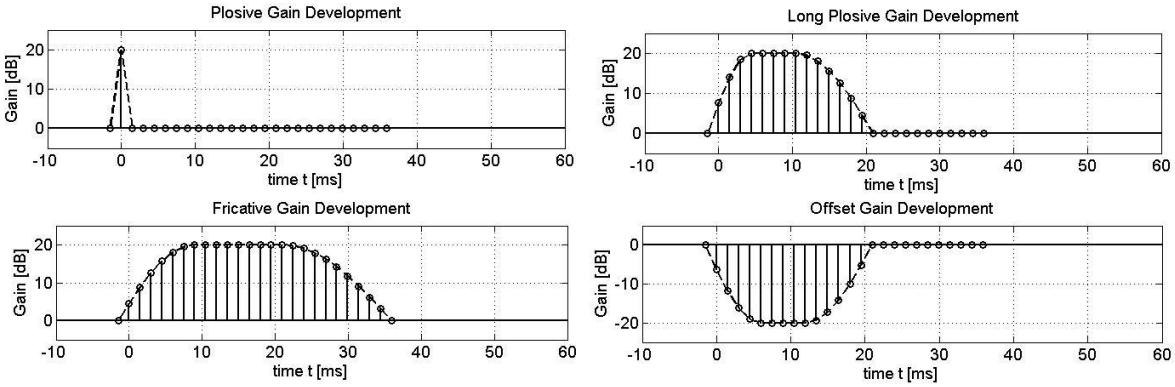


Figure 5. 14 New time development gain functions, in milliseconds, for the four adjusted landmark labels.

As a consequence of altering the time gain development of the landmark labels, all landmark transcriptions were checked and corrected accordingly. Figure 5.15 demonstrates how some landmarks originally labelled as **P** were changed to **Pn** and an **Off** label inserted at the end of aspiration. In addition to reclassifying some **P** labels as **Pn,** re-examining the spectrograms and their corresponding landmark transcriptions afforded an opportunity to check the accuracy of the hand-annotation. Figure.16 demonstrates how during this process some landmark labels were adjusted (in time) and some were introduced, such as for the aspirated sound /h/ which was previously not labelled but for the following experiment was classified as a fricative (unvoiced turbulent noise).

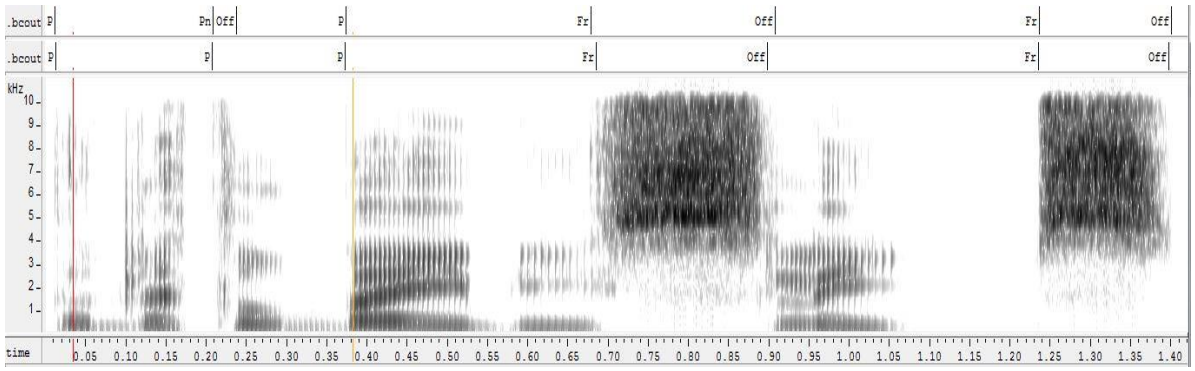


Figure 5. 15 Broadband spectrogram for the sentence "The little baby sleeps". The top transcription panel shows the corrected landmark labels for the present experiment whereas the bottom transcription panel shows the original hand-generated labels from experiment III.

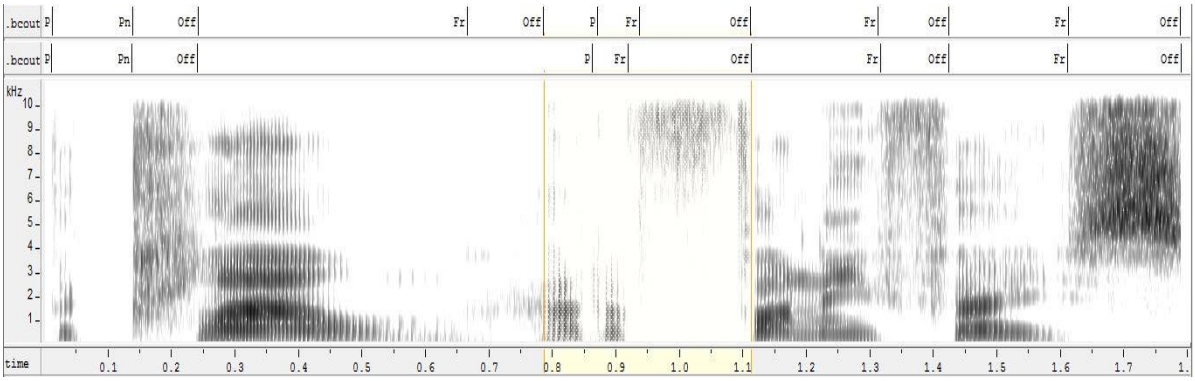


Figure 5. 16 Broadband spectrogram for the sentence "The clown had a funny face". The top transcription panel shows the corrected landmark labels for the present experiment whereas the bottom transcription panel shows the original hand-generated labels from experiment III.

5.6.2 Method

Nine NH listeners, aged 20-29, participated in this study. None of the participants had previous experience with vocoder simulated speech. All subjects were paid for their participation. Subjects completed two listening tasks, the first tested their open-set speech recognition scores and the second was a consonant identification task using VCV stimuli. The VCV task was included so that the effect of the individual boost functions could be analysed and be used to guide further improvements to the LBP. Landmarks label files (.bcout) were therefore generated for each of the VCV stimulus recordings used in this experiment.

Landmark enhancement has so far not shown any benefit for speech recognition in the SNRs tested. The present experiment therefore wanted to explore whether an improvement might be seen at more positive SNRs, where there is less masking by the noise and potentially less negative effects of also boosting the noise. CI users typically struggle in more positive SNRs

(Spahr et al., 2007 and Oxenham and Kreft) and therefore it would be advantageous to show benefit of landmark enhancement at these noise levels. However, pilot experiments using the new enhancement parameters in more positive SNRs found that ceiling performance was being reached with the 12-*of*-22 ACE strategy that was used in previous experiments. Further pilot testing found that a 3-*of*-22 ACE strategy was needed in order to restrict performance in the baseline condition to be more in line with actual CI user performance at the SNRs used. This agrees with the findings of Loizou et al. (1999) who showed that NH listeners were able to achieve high levels of speech recognition with as few as five channels. Although spectral smearing can be used to help limit performance in vocoder simulation, restricting information transmission by simply reducing the number of channels may provide a better prediction of actual CI user performance (Verschuur, 2009; Whitmal et al., 2007; Oxenham and Kreft, 2014).

For both tasks, each participant was tested in +10 and +5 dB SNRs, with and without enhancement. Participants were tested in SS noise only to reduce the number of overall test conditions in the VCV task (this was to avoid fatigue effects). As level of boost has so far shown no significant effect on speech recognition scores, a single level of boost, 9dB, was investigated in the present study. This was selected based on evidence from the study by Ortega et al. (2000), who applied a 9 dB boost to bursts, aspirations and fricatives.

Sentence task

Participants attended a two hour session, allowing time for training and for regular breaks. As in experiment III, participants were scored on the number of key words correctly identified for each sentence. For each of the four conditions, participants listened to and repeated three full lists from the BKB sentence lists, meaning that participants listened to a total of 12 different lists. Participants listened to three lists per condition in the present study because in experiment III participant's scores seemed to vary greatly between conditions and it was considered that this might be a reflection of participants finding some lists harder than others. Participants were therefor tested with three lists per condition in an attempt to balance out scores if a participant found a particular list harder than others and to account for any learning effects within a condition. Although one list should be sufficient to account for random variation within and between lists, swapping between SNRs can take the brain time to adjust (Li and Loizou, 2009). Therefore, participants completed all three lists for a single condition in succession.

VCV task

It was decided to introduce this task to explore in more detail the effects of the LBP on specific consonant stimuli, and allow for information transmission analysis to be examined. This may help to give some indication whether the LBP was beneficial in improving the transmission of certain features and recongition of specific speech sound and/or detrimental to the transmission and recongition of others. For each of the four conditions, participants listened to two recordings of each of the 20 consonants (produced by a female speaker), repeated three times and for the three vowel envrionments: aCa, iCi and uCu. Testing was split into three blocks (for the different vowel environments) and subdivided into the four different test conditions. There were 120 stimulus presentations per subdivision, equating to 480 stimulus presentations per vowel environement. This equalled 1440 stimulus presentations in total. Each subdivision took around five minutes to complete and therefore testing for each block took about 20 mintues in total. Breaks were given between each block and/or when requested by the participant. Test order for blocks, subdivisions and consonant presentation within each test condition were randomised.

Training

Participants completed a more extensive training session than in previous experiments. This was to try and limit possible learning effects during the actual test. Lists four to 16 of the IHR sentences were used for training in the sentence recognition task. Participants first listened to practice lists in quiet to acclimatise them to the quality of the vocoder simulation. For the first practice list, participants listened to each sentence, first without and then with vocoder processing. Participants were not required to respond to the sentences. The second list consisted of only vocoded versions of each sentence and participants were asked to repeat each sentence back to the tester; the participant was able to listen to each sentence several times if necessary. If the participant could not correctly repeat the whole sentence then they were informed by the tester of the correct sentence and it was played once more. Following this list, a final list was played which replicated the test condition (no repetitions and no feedback). Participants received training in quiet and for both SNRs used during the actual test. All practice stimuli were presented without enhancement.

Participants also received training for the VCV task so as to familiarise themselves with the test stimuli and their corresponding symbols. Participants were able to listen to each of the 20

consonants with and without vocoder processing in quiet and for both levels of SNR in speech-shaped noise. Participants listened to each recording as many times as they liked.

5.6.3 Results

Sentence test

Mean percent correct scores are shown in figure 5.17. From visually inspecting the raw data there did not appear to be any trend of learning across the three lists for any condition (i.e. scores are not necessarily higher in the final list of a condition). Variation in scores for each of the four conditions were similar, with standard deviations of 7.1 and 6.5 for the baseline and enhanced conditions at +5 dB SNR and 7.9 and 7.8 for the same conditions at +10 dB SNR. This would suggest that any difference in scores between conditions would be the result of the processing applied and not because of random variations in the scores for each list.

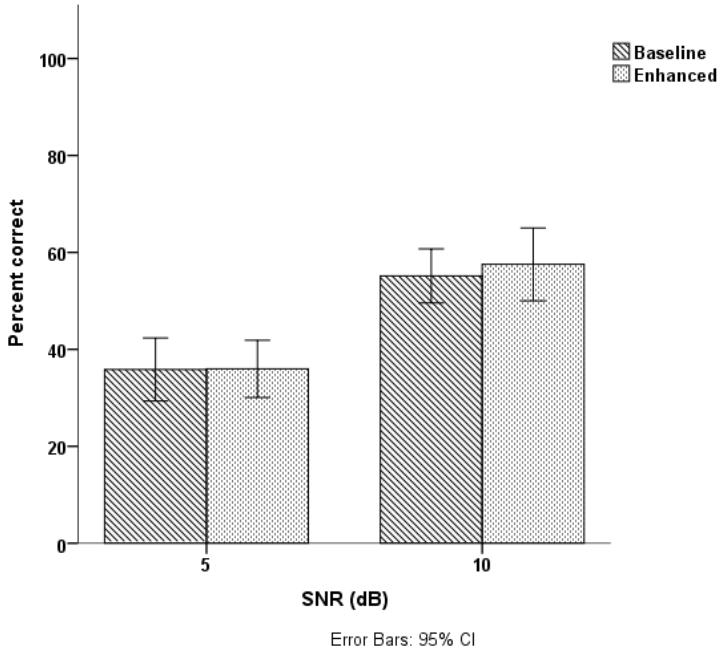


Figure 5. 17 Mean percent correct scores for enhanced (9 dB) and baseline conditions at 5 and 10 dB SNRs.

A two-way repeated measures ANOVA was performed with two factors (SNR and enhancement), each with two levels (5 and 10 dB for SNR and baseline and 9 dB boost for enhancement). Results indicated a significant effect of SNR ($F(1, 26)$= 53.02, p=.00) but no significant effect of enhancement and no significant interaction between SNR and

enhancement. Although a significant effect of processing was not observed for the average data, figure 5.18 demonstrates that four of the nine participants did achieve higher scores in the enhanced condition at 10 dB SNR (although an equal number scored poorer in the enhanced condition).

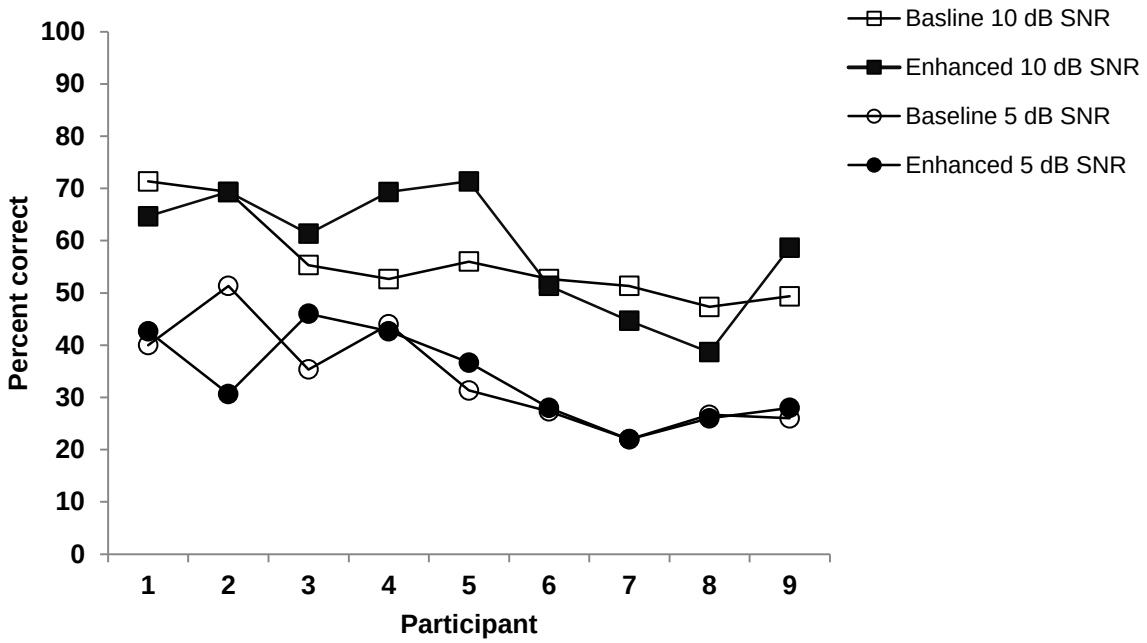


Figure 5. 18 Individual percent correct scores for the nine participants in the BKB sentence test.

VCV test

Mean percent correct scores for each vowel environment in both levels of SNR are shown in figure 5.19. A three-way repeated measures ANOVA was performed with three factors (SNR, vowel context and level of enhancement) with two levels for SNR (+5 and +10 dB) and level of enhancement (baseline and 9 dB boost) and three levels for vowel context (/a/, /i/ and/u/). Results revealed a significant effect of SNR ($F(1, 98)$=23.7, p=.00) and vowel context ($F(2, 97)$=52.7, p=.00), a non-significant effect of boost ($F(1, 98)$=1.6, p= 0.21) and no significant interactions. Sidak's *post-hoc* test showed the following significant trend in percent correct scores: /a/ > /u/ > /i/. Consonant confusion matrices were not computed analysed for the VCV task for reasons which are discussed in the following section.

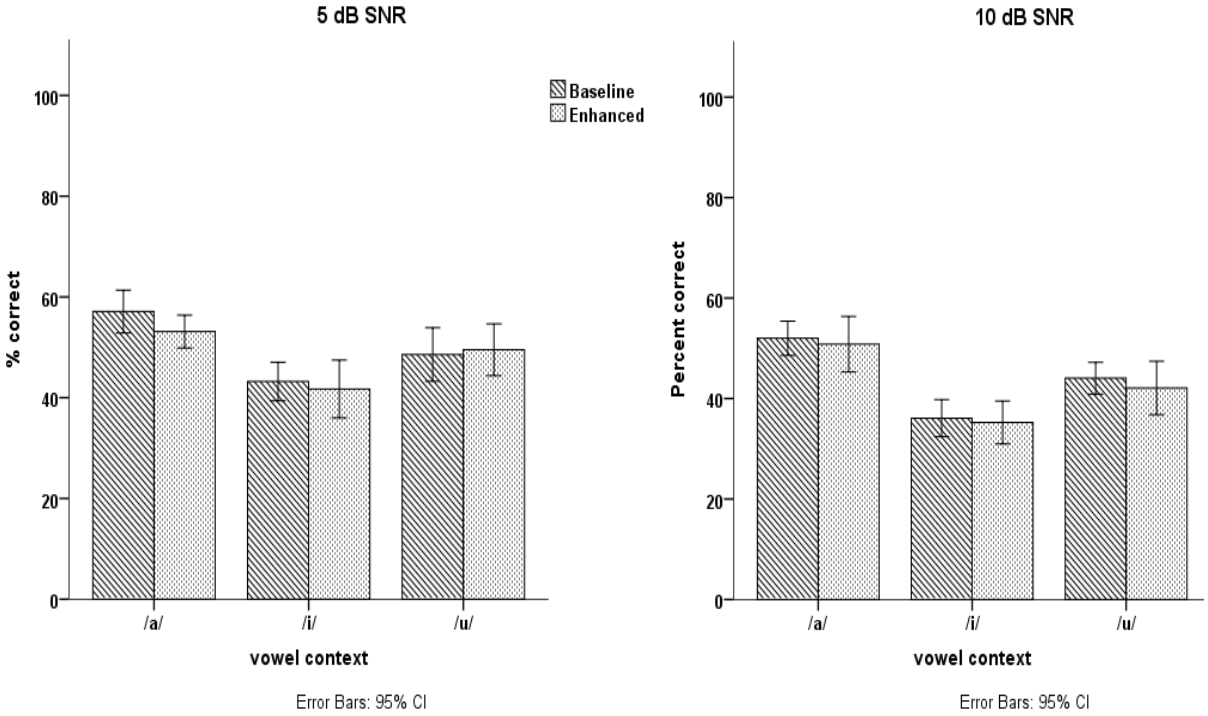


Figure 5. 19 Percent correct scores for each vowel in environment, with and without enhancement, in 5 dB SNR SS noise (left panel) and 10 dB SNR SS noise (right panel).

5.6.4 Discussion

Results from the present experiment do not show any benefit of obstruent landmark enhancement, in either the sentence or VCV task. However, individual scores from the sentence recognition task suggested that some participants may be able to make use of the enhanced obstruent landmark information (at least in the +10 dB SNR condition) as they scored between six to 16 percent higher than in the baseline condition.

Given that once again, enhancement of obstruent ALs with the LBP did not help to improve speech perception, the present author considered whether the information transmitted with the two strategies (standard ACE processing and ACE with LBP applied prior to channel selection) was in fact different enough to affect test outcome. Figure 5.20 shows broadband spectrograms for the original and processed versions of the sentence ‘The wife helped her husband’ and highlights regions where enhancement has been applied. Below this in figure 5.21 are the corresponding electrodograms which show only marginal differences in the electrode activation pattern between the original noise-mixed sentence and the enhanced sentence (shown by the blue ellipses). Further to this, activation patterns for the VCV stimuli

showed even less variation for the two strategies (demonstrated in figure 5.22). As, on average, scores between the two strategies did not differ significantly, this would suggest that differences in channel activation seen with the LBP did not provide the listeners with any new or more useful information about the target stimuli. Therefore, the seemingly large differences in scores observed for some participants between the baseline and enhanced condition for the sentence recognition task are likely only to be due to random variation and not a result of the processing applied by the LBP. This variation could reflect an unequal difficulty across the sentence lists, however, examination of the results for lists which had been heard by more than one listener (in the same condition or in a different condition) did not indicate that any particular lists were easier or harder than the others; scores for these lists varied between participants. This is not to say, however, that an individual participant did not find a particular list easier/harder than others. It is also possible that the random variation in scores could actually reflect a participant's attention during the test, with lower scores obtained when the participants were less engaged with the task.

For reasons why there was little evidence of the enhancement in the vocoded stimuli it was firstly considered that the level of boost applied was not enough to significantly affect the channel selection process in the n-*of*-m strategy used; channels which were highest in amplitude in the original noise-mixed sentence remained highest in amplitude following the application of the boost. It was also considered that the normalisation stage of the LBP applied to the boosted, noise-mixed sentences and VCV tokens may have contributed to the general lack of enhancement seen in the vocoder output. The normalisation stage reduced the overall level of the unprocessed segments of the stimuli (i.e. where a landmark did not occur) thus reducing the overall effect of the boost. This is evident from figure 5.20 where the intensity of the sonorant segments is less in the spectrogram for the enhanced sentence than in the spectrogram for the original sentence; indicated by lighter shading during these segments. It was therefore decided that a final experiment should be conducted in which the normalisation stage of the LBP would be removed and the level of boost would first be piloted.

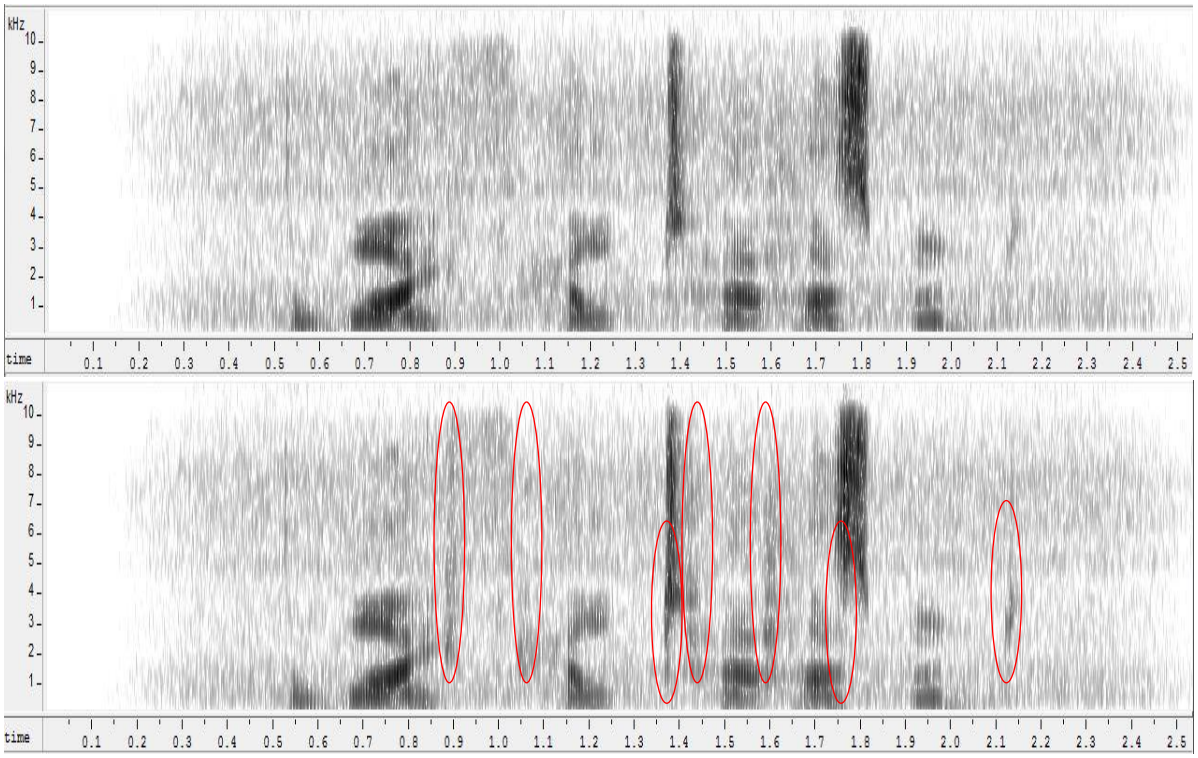


Figure 5. 20 Broadband spectrogram for the original sentence (top panel) "The wife helped her husband" mixed with speech-shaped noise at 10 dB SNR. In the bottom panel is the broadband spectrogram for the same sentence but with 9 dB boost applied at corresponding landmark labels. These areas of boost are highlighted by the red ellipses.

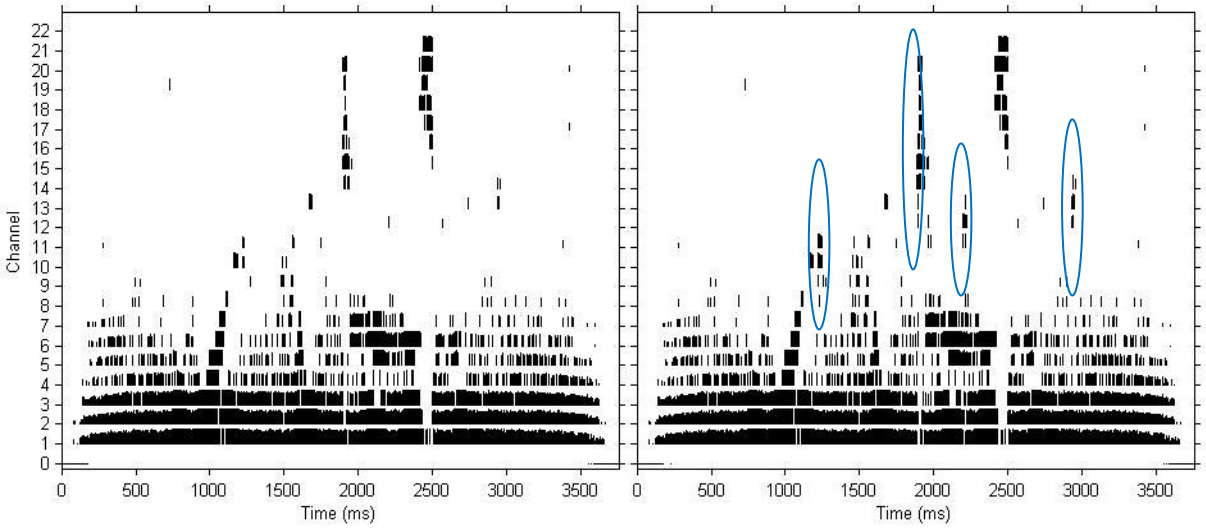


Figure 5. 21 Electrograms for the sentences "The wife helped her husband", corresponding to the original sentence at 10 dB SNR (left panel) and the noise-mixed sentence with enhancement applied (right panel). The electrodograms represent the resulting stimulation pattern of a 3-of-22 ACE strategy.

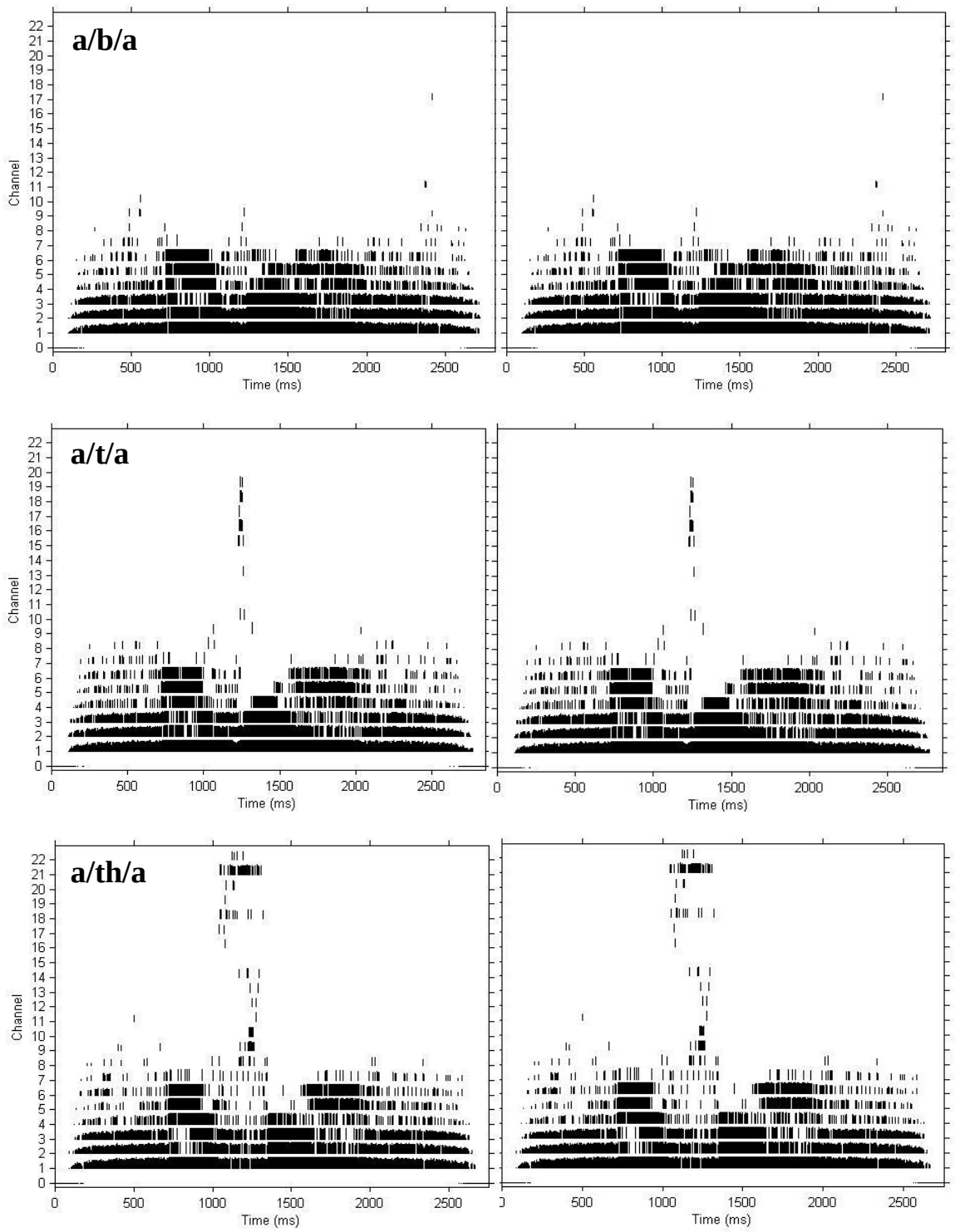


Figure 5. 22 Electrodograms for the unprocessed (left panel) and processed (right panel) versions of some VCV stimuli in the context a/C/a.

5.7 Experiment V- Removal of normalisation stage from the LBP

5.7.1 Introduction

This experiment was conducted in a final attempt to determine whether obstruent landmark enhancement with the proposed LBP could be used to help improve speech recognition in noise for NH listeners listening to CI simulated speech. The normalisation stage of the LBP was removed from the processing and all other parameters (i.e. time and frequency development of the boost functions) remained the same. A small pilot study was first conducted to choose an appropriate level of boost to apply at the landmark labels as this was also considered as a factor that may have influenced results in the previous experiment. Five experienced listeners were asked to listen to IHR sentences (in noise) for which the level of boost was varied between 10 and 20 dB. The participants were asked to select the level of boost which they felt produced the most intelligible version of sentence. Percent correct scores were not measured in the pilot study due to familiarity with the sentence material by the participants. Nonetheless, all the pilot subjects chose either the +12 or +15 dB boost and reported it sounded more intelligible than the original sentence (all done with vocoder processing).

5.7.2 Method

It was intended that the same participants from experiment IV would be recruited for the final experiment so that their scores from the two experiments could be compared; however, only five of the original participants were able to take part. A further three participants were recruited for the study, giving a total of eight NH participants. Subjects were paid for their participation.

During pilot testing it was also observed that the removal of the normalisation stage made the simulation in the baseline condition easier, therefore, to avoid ceiling effects participants were tested in 0 and +5 dB SNRs. Once again participants completed a sentence recognition task and a VCV task and were tested in four conditions: a baseline condition (no enhancement) and 15 dB boost at landmarks in 0 and +5 dB SNRs, all processed to simulated a 3-*of*-22 ACE strategy. Due to the limited number of sentence lists in the BKB corpus it was decided to use only two sentence lists per condition for each subject so as to avoid repeating lists that participants had heard in the previous experiment. Participants were given a refresher training session if they had taken part in experiment IV and the three new participants repeated the full training session as detailed in section 5.6.2.

Spectrograms and electrodograms for the enhanced stimuli were visually inspected prior to testing to check for evidence of the landmark enhancement being carried through to the vocoder simulation. Figure 5.23 gives the spectrograms for the sentence ‘The machine was quite noisy”’ with and without landmark enhancement, and figure 5.24 shows their corresponding electrodograms. The effect of the LBP is much more evident in these examples than those shown in section 5.6.4 and the electrodograms clearly demonstrate that it does have an effect on the resulting stimulation pattern derived from the vocoder processing.

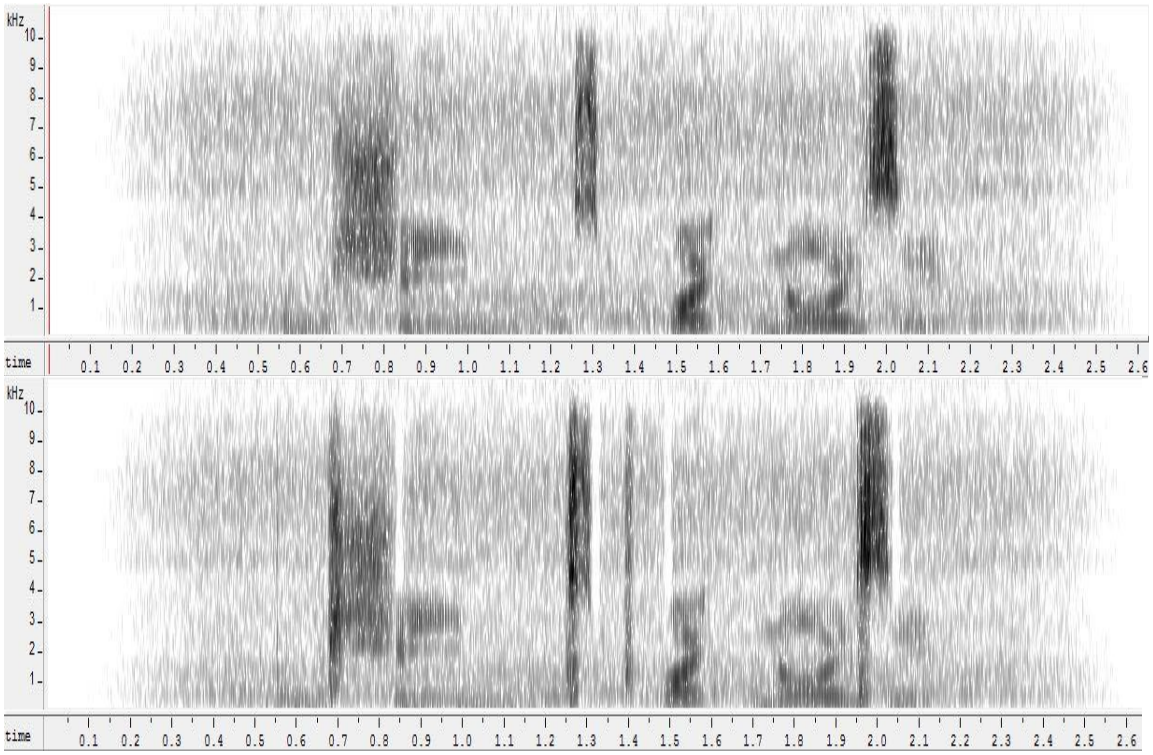


Figure 5. 23 Broadband spectrogram for the original sentence (top panel) “The machine was quite noisy” mixed with speech-shaped noise at 5 dB SNR. In the bottom panel is the broadband spectrogram for the same sentence but with 15 dB boost applied at corresponding landmark labels.

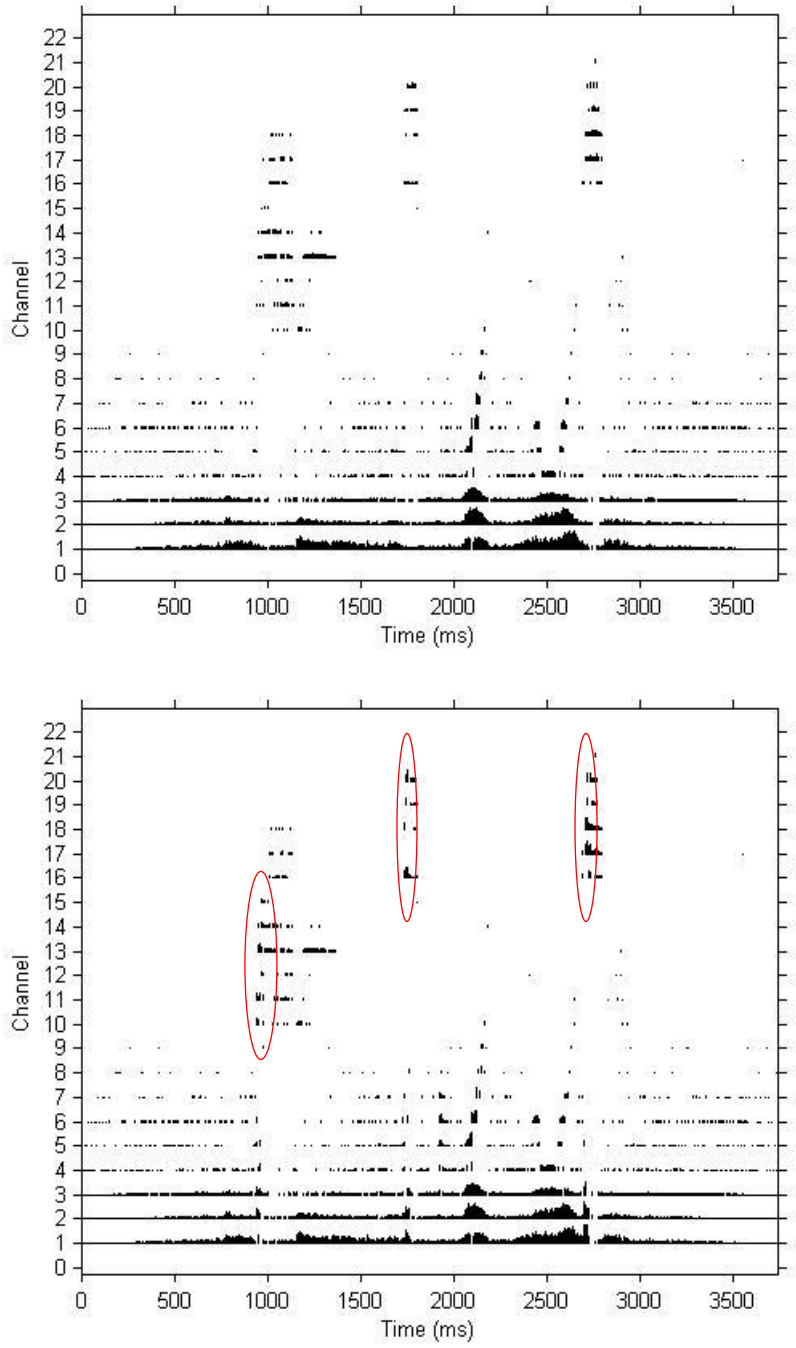


Figure 5. 24 Electrodograms for the unprocessed, noise-mixed sentence "The machine was quite noisy" (top panel) and the enhanced sentence (bottom panel). The red ellipses indicate regions in the high frequency channels where there has been an increase in stimulus activation, corresponding to boosted landmarks.

5.7.3 Results

Sentence test

Mean percent correct scores for each condition are shown in figure 5.25. Results from a two-way repeated measures ANOVA showed a significant effect of SNR ($F(1, 23)$= 146.4, p=.00), no significant effect of enhancement and no significant interaction between SNR and enhancement. When compared with results from experiment IV, baseline scores at +5 dB SNR improved by 43 percentage points with the normalisation stage of the LBP processing removed. On average, scores were 18 percentage points higher in the +5 dB SNR baseline condition in experiment V than for the +10 dB SNR baseline condition in experiment IV. Nonetheless, the pattern of results was similar for the two experiments.

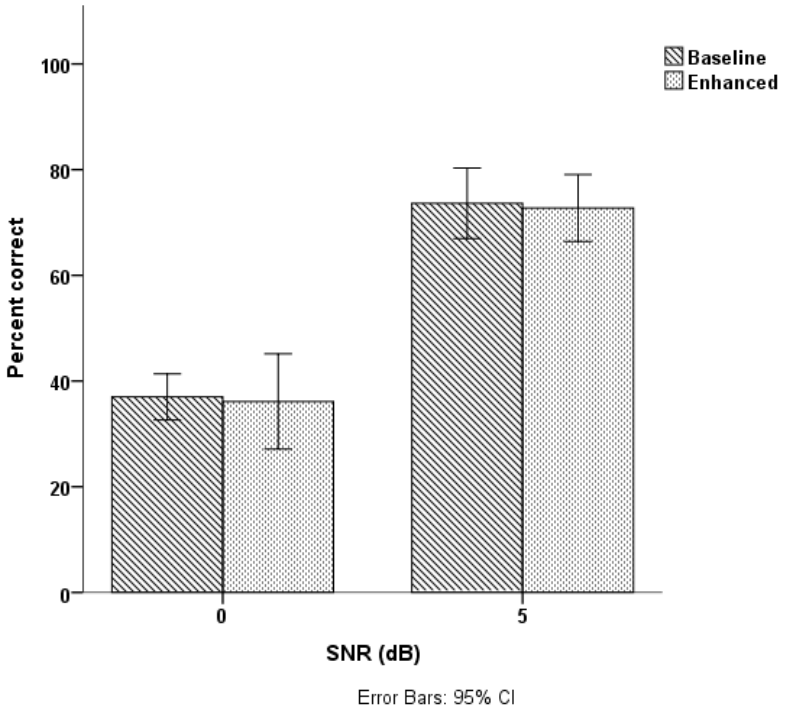


Figure 5. 25Mean percent correct scores for baseline and enhanced (15 dB boost) conditions at 0 and 5 dB SNRs.

VCV test

Mean percent correct scores for three vowel environments at both SNRs are shown in figure 5.26. A three-way repeated measures ANOVA was performed with three factors (SNR, vowel context and level of enhancement) with two levels for SNR (+5 and 0 dB) and level of

enhancement (baseline and 15 dB boost) and three levels for vowel context (/a/, /i/ and/u/). A significant effect was observed for SNR ($F(1, 87)$= 49.7, p= 0.00) and vowel context ($F(2, 86)$= 61.1, p= 0.00) but not for level of enhancement ($F(1, 87)$= 0.5, p= 0.5) and there were no significant interactions. Sidak's *post-hoc* test showed the same significant trend in percent correct scores as in experiment IV : /a/ > /u/ > /i/. Unlike for the sentence recognition task, removal of the normalisation stage of the processing did not appear to improve baseline results in the +5 dB SNR condition when compared with results from experiment IV.

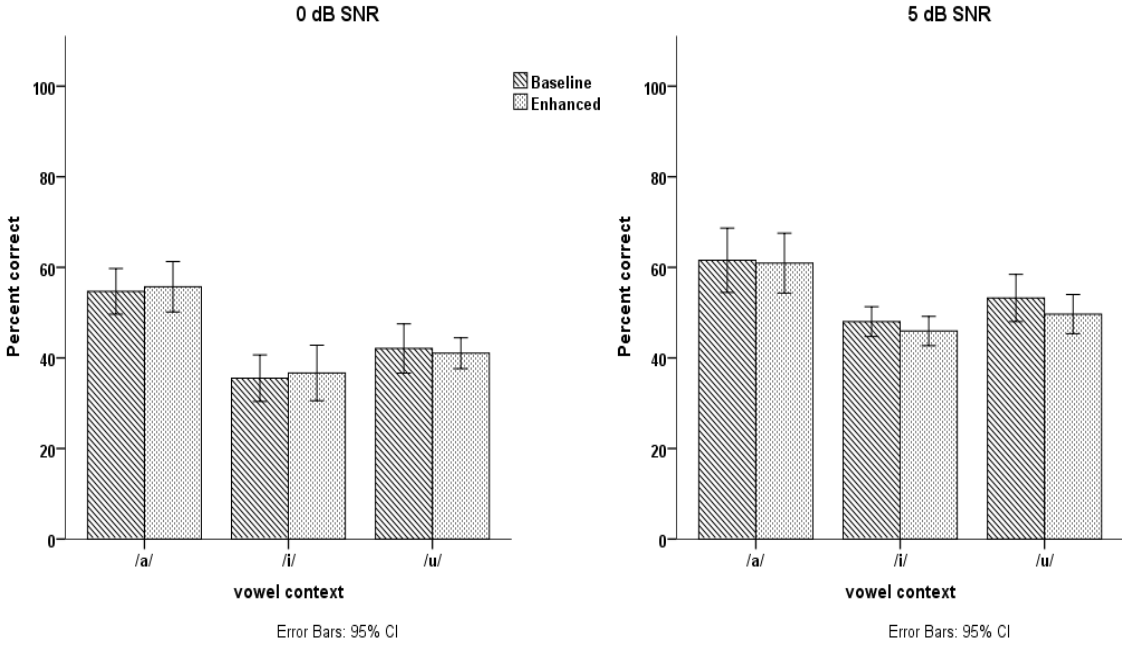


Figure 5. 26 Percent correct scores for each vowel in environment, with and without enhancement, in 5 dB SNR SS noise (left panel) and 10 dB SNR SS noise (right panel).

Consonant confusion matrices were computed for each of the conditions and subjected to information transmission analysis for the features voicing, place, manner, plosive and fricative. Percentage information transmission for the voicing, place and manner are shown for the three vowel environments in both levels of noise in figure 5.27 and 5.28 respectively. Similar plots are shown for plosive and fricative features in figure 5.29 and 5.30.

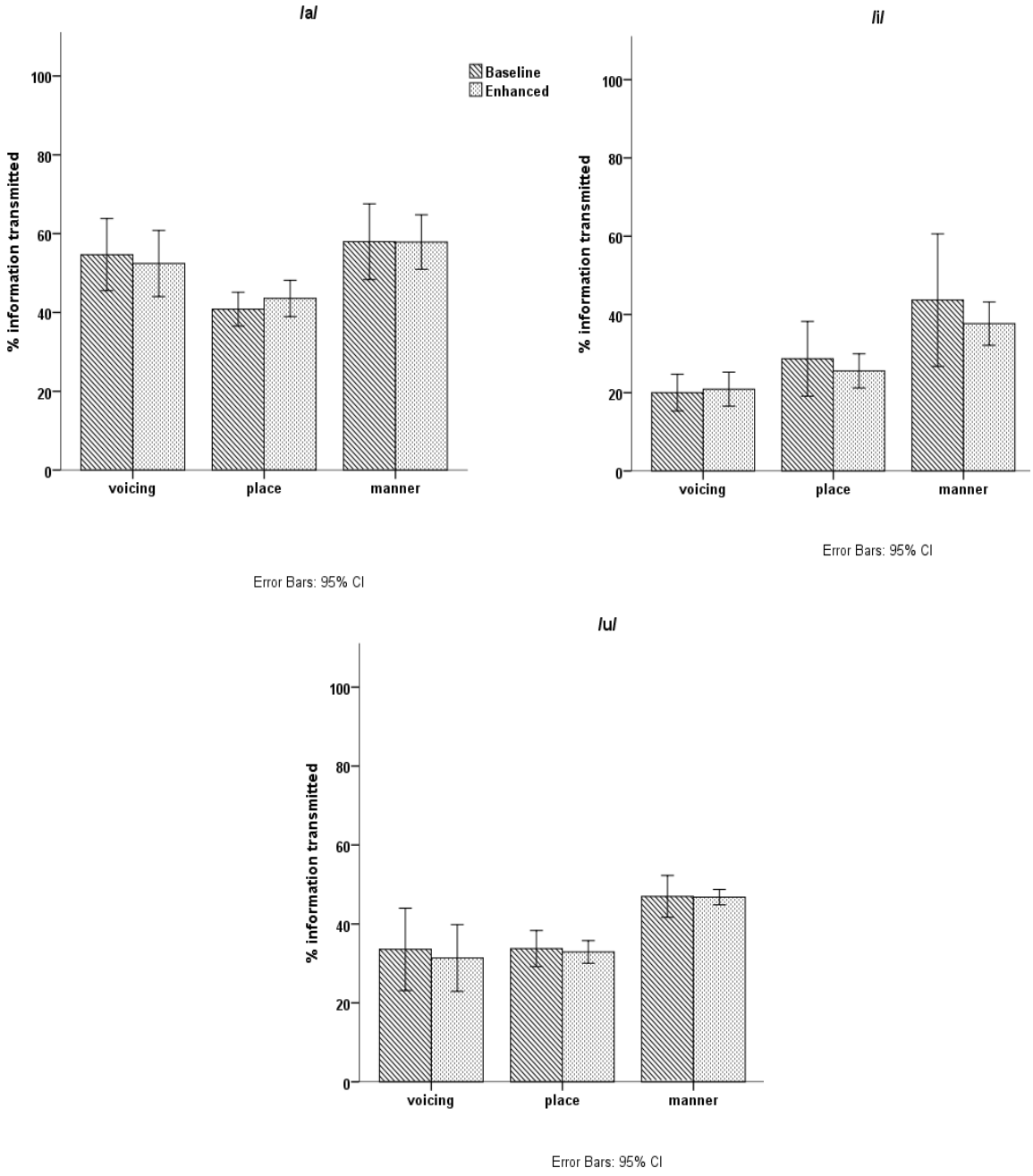


Figure 5. 27 Percentage information transmitted for the features voicing, place and manner for the vowel contexts /a/ (top left panel), /i/ (top right panel) and /u/ (bottom panel) at 0 dB SNR.

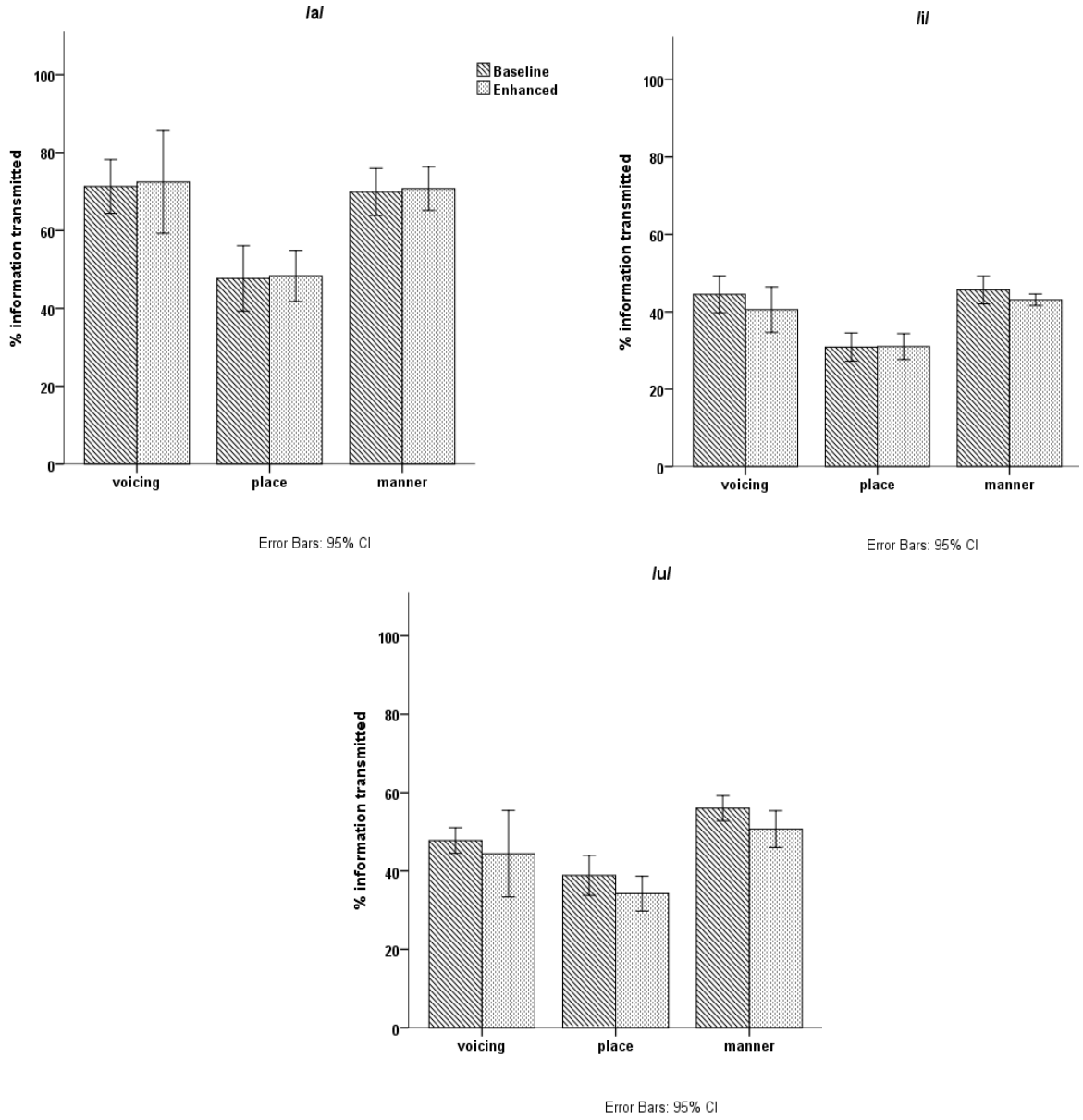


Figure 5. 28 Percentage information transmitted for the features voicing, place and manner for the vowel contexts /a/ (top left panel), /i/ (top right panel) and /u/ (bottom panel) at 5 dB SNR.

A MANOVA was performed with five dependent variables (percentage information transmission for the features voicing, place, manner, plosive and fricative), with three factors (SNR, vowel context and level of enhancement) and two levels for SNR (0 and +5 dB) and level of enhancement (baseline and 15 dB boost) and three levels for vowel context (/a/, /i/, /u.). Analysis showed a significant effect of SNR and vowel context on percentage information transmitted for all features (p<0.05), no significant effect of enhancement on any feature but a significant interaction between SNR and vowel context for the plosive feature.

Sidak's *post-hoc* test again revealed a significant trend that information transmission was significantly higher in the /a/ context compared to the /i/ and /u/ contexts and that scores in the /u/ context were significantly higher than in the /i/ context. This was true for all features analysed accept for plosive. A two-way repeated measures ANOVA for 0 dB SNR showed no significant effect of boost on the plosive feature ($F(1, 39)$= 1.6, p = 0.21) but a significant effect of vowel context ($F(2, 38)$= 4.7, p= 0.02). Sidak's *post-hoc* test revealed that results in the /a/ context were higher than for the /i/ context only. A two-way repeated measures ANOVA for +5 dB SNR again only showed a significant effect of vowel context on information transmission for the plosive feature. Sidak's *post-hoc* test showed that results in the /a/ context were significantly higher than for the /i/ and /u/ contexts but results with the /i/ and /u/ contexts did not differ significantly from one another.

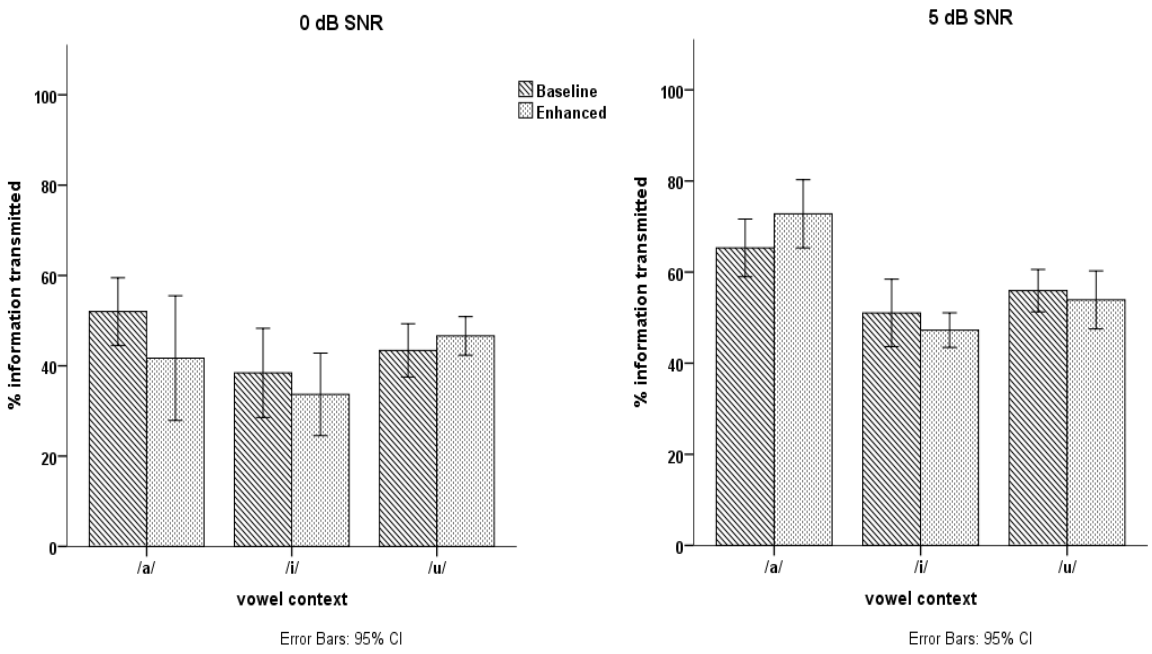


Figure 5. 29 Percentage information transmitted for the feature plosive in the vowel contexts /a/, /i/ and /u/ at 0 dB SNR (left panel) 5 dB SNR (right panel).

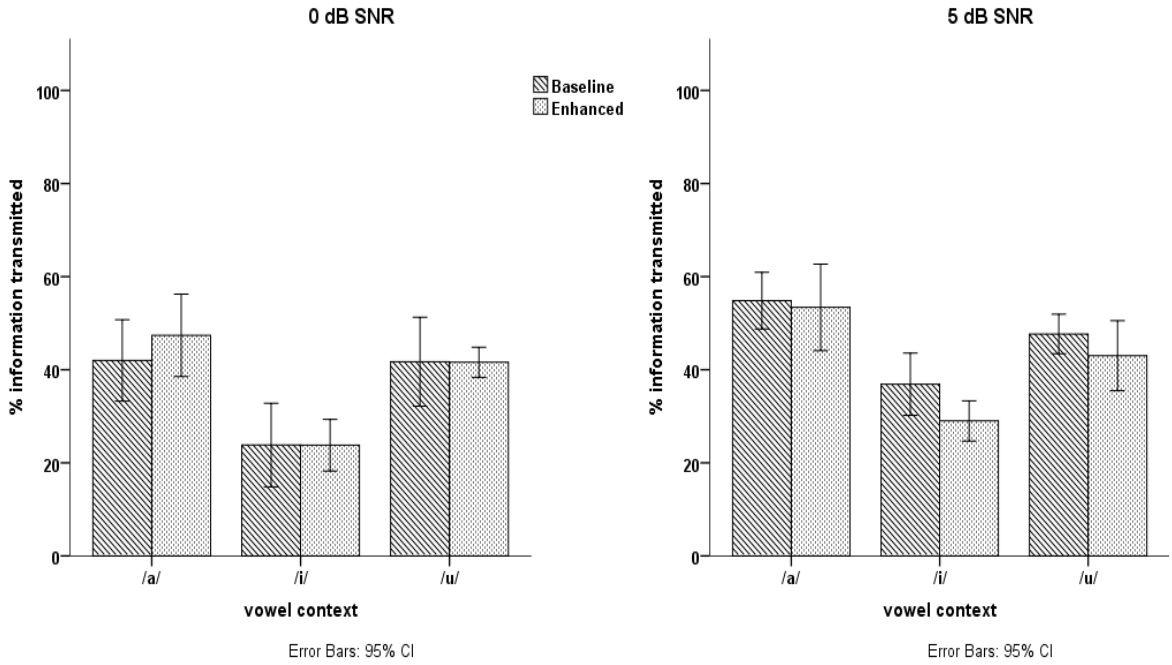


Figure 5. 30 Percentage information transmitted for the feature fricative in the vowel contexts /a/, /i/ and /u/ at 0 dB SNR (left panel) 5 dB SNR (right panel).

5.7.4 Discussion

Results once again showed no improvement in speech recognition scores or in the transmission of any specific features for NH listeners listening to vocoded speech with obstruent landmark enhancement applied. However, in general, obstruent landmark enhancement did not have a negative effect on speech recognition abilities. As per experiment III, the effect of experience on scores was considered, however, scores in the sentence recognition task were similar for the experienced and non-experienced participants. Average percent correct scores for the two groups for each of the four conditions are given in table 5.5. However, it is important to consider the small number of participants in each group and the fact that the experienced listeners from experiment III had participated in more than one vocoder study prior to taking part in the experiment whereas the experienced group in the present study had only had experience from the previous experiment.

It is possible that although participants were able to distinguish between the enhanced stimuli and the unprocessed stimuli, they did not know how to use the additional information. For all experiments, the participants only received training with unprocessed versions of the vocoded sentences/VCV stimuli and it is therefore possible that further training with enhanced

versions of the stimuli may result in listeners being able to make more use of boosted landmark information.

	Baseline 0 dB SNR	Enhanced 0 dB SNR	Baseline +5 dB SNR	Enhanced +5 dB SNR
Experienced	36.2	35.8	74	74.8
Non-experienced	37.1	36.1	73.6	72.6

Table 5. 5 Mean percent correct scores in the sentence recognition task for the experienced listeners and non-experienced listeners.

The VCV task showed that information transmission was always greatest for the feature manner and looking at the consonant confusion matrices for each participant it was evident that confusions tended to be made within manner categories. Low energy fricatives, such as /f/, /θ/ and /v/ were commonly misperceived in all vowel environments, whereas fricatives such as /ʃ/, /s/ and /z/ where generally well perceived. This is likely the result of the method used to mix the stimuli with the noise signal. The SNR was calculated based on the average RMS as calculated for all VCV stimuli (the same method was used to generate the noise-mixed sentences) and not on a case-by-case basis. Although all of the VCV stimuli were altered so that their average RMS level was the same as that of the groups average RMS level, this method would still have resulted in the local SNR for plosives and fricatives with greater amplitude being larger than for those which are generally considered as being weaker.

Percent correct and percentage information transmitted (for all features) was highest in the /a/ vowel environment. In fact, it is possible that in the 0 dB SNR condition, participants were scoring near chance level for the /i/ vowel context (close to 30 percent correct for some conditions). Percentage information transmission for the feature nasal was also analysed and it was found that this feature was very poorly transmitted by the vocoder processing, with scores no higher than 15 percent transmitted for /i/ and /u/ and between 15 and 24 percent transmitted for /a/. The sonorant consonants /r, j, l, m, n, w/ were, on the whole, very poorly perceived and commonly confused with one another. It was decided to include these consonant sounds, even though no enhancement was applied to them, as a control in order to test whether any potential improvement in the identification of obstruent consonants with enhancement was not due to chance. The introduction of the sonorant consonants also meant

that participants were able to make manner errors other than for plosive, fricative and affricate. Constraining the options to simply obstruent consonants may have artificially inflated manner and place scores, however, the inclusion of the sonorant consonants also increased the length of testing and therefore also increased the possibility of fatigue effects; thus introducing more errors. Nonetheless, the poor identification of the sonorant sounds during the test would suggest that they should be included in any landmark enhancement strategy.

In summary, this chapter has shown the development of an obstruent landmark enhancement strategy for use with an n-*of*-m style cochlear implant processing strategy. Results from vocoder studies however, do not show any improvement in speech recognition scores in noise with obstruent landmark enhancement applied. This finding is discussed further in chapter 6.

Chapter 6- Discussion

6.1 Summary of thesis and overall findings

The present thesis looked to investigate the benefit of enhancing ALs for CI users when listening in noise. The main aims of this research were to:

1. Develop a clear definition of acoustic landmarks
2. Investigate the role of ALs in speech perception, particularly in noise
3. Understand the effect current CI processing may have on AL transmission
4. Develop and evaluate a method of automatically detecting obstruent ALs that can be used with CI processing
5. Develop and evaluate a method of enhancing obstruent ALs

Evidence from past literature suggests that current n-*of*-m strategies that base channel selection on amplitude may not be effectively transmitting the elements of the speech signal that are important for speech perception. Initially, this thesis explored the feasibility of an n-*of*-m strategy which based channel selection on channel-specific landmark probabilities. Channel-specific landmark detectors were developed from the landmark detection algorithm of Juneja and Espy-Wilson (2008) and are outlined in section 4.3.2. Initial testing with channel-specific AL detectors suggested they may not be robust enough in the presence of background noise and it was therefore concluded that further development is required. As it was not within the scope of this thesis to continue development on the channel-specific landmark detectors the focus therefore changed to enhancement of landmarks with current CI processing techniques.

Li and Loizou (2008a, 2009 and 2010) and Chen and Loizou (2010) found benefit in providing NH listeners listening to CI simulated speech, and CI users, with clear obstruent consonant information when listening to speech in noise. The obstruent consonants are more affected by the addition of noise than sonorant sounds (Parikh and Loizou, 2005), are most commonly confused by CI users (Munson and Nelson, 2005), and landmarks relating to acoustically abrupt changes in the speech signal account for almost 70 percent of all landmarks in speech (Liu, 1996). For these reasons, the present thesis focused on the

development of an enhancement strategy that would improve the transmission of information relating to obstruent consonants.

Chapter 5 outlined the development and evaluation of such a strategy. Initially the obstruent landmark enhancement strategy was guided by the output an automatic landmark detection algorithm, again adapted from method proposed by Juneja and Espy-Wilson (2008), and applied a simple boost at the times of plosive bursts and fricative labels. Unfortunately, as was found to be the case with the channel-specific detectors, the accuracy of the algorithm was poor after the addition of even moderate levels of background noise. This resulted in the insertion of incorrect plosive and fricative labels, and led to unwanted boosting of the noise signal. This had a detrimental effect on speech recognition scores. Further work with the enhancement strategy was based on obstruent landmark labels which were manually generated by the author of this thesis in an attempt to mimic perfect landmark detection. However, despite several modifications to the timing and frequency characteristics of the boost functions, the proposed obstruent landmark enhancement strategy failed to generate any improvement in speech recognition scores when listening in noise (NH listeners listening to a cochlear implant simulation). The following sections discuss reasons for why an improvement was not seen.

6.2 Methodological considerations

6.2.1 Vocoder simulation

Cochlear implant simulations are often used as a method for predicting CI user performance and allow for new speech processing techniques to be evaluated in the absence of confounding variables, such as number of channels available for stimulation, stimulation degree of current spread etc. The use of vocoder simulations in CI research and the different carrier signals that can be used was discussed in section 2.4. Typically, CI simulations have been found to better represent the performance of the best performing CI users (Fu et al., 2004; Fu and Nogaki, 2004; Oxenham and Kreft, 2014). Although a number of factors may contribute to the wide variation in speech recognition in noise observed between implant listeners (for example, duration of deafness prior to implantation and number and pattern of surviving neurons), it is likely that the better performing users are able to make use of, or "extract", more information from the restricted signal which they receive. This may include better access to or extraction of landmark information.

As the present author wanted to test the proposed landmark enhancement strategy in SNRs which are representative of situations in which typical CI users have difficulty, a noise-band carrier was used so as to limit performance. Noise-band vocoders also stimulate a larger portion of the basilar membrane and this is closer to the stimulation of neurons with electric hearing than that produced by a sine vocoder.

In the earlier experiments conducted as part of this thesis (I-III), NH participants listened to speech processed in line with a 12-*of*-22 ACE strategy. However, with this relatively high number of stimulation channels, the participants were typically scoring around 80 percent correct in the BKB sentence task at SNRs as low as +3 dB. It was therefore considered that this was likely over-estimating the majority of CI users' performance in similar conditions, and that the NH listeners were able to resolve a significant amount of the information present in the coded signal, including information relating to obstruent acoustic landmarks. Munson et al. (2003) suggested that simulation studies exploring the effects of enhancement techniques should not worry about whether the results are a better predictor of best versus poorer performing users, as the trends in error patterns are the same for both groups. However, poorer performing CI users may receive less or are unable to extract as much information from the transmitted signal, so may therefore benefit more from enhancement. In order to be able to test the effects of obstruent landmark enhancement at SNRs more representative of those in which CI users struggle and without over-estimating performance, channel selection was restricted to a 3-*of*-22 strategy in experiments IV and V. Although this method was successful in limiting baseline performance in the sentence recognition task, it was possibly too hard for some of the vowel environments in the VCV task, with performance at or around chance level.

The goal of boosting the signal at times of obstruent landmarks is to increase the chances of the corresponding channels being selected for stimulation with an n-*of*-m strategy. Comparison of the electrograms for the unprocessed and enhanced stimuli demonstrated increased activation in some channels at the onsets of obstruent segments (see section 5.7.2); however, the selection of channels remained relatively unchanged. For an amplitude-based n-*of*-m strategy with so few channels selected in stimulation in each frame, it is possible that channel selection is not likely to differ much between unprocessed and enhanced stimuli and that the only difference that may be observed will be one of increased stimulus level at the times of the applied enhancement. The increase in stimulus amplitude may help listeners to distinguish between segments or identify that a new sound has begun, however, without

changing or improving the information transmitted around these events (by changing the channels stimulated) then speech sound identification or word identity will not improve. In addition, the use of a noise-band vocoder with only a small number of channels will convey primarily temporal envelope cues rather than spectral cues. Whitmal et al. (2007) argued, however, that the inherent fluctuations of a noise carrier may interfere with envelope cues and make it difficult for a listener to extract useful information form the target speech signal by masking modulations within the envelope.

Testing whether increasing spectral resolution also helps to improve the transmission of acoustic landmarks, and ultimately helps to improve speech recognition in noise, would require using much lower SNRs to avoid ceiling performance. Figure 6.1 shows example electrodograms for the VCV token a/θ /a, with and without landmark enhancement, at 5 dB SNR for a 3-*of*-22 ACE strategy (top panel) and 12-*of*-22 ACE strategy (bottom panel). The figure demonstrates that both strategies show increased activation across channels in the analysis frames corresponding to the parameters of the landmark boost function applied (as per experiment V). When compared with electrodograms for the unprocessed version of each of the speech tokens, it appears that channel selection does not differ greatly for the enhanced token with either strategy. What is evident, however, is that there is more activation across a greater number of channels for the enhanced token processed to mimic the 12-*of*-22 strategy.

It is possible that by providing enhancement over a larger number of channels listeners would have access to more landmark information and possibly even more information about the surrounding cues to consonant identity. However, a further observation that can be made from these diagrams is that there is more activation across channels in the 12-*of*-22 strategy as a result of the background noise that has been added to the signal. Nonetheless, increasing channel number, and by extension spectral resolution, has been shown to improve speech recognition in noise, at least for NH listeners listening to vocoded speech (Shannon et al., 1995; Dorman et al., 1998; Fu et al., 1998). The question still remains however, as to whether this increased activation, alongside increased spectral resolution, would result in "useable" information. Having to test in lower SNRs to avoid ceiling effects could mean that the important landmark information is in fact masked further as a result of the LBP also applying the boost to the noise signal; meaning that listeners are still not able to access the landmark information. The effect of boosting the noise is considered further in section 6.2.3.

For the final experiments it was considered that spectral smearing could be used to limit performance in the desired SNRs without having to simply restrict the number of channels available for activation. This approach is more representative of what happens when electrodes are stimulated in an implanted cochlea, as detailed in section 2.3.3, however, there was not enough time to introduce a model of channel interaction into the LBP coding prior to the undertaking of the last two experiments, and Oxenham and Kreft (2014) and Whitmal et al. (2007) have demonstrated restricting channel number is possibly as good a predictor of CI user performance as vocoder simulations which model some degree of channel interaction. It is therefore recommended that future work should investigate the potential benefit of obstruent landmark enhancement with increased spectral resolution, with at least some degree of spectral smearing applied to model channel interaction.

In addition to the limitations of the reduced spectral resolution used in the final experiments, the default settings of the NMT also resulted in simulations with 100 percent dynamic range. This may have contributed to the high levels of speech recognition obtained for spectral resolutions greater than the three channels used. Reduced amplitude resolution has been shown to impair the transmission of rapid amplitude and spectral changes important for distinguishing between segments (Loizou et al., 2000a). The CI simulation as generated by the NMT for the present thesis would have had both a wide input DR and, due to the healthy cochleae tested during this study, a wide output DR. A better model of CI processing and user performance would be to restrict the DR, potentially investigating the effects of obstruent landmark enhancement with a number of different DRs.

A further consideration is that by using simulation experiments to predict CI user performance, the NH participants recruited have, by extension, normal functioning cochleae. Section 3.3.1 considered the importance of adaptation of auditory neurons to sudden changes in amplitude and/or frequency for coding events in the speech signal which indicate new information. This adaptation is lost in an implanted ear and it was therefore suggested that enhancing obstruent landmarks, which are signalled by sudden amplitude and spectral changes, may, in part be able to compensate for this. NH listeners may already be able to resolve sudden spectral and amplitude changes from within the vocoded signal and therefore, landmark enhancement did not improve their speech perception scores for this very reason. The only way to determine whether landmark enhancement could be used to improve neural adaptation in an implanted ear would be to measure the firing rates of ANFs. This is a fairly invasive procedure and would likely require the use of an animal model.

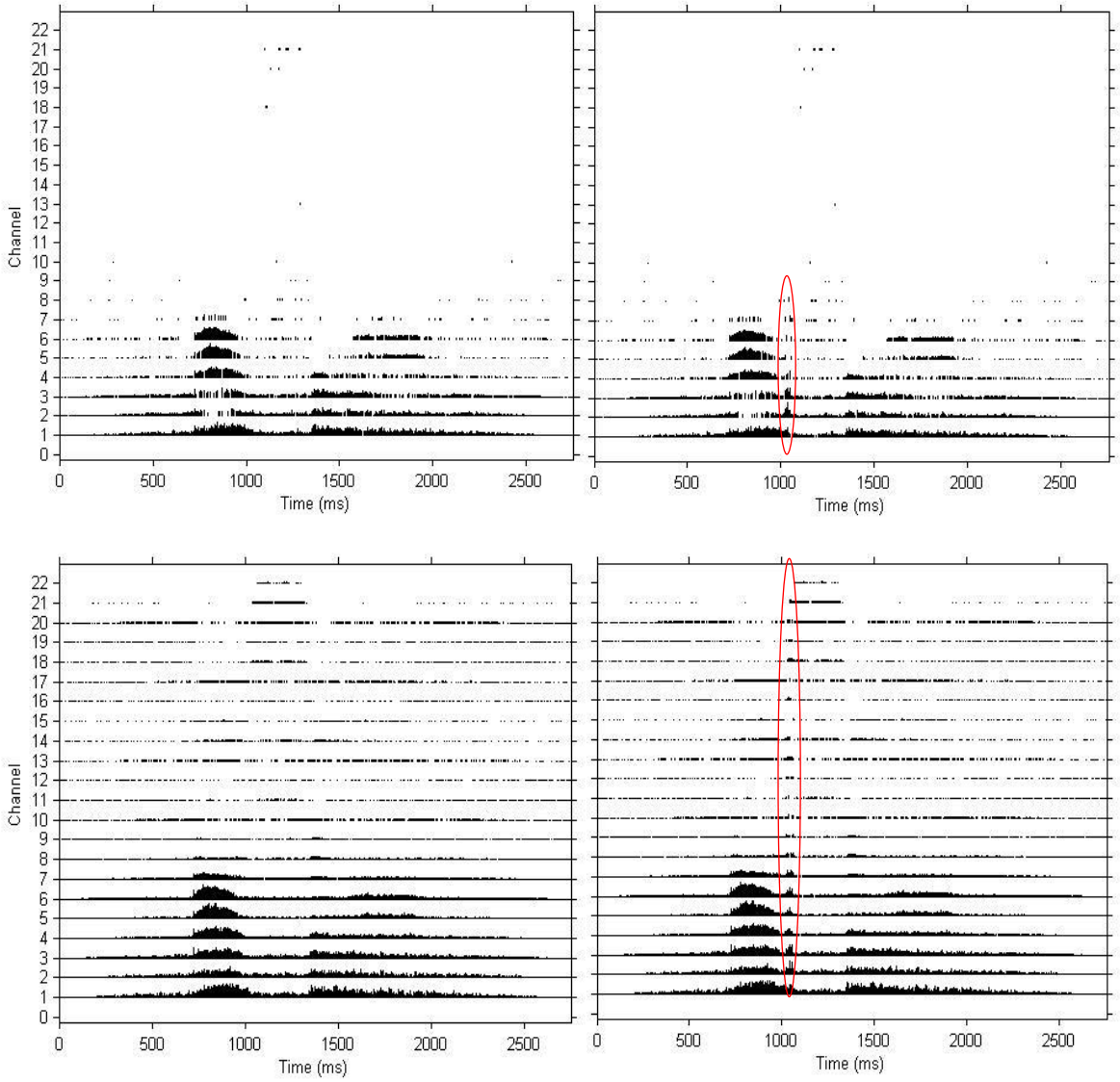


Figure 6. 1 Electrodograms for the unprocessed (left panels) and enhanced (right panels) versions of the VCV token a/θ/a as processed by a 3-of-22 (top panel) and 12-of-22 (bottom panel) ACE strategy. The red ellipses indicate regions of increased activation in the enhanced condition.

6.2.2 Speech material

Open-set speech recognition scores were tested as they have a larger degree of coarticulation and context variability than other, shorter speech tokens, such as words or VCV stimuli. The BKB sentences were chosen because they are commonly used to measure the speech recognition abilities of both prospective and existing CI users. Equivalent sentence lists also exist for different languages and dialects meaning that results from this study are comparable with studies utilising these lists.

Evidence suggested that one list per test condition should be sufficient to account for the random variation within the test material (Thornton and Raffin, 1978), however, in the initial experiments undertaken as part of this thesis scores between listeners seemed to vary greatly. It was not clear whether this variation was the result of individual differences in the ability of participants to extract information from the vocoded signal or whether it was because particular lists were harder or easier than others. In the final two experiments, participants listened to more than one sentence list per condition and scores even varied greatly within subjects' scores. BKB lists 2, 12, 17 and 21 are known to produce scores which are significantly different from the overall mean score across all lists (Bench and Bamford, 1979). As some participants listened to some of the same lists (either for the same condition or in a different processing condition) the pattern of scores for the lists were scrutinised, however, there was no clear indication that participants found a particular list harder or easier.

One possible explanation is that participants' attention levels varied throughout testing and that this had a considerable influence on scores. Another potential cause of the variation observed is the vocabulary used in the sentences. The BKB sentences were originally developed to test the speech recognition abilities of children and therefore they use fairly simple language and are possibly somewhat predictable. There are common key words used across the sentence lists, such as "she", "he", "they", "they're", "dog", "girl" and "boy" and participants may score higher for sentences which contain these more common words. There are also some sentences which appear to stand out from the others as perhaps being less predictable, such as, "Some animals sleep on straw", "Lemons grow on trees" and "The green tomatoes are small". Although the actual responses from the participants were not recorded, during testing it was noted that scores for these "less predictable" sentences tended to be lower.

The BKB sentence recordings used were from a single male talker and have been produced in a clear speaking style. This may naturally serve to improve the transmission of certain speech cues in the presence of a competing noise source. Hazan and Simpson (2000) showed a significant effect of speaker on the amount of improvement seen with a landmark enhancement strategy (for NH listeners listening to VCV tokens). It is therefore possible that an improvement was not seen with the landmark enhancement strategy proposed in this thesis because landmark information was already naturally enhanced as a result of the speaker's clear pronunciation of each sentence. It may then be beneficial to test the proposed obstruent

landmark enhancement strategy using speech material that is more representative of conversational speech.

As sentence tests measuring key word identification do not represent any improvement in the transmission of specific sounds, a VCV task was included in the final two experiments to explore the effects of obstruent landmark enhancement on the transmission of specific features. VCV tasks are inherently tiring and boring, therefore the number of repetitions used must be carefully considered so as to avoid fatigue effects. This can also be helped by the introduction of regular breaks. A minimum of three repetitions is usually considered sufficient to account for random variation and potential errors in identity selection (i.e. accidentally selecting the wrong icon from the GUI). Considering that two versions of each VCV token were selected (to consider variations in production) and that participants were tested in three vowel environments, it was decided that no more than the minimum three repetitions should be used for each speech token.

It is difficult to compare the pattern of results obtained for the VCV task in experiment V with other CI simulation or CI user studies as there are very few studies which have measured consonant confusions and feature information transmission in noise. As far as the author is aware, there are also no studies which have measured VCV scores for either CI users or NH listeners listening to vocoded speech in SS noise, for the SNRs and vowel environments used in the present study. However, CI user studies which have explored information transmission in quiet, have shown that manner and voicing features are significantly better transmitted than the place feature (Dorman, 1995; Munson et al., 2003) in the a/C/a environment. Verschuur (2009) showed a similar pattern of results for a group of nine Nucleus 24 CI users listing to VCV stimuli in the presence of SS noise at 10 dB SNR in the vowel context i/C/i. This finding of the place feature being most poorly transmitted with CI processing was also found in the present study but for all vowel environments.

Hazan and Simpson (2000) and Geurts and Wouters (1999) have both shown a significant trend that /a/>/i/>/u/, even though the former measured performance in noise for NH listeners and the latter measured performance for CI users in quiet. This is in contrast with the trend shown in the current thesis, that /a/>/u/>/i/ (in both experiment IV and V). The inclusion of the sonorant consonants in the present study may have contributed to the different pattern of results seen for the vowel environments, as Hazan and Simpson (2000) included the nasal consonants in their study but no other sonorant sounds and Geurts and Wouters (1999)

included the stop consonants only. In general, the transmission of the nasal feature was found to be extremely poor for all vowel environments in the final VCV task, and the sonorant consonants were often confused with one another. This is likely the result of the limited number of spectral channels used in the present study, with listeners not being able to resolve important low-frequency information relating to the nasal murmur or important formant transitions in the region of F2 and F3. Although no processing was applied to the sonorant consonants they were still included in the VCV task (see section 5.7.4) and the results indicate that the sonorant consonants may also benefit from some form of landmark enhancement.

Finally, the method in which the noise was added to the speech stimuli may have affected scores in both tasks. Prior to applying any processing to the stimuli the average RMS, as measured across all utterances, was computed for the BKB, IHR and VCV stimuli. Each recording was then altered to have the same average RMS level as the group mean level. The SNR was then calculated based on the average RMS as measured across all the unprocessed stimuli, rather than on a case-by-case basis. This may have resulted in the local SNR for some speech segments or consonant tokens being higher than for others, thus making them easy to detect. This could potentially account for the wide variation seen between some of the BKB sentence lists and for why certain consonant sounds were commonly well perceived, regardless of SNR or vowel environment.

6.2.3 Enhancement of obstruent landmarks

The literature indicates that the transmission of obstruent acoustic landmark information is impaired by current CI processing techniques, and supports the notion that CI users may achieve higher speech recognition scores in noise if they have better access to acoustic landmarks, and in particular, those pertaining to the obstruent consonants. So far this chapter has considered the potential limiting factors of the simulation parameters used in the experiments and the impact they may have on the potential benefit of the obstruent landmark enhancement strategy proposed in this thesis. However, it is possible that the method for enhancing landmarks, as outlined in chapter 5, does help to improve the transmission of obstruent AL information.

Section 6.2.1 considered the differences in electric versus acoustic stimulation of the auditory pathway and suggested that NH listeners are not able to make use of the boosted landmark information because they are already able to resolve landmarks from the unprocessed stimuli.

It may also be the case the CI users are not able to make use of the boosted landmark information, if the proposed method does not actually change/improve the information transmitted. Chapter 2 argued that speech perception with current implant design could be improved if we make more efficient use of the capacity available and select only the important elements of the signal to be transmitted by the implant. The cue enhancement studies explored in section 4.3.3 suggested there may be benefit in boosting regions of the speech signal which contain information relating to acoustic landmarks; however, this approach involved applying enhancement to the signal *before* the noise was added and was only found to be beneficial for negative SNRs. The method of obstruent landmark enhancement proposed in this thesis applies the enhancement to the speech signal *after* it has been mixed with the noise.

This approach was used because it is representative of the signal that would be available to the speech processor of a CI and is therefore more likely to give a realistic prediction of the benefit of the proposed enhancement strategy. However, this results in the noise signal being boosted alongside the desired landmark information, and may mean obstruent landmarks are no more accessible or intelligible than in the unprocessed stimulus. Boosting background noise may also result in forward masking effects that impact on the ability of listeners to resolve the surrounding important formant transitions. The introduction of background noise may also effect the COG calculation and alter burst and frication cues, especially for low level speech sounds which are more dominated by the noise. This effect may be more pronounced for more modulated speech maskers (e.g. two-four person babble) which will contain conflicting speech cue information, resulting in energetic masking. However, the burst portions of plosive sounds are generally less affected by the addition of background noise than the relatively weak closure period (Hu and Wang, 2003). This closure is represented by a period of silence in the spectrum, the duration of which is an important cue for plosive identification (section 3.1.3); however this cue is not enhanced by the proposed LBP.

Figure 6.2 considers the effect of applying obstruent landmark enhancement prior to the noise signal being mixed with the quiet sentence. The figure shows broadband spectrograms for the example sentence “The machine was quite noisy”, which was used to demonstrate the effects of the LBP in experiment V. The top panel represents the original noise-mixed sentence and the panels below represent the same sentence with obstruent landmark enhancement applied *after* the noise was added and *before* the noise was added. The red ellipses shown on the

spectrogram for the sentence with enhancement applied *after* the noise had been added (middle panel) indicate regions where the COG calculation may have been influenced by the background noise, as the enhancement is applied across a greater frequency region than is observed for the same sentences with enhancement applied *before* the noise was added (bottom panel). Although the enhancement strategy investigated in this thesis did not result in poorer speech recognition scores, it is possible that boosting the noise-mixed signal introduced conflicting burst and place of articulation cues for plosives and fricatives if the spectral region of maximum intensity was altered.

This boosting of lower frequency information is evident in the corresponding electrodogram (middle panel, figure 6.3), with increased stimulation of apical electrodes. The bottom panel of figure 6.3 represents the electrodogram for the same sentence as displayed in figure 6.2, with enhancement applied *before* the noise was added. Point C on the electrodogram demonstrates regions of enhancement where low-frequency channel activation is less than for the sentence with enhancement applied *after* the noise was added. In fact, the activation pattern in the low-frequencies for the enhancement *before* adding noise is similar to that of the unprocessed, noise-mixed sentence (top panel of figure 6.3). Without the increased activation in the low-frequency channels, this therefore means that greater current amplitude can be applied to the mid-high frequency channels and is demonstrated by points A and B in figure 6.3.

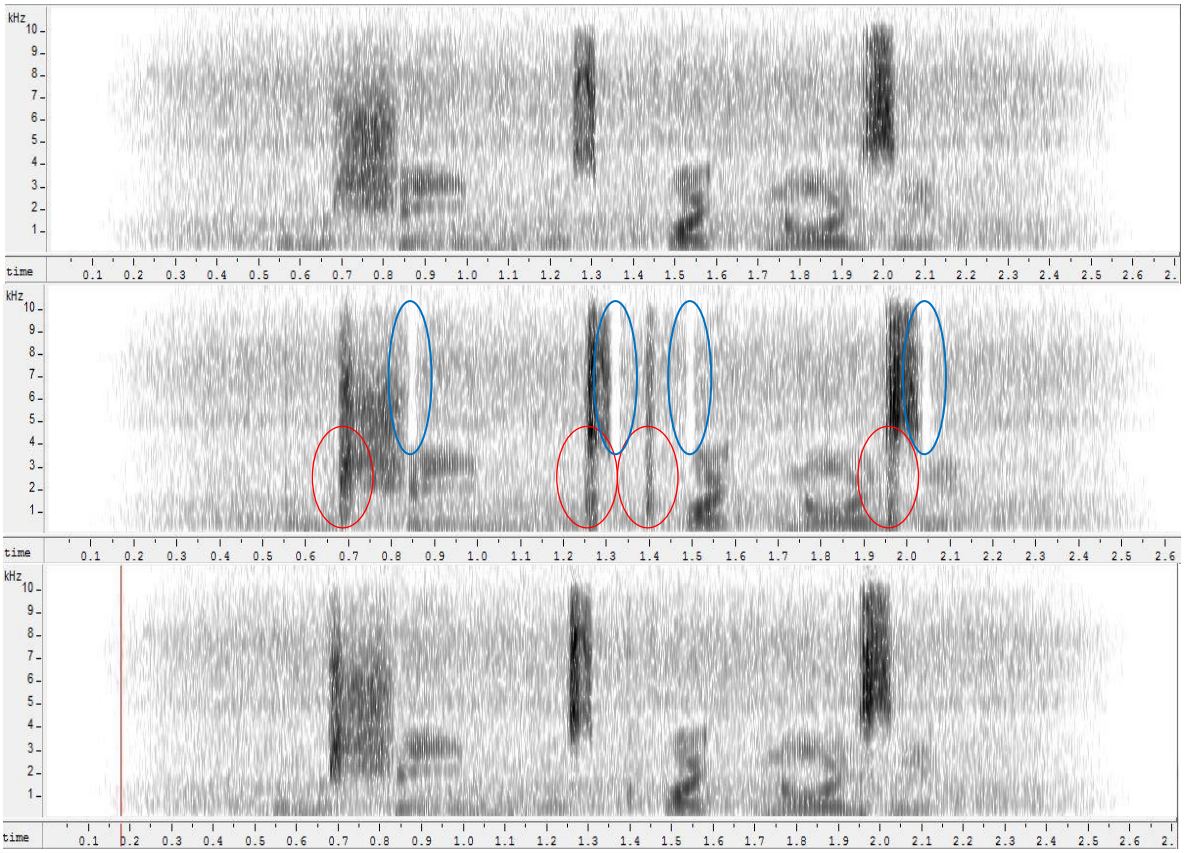


Figure 6. 2 Broadband spectrograms for the BKB sentence "The machine was quite noisy" as mixed with SS noise at 5 dB SNR. The top panel represents the original, unprocessed noise-mixed sentence, the middle panel represents the sentence with 15 dB level of enhancement applied <u>after</u> the noise was added and the bottom panel represent the sentence with 15 dB level of enhancement applied <u>before</u> the noise was added.

One of the aims of landmark enhancement is to restore the natural amplitude fluctuations in speech and improve VC contrasts. Hazan and Simpson (2000) used raised cosine ramps to blend enhanced regions of the signal with preceding and successive segments in order to avoid waveform discontinuities. The boost functions applied to the signal at the different landmark labels in the present study were designed to retain the abrupt nature of these events and therefore generally had fast rise times. When listening to the non-vocoded versions of the enhanced stimuli (noise added before enhancement), the boosted regions of the signal do sound like artefacts, however, this is less pronounced in the vocoded versions of the utterances and results do not indicate this had an effect on intelligibility.

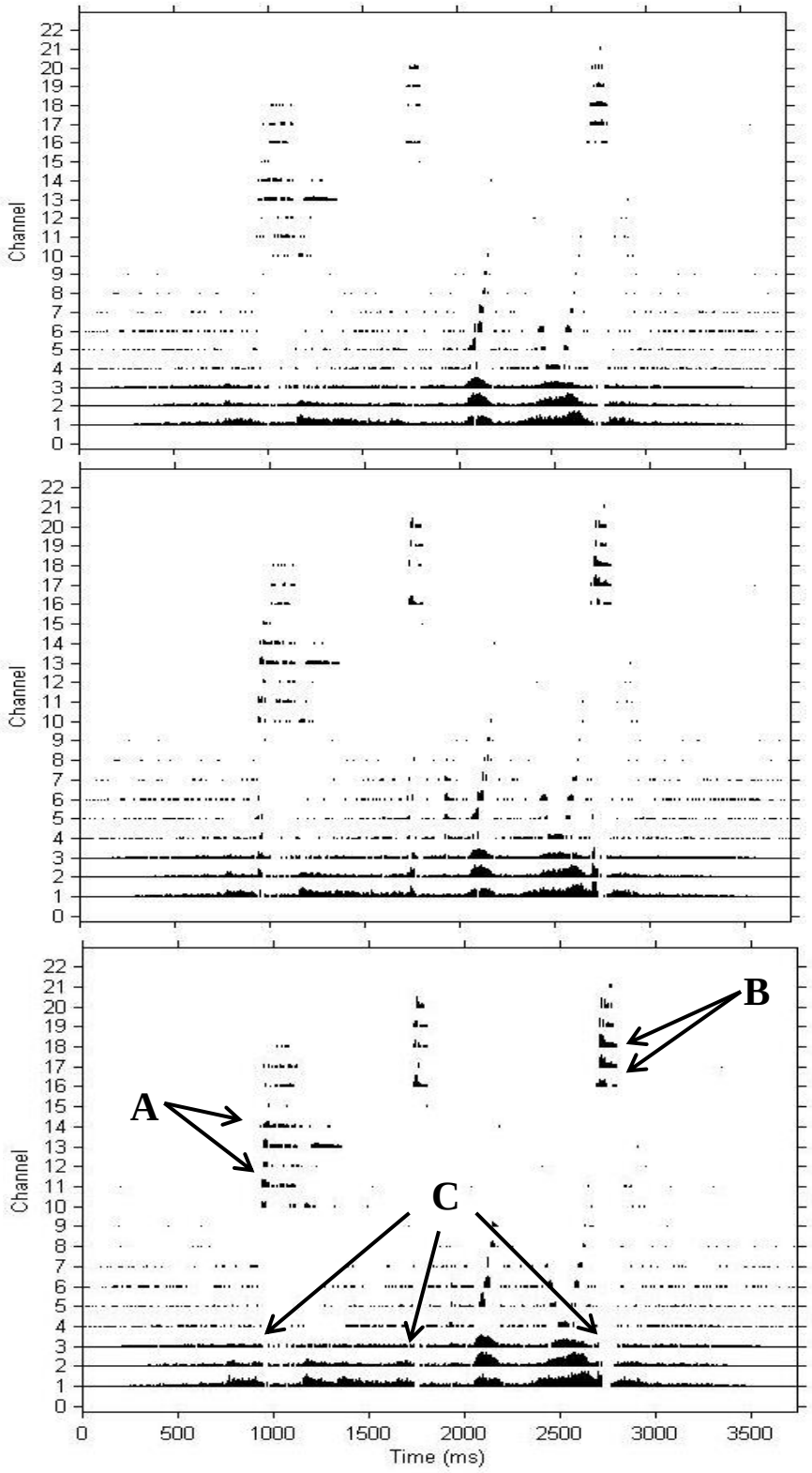


Figure 6. 3 Electrodograms for the BKB sentence "The machine was quite noisy" as mixed with SS noise at 5 dB SNR. The top panel represents the original, unprocessed noise-mixed sentence, the middle panel represents the sentence with 15 dB level of enhancement applied <u>after</u> the noise was added and the bottom panel represent the sentence with 15 dB level of enhancement applied <u>before</u> the noise was added.

The non-vocoded and vocoded versions of the BKB sentences, with enhancement added before the noise, sound more natural. In addition, the activation of the mid-high frequency electrodes at the onsets of the obstruent consonants (as highlighted by points A and B in the electrodogram) are more temporally aligned for the sentence in which landmark enhancement was applied *before* the noise was added than for the unprocessed sentence and even for the processed sentence with enhancement applied *after* the noise was added. This suggests that the abrupt nature of the boost functions is appropriate and can lead to better representation of the obstruent landmarks in the stimulation pattern of the electrode array. The blue ellipses in figure 6.2 demonstrate that when enhancement is applied to the clean signal prior to adding the noise the effect of the processing applied at **Off** labels is lost; the region is filled with the noise signal. However, when the electrodograms for the two processing methods are compared (figure 6.3), there is little evidence that the processing applied at **Off** labels has any effect on channel selection for segments following an obstruent consonant. The other landmark labels are considered in section 6.3.1.

Unlike for vocoded versus non-vocoded speech and for the change in SNR, participants did not comment on the relative difficulty or subjective sound quality of the enhanced versus unenhanced conditions. As it was not directly assessed during the studies, it is not possible to determine whether this was because the participants could not perceive any difference between the enhanced and unenhanced stimuli or if they could hear a difference but it made no difference to intelligibility. The participants that took part in the pilot study of experiment V were asked to choose the level of enhancement they felt sounded most intelligible, whilst also remaining as natural as possible. In addition, when listening to the processed stimuli, the author (who of course is not a naïve listener) was able to clearly perceive the change (particularly for experiment V); with processed sentences sounding slightly crisper than the unprocessed sentences.

Sound quality and listening effort are both important aspects of speech perception and are often neglected in the initial evaluations of new speech processing strategies. However, a number of the studies reported in chapter 2 have attempted to explore the issue of sound quality by asking CI users which of the speech processing strategies evaluated they preferred (Skinner et al., 2002a; James et al., 2002; Holden et al., 2005). These studies have demonstrated that processing strategies which do not improve speech intelligibility are still sometimes preferred by CI users because they may improve sound quality and/or listening effort. Therefore, future work with the LBP should investigate a range of levels of

enhancement and evaluate their overall intelligibility as well as subjective sound quality and listening effort, and compare these with existing speech processing strategies.

The present thesis chose to focus on the enhancement of obstruent landmarks only, however, results from both experiments I and V suggest there may also be benefit from enhancing the sonorant consonants, especially nasals, as they were generally very poorly perceived by listeners and often confused with one another. Future work with the LBP should therefore look at ways for enhancing sonorant landmarks as well. Ultimately, the continued development of the LBP would also require development of an ALD method which is more noise-robust as results from early experimental work in this thesis suggest that current methods are not accurate enough in the SNR levels required for testing.

Finally, experiment I considered whether landmark enhancement should be applied at specifically defined landmark events. The modulation enhancement strategy explored in this experiment did not apply the processing at defined landmark regions but rather attempted to target the obstruent landmarks by amplifying modulation frequencies associated with rapid spectral changes (3-100 Hz). There is a strong relationship between speech intelligibility and the modulation spectrum (Hodashima et al., 2002), with modulations in the region of 3-20 Hz considered as most important for speech perception; this is because they reflect syllable structure, and the dynamic changes as the result of changes in lip, jaw and tongue position (Greenberg et al., 2003; Greenberg and Arai, 2004).

Results from the experiment did show a small, but significant improvement in the transmission of the manner, plosive and nasal features with vocoder processing. However, this improvement was not seen for all levels of enhancement and was only observed for babble noise. Speech intelligibility scores were measured using a VCV task, using a single vowel context (aCa) with stimuli processed by a single talker. To determine whether modulation enhancement may be a more effective way of improving landmark transmission, and consequently improving speech recognition scores in noise, further work should include measuring scores for sentence level material as well as exploring more vowel contexts.

6.2.4 Landmark labels

The definition of acoustic landmarks developed during the experimentation phase of this thesis. At the end of chapter 3, landmarks were defined as either “high-frequency” (or obstruent landmarks), relating to the rapid onsets and offsets of a segment of aperiodic noise,

or “low-frequency” which describes the slowly varying, periodic segments of vowels and sonorant consonants. The output of the ALD sub-divided the obstruent landmarks into plosive and fricative landmarks and allowed for different processing to be applied at each label, i.e. rise time and duration of boost. A series of hand-generated labels were then created for the BKB and IHR sentence lists in which the plosive landmarks were then sub-divided in to short-duration plosives (**P**) and longer duration plosives (**Pn-** with a period of aspiration following the burst). However, as no effect of enhancement was seen, even for the VCV task, there is no evidence to suggest that there is any benefit of distinguishing between long and short-duration plosives. This also suggests that duration may be an unreliable cue for distinguishing between fricatives and plosives.

Consequently any future developments of the LBP might consider returning to the original two landmark labels, plosive and fricative, as used in experiment II. This would be less computationally complex than developing automatic landmark detectors that can distinguish between very short duration plosives and those with aspiration following the burst. If such landmark detectors were to be developed then they would require a longer analysis period over which to make the classification and this would lead to an increased processing delay. Ortega et al. (2000) found that with automatically identified landmark labels, the accuracy of distinguishing between bursts, aspiration and fricatives was not high enough to apply different processing techniques so they were grouped under a single label. In fact, in Stevens’ (2002) definition of landmarks, he does not use different landmark types to distinguish between obstruent manner categories but rather uses a simple “onset” and “offset” landmark distinction for obstruent consonants. In his LAFF model, Stevens uses articulator-bound features to help distinguish between manner categories, and these features are “uncovered” by analysing the signal in the vicinity of an onset or offset landmark. However, the important cues that define these articulator-bound features are often only within 10ms of the corresponding landmark and are therefore difficult to separate from the landmark event itself when developing a landmark enhancement strategy. These cues are also often relatively low in amplitude and are likely to require some form of enhancement in order for them to be better transmitted with current CI processing. As such, it is recommended that further development of the LBP should explore whether it is beneficial to distinguish between manner categories when labelling landmarks (either through automatic detection or hand-labelling) or whether a simple “onset”/”offset” labelling system should be used to guide enhancement.

Lastly, given the potential disadvantages of boosting the noise signal in addition to the target speech (as discussed in section 6.2.3) it should be considered whether hand labelling of the sentences should have been done using the spectrograms of the noise-mixed stimuli, rather than those for the clean versions of the sentences. In this instance, labels would only be included if the landmark event was clearly visible in the noise-mixed utterance. Hu and Loizou (2008) argued that selecting and stimulating channels dominated by the masker will make it harder for the listener to segregate the target speech signal from the masker. Thus there would be no benefit of boosting a landmark region which is heavily dominated by the background noise. This may also be more representative of the output of an automatic landmark detector when applied to a noise-corrupted signal and should be considered when evaluating the detection accuracy of the algorithm. If using hand-generated labels, these would need to be produced for each sentence in a range of SNRs and for different types of competing maskers.

Figure 6.4 demonstrates how labelling of a sentence (using the same labels used in experiment V) might differ for a sentence in quiet and the same sentence at 5 dB SNR as mixed with either SS noise or eight-talker babble. There appear to be far fewer visible landmarks in the sentence mixed with the SS noise than for the eight-talker babble at the same SNR. In particular, the low-intensity unvoiced fricative segments, such as occur for the sound /f/ in “wife” and /h/ (classed as a fricative for labelling) in “helped” and “husband”, are more easily masked by the SS noise masker. This might suggest that there is a greater chance of benefit to be seen for landmark enhancement in babble noise than for SS noise, and future experiments should explore this. However, it should be noted that although the landmarks are not easily visible in the spectrogram of the SS noise-mixed sentence, the /f/ and /h/ are still perceptible; this is likely the result of phonemic restoration, a phenomenon that was discussed in section 3.2. In the final experiments of this thesis, speech recognition with the LBP was only measured for SS noise; this was due to the limited number of sentence lists available for testing (which in reality would only have been an issue if a larger number of participants had been recruited), and the fact that SS noise has commonly been used in the relevant literature and would therefore allow for easy comparison of results.

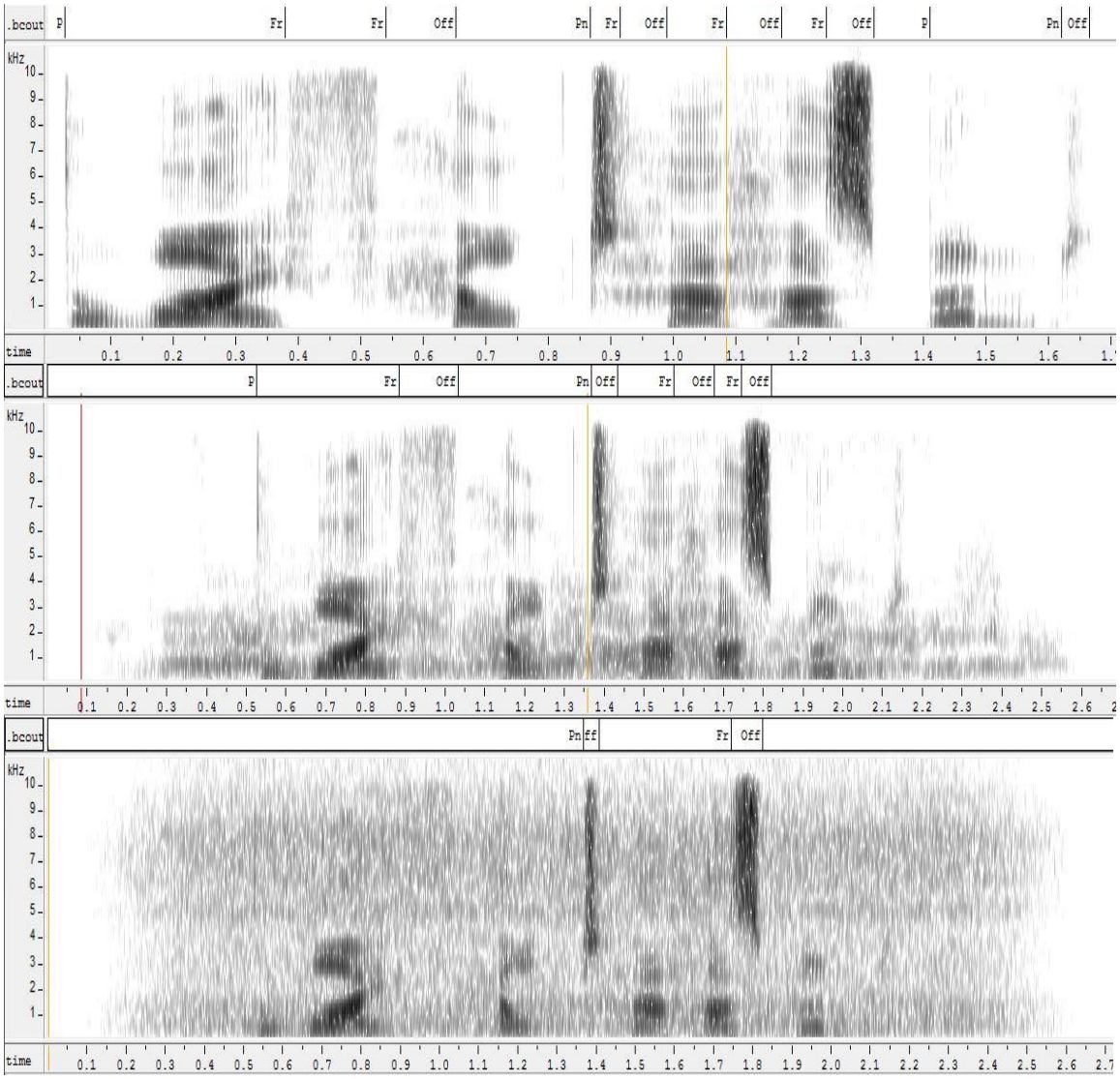


Figure 6. 4 Broadband spectrograms for the BKB sentence "The wife helped her husband" and their corresponding landmark label transcriptions for the sentence in quiet (top panel), the sentence mixed with eight-talker babble at 5 dB SNR (middle panel) and the sentence mixed with SS noise at 5 dB SNR (bottom panel). Note that the noise-mixed sentences have been zero padded at the beginning and end of the recording.

6.3 Role of acoustic landmarks in speech perception

One of the aims of the current thesis was to further explore the role of acoustic landmarks for speech perception in noise. Although the literature (as discussed in chapter 3) argues that sudden amplitude and spectral changes, such as those that occur at obstruent landmarks, are important for helping listeners to distinguish between speech sounds, there is little research that has explored their role for listening to speech in noise. Further to this, studies which have investigated enhancing landmarks or transient events in speech, have shown varying degrees of benefit, and only at negative SNRs (Hazan and Simpson, 1998; 2000; Ortega et al., 2000; Yoo et al., 2007). This might suggest that for normal hearing listeners, acoustic landmarks are

readily accessible in the noise-mixed stimuli for more positive SNRs, or that they only become important when many of the other cues are heavily masked by the noise.

One consideration for why the landmark enhancement strategy evaluated in this thesis did not help to improve speech perception in noise is that the role of seemingly important events has been over-stated or misunderstood. Chapter 3 discussed how ALs signal the presence of new linguistic information in the speech signal and that it is around these events that further processing takes place to uncover important cues to speech sound identity. Therefore, it is possible that landmarks are not important for accurate phoneme/word identification but rather are more important for segmentation. For that reason, it would be interesting to know whether the participants that took part in the experiments in chapter 5 were able to determine word and syllable boundaries better for the enhanced stimuli, compared with the unprocessed stimuli. Unfortunately, the methodology used in these experiments does not allow for this sort of analysis and it is therefore recommended that an appropriate test should be developed in order to test this, both for NH listeners and CI users.

Section 2.6.2 also considered that the enhancement of specific speech elements with CI processing may only be beneficial for softer input levels, for example, 55 dB SPL (Holden et al., 2005; Ali et al., 2014). The present study presented stimuli to the participants at a comfortable and clearly audible level of 65 dB SPL, and therefore any future investigations with LBP should consider presentation levels quieter than this to test this hypothesis.

6.3.1 Transmission of landmarks with current CI processing

In the discussion regarding the limitations of using vocoder simulation studies to predict CI user performance it was considered that ALs do not contribute much to speech perception when spectral resolution is severely restricted. Li and Loizou (2009) found that when NH listeners listening to vocoded speech were presented with clean obstruent consonant information the amount of improvement observed was smaller for the six channel condition than for the 12 and 22 channel condition. The authors concluded that when the SNR is reasonably high and spectral resolution is “reasonable” that listeners are less reliant on ALs, as they have access to a number of other cues (e.g. formant transitions). Results from the present study, however, would perhaps suggest that obstruent landmarks are not accessible when spectral resolution is poor. Section 6.2.1 considered whether increasing spectral resolution helps to improve the transmission of high-frequency obstruent landmarks and figure 6.1 compared the electrode activation patterns for an unprocessed and enhanced

version of the VCV token a/th/a with a 3-*of*-22 and 12-*of*-22 ACE strategy. Increased activation of mid-high frequency channels was visible in the electrodogram for the enhanced speech token in the 12-*of*-22 condition, however, further experimental work is required to determine whether this equates to useable and useful information. It may only be possible to test this with actual CI users, as vocoder simulations using reasonable levels of spectral resolution require listeners to be tested in more difficult SNRs so as to avoid ceiling effects. However, the present study and the study by Li and Loizou (2009) do not consider the effects of channel interaction on the transmission of landmarks, and the potential benefit that might be seen with landmark enhancement, and so future vocoder experiments might be feasible if some degree of spectral smearing is applied in the CI simulation. It would also be interesting to explore the abilities of poorer and better performing CI users to make use of enhanced acoustic landmark information. The participants would need to be recruited with care however, so that results could be analysed for patterns that might be linked to specific factors, for example, dynamic range, duration of deafness before implantation, number of channels available for stimulation etc.

The present thesis has been focused on the specific enhancement of obstruent information only and paid little attention to sonorant or vowel segments. As of yet, there have been no studies which have looked at the transmission of sonorant or vowel landmarks with current CI processing, or considered ways in which to improve CI users' access to them. Although discussions in section 6.2.3 suggested that the abrupt boost applied at the time of obstruent landmarks was appropriate, it is also important not to ignore the importance of contextual information held by the adjacent vowel segments. Section 3.2 explored how acoustic cues for vowels and consonants overlap at their boundaries, particularly for information relating to important formant transitions. It may therefore be appropriate to apply some form of enhancement to the ends and beginnings of adjacent vowel segments, perhaps in a similar way to the ramping method used by Hazan and Simpson (2000) to blend successive segments. However, specific enhancement of landmarks for CI users has so far not shown any improvement in speech recognition scores. Future studies may therefore wish to further explore the potential of altering the compression function, as proposed by Li and Loizou (2010), as a method for improving the transmission of landmarks and the information held at V-C boundaries.

6.4 Amplitude-based n-of-m strategies

The landmark enhancement strategy proposed in this thesis can be considered as a form of channel-specific gain. This is similar to the approach used in strategies such as TESM (Vandali, 2001), ADRO (Blamey et al., 1999) and EECIS (Koning and Wouters, 2012), whereby a boost is applied to the signal after the filterbank, but prior to sampling and selection stage. As has been discussed throughout this chapter, it is possible that these strategies are limited by the fact that the enhancement applied also boosts the noise, and therefore does not always improve the transmission of the information being targeted by the processing. Strategies which focus on minimising the transmission of segments which are dominated by noise, such as IBM and SPARSE, may therefore be a more efficient use of the bandwidth available with current CI devices. Although these noise reduction strategies are not yet fully implementable in CI speech processors, preliminary results suggest that they could provide significant benefit to CI users when listening in noise.

Nonetheless, with the exception of EECIS, all the aforementioned strategies still use an n-of-m channel selection process that selects *n* based on the channels with the highest amplitude. Yoo et al. (2007) demonstrated that the loudest parts of speech are not necessarily the most important for speech perception and therefore n-of-m strategies based on amplitude may restrict the transmission of the more important elements of speech. Channel-specific gain strategies, such as the proposed landmark strategy, aim to get round this by boosting these weaker, yet important, speech cues so that they have a greater chance of being selected in an amplitude-based n-of-m strategy. However, as had already been discussed, this technique has yet to show any improvement in the speech recognition scores of CI users, in noise or in quiet. The question therefore remains: what is the best method for improving the transmission of landmark information with CI processing?

In chapter 4, a new method for selecting *n* was proposed; based on the probability of each channel containing a landmark. However, the channel-specific landmark detectors that were developed were not considered accurate enough and further development of the detectors was not possible in the time-frame of this thesis. It would therefore be interesting to continue this line of research and explore whether it can in fact be used to improve both the transmission of landmarks (and not just obstruent landmarks) and also speech recognition scores in noise for CI users. Stilp et al. (2013) and Stilp (2014) consider that an approach to CI channel selection based on information bearing regions of spectral change (linked to high levels of CSE) may

make better use of current CI processing bandwidth. Both studies demonstrated that NH listeners are able to exploit regions of relative change in the perception of speech, even when the signal has been vocoded; suggesting that CI users may also be able to make use of the spectral change in speech, even though they received a considerably restricted signal. However, so far, research into CSE and cochlear implant processing has only investigated the effects of replacing regions of low versus regions of high CSE with noise for noise-band vocoded sentences in quiet. Further work is still required to gain a better understanding of the ability of CI users to make use of information-bearing regions of spectral change before developing a potential new n-of-m strategy based on a measure of CSE.

6.5 Suggestions for future work and development

Obstruent landmark enhancement using the proposed method does not provide any speech intelligibility benefit when listening to speech in noise as processed through a CI simulation. As CI simulations are often used to predict CI users' performance, the results from this thesis would seem to suggest that there would be little benefit observed with actual CI users. However, a number of methodological limitations have been identified in this chapter that consider whether evaluation of this landmark enhancement strategy using NH listeners listening to vocoded speech is appropriate. Therefore, in order to completely discount the proposed method as a strategy for improving obstruent landmark transmission and by extension, speech intelligibility in noise for CI users, it is recommend that any further developments of the LBP should be tested using either a more realistic vocoder simulation (incorporating channel interaction and a variable DR) or actual CI users. Future development of the LBP may include:

- Investigating whether there is need to differentiate between manner categories in the labelling and subsequent boosting of obstruent consonants
- Development of appropriate boost functions for sonorant and maybe even vowel landmarks
- Determining the most appropriate level of boost to apply at landmark labels
- Investigating the effects of noise type on the potential benefits of landmark enhancement

- Investigating whether landmark enhancement is more beneficial for poorer performing CI users, and more specifically the effects of varying spectral resolution and DR on the transmission of obstruent landmarks

If landmark enhancement using the proposed method was found to be beneficial then further work would also be required to develop automatic landmark detectors that are more noise-robust. However, given that similar channel-specific gain strategies have failed to realise any improvement for speech recognition in noise for CI users, the author of the present thesis would recommend that future research into improving access to ALs for CI users (in the hope of also improving speech recognition) should focus either on the benefit of applying selective compression (as per Li and Loizou, 2010) or the development of a landmark-based n-of-m strategy.

Chapter 7- Conclusions

The main conclusions of this thesis are:

1. The accuracy of existing automatic landmark detection algorithms is not yet satisfactory for use in a CI speech processor.

2. It is possible to enhance obstruent landmarks using the proposed method; however, it is not yet clear whether information transmitted by these enhanced regions is perceptible or useful for CI users.

3. The role of acoustic landmarks for speech perception in noise is still not yet fully understood. Their role may be dependent on spectral resolution and perhaps only help listeners to segment the speech signal, rather than increasing cue transmission.

4. Enhancement strategies that apply a boost to the signal prior to channel selection with an n-*of*-m strategy have not yet proven successful for improving speech perception scores, in quiet or in noise.

5. Future work into improving the transmission of landmark information with CIs should re-consider developing a new n-*of*-m approach that selects channels based on the probability of them containing information relating to an acoustic landmark.

Appendices

Appendix 1

The BKB sentence lists. Key words are in bold and underlined.

List 1

1. The **clown** had a **funny** **face**
2. The **car** **engines** **running**
3. **She** **cut** with her **knife**
4. **Children** **like** **strawberries**
5. The **house** had **nine** **rooms**
6. **They're** **buying** some **bread**
7. The **green** **tomatoes** are **small**
8. **He** **played** with his **train**
9. The **postman** **shut** the **gate**
10. **They're** **looking** **at** the **clock**
11. The **bag** **bumps** on the **ground**
12. The **boy** **did** a **handstand**
13. A **cat** **sits** **on** the **bed**
14. The **lorry** **carried** **fruit**
15. The **rain** **came** **down**
16. The **ice** **cream** was **pink**

List 2

1. The **ladders** **near** the **door**
2. **They** had a **lovely** **day**
3. The **ball** **went** **into** the **goal**
4. The **old** **gloves** are **dirty**
5. **He** **cut** his **finger**
6. The **thin** **dog** was **hungry**
7. The **boy** **knew** the **game**
8. **Snow** **falls** at **Christmas**
9. **She's** **taking** her **coat**
10. The **police** **chased** the **car**
11. A **mouse** **ran** **down** the **hole**
12. The **lady's** **making** a **toy**
13. Some **sticks** were **under** the **tree**
14. The **little** **baby** **sleeps**
15. **They're** **watching** the **train**
16. The **school** **finished** **early**

List 3

1. The **glass** **bowl** **broke**
2. The **dog** **played** with a **stick**
3. The **kettles** **quite** **hot**
4. The **farmer** **keeps** a **bull**
5. **They** **say** some **silly** **things**
6. The **lady** **wore** a **coat**
7. The **children** are **walking** **home**
8. **He** **needed** his **holiday**
9. The **milk** **came** in a **bottle**
10. The **man** **cleaned** his **shoes**
11. **They** **ate** the **lemon** **jelly**
12. The **boys** **running** **away**
13. **Father** **looked** at the **book**
14. **She** **drinks** from her **cup**
15. The **rooms** **getting** **cold**
16. A **girl** **kicked** the **table**

List 4

1. The **wife** **helped** her **husband**
2. The **machine** was **quite** **noisy**
3. The **old** **man** **worries**
4. A **boy** **ran** down the **path**
5. The **house** had a **nice** **garden**
6. **She** **spoke** **to** her **son**
7. **They're** **crossing** the **street**
8. **Lemons** **grow** on **trees**
9. **He** **found** his **brother**
10. Some **animals** **sleep** **on** **straw**
11. The **jam** **jar** was **full**
12. **They're** **kneeling** **down**
13. The **girl** **lost** her **doll**
14. The **cooks** **making** a **cake**
15. The **child** **grabs** the **toy**
16. The **mud** **stuck** on his **shoe**

List 5

1. The **bath** **towel** was **wet**
2. The **matches** **lie** on the **shelf**
3. **They're** **running** **past** the **house**
4. The **train** had a **bad** **crash**
5. The **kitchen** **sinks** **empty**
6. A **boy** **fell** from the **window**
7. **She** **used** her **spoon**
8. The **park's** **near** the **road**
9. The **cook** **cut** some **onions**
10. The **dog** **made** an **angry** **noise**
11. **He's** **washing** his **face**
12. **Somebody** **took** the **money**
13. The **light** **went** **out**
14. **They** **wanted** some **potatoes**
15. The **naughty** **girl's** **shouting**
16. The **cold** **milks** in a **jug**

List 6

1. The **paint dripped** on the **ground**
2. The **mother stirs** the **tea**
3. **They laughed** at his **story**
4. **Men wear long trousers**
5. The **small boy** was **asleep**
6. The **lady goes to** the **shop**
7. The **sun melted** the **snow**
8. The **fathers coming home**
9. **She** had her **pocket money**
10. The **lorry drove** up the **road**
11. **He's bringing** his **raincoat**
12. A **sharp knifes dangerous**
13. **They took** some **food**
14. The **clever girls** are **reading**
15. The **broom stood** in the **corner**
16. The **woman tidied** her **house**

List 7

1. The **children dropped** the **bag**
2. The **dog came back**
3. The **floor looked clean**
4. **She found** her **purse**
5. The **fruit lies** on the **ground**
6. **Mother fetches** a **saucepan**
7. **They washed** in **cold water**
8. The **young people** are **dancing**
9. The **bus went early**
10. **They** had **two empty bottles**
11. A **balls bouncing along**
12. The **father forgot** the **bread**
13. The **girl** has a **picture book**
14. The **orange** was **quite sweet**
15. **He's holding** his **nose**
16. The **new** r**oad's** on the **map**

List 8

1. The **boy forgot** his **book**
2. A **friend came** for **lunch**
3. The **match boxes** are **empty**
4. **He climbed** his **ladder**
5. The **family bought** a **house**
6. The **jug stood** on the **shelf**
7. The **ball broke** the **window**
8. **They're shopping** for **cheese**
9. The **pond waters dirty**
10. **They heard** a **funny noise**
11. **Police** are **clearing** the **road**
12. The **bus stopped suddenly**
13. **She writes** to her **brother**
14. The **footballer lost** a **boot**
15. The **three girls** are **listening**
16. The **coat lies on** a **chair**

List 9

1. The **book tells** a **story**
2. The **young boy left home**
3. **They're climbing** the **tree**
4. **She stood** near her **window**
5. The **table** has **three legs**
6. A **letter fell** on the **mat**
7. The **five men** are **working**
8. **He listens to** his **father**
9. The **shoes** were **very dirty**
10. **They went** on **holiday**
11. **Baby broke** his **mug**
12. The **lady packed** her **bag**
13. The **dinner plates hot**
14. The **train's moving fast**
15. The **child drank** some **milk**
16. The **car hit** a **wall**

List 10

1. A **tea towel's** by the **sink**
2. The **cleaner used** a **broom**
3. **She looked in** her **mirror**
4. The **good boys helping**
5. **They followed** the **path**
6. The **kitchen clock** was **wrong**
7. The **dog jumped on** the **chair**
8. **Someone's crossing** the **road**
9. The **postman brings** a **letter**
10. **They're cycling along**
11. **He broke** his **leg**
12. The **milk** was by the **front door**
13. The **shirts hang** in the **cupboard**
14. The **ground** was **too hard**
15. The **buckets hold water**
16. The **chicken laid** some **eggs**

List 11

1. The **sweet shop** was **empty**
2. The **dogs go** for a **walk**
3. **She's washing** her **dress**
4. The **lady stayed** for **tea**
5. The **driver waits** by the **corner**
6. **They finished** the **dinner**
7. The **policeman knows** the **way**
8. The **little girl** was **happy**
9. **He wore** his **yellow shirt**
10. **They're coming** for **Christmas**
11. The **cow gave** some **milk**
12. The **boy got** =into **bed**
13. The **two farmers** are **talking**
14. **Mother picked** some **flowers**
15. A **fish lay** on the **plate**
16. The **father writes** a **letter**

List 12

1. The **food cost** a **lot**
2. The **girls washing** her **hair**
3. The **front garden** was **pretty**
4. **He lost** his **hat**
5. The **taps** are **above** the **sink**
6. **Father paid at** the **gate**
7. **She's waiting** for her **bus**
8. The **bread vans coming**
9. **They** had some **cold meat**
10. The **football games over**
11. **They carry** some **shopping bags**
12. The **children help** the **milkman**
13. The **picture came** from a **book**
14. The **rice pudding** was **ready**
15. The **boy** had a **toy dragon**
16. A **tree fell** on the **house**

List 13

1. The **fruit came** in a **box**
2. The **husband brings** some **flowers**
3. **They're playing** in the **park**
4. **She argued** with her **sister**
5. A **man told** the **police**
6. P**otatoes** g**row** in the **ground**
7. **He's cleaning** his **car**
8. The **mouse found** the **cheese**
9. **They waited** for **one hour**
10. The **big dog** was **dangerous**
11. The **strawberry jam** was **sweet**
12. The **plant hangs above** the **door**
13. The **children** are **all eating**
14. The **boy** has **black hair**
15. The **mother heard** her **baby**
16. The **lorry climbed** the **hill**

List 14

1. The **angry man shouted**
2. The **dog sleeps** in a **basket**
3. **They're drinking tea**
4. **Mother opens** the **drawer**
5. An **old woman** was at **home**
6. **He dropped** his **money**
7. **They broke all** the **eggs**
8. The **kitchen window** was **clean**
9. The **girl plays** with the **baby**
10. The **big fish got away**
11. **She's helping** her **friend**
12. The **children washed** the **plates**
13. The **postman comes early**
14. The **sign showed** the **way**
15. The **grass** is **getting long**
16. The **match fell** on the **floor**

List 15

1. A **man's turning** the **tap**
2. The **fire** was **very hot**
3. **He's sucking** his **thumb**
4. The **shop closed** for **lunch**
5. The **driver starts** the **engine**
6. The **boy hurried** to **school**
7. Some **nice people** are **coming**
8. **She bumped** her **head**
9. **They met some friends**
10. **Flowers grow** in the **garden**
11. The **tiny baby** was **pretty**
12. The **daughter laid** the **table**
13. **They walked across** the **grass**
14. The **mother tied** the **string**
15. The **train stops** at the **station**
16. The **puppy plays** with a **ball**

List 16

1. The **children wave** at the **train**
2. **Mother cut** the **Christmas cake**
3. **He closed** his **eyes**
4. The **raincoats very wet**
5. A **lady buys** some **butter**
6. **They called** an **ambulance**
7. **She's paying** for her **bread**
8. The **policeman found** a **dog**
9. Some **men shave** in the **morning**
10. The **driver lost** his **way**
11. **They stared** at the **picture**
12. The **cat drank** from a **saucer**
13. The **oven door** was **open**
14. The **cars going too fast**
15. The **silly boys hiding**
16. The **painter used** a **brush**

List 17

1. The **apple pies cooking**
2. **He drinks** from his **mug**
3. The s**ky** was **very blue**
4. **They knocked** on the **window**
5. The **big boy kicked** the **ball**
6. **People** are **going home**
7. The **baby wants** his **bottle**
8. The **lady sat** on her **chair**
9. **They** had some **jam pudding**
10. The **scissors** are **quite sharp**
11. **She's calling** her **daughter**
12. Some **brown leaves fell** off the **tree**
13. The **milkman carried** the **cream**
14. A **girl ran along**
15. The **mother reads** a **paper**
16. The **dog chased** the **cat**

List 18

1. The **cake shops opening**
2. **They like orange marmalade**
3. The **mother shut** the **window**
4. **He's skating with** his **friend**
5. The **cheese pie** was **good**
6. **Rain falls** from **clouds**
7. **She talked** to her **doll**
8. **They painted** the **wall**
9. The **towel dropped** on the **floor**
10. The **dogs eating** some **meat**
11. A **boy broke** the **fence**
12. The **yellow pears** were **lovely**
13. The **police help** the **driver**
14. The **snow lay** on the **roof**
15. The **lady washed** the **shirt**
16. The **cup hangs** on a **hook**

List 19

1. The **family like fish**
2. **Sugars very sweet**
3. The **baby lay** on a **rug**
4. The **washing machine broke**
5. **They're clearing** the **table**
6. The **cleaner swept** the **floor**
7. A **grocer sells butter**
8. The **bath water** was **warm**
9. **He's reaching** for his **spoon**
10. **She hurt** her **hand**
11. The **milkman drives** a **small van**
12. The **boy slipped on** the **stairs**
13. **They're staying** for **supper**
14. The **girl held** a **mirror**
15. The **cup stood** on a **saucer**
16. The **cows went** to **market**

List 20

1. The **boy got** into **trouble**
2. **They're going out**
3. The **football hit** the **goalpost**
4. **He paid** his **bill**
5. The **teacloths quite wet**
6. A **cat jumped off** the **fence**
7. The **baby** has **blue eyes**
8. **They sat** on a **wooden bench**
9. **Mother made** some **curtains**
10. The **ovens** too **hot**
11. The **girl caught** a **cold**
12. The **raincoats hanging up**
13. **She brushed** her **hair**
14. The **two children** are **laughing**
15. The **man tied** his **scarf**
16. The **flower stands** in a **pot**

List 21

1. The **pepper pot** was **empty**
2. The **dog drank** from a **bowl**
3. A **girl came** into the **room**
4. **They're pushing** an **old car**
5. The **cat caught** a **mouse**
6. The **road goes up** a **hill**
7. **She made** her **bed**
8. **Bananas** are **yellow fruit**
9. The **cow lies** on the **grass**
10. The **egg cups** are on the **table**
11. **He frightened** his **sister**
12. The **cricket teams playing**
13. The **father picked** some **pears**
14. The **kettle boiled quickly**
15. The **man's painting** a **sign**
16. **They lost** some **money**

Appendix 2

The IHR sentence lists. Key words are in bold and underlined.

List 1

1. **They moved** the **furniture**
2. **He's wiping** the **table**
3. **He hit** his **head**
4. The **yellow leaves** are **falling**
5. The **cat played** with some **wool**
6. The **bag** was **very heavy**
7. The **towel dripped** on the **carpet**
8. The **bull chased** the **lady**
9. The **man dug** his **garden**
10. The **room** has a **lovely view**
11. The **girl helped** in the **kitchen**
12. The **old shoes** were **muddy**
13. **Father's hiding** the **presents**
14. The **milk boiled over**
15. The **neighbour knocked** at the **door**

List 2

1. **He tore** his **shirt**
2. **They finished** the **jigsaw**
3. **She brought** her **camera**
4. The **lady watered** her **plants**
5. The **salt cellars full**
6. The **boy hit** his **thumb**
7. The **mother shook** her **head**
8. The **snow lay** on the **hills**
9. The **father used** a **towel**
10. The **tree** was in the **back garden**
11. The **yacht sailed past**
12. The **lady pushed** the **pram**
13. **They're leaving today**
14. The **picture hung** on the **wall**
15. The **children sit** under the **tree**

List 3

1. The **lunch** was **very early**
2. The **dirty boy** is **washing**
3. **He hid** his **money**
4. The **curtains** were **too short**
5. The **knife cut** the **cake**
6. **They emptied** their **pockets**
7. The **new shoes** were **tight**
8. The **coat hangs** in a **cupboard**
9. The **sun shone** through the **clouds**
10. **She took** her **purse**
11. The **team lost** the **match**
12. The **shirt caught** on a **nail**
13. **They picked** some **raspberries**
14. The **man climbed** the **mountain**
15. The **lady hurt** her **arm**

List 4

1. The **old clothes** were **dirty**
2. **He carried** a **stick**
3. **She read** her **book**
4. The **new house** was **empty**
5. The **thief brought** a **ladder**
6. The **horse stands** by the **gate**
7. **They're heading** for the **park**
8. The **gardener trimmed** the **hedge**
9. **They're standing up**
10. **Someone's hiding** in the **bushes**
11. The **waiter lit** the **candles**
12. The **baker iced** the **cake**
13. The **woman slipped** on the **ice**
14. The **small puppy** was **scared**
15. The **lady changed** her **mind**

List 5

1. The **daughter closed** the **box**
2. **He broke** into the **safe**
3. The **doctor carries** a **bag**
4. The **new game** was **silly**
5. The **little boy** was **tired**
6. **They saw** the **sign**
7. **She's wrapping** the **parcel**
8. The **children laughed** at the **clown**
9. The **apple pie** was **hot**
10. The **ship sailed** up the **river**
11. The **house** had a **lovely garden**
12. The **noisy dog** is **barking**
13. **They bought** some **tickets**
14. The **man goes** to the **bank**
15. The **nurse helped** the **child**

List 6

1. The **girl knew** the **story**
2. **He reached** for a **cup**
3. They **lady** was **quite cross**
4. The **rope** was t**oo short**
5. **She's listening** to the **radio**
6. The **husband cleaned** the **car**
7. The **postman leaned** on the **fence**
8. The **china vase** was **broken**
9. The **other team won**
10. **They locked** the **safe**
11. The **leaves dropped** from the **trees**
12. The **men watched** the **race**
13. The **birds building** a **nest**
14. The **woman called** her **dog**
15. **They're waving** at the **train**

List 7

1. The **cat** **scratched** the **chair**
2. **She** **tapped** at the **window**
3. The **man** **painted** the **gate**
4. **He** **slid** on the **floor**
5. **They're** **lifting** the **box**
6. The **woman** **listened** to her **friend**
7. The **driver** **hooted** his **horn**
8. The **cake** **tasted** **nice**
9. The **sailor** **stood** on the **deck**
10. The **young** **girls** were **pretty**
11. **They** **painted** the **ceiling**
12. The **back** **door** was **shut**
13. The **tree** **lost** its **leaves**
14. The **boy** **eats** with his **fork**
15. The **young** **mother's** **shopping**

List 8

1. The **girl** **sharpened** her **pencil**
2. **She** **closed** her **eyes**
3. The **puppy** **licked** his **master**
4. The **plant** **grows** on the **wall**
5. The **family's** **having** a **picnic**
6. The **train** **arrived** on **time**
7. **They** **won** the **game**
8. The **lady** **waited** for her **husband**
9. The **post** **office** was **near**
10. **They** **rowed** the **boat**
11. The **old** **fox** was **sly**
12. The **baby** **lost** his **rattle**
13. **He** **dug** with his **spade**
14. The **boiled** **egg** was **soft**
15. The **two** **ladies** were **watching**

List 9

1. The **car** **engine's** **running**
2. **They** **parked** by the **station**
3. The **lemons** were **quite** **bitter**
4. **They're** **cutting** the **grass**
5. The **woman** **called** a **doctor**
6. The **man** **shaved** with a **razor**
7. **He** **tied** his **shoelaces**
8. The **bus** is **leaving** **early**
9. **She's** **sewing** on a **button**
10. The **horse** **kicked** the **rider**
11. The **yellow** **bananas** are **ripe**
12. The **lady** has a **fur** **coat**
13. The **cat** **jumped** onto the **table**
14. The **book** **sits** on the **shelf**
15. The **boy** t**old** a **joke**

List 10

1. **She** **sings** in the **bath**
2. The **meat** was **too** **tough**
3. The **child** **ate** some **jam**
4. **They're** **stealing** the **apples**
5. The **children** **dried** the **dishes**
6. The **paper** **boy** was **cheeky**
7. The **little** **car** was **slow**
8. The **bath** **taps** are **dripping**
9. **They** **came** at **Easter**
10. **He's** **wearing** a **tie**
11. The **new** **towel** was **clean**
12. The **water** **poured** from a **jug**
13. The **red** **apples** were in a **bowl**
14. The **bus** **stopped** at the **shops**
15. The **man** **drew** with a **pencil**

List 11

1. The **lady** **cut** her **finger**
2. The **horses** **stood** under the **tree**
3. **Mother's** **talking** to the **milkman**
4. **She** **polished** her **shoes**
5. Some **friends** **stayed** for **supper**
6. The **pudding** was **very** **good**
7. The **apples** **came** in a **bag**
8. The **greedy** **boy** was **hungry**
9. The **three** **men** were **angry**
10. The **children** **cleared** the **table**
11. The **man** **forgot** his **change**
12. The **clothes** are **covered** in **mud**
13. The **raincoat's** **wet** **through**
14. The **three** **friends** are **cycling**
15. **They** **tied** the **rope**

List 12

1. **Christmas** is **coming** **soon**
2. The **tall** **man** was **thin**
3. The **girl** **broke** a **vase**
4. The **other** **team** are **losing**
5. The **girl's** **playing** **tennis**
6. They **lady** **spoke** to the **driver**
7. The **noise** **scared** the **sheep**
8. **She's** **sitting** on the **swing**
9. The **red** **ball's** **bouncing**
10. The **children** **heard** the **doorbell**
11. **They** **worked** in the **rain**
12. The **children** **carried** the **suitcase**
13. The **new** **teacher's** **nice**
14. The **traffic** **lights** are **green**
15. **They're** **going** to the **seaside**

List 13

1. The **story's very exciting**
2. **He's kicking** the **door**
3. The **pool** was **very deep**
4. **Mother served** the **soup**
5. The **woman used** her **key**
6. The **red dress** was **pretty**
7. The **pears** were **too hard**
8. **He turned** on the **taps**
9. **She tore** her **dress**
10. **Mother's filling** the **kettle**
11. The **lady writes** to her **sister**
12. **They're looking** at the **clock**
13. The **farmer's buying** some **pigs**
14. The **old man** is **leaving**
15. The **boy ate** his **lunch**

List 14

1. The **chocolate box** was **empty**
2. The **boy filled** the **buckets**
3. The **lorry drove** up the **hill**
4. **They called** the **police**
5. The **lady wore** her **coat**
6. The **policeman stopped** the **traffic**
7. The **dog heard** a **noise**
8. The **rose bush** was **blooming**
9. The **cows grazed** in the **field**
10. The **sun came out**
11. **He's starting** the **engine**
12. **He's visiting** his **uncle**
13. The **jam sticks** to the **plate**
14. **They're bringing** some **pears**
15. The **garden's very neat**

List 15

1. The **red bus** was **late**
2. **They're leaning** on the **ladder**
3. The **ice cream** is **melting**
4. The **green apples** were **sour**
5. The **family ate supper**
6. The **horse** was **quite old**
7. The **towel's quite dry**
8. The **birds sang** from the **tree**
9. The **lion escaped** from the **zoo**
10. The **man took** a **picture**
11. **They're buying** some **lunch**
12. **He's combing** his **hair**
13. **They watched** the **sunset**
14. **Someone's listening** at the **door**
15. The **girl told** her **mother**

List 16

1. **They did** their **homework**
2. The **sister hurt** her **leg**
3. The **two boys** are **laughing**
4. The **big needle** was **sharp**
5. The **hill** was **very steep**
6. **Roses grow** in the **garden**
7. The **puppy chased** the **ball**
8. The **father answered** the **door**
9. The **friends came** for **tea**
10. The **old man** was **poor**
11. **They ate some plums**
12. The **mice saw** the **trap**
13. The **car hit** the **tree**
14. **They're hanging** up their **coats**
15. **She's holding** a **brush**

List 17

1. The **man climbed** the **ladder**
2. **Mother** c**ooked** the **dinner**
3. The **ice rink** was **closed**
4. The **ice cream** was **cold**
5. **They closed** the **curtains**
6. The **doctor came quickly**
7. The **ice melted** in the **sun**
8. **They ran** up the **hill**
9. The **milk** was **very cold**
10. The **father si**gned a **letter**
11. **He remembered** the **way**
12. The **loud noise** was **sudden**
13. The **boy grazed** his **knee**
14. The **lady** has **small feet**
15. **They helped** with the **dishes**

List 18

1. **They watched** the **cricket**
2. The **girl carried** a **basket**
3. The **baby sleeps** in a **cot**
4. The **daughter drank** some **lemonade**
5. The **little girl** was **staring**
6. **They're living** by the **sea**
7. The **church stood** on the **hill**
8. The **farmer sowed** some **seeds**
9. The **east wind** was **cold**
10. The **blue towel** was **damp**
11. The **man forgot** his **hat**
12. **Mother's reading** a **story**
13. **She ironed** her **skirt**
14. The **floor** was **quite slippery**
15. The **biscuit tin** was **empty**

Appendix 3

Praat script file used for the presentation and scoring of the BKB sentences

"ooTextFile"
"ExperimentMFC 2"
fileNameHead = ""
fileNameTail = ".wav"
carrierBefore = ""
carrierAfter = ""
initialSilenceDuration = 0.5 seconds
interStimulusInterval = 0.5 seconds
numberOfDifferentStimuli = 16
"1"
"2"
"3"
"4"
"5"
"6"
"7"
"8"
"9"
"10"
"11"
"12"
"13"
"14"
"15"
"16"
numberofReplicationsperStimulus = 1
breakafterEvery = 300 seconds
randomize = <CyclicNonRandom>
startText= "Touch here to start."
runText = ""
pauseText= ""
endText = "Well done - the condition is finished."
numberOfResponseCategories = 5
0.3 0.7 0.85 1.0 "No key words correct" "0"
0.3 0.7 0.65 0.8 "One key word correct" "1"
0.3 0.7 0.45 0.6 "Two key words correct" "2"
0.3 0.7 0.25 0.4 "Three key words correct" "3"
0.3 0.7 0.0 0.2 "Four key words correct" "4"
numberOfGoodnessCategories = 0

Appendix 4

Praat script file used for the presentation and scoring of VCV stimuli (example is for the /aCa/ condition).

"ooTextFile"

"ExperimentMFC 2"

fileNameHead = ""

fileNameTail = ".wav"

carrierBefore = ""

carrierAfter = ""

initialSilenceDuration = 0.5 seconds

interStimulusInterval = 0.5 seconds

numberOfDifferentStimuli = 40

"ABA1"

"ABA2"

"ACHA1"

"ACHA2"

"ADA1"

"ADA2"

"AFA1"

"AFA2"

"AGA1"

"AGA2"

"ADJA1"

"ADJA2"

"AKA1"

"AKA2"

"ALA1"

"ALA2"

"AMA1"

"AMA2"

"ANA1"

"ANA2"

"APA1"

"APA2"

"ARA1"

"ARA2"

"ASA1"

"ASA2"

"ASHA1"

"ASHA2"

"ATA1"

"ATA2"

"ATHA1"

"ATHA2"

"AVA1"

"AVA2"

"AWA1"

"AWA2"

"AYA1"

"AYA2"

"AZA1"

"AZA2"

numberofReplicationsperStimulus = 3

breakafterEvery = 300 seconds

randomize = <PermuteBalancedNoDoublets>

startText= "If you are ready please click to start."

runText = ""

pauseText= ""

endText = "Well done- you've completed the condition."

numberOfResponseCategories = 20

0.0 0.2 0.75 1.0 "b" "b"

0.2 0.4 0.75 1.0 "ch" "ch"

0.4 0.6 0.75 1.0 "d" "d"

0.6 0.8 0.75 1.0 "f" "f"

0.8 1.0 0.75 1.0 "g" "g"

0.0 0.2 0.5 0.75 "j" "dj"

0.2 0.4 0.5 0.75 "k" "k"

0.4 0.6 0.5 0.75 "l" "l"

0.6 0.8 0.5 0.75 "m" "m"

0.8 1.0 0.5 0.75 "n" "n"

0.0 0.2 0.25 0.5 "p" "p"

0.2 0.4 0.25 0.5 "r" "r"

0.4 0.6 0.25 0.5 "s" "s"

0.6 0.8 0.25 0.5 "sh" "sh"

0.8 1.0 0.25 0.5 "t" "t"

0.0 0.2 0.0 0.25 "th" "th"

0.2 0.4 0.0 0.25 "v" "v"

0.4 0.6 0.0 0.25 "w" "w"

0.6 0.8 0.0 0.25 "y" "y"

0.8 1.0 0.0 0.25 "z" "z"

numberOfGoodnessCategories = 0

Appendix 5

Participant questionnaire (*Version1*)

Study title: Obstruent landmark enhancement for cochlear implant users: A simulation study

Researcher name: Cherith Webb
Study reference:
Ethics reference:

Subject identification number:

Subject ear to be tested: L R

Subject details-

Gender: M F

Age:

Are you a native English speaker? Y N

Have you ever participated in a cochlear implant simulation experiment before?

Y N

Please give details if you suffer from and/or have recently received treatment, or are currently undergoing treatment, for any of the conditions listed below:

Troublesome tinnitus: Y N

Current ear disease (e.g. persistent ear pain, ear infection or ear discharge): Y N

Operations on your ears: Y N

Difficulty hearing in background noise: Y N

Have you been exposed to excessive loud noise in the past 24 hours? Y N

Are there any other factors that may prevent you from carrying out this experiment? Please give details in the space below.

Appendix 6

CONSENT FORM (*Version1*)

Study title: Obstruent landmark enhancement for cochlear implant users: A simulation study

Researcher name: Cherith Webb
Study reference:
Ethics reference:

Please initial the box(es) if you agree with the statement(s):

I have read and understood the information sheet (insert date /version no. of participant information sheet) and have had the opportunity to ask questions about the study.	☐
I agree to take part in this research project and agree for my data to be used for the purpose of this study	☐
I understand my participation is voluntary and I may withdraw at any time without my legal rights being affected	☐
I am happy to be contacted regarding other unspecified research projects. I therefore consent to the University retaining my personal details on a database, kept separately from the research data detailed above. The 'validity' of my consent is conditional upon the University complying with the Data Protection Act and I understand that I can request my details be removed from this	☐

Data Protection
I understand that information collected about me during my participation in this study will be stored on a password protected computer and that this information will only be used for the purpose of this study. All files containing any personal data will be made anonymous.

Name of participant (print name)……

Signature of participant……..

Date……

Appendix 7

Feature matrix used to compute consonant confusion matrices for the VCV experiments outlined in Chapter 5.

	m	b	p	v	f	n	z	d	t	s	l	r	sh	ch	k	g	th	w	dj	y
voic	+	+	-	+	-	+	+	+	-	-	+	+	-	-	-	+	-	+	+	+
fric	0	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0	1	0	1	0
nasl	y	n	n	n	n	y	n	n	n	n	n	n	n	n	n	n	n	n	n	n
plsv	no	yes	yes	no	no	no	no	yes	yes	no	no	no	no	yes	yes	yes	no	no	yes	no
place	bil	bil	bil	lad	lad	alv	alv	alv	alv	alv	alv	ret	ret	ret	vel	vel	den	bil	ret	alv
manner	nasl	plsv	plsv	fric	fric	nasl	fric	plsv	plsv	fric	con	con	fric	aff	plsv	psl v	fric	con	aff	con

Appendix 8

MATLAB participant batch processing script for the final version (1.5) of the LBP as used in experiment V.

```
% This script allows for the use of landmark processing in a
% batch processing framework

% ~~~~~~~~~~ Written by ~~~~~~~~~~~~ %
% author: Falk-Martin Hoffmann       %
% email: falk-martin.hoffmann@rub.de %
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
% VERSION 1.5

clear all;
parent = cd;
p1 = genpath(pwd);
addpath(p1);
cd ..;
p2 = genpath([cd,'/NMT_Matlab']);
addpath(p2);

% *** Define participant and processing *** %
P_name = 'BKB_clean_12_channels'; % participant's name
SNR_cond = 5; % desired SNR-condition in dB
n_type = 'babble'; % desired noise type ('babble','speech-shaped' or
'modulated')
lm_boost = 15; % boost of the landmarks in dB
channels = 12; % number of channels to be excited, default = 12
strategy = 1; % set to 1 if landmark processing should be applied to clean
files, else set to 0
% *** END section *** %

% *** Define database *** %
database = 'BKBQ'; % choose either BKBQ, ASLQ (IHR-Sentence Lists) or VCV
gender = 'MALE'; % choose either 'FEMALE' or 'MALE', if you are using the
VCV files
first_list_no = 4; % number of the first list to be processed
last_list_no = 4; % number of the last list to be processed
sents = 3; % number of sentences in each list
% *** END section *** %

% *** Define output flags *** %
wb_unprocessed_out = 1; % set 1, if unprocessed non-vocoder sound files
shall be written, else 0
wb_processed_out = 1; % set 1, if processed non-vocoder sound files shall
be written, else 0
voc_original_out = 1; % set 1, if unprocessed vocoder sound files shall be
written, else 0
voc_processed_out = 1; % set 1, if processed vocoder sound files shall be
written, else 0
% *** END section *** %

% *** Define boost and algorithm parameters *** %
% framesize in seconds
framesize = 0.003; % formerly 12ms
% frameshift in seconds
frameshift = 0.0015; % formerly 6ms
% lower edge frequency for Plosives
```

```
leb_P = 1000;
% higher edge frequency for Plosives
heb_P = 8000;
% lower edge frequency for long Plosives
leb_Pn = 1000;
% higher edge frequency for long Plosives
heb_Pn = 8000;
% lower edge frequency for Fricatives
leb_Fr = 1000;
% higher edge frequency for Fricatives
heb_Fr = 8000;
% lower edge frequency for Offsets
leb_Off = 2500;
% higher edge frequency for Offsets
heb_Off = 8000;

% preparing parameter struct for TF-analysis
par = struct('framesize',framesize,...
             'frameshift',frameshift,...
             'strategy',strategy,...
             'leb_P_f',leb_P,...
             'heb_P_f',heb_P,...
             'leb_Pn_f',leb_Pn,...
             'heb_Pn_f',heb_Pn,...
             'leb_Fr_f',leb_Fr,...
             'heb_Fr_f',heb_Fr,...
             'leb_Off_f',leb_Off,...
             'heb_Off_f',heb_Off);
clear framesize frameshift;
% *** END Section *** %

P_dir = [P_name,'_SNR=',num2str(SNR_cond),'dB'];
c_fold = pwd; % capture current folder
mkdir(c_fold,P_dir); % create participant's individual directory

for v = first_list_no:last_list_no % for all lists
    for w = 1:sents % for all sentences
        switch database
            case {'BKBQ','ASLQ'} % sentence data
                if (v<10)
                    if (w<10)
                        filename =
[database,'0',num2str(v),'0',num2str(w),'.wav'];
                    else
                        filename =
[database,'0',num2str(v),num2str(w),'.wav'];
                    end
                else
                    if (w<10)
                        filename =
[database,num2str(v),'0',num2str(w),'.wav'];
                    else
                        filename = [database,num2str(v),num2str(w),'.wav'];
                    end
                end

                % read landmarks
                cd([database,'-Landmarks']);
                landmark_data = ReadLandmarks([filename,'.bcout']);
                cd ..;
```

```
                % read from database folder
                cd(['equalRMS_',database]);
                [out,original,param,COGs] =
ProcessLandmarks2(par,filename,landmark_data,SNR_cond,n_type,lm_boost);
                cd ..;
            case 'VCV' % VCV data
                VCVs =
{'ABA','ACHA','ADA','ADJA','AFA','AGA','AKA','ALA','AMA','ANA','APA',...

'ARA','ASA','ASHA','ATA','ATHA','AVA','AWA','AYA','AZA',...

'IBI','ICHI','IDI','IDJI','IFI','IGI','IKI','ILI','IMI','INI','IPI',...

'IRI','ISHI','ISI','ITHI','ITI','IVI','IWI','IYI','IZI',...

'UBU','UCHU','UDJU','UDU','UFU','UGU','UKU','ULU','UMU','UNU','UPU',...

'URU','USHU','USU','UTHU','UTU','UVU','UWU','UYU','UZU'};
                filename = [VCVs{v},num2str(w),'.wav'];

                % read landmarks
                cd([database,'-Landmarks']);
                landmark_data = ReadLandmarks([filename,'.bcout']);
                cd ..;
                % read from database folder
                cd(['equalRMS_',database]);
                cd(gender);
                [out,original,param,COGs] =
ProcessLandmarks2(par,filename,landmark_data,SNR_cond,n_type,lm_boost);
                cd ..;
                cd ..;
        end

        cd([c_fold,'/',P_dir]);

        if wb_unprocessed_out
            % write unprocessed output wav-file

wavwrite(original,param.fs,['SNR_',num2str(SNR_cond),'dB_',filename]);
        end

        if wb_processed_out
            % write processed output wav-file

wavwrite(out,param.fs,['new_SNR_',num2str(SNR_cond),'dB_',filename]);
        end

        if voc_original_out
            % prepare unprocessed signal for vocoder by disposal of the DC
component
            buffer = original - mean(original);
            % normalise 'buffer' to RMS = 1 and produce vocoder signal
            [Voc,OriginalSequence] =
VocoderOut(buffer/std(buffer),'noise',channels);

            % write wav-files containing vocoder signal
%
wavwrite(Voc/max(abs(Voc)),param.fs,['Voc_SNR_',num2str(SNR_cond),'dB_',fil
ename]);
```

```
wavwrite(Voc,param.fs,['Voc_SNR_',num2str(SNR_cond),'dB_',filename]);
        end

        if voc_processed_out
            % prepare processed signal for vocoder by disposal of the DC
component
            buffer = out - mean(out);
            % normalise 'buffer' to RMS = 1 and produce vocoder signal
            [VocNew,ProcessedSequence] =
VocoderOut(buffer/std(buffer),'noise',channels);

            % write wav-files containing vocoder signal
%
wavwrite(VocNew/max(abs(VocNew)),param.fs,['VocNew_SNR_',num2str(SNR_cond),
'dB_',filename]);

wavwrite(VocNew,param.fs,['VocNew_SNR_',num2str(SNR_cond),'dB_',filename]);
        end

        info = [num2str(voc_original_out) num2str(voc_processed_out)];

        if par.strategy
            type = 'Processing applied to clean signals';
        else
            type = 'Processing applied to noisy signals';
        end

        switch info
            case '11'
                % saving COG information and sequence data into file

save([filename,'_INFO.mat'],'COGs','type','OriginalSequence','ProcessedSequ
ence');
            case '10'
                % saving COG information and sequence data into file

save([filename,'_INFO.mat'],'COGs','type','OriginalSequence');
            case '01'
                % saving COG information and sequence data into file

save([filename,'_INFO.mat'],'COGs','type','ProcessedSequence');
            case '00'
                % saving COG information into file
                save([filename,'_INFO.mat'],'COGs','type');
            otherwise
                disp('Vocoder output flags set to invalid values');
        end
        cd(c_fold);
    end
end
cd(parent);
```

Appendix 9

Confirmation of Ethics approval for each experiment

Experiment I

UNIVERSITY OF SOUTHAMPTON

INSTITUTE OF SOUND AND VIBRATION RESEARCH

HUMAN EXPERIMENTATION SAFETY AND ETHICS COMMITTEE

Human Experimentation Safety and Ethics Application Number: 1259

Title of Experiment: Enhancing fast modulation in speech associated with obstruent acoustic landmarks

submitted by: Cherith Web **on** 09 January 2012

The Human Experimentation Safety and Ethics Committee has found the planned study satisfactory and confirm that it can proceed.

Please observe the following:

1. ***Record the Human Experimentation Safety and Ethics Approval Number on the subject consent forms. This number can be found at the top of this page.***
2. ***The subject consent forms, to be completed by all the subjects who participate in this study, must be provided to the Secretary of the Human Experimentation Safety and Ethics Committee by 09 May 2012.***
3. ***You must not make any changes to the study without obtaining the approval of the Human Experimentation Safety and Ethics Committee.***
4. ***You must inform the Human Experimentation Safety and Ethics Committee immediately if any subject experiences a problem or makes a complaint related to this study.***

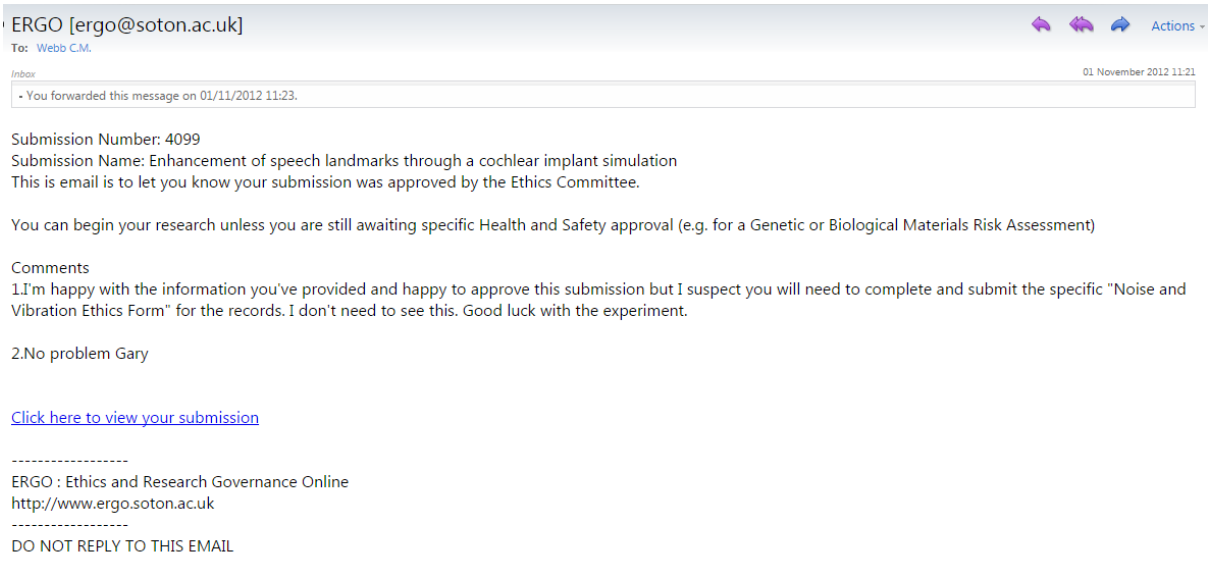
Date: 09 February 2012

Signed: .....
pp
Professor M.J. Griffin
Chair, Human Experimentation Safety and Ethics Committee

cc.Supervisor: Dr Carl Verschuur

Safety and Ethics Approval Form Issue 1.0 06/02/2006

Experiment II



ERGO [ergo@soton.ac.uk]

To: Webb C.M.

Actions

Inbox

01 November 2012 11:21

- You forwarded this message on 01/11/2012 11:23.

Submission Number: 4099
Submission Name: Enhancement of speech landmarks through a cochlear implant simulation
This is email is to let you know your submission was approved by the Ethics Committee.

You can begin your research unless you are still awaiting specific Health and Safety approval (e.g. for a Genetic or Biological Materials Risk Assessment)

Comments
1.I'm happy with the information you've provided and happy to approve this submission but I suspect you will need to complete and submit the specific "Noise and Vibration Ethics Form" for the records. I don't need to see this. Good luck with the experiment.

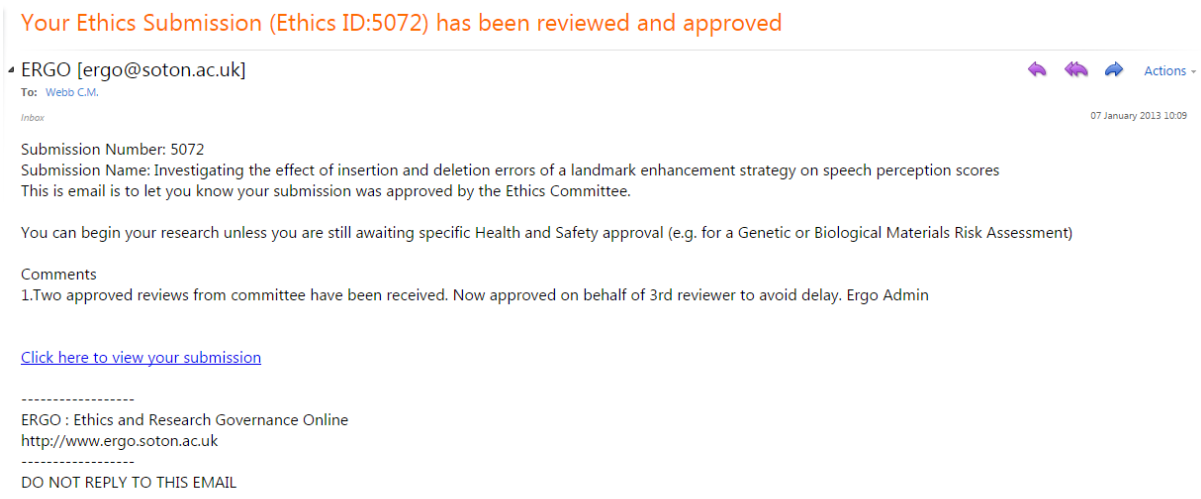
2.No problem Gary

Click here to view your submission

ERGO : Ethics and Research Governance Online
http://www.ergo.soton.ac.uk

DO NOT REPLY TO THIS EMAIL

Experiment III



Your Ethics Submission (Ethics ID:5072) has been reviewed and approved

ERGO [ergo@soton.ac.uk]

To: Webb C.M.

Actions

Inbox

07 January 2013 10:09

Submission Number: 5072
Submission Name: Investigating the effect of insertion and deletion errors of a landmark enhancement strategy on speech perception scores
This is email is to let you know your submission was approved by the Ethics Committee.

You can begin your research unless you are still awaiting specific Health and Safety approval (e.g. for a Genetic or Biological Materials Risk Assessment)

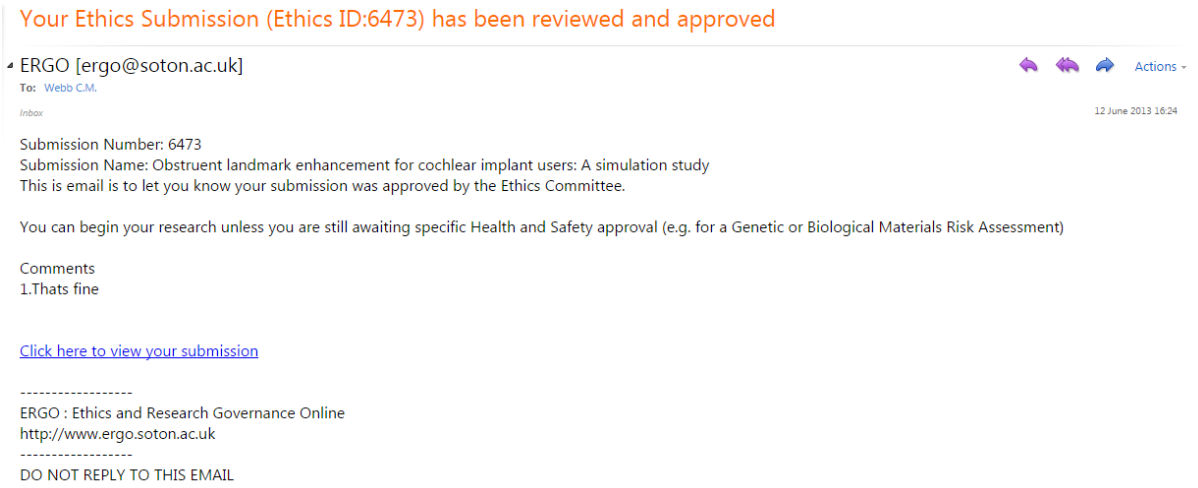
Comments
1.Two approved reviews from committee have been received. Now approved on behalf of 3rd reviewer to avoid delay. Ergo Admin

Click here to view your submission

ERGO : Ethics and Research Governance Online
http://www.ergo.soton.ac.uk

DO NOT REPLY TO THIS EMAIL

Experiment IV



Your Ethics Submission (Ethics ID:6473) has been reviewed and approved

ERGO [ergo@soton.ac.uk]

Actions

To: Webb C.M.

Inbox

12 June 2013 16:24

Submission Number: 6473
Submission Name: Obstruent landmark enhancement for cochlear implant users: A simulation study
This is email is to let you know your submission was approved by the Ethics Committee.

You can begin your research unless you are still awaiting specific Health and Safety approval (e.g. for a Genetic or Biological Materials Risk Assessment)

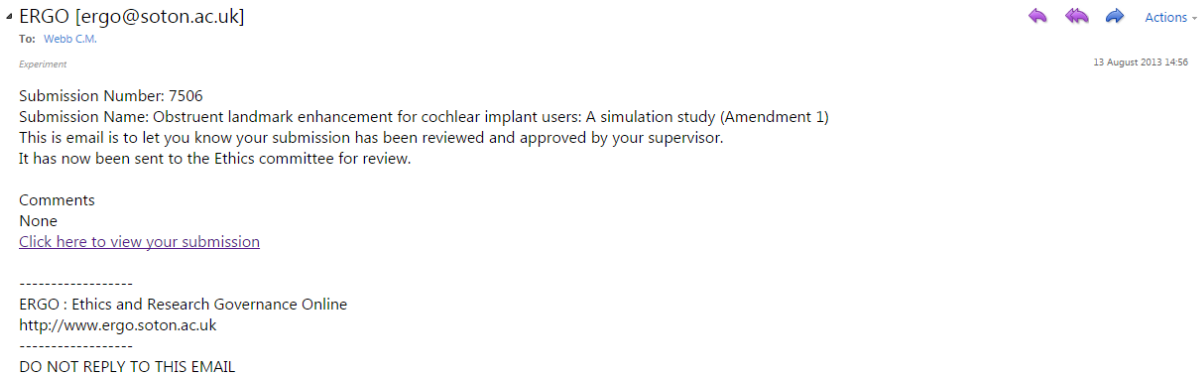
Comments
1.Thats fine

Click here to view your submission

ERGO : Ethics and Research Governance Online
http://www.ergo.soton.ac.uk

DO NOT REPLY TO THIS EMAIL

Experiment V (application submitted as an Amendment to experiment IV)



ERGO [ergo@soton.ac.uk]

Actions

To: Webb C.M.

Experiment

13 August 2013 14:56

Submission Number: 7506
Submission Name: Obstruent landmark enhancement for cochlear implant users: A simulation study (Amendment 1)
This is email is to let you know your submission has been reviewed and approved by your supervisor.
It has now been sent to the Ethics committee for review.

Comments
None
Click here to view your submission

ERGO : Ethics and Research Governance Online
http://www.ergo.soton.ac.uk

DO NOT REPLY TO THIS EMAIL

References

ABBAS, P. J. & MILLER, C. A. (2004). Biophysics and Physiology. *In:* BACON, S. P., FAY, R. R. & POPPER, A. N. (eds.) *Compression: From Cochlea to Cochlear Implants.* New York: Springer

ALBERT, E. S., BEC, J. M., DESMADRYL, G., CHEKROUD, K., TRAVO, C., GABOYARD, S., BARDIN, F., MARC, I., DUMAS, M., LENAERS, G., HAMEL, C., MULLER, A & CHABBERT, C. (2012). TRPV4 channels mediate the infrared laser-evoked response in sensory neurons. *Neurophysiol*, 107: 3227-3234.

ALI, H., HAZRATI, O., TOBEY, E. A. & HANSEN, J, H. L. (2014). Evaluation of adaptive dynamic range optimization in adverse listening conditions for cochlear implants. *J Acoust Soc Am*, 136 (3): El242-El248.

ALLUM, D. J. (1996). *Cochlear Implant Rehabilitation in Children and Adults,* London, Whurr Publishers Ltd.

ASHBY, M. & MAIDMENT, J. A. (2005). *Introducing phonetic science,* Cambridge, Cambridge University Press

BACON, S. P. (2004). Overview of auditory compression. *In:* BACON, S. P., FAY, R. R. & POPPER, A. N. (eds.). *Compression: From Cochlea to Cochlear Implants.* New York: Springer

BACON, S. P., FAY, R. R. & POPPER, A. N. (eds.) (2004). *Compression: From Cochlea to Cochlear Implants.* New York: Springer

BARLOW, H. B. (1959). Sensory mechanisms, the reduction of redundancy and intelligence, *In:* National Physical Laboratory symposium No. 10.

BENCH, J., KOWAL, A. & BAMFORD, J. (1979) The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially Hearing Children. *British Journal of Audiology*, 13: 108-112.

BENKI, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*. 29: 1-22.

BERENSTEIN, C. K., MENS, L. H. M, MULDER, J. J. S. & VANPOUCKE, F.J. (2008). Current Steering and Current Focusing in Cochlear Implants: Comparison of Monopolar, Tripolar, and Virtual Channel Electrode Configurations. *Ear Hearing*, 29 (2): 250-260.

BESKOW, J. & SJÖLANDER, K. (2004). *Wavesurfer* [computer program]. Version 1.8.5, Available: http://www.speech.kth.se/wavesurfer/ [Accessed 30 December 2011].

BITAR, N. N. (1997). *Acoustic Analysis and Modelling of Speech Based on Phonetic Features.* PhD Thesis, Boston University.

BITAR, N. N. & ESPY-WILSON, C. (1995). Speech parameterization based on phonetic features: application to speech recognition. *4th European Conference on Speech Communication and Technology EUROSPEECH '95.* Madrid, Spain.

BLAMEY, P. J. (2005). Adaptive Dynamic Range Optimization (ADRO): A Digital Amplification Strategy for Hearing Aids and Cochlear Implants. *Trends Amlif*, 9 (2): 77-98.

BLAMEY, P. J., DOWELL, R. C., BROWN, A. M., CLARK, G. M. & SELIGMAN, P. M. (1987). Vowel and Consonant Recognition of Cochlear Implant Patients Using Formant-Estimating Speech Processors. *J Acoust Soc Am,* 82 (1)**:** 48-57.

BOOTHROYD, A., ERICKSON, F. N. & MEDWETSKY, L. (1994) The hearing aid input: a phonemic approach to assessing the spectral distribution of speech. *Ear Hear*, 15 (6): 432-442.

BOOTHROYD, A., MULHEARN, B., GONG, J. & OSTROFF, J. (1996). Effects of spectral smearing on phoneme and word recognition. *J Acoust Soc Am,* 100 (3)**:** 1807-1818.

BORDEN, G. J., HARRIS, K. S. & RAPHAEL, L. J. (2007). *Speech science primer : physiology, acoustics, and perception of speech,* Philadelphia, Lippincott Williams & Wilkins.

BOERSMA, P. & WEENINK, D. (2011). *Praat, a system for doing phonetics by computer* [computer program].Version 5.3.32, Available: http://www.praat.org/ [Accessed 21 November 2011].

BREGMAN, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organisation of Sound,* London, MIT Press.

BRUNGART, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am,* 109 (3): 1101-1109.

BUSS, E., PILLSBURY, H., BUCHMAN, C., PILLSBURY, C. & CLARK, M. (2008). Multicenter U.S. bilateral Med-El cochlear implantation study: Speech perception over the first year of use. *Ear Hear,* 29 (1): 20-32.

CHATTERJEE, M. & SHANNON, R. V. (1998) Forward masked excitation patterns in multielectrode electrical stimulation. *J Acoust Soc Am,* 103 (5): 2565-2572

CHEN, F. & LOIZOU, P. C. (2010). Contribution of Consonant Landmarks to Speech Recognition in Simulated Acoustic-Electric Hearing. *Ear Hear* 31 (2)**:** 259-267.

CHINCHILLA, S. S. & FU, Q. J. (2003 Voice gender discriminations and vowel recognition in normal-hearing and cochlear implants users. *In* Absratcts of the *Conference on Implantable Auditory Prostheses,* Asilomar, CA.

CHOMSKY, N. & HALLE, M. (1968). *The sound pattern of English,* New York, Harper & Row.

CHRISTOVICH, ET AL. (1982). *In*: DELGUTTE, B. (1997). Auditory Neural Processing of Speech. *In:* HARDCASTLE, W. J. & LAVER, J. (eds.) *The Handbook of Phonetic Sciences.* Oxford: Blackwell.

COCHLEAR LTD. (1993). *Nucleus Technical References 3: Programming fundamentals.*

COCHLEAR LTD. (2015). *How it works* [Online]. Available: http://www.cochlear.com/wps/wcm/connect/uk/home/discover/cochlear-implants/how-it-works [Accessed 25/11/1010/04/15].

COLE, R. A., YAN, Y., MAK, B. & BAILEY, T. (1996) The contribution of consonants versus vowels to word recognition in fluent speech. *In:* Acoustics, *IEEE International Conference on Speech and Signal Processing*. 5: 853-856.

COOPER, F., DELATTRE, P., LIBERMAN, A., BORST, J. & GERSTMAN, L. (1952). Some experiments on the perception of synthetic speech sounds. *J Acoust Soc Am* 24: 597–606.

COSENDAI, G. & PELIZZONE, M. (2001). Effects of the Acoustical Dynamic Range on Speech Recognition with Cochlear Implants. *Audiology*, 40 (5): 272-281

COVER, T. M. & THOMAS, J. A. (1991). *Elements of information theory,* New York, John Wiley.

DAU, T., VERHEY, J. & KOHLRAUSCH, A. (1999). Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers. *J Acoust Soc Am,* 106: 2752-2760.

DELGUTTE, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *J Acoust Soc Am,* 68 (3)**:** 843-857.

DELGUTTE, B. (1997). Auditory Neural Processing of Speech. *In:* HARDCASTLE, W. J. & LAVER, J. (eds.) *The Handbook of Phonetic Sciences.* Oxford: Blackwell.

DELGUTTE, B. & KIANG, N. Y. S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *J Acoust Soc Am,* 75 (3)**:** 897-907.

DENES, P. B. & PINSON, E. N. (1993). The speech chain : the physics and biology of spoken language, New York, Freeman.

DIEHL, R. L. & LINDBLOM, B. (2004). Explaining the Structure of Feature and Phoneme Inventories: The Role of Auditory Distinctiveness. *In:* GREENBERG, S., AINSWORTH, W. A., POPPER, A. N. & FAY, R. R. (eds.) *Speech Processing in the Auditory System.* New York: Spinger.

DORMAN, M. F., SOLI, S., DANKOWSKI, K., SMITH, L. M., MCCANDLESS, G. & PARKIN, J. (1990). Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant. *J Acoust Soc Am*, 88: 2074– 2079.

DORMAN, M. F., LOIZOU, P. C. & RAINEY, D. (1997) Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J Acoust Soc Am*, 102 (4): 2403- 2411

DORMAN, M. F. & LOIZOU, P. C. (1998). The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels. *Ear Hear,* 19 (2)**:** 162-166

DORMAN, M. F., LOIZOU, P. C., FITZKE, J. & TU, Z. M. (1998a). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels. *J Acoust Soc Am,* 104 (6)**:** 3583-3585.

DORMAN, M. F., LOIZOU, P. C. & FITZKE, J. (1998b). The identification of speech in noise by cochlear implant patients and normal-hearing listeners using 6-channel signal processors. *Ear Hear,* 19 (6)**:** 481-484.

DORMAN, M. F., LOIZOU, P. C., SPAHR, A. J. & MALOFF, E. (2002). A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants. *J Speech Lang Hear R,* 45 (4).

DOWELL, R. C., SELIGMAN, P. M., BLAMEY, P. J. & CLARK, G. M. (1987). Speech-Perception Using a 2-Formant 22-Electrode Cochlear Prosthesis in Quiet and in Noise. *Acta Oto-Laryngol,* 104 (5-6)**:** 439-446.

DRULLMAN, R. (1995). Temporal Envelope and Fine-Structure Cues for Speech-Intelligibility. *J Acoust Soc Am,* 97 (1)**:** 585-592.

DRULLMAN, R., FESTEN, J. M. & PLOMP, R. (1994). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am,* 95 (2): 1053-1064.

DUDLEY, H. (1939). Remaking speech. *J Acoust Soc Am*, 11: 169-177.

DURLACH, N. I., MASON, C. R., SHINN-CUNNINGHAM, B. G., ARBOGAST, T. C., COLBURN, H. S. & KIDD, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity a). *J Acoust Soc Am*, 114 (1): 368-379.

ESPY-WILSON, C. Y., PRUTHI, T., JUNEJA, A. & DESHMUKH, O. (2007). Landmark-based Approach to Speech Recognition: An Alternative to HMMs. *In:* Interspeech: 8th Annual Conference of the International Speech Communication Association. 2516-2519.

FANT, G. (1964). Auditory patterns of speech. *In*: Department for Speech Music and Hearing: *Quarterly Progress and Status Report* [Online]. Available: http://www.speech.kth.se/qpsr/ [Accessed 04/05/11].

FANT, G. (1973). *Speech sounds and features*, Cambridge, MIT Press.

FOGERTY, D. & KEWLEY-PORT, D. (2009). Perceptual contributions of the consonant-vowel boundary to speech intelligibility [a)]. *J Acoust Soc Am*, 126 (2): 847-857.

FREYMAN, R. L., BALAKRISHNAN, U. & HELFER, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech. *J Acoust Soc Am*, 115 (5): 2246-2256.

FRIESEN, L. M., SHANNON, R. V., BASKENT, D. & WANG, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am,* 110 (2)**:** 1150-1163.

FRIESEN, L. M., SHANNON, R. V. & CRUZ, R. J. (2005). Effects of Stimulation Rate on Speech Recognition with Cochlear Implants. *Audiol Neurotol*, 10: 169-184.

FU, Q. J. & NOGAKI, G. (2004). Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. *Jaro-J Assoc Res Oto,* 6 (1)**:** 19-27.

FU, Q. J. & SHANNON, R. V. (1998). Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners [a)]. *J Acoust Soc Am*, 104 (5): 2570-2577.

FU, Q. J. & SHANNON, R. V. (1999a). Effect of acoustic dynamic range on phoneme recognition in quiet and noise by cochlear implant users. *J Acoust Soc Am*, 106 (6): L65-L70.

FU, Q. J. & SHANNON, R. V. (1999b). Phoneme recognition by cochlear implant users as a function of signal-to-noise ratio and nonlinear amplitude mapping. *J Acoust Soc Am*, 106 (2): L18- L23.

FU, Q. J. & SHANNON, R. V. (1999c). Effects of electrode location and spacing on phoneme recognition with the nucleus-22 cochlear implant. *Ear Hear*, 20 (4)**:** 321-331.

FU, Q. J. & SHANNON, R. V. (1999d). Effects of electrode configuration and frequency allocation on vowel recognition with the Nucleus-22 cochlear implant. *Ear Hear*, 20 (4): 332.

FU, Q. J., SHANNON, R. V. & WANG, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *J Acoust Soc Am*, 104 (6): 3586-3596.

FU, Q. J. & SHANOON, R. V. (2000). Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners. *J Acoust Soc Am*, 107 (1): 589-597.

FU, Q. J., CHINCHILLA, S. & GALVIN, J. J. (2004). The Role of Spectral and Temporal Cues in Voice Gender Discrimination by Normal-Hearing Listeners and Cochlear Implant Users. *JARO*, 5: 253-260.

FURUI, S. (1986). On the role of spectral transition for speech perception. *J Acoust Soc*, 80 (4): 1016-1025.

GALVIN, J. J. & FU, Q. J. (2005). Effects of Stimulation Rate, Mode and Level on Modulation Detection by Cochlear Implant Users. *JARO*, 6: 269-279.

GELFAND, S. A. (2004). *Hearing : an introduction to psychological and physiological acoustics*, New York, Marcel Dekker.

GEURTS, L & WOUTERS, J. (1999) Enhancing the speech envelope of continuous interleaved sampling processors for cochlear implants. *J Acoust Soc Am*, 105 (4): 2476-2484.

GLASS, J., CHANG, J. & MCCANDLESS, M. (1996). A probabilistic framework for feature based speech recognition. *In*: Fourth International Conference on Spoken Language, ICSLP. 4: 2277-2280.

GONZALEZ, J. & OLIVER, J. C. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *J Acoust Soc Am*, 118 (1): 461-470.

GREEN, T., FAULKNER, A. & ROSEN, S. (2012). Variations in Carrier Pulse Rate and the Perception of Amplitude Modulation in Cochlear Implant Users. *Ear Hear*, 33 (2): 221-230.

GREENBERG, S. & AINSWORTH, W. A. (2004). Speech Processing in the Auditory System: An overview. *In:* GREENBERG, S., AINSWORTH, W. A., POPPER, A. N. & FAY, R. R. (eds.) *Speech Processing in the Auditory System.* New York: Springer.

GREENBERG, S., CARVEY, H., HITCHCOCK, L. & CHANG, S. (2003). Temporal properties of spontaneous speech- a syllable-centric perspective. *Journal of Phonetics*, 31: 465-485.

HANEKOM, J. J. & SHANNON, R. V. (1998). Gap detection as a measure of electrode interaction in cochlear implants. *J Acoust Soc Am*, 104 (4): 2372-2384.

HASEGAWA-JOHNSON, M., BAKER, J., BORYS, S., CHEN, K., COOGAN, E., GREENBERG, S., JUNEJA, A., KIRCHHOFF, K., LIVESCU, K., MOHAN, S., MULLER, J., SONMEZ, K. & TANG, T. Y. (2005).

Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. *2005 Ieee International Conference on Acoustics, Speech, and Signal Processing, Vols 1-5***:** 213-216.

HASEGAWA-JOHNSON, M. (2001). Finding the best acoustic measurements for landmark-based speech recognition. (Unpublished) Urbana: University of Illinois. Available from www.illinois.edu.

HAWKINS, S. (1999). Looking for invariant correlates of linguistic units: Two classical theories of speech perception. *In:* PICKETT, J. M. (ed.) *The Acoustics of Speech Perception: Fundamentals, speech perception theory, and technology.* Boston: Allyn and Bacon.

HAZAN, V & SIMPSON, A. (1998) The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24: 211-226

HAZAN, V & SIMPSON, A. (2000) The Effect of Cue-Enhancement on Consonant Intelligibility in Noise: Speaker and Listener Effects. *Language and Speech*, 43 (3): 273-294

HAZRATI, O., LOIZOU, P. C. (2013). Comparison of two channel selection criteria for noise suppression in cochlear implants. *J Acoust Soc Am*, 133 (3):1615-1624.

HENRY, B. A., TURNER, C. W. & BEHRENS, A. (2005). Spectral peak resolution and speech recognition in quiet: Normal hearing, hearing impaired, and cochlear implant listeners. *J Acoust Soc Am*, 118 (2): 1111-1121.

HODOSHIMA, N., TAKAYUKI, A. & AKIKO, K. (2002). Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments. *Forum Acusticum Sevillai*

HOLDEN, L. K., SKINNER, M. A., HOLDEN, T. A. & DEMOREST, M. E., (2002). Effects of Stimulation Rate with the Nucleus 24 ACE Speech Coding Strategy. *Ear Hear*, 23 (5): 464-476.

HOLDEN, L. K., VANDALI, A. E., SKINNER, M. W., FOURAKIS, M. S. & HOLDEN, T. A. (2005). Speech recognition with the advanced combination encoder and transient emphasis spectral maxima strategies in nucleus 24 recipients. *J Speech Lang Hear R*, 48 (3)**:** 681-701.

HOLMES, J. & HOLMES, W. (2001). *Speech synthesis and recognition*, New York, Taylor and Francis.

HONG, R. S., RUBINSTEIN, D. W. & HORN, D. (2003) Dynamic Range Enhancement for Cochlear Implants. *Otology and Neurotology*, 24: 590-595.

HOWELL, D. C. (2012). *Statistical methods for psychology (8th Ed.).* Belmont, CA: Wadsworth.

HU, Y. & LOIZOU P. C. (2008). A new sund coding strategy for supressing noise in cochlear implants. *J Acoust Soc Am*, 124 (1): 498-509.

HU. H., LI, G., CHEN, L., SANG, J., WANG, S., LUTMAN, M. E. & BLEECK, S. (2011). Enhanced sparse speech processing strategy for cochlear implants. *In*: 19th European Signal Processing Conference, Barcelona, Spain.

HU. H., MOHAMMADIHA, N., TAGHIA, J., LEIJON, A., LUTMAN, M. E. & WANG, S. (2012). Sparsity level in a non-negative matrix factorization based speech strategy in cochlear implants. *In:* 20th European Signal Processing Conference, Bucharest, Romania.

HUCKVALE, M. (1997). A syntactic pattern recognition method for the automatic location of potential enhancement regions in running speech. *Speech, Hearing and language: UCL* [Online]. Available: http://www.phon.ucl.ac.uk/home/shl10/huckvale/shlmark.htm. [Accessed 01 January 2015].

JAMES, C. J., BLAMEY, P. J., MARTIN, L., SWANSON, B., JUSt, Y. & MACFARLANE, D. (2002). Adaptive Dynamic Range Optimization for Cochlear Implants: A Preliminary Study. *Ear Hear*, 23 (1S): 49S-58S.

JAMES, C. J., SKINNER, M. W., MARTIN, L. F. A., HOLDEN, L. K., GALVIN, K. L., HOLDEN, T. A. & WHITFORD, L. (2003). An Investigation of Input Level Range for the Nucleus 24 Contour Implant System: Speech Perception Performance, Program Preference, and Loudness Comfort Ratings. *Ear Hear*, 24 (2): 157-174.

JANSEN, A. & NIYOGI, P. (2008). Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition. *J Acoust Soc Am,* 124 (3)**:** 1739-1758.

JUNEJA, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks.* PhD Thesis, University of Maryland

JUNEJA, A. & ESPY-WILSON, C. (2003). Speech Segmentation Using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines. *In: Proceedings of the International Joint Conference on Neural Networks.* 1: 675-679.

JUNEJA, A. & ESPY-WILSON, C. (2004). Significance of Invariant Acoustic Cues in a Probabilistic Framework for Landmark based Speech Recognition. *In:* From Sound to Sense, MIT. 151-156.

JUNEJA, A. & ESPY-WILSON, C. (2008). A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *J Acoust Soc Am,* 123 (2)**:** 1154-68.

KEIDSER, G., DILLON, H., CONVERY, E. & O'BRIEN, A. (2011). Differences Between Speech-Shaped Test Stimuli in Analyzing Systems and the Effect on Measured Hearing Aid Gain. *Ear Hear*, 31 (3): 437–440.

KEWLEY-PORT, D., BURKE, T. Z. & LEE, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners [a)]. *J Acoust Soc Am*, 122 (4): 2365-2375.

KIM, K. H., CHOI, S. J., KIM, J. H. & KIM, D. H. (2009). An Improved Speech Processing Strategy for Cochlear Implants Based on an Active Nonlinear Filterbank Model of the Biological Cochlea. *IEEE Transactions on Biomedical Engineering*, 56 (3): 828-836

KHING, P. P., SWANSON, B. A. & AMBIKAIRAJAH, E. (2013). The Effect of Automatic Gain Control Structure and Release Time on Cochlear Implant Speech Intelligibility. *PLOS One*, 8 (11): 1-11

KIEFER, J., HOHL, S., STÜRZEBECHER, E., PFENNIGDORFF, T. & GSTÖETTNER, W. (2001). Comparison of Speech Recognition with Different Speech Coding Strategies (SPEAK, CIS, and ACE) and Their Relationship to Telemetric Measures of Compound Action Potentials in the Nucleus CI 24M Cochlear Implant System. *Audiology*, 40 (1): 32-42.

KLEUNER, K. R., COADY, J., A. & KIEFTE, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, 41: 59-69.

KOCH, D. B., DOWNING, M., OSBERGER, M. J. & LITVAK, L. (2007). Using Current Steering to Increase Spectral Resolution in CII and HiRes 90K Users. *Ear Hear*, 28 (2): 38S-41S.

KOHLRAUSCH, A., FASSEL, R. & DAU, T. (2000). The influence of carrier level and frequency modulation and beat-detection thresholds for sinusoidal carriers. *J Acoust Soc Am*, 108: 723-734.

KONING, R & WOUTERS, J. (2012) The potential of onset enhancement for increased speech intelligibility in auditory prostheses. *J Acoust Soc Am*, 132 (4): 2569-2581.

KONG, Y. Y., DEEKS, J. M., AXON, P. R. & CARLYON, R. P. (2009). Limits of temporal pitch in cochlear implants. *J Acoust Soc Am,* 125 (3)**:** 1649-1657.

LAMEL, L. F., ET AL. (1986). Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus *In:* COLE, R. A., YAN, Y., MAK, B. & BAILEY, T. (1996) The contribution of consonants versus vowels to word recognition in fluent speech. In: *IEEE Conference Proceedings for the International Conference on Acoustics, Speech, and Signal Processing*. 2: 853-856.

LANDSBERGER, D. M., & SRINIVASAN, A. G. (2009). Virtual channel discrimination is improved by current focusing in cochlear implant recipients. *Hear Res*, 254: 34–41.

LANDSBERGER, D. M., PADILLA, M. & SRINIVASAN, A. G. (2012). Reducing current spread using current focusing in cochlear implant users. *Hear Res*, 284: 16–24.

LANEAU, J., MOONEN, M. & WOUTERS, J. (2006). Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants. *J Acoust Soc Am*, 119 (1): 491-506.

LEAKE, P. A. & REBSCHER, S. J. (2004). Anatomical Considerations and Long-Term Effects of Electrical Stimulation. *In*: ZENG, F. G., POPPER, A. N. & FAY, R. R. (eds.) *Cochlear Implants: Auditory Prostheses and Electric Hearing.* New York: Springer

LEEK, M. R. & SUMMERS, V. (1996). Reduced frequency selectivity and the preservation of spectral contrast in noise. *J Acoust Soc Am,* 100 (3): 1796-1806.

LEVITT, H. (2001). Noise reduction in hearing aids: A review. *J Rehabil Res Dev,* 38 (1): 111–121.

LEWICKI, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience,* 5 (4): 356-363.

LI, N. & LOIZOU, P. C. (2007). Factors influencing glimpsing of speech in noise. *J Acoust Soc Am,* 122 (2)**:** 1165-1172.

LI, N. & LOIZOU, P. C. (2008a). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *J Acoust Soc Am,* 124 (6)**:** 3947.

LI, N. & LOIZOU, P. C. (2008b). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *J Acoust Soc Am*, 123 (3): 1673-1682.

LI, N. & LOIZOU, P. C. (2008c). Effect of spectral resolution on the intelligibility of ideal binary masked speech. *J Acoust Soc Am*, 123 (4): EL59-El64.

LI, N. & LOIZOU, P. C. (2009). Factors affecting masking release in cochlear-implant vocoded speech. *J Acoust Soc Am,* 126 (1)**:** 338-346.

LI, N. & LOIZOU, P. C. (2010). Masking release and the contribution of obstruent consonants on speech recognition in noise by cochlear implant users. *J Acoust Soc Am,* 128 (3)**:** 1262-1271.

LI, G. & LUTMAN, M. E. (2008). Sparse stimuli for cochlear implants. *In:* 16th European Signal Processing Conference, Lausanne, Switzerland.

LITTLEFIELD, P. D., VUJANOVIC, I., MUNDI, J., MATIC, A. I. & RICHTER, C. P. (2010). Laser stimulation of single auditory nerve fibers. *Laryngoscope,* 120 (10): 2071–2082.

LIU, S. A. (1996). Landmark detection for distinctive feature-based speech recognition. *J Acoust Soc Am,* 100 (5)**:** 3417-3430.

LOIZOU, P. C. (1998). Mimicking the human ear: An overview of signal processing techniques for converting sound to electrical signals in cochlear implants. *IEEE Signal Process Mag*, 15: 101–130.

LOIZOU, P. C. (2006). Speech processing in vocoder-centric cochlear implants. *Adv Otorhinolaryngol*, 64: 109-

LOIZOU, P. C. (2007). *Speech Enhancement: Theory and Practice,* Boca Raton, Taylor and Francis Group.

LOIZOU, P. C., DORMAN, M. & TU, Z. (1999). On the number of channels needed to understand speech. *J Acoust Soc Am*, 106 (4): 2097-2103.

LOIZOU, P. C., DORMAN, M. & FITZKE, J. (2000a). The Effect of Reduced Dynamic Range on Speech Understanding: Implications for Patients with Cochlear Implants. *Ear Hear*, 21 (1): 25-31.

LOIZOU, P. C. & POROY, O. (2001). Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners. *J Acoust Soc Am,* 110 (3)**:** 1619-1627.

LOIZOU, P. C., POROY, O. & DORMAN, M. (2000b). The effect of parametric variations of cochlear implant processors on speech understanding. *J Acoust Soc Am*, 108 (2): 790-802.

LOIZOU, P. C., LITOVSKY, R., YU, G., PETERS, R., LAKE, J. & ROLAND, P. (2009). Speech recognition by bilateral cochlear implant users in a cocktail-party setting. *J Acoust Soc Am,* 125 (1): 372-383.

LORENZI, C., GATEHOUSE, S. & LEVER, C. (1999). Sound localization in noise in hearing-impaired listeners [a)]. *J Acoust Soc Am*, 105 (6): 3454-3463.

LUTMAN, M. E. (1997). Speech tests in quiet and noise as a measure of auditory processing. *In:* MARTIN, M. (Ed). *Speech Audiometry* (2nd Ed), London, Wurr Publishers Ltd.

MARTIN, M. (Ed). (1997). *Speech Audiometry* (2nd Ed.), London, Wuur Publishers Ltd.

MCLEOD, A. & SUMMERFIELD, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology*, 24: 29-43.

MCALPINE, D. (2011). Developing a neuro-centric perspective to cochlear implantation. *Cochlear Implants International*, 12 (S1): S40-S43.

MCDERMOTT, H. J., HENSHALL, K. R. & MCKAY, C. M. (2002). Benefits of Syllabic Input Compression for Users of Cochlear Implants. *J Am Acad Audiol*, 13 (1): 14-24.

MIDDLEBROOKS, J. C. & GREEN, D. M. (1991). Sound Localization by human listeners. *Annu Rev Psychol*, 42: 135-159.

MIDDLEBROOKS, J. C. & SNYDER, R. L. (2007). Auditory prosthesis with a penetrating nerve array. *J Assoc Res Otolaryngol*, 8: 258–279.

MIDDLEBROOKS, J. C. & SNYDER, R. L. (2010). Selective electrical stimulation of the auditory nerve activates a pathway specialized for high temporal acuity. *J Neurosci*, 30: 1937–1946.

MILLER, G. A. & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants *J Acoust Soc Am,* 77 (2)**:** 338-352.

MILLER, M. I. & SACHS, M. B. (1983). Representation of Stop Consonants in the Discharge Patterns of Auditory-Nerve Fibers. *J Acoust Soc Am,* 74 (2)**:** 502-517.

MOORE, B. C. J. (2003a). Speech processing for the hearing impaired: successes, failures, and implications for speech mechanisms. *Speech Communication*, 41: 81-91.

MOORE, B. C. J. (2003b). Coding of Sounds in the Auditory System and Its Relevance to Signal Processing and Coding in Cochlear Implants. *Otology & Neurology*, 24 (2): 243-254.

MORRIS, A. C., ESCUDIER, P. & SCHWARTZ, J. L. (1991). On and Off Units Detect Information Bottle-necks for Speech Recongition. *In:* 2nd European Conference on Speech Communication and Technology EUROSPEECH, Genova, Italy. 1441-1444.

MÜLLER, J., SCHON, F. & HELMS, J. (2002). Speech understanding in quiet and noise in bilateral users of the Med-El Combi 40 401 cochlear implant system. *Ear Hear*, 23 (3): 198–206.

MUNSON, B. & NELSON, P. B. (2005). Phonetic identification in quiet and in noise by listeners with cochlear implants. *J Acoust Soc Am,* 118 (4)**:** 2607-2617.

MUNSON, B., DONALDSON, G. S., ALLEN, S. L., COLLISON, E. A. & NELSON, D. A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *J Acoust Soc Am,* 113 (2)**:** 925-935.

NELSON, D. A., SCHMITZ, J. L., DONALDSON, G. S., VIEMEISTER, N. F. & JAVEL, E. (1996). Intensity discrimination as a function of stimulus level with electric stimulation. *J Acoust Soc Am*, 100 (4): 2393-2414.

NAGARAJAN, S. S., WANG, X., MERZENICH, M. M., SCHREINER, C. E., JOHNSTON, P., JENKINS, W. M., MILLER, S. & TALLAL, P. (1998). Speech modifications algorithms used for training language learning-impaired children. *IEEE transactions on Rehabilitation Engineering,* 6 (3)**:** 257-268.

NELSON, P. B., JIN, S. H., CARNEY, A. E. & NELSON, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *J Acoust Soc Am,* 113 (2)**:** 961-968.

NUETZEL, J. M. & HAFTER, E. R. (1981). Discrimination of interaural delays in complex waveforms: Spectral effects. *J Acoust Soc Am,* 69 (4)**:** 1112-1118.

National Institute of Health and Clinical Excellence [NICE] (2009). Cochlear implants for children and adults with severe to profound deafness, London.

OLSHAUSEN, B. A. & FIELD, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14: 481-487.

ORTEGA, M, HAZAN, V. & HUCKVALE, M. (2000) Automatic cue-enhancement of natural speech for improved intelligibility. *Speech, Hearing and Language: work in progress*, 12: 42-56.

OWREN, M. J. & CARDILLO, G. C. (2006). The relative role of vowels and consonants in discriminating talker identity versus word meaning [a)]. *J Acoust Soc Am*, 119 (3): 1727-1739.

OXENHAM, A. J. & KREFT. H. A. (2014). Speech Perception in Tones and Noise via Cochlear Implants Reveals Influence of Spectral Resolution on Temporal Processing. *Trends in Hearing*, 18: 1-14.

PALMER, A. & SHAMMA, S. (2004). Physiological Representations of Speech. *In:* GREENBERG, S., AINSWORTH, W. A., POPPER, A. N. & FAY, R. R. (eds.) *Speech Processing in the Auditory System.* New York: Springer.

PARIKH, G. & LOIZOU, P. C. (2005). The influence of noise on vowel and consonant cues. *J Acoust Soc Am,* 118 (6)**:** 3874-3888.

PHONAK. (2014). *HearingLikeMe: Now Hear This*. [Online]. Available: http://www.hearinglikeme.com/facts/what-hearing-loss/now-hear. [Accessed 07 October 2014].

PICKETT, J. M. (1999). *The acoustics of speech communication: fundamentals, speech perception theory, and technology*, Boston, Allyn and Bacon.

PISONI, D. B. & REMEZ, R. E. (eds.) (2005). *The handbook of speech perception,* Malden, MA ; Oxford: Blackwell.

PLANT, G. & SPENS, K. (1995). *Profound Deafness and Speech Communication,* London, Whurr Publishers Ltd.

PORT, R. (2010). Forget about phonemes: Language processing with rich memory. [Unpublished] Indiana: Indiana University. Available from www.cs.indiana.edu

RAJGURU, S. M., MATIC, A. I., ROBINSON, A. M., FISHMAN, A. J., MORENO, L. E., BRADLEY, A., VUJANOVIC, I., BREEN, J., WELLS, J. D., BENDETT, M. & RICHTER, P. (2010). Optical cochlear implants: evaluation of surgical approach and laser parameters in cats. *Hear Res,* 269 (1-2): 102–111.

RAPHAEL, L. J. (2005) Acoustic cues to the perception of segmental phonemes. *In:* PISONI, D. B. & REMEZ, R. E. (eds.) *The handbook of speech perception,* Malden, MA ; Oxford: Blackwell.

ROSEN, S. (1992). Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects. Phil *Trans R Soc Lond* , 336 (1278): 367-373.

REETZ, H. & JONGMAN, A. (2008). *Phonetics: Transcription, Production, Acoustics and Perception*, Chichester, Blackwell Publishing.

SALOMON, A., ESPY-WILSON, C. Y. & DESHMUKH, O. (2004). Detection of speech landmarks: Use of temporal information. *J Acoust Soc Am,* 115 (3)**:** 1296-1305.

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

SHANNON, R. V. (1992). Temporal modulation transfer functions in patients with cochlear implants. *J Acoust Soc Am*, 91 (): 2156–2164.
SHANNON, R. V., ZENG, F. G., KAMATH, V., WYGONSKI, J. & EKELID, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, 270 (5234): 303-304.

SHANNON, R. V., FU, Q. U., GALVIN, J. & FRIESEN, L. (2004). Speech Perception with Cochlear Implants. *In*: ZENG, F. G., POPPER, A. N. & FAY, R. R. (eds.) *Cochlear Implants: Auditory Prostheses and Electric Hearing.* New York: Springer

SHATTUCK-HUFNAGEL, S. & VEILLEUX, N. M. (2007). Robustness of Acoustic Landmarks in Spontaneously-Spoken American English. *In:* 16th International Congress of Phonetic Sciences, Saarbrücken, Germany. 925-928.

SHEPHERD, R. K., HATSUSHIKA, S. & CLARK, G. M. (1993). Electrical-Stimulation of the Auditory-Nerve - the Effect of Electrode Position on Neural Excitation. *Hearing Res,* 66 (1)**:** 108-120.

SKINNER, M. W., FOURAKIS, M. S., HOLDEN, T. A., HOLDEN, L. K. & DEMOREST, M. E. (1996). Identification of speech by cochlear implant recipients with the multipeak (MPEAK) and spectral peak (SPEAK) speech coding strategies .1. Vowels. *Ear Hear,* 17 (3)**:** 182-197.

SKINNER, M. W., HOLDEN, L.K., WHITFORD, L. A., PLANT, K. L., PSARROS, C. & HOLDEN, T. A. (2002a) Speech Recognition with the Nucleus 24 SPEAK, ACE, and CIS Speech Coding Strategies in Newly Implanted Adults. *Ear Hear*, 23 (3): 207-223.

SKINNER, M. W., ARNDT, P. L. & STALLER, S. J. (2002b) Nucleus® 24 Advanced Encoder Conversion Study: Performance versus Preference. *Ear Hear*, 23 (1S): 2S-17S.

SMITH, L. S. 1995. Using an Onset-based Representation for Sound Segmentation. *NEURAP95*. Marseilles, France.

SMITH, E. C. & LEWICKI, M. S. (2006). Efficient auditory coding. *Nature*, 439: 978- 982.

SLIFKA, J., STEVENS, K. N., MANUEL, S. & SHATTUCK-HUFNAGEL, S. (2004). A landmark-based model of speech perception: history and recent developments. *In:* From Sound to Sense, MIT. 85-90.

SOUZA, P. & ROSEN, S. (2009). Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech. *J Acoust Soc Am*, 126 (2): 792-805.

SPAHR, A. J., DORMAN, M. F. & LOISELLE, L. H. (2007). Performance of Patients Using Different Cochlear Implant Systems: Effects of Input Dynamic Range. *Ear Hear*, 28 (2): 260-275.

SRINIVASAN, A. G., SHANNON, R. V. & LANDSBERGER, D. M. (2012). Improving virtual channel discrimination in a multi-channel context. *Hear Res, 286*: 19–29.

STEVENS, K. N. (1998). *Acoustic phonetics*, Cambridge (Mass.), MIT Press.

STEVENS, K. N. (2000). Diverse acoustic cues at consonantal landmarks. *Phonetica*, 57 (2-4)**:** 139-151.

STEVENS, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J Acoust Soc Am*, 111 (4)**:** 1872-1891.

STEVENS, K. N. (2005). Features in Speech Perception and Lexical Access. *In:* PISONI, D. B. & REMEZ, R. E. (eds.) *The Handbook of Speech Perception.* Malden, MA ; Oxford: Blackwell

STEVENS, K. N. & BLUMSTEIN, S. E. (1978). Invariant Cues for Place of Articulation in Stop Consonants. *J Acoust Soc Am*, 64 (5)**:** 1358-1368.

STICKNEY, G. S., ZENG, F. G., LITOVSKY, R. & ASSMANN, P. (2004). Cochlear implant speech recognition with speech maskers. *J Acoust Soc Am*, 116 (2)**:** 1081-1091

STILP. C.E. & KLEUNDER, K. R. (2010) Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *PNAS*, 107 (27): 12387–12392.

STRANGE, W., VERBRUGGE, R. R., SHANKWEILER, D. P. & EDMAN, T. R. (1976). Consonant environment specifies vowel identity. *J Acoust Soc Am*, 60 (1): 213-224.

STRANGE, W., JENKINS, J. J. & JOHNSON, T. L. (1983). Dynamic specification of coarticulated vowels. *J Acoust Soc Am*, 74 (3): 695- 705.

STUDEBAKER, G. (1985). A 'rationalized' arcsine transform. *J Speech Hear Res*, 28: 455–462.

STUDEBAKER, G. A. & SHERBECOE, R. L. (2002). Intensity-importance functions for bandlimited monosyllabic words. *J Acoust Soc Am*, 111(3): 1422-1436.

SWANSON, B. A. (2008). Pitch perception with cochlear implants. PhD Thesis, University of Melbourne.

SWANSON, B. & MAUCH, H. (2008). *Nucleus Matlab Toolbox* [computer program].Version 4.31.

THORNTON, A. & RAFFIN, M. (1978). Speech discrimination scores modelled as a binomial variable. *Journal of Speech and Hearing Research*, 21: 507-518.

TURNER, C. W., SOUZA, P. E. & FORGET, L. N. (1995). Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners. *J Acoust Soc Am*, 97 (4): 2568-2576.

TURNER, C. W., GANTZ, B. J., VIDAL, C., BEHRENS, A. & HENRY, B. A. (2004). Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing. *J Acoust Soc Am,* 115 (4)**:** 1729-1735.

TYE-MURRAY, N., LOWDER, M. & TYLER, R. S. (1990). Comparison of the Fof2 and Fof1f2 Processing Strategies for the Cochlear Corporation Cochlear Implant. *Ear Hear,* 11 (3)**:** 195-200.

TYLER, R. S., GANTZ, B. J., RUBINSTEIN, J. T., WILSON, B. S., PARKINSON, A. J., WOLAVER, A., PREECE, J. P., WITT, S. & LOWDER, M. W. (2002). Three-Month Results with Bilateral Cochlear Implants. *Ear Hear,* 23 (1): 80S-89S.

VAN HOESEL, R. J. M. & CLARK, G. M. (1995). Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients. *J Acoust Soc Am,* 97(4):2498-2503.

VAN HOESEL, R. J. M. & CLARK, G. M. (1997). Psychophysical studies with two binaural cochlear implants subjects. *J Acoust Soc Am,* 102 (1)**:** 495-507.

VAN HOESEL, R. J. M., TONG, Y. C., HOOLOW, R. D. & CLARK, G. M. (1993). Psychophysical and speech perception studies: A case report on a binaural cochlear implant subject. *J Acoust Soc Am,* 94 (6)**:** 3179-3189.

VAN WIERINGEN, A. & WOUTERS, L. (2008). LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. *Int J Audiol,* 47: 348–355.

VANDALI, A. E. (2001). Emphasis of short-duration acoustic speech cues for cochlear implant users. *J Acoust Soc Am,* 109 (5)**:** 2049-2061.

VANDALI, A. E., WHITFORD, L. A., PLANT, K. L. & CLARK, G. M. (2000). Speech Perception as a Function of Electrical Stimulation Rate: Using the Nucleus 24 Cochlear Implant System. *Ear Hear,* 21 (6): 608-624.

VERSCHUUR, C. (2009). Modelling the effect of channel number and interaction on consonant recognition in a cochlear implant peak-picking strategy. *J Acoust Soc Am,* 125 (3)**:** 1723-1736.

WARREN, R. M. & SHERMAN, G. L. (1974). Phonemic restorations based on subsequent context. *Perception and Psychophysics,* 16 (1): 150-156.

WANG, M. D. & BILGER, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *J Acoust Soc Am* 54: 1248-1266

WANG, D. L., KJEMS, U., PEDERSEN, M. S., BOLDT, J. B. & LUNNER, T. (2008). Speech perception of noise with binary gains. *J Acoust Soc Am,* 124 (4): 2303-2307.

WANG, D. L., KJEMS, U., PEDERSEN, M. S., BOLDT, J. B. & LUNNER, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *J Acoust Soc Am,* 125 (4)**:** 2336-2347.

WHITMAL, N. A., POSSANT, S. F., FREYMAN, R. L. & HELFER, K. S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *J Acoust Soc Am,* 122 (4): 2376-2388.

WILSON, B. S. (2004). Engineering Design of Cochlear Implants. *In:* ZENG, F. G., POPPER, A. N. & FAY, R. R. (eds.) *Cochlear Implants: Auditory Prostheses and Electric Hearing.* New York: Springer

WRIGHT, R. (1997). Basic Properties of speech. *In*: MARTIN, M. (Ed). *Speech Audiometry* (2nd Ed.), London, Wurr Publishers Ltd.

WOUTERS, J. & BERGHE, J. (2001). Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system. *Ear Hear,* 22(5): 420-430.

YANG, L. & FU, Q. J. (2005). Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *J Acoust Soc Am*, 117(3): 1001-1004.

XU, L. & ZHENG, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *J Acoust Soc Am*, 122 (3): 1758-1764.

YOO, S., BOSTON, J. R., DURRANT, J. D., KOVACYK, K., KARN, S., SHAIMAN, S., EL-JAROUDI, A. & LI, C. C. (2005) speech enhancement based on transient speech information. *In:* IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, 62-65.

YOO, S., BOSTON, J. R., DURRANT, J. D., KOVACYK, K., KARN, S., SHAIMAN, S., EL-JAROUDI, A. & LI, C. C. (2007). Speech signal modification to increase intelligibility in noisy environments. *J Acoust Soc Am*, 122 (2): 1138-1149.

ZENG, F. G. (2004a). Compression and cochlear implants. *In:* BACON, S. P., FAY, R. R. & POPPER, A. N. (eds.) *Compression: From Cochlea to Cochlear Implants.* New York: Springer

ZENG, F. G. (2004b). Trends in Cochlear Implants. *Trends Amplif*, 8 (1): 1-34.

ZENG, F. G. (2004c). Auditory Prostheses: Past, Present, and Future. *In*: ZENG, F. G., POPPER, A. N. & FAY, R. R. (eds.) *Cochlear Implants: Auditory Prostheses and Electric Hearing.* New York: Springer

ZENG, F. G. & GALVIN, J. J. (1999). Amplitude Mapping and Phoneme Recognition in Cochlear Implant Listeners. *Ear Hear*, 20 (1): 60-74.

ZENG, F. G., GRANT, G., SHANNON, R., OPIE, J. & SEGAL, P. (2002). Speech dynamic range and its effect on cochlear implant performance. *J Acoust Soc Am*, 111 (1): 377-386.

ZENG, F. G., NIE, K., STICKNEY, G. S., KONG, Y. Y., VONGPHOE, M., BHARGAVE, A., WEI, C. & CAO, K. (2005) Speech recognition with amplitude and frequency modulations. *PNAS*, 102 (7): 2293–2298.

ZENG, F. G., POPPER, A. N. & FAY, R. R. (eds.) (2004) *Cochlear Implants: Auditory Prostheses and Electric Hearing.* New York: Springer

ZUE, V., GLASS, J., GOODINE, D., PHILLIPS, M. & SENEFF, S. (1990). The Summit Speech Recognition System - Phonological Modeling and Lexical Access. *Icassp 90 (1-5)***:** 49-52.

ZUE, V., GLASS, J., PHILLIPS, M. & SENEFF, S. (1989). The MIT SUMMIT Speech Recognition system: a progress report. *HLT '89 Proceedings of the workshop on Speech and Natural Languag*

Websites accessed

COCHLEAR: HOW IT WORKS, Available: http://www.cochlear.com/wps/wcm/connect/uk/home/discover/cochlear-implants/how-it-works. [Accessed: 24 March 2015].

THE EAR FOUNDATION. Available: www.earfoundation.org.uk/hearing-technologies/cochlear-implants).[Accessed: 01 January 2015].

NATIONAL INSTITUE ON DEAFNESS AND OTHER COMMUNICATION DISORDERS. Available: http://www.nidcd.nih.gov/health/hearing/pages/coch.aspx. [Accessed: 01 January 2015].