UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author KUO, Yu-Chung

COCHLEAR IMPLANTS IN A TONE LANGUAGE

MANDARIN CHINESE

YU-CHING KUO

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

Department of Phonetics and Linguistics

University College London

July 2006

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Cochlear implantation, as a means of restoring hearing to profoundly and totally deaf people, has now become a routine clinical procedure. Implant users can perform remarkably well in many aspects of speech perception. However, current implant devices do not provide all speech information equally well. One major limitation is in providing voice fundamental frequency (f0), known to be problematic even for non-tonal languages in which intonation plays an important role in speech communication. This causes even more difficulty in tonal languages, such as Mandarin Chinese or Thai, in which pitch variations are used to convey lexical meanings. This thesis is mainly concerned with how implant users perceive and use tonal information. Studies were first conducted in normal-hearing listeners to investigate the nature of tone and the importance of voice f0 in understanding running speech. Three acoustic cues (f0, amplitude envelope, and duration) were examined for their contributions to tonal perception in syllables. The results clearly demonstrated voice f0 to be the dominant cue. To determine the effect of explicit f0 in sentence recognition, vocoded stimuli with various degrees of spectral information were presented to four age groups of listeners (aged 6, 9, 12, and 20). Information about natural f0 variations enhanced sentence recognition significantly even when spectral information was severely degraded, and the effect was strong across all ages. The investigation in implanted children first examined tone recognition performance and the acoustic cues used for recognising tonal contrasts, especially the use of amplitude envelope. These implanted children appeared to make some use of amplitude changes in recognising tonal contrasts, though the overall effect was rather small. They also showed some evidence for the use of duration and temporal pitch information. For the effect of f0 in

sentences, no significant difference was found in performance on sentences with their original f0 contours and those with uninformative f0 contours. This indicates that the voice pitch information provided by current implant devices is too limited to allow listeners to take advantage of the presence of natural f0. In the light of the significance of f0 both in signalling tonal identity in syllables and in perceiving sentences, it is likely that implant users will further benefit with a better representation of voice pitch.

Acknowledgements

I would like to acknowledge firstly my primary supervisor, Professor Stuart Rosen, for his tireless help and support from the very beginning throughout the entire period of my PhD. His great gift in explaining complex concepts in plain words, full of inspiring ideas, encouragement, and enthusiastic attitude made the time working with him so enjoyable. I also want to thank my second supervisor, Dr Andrew Faulkner, for his clever comments all the time and kindly assistance whenever I needed help.

I wish to thank all the implanted children and their parents I have worked with, especially those who attended the research presented in this thesis. I still remember the cheerful but monotone voice that a little implanted girl used to greet me, 'Hello, Miss Kuo', every time she walked into the therapy room when I worked as a speech therapist in the hospital many years ago. My wondering about her lack of intonation and what we could do about it inspired this work.

This work would hardly be possible to complete without the support of my husband Shih-kuen, who has always been so patient and encouraging me to fulfill my dream. I also thank my mum, Ni-Wen, Hsu-Yin, Bing-Zang, and my parents in law for their understanding and constant support all the way. I would also like to thank Claudia and her family for warmly welcoming me into their house. While I thought I had found a place simply to stay in London for the period doing writing-up, instead I found another home.

Lastly, I would like to thank the Chiang Ching-kuo Foundation for International Scholarly Exchange (Taiwan) for the Ph.D. dissertation fellowship, and the generous assistance of Defeating Deafness (UK) for traveling to conferences. Thanks also go to the department of Otolaryngology at Chi-Mei Medical Centre, the department of Otolaryngology at National Cheng Kung University Hospital, and the Children's Hearing Foundation, for the assistance in subject recruitment.

Contents

Figures

Tables

Chapter 1

General introduction

1.1 Cochlear implants and tone languages

In the normal-hearing ear, sound is transmitted through the middle ear to the inner ear, and then causes vibrations on the basilar membrane. The hair cells in the cochlea translate mechanical vibrations of the basilar membrane into electrical potentials. The majority of profoundly deaf people who can not benefit from traditional hearing-aids suffer from the absence or degeneration of the sensory hair cells in the inner ear. Cochlear implant systems bypass the damaged hair cells, stimulating the auditory nerve directly by electrical current. These devices have been used widely in profoundly hearing-impaired people around the world, and have helped them in the successful restoration of some hearing. Today, cochlear implant systems have been the most successful sensory prosthesis. Many implant users are able to communicate with auditory information only, even carry on a telephone conversation.

Although there has been great progress since cochlear implants were introduced, there is still much room for improvement. One essential limitation in current implant devices is in providing voice f0 information, which is particularly important for users of a tone language. The distinctive feature of tone languages is the use of pitch pattern to convey lexical meanings. The same syllable with different pitch contours represents different words. For instance, in Mandarin, the syllable *yi* produced with constant, rising, falling-rising, and falling pitch contours represents 'one', 'to move', 'chair',

and 'meaning' respectively. Studies in Mandarin and Cantonese have shown that implant users seem able to obtain some information about tonal contrasts, though the benefit remains quite limited. While many implant users have difficulty in perceiving tonal contrasts, a few implant users perform surprising well. This indicates that some information other than voice pitch might be used by those extremely good performers, which can be transmitted relatively well by cochlear implants. With the better understanding the nature of tone, and how tonal information is transmitted by cochlear implants, we may be able to help those with a low level of performance.

1.2 The aims of the thesis

The studies reported in this thesis address the following three issues related to the use of cochlear implants in tone language users:

(1) The extent to which tonal contrasts can be recognised on the basis of different acoustic cues.

While the four tones in Mandarin are mainly distinguished by their f0 patterns, they can also be identified correctly to a certain extent by other acoustic cues. Previous studies have shown that amplitude envelope and duration tend to vary systematically with tonal identity. The contributions of the three main cues to Mandarin tone (f0, amplitude envelope, and duration) were examined using simplified stimuli with all combinations of one, two and three of these cues.

(2) The extent to which tonal contrasts in a single syllable can be perceived by implanted children, and whether the amplitude envelope and duration help in this.

Given that only very weak information about f0 is provided, and that other acoustic cues, amplitude envelope and duration, can be signalled by cochlear implants, it is likely that implanted children would make the most of them. The effect of amplitude envelope was first investigated in a group of implanted children with different speech coding strategies. The effect of duration was further controlled in a sub-experiment and examined in 2 implanted children.

(3) How important is voice pitch information for speech understanding in a tone language? How much can implant users take advantage of natural f0 variations?

The effect of voice f0 on sentence recognition was examined in normal-hearing listeners using vocoder techniques and in implanted listeners. The direct manipulation of pitch contours in sentences allowed us to examine the extent to which voice f0 variations could be used to enhance speech understanding.

1.3 Overview of the thesis

This thesis is organised into six chapters. An introduction about cochlear implants is given first in chapter 2. The experiments conducted in this thesis are reported in chapters 3, 4, and 5; each is a self-contained paper. The last chapter then gives a summary and general discussion.

Chapter 2 - Literature Review is a review chapter giving a general introduction to cochlear implants. This chapter includes (1) the cochlear implant system, (2) a summary of different speech coding strategies that are used to translate acoustic signals into electrical stimulations, (3) a summary of previous studies focusing on several variables possibly related to implant performance, and (4) limitations of current implant devices.

Chapter 3 - Acoustic Cues to Tonal Contrasts examines the nature of tonal contrasts in Mandarin. Simplified stimuli were imposed with f0, amplitude envelope, and duration in isolation, in pairs, or all together to examine their contributions to tonal contrasts. The f0 information was represented both by a clear indication with a sawtooth carrier created period by period to match the f0, and by temporal fluctuations using a noise carrier.

Chapter 4 - Effect of Voice F0 in Sentence Recognition. This chapter investigates the effect of voice f0 in sentences using vocoder techniques. The f0 information was manipulated by using either an f0-controlled pulse carrier or a pulse carrier with a slightly falling pitch contour for voiced speech. The number of frequency channels was also varied to examine the effect of natural f0 with various degrees of spectral information.

Chapter 5 - Speech Perception in Implanted Children. The experiments carried out in this chapter attempt to investigate acoustic features used to recognise tonal

contrasts, and to perceive sentences in the presence/absence of natural f0 variations. The perception of tone was first examined in twenty-one implanted children, considered 'good' performers, and aged between 6 and 15. Only tones 1, 2, and 4 (level, rising, and falling) were explored here. Tone 3 was excluded due to its variable realisation. Stimuli with and without natural amplitude variations were used to investigate the use of amplitude envelope. Both amplitude envelope and/or duration were further investigated in a sub-experiment conducted in two ACE users.

In the sentence recognition study, sentences with natural f0 contours were compared to sentences whose contours had been manipulated to be slightly falling. The results from implant users were compared with the data from normal-hearing subjects listening to vocoded speech in chapter 4. The comparison between the results from implant users and the data from normal-hearing subjects listening to vocoded speech can be used to indicate the benefit from current devices and what can be achieved if sufficient information about f0 was provided.

Chapters 6 Summary and General Discussion summarises the main findings reported in this thesis and further discusses the implications of the current study, especially for prelingually implanted children during the process of spoken language acquisition.

Chapter 2

Literature review: cochlear implants and speech perception through an implant

The literature review in this chapter will give a general introduction of cochlear implants which is relevant to the thesis as a whole. More detailed reviews specifically related to each of three main experiments will be given at the start of Chapters 3, 4, and 5.

This chapter introduces the multi-channel cochlear implant system first, and then reviews the speech coding strategies used in early devices and those available in current implant systems. The following section discusses several variables which have some effects on speech performance in implant users, including the number of channels, stimulation rate, and insertion depth. Finally, limitations in current implant systems and future directions are discussed.

2.1 A cochlear implant

The multiple-channel implant system comprises a set of external components and a set of internal components (Figure 2.1.A). The external components include a directional microphone, a speech processor, and a transmitter coil. The internal components are placed inside the body by surgery, including a receiver- stimulator and an electrode array. Figure 2.1.B illustrates where these components are placed on implanted patients. Sounds are first picked up by the microphone, and sent to the

speech processor, in which incoming sounds are analysed and converted into electrical signals. The electrical signals are transmitted by radio waves from the transmitter coil to the receiver under the skin. This signal is then decoded to determine the electrical currents sent to individual electrodes.

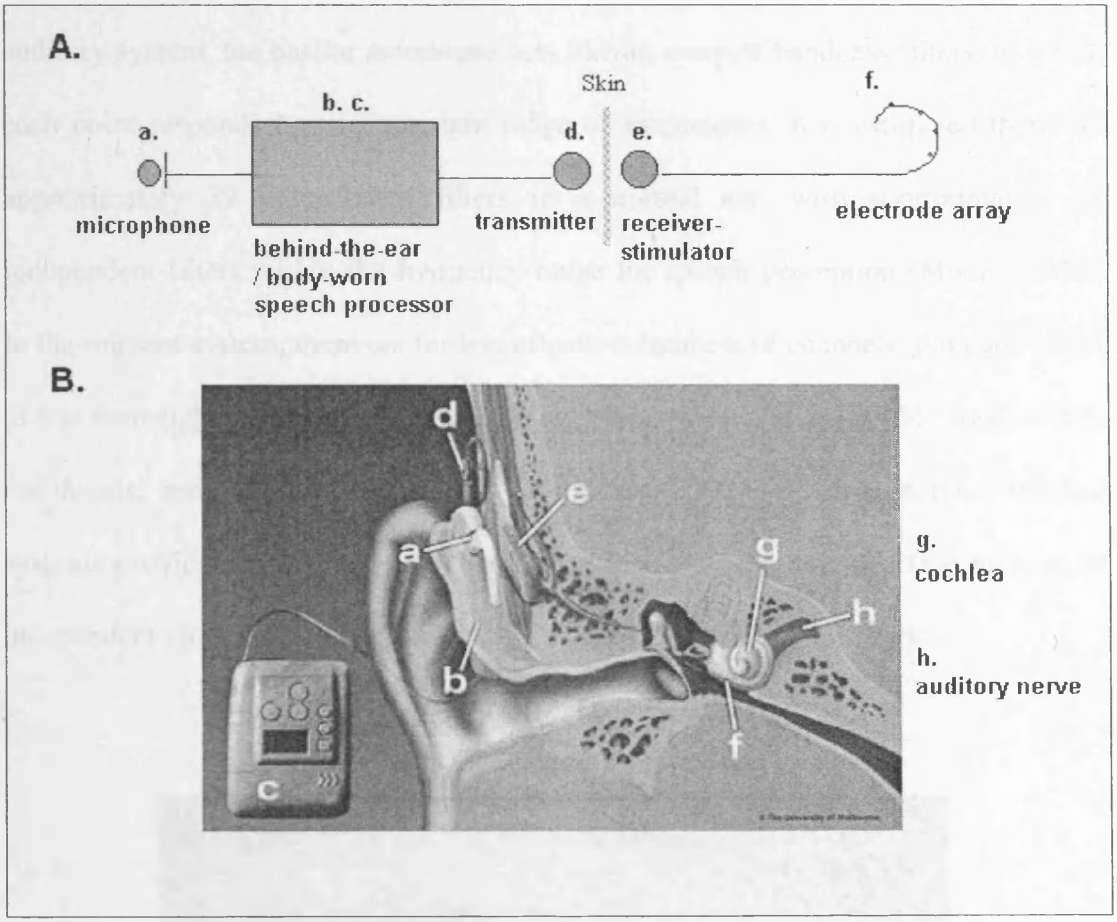

Figure 2.1.A: Components of the multi-channel cochlear implant system (modified from Loizou, 1998). B: An implant system and its placement in an implant user (from the website of the Bionic Ear Institute; the implant device shown here is the Nucleus device).

The electrode array is inserted through the round window into the scala tympani, and placed around the first turn of the cochlea. Figure 2.2 shows an electrode array placed in the cochlea. To mimic the place mechanism for coding frequencies in the

cochlea, electrodes near the base are stimulated with high frequency signals, and those near the apex stimulated with low frequency signals. Despite individual differences, the majority of implant users demonstrate a tonotopic order of pitch percepts as in the normal ear (e.g. Busby *et al.*, 1994; Busby & Clark, 2000; Nelson *et al.*, 1995; Tong *et al.*, 1982; Tong & Clark, 1985). However, the place coding in the implant systems is not comparable to frequency resolution in the normal auditory system. In the auditory system, the basilar membrane acts like an array of bandpass filters, in which each point responds best to a certain range of frequencies. It is estimated there are approximately 39 independent filters in a normal ear, with approximately 28 independent filters within the frequency range for speech perception (Moore, 2003). In the implant system, there are far less effective numbers of channels, typically about or less than eight (e.g. Fu, Shannon & Wang, 1998; Friesen *et al.*, 2001; Moore, 2003; for details, see 'number of channels' in session 2.4). Even though most implant systems provide more frequency channels in their devices, the effective number of independent channels is often limited by the interaction among electrodes.

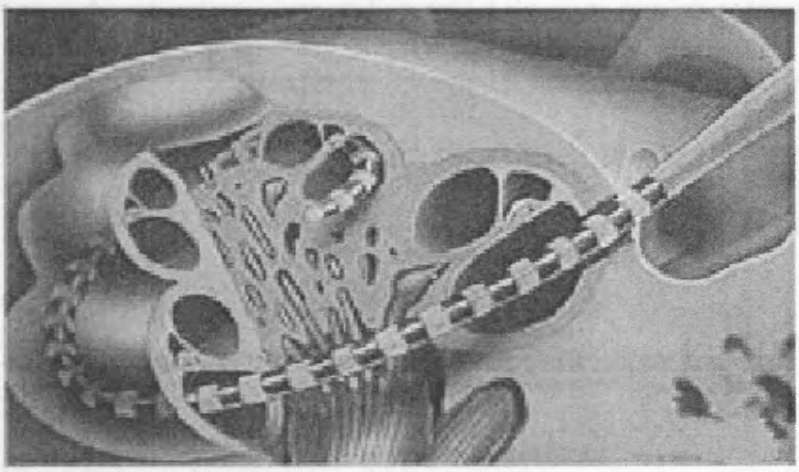

Figure 2.2 An electrode array inserted into the cochlea.

Some of the latest implant devices use a perimodiolar electrode array (or so-called 'modiolus hugging'). These modiolus hugging electrodes, instead of lying anywhere in the scala tympani, are designed to be closer to the nerve fibres. Therefore, it is expected that lower current levels are required for the implant with perimodiolus electrodes, and that the interaction between electrodes will be reduced as well (Gstoettner *et al.*, 2001; Tykocinski *et al.*, 2001). Animal experiments (Shepherd *et al.*, 1993) and computational models (Frijns *et al.*, 1995; Frijns *et al.*, 1996) both confirm that the electrode position in the scala tympani has some effect. Studies have reported some advantages for perimodiolar electrodes, especially in the basal turn (e.g. Tykocinski *et al.*, 2001; Frijns *et al.*, 2001). This is due to the fact that in humans the distance from the medial wall of the scala tympani to the nerve bundle in the modiolus is much larger in the basal turn than in the middle and apical turns.

Figure 2.3 shows the place coding of frequency information in the normal auditory system, and the number represents the frequency which each point along the basilar membrane responds best. As mentioned before, an electrode array is often inserted around the first turn of the cochlea, therefore, the most apical electrode might be located at the place normally responding to 1 kHz or even higher frequencies (e.g. Ketten *et al.*, 1998). The place/frequency mismatch would cause speech sounds to be up-shifted to a higher frequency range, and became less intelligible (e.g. Dorman *et al.*, 1997b; Shannon *et al.*, 1998). The effect of frequency up-shifting will be discussed more in 'insertion depth' in section 2.4.

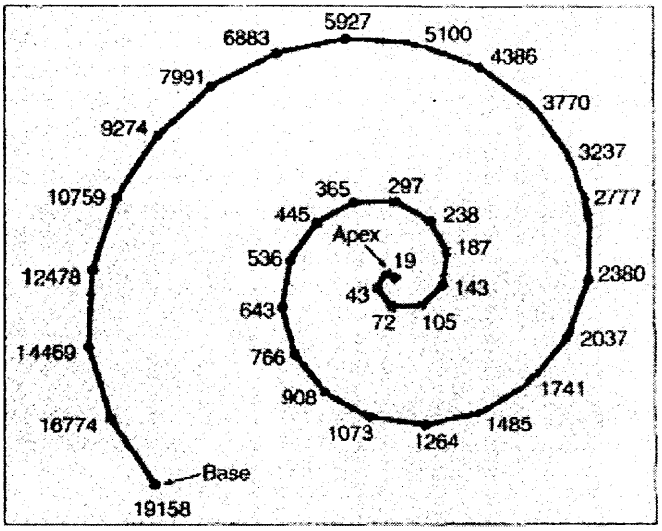

Figure 2.3 Frequency encoding in the cochlea. The numbers represent the frequency of a sinusoid that each place along the basilar membrane responses the best (resulting in the maximum displacement) (from Loizou, 1998).

2.2 Speech coding strategy

Table 2.1 summarises speech coding strategies for representing speech information in different multi-channel cochlear implants

Strategy	Nucleus	Med-El	Clarion	Ineraid
feature-extraction	F0/F2 F0/F1/F2 MPEAK			
filterbank	SPEAK(SMSP) ACE CIS	CIS n-of-m	CIS HiRes	
analog			CA SAS	CA

Table 2.1 Summary of speech coding strategies used in different implant systems.

Feature extraction approach

F0/F2, F0/F1/F2, MPEAK

Feature extraction strategies were used in the early versions of the University of Melbourne /Nucleus Limited multi-channel implant system (Clark, 1987; Patrick & Clark, 1991). To avoid sensation overload in the brain, this approach extracted essential speech features, such as formant frequency and voicing/fundamental frequency. Instead of transmitting all the information in speech, only simple patterns of stimulation were presented. The initial version of feature extraction strategies was ***F0/F2***, which transmitted only voicing information and the frequency and amplitude of the second formant (F2). The frequency of F2 was used to select the electrode to stimulate on a place coding basis, and the amplitude of the spectral peak of the F2 was used to determine the current level used to stimulate the electrode. F0 was used to determine the stimulation rate for voiced speech, with one pulse presented in each period. A quasi-random low frequency rate, averaging around 100 Hz, was used for voiceless speech. Some of F0/F2 users were able to perform open-set speech recognition task with hearing alone (Clark, 1987; Patrick & Clark, 1991).

The next feature-extraction strategy was ***F0/F1/F2***, which further included information for the first formant (F1). Similar to the F0/F2, a low pulse rate was used for voiceless speech, and a rate equal to f0 was used for voiced speech except with two pulses presented in each period. The frequencies of F1 and F2 were used to select two stimulating electrodes, with lower frequency (F1) in a more apical position and higher frequency (F2) in a more basal position. The later strategy, ***MultiPeak (MPEAK)***, added more information about high frequency. Three high frequency bands (bands 3, 4, and 5 with frequency range at 2-2.8kHz, 2.8-4 kHz, and 4-6k Hz

respectively) were stimulated on three fixed electrodes at the most basal end. Four pulses were delivered in base-to-apex order to the four electrodes for F1, F2 and bands 3 and 4 at the rate of f0 during voiced speech, and to four electrodes for F2 and bands 3, 4, and 5 at a quasi-random rate between 200 -300 Hz during the voiceless speech.

In general, implant users showed significant improvement in their speech performance with the later versions of feature extraction strategies (MPEAK > F0/F1/F2 > F0/F2) (e.g. Dowell *et al.*, 1987a; 1987b; Tye-Murray *et al.*, 1990; Dowell *et al.*, 1991; Skinner *et al.*, 1991, Whitford, 1993; Clark *et al.*, 1997). A major problem of these feature extraction strategies was error when extracting f0 and formant information, especially in noise (Loizou, 1998). Although implant users with these strategies achieved some substantial performance, the feature-extraction strategies were gradually superseded by filter-bank strategies.

Filter-bank approach

CIS (Continuous Interleaved Sampling) strategy

In the CIS strategy, input signals are analysed with a fixed number, usually 6-12, of bandpass channels. In each channel, the amplitude envelope is extracted by rectification and low-pass filtering, and then used to modulate high frequency pulse trains. The cut-off frequency of the low-pass filter for envelope smoothing is typically at 400Hz, and the frequency of pulse trains in each channel is normally over 800Hz. Figure 2.4 shows examples of two speech waveforms, a voiced sound /ɔ/ and a voiceless sound /t/, and the output of each channel after passing through a four-channel CIS processor. To mimic the tonotopic organisation in the cochlea, the

modulated pulse trains derived from lower frequency channels are directed to electrodes near the apex and those derived from higher frequency channels are directed to electrodes near the base.

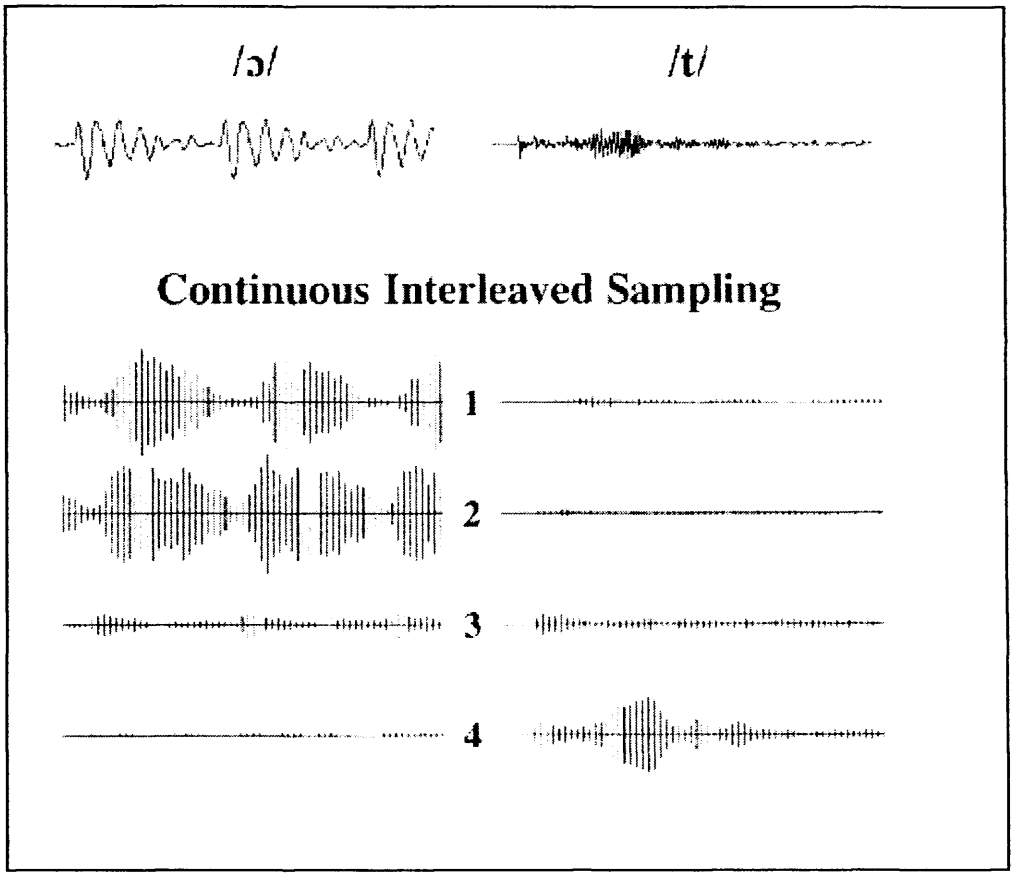

Figure 2.4 Output waveforms for a voiced speech sound /ɔ/ and a voiceless speech sound /t/ after passing through a four-channel processor with the CIS strategy. Original speech waveforms are shown in the top panel, and pulses from each of the four channels produced by a CIS processor are shown in the remaining panels. The numbers 1 to 4 indicate from the most apical electrode to the most basal one (Adapted from Wilson et al., 1991).

Two essential features of the CIS strategy are: (1) interleaved nonsimultaneous stimulation, in which pulses are present on only one electrode at any time, so as to minimize interactions among channels (This feature is also used in 'n-of-m', ACE, and SPEAK strategies); (2) a relatively high stimulation rate on each channel (normally over 800 Hz), which allows the preservation of rapid temporal variations in speech (Wilson *et al.*, 1991; Wilson, 1993).

n-of-m and ACE (Advanced Combination Enbcoders) strategies

Another popular speech coding strategy in current implant devices is n-of-m strategies (Wilson *et al.*, 1988). In the 'n-of-m' strategy, speech is divided into a relative large number of band-pass filters (*m*) and only a small number of bands (*n*), typically 6 to 8, with maximum amplitude is chosen, with the electrodes corresponding to spectral peaks stimulated. Theoretically, the 'n-of-m' strategy would be an excellent method for coding speech signals. This peak-picking approach mimics the response of the basilar membrane to spectral peaks, and therefore is expected to provide relatively better frequency resolution than the CIS strategy. In CIS, all the electrodes are stimulated, and spectral peaks must be inferred from the relative amplitudes of adjacent electrodes. This peak-picking approach is implemented in Med-EL devices and as the SPEAK and ACE strategies in Nucleus devices. The design of n-of-m and ACE strategies are very similar, but the SPEAK strategy is different in some aspects and will be described later.

One essential issue for the n-of-m strategy would be the choice of the numbers for *m* and *n* to achieve the best performance. In the Nucleus-24 systems, the m is typically 20 and the n is 6-10, while in the Med-El Combi 40+ systems, 6 channels are typically selected out of 12 analysis channels. However, whether this would lead to the best performance for implant users remains unclear. In theory, the ***m*** should lead to filters that are sufficiently dense, and the ***n*** should be sufficient to include essential information but not too redundant, so as not to reduce overall stimulation rate. The choice of the number for *n* is important due to the limited capacity of implant devices in terms of overall pulse rate. There is a trade-off between the maximum pulse rate per channel and number of stimulated channels (greater number of channels gives a

lower stimulation rate in each channel, whereas smaller number of channels allows a higher stimulation rate per channel).

Results from acoustic simulations on normal-hearing listeners have demonstrated that high levels of speech perception performance can be achieved with a small number of channels which have maximum outputs from 16-20 analysis channels. For instance, recognition scores for NU6 words were 73, 87, and 90% with 2, 4, and 6 channels, out of 20, respectively (Dorman, 2000).

Speech performance with the n-of-m/ACE and CIS strategies was generally similar, or sometimes better with the n-of-m strategy (e.g. Arndt *et al.*, 1999, cited in Clark, 2003; Kiefer *et al.*, 2001; Lawson *et al.*, 1996; Ziese *et al.*, 2000). For instance, Arndt *et al.* (1999, cited in Clark, 2003) reported results from a group of post-lingually deafened adults which showed that speech perception with the ACE strategy was significantly higher than the CIS, and no difference between ACE and SPEAK.

SPEAK (Spectra Peak) or SMSP (Spectral Maxima Sound Processor) strategy

The SPEAK strategy adaptes the n-of-m approach and is implemented in the Nucleus-22 devices. The input signals are filtered into 20 frequency channels, and the envelope of each channel is derived as in CIS and n-of-m, except extracted at 200Hz. Six to ten channels with maximum energy are selected. The stimulation rate is between 180 and 300 pps, depending on the number of electrodes selected. For instance, a 6-channel stimulation would have a stimulation rate of about 400 pps, which is still much lower than those in the CIS or ACE strategies. The use of a relatively low stimulation rate means that aliasing and other distortions not be prevented in SPEAK.

CA and SAS strategies

In the CA strategy, speech sounds are compressed to a narrow dynamic range firstly, then filtered into a small number of frequency bands, normally 4 to 6, and presented simultaneously to the electrodes. Figure 2.5 shows outputs processed by a CA processor. The major difference of the CA strategy from the CIS strategy and other strategies described above is the use of analog (or continuous) waveforms, instead of interleaved pulses (see both Figure 2.4 and 2.5 for the comparison of CA and CIS). Spectral and temporal information is provided through both the amplitude changes across channels and the temporal variations with channels. Therefore, more speech details are provided by the CA strategy. Although there are more temporal details provided in CA, implanted listeners may not be able to use this temporal information (Wilson *et al.*, 1990; Wilson, 1993). Several studies have shown that most implant users could only perceive changes in frequency up to about 300Hz (e.g., Shannon, 1983; Tong *et al.*, 1982; Zeng, 2002). Furthermore, simultaneous stimulation may cause more channel interactions and distort spectral cues for speech. The SAS strategy is adapted from the CA strategy (Eddington, 1980; Merzenich *et al.*, 1984), but with much improvement on some major limitations found in the CA strategy, such as channel interaction and speech distortion (e.g. Zwolan *et al.*, 2001; Frijns *et al.*, 2002; Wilson, 2004).

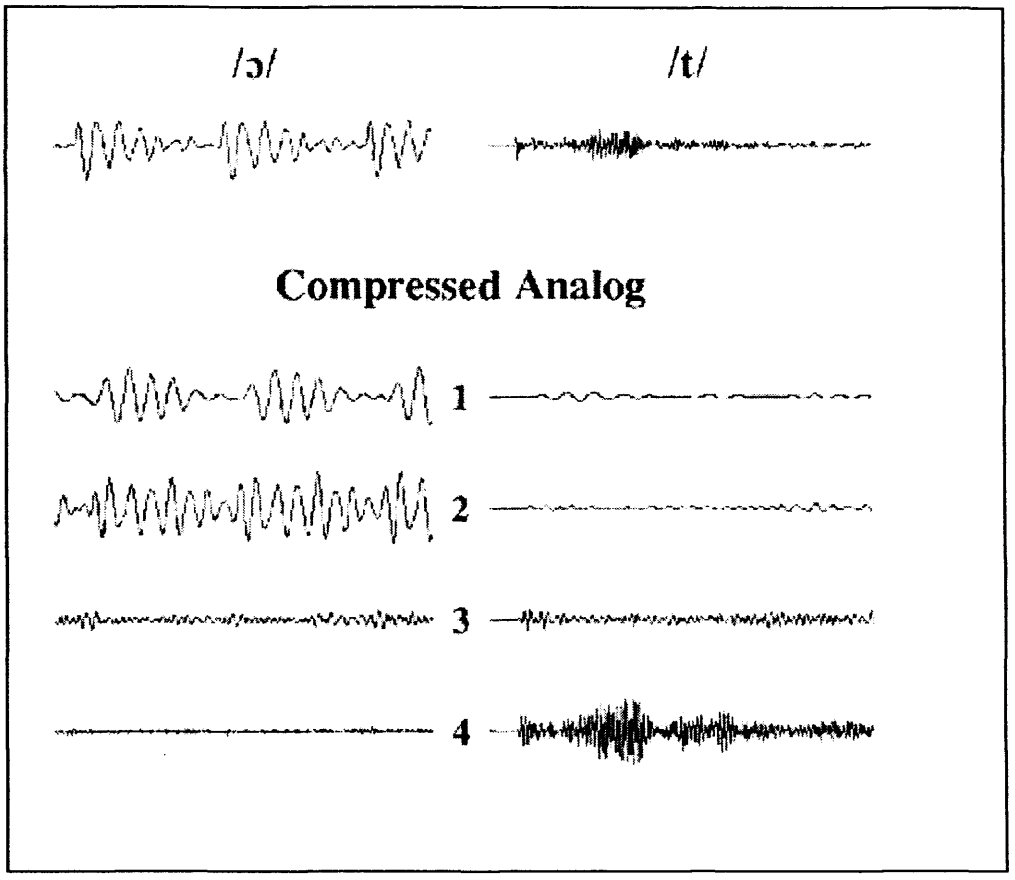

Figure 2.5 Output waveforms for a voiced speech sound /ɔ/ and a voiceless speech sound /t/ after passing through a four-channel processor with the CA strategy. This figure is organised in the same way as figure 2.4 for the CIS strategy (original speech waveforms shown at the top and outputs of the four channels are shown below. Number 1 represents the lowest frequency channel, and number 4 represents the highest frequency channel) (Adapted from Wilson et al., 1991).

Early studies comparing the CA and CIS strategies consistently reported significantly better performance for CIS than for CA (e.g. Wilson *et al.*, 1991; Boex *et al.*, 1996; Dorman & Loizou, 1997; Pelizzone *et al.*, 1999). However, recent comparisons of these two strategies by users of the Clarion devices showed variable results across studies (Battmer *et al.*, 1999; Osberger & Fisher 2000; Zwolan *et al.*, 2001; Frijns *et al.*, 2002). For instance, Battmer *et al.* (1999) found that Clarion users had comparable performance for the CIS and SAS strategies with no overall preference for either strategy. Osberger & Fisher (2000) reported that, although fewer implant users preferred SAS, some users had superior performance with it. Another

study by Zwolan *et al.* (2001), comparing SAS with CIS and PPS[1] (Paired pulsatile stimulation, a variation of the CIS), found that more users of the Clarion CI implant preferred CIS while more users of the Clarion CII implant preferred SAS. This was possibly due to the use of the 'modiolus hugging' electrode in the CII device. Because of a closer position of electrodes to the inner wall, less electrode interactions were expected, and might allow the simultaneous strategy, SAS, to function better. However, the preference of SAS for CII users was not shown in the study by Frijns *et al.* (2002), who reported nine out of ten CII users preferred CIS, one user preferred PPS, but none of them preferred SAS.

'HiRes' strategy

HiRes is a variation of CIS, and has been implemented in the Clarion CII system. In comparison with CIS typically using 8 channels and a carrier frequency at around 800 pps per electrode, HiRes can support up to 16 channels, and present pulses at a rate of up to 90,000 across electrodes. The HiRes is typically fitted with 16 channels and a carrier rate of 2800-5600 pps per electrode. A clinical trial compared HiRes with the preferred strategy (CIS, PPS, or SAS) for 3 months, and reported a significantly higher performance for HiRes on monosyllabic words, CID and HINT sentences (Osberger *et al.*, 2002, cited in Wilson, 2004). The greatest improvement was for performance in noise, with an increase of mean score from 47% to 61%. However, as Wilson (2004) pointed out, since all subjects used the strategy they preferred first for 3 months and then the HiRes later for another 3 months, it is also

[1] PPS (Paired pulsatile stimulation), also called MPS (multiple pulsatile sampler), is a variation of the CIS strategy. This strategy doubles the stimulation rate but minimizies electrode interaction by stimulating pairs of distant electrodes simultaneously, with stimulation of the pairs in a non-simultaneous sequence.

possible that the initial experience with the prior strategy favoured HiRes in this study.

Another study by Frijns *et al.* (2003) compared typical CIS processors using 833 pps per electrode with processors using 1400 pps per electrode. Typically 8-channel was used for the former processors with lower rate, and 8-, 12-, and 16-channel was fitted for the later trial processors. Results showed a great variation between subjects. While some implanted listeners achieved the highest recognition scores with the higher rate and 8-channel, others had the best performance with the higher rate and 12- or 16-channels. Overall, consistent with the study by Osberger *et al.* (2002, cited in Wilson, 2004), a higher stimulation rate generally have shown a significant improvement in subject performance.

2.3 Pitch perception in acoustic and electric hearing

Pitch and pitch coding in the normal auditory system

Many models have been proposed to account for how normal-hearing listeners perceive the pitch of complex sounds such as speech. However, it is still not yet completely clear about how a pitch percept is evoked in the auditory system. For a pure tone, pitch is closely related to its frequency. A sinewave creates a maximum excitation at a certain place along the basilar membrane corresponding to its frequency. This, known as tonotopic organisation, formed the basis of the classic ***place theory***. Place coding generally works well for a pure tone, but has difficulty in accounting for a complex sound. The pitch of a periodic complex sound is generally related to its fundamental frequency. A complex sound would create excitation

patterns that show a distribution of many maxima corresponding to frequency components of the complex sound. However, the place with the maximum displacement is not always the place corresponding to the frequency of the fundamental. Furthermore, the existence of the fundamental frequency is not necessary for pitch to be perceived. This is known as the "phenomenon of the missing fundamental". Pitch perception remains unaltered for a complex sound even if its fundamental frequency is eliminated, despite a change in the timbre of the sound.

Several models have been proposed to account for the pitch of complex sounds. The early models may be grouped into two classes: temporal models and pattern recognition models. In the ***temporal models***, the pitch of a sound is related to the time pattern of the neural response evoked by the incoming sound (e.g. Schouten *et al.*, 1962). Neurophysiologic studies have shown that nerve spikes tend to synchronize with a particular phase of the stimulating waveform (phase locking). The time intervals between successive spikes approximate integer multiples of the period of the input waveform, and therefore, the time intervals of the neural spikes could be used to code pitch information.

Figure 2.6 illustrates the response of the basilar membrane to a periodic complex sound. The input is a pulse train with a 200-Hz repetition rate. The output waveforms show the responses observed at different places along the basilar membrane. The lower harmonics, at least up to the first five or so, are effectively resolved, or separated. Each of the lower harmonics produces a local peak at the place where the neurons have their characteristic frequencies, close to the integer multiples of the fundamental frequency. The timing of the neural spikes is related to the frequency of the individual harmonic, rather than to the fundamental frequency of the whole waveform (For instance, the interspike intervals in 400 and 600 Hz places are close to

integer multiples of 2.5 and 1.67 ms, respectively). For the places responding to higher frequencies, the harmonics are not resolved. The excitation pattern does not show local peak corresponding to individual harmonics. The waveforms at the higher frequency ranges results from the interaction of several harmonics, but with the same periodicity as the input waveform. The timing of neural spikes in these regions is also related to the fundamental frequency of the input signal. In the temporal models, the pitch of a complex sound is derived from the time intervals between successive spikes at a place on the basilar membrane responding to the higher frequency ranges where harmonics are interfering with each other and not resolved.

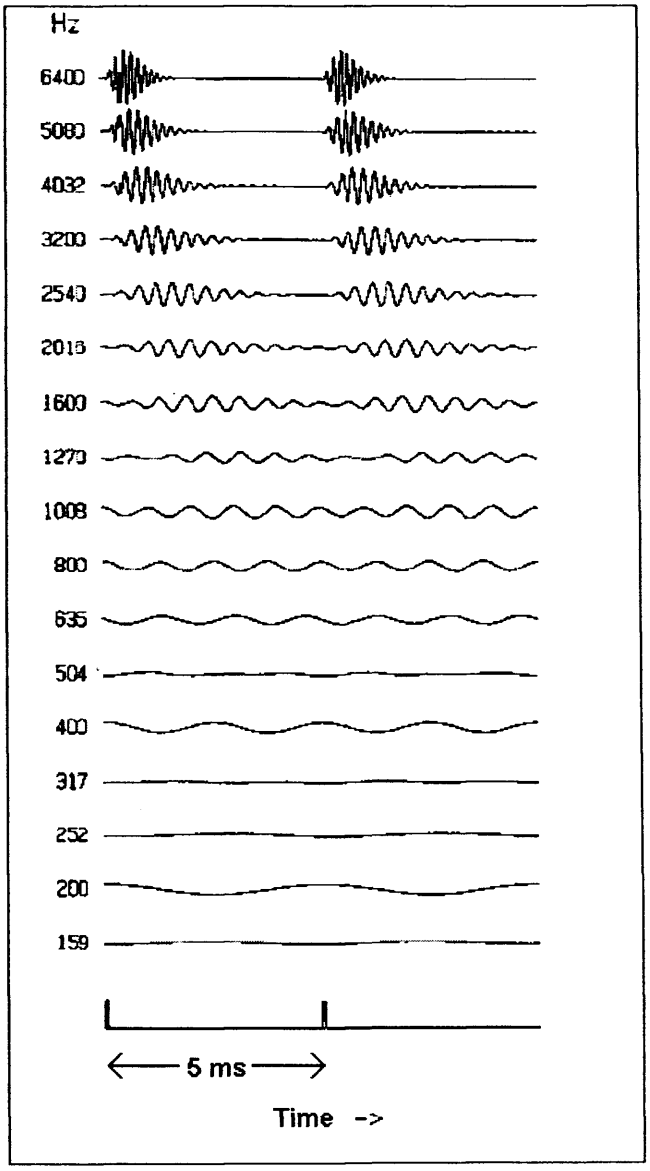

Figure 2.6 Simulation of the response on the basilar membrane to periodic impulses at a rate of 200 pulses per second. The waveforms represent what would be observed at different points along the basilar membrane. The numbers on the left represent the frequencies which result in maximum displacement on the basilar membrane. The input pulses are shown on the bottom (modified from Moore, 2003).

Another class of pitch models, the ***pattern recognition models***, focus on the pattern of the frequencies extracted from the incoming sound (e.g. Terhardt, 1974; Goldstein, 1973). In these models, there are two stages involved in determining the pitch of a complex sound: a frequency analysis is conducted first to determine the frequencies of some of the individual frequency components in a complex sound; then

a central auditory mechanism operates on neural signals to determine the pitch of the complex sound from the frequencies of the resolved components. The way to compute the pitch is possibly by trying to find a fundamental frequency with harmonics matching the frequencies of those resolved components as closely as possible (For instance, for a complex sound containing individual components at 1800, 2000, and 2200Hz, the best-fitting fundamental frequency is 200Hz). The pattern recognition models assume the pitch of a complex sound is derived from the pitches of the individual harmonics, and require at least some individual harmonics to be resolved. For these models, lower resolved harmonics are essential in determining pitch information.

The temporal and pattern recognition models both are supported by some experimental evidences, but none of them can explain all the experimental data. For instance, psychoacoustic studies have shown that a stronger and more salient pitch is heard when lower resolved harmonics are present (Moore & Peters, 1992; Plomp, 1967; Ritsma, 1967). Ritsma (1967) reported that the 3^{rd}, 4^{th}, and 5^{th} harmonics tend to dominate pitch perception. The dominance of lower resolvable harmonics for determining pitch supports the pattern recognition model. On the other hand, the vibration patterns resulting from the higher unresolved harmonics would also provide sufficient information about temporal periodicity. Moore and Rosen (1979) found that the pitch of a periodic pulse train remains the same even when all lower resolved harmonic are removed. The pattern recognition models can not account for the pitch derived from higher harmonics that are not resolved, while the temporal models can explain this well.

A number of models incorporate the features of temporal and pattern recognition theories (e.g. Assmann and Summerfield, 1990; Meddis and Hewitt, 1991; Moore,

2003). These ***combined models*** assume information both from resolved and unresolved harmonics is used in determining the pitch of a complex sound. Figure 2.7 illustrates one of such models proposed by Moore (2003). The input signal first passes through a bank of bandpass filters with overlapping passbands. The outputs of the auditory filters would be as the waveforms shown in Figure 2.6. The lower harmonics are resolved, and the higher harmonics show complex waveforms but with a repetition rate corresponding to that of the input signal. In the next neural transduction stage, the outputs of bandpass filters evoke neural spikes in neurons with corresponding CFs. The firing pattern in neurons reflects the temporal properties of the stimulating waveform (for instance, the interspike intervals for a 400 Hz harmonic are multiples of the period of the 400 Hz sound, i.e., 2.5, 5.0, 7.5 ms and so on). The temporal pattern of neural activity in each channel is then analyzed. The next stage compares the intervals between successive neural impulses measured in different channels and searches for common time intervals. Finally, the prominent time intervals present across channels are fed to a decision mechanism in which one interval is then selected. The reciprocal of the final interval selected determines the pitch which has been perceived.

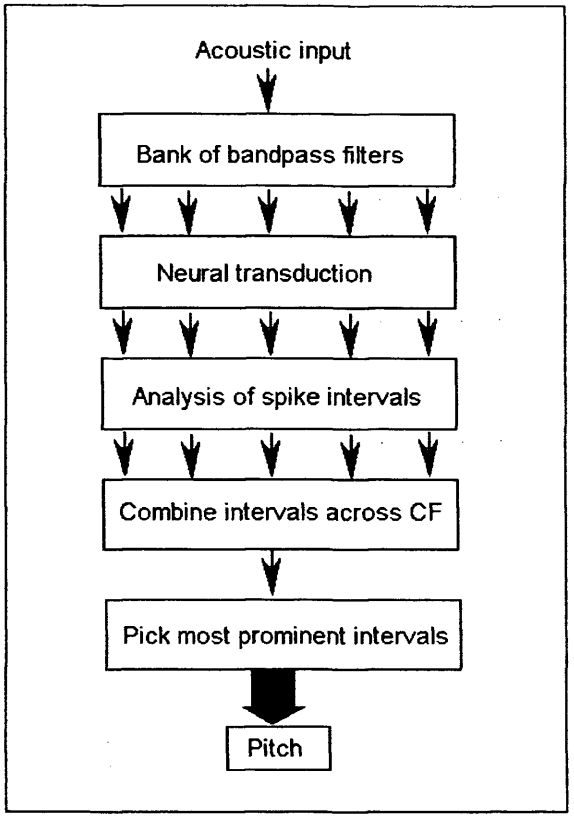

Figure 2.7 A schematic model for the pitch of a complex sound (modified from Moore, 2003).

Other models propose different methods for computing the pitch value. Some combined models are specified more quantitatively. For instance, the model proposed by Meddis and Hewitt (1991) uses an autocorrelation function for determining the pitch of the incoming signals. The first two stages are similar to those in Moore's model; the sounds first pass through a bank of bandpass filters, and the mechanical movement is converted into neural firing. An autocorrelation function of the neuron firing probability is generated for each channel. Then, the autocorrelation functions from different channels are summed up. The pitch is determined by the peak in the summary autocorrelation function. In general, these combined models are able to account for many important features of pitch perception: for instance, the dominance

of lower resolved harmonics in pitch perception and the relatively weak pitch percept evoked from higher unresolved harmonics only.

Pitch and pitch coding in cochlear implant systems

In general, implant users exhibit a tonotopic aspect of pitch - pitch is perceived as sharp to dull when the place of stimulation varies from the base to the apex. However, as described in section 2.2, the place pitch perceived by implanted listeners is not comparable to that perceived by normal-hearing listeners. This is mainly due to the limited number of effectively channels in implant devices (typically about 8, or less, in implanted system, with approximately 39 in the normal hearing system).

For the pitch of a complex sound, the information from resolved harmonics, which is used primarily for determining pitch in normal hearing, is not available for implanted listeners. In most multichannel implant systems, speech signals are analysed into a number of frequency channels, and the amplitude envelope in each channel is extracted and used to modulate a pulse train carrier. Since there is only a relatively small number of effectively channels in implant devices, the lower harmonics of input signals usually are not resolved. Therefore, implant users almost entirely rely on the temporal periodicity of unresolved harmonics to extract pitch information. The pitch perceived by implanted listeners might be seen as what normal-hearing subjects could perceive when listen to a sound containing only higher harmonics. The pitch may still be perceived, but is rather weak.

2.4 Variables affecting speech performance of implanted listeners

System variables

In current implant systems, there are many parameters which can be adjusted easily to achieve the optimal performance for individual implant users. Here, some factors which have been investigated in numerous studies and reported to be associated with speech perception will be described (number of channels, insertion depth, and stimulation rate).

Results from clinical studies often show a large variation in the speech perception performance among implant users, ranging from nearly zero up to almost perfect. There are many factors which may affect subject performance, such as the number of surviving neurons, the depth of electrode insertion, previous hearing or training experience etc. These individual differences make it hard to assess the effect of a particular factor. The use of acoustic simulations in normal-hearing listeners allows at least some of these individual factors (e.g. neuron survival) can be controlled, so that the effect of other variables to be investigated (e.g. the number of frequency channels or insertion depth). Although results of simulation studies can not reflect exactly how implant users behave, they do provide us with some idea about the amount of information implanted listeners may obtain from their implants and represent the optimal performance which can be achieved when other factors are held equal (Dorman & Loizou, 1998). In simulation studies, speech signals are processed in the same manner as cochlear implants: sounds are passed through a bank of band-pass filters, and the speech envelope extracted from each frequency band is used to modulate a noise or sine-wave carrier. The outputs of all the frequency bands are then

summed up. Although the sound quality is completely different for noise and sine-wave carriers, there is no significant difference in speech performance between the two types of carrier (Dorman *et al.*, 1997a).

Number of channels

One important feature of the multichannel implant system is the use of multiple electrodes to restore some place coding of frequency information in the cochlea. There are different numbers of frequency channels used in different implant devices, ranging from 4 or 6 in the Ineraid implant system to over 20 channels in the Nucleus implant system. In theory, more frequency channels would provide better frequency resolution, and therefore would provide more benefit for implanted listeners in understanding speech signals. There have been a number of studies investigating the number of channels necessary to achieve high levels of speech perception performance. Studies using normal-hearing subjects listening to acoustic simulations of cochlear implants have shown that speech can be recognised to a great accuracy with a relatively small number of channels, 4 to 6, at least in good listening situations and with simple highly predictable sentences (e.g. Shannon *et al.*, 1995; Dorman *et al.*, 1997a; Loizou *et al.*, 1999; Shannon *et al.*, 2004). For instance, Shannon *et al.* (1995) examined speech performance with one to four noise-modulated bands, and reported that subject perception performance improved dramatically as the numbers of channels increased from one to four. Nearly perfect speech performance was achieved with 4-channels of spectral information (over 90% correct for all speech tasks, including recognition of vowels, consonants and sentences). Dorman *et al.* (1997a) also used acoustic models to investigate the effect of the number of channels on speech intelligibility, but constructed speech with both sine waves and noise bands.

They reported similar results to Shannon *et al.* (1995), and subject performance was very similar for the sine wave processor and the noise band processor, though their sound quality was completely different.

The number of channels for optimum performance has also been found to vary with test materials. For simple sentences produced by one male speaker, 4-5 channels allow a high level of performance (e.g. Shannon *et al.*, 1995; Dorman *et al.*, 1997a). For more difficult materials, such as vowels or sentences produced by more than one speaker, or complex sentences, more frequency channels are required (e.g. Loizou *et al.*, 1999; Zeng *et al.*, 2005). Loizou *et al.* (1999) reported that 5 channels were necessary for recognising sentences produced by multiple talkers. Zeng *et al.* (2005) compared the performance of the same subjects for three sentence tests, the CUNY, HINT, and IEEE sentences, using 4-channel signal processing, and reported a significant effect of speech materials (more than 80% correct for the CUNY and HINT tests, but only around 40-50% correct for the IEEE test). The difference between the IEEE test and the other two tests was due to the IEEE sentences having more complicated sentence structure, while the CUNY sentences were topic-related and the HINT sentences were short and declarative[2].

Studies in implanted patients have shown that the number of channels required for optimal performance was very similar to results from acoustic models (e.g. Lawson *et al.*, 1996; Fishman *et al.*, 1997). Fishman *et al.* (1997) examined the effect

---

[2] Sentence examples for the three sentence tests (given in Zeng *et al.*, 2005):

The CUNY (City University of New York) sentences, e.g., "Make my steak well done."

The HINT (Hearing in Noise Test) sentences, e.g., "A boy fell from the window."

The IEEE (Institute of Electrical and Electronic Engineers) sentences, e.g., "The kite dipped and swayed, but stayed aloft."

of channel number on speech performance in 11 Nucleus-22 users with the SPEAK strategy. Implant users were given two days to familiarize themselves with each new electrode setting. On average, subject performance increased dramatically as the number of channels increased from one to four, and reached asymptotic levels with seven channels. The performance of implanted listeners was further compared with the results in Shannon *et al.* (1995), revealing that, when channel number was restricted to a small number (from 1 to 4 channels), the best performance of the implanted listeners was similar to the average performance of normal-hearing listeners. No significant difference was found on any speech performance between conditions with 7, 10, and 20 channels. This indicated that, despite up to 20 channels available in the Nucleus implant system, these implanted listeners did not make use of all the information provided.

Speech recognition in noise

Although speech can be recognised with a relatively small number of frequency channels, the results mentioned above represent what could be achieved in ideal, quiet listening conditions. In real everyday situations, speech often occurs with some background noise. In more difficult listening conditions, such as in noise or the presence of a competing speaker, more frequency channels are required to achieve a high level of speech recognition (e.g. Dorman *et al.*, 1998b; Faulkner *et al.*, 2001; Fu *et al.*, 1998a). Dorman *et al.* (1998b) used acoustic simulations to investigate the effect of different noise levels on the number of channels needed for good sentence recognition. While the optimal performance was reached with only 5 channels in quiet, up to 12 and 20 channels was required at +2dB SNR (signal-to-noise rate) and at -2dB SNR, respectively. Friesen *et al.* (2001) investigated speech recognition by implanted and normal-hearing listeners in a number of different SNRs. The results

showed that, while the performance of normal-hearing listeners continued to improve as the number of channels increased, the performance of implanted listeners reached asymptotic level at seven to ten channels. This indicated that, regardless of the number of channels provided, implanted listeners appear to use only four to seven channels of spectral information.

Hearing-impaired listeners have been found to have more reduction on their speech performance in noise, compared to normal-hearing listeners (e.g., Van Tasell & Yanz, 1987; Van Tasell *et al.*, 1988). This noise susceptibility in hearing-impaired listeners could result from reduced frequency selectivity, loudness recruitment, or other factors (Moore, 1996). To investigate if the noise susceptibility of implant users is, at least partially, due to the loss of spectral information caused by limited frequency channels in implants, Fu *et al.* (1998a) compared phoneme recognition performance from implanted users to that of normal-hearing listeners. The best implant user had similar patterns of performance as normal-hearing subjects did both in quiet and in noise. This suggested that the susceptibility to noise of implant users was, at least partially, due to the loss of spectral information. Performance of implant users in noise would improve with a greater number of effective channels.

Speech recognition by children

The number of channels necessary for high levels of speech recognition performance has also been investigated in children. In general, young children require more channels than adults to reach a given level of performance (Dorman *et al.*, 2000; Eisenberg *et al.*, 2000; Eisenberg *et al.*, 2002). Dorman *et al.* (2000) examined the effect of degraded spectral information on a group of young children, aged 3 to 5

years, and adults. Two word lists, one containing lexically 'easy' words and another containing lexically 'hard' words[3], were processed into 4 to 20 channels with sine wave outputs (4-, 6-, 8-, 10-, and 12-channel CIS processing and 6-out-of-20 channel n-of-m processing). The results demonstrated that young children generally had lower performance than adults, and needed more channels to achieve the same level of performance. For instance, children required about four more channels than adults to achieve 80-90% correct (10 channels required to reach around 80% correct for children but only 6 channels needed for adults). These young children were able to match the level of adult performance for easy words at 12 channels, which could be due to both reach the ceiling level near 100%. However, they did not reach adult performance for hard words in any condition, even with 20 analysis channels (the 6-out-of-20 peak-picking processing). Since all the words used were in the vocabulary of children, the results indicated that, apart from the level of lexical knowledge, younger children required more detailed sensory information than adults. The fact that degraded spectral information affects younger children more than older ones has been reported previously, thought the underlying mechanism is still not completely clear yet. For instance, Elliott (1979) reported that, while 11- and 17-year-olds did not perform differently in recognising words in quiet, 11- and 13-year-olds had significantly lower performance than 15- and 17-year-olds for the same test materials in noise.

Eisenberg *et al.* (2000) conducted similar investigation on two groups of children (5-7 and 10-12 years old) and adults and reported similar results. Speech

[3] The easy words were high in word frequency and low in 'neighbourhood density', and the hard words had opposite characteristics (low in word frequency and high in neighbourhood density). Neighbourhood density refers to the number of words generated by adding, substituting or deleting extra phonemes.

materials were processed into 4 to 32 frequency channels using noise carriers for outputs. They found that older children and adults had no significant differences in performance, but younger children were significantly lower. Further analysis of context effects showed that the 5-7 year olds were less capable of making use of sentence context to help to recognise words.

Dorman *et al.* (2000) also examined the recognition of the same word lists in 56 prelingually-deafened children with Nucleus-22 implant devices, and compared their performance to the simulation results from normal-hearing listeners. The average performance of these implanted children was close to the performance of young children listening to 6-channel simulations, or that of adults listening to 4-channel simulations. These results indicated that the prelingually implanted children, just like postlingually implanted adults mentioned before, could only extract information from a relatively small number of channels, despite up to 20 channels provided in the Nucleus devices. Both Dorman *et al.* (2000) and Eisenberg *et al.* (2000) reported that young children had great difficulty in identifying six or less channels of vocoded speech, suggesting that the development of spoken word recognition in prelingually implanted children might be much more difficult and take a longer period to complete given only the relatively degraded information available.

Insertion depth

The electrode arrays are typically inserted 22-30 mm into the cochlea, depending on different implant systems and the state of the patient's cochlea (Loizou, 1998). Because the electrode array is not placed in the whole range of the cochlea, spectral information is often presented to the 'wrong' place for implanted listeners. Ketten *et al.* (1998) measured the electrode position in a group of implant users with

Nucleus-22 devices, and reported that, on average, the most apical electrode was about 20 mm from the base. According to the Greenwood equation (1990)[4], this would place the most apical electrode at around the 1 k Hz place, while the centre frequency of the most apical channel in an implant is normally just a few hundred Hz.

To investigate the effect of the frequency/place mismatch on speech performance, several studies have been conducted in implant users and normal-hearing subjects listening to acoustic simulations (e.g. Dorman *et al.*, 1997b; Dorman & Ketten, 2003; Faulkner *et al.*, 2003; Fu & Shannon, 1999; Shannon *et al.*, 1998). In general, listeners were able to compensate for the frequency shift to some extent. For a frequency shift corresponding to a 3-mm electrode shift basalward, both implant users and normal-hearing subjects had no significant drop in speech performance or some decrement but back to the baseline performance after a relatively short period. However, severe reductions were found in speech performance with a 5-mm or more frequency shift.

Although the spectral shift had a devastating effect on speech performance initially, Rosen *et al.* (1999) reported that listeners were able to adapt to the spectral shift quickly to some extent. For instance, performance on word recognition in 4-channel sentences, with a shift equivalent to 6.5-mm, improved from 1% to 30% correct after only 3 hours of training (64% correct for un-shifted sentences). However, it is not clear whether a complete accommodation to spectral shift will be possible after further training, or how long a period of training will be necessary. Fu *et al.* (2002) investigated the adaptation of spectrally shifted speech in three implanted listeners with Nucleus-22 devices over a 3-month period. All three subjects showed

[4] $frequency = 165.4(10^{35-0.06p} - 0.88)$ where p is the position on the basilar membrane (in mm) from the base.

significant improvements after 3 months, indicating they were capable of learning the spectrally distorted speech. However, performance of most speech materials remained significantly lower than the baseline performance. The authors, therefore, suggested a full adaptation for spectral-shifted speech would take a longer time, at least more than 3 months, and the degree of spectral shift might also affect the time needed for a full adaptation. However, because the frequency-shift was done by the use of different frequency allocation tables in implant systems, the poorer performance on shifted speech might be confounded by the fact that spectral information was actually presented in different places of the cochlea. For instance, the information around the 1.5k Hz frequency range, which was crucial for vowel recognition, was presented on Electrode 8 or 9 for the two subjects using Table 7 and on Electrode 7 for another subject using Table 9. Over the experimental period, this information was basalward-shifted to Electrode 12 when the frequency tables of all the three subjects were changed to Table 1. Because the degeneration of neurons tends to be more severe in basal rather than apical locations, the number of surviving ganglion cells in different sites of the cochlea might also contribute to why these implanted subjects were unable to fully compensate for the deficit in performance.

Stimulation rate

The stimulation rate refers to the number of pulses delivered to each electrode per second, and is expressed as pps (pulses per second). The stimulation rate used in currently commercial implant devices varies from few hundred to over 2000 pps. Theoretically, higher stimulation rates are considered to be more beneficial due to the better representation of temporal envelope (e.g. Loizou *et al.*, 2000b), and more normal stochastic firing patterns in auditory nerves (Rubinstein *et al.*, 1999; Wilson *et*

al., 1998). However, it is not yet clear whether the better temporal information and/or the stochastic response would lead to better speech performance in implant users. Contradictory results have been reported across studies. Fu and Shannon (2000) examined the effect of the stimulation rate, ranging between 50 and 500pps, in 6 Nucleus-22 users with an experimental 4-channel CIS strategy. Their results showed no significant difference in phoneme recognition performance for stimulation rates between 150 to 500 pps. However, performance reduced significantly when the stimulation rate was reduced to less than 150 pps. Although subjects showed no difference on their performance, they reported some difference in the quality of sound, more machinelike and weird, for stimulation rates at 150 and 200pps. Vandali *et al.* (2000) examined the effect of stimulation rate using rates of 250, 807, and 1615 pps in 5 Nucleus-24 users with the ACE strategy. They also reported no significant improvement on group performance with stimulation rates higher than 250pps.

In contrast, other studies have found significant improvements in performance for higher stimulation rates (e.g. Brill *et al.*, 1997; Kiefer *et al.*, 1999; Loizou *et al.*, 2000b). Loizou *et al.* (2000b) examined speech recognition with 4 different stimulation rates (400, 800, 1400, and 2100pps) in 6 Med-El CIS users. They found that higher stimulation rates (2100 pps) produced better performance in word and consonant recognition than lower ones (less than 800pps). There was no significant effect of stimulation rate on vowel recognition. The lack of the effect of stimulation rate on vowel recognition was likely due to vowels being characterised by relatively slow changes in the spectrum, which was not affected by the change of temporal information with different stimulation rates (Loizou *et al.*, 2000b). Kiefer *et al.* (1999) examined the effect of stimulation rate, varying from 600 pps to the maximum available rate (depending on the number of active channels), in 13 Med-El CIS users

and also reported similar results. For instance, with eight active channels, speech performance was reduced consistently from the standard rate (about 1500-1700 pps) to 1200 pps and to 600 pps, with the most reduction in monosyllables by 10% and then in consonants by 6%.

Many studies have reported great variation for the optimal stimulation rate across subjects. While some subjects received significant benefit with a high stimulation rate, some showed no difference with different stimulation rates. One possible explanation for no effect on stimulation rate may be due to some implant users not being able to make use of the temporal information provided by higher stimulation rates (Fu & Shannon, 2000). Higher stimulation rates have even shown negative effects in some subjects, both in terms of performance (e.g. Holden *et al.*, 2002; Vandali *et al.*, 2000; Verschuur, 2005), and, occasionally, of increased tinnitus (Vandali *et al.*, 2000). Thus, to optimise individual performance, it appears that different stimulation rates should be tried for different implant users.

Patient variables

Several subject-related factors have been suggested to affect speech perception performance in implant users, including etiology, survival of spiral ganglion cells, duration of profound hearing impairment, age of onset of deafness, age at implantation, and duration of implant use. For postlingually-deafened implanted adults, Blamey *et al.* (1996) reported a strong negative effect for duration of deafness and a strong positive effect for duration of implant use. Only slight effects were reported for age at onset of deafness and age at implantation. The effect of etiology of deafness was also weak, despite the idea that the number of surviving ganglion cells

had been expected to be positively related to subject performance. For prelingually-deafened implanted children, two crucial factors seem to be the age of implantation and duration of implant use. Performance of implanted children has been found to be positively correlated with the duration of implant experience, and negatively related with the age of implantation (Fryauf-Bertschy *et al.*, 1997; Rubinstein, 2002; Tyler *et al.*, 2000). Tyler *et al.* (2000) reported that children implanted at a younger age had higher scores on recognition of phonetically balanced kindergarten (PBK) words than those implanted at an older age. The group of children who were implanted as young as 2 to 3 years old had the best performance compared to the other age groups. Dorman *et al.* (2000) compared the performance on word recognition by early- and late-implanted children with Nucleus 22 devices (mean age of implantation was 3.3 and 5.4 years, respectively), and reported that the average performance of early-implanted children was slightly better than late-implanted children (51 and 45% correct, respectively). Furthermore, while the performance of late-implanted children varied from near 0 to over 90%, none of the early-implanted children fell into the low-end of performance.

2.5 Limitations in current implant devices and future directions

One of the major limitations in currently implant systems is that implanted users are not able to use all channels of spectral information. At present, implant users are only able to make use of around 4 to 8 channels of spectral information, which is not always sufficient except in good listening situations and with simple predictable

sentences. It is necessary to determine the cause of this limitation before next generation implants can provide a greater number of efficient channels. Another major limitation of current implant devices is that users can only perceive coarse information for voice pitch. As a result, implant users often perform poorly in speaker identification, melody recognition, and recognising paralinguistic information. For tone language users with an implant, the difficulty in recognising tonal contrasts has been reported in speakers of Mandarin and Cantonese. To improve the performance of implant users, better pitch information will need to be encoded in newer implant devices.

Another challenge for cochlear implant research is the great variation in subject performance. While some implant users demonstrate excellent performance with their implants, others receive only limited benefits from the same implant systems. A better understanding of the causes of subject variation will assist both in the development of better implant systems, and will allow for better fitting of device; that are more effectively programmed and customised for individual user so as to further help implanted listeners improve their speech perception ability.

Chapter 3

Acoustic cues to tonal contrasts in Mandarin:

Implications for cochlear implants

This chapter aims to investigate the contribution of fundamental frequency, amplitude envelope and duration to the perception of Mandarin tone. The experiment described in this chapter systematically manipulated these three cues in isolation, all together, and in combination of any two cues. The study by Lin (1988) reported that amplitude envelope and duration had only negligible effect on tone recognition in the presence of f0 information. For instance, he imposed either amplitude envelope or duration of other tones to a tone containing its original f0, and found the recognition rate only dropped by 1.2 and 3 percentage points, respectively. Some more recent studies, on the other hand, have reported some significant effect for these two cues when f0 information was relatively weak or completely absent (e.g. Whalen and Xu, 1992; Fu and Zeng, 2000). To see how different degrees of f0 information interact with amplitude envelope and duration, and influence tone recognition, both explicit f0 and less salient f0 cues are investigated in this study.

Before describing the main experiment, an introduction to Mandarin tone will be given, followed by a review of previous studies investigating acoustic cues for recognising tonal contrasts. While cues other than explicit f0 are less important for normal-hearing listeners, they may play a more important role for cochlear implant users with current implant devices. The implications of the present study for tone perception by implant users will be discussed in the end.

Fundamental frequency (f0), pitch, and tone

In the discussion of lexical tone, the three terms, f0, pitch, and tone, are used frequently. Each term refers to a specific aspect, from the purely acoustic term to the truly linguistic one (Yip, 2002), and they are not inter-changeable. ***F0*** is an acoustic term, referring to the frequency of glottal vibration when speech signals are produced. ***Pitch*** is a perceptual term. The pitch percept is related to f0. A high repetition frequency is associated with a 'high' pitch percept, and vice versa. However, the pitch percept does not change linearly with increasing f0. ***Tone*** is a linguistic feature used to indicate a phonological category in languages which use pitch to distinguish two words.

3.1 Introduction

Tone system in Mandarin

In Mandarin Chinese, four tones are used as one of the phonetic determinants of lexical meaning (e.g. Chao, 1948; 1968; Yip, 1980). These four tones are mainly distinguished by their variations in fundamental frequency (e.g. Howie, 1976; Lin, 1988). Table 3.1 summarises the four tones in Mandarin. The four tonal patterns are often described in terms of the pitch range of the particular speaker. Tone 1, the level tone, starts within a speaker's high pitch range and maintains approximately the same pitch to the end of the syllable. Tone 2, the rising tone, starts in a middle pitch range, then shows a rise, which is sometimes preceded by a small dip. Tone 3, the falling-rising tone, also starts in a middle pitch range, falls gradually toward a low pitch and then rises. Tone 4, the falling tone, starts at a high pitch, and drops to a low one. A

syllable with different tones represents different words. For example, the syllable *ba* expressed with tones 1, 2, 3, and 4 means 'eight', 'to pull', 'to hold' and 'father' respectively.

Tone	Tone Description	Example	Chinese Character	Meaning
1	level tone	bā	八	eight
2	rising tone	bá	拔	to pull
3	falling-rising tone	bǎ	把	to hold
4	falling tone	bà	爸	father

Table 3.1 The four tones in Mandarin

Although tone 3 is often referred to as a falling-rising pattern, it also appears as other patterns (e.g. Chao, 1948; 1968; Shih, 1988; Chen, 2000). This variation is due to phonological processes that are described by the third tone sandhi rule. A falling-rising pattern appears in isolated words and in sentence-final position. However, in non-final positions, a low-falling pattern occurs. In addition, if a syllable with tone 3 is followed by another syllable with the same tone, the first syllable will be pronounced as tone 2, a rising tone. Therefore, tone 3 appears in three patterns: the falling-rising pattern, the low-falling pattern and the rising pattern. Previous studies have shown that tone 3 is sometimes mis-identified as the falling tone (tone 4) or the rising tone (tone 2) (Garding *et al.*, 1986; Shen and Lin, 1991)

The usage of the falling-rising and the low-falling patterns also varies according to accent (Shih, 1988). Northern Chinese speakers use the falling-rising pattern in

sentence-final position in all speech. By contrast, southern speakers use the low-falling pattern even in the final position in casual speech, and use the falling-rising pattern in emphatic speech and in yes-no questions. Dialect may also influence Mandarin tone. Fon and Chiang (2000) reported that Taiwan Mandarin showed a narrower tonal range, lower tonal heights and more conservative tonal contours compared to Beijing Mandarin. This could be attributed to an influence from Taiwanese, which has a lower tonal register than Mandarin (Lin and Repp, 1989).

Acoustic cues to Mandarin tone

While the four tones are mainly distinguished by fundamental frequency, other acoustic characteristics such as overall intensity and duration tend to vary systematically with tone (Howie, 1976; Zee, 1978; Blicher *et al.*, 1990; Tseng, 1990; Whalen and Xu 1992; Fu and Zeng, 2000). Tone 4 is often the most intense one and has the shortest duration, while tone 3 is the least intense and has the longest duration. However, different patterns of tone 3 differ in duration as well as in pitch contour. While the falling-rising pattern shows the longest duration among the four tones, the low-falling pattern has the shortest duration (Shih, 1988). Amplitude envelope has also been found to be highly correlated with f0 contour for tones 3 and 4 (Whalen and Xu, 1992; Fu and Zeng, 2000).

Several studies have investigated tone perception from purely temporal cues using various noise stimuli modulated by envelopes derived from natural speech stimuli. Three temporal envelope cues were defined by Fu and Zeng (2000): periodicity, amplitude contour and duration. As in Rosen (1992), 'periodicity' refers to fluctuations in the overall amplitude at a rate between 50 and 500 Hz, whereas

'amplitude contour' refers to fluctuations at a rate between 2 and 50 Hz. Although percepts of pitch are strongest for truly periodic sounds, they can also be elicited by temporal fluctuations in amplitude-modulated noise (Burns and Viemeister, 1976). When the temporal envelope of speech is derived from an envelope-smoothing filter which includes the voice f0 range, listeners can perceive the quasi-periodic amplitude modulations imposed on a noise carrier as pitch changes. The temporal regularities in the modulated noise can be referred to as 'periodicity' information (Rosen, 1992). However, when an envelope-smoothing filter is below the fundamental frequency range, the periodicity information is not included in the envelope signal.

Whalen and Xu (1992) used signal-correlated-noise (SCN) stimuli to investigate the extent to which tone recognition can be aided by amplitude contour and duration. Their results showed high recognition scores for tones 2, 3 and 4, averaging 87.6 %, but only 38.5% correct for tone 1. To further examine the contribution of amplitude contour, they used stimuli with controlled duration, and found that recognition scores were still well above chance in the absence of the duration cue (45.0%, 55.3%, 69.5% and 92.3% for tones 1, 2, 3 and 4, respectively). However, the SCN stimuli used in the study did not only contain amplitude contour and duration cues, but also the periodicity of the natural tokens on which they were based, albeit in weakened form. Green *et al.* (2002) showed listeners to have some ability to label glides in fundamental frequency at low frequency ranges in stimuli that were similar to, although not identical to, SCN. Thus, periodicity information, directly related to the change in fundamental frequency, was likely to influence subject performance as well (Rosen, 1992; Van Tasell *et al.*, 1987).

Fu and his colleagues (1998) investigated the importance of periodicity and amplitude envelope to tone recognition by using amplitude-modulated noise. In this

study, duration was not examined and retained its natural variations. When no spectral information was available (1-band condition), recognition scores of up to 80% could be achieved in a condition that preserved amplitude envelope, periodicity, and duration information, while accuracy was about 67% in a condition preserving only amplitude envelope and duration information. Tones 3 and 4 were identified best, nearly twice as well as tone 1 and tone 2.

In another study, Fu and Zeng (2000) used signal-correlated-noise to further explore the contribution of periodicity, amplitude contour and duration to tone recognition. These three cues were manipulated in isolation, in pairs and all together. Results showed that the highest score, nearly 70%, was achieved in the condition with all three cues. Around 55% correct recognition was shown in conditions that preserved either the amplitude contour or periodicity cues. A condition preserving only the duration cue had the lowest recognition score of about 35% correct. The duration cue was found to contribute to recognition of tone 3, and the amplitude cue contributed to the recognition of both tones 3 and 4. The periodicity cue contributed to recognition of all four tones.

In a recent study by Xu *et al.* (2002), amplitude envelope and periodicity were manipulated by varying the low pass cutoff frequency of an envelope-smoothing filter between 1 to 512 Hz in 1-octave steps. Tone recognition improved slightly but consistently as cutoff frequency increased in the 1-channel condition in which there was no spectral information available. Consistent with the results of Fu *et al.* (1998), around 50 % correct recognition was achieved with only amplitude envelope and duration cues (when the envelope cutoff frequency was as low as 1 or 2 Hz), while accuracy approached 65 % when periodicity cues were also available (with an envelope cutoff frequency at 512 Hz). When the duration cue was removed, scores

dropped significantly, with tones 3 and 4 affected the most (the decreases were by 12.5, 8.5, 19.5, and 19.7 percentage points for tones 1, 2, 3 and 4, respectively). These decreases in performance indicated that duration did show a significant effect on tone recognition, mainly on tones 3 and 4.

To sum up, results from the above studies confirm that tone recognition can be assisted by temporal envelope cues. Although the pitch from modulated noise is considerably less salient than the pitch of harmonic signals for which there are spectral cues to pitch (Burns and Viemeister, 1976), listeners can perceive the periodicity information in the modulated noise and use it to recognize tone. The amplitude and duration cues also play significant roles for tone recognition, and contribute to tones 3 and 4 the most. The studies showed an inconsistency of effects of duration. Duration contributed to the identification of both tones 3 and 4 in Xu *et al* (2002), but only affected tone 3 in Fu and Zeng (2000). It appears that duration did not consistently contribute to tone 4 recognition in the latter study because tone 4 was not significantly different from tone 1 in duration in the latter study. In the former study, tone 4 was the shortest tone. Therefore, although duration can sometimes be used to aid tone recognition, it varies in natural speech so may not always be a reliable cue.

3.2 Experiment I: Contribution of f0, amplitude envelope, and duration to Mandarin tone recognition

Aims and experimental predictions

The aim of the study is to investigate the extent to which tonal contrasts can be recognised with different acoustic cues. Although these three main cues to tonal contrasts in Mandarin (f0, amplitude envelope, and duration) have been investigated in several studies, only the study by Fu and Zeng (2000) systematically combined different cues to examine their contributions. Modulated noises were used to carry all possible combinations of the three cues. However, the pitch information was conveyed by the temporal fluctuations in the speech envelope which is a far less salient cue than that which arises from true periodic sounds. Here, we introduce a sawtooth waveform created period by period to match the f0 of original speech, and used this to carry pitch information. The f0-controlled sawtooth waveform contained both spectral and temporal cues to pitch, and will give a clearer indication of voice pitch than what would be possible from the temporal envelope of noise alone. The same three cues in all combinations were conveyed both by sawtooth and noise carriers (the latter as in Fu and Zeng, 2000, but with some modifications in signal processing). With the same testing materials, the results of the present study allow the comparisons of the level of correct tone recognition with different combinations of acoustic cues.

Stimuli with explicit f0 information (f0 carried by sawtooth carriers) were expected to be recognised the best among all stimuli. For stimuli with the relatively weak temporal f0 information (f0 carried by temporal fluctuations of noise carriers),

listeners were expected to be able to identify some tonal information, but their performance would not be as good as the performance when f0 was represented explicitly. It is well known that amplitude-modulated noise stimuli, with no regularity in fine structure, lead to weak pitch sensations. According to Lin's (1988) study, when explicit f0 is presented, amplitude envelope and duration are unlikely to contribute much to the recognition of tonal contrasts. However, when f0 information is absent or presented in a rather weak form, as in Whalen and Xu (1992) and Fu et al. (1998), amplitude envelope and/or duration could be of use. Even so, the effect of amplitude envelope and/or duration would not be huge.

3.2.1 Method

1. Speech stimuli

Four syllables (/i/, /ba/, /fu/ and /tɕ^{h}i/) with each of four tones were used as stimuli. Natural models for the stimuli were produced by two young adults, one male and one female, who were native speakers of Mandarin from Taiwan. All stimuli were randomized and produced in the carrier phrase "qǐ tiao chu –" ("Please choose – ") to avoid inconsistency in pitch range. Both speech and laryngograph signals (Lx) were recorded in an anechoic chamber. Two tokens were selected for each tone for each of the four syllables, resulting in a total of 64 speech stimuli (4 syllables ×4 tones ×2 tokens ×2 speakers). These speech stimuli were digitized using a Sony DAT recorder at a 44.1 kHz sampling rate.

Acoustic analysis

Table 3.2 shows fundamental frequencies, mean RMS amplitudes and duration for the four tones.

		Fundamental Frequency (Hz)		**RMS Amplitude** (dB re 1)	**Duration** (ms)
		Male	**Female**		
TONE 1	Average	158 (6)	257 (12)	-23.5 (3.5)	308.8 (38)
TONE 2	Initial	121 (16)	201 (11)	-27.7 (2.9)	347.7 (47)
	Final	143 (15)	222 (15)		
TONE 3	Initial	120 (13)	198 (15)	-30.7 (2.5)	323.7 (66)
	Turning Point	91 (9)	113 (41)		
	Final	119 (9)	147 (37)		
TONE 4	Initial	171 (32)	264 (18)	-25.0 (3.0)	256.2 (72)
	Final	108 (14)	150 (42)		
Average				-26.7 (3.6)	309 (65)

Table 3.2 Fundamental frequency (Hz), RMS amplitude (dB re 1, where all amplitudes are within the range ± 1), and duration for the four tones (standard deviations in parentheses).

F0 contour

Figure 3.1 shows pitch contours for all speech tokens by the male and female speakers. Note that the pitch contours of tone 3 for the female appeared to be more variable. While the male speaker used the falling-rising pattern consistently, the female speaker used both falling-rising and low-falling patterns. Some of the pitch contours kept falling throughout while others rose slightly at the end.

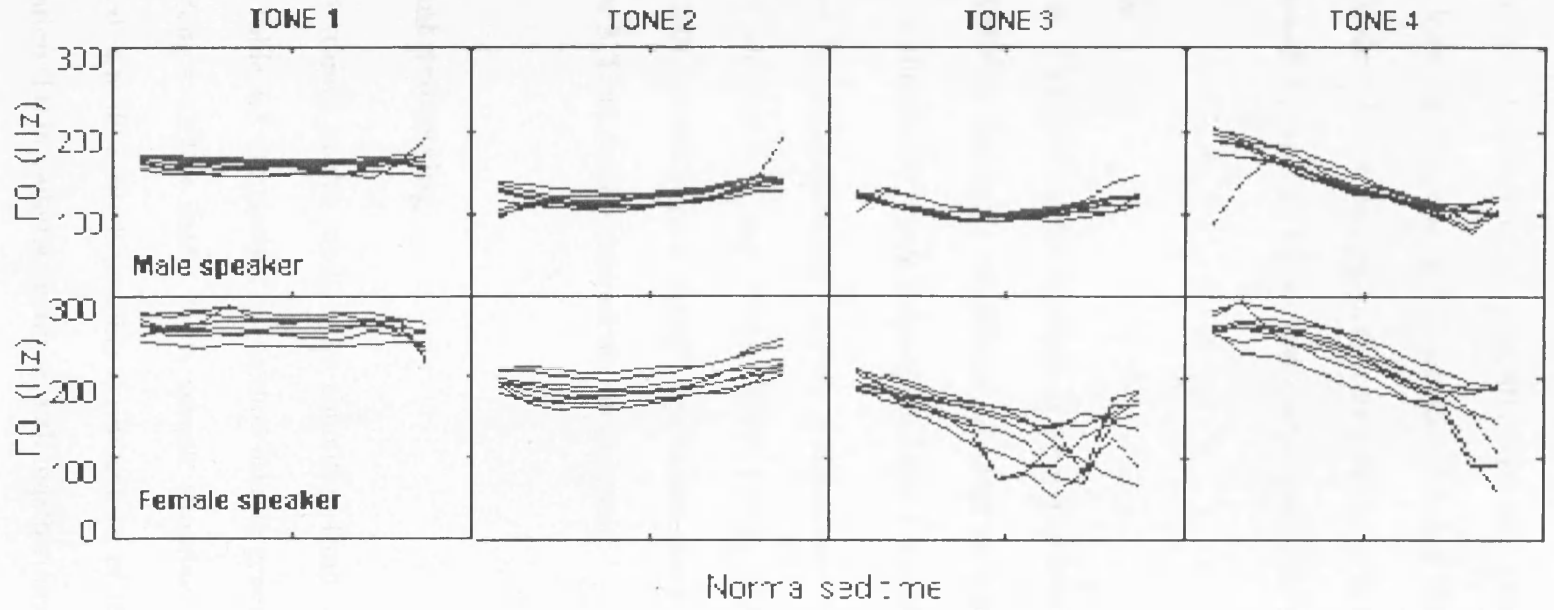

Figure 3.1 F0 contours for speech stimuli from the two speakers (all the pitch contours were normalised to give the same duration).

Overall intensity

An ANOVA showed a significant difference of mean RMS-amplitude of the four tones [F (3, 60) = 17.22, p<0.001]. *Post-hoc* comparisons revealed that tone 3 was the least intense tone; it had significant lower RMS-amplitude than other three tones. Tones 1 and 4 had no significant difference between their overall intensities, and both had significnant higher RMS-amplitudes than tone 2.

Duration

Tone 2 had the longest duration, followed by tone3, then tone 1, and finally tone 4. An ANOVA showed a significant effect of duration [F (3, 60) =7.3, p<0.001]. *Post-hoc* comparisons revealed that tones 2 and 3 was significantly longer than tone 4.

Note that unlike previous studies, in which tone 3 was often the longest and tone 4 was the shortest tone (e.g. Howie, 1976; Tseng, 1990; Whalen & Xu, 1992; Fu & Zeng, 2000), for the speech stimuli used in this study, tone 3 had a shorter duration than tone 2. Tone 4 still remained the shortest tone.

2. Signal processing

All stimuli in this study were simplified from the originally recorded speech stimuli. Table 3.3 summarises the method used to generate these stimuli, compared to Fu & Zeng's (2000) study. Each speech stimulus was used to produce fifteen simplified stimuli with all possible combinations of f0 (F), amplitude envelope (A) and duration (D) in isolation, pairs, and all together, and carried by either sawtooth or noise carriers. Figure 3.2 shows some examples of the simplified stimuli.

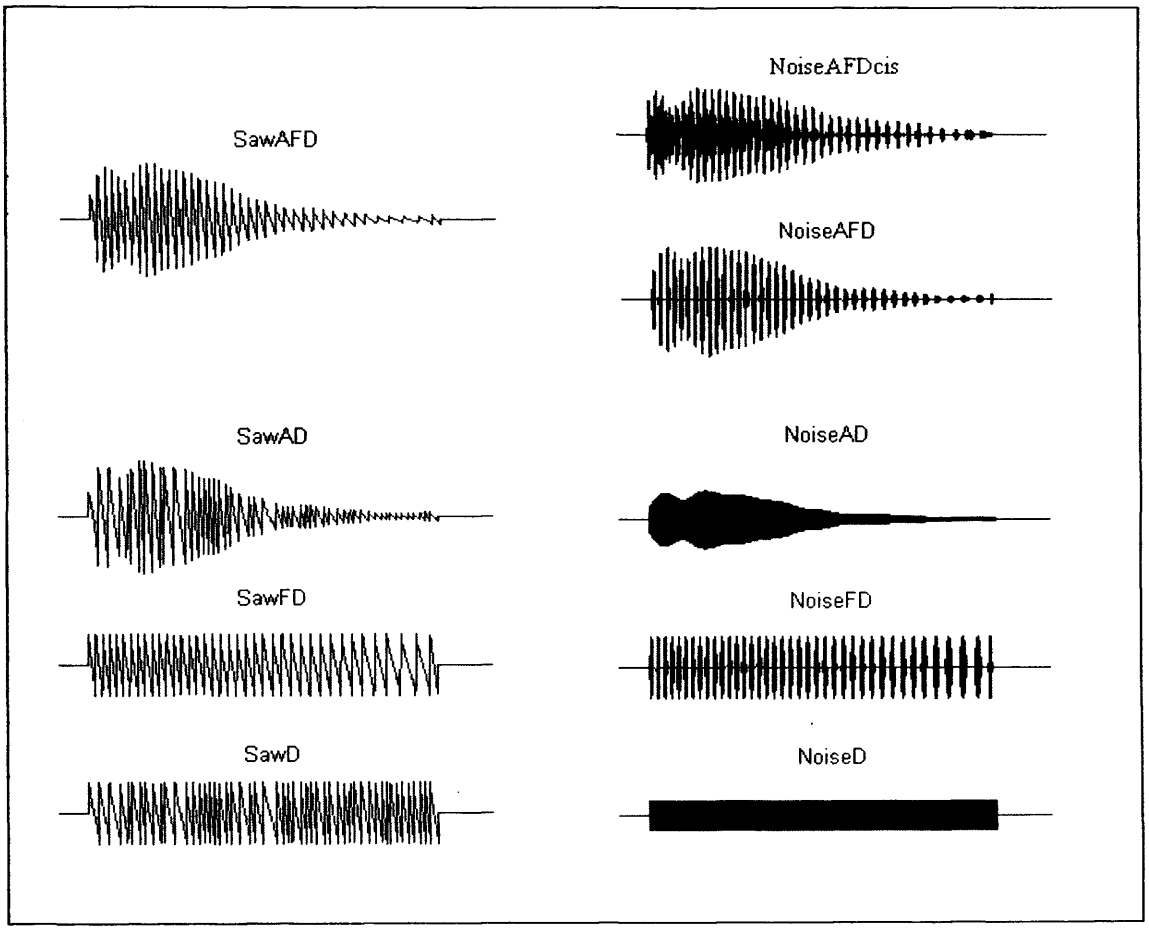

Figure 3.2 Examples of some simplified stimuli (conditions AFD, AD, FD, and D, using both sawtooth and noise carriers and NoiseAFDcis) for the syllable /ba/ with tone 4 produced by the male speaker. Other simplified stimuli AF, A, and F are similar to AFD, AD, and FD, respectively, except with duration fixed at 309ms.

Carrier / Cues		Present study		Fu & Zeng 2000
		Sawtooth carrier	**Noise** carrier	**Noise** carrier
Natural duration	**AFD**	f0-controlled sawtooth carrier x envelope extracted at 30 Hz	**AFDcis:** noise carrier x envelope extracted at 400 Hz **AFD:** f0-controlled sinusoidal modulated noise x envelope extracted at 30 Hz	noise carrier x envelope extracted at 500 Hz
	AD	random-frequency sawtooth carrier x envelope extracted at 30 Hz	noise carrier x envelope extracted at 30 Hz	noise carrier x envelope extracted at 50 Hz
	FD	f0-controlled sawtooth carrier	f0-controlled sinusoidal modulated noise	f0-controlled 100% amplitude modulated noise (no further details provided)
Duration fixed at 309ms	**AF**	sawtooth carrier controlled by the 309-ms time-scaled f0 contour x time-scaled envelope extracted at 30 Hz	time-scaled f0-controlled sinusoidal modulated noise x time-scaled envelope extracted at 30 Hz	linear interpolation was applied to AFD-stimuli to give a fix duration at 400ms (the range of f0 variation may change although the same percentile of f0 change remain)
	A	309-ms random-frequency sawtooth carrier x time-scaled envelope extracted at 30 Hz	noise carrier x time-scaled envelope extracted at 30 Hz	linear interpolation was applied to AD-stimuli to give a fix duration at 400ms (the range of f0 variation may change)
	F	sawtooth carrier controlled by the 309-ms time-scaled f0 contour	time-scaled f0-controlled sinusoidal modulated noise	linear interpolation was applied to FD-stimuli (the range of f0 variation may change)
Duration cue only	**D**	random-frequency sawtooth carrier with original duration	noise carrier with original duration	noise carrier with original duration

*Table 3.3 Summary of stimuli used in the present study and the study of Fu & Zeng (2000). Cues **A**, **F**, and **D** represent amplitude contour, f0, and duration, respectively.*

1. ***AFD*** (amplitude envelope, f0, and duration)

Sawtooth carrier: The times between successive regular vocal fold closures (referred to as Tx) were measured from laryngograph signals, and used to generate a sawtooth waveform period by period. The f0-controlled sawtooth waveform preserved the f0 contour of the original speech with constant amplitude contour. The amplitude envelope was extracted by full-wave rectification, followed by forward and backward filtering with a 30 Hz cut-off frequency fourth-order elliptical lowpass filter. The f0-controlled sawtooth carrier was then multiplied by the amplitude envelope to generate the ***SawAFD***-stimulus.

Noise carrier: In the present study, the signal processing used to generate the AFD-stimulus by a noise carrier was different from that used in Fu and Zeng's study (2000). A sinusoidal waveform was generated from Tx values period by period, then half-wave rectified, and then used to modulate a noise carrier. The ***NoiseAFD***-stimulus contained all three cues, but the temporal cue to pitch was considered to be rather weak. Another condition, which was almost the same as the one used by Fu and Zeng (2000), except using a slightly different cut-off frequency (see Table 3.3), was also generated. A noise carrier was multiplied by a speech envelope extracted by half-wave rectification and low-pass filtering with a 400 Hz cut-off frequency to generate the ***NoiseAFDcis***-stimulus. Since the range of voice pitch was included, the envelope with 400-Hz cut-off frequency would be expected to preserve some pitch information. The NoiseAFD-stimulus, because of its clearer pattern of f0-related modulations, might contain slightly better pitch information than the NoiseAFDcis-stimulus (Green *et al.*, 2002). However, it would be expected that there was not much difference in the information carried by these two conditions, and this sinusoidal modulator allowed condition NoiseAFD be more comparable to the condition NoiseFD.

2. ***AD*** (amplitude envelope and duration)

Sawtooth carrier: A random-frequency sawtooth carrier was generated by taking Tx randomly from the Tx values of the 64 original speech signals across four tones and two speakers, with the same duration as the original speech. The random-frequency sawtooth carrier eliminated the f0 cue to tonal contrasts. A fixed frequency sawtooth carrier was not used because a constant f0 would sound like tone 1, the level tone. The random-frequency sawtooth carrier was then multiplied by the amplitude envelope to generate the ***SawAD***-stimulus.

Noise carrier: A noise carrier was multiplied by the amplitude envelope extracted at 30 Hz to generate the ***NoiseAD***-stimulus.

3. ***FD*** (f0 and duration)

Sawtooth carrier: The ***SawFD***-stimulus was the f0-controlled sawtooth carrier, which preserved the f0 contour and the duration of the original speech but no variation in amplitude.

Noise carrier: The ***NoiseFD***-stimulus was the f0-controlled sinusoidally modulated noise. This stimulus contained the temporal cue to pitch with constant amplitude envelope.

4. ***AF*** (amplitude envelope and f0)

Sawtooth carrier: A linear time scaling was applied to the f0 contour so as to give a constant duration of 309 ms (the average duration of the voiced parts of all original speech signals), thus eliminating the duration cue. To create a sawtooth waveform with the original f0 contour but a fixed duration at 309 ms, the following procedure

was used. The fundamental periods measured from the laryngograph signal were converted to a pitch contour first. Then, a linear time scaling was applied to the pitch contour to produce a new pitch contour with the same shape but duration fixed at 309 ms. The new pitch contour was converted back to a new set of Tx periods. The first Tx period was the same as that in the original speech. The second Tx period was determined by the f0 value in the time-scaled pitch contour at the end of this first period. Similarly, the third and subsequent Tx periods were determined from the f0 value of the time-scaled contour at the end of the immediately preceding period. The new sets of Tx values preserved the original pitch contours without changing the overall f0 range.

The amplitude envelopes extracted at 30 Hz were also scaled in time to give the 309 ms duration, and then multiplied against the sawtooth carriers controlled by time-scaled f0 contour. The ***SawAF***-stimulus preserved both the f0 and amplitude contours in the original speech but with duration fixed at 309ms.

Noise carrier: A sinusoidal waveform was generated from the time-scaled Tx values and half-wave rectified, and then used to modulate a noise carrier. The time-scaled f0-controlled sinusoidal waveform was then multiplied by the time-scaled amplitude envelopes extracted at 30 Hz. The ***NoiseAF***-stimulus contained the same temporal pitch and amplitude information as in NoiseAFD-stimulus but without the duration cue.

5. ***A*** (amplitude envelope only)

Sawtooth carrier: A random-frequency sawtooth carrier with duration fixed at 309ms was multiplied by the time-scaled amplitude envelope extracted at 30 Hz, resulting in the ***SawA***-stimulus.

Noise carrier: A noise carrier was multiplied by the time-scaled amplitude envelope extracted at 30 Hz, resulting in ***NoiseA***-stimulus.

6. ***F*** (f0 only)

Sawtooth carrier: The ***SawF***-stimulus was the constant amplitude sawtooth carrier generated from the time-scaled Tx values.

Noise carrier: The ***NoiseF***-stimulus was the time-scaled f0-controlled sinusoidal modulated noise. This stimulus contained the temporal cue to pitch with a constant amplitude envelope and a fixed duration.

7. ***D*** (duration only)

Sawtooth carrier: The ***SawD***-stimulus was a fixed-amplitude random-frequency sawtooth carrier with duration varied as in the original speech.

Noise carrier: The ***NoiseD***-stimulus was a noise carrier with the duration of the original speech.

Pitch information available in different stimuli

For stimuli in which f0 information is conveyed by sawtooth carriers (SawAFD, SawAF, SawFD, and SawF), pitch percept can be evoked from both resolved lower harmonics and unresolved higher harmonics. For stimuli with f0 carried by noise carriers (NoiseAFD, NoiseAF, NoiseFD, NoiseF, and NoiseAFDcis), pitch can only be derived from regular fluctuations in the amplitude of the noise Since pitch perception in the normal auditory system is mainly determined by resolved lower frequency harmonics (e.g. Moore & Perter, 1992; Ritsma, 1967), stimuli SawAFD,

SawAF, SawFD, and SawF would be expected to give the most clear information about tonal contrasts.

3. Subjects

Eight normal hearing adults, three male and five female, participated in the experiment. All were native speakers of Mandarin from Taiwan, aged between 27 and 35.

4. Procedure

Stimuli were blocked by condition, so were presented in 15 blocks. A graphical user interface (GUI) built in MATLAB was used for running the experiment, and some examples of the interfaces are illustrated in Figure 3.3. In each block, a learning session and a training session were given before a testing session started. Subjects were given sample stimuli produced by the male and female speaker in the learning session, and these stimuli were played randomly in the training session with feedback given. Subjects were allowed to spend as much time as desired in these two sessions. No stimuli used for familiarization were used in the testing session. In the testing session, the four corresponding Mandarin characters of the stimulus were shown first, and the stimulus was played. The Mandarin characters with the Chinese phonetic alphabet, the jùyīnfúhaù, were also shown on the screen. Four push buttons labelled with Tone 1 to Tone 4 were presented below the corresponding characters. Subjects made their identification response using a computer mouse to click one of the four buttons. Then the four corresponding characters of the next stimulus were shown, and the next stimulus was played. No feedback was given in the testing session.

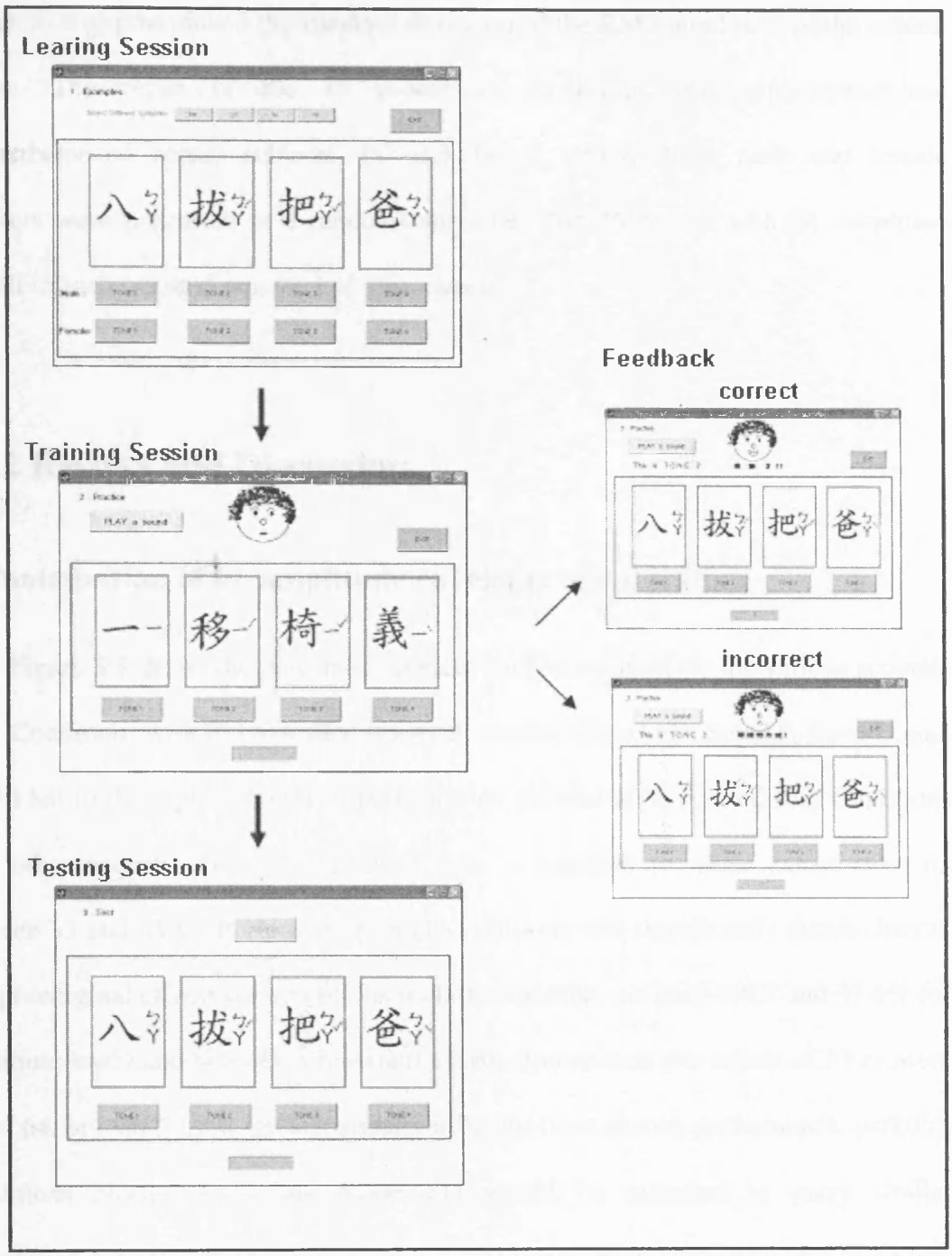

Figure 3.3 Examples of the Matlab GUI. The syllables /ba/ with different tones are shown here.

Stimuli were presented through Sennheiser HD 414X headphones, with intensity varied randomly within a 3dB range around the original level on each trial to eliminate cues derived from the overall intensity of the stimuli. The 3dB range was

chosen as it approximated the standard deviation of the RMS-amplitude of the natural tokens. The order of the 15 processing conditions was randomized and counterbalanced across subjects. In each block, stimuli from male and female speakers were presented in a randomized order. The 15 blocks with 64 simplified stimuli in each resulted in a total of 960 stimuli.

3.2.2 Results and Discussion

1. Contribution of f0, amplitude envelope, and duration

Figure 3.4 shows the percentage correct tone recognition for the various acoustic cues. Conditions with f0-controlled sawtooth carriers (SawAFD, SawAF, SawFD, and SawF) led to the highest levels of performance (around 90% correct), but conditions with other acoustic cues also allowed tone recognition to some extent (varying between 33 and 65%). Performance on all conditions was significantly above chance, except marginal effects for conditions with the duration cue only (36.3 and 33.6% for condition SawD and NoiseD; a binomial distribution reveals that scores of 23 or more out of 64, or over 35.9%, are statistically different from chance performance, $p<0.05$). Conditions NoiseAFDcis and NoiseAFD would be expected to carry similar information for tonal contrasts, and an *a priori* comparison confirmed that there was no significant difference between performances on these two conditions. *A priori* comparisons were also carried out to examine the effect of periodicity, amplitude envelope, and duration, by using contrasts of conditions with and without one of these three cues (for instance, comparison of conditions NoiseAFD/NoiseAF/NoiseFD vs. NoiseAD/NoiseA/NoiseD was used to examine the effect of periodicity; contrasts of NoiseAFD/AF/AD vs. NoiseFD/F/D, and NoiseAFD/FD/AD vs. NoiseAF/F/A were

used for amplitude envelope and duration, respectively). These revealed that there were significant differences for the periodicity and amplitude envelope cues ($p<0.01$), but not for the duration cue.

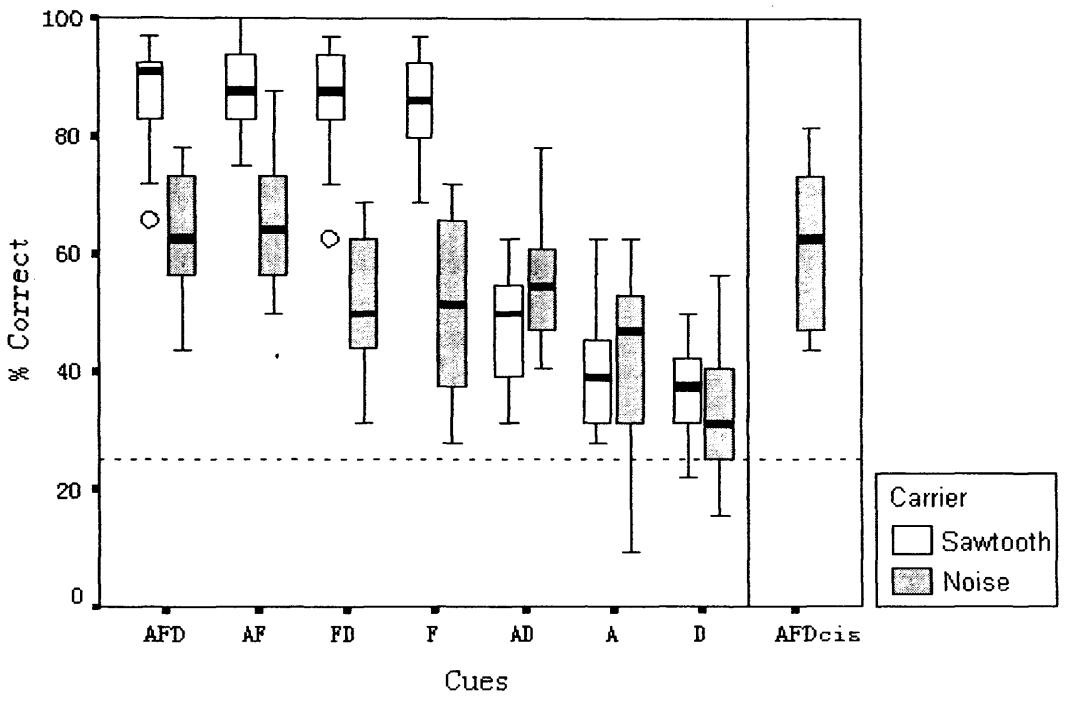

Figure 3.4 Boxplots of percentage of correct tone recognition across conditions of different acoustic cues carried by sawtooth or noise carriers. The box represents the 25- to 75-percentile range of the data over subjects and speakers, and the bar within each box represents the median. The whiskers represent the range of data points, except for outliers which are shown as asterisks (more than 3 box lengths from the box edge) or open circles (more than 1.5 box lengths). The dashed line represents chance performance (25% correct).

A repeated-measures analysis of variance (ANOVA) was performed for factors of cue, carrier, and speaker. This showed significant effects of carrier [$F(1, 7) = 131.3$, $p<0.001$] and cue [$F(6, 42) = 51.3$, $p<0.001$], and significant interactions of carrier by cue [$F(6, 42) = 27.3$, $p<0.001$] and speaker by carrier [$F(1, 7) = 15.0$, $p<0.05$]. No other two-way or three-way interaction was significant.

Interaction between carrier and cue

The interaction of carrier by cue was almost certainly due to performance with the two carriers differing greatly only in those conditions in which f0 information was presented (AFD, AF, FD, and F). Bonferroni-corrected *post-hoc* comparisons, with an alpha of 0.05, were used to examine the effect of the two different carriers on each of seven conditions, revealing that this was indeed the case. All conditions involving f0 information (AFD, AF, FD, and F) were significantly different, whereas conditions which had no f0 information (AD, A, and D) were statistically equal. This confirmed our expectation that a sawtooth carrier generated period by period from f0 provided better information for tone recognition than could possibly be conveyed by a modulated noise, with no significant difference for using a random-frequency sawtooth carrier and a noise carrier for conveying information about amplitude envelope and duration.

To examine the relative importance of the different acoustic cues to tonal contrasts, the effect of different combinations of cues on tone recognition was also compared, with sawtooth and noise carriers separately. Table 3.4 summarises the results. For the sawtooth carrier, no significant difference was found between the four conditions with salient f0 (SawAFD, SawAF, SawFD and SawF), and they were all significantly higher than the three conditions without f0 information. There was no significant difference between the two conditions with amplitude envelope cue presented (SawAD and SawA), and they were significantly higher than the condition with duration only (SawD). This indicated that information about f0 conveyed by a periodic sound contributed to tone recognition the most, and neither amplitude envelope nor duration affected tone recognition performance while f0 was presented. In the absence of salient f0, amplitude envelope appeared to contribute more to tone recognition than duration.

For noise carriers, only a few comparisons showed significant effects, even though several pairs of conditions had considerably different means (for instance, the mean score of NoiseAFD was more than 20% higher than that of NoiseA, 63 and 42% correct, respectively, yet, not significant). This was likely due to there being larger variations in some conditions than others, and the Bonferroni correction leads to relatively conservative tests.

	Sawtooth Carrier							Noise Carrier						
	AFD	**AF**	**FD**	**F**	**AD**	**A**	**D**	**AFD**	**AF**	**FD**	**F**	**AD**	**A**	**D**
AFD														
AF	-							-						
FD	-	-						-	*					
F	-	-	-					-	*	-				
AD	*	*	*	*				-	-	-	-			
A	*	*	*	*	-			-	-	-	-	-		
D	*	*	*	*	*	*		*	*	-	-	-	-	

*Table 3.4 Results of Bonferroni post hoc comparisons for pairs of conditions with different acoustic cues. The * symbol indicates a significance level of 0.05.*

Interaction between speaker and carrier

Figure 3.5.A shows the interaction of carrier and speaker gender. Bonferroni-corrected *post-hoc* comparisons revealed that performance for the male speaker was significantly higher than for the female speaker in conditions with noise carriers, but not with sawtooth carriers. Figure 3.5.B shows results from noise carriers only, with male and female separately. Subjects showed significant better performance for the male speaker in most conditions containing f0 information (three out of four, including NoiseAFD, NoiseAF, and NoiseF), but not in any condition without f0.

This is most likely due to the fact that temporal cues to voice pitch are less effective for a higher frequency range (e.g. Shannon, 1992; Green *et al.*, 2002; Green *et al.*, 2004). For instance, Green *et al.* (2004) used synthetic vowel glides, with central frequency approximately corresponding to male and female voice pitch, to investigate the perception of modulation frequency in noise carriers. Their results showed that the performance was clearly reduced as f0 increased.

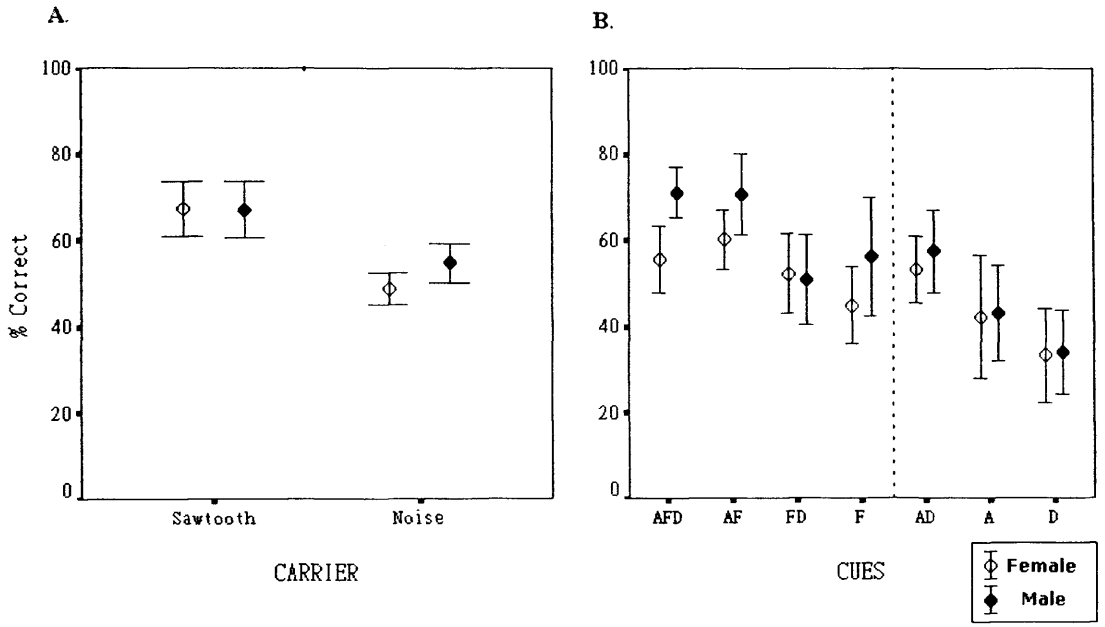

Figure 3.5 A: The interaction of carrier and gender. B: Recognition scores for male and female speakers across conditions with noise carriers. Error bars represent 95% confidence interval for the means.

2. Performance for the four tones

Studies have shown that the recognition for the four tones in Mandarin can vary with different acoustic cues available (e.g. Whalen & Xu, 1992; Fu *et al.*, 1998; Fu *et al.*, 2000). Here, performance for the four tones in conditions with different combinations of cues was further examined. To give an unbiased measure of

identification performance for individual tones, information transfer scores were computed from tone confusion matrices (e.g., a 2x2 matrix classifying stimuli as tone 1 vs. all other tones, and responses in the same way). The percentage of information transferred for the four tones across conditions is shown in Figure 3.6.

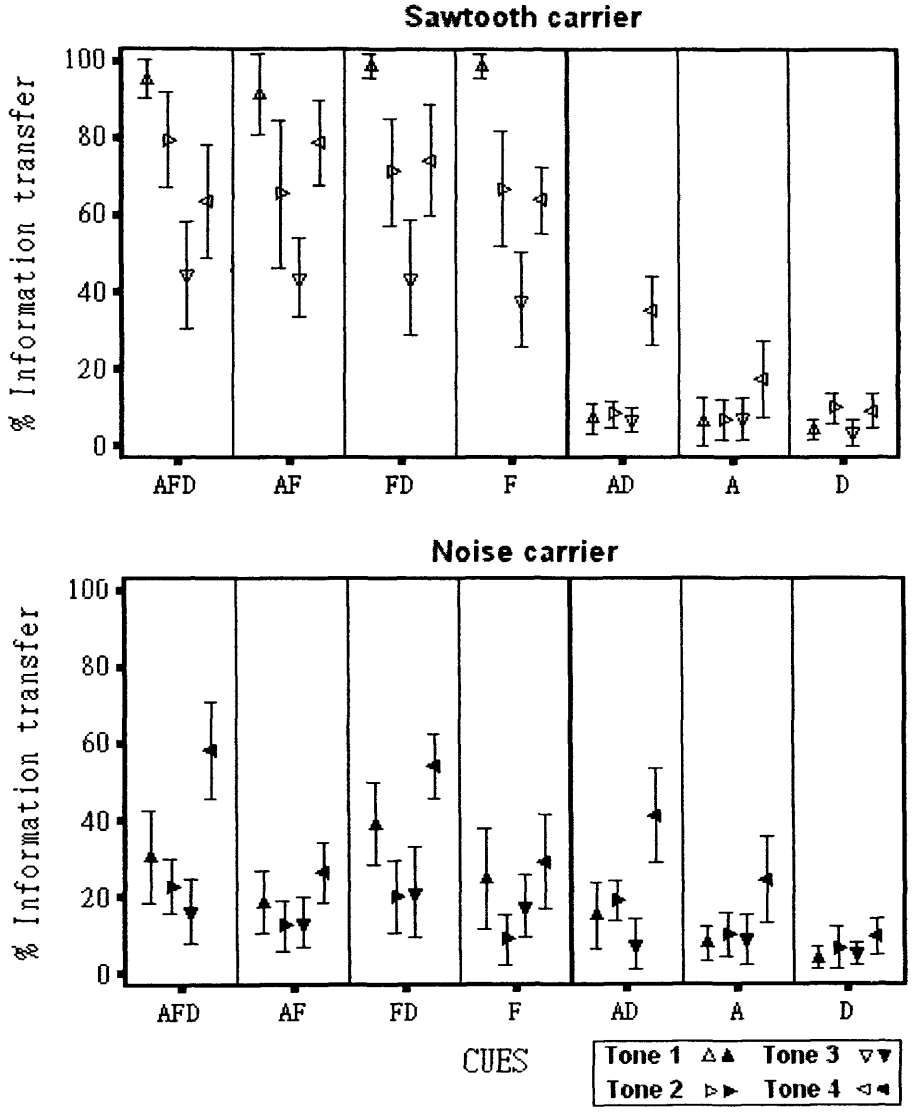

Figure 3.6 Percentage of information transfer score of tone recognition for four tones across conditions with different combinations of three acoustic cues. Performance on conditions using the sawtooth carrier is displayed in the upper panel, and performance on conditions using the noise carrier is displayed in the lower panel. Error bars indicate the 95% confidence interval for the means averaging over 16 data points (8 subjects x 2 speakers).

A repeated-measures ANOVA was performed for factors of tone and condition, demonstrating a significant interaction [$F(39, 273) = 18.2$, $p<0.001$], and significant main effects for condition [$F(13, 91) = 96.5$, $p<0.001$] and tone [$F(3, 21) = 65.7$, $p<0.001$]. Bonferroni-corrected *post-hoc* comparisons were used to examine performance of the four tones on conditions with different combinations of cues. For conditions with explicit f0 included (SawAFD, SawAF, SawFD and SawF), information transfer scores for the four tones showed similar patterns: tone 1 was significantly higher than tones 2 and 4, and there was no significant difference between these two tones. All these three tones were significantly higher than tone 3. Tone 1 might be recognised best because it had a higher and more distinct frequency range compared to the other three tones. As for tone 3, the variation of its realisation presumably was responsible for its relatively low scores. In the two conditions with the duration cue only (SawD and NoiseD), information transfer scores were all very low, with not much difference among the four tones. For the rest of conditions with little or no f0 presented, tone 4 was generally recognised well compared to the other three tones.

3. The use of amplitude envelope for tone recognition

Table 3.5 shows subject response for stimuli with the amplitude envelope cue only. The confusion matrices were summed over conditions SawA and NoiseA for the male and female speaker separately. Percent correct for each of the four tones is shown along the diagonal of each table. Results from previous studies have shown that listeners were able to use the amplitude cue to label tones to some extent, though the performance based on the amplitude envelope cue was often highly variable

across tones and speakers. Whalen and Xu (1992) suggested that listeners either recognized the consistent correlation between pitch and amplitude contours or interpreted amplitude change as f0 change. The variation in performance might arise from the fact that a significant correlation between pitch and amplitude contours was only found for certain tones and speakers (Whalen & Xu, 1992; Fu & Zeng, 2000). Figure 3.7 shows the average pitch and amplitude contours of the four tones for the male and female speaker for stimuli used in present study (each pitch/amplitude contour was averaged over eight normalised contours with the duration of 309ms). There were some general similarities between the f0 and amplitude contours. To examine if subject response could be explained by the similarity of pitch and amplitude contours, the correlations between pitch and amplitude contours were calculated and compared with the frequency of subject response for stimuli with the amplitude envelope cue only.

Stimulus	Response to **male** speaker				Response to **female** speaker			
	1	2	3	4	1	2	3	4
TONE 1	**31.3**	43.0	16.5	9.4	**35.9**	28.9	16.4	18.8
TONE 2	15.6	**48.4**	25.8	9.4	19.5	**42.2**	26.6	11.7
TONE 3	7.8	44.5	**38.3**	9.4	17.2	19.5	**25.8**	37.5
TONE 4	23.4	10.2	13.3	**53.1**	17.2	9.4	16.4	**57.0**

Table 3.5 Response matrices (in percentage) for stimuli with the amplitude envelope cue only (Results were summed over conditions SawA and NoiseA).

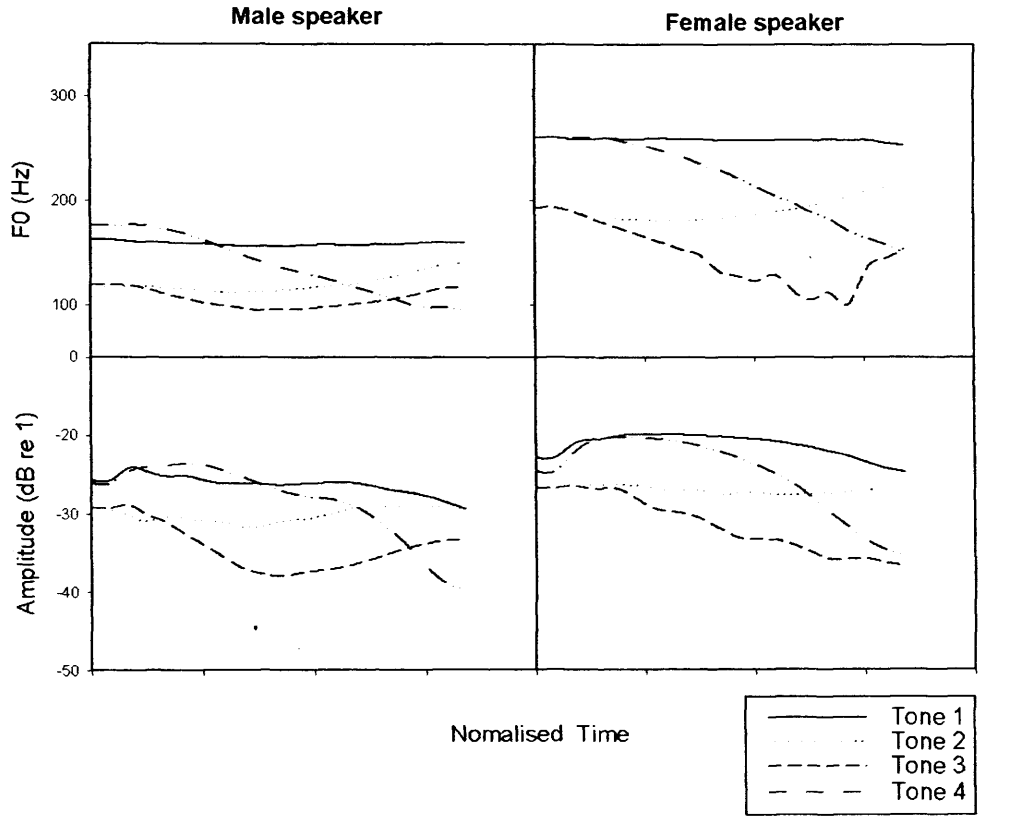

Figure 3.7 Average pitch and amplitude contours for the four tones for the four speakers. Each line was averaged over 8 speech tokens after a linear time scaling.

The similarity of the amplitude contour of one tone to the pitch contour of each of the four tones was examined by calculating a cross-correlation. Normalized amplitude and pitch contours with the duration of 309ms were used. To obtain approximately independent samples, points along a contour were sampled separated by an interval determined by calculating autocorrelation coefficients for pitch and amplitude contours of several sentences. Since a sample point and its neighbor become less related to each other when they are further apart, the autocorrelation coefficients were often not significant when the time lag was over 55ms. Therefore, six samples for each of normalized amplitude and pitch contours were calculated by dividing one contour into six sections and taking the average value in each section. Correlations for the lags of –1, 0, and +1 were calculated, and the maximum value

was selected. Table 3.6 shows the average correlation between the amplitude contour of one tone and the pitch contours of the four tones. The amplitude contour of a tone did not always correlate most highly with the pitch contour of that same tone. For instance, the amplitude contour of tone 1 for the male speaker was highly correlated to the pitch contour of tone 4 (r =0.69) rather than that of tone 1 (r = 0.38).

	male speaker				**female** speaker			
	F0				F0			
Amp	1	2	3	4	1	2	3	4
TONE 1	**0.38**	-0.07	0.22	0.69	**0.62**	-0.01	0.69	0.72
TONE 2	0.34	**0.68**	0.45	0.04	0.33	**0.45**	0.33	0.25
TONE 3	0.43	0.40	**0.78**	0.51	0.43	-0.02	**0.79**	0.85
TONE 4	0.41	-0.01	0.37	**0.77**	0.49	-0.19	0.88	**0.89**

Table 3.6 Cross correlations between amplitude contours and pitch contours for male and female speakers.

To examine whether response frequencies to stimuli with only the amplitude envelope cue (Table 3.5) could be explained by the similarity between pitch and amplitude contours (Table 3.6), a scatterplot of subject responses (y-axis) paired with the correlation of pitch and amplitude (x-axis) of each tone is shown in Figure 3.8. Generally speaking, high response frequencies were associated with high correlation coefficients. Although a high correlation coefficient did not always lead to a high proportion of subject responses, a low correlation coefficient seldom did. This indicated that a certain relationship between identification responses and the correlation between amplitude and pitch contour existed. Note that, as shown clearly in the figure, this varied with different tones. Since the pattern of responses for the four tonal labels was significantly different ($p<0.001$), regression lines were plotted

separately for the four tones. The response to tones 2 and 4 showed highly positive correlation, while the response to tone 3 had only a slightly positive correlation. The response for tone 1, however, showed a weak negative correlation. For tones 2 and 4, the correlation between amplitude and pitch contours may account for the role of the amplitude envelope cue in the recognition of these two tones.

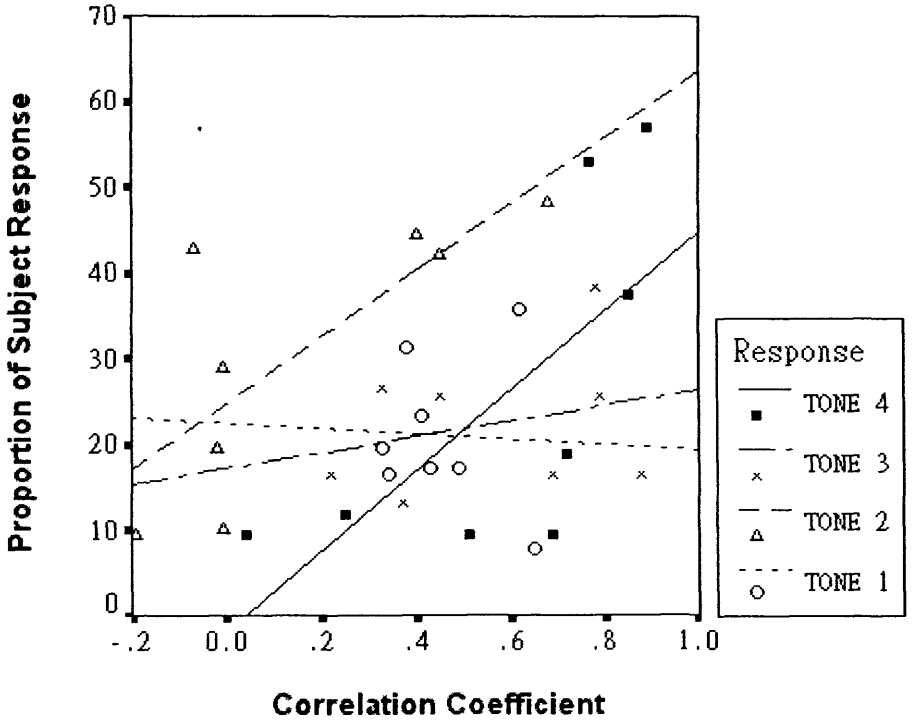

Figure 3.8 Scatterplot of the proportion of responses for each possible tone label against the correlation between pitch and amplitude contours. Each point was made by a value in Table 3.5 (y-axis) paired with a value in Table 3.6 (x-axis). For instance, the four points of the responses for tone 1 for the male speaker were the values in the first column in the former table paired with the values in the first column in the latter table. There were 32 points in total, 8 points for each of the four tones. The four regression lines were fit for each tone label. The Pearson correlation coefficients for tones 1, 2, 3, and 4 are -0.04, 0.75, 0.28, and 0.69, respectively. Correlations for tones 2 and 4 are statistically significant ($p<0.01$)

4. Implications for cochlear implants

While lexical tones are mainly distinguished by their f0 patterns, early studies often suggested that other acoustic cues such as amplitude and duration were of little importance for tone recognition. Recent studies have reported that these temporal cues have a more substantial contribution when f0 information is weak or completely absent. However, results from the present study clearly demonstrated that performance with any combination of these temporal cues was still much lower than with explicit f0 information. The results here might be used to indicate the extent to which information about tonal contrasts could possibly be obtained by current cochlear implant users with CIS-like speech processors, and what may possibly be achieved in future devices. While voice f0 is indubitably the most essential cue for recognising tonal contrasts, it is not transmitted sufficiently well through current implants. The f0 information available in implant systems is very similar to that conveyed by temporal fluctuations of a noise carrier in the present study (see Chapter 2 for more details about the signal processing in cochlear implants). Without the presence of explicit f0 information, this temporal pitch information could help tone recognition, as could amplitude envelope and duration. However, even with all available acoustic cues, tone recognition performance was hardly above 70% correct. This is consistent as shown in clinical results, which have reported that users of tonal languages often encounter great difficulty in tone recognition.

While these temporal cues play a more important role for implant users than normal-hearing listeners, none of them is always reliable. Listeners often vary greatly in their ability to make use of these temporal cues, as shown in this study (see Figure 3.4) and in previous research. For instance, in Green *et al.* (2002), glide labelling performance based on temporal envelope cues was limited and varied greatly across

subjects. Also, the amplitude envelope and duration cues in Mandarin have been reported to be highly variable across speakers and different tones (Fu & Zeng, 2000). The performance in those conditions with noise carriers may represent what can be achieved when listeners are able to perceive one, two, or all of these three cues. The average performance across conditions varies greatly between 35 and 65%, and this is about the level of performance observed in many implant users. While these results are far from satisfactory, in the real life situation, these temporal cues are likely to be even less salient.

3.3 Summary

- As would be expected, explicit f0 cue contributed to tone recognition the most, irrespective of the presence of amplitude envelope and/or duration cues.
- The temporal f0, amplitude envelope and duration also contributed to tone recognition to some extent in the absence of the explicit f0, but these cues were highly variable across speakers and tones.
- The contribution of the amplitude envelope cue to the identification of certain tones, especially for tone 4, arose from the relatively high correlation between their amplitude and pitch contours.

Chapter 4

Effects of voice f0 on sentence recognition:

Implications for cochlear implants

Voice f0 is essential, for instance, to recognise statement/question contrasts, to identify stressed words, to appreciate music melody, and to identify paralinguistic information such as speaker's age and gender. It is also important for young children in the early development of spoken language (e.g. Fernald & Simon, 1984; Jusczyk, 1997; Snow and Ferguson, 1977). For users of a tone language, voice f0 plays an even more significant role in conveying lexical meanings. The first study reported in chapter 3 demonstrated that performance on tone recognition with a clear indication of voice pitch is much higher than with other acoustic cues. In this chapter, the role of voice f0 in understanding running speech was further investigated using vocoder techniques. The amount of spectral information was manipulated systematically to examine the effect of voice f0 with various degrees of spectral information. The results allow us to determine the importance of voice f0 in a tone language when spectral information is degraded. Furthermore, the vocoded speech is processed in a similar manner to current cochlear implants, and has been used in normal-hearing listeners to simulate what implanted listeners can possibly achieve from the information obtained from their devices (e.g. Shannon *et al.*, 1995; Dorman *et al.*, 1997a; Faulkner *et al.*, 2000). This acoustic model has also indicated how implant users of a tone language might benefit if better voice pitch information is provided (e.g. Lan *et al.*, 2004; Xu & Pfingst, 2003).

Here, the information about voice f0 in current cochlear implants will be discussed first, and recent studies focusing on f0 enhancement in implant systems will be reviewed. The main experiment of the effect of f0 will then be described, followed by a sub-experiment with further control of amplitude variations. Finally, the contribution of natural f0 in a tone language and its implications for cochlear implants will be discussed.

4.1 Introduction

Temporal pitch information in CIS processing

Cochlear implants have been used widely in profoundly hearing-impaired people around the world, and help them in the successful restoration of some hearing. One of many limitations in current implant devices is in providing information about voice f0. There is only limited information about f0 provided by the most popular CIS speech coding strategy (Continuous Interleaved Sampling, Wilson *et al.*, 1991). In CIS, speech signals are analysed into a number of frequency channels. The amplitude envelope in each channel is extracted and used to modulate a high frequency pulse train carrier. Because of the relatively large frequency range assigned to each channel, the lower harmonics of input signals are unresolved. Those spectral cues from resolved lower harmonics, which are used to perceive pitch information in normal hearing, are thus no longer available. Voice f0 information can still be encoded in implant systems by temporal fluctuations of speech envelopes if the following two conditions are met: (1) the lowpass filters for envelope extraction include the voice f0 range; (2) the carrier frequency is high enough, normally 4 to 5 times higher than the

modulation frequency (e.g. Busby *et al.*, 1993; Wilson, 1997). However, this temporal pitch information is considered to be relatively weak (Green *et al.*, 2004).

The contribution of temporal pitch information to speech perception has been investigated in a number of studies. Shannon *et al.* (1995) used noise-excited vocoder simulations to examine the contribution of spectral and temporal cues to speech recognition by varying the number of frequency bands and the cutoff frequency of envelope filters. The number of noise bands was varied from 1 to 4, and combined with cutoff frequencies of envelope-smoothing filters at 16, 50, 160, and 500 Hz. The number of noise bands showed a significant effect; recognition scores for vowels, consonants, and sentences increased significantly with the number of noise bands. Those conditions with a cutoff frequency at 50, 160, and 500 Hz showed significantly higher recognition scores in consonant and sentence tasks than the conditions with a cutoff frequency at 16 Hz. No significant difference was found among conditions with 50, 160, and 500 Hz cutoff frequencies, although conditions with higher cutoff frequencies, such as 160 and 500 Hz, were considered to contain more information for voice pitch.

Fu *et al.* (1998) applied similar signal processing as in Shannon *et al.* to a study with Mandarin speakers, using only two cutoff frequencies, 50 and 500 Hz, for envelope-smoothing filters to control the information about voice pitch. Consistent with the results in Shannon *et al.*, the cutoff frequency did not show a significant effect in consonant and vowel recognition. However, in sentence and tone recognition, significantly higher recognition scores were found in conditions with a 500 Hz cutoff frequency. For instance, with the voice f0 range included in the envelope filter, recognition scores for tones and sentences increased by around 10 and 15 percentage points, respectively, in the four-band condition. They also reported that,

without spectral variations (1-band condition), sentences were recognised about 11% correct in Mandarin, compared to only 3% correct in English. A further examination of the relationship between performance for tones and sentences suggested that sentence recognition in Mandarin was contributed to by a high level of tone recognition. Another study by Xu *et al.* (2002) further confirmed the significant effect of envelope cutoff frequency to tone recognition. In their study, the cutoff frequency of the envelope-smoothing filter was systematically varied between 1 to 512 Hz in 1-octave steps. Performance in tone recognition improved consistently as the cutoff frequency of the envelope filter increased.

The results from Shannon *et al.* (1995) and Fu *et al.* (1998) indicated that although temporal pitch information had no significant effect on speech recognition in English, it clearly made a significant contribution to Mandarin. However, this does not necessarily mean that voice pitch has no effect at all in English, but rather that the speech tasks used are insensitive to the change of f0 information carried by temporal fluctuations. A study by Hillenbrand (2003) reported some effects of pitch contour on sentence intelligibility in English. Synthetic sentences were imposed with three different pitch contours (original, monotone, or inverted) generated from a source-filter synthesizer. Compared to sentences with original pitch, subject performance had a small but significant decrease for sentences with monotone and inverted pitch with no difference between the latter two conditions. Even larger reductions were found when spectral information was removed by low-pass filtering at 2 kHz, and sentence intelligibility reduced more for inverted pitch than flat pitch.

In general, pitch information carried by temporal fluctuations is rather weak, compared to explicit pitch information in truly periodic sounds (Burn & Viemeister, 1976; Green *et al.*, 2004). To examine how voice pitch information would affect

speech perception in English, especially the effect of explicit pitch, Faulkner *et al.* (2000) constructed a number of 4-channel simulations of CIS processors with different degrees of pitch and periodicity information. Two processors were noise-excited processors with envelope-smoothing at 32 and 400 Hz, and another three used either f0-controlled pulse trains or fixed-frequency pulse trains for the voice source. The most salient pitch and periodicity information was carried by the processor with f0-controlled pulse trains for voiced speech and a noise carrier for voiceless speech (FxNx). The noise excitation processor with 400 Hz cutoff frequency for the envelope-smoothing filter (Noise 400) contained also both pitch and periodicity, but in a relatively weak form. The other three conditions contained only periodicity information (VxNx) or neither pitch nor periodicity information (Mpulse and Noise32). No significant difference was found among different processors in recognition of vowels, consonants, sentences, and connected discourse tracking but subjects showed significantly better performance on pitch glide labelling with the FxNx processor. Although no significant effect of voice pitch was found in the speech tests, as the authors pointed out, it should not be interpreted that voice pitch is unimportant. It rather reflected the fact that these speech intelligibility tests lacked sensitivity to essential information such as voice pitch.

Green *et al.* (2002) further investigated the perception of the pitch movement for sawtooth glides and synthetic diphthong glides by using one- and four-band noise-excited vocoders with envelope-smoothing filters at 32 and 400 Hz. Subjects performed better in identifying the pitch movement of sawtooth waveforms than that of synthesized diphthongs. It was likely due to the formant movement of the diphthongs obstructed the perception of pitch movement, and this suggested that the spectral movement in real speech would make voice pitch even harder to perceive.

Enhancing voice pitch in CIS processing in implant users

Since the pitch information provided by CIS processing is rather weak, therefore, those above studies found only limited effect with the presence of voice pitch information. Some studies have attempted to improve voice f0 information more directly in CIS speech processors. One approach is to enhance temporal cues to voice pitch. Geurts and Wouters (2001) extracted envelopes by half-wave rectification, which might lead to better phase-locked activity in the nervous system, and forward and backward low-pass filtering so as to eliminate phase distortion. The fluctuating envelope extracted by 400 Hz low-pass filtering was then subtracted by the relative flat envelope extracted by 50 Hz low-pass filtering, so as to increase the modulation depth. This condition, called F0 CIS, was compared with two conditions using the default 400 Hz filtering for the envelope (CIS) and 50 Hz filtering for the envelope (FLAT CIS). Four LAURA cochlear implant users, who were able to discriminate to some extent of the change of modulation frequency of a SAM (sinusoidally amplitude modulated) pulse train on one electrode pair, were presented with two synthetic vowels and asked to identify the one with the higher f0. They found no difference between the F0 CIS and the standard CIS, but they were both significantly better than the FLAT CIS.

Green *et al.* (2004) enhanced temporal pitch cues by decomposing the envelope into two components: one with slow rate information conveying the dynamic changes in spectral shape and another with periodicity information carried by a simplified waveform. Both normal-hearing subjects, listening to vocoder simulations, and implant users, were asked to identify the pitch direction of synthetic diphthong glides, and their results showed a similar pattern. Performance in the modified CIS

processing was significantly better than in the standard CIS processing. However, the improvement was rather small. Subjects showed better performance for stimuli in a lower frequency range, and their performance decreased with increasing glide centre frequency.

Another approach to improve voice pitch is to achieve better *place*-coding of f0 in an implant. Geurts and Wouters (2004) constructed a new filter bank for signal processing, in which the first harmonic of a complex sound was always resolved in two adjacent filters. Based on the study of McDermott and McKay (1994), the frequency of the first harmonic could be perceived from the relative output levels of the two adjacent electrodes. Synthesized vowels with different f0 were used to examine four LAURA implant users for their ability to discriminate the smallest difference in f0. In general, thresholds for detecting f0 differences were lower with the modified filter bank, and decreased more in the absence of temporal cues. However, Green *et al.* (2005) pointed out that it remains unclear if this advantage would remain in more natural speech. The effectiveness of place coding to f0 might be obscured by spectral variations in real speech.

In short, information for voice f0 in current implant systems is limited. Even though some significant improvement has been reported in studies with manipulations of the signal processing algorithms for current implant devices, the benefit is often small.

4.2 Experiment II: Role of voice f0 on sentence understanding by child and adult tone language users

Aims and experimental predictions

In contrast to the results in English, which showed little or no effect of f0 in recognising running speech, voice f0 plays a much more significant role in tone languages. The present study adapted the method used in Faulkner et al. (2000) to examine the effect of voice f0 in sentence recognition in Mandarin. Vocoded speech was generated with either a source carrier preserved (the original voice f0 contour of natural speech), or a source carrier with a slightly falling f0 contour, irrelevant to natural f0 variations (FxNx and VxNx carriers, respectively; more details about signal processing can be found in section 4.2.2). Given that voice f0 was the most important cue to signal tonal contrasts, it is expected that sentences contain the original f0 information (FxNx sentences) would be recognised better than those with neutralised slightly-falling ones (VxNx sentences).

4.2.1 Speech stimuli

A total of 240 sentences were selected from the BKB Standard Sentences (Bamford-Kowal-Bench sentences; Bench & Bamford, 1979) and the IHR sentences (MacLeod & Summerfield, 1990)[5], and translated into Mandarin. Some sentences were modified so as to be closer to language usage and culture in Taiwan. A male and

[5] BKB lists comprise 16 simple sentences with either 3 or 4 key words. Each BKB list includes 50 key words. IHR list comprise 15 sentences, each with three key words, giving a total of 45 key words per list. IHR sentences are developed based closely on the BKB sentences. The BKB and IHR sentences are therefore highly equivalent.

a female who were native speakers of Mandarin from Taiwan recorded the sentences. The 120 sentences produced by the male speakers were from lists 1, 3, 5, 7, 9, 11, 13, 15, 17, and 19 in the BKB and lists 1, 3, 5, 7, and 9 in the IHR, and the 120 sentences produced by the female speaker were from Lists 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 in the BKB and lists 2, 4, 6, 8, and10 from IHR.

4.2.2 Signal processing

Figure 4.1 illustrates the signal processing procedure. Speech signals were passed through a bank of bandpass filters (third-order Butterworth filters), in which the amplitude envelope of each channel was extracted by full-wave rectification and low-pass filtering with a 30Hz cutoff frequency fourth-order Butterworth filter. The number of channels was **2**, **4**, **8**, and **16**, and the cutoff frequencies for each channel were calculated by the Greenwood equation (1990). The overall frequency range was from 100 to 5500 Hz. Table 4.1 shows frequency ranges and centre frequencies of analysis filters for processors with different number of frequency channels.

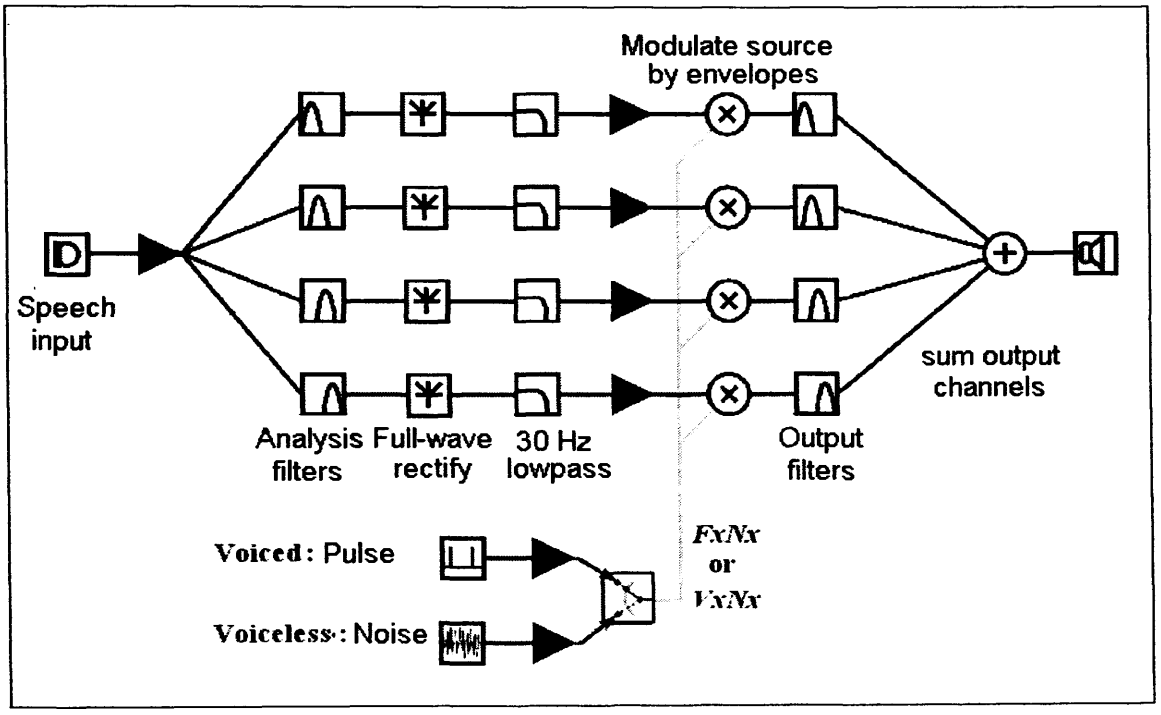

Figure 4.1 Signal processing for 4-channel acoustic simulations.

Channel number	Channel	Centre freq.	Lower freq.	Higher freq.
2 channels	1	*405*	100	1061
	2	*2470*	1061	5500
4 channels	1	*224*	100	405
	2	*671*	405	1061
	3	*1632*	1061	2470
	4	*3699*	2470	5500
8 channels	1	*156*	100	224
	2	*306*	224	405
	3	*525*	405	671
	4	*847*	671	1061
	5	*1319*	1061	1632
	6	*2011*	1632	2470
	7	*3026*	2470	3699
	8	*4514*	3699	5500
16 channels	1	*127*	100	156
	2	*188*	156	224
	3	*263*	224	306
	4	*353*	306	405
	5	*462*	405	525
	6	*595*	525	671
	7	*755*	671	847
	8	*949*	847	1061
	9	*1184*	1061	1319
	10	*1468*	1319	1632
	11	*1813*	1632	2011
	12	*2230*	2011	2470
	13	*2735*	2470	3026
	14	*3346*	3026	3699
	15	*4087*	3699	4514
	16	*4983*	4514	5500

Table 4.1 Frequency range and centre frequency in each channel for 2- to 16-channel processing

The presence of voice pitch information was controlled by using either an f0 controlled pulse carrier or a pulse carrier with a slightly falling pitch contour for voiced speech. For voiceless speech, a random noise was used as carrier. The pulse and noise carriers had the same rms level. The condition with the f0 controlled pulse carrier was referred to as **FxNx**, and the other as **VxNx**. These conditions were similar to two used in the study of Faulkner *et al* (2000). The f0 of speech inputs was measured by the STRAIGHT programme (Kawahara). A slightly falling pitch contour for the VxNx pulse train was intended to provide a more similar intonation to natural

speech (Tseng, 1990). The pitch values of a falling contour were calculated according to the pitch range of each speaker. The initial frequency of the falling pitch contour was a random value from the range of approximately one standard deviation around the mean f0 for each speaker (the range was140 $\pm$8 Hz for the male speaker and 240 $\pm$ 14 Hz for the female speaker). The end frequency of the falling pitch contour was the initial frequency decreased by 6%. The transition was linear in frequency. Figure 4.2 shows f0 contours for FxNx and VxNx sentences processed from the same male-spoken sentence.

Pitch information available in different stimuli

Both FxNx and VxNx vocoded sentences provided clear information about pitch and periodicity. Listeners should be able to perceive the pitch variations of sentences from both resolved and unresolved harmonics. However, the FxNx sentences contained natural pitch variations which provided clear information about tonal contrasts, whereas the VxNx sentences constantly had slightly falling pitch contours which were irrelevant to the original tonal contrasts.

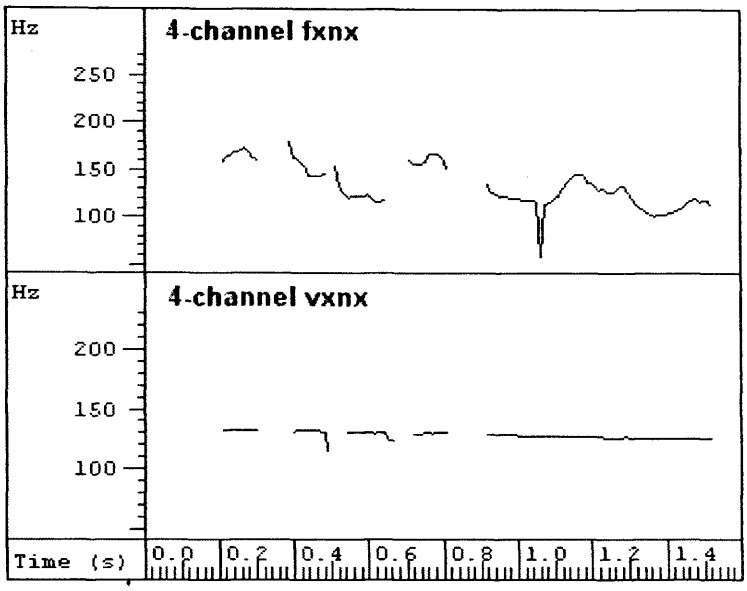

Figure 4.2 Examples of f0 contours for sentences with natural f0 contour (FxNx sentence) and with slightly falling f0 contour (VxNx sentence).

There were eight conditions in this study: four different numbers of channels (2-, 4-, 8-, and 16-channel) combined with two carrier sources (FxNx and VxNx). Both male and female speech was investigated in this study. Therefore, eight male-spoken and eight female-spoken lists were needed. To obtain 16 homogeneous lists, the following calibration was carried out.

4.2.3 Calibration

A. Stimuli

Pilot studies showed low performance with 4-channel simulations, so only 8-channel simulations were used for the calibration study. All the 240 sentences were processed using 8-channel acoustic simulations with FxNx and VxNx carrier source,

resulting in 480 processed sentences. The VxNx and FxNx sentences processed from the same sentence were tested by different subject groups.

B. Subjects

All children were recruited in Taiwan. 10-year-old normal-hearing children, who were native speakers of Mandarin and had no known history of hearing problems or language impairment, were recruited for the calibration. Group testing was conducted in order to obtain sufficient data for the calibration. Each group of children attended the study for one hour, and sixty sentences were given (30 VxNx and 30 FxNx sentences). All 480 sentences were completed by 8 different groups, with each consisting of between 27 and 32 children.

C. Procedure

The testing was conducted in a classroom. Sentences were presented through a loudspeaker, and children were asked to write down what they heard on provided answer sheets. They were encouraged to guess. Some practice sentences were given before the testing to familiarise the children with the stimuli. In the testing, all the VxNx sentences were given first, then the FxNx sentences. A sentence was played only when all children finished their writing for the previous sentence. Several short breaks were given during the testing.

D. New Sets of Sentence Lists

To obtain homogeneous sentence lists, sentences were assigned to lists by the following method, with male- and female-spoken sentences rearranged separately. Firstly, sentences were sorted by their scores in the FxNx condition. The first eight

sentences with the 1st, 2nd, 3rd, ..., 8th highest scores were assigned to lists 1 to 8 respectively. Then, the next eight sentences with the 9th, 10th, ..., 16th highest scores were assigned to lists 8 to 1 respectively. Similarly, the rest of sentences were allocated to lists, resulting in eight lists with approximately the same FxNx scores. Finally, some sentences were moved among lists, so as to achieve each list with roughly the same VxNx scores and to avoid similar sentences appearing in one list.

The above procedure produced sixteen new sentence lists with approximately equal average scores for FxNx and VxNx sentences. These new sets of sentences are listed in Appendix 2. Figure 4.3 shows the percent recognition scores for these new sentence lists, eight produced by a female speaker (f1 to 8) and eight produced by a male speaker (m1 to 8).

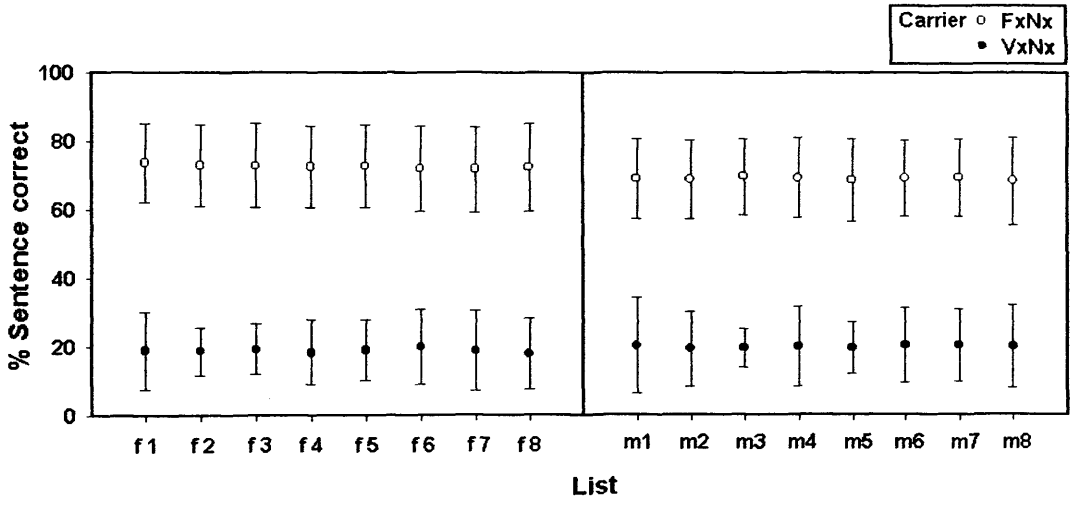

Figure 4.3 The percent recognition scores for new sets of sentence lists, 8 male-spoken lists and 8 female-spoken lists. Error bars show 95% confidence for the means.

E. Statistical Analysis for New Sets of Sentence Lists

A logistic regression was performed on the total number of keywords perceived correctly summed over the set of listeners for each sentence. The saturated model included the factors of speaker (male, female), f0 (the presence/absent of f0 variations) and list (1-8) plus all the interactions. None of the interactions involving list, nor list as a main effect, was significant. The interaction of speaker and f0 was highly significant, however ($p<0.0001$), leading to the following model (Table 4.2):

parameter	estimate
1	0.98
speaker	-0.17
f0	-41.66
speak * f0	0.25

Table 4.2 The parameters and their coefficients for the final model

Inspection of the data suggests that the interaction term arises because the male speaker is slightly less intelligible than the female for condition FxNx, but not condition VxNx.

The fact that there was no significant effect of list in the final model, indicates that these sentence lists were equal in their difficulty both in conditions with and without voice pitch information.

4.2.4 Subjects

All children were recruited in Taiwan. Four age groups of normal-hearing subjects, ages 6, 9, 12, and 20, were recruited in the study. There were ten subjects in each group. All of them were native speakers of Mandarin with no known history of hearing or language problems.

4.2.5 Procedure

All subjects were tested individually. For the three groups of children, testing was carried out in a quiet room in school. There were curtains and carpets in the testing environment which would absorb some reverberation. Sentences were presented through a loudspeaker which was placed in front of the child at a distance of around 1 metre and at the height of around 1 metre above the ground. Children were asked to repeat what they heard and their responses recorded on tapes. For adult subjects, a graphical user interface (GUI) built in MATLAB was used for the testing, and subjects were allowed to proceed at their own speed. Sentences were presented through Sennheiser HD414X headphones, and subjects were asked to write down what they heard on provided answer sheets. All subjects were encouraged to say or write down as many words as possible for their answer, and to guess. Sentences were presented in a randomised order. A training session and a practice session were given before the testing started.

4.2.6 Scoring method

The 'loose keyword' method used for scoring the BKB and IHR sentences was applied to score subject's response. This method scores the answer as correct when keywords are reported correctly, ignoring errors such as declension. For instance, to the stimulus "The postman brings a letter" (key words were underlined), all three key words would be scored correctly for the answer "Postmen bring letters', and two would be scored for the answer "The postman has a letter". Chinese characters are orthographic, and words are mainly made of one to three characters. Key words were still scored correct for small mistake in writing a character. A score was also given for

a few words which did not change its meaning with a different suffix. For example, the response "dài lè" was scored as correct to the key word "dài tze" since they meant the same, *brought*, in the sentence 2.13 in List f2 ("The thief brought a ladder")

4.2.7 Results and discussion

Logistic regression was used to determine a statistically adequate model for describing the way in which performance depended upon the age of the listener, the number of bands, the speaker, and the presence or absence of variations in fundamental frequency. Both age and number of bands were treated as continuous variables. It seemed likely that these variates would lead to better predictions with some kind of compressive transformation (e.g., there is good evidence that performance increases more or less linearly with the logarithm of the number of bands, and improvement over age would be expected to be much greater at younger ages than at older). Therefore, the Box-Cox transformation ($\frac{x^{\lambda}-1}{\lambda}$, a useful family of transformations, including logarithmic, controlled by a single parameter) was applied to both age and bands, varying the single parameter to find the best fit for a logistic regression using a saturated model (all 4 main effects and interactions). These values were then used for all further regressions (λ = -2.5 for age and -0.3 for bands).

Model fitting proceeded from a fully saturated model, excising terms sequentially, that were not significant at the $p<0.05$ level using changes in the deviance. Because even the saturated model displayed a greater variability than would be expected from a binomial model, methods appropriate for such so-called overdispersion were applied (as detailed in Collett, 2003, pp. 206-210).

The final model included the terms in Table 4.3. Also given are the value of the coefficients and their standard errors. The final deviance had a value of 1229.3 with df = 630. The predicted performance can be calculated by

$$P(z) = \frac{1}{1+e^{-z}}, \qquad \text{where } \mathbf{z} \text{ is the linear predictor}$$

('age' and 'bands' need to be transformed first; 'fx' = 0 or 1 for the absence or presence of f0 variations; 'speaker' = 0 for a female speaker and 1 for a male speaker.)

parameter	estimate	(s.e.)
1	-134.7	(7.54)
bands	3.09	(0.108)
age	326.6	(18.87)
speaker	-41.66	(11.29)
f0	0.98	(0.21)
speaker * bands	1.66	(0.18)
speaker * age	99.30	(28.23)
f0 * bands	0.55	(0.15)
speaker * f0	1.05	(0.33)
speaker * f0 * bands	-1.07	(0.24)

Table 4.3 The parameters and their coefficients for the best fit model. The standard errors are presented in parentheses. These values have also been adjusted for overdispersion in the saturated model (Collett, 2003).

Subject performance for the four age groups and the prediction of the model are shown in Figure 4.4. All four factors – age, bands, f0 variations, and speakers - had significant effects on subject performance. As would be expected, sentence recognition scores improved significantly as the number of bands and/or the age of listeners increased. Young children often performed worse than older children and adults. For instance, even in the 16-band condition with natural f0 variations, 6-year-

olds had around 85% correct sentence recognition, compared to nearly perfect scores for the other three age groups (over 95%). In the present study, we were particularly interested in the effect of a natural f0 in sentences in Mandarin. The results here clearly demonstrated the importance of f0 for users of a tone language. Sentences were recognised significantly better with the presence of natural f0 variations and the effect was strong across all ages and different numbers of frequency bands (whenever performance had not reached ceiling). The effect of speaker was also highly significant. In general, subjects performed better for female speech. However, the speaker effect also significantly interacted with other factors (see those two- and three-way interactions in Table 4.3), and performance for the male and female varied with condition. Since there was only one male and one female in the present study, it was impossible to determine whether the speaker effect resulted from the difference in gender or was due to some other difference (for example, some speakers are known to produce more intelligible speech than others).

The final model shows a significant three-way interaction of the number of bands by f0 by speaker, indicating that the increase of subject performance with increasing numbers of bands depends on both the presence/absence of natural f0 and who the speaker was. One point worthy of note is that the effect of age only interacted with speaker, but not with any other factors. Therefore, the ability to use f0 information, or to take advantage of increasing degrees of spectral information (bands) was the same at all ages.

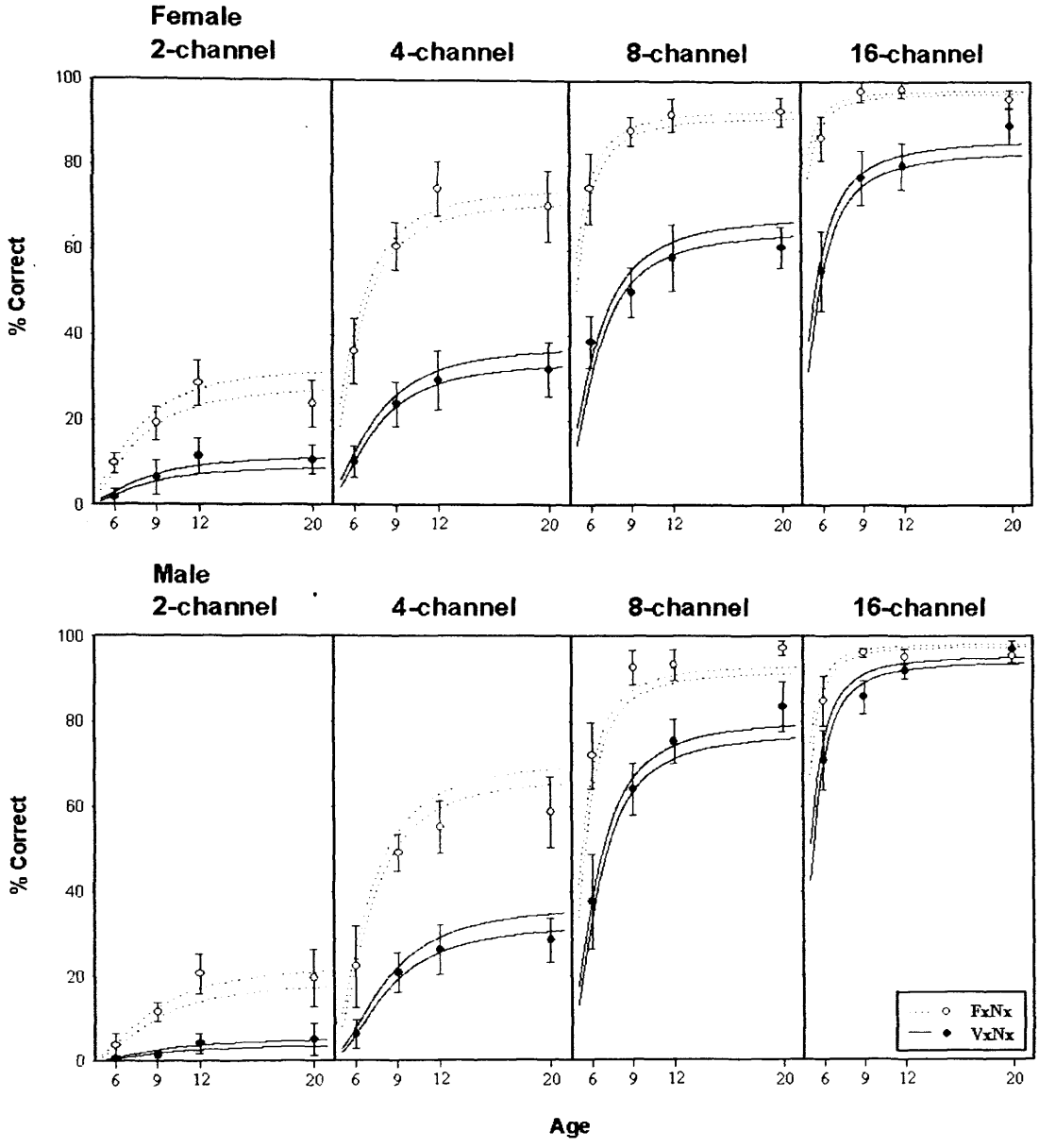

Figure 4.4 Percentage of words in sentences correctly recognised from the observed data and the model prediction, displayed as a function of the age of listeners. The upper panel presents results for the female speaker and the lower panel for the male speaker. From left to right the panels present performance for different numbers of channels. In each graph, open symbols represent subject performance for sentences with natural f0 variations (using a FxNx carrier) and filled symbols for sentences without f0 variations (using a VxNx carrier). Error bars show 95% confidence intervals for the means from ten subjects in each group. The curves represent the 95% confidence level of performance predicted from the model: dashed ones for sentences with f0 variations and solid ones for those without f0 variations.

4.3 Experiment IIA: Effects of voice f0 on sentences with controlled rms-amplitude contours

Aims and experimental predictions

While experiment II demonstrated a very strong effect for the presence of natural f0 on sentence recognition in Mandarin, it cannot be guaranteed that this effect was not contributed to partially from amplitude cues. Although pulses in the carrier source had a constant amplitude, the rms amplitude for a varying frequency pulse train would be positively related to the values of the f0 (i.e., more pulses per unit time for higher f0s, and vice versa). Therefore, the FxNx carrier contained not only different f0 information from the VxNx carrier, but also with an amplitude contour which was highly correlated with the change of f0, even though the effect of amplitude is often considered to be rather small or negligible in the presence of f0. To confirm whether this amplitude change had any effect on differences of performance between FxNx and VxNx conditions, two conditions with adjusted rms amplitude were added. Four frequency channels were used for acoustic simulations to avoid either floor or ceiling effects for subject performance, so that any possible effect resulting from the change of rms-amplitude with the f0 of pulse carrier could be examined. The results in Experiment I have shown that amplitude envelope did not contribute much to tone recognition when explicit f0 was presented. It is therefore predicted that listeners would not have much difference on their performance for sentences with their rms-amplitude being further controlled or not.

4.3.1 Signal processing

The vocoder-like speech was generated by the same technique described before. Four different source carriers were investigated: the two carriers that were used before, **FxNx** and **VxNx**, and another two with adjusted rms values, **FxNx_flatRMS** and **VxNx_flatRMS**. In the latter two conditions, to give a constant rms amplitude envelope for pulse carriers, the amplitude of individual pulse for voiced speech was adjusted according to the length of each period so as to keep a constant rms level within each cycle. The amplitude of a pulse was calculated by squaring the fixed value we intended to impose in a period, divided by the number of samples in the period, and then taking a square root. This adjustment eliminated the covariation between f0 and rms level of the pulse carrier. Figure 4.5 shows the FxNx carrier, and the carrier with adjusted rms level, FxNx_flatRMS. Since there was only a very small change in the f0 in the whole sentence for the VxNx carrier (final frequency of a sentence was 6% lower than the initial frequency), the rms level only varied over a very small range.

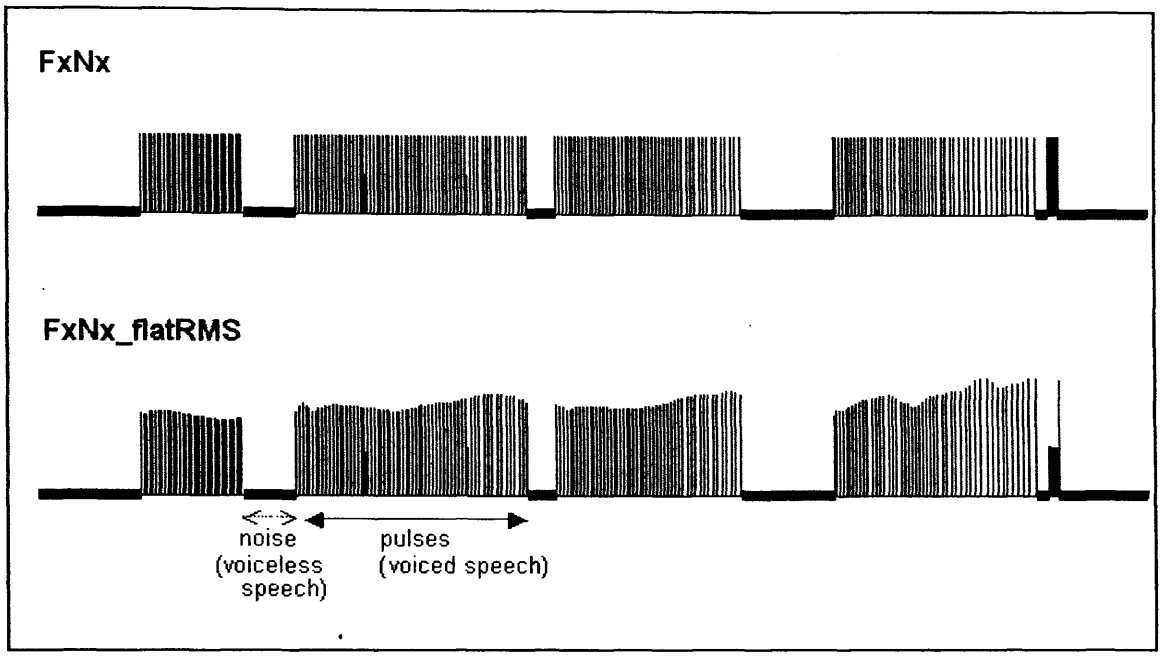

Figure 4.5 Examples of FxNx and FxNx_flatRMS carriers. The FxNx carrier was generated by pulses with constant amplitude. The FxNx_flatRMS carrier was generated by pulses with adjusted amplitude to give a constant rms value for each period. The longer the period, the larger the amplitude value, and vice versa.

4.3.2 Subjects

Eight normal hearing adults, three male and five female, participated in the experiment. All were native speakers of Mandarin from Taiwan, aged between 27 and 35.

4.3.3 Procedure

A GUI built in MATLAB was used to run the experiment, and subjects were allowed to proceed at their own speed. Sentences were presented through Sennheiser HD 414 headphones. A training session was given before the testing session started. Sixteen sentences (8 from the male speaker and 8 from the female speaker), each with their original sentence and processed in to four conditions, were given to familiarize

subjects with vocoder-like speech. Subjects could listen to sample sentences by clicking buttons on screen and were allowed to listen as many times as desired. In the testing session, each sentence was played only once, and subjects were asked to write down what they heard on provided answer sheets. No feedback was given in the testing session. There were 120 sentences in total (15 sentences x 2 speakers x 4 conditions), and they were presented in a randomised order. The sentence lists used for the four processing conditions were counterbalanced across subjects.

4.3.4 Results and discussion

Figure 4.6 shows the percentage correct sentence recognition across conditions. Subjects had similar performance on conditions FxNx_flatRMS and FxNx (74.3 and 71.7%, respectively), and similar performance on conditions VxNx_flatRMS and VxNx (31.7 and 31.0%, respectively). A logistic regression was performed on the total number of keywords each listener perceived correctly for the 15 sentences from each speaker (a total of 45 key words). Again, overdispersion in the data was accounted for. The saturated model included the factors speaker (male, female), f0 (the presence/absent of f0 variations) and rms (controlled or not) plus all the interactions. None of the interactions was significant, nor was the effect of rms, leaving an adequate model to be (Table 4.4):

parameter	estimate
1	-0.63
speaker	-0.31
f0	1.79

Table 4.4 The parameters and their coefficients for the final model

The effect of speaker and f0 were both highly significant ($p<0.0005$ and $p<0.0001$, respectively). The final model indicated that, consistent with the previous experiment, performance for sentences with natural f0 was significantly higher than those without natural f0, and female speech was generally recognised better than male speech. There was no significant effect of rms. This confirmed that the advantage of natural f0 on sentence recognition was equally strong when rms-amplitude was kept constant.

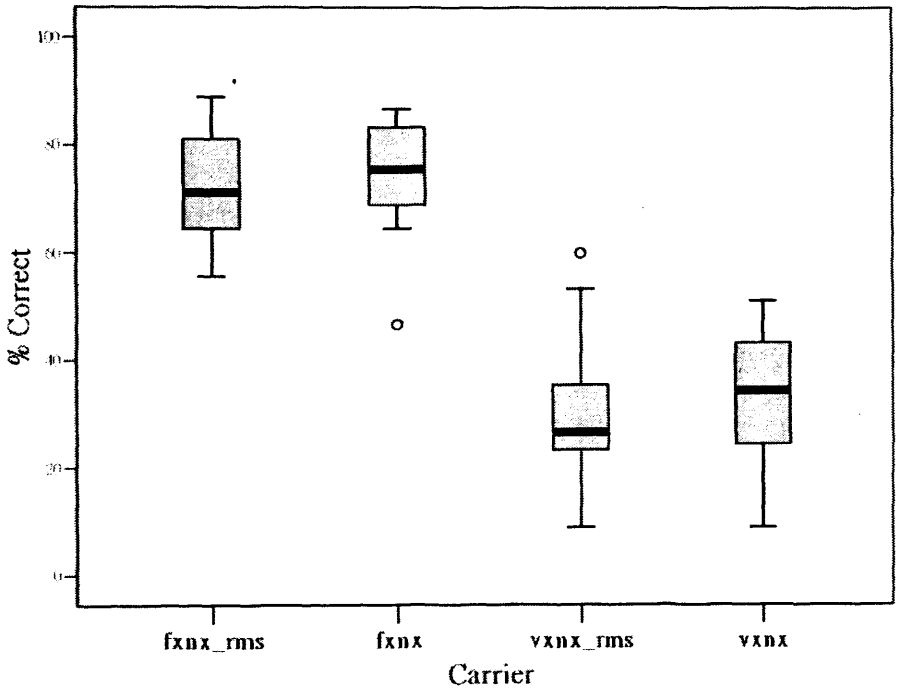

Figure 4.6 Boxplots of percentage of correct sentence recognition across conditions of different carriers. Open circles represent outliers which are more than 1.5 box lengths from the median (boxlength represents the 25- to 75-percentile range of 16 data points, 8 subjects x 2 speakers).

Note that recognition scores for FxNx sentences here were somewhat better than those for 4-channel sentences in the previous experiment (74.30 and 64.55 % for the present and previous studies, respectively), with, however, no difference for VxNx sentences (30.97 and 30.22 %, respectively). This difference for FxNx sentences may

be due to the fact that only 4-channel simulations were used in this experiment, compared to the mixed conditions with four different numbers of channels before, and subjects might have more practice to familiarise 4-channel vocoded speech. Also, different subjects were used in the two experiments.

4.4 General discussion

Contribution of natural f0 in tone languages

While previous studies in English speakers found that voice f0 has little or no effect on speech recognition (e.g. Faulkner *et al.*, 2000; Hillenbrand, 2003; Shannon *et al.*, 1995), Fu *et al.* (1998) reported that temporal cues to voice f0 showed some effect on the recognition of Mandarin. The present study adapted the method used in the study by Faulkner *et al.* (2000), conveying voice f0 explicitly by f0-controlled pulses trains, to determine the maximal effect of voice f0 on sentence recognition in a tone language. In contrasts to the results reported for English speakers, the results here demonstrate clearly a significant contribution of voice f0 to the recognition of Mandarin. In the presence of f0, subjects were able to achieve a high level of performance with a small number of frequency bands. On a rough estimate, the effect of a natural f0 contour was very close to the effect of doubling the number of frequency bands (this rule being approximately true for channel numbers from 2 to 8). For instance, the performance for the 2-channel FxNx was roughly equal to the 4-channel VxNx; and the 4-channel FxNx was roughly equal to 8-channel VxNx.

The significant contribution of f0 to sentence intelligibility in tone languages, especially when spectral content is reduced greatly, has also been reported in some

studies. Fok Chan (1984) found tonal information significantly enhanced speech understanding in Cantonese[6] in noise. When signal-to-noise ratio increased from -6 to -12dB, speech intelligibility for sentences produced with a monotone fell significantly (from 45% to 17%) whereas performance for sentences produced with natural tone remained high and at the same level (85%). Another study by Lan *et al.* (2004), with some similarity to the present study, investigated the effect of incorporating explicit f0 in CIS processing in Mandarin. They constructed a novel speech coding strategy, in which both envelopes and f0 were extracted and used to modulate the amplitude and the centre frequency of pulse carriers. The novel strategy was compared with the conventional CIS strategy. Tones, phrases, and sentences were processed into 4-, 6-, and 8-channel vocoded speech, and presented to 20 normal-hearing adults for identification. With an extra coding of explicit f0, listeners performed significantly better with the new strategy. Our results and previous studies emphasis the importance of voice pitch in speech communication for tone language users. Pitch variations are thought to be essential for isolated syllables but less important or even redundant in running speech when context is available (Fok Chan, 1984). However, when spectral information is degraded by noises or limited frequency channels, a clear indication of voice f0 makes a significant contribution to sentence recognition.

6 There are six contrastive tones in the Cantonese tonal system, which are distinguished exclusively by their f0 patterns only (e.g. Fok Chan, 1974; Vance, 1976).

Implications for cochlear implants

The results of present study have significant implications for implant users. Despite the greater number of frequency channels provided in current implant systems (for instance, up to 20 frequency channels in Nucleus), there is some evidence that implanted listeners seem not to be able to make use of more than about eight channels of spectral information (Fishman *et al.*, 1997; Fu *et al.*, 1998a; Friesen *et al.*, 2001; Moore, 2003). It remains unclear why implant users are not able to make use of all spectral channels in a cochlear implant. Although studies have shown that high levels of performance can be achieved with relatively small number of channels in good listening conditions (e.g. Shannon *et al.*, 1995; Dorman *et al.*, 1997a; Dorman & Loizou, 1998; Fishman *et al.*, 1997), more spectral channels were required for a more difficult listening condition such as in noise or the presence of a competing speaker (Fu *et al.*, 1998; Friesen *et al.*, 2001). The results of our study have demonstrated that the presence of a clear indication of f0 significantly improved speech understanding, and this enhancement was especially important with limited spectral resolution. At present, implant users of a tone language obtain very limited information about tonal contrasts through their implants (e.g. Barry *et al.*, 2002; Ciocca *et al.*, 2002; Lee *et al.*, 2002; Peng *et al.*, 2004; Wei *et al.*, 2004; Wu & Yang, 2003), which might subsequently affect their speech understanding (Fu *et al.*, 1998). If voice f0 can be conveyed explicitly in future implant devices, implant listeners, at least tone language users, are likely to benefit significantly.

Furthermore, cochlear implants have been applied to deafened children at a very young age. Given the essential role of f0 during speech development, information about voice f0 may play a more significant role in implanted children of all languages. Studies in normal-hearing children listening to acoustic simulations of cochlear

implants have shown that younger children often perform worse than older children or adults, and require more frequency channels to achieve a given level of performance (Dorman *et al.*, 2000; Eisenberg *et al.*, 2000; Eisenberg *et al.*, 2002). The underlying mechanism explaining the difference between young children and adults is not yet completely clear. In addition to different linguistic knowledge and central cognitive processing, young children were also less capable of making use of sentence context (Eisenberg *et al.*, 2000). One important finding in our study was that young children had the same ability as adults in terms of using f0 information. To achieve the same level of performance as adult subjects, young children required more information, either by a greater number of channels or better voice pitch. Therefore, providing better voice pitch information for implanted children is likely further improve their performance.

4.5 Summary

- Natural f0 enhanced sentence recognition significantly in Mandarin when spectral information was degraded. The effect of f0 was strong across all age groups of listeners from 6 years to adults.
- Young children required more information either from voice pitch or from increased spectral resolution to achieve the same level of performance as older children and adults.
- The effect of age did not interact with f0 or channel number, indicating that listeners of all ages were the same in terms of taking advantage of natural f0 or increased spectral information.
- Providing better voice f0 information in cochlear implant will improve speech perception in implant users of a tone language.

Chapter 5

Perception of lexical tone and sentence recognition in children with cochlear implants

A number of studies have examined the perception of lexical tones in Mandarin and Cantonese speakers receiving an implant. In general, implant users with current devices are able to obtain some information about tonal contrasts through their implants. However, the benefit is quite limited. This is hardly surprising due to the fact that information about voice f0 in current implant systems can only be perceived through temporal fluctuations of speech inputs. This temporal pitch information is considered to be rather weak, and only represented reliably by a speech coding strategy using a sufficiently high stimulation rate. The study reported in chapter 3 has shown that, on the basis of the periodicity information only, Mandarin tonal information can be recognised correctly to a certain extent (averaging around 50% correct; chance 25%). However, this periodicity information is unlikely to be transmitted properly for implant users with the SPEAK strategy due to its use of a relatively slow stimulation rate. Another two acoustic cues, amplitude envelope and duration, have also demonstrated some substantial contribution to tone recognition (averaging about 45% and 35% correct, respectively), and all recent implant devices are able to transmit these cues relatively well. The further analysis in Chapter 3 suggests that the contribution of amplitude envelope may arise from its similarity to the f0 contour, and this is especially true for tones 2 and 4. Here, the first study in the current chapter aims to examine the perception of lexical tones in implanted children,

and whether amplitude envelope plays some role in recognising tonal contrasts in these implant users. The effect of amplitude envelope was examined by imposing a relatively flat amplitude envelope to speech stimuli so as to see if there was any effect on subject performance due to the lack of the correlation between pitch and amplitude contours as in natural speech. Implant users with three different speech coding strategies (SPEAK, ACE, and CIS) were recruited in this study in order to determine if the reliance on amplitude envelope varied with speech coding strategy. A more intensive investigation on the effect of both amplitude envelope and duration cues to tonal contrasts in implanted listeners was carried out in two children with Nucleus ACE devices. Either or both of the two cues was neutralised so as to examine if there was any effect on tone recognition performance. The condition with both amplitude envelope and duration cues removed also allowed determination of the level of performance which could be achieved by temporal pitch information alone.

Given the significant role of voice f0 in a tone language, the third study in this chapter investigated the role of voice f0 information in running speech in implanted children. Results of acoustic simulations in chapter 4 have clearly demonstrated the significant contribution of natural f0 variations in understanding running speech, especially when spectral information was degraded greatly. For instance, the performance for 4-channel vocoded speech with natural f0 variations was comparable to that achieved by 8-channel vocoded speech with a relatively constant f0 contour. Here, voice f0 in sentences was manipulated so as to examine its effects on sentence recognition in implanted listeners.

Before reporting these experiments, previous studies related to tone perception in implant users who speak a tone language will be introduced first.

5.1 Introduction

Tone perception in implant users of a tone language

Table 5.1 summaries results from previous studies investigating the perception of lexical tones in speakers of Mandarin or Cantonese using implants.

Authors (Year)	Subjects	Device/ Strategy	Language	Task	Results
Single-channel implant					
Tang et al. (1990)	4 adults	House/3M	Cantonese	6-alternative (**17%** chance)	● Group mean: **30%** (10% before implantation) ● Individual performance: 27 - 47%
Kwok et al. (1991)	8 adults	House/3M	Cantonese	6-alternative (**17%** chance)	● Group mean: **39%** (13 % with hearing aids) ● Individual performance: chance - 67%
Multi-channel implant					
Feature-extraction strategy					
Xu et al. (1987)	1 adult	Nucleus 22, f0 as stimulation rate	Mandarin	4-alternative (**25%** chance)	● **100%** correct
Huang et al. (1996)	4 adults	Nucleus MPEAK	Mandarin	4-alternative (**25%** chance)	● Group mean: **68%** (35% before implantation) ● Individual performance: chance - over 90%
Liu et al. (1997)	5 adults	comparison between MPEAK vs. SPEAK	Mandarin	4-alternative (**25%** chance)	● Group means: **65** and **66**% (MPEAK/ SPEAK) ● Individual performance: 55/60, 62/69*,60/84, 70/60*, 76/56 (* sig.)
Sun et al. (1998)	6 adults	5 MPEAK & 1 SPEAK	Mandarin	4-alternative (**25%** chance)	● Group mean: about **57%** ● Individual performance: chance - 90 % ● 4 MPEAK users performed above chance
Filter-bank strategy					
Tong et al. (2000)	11 adults 12 children, aged 3-14	Nucleus SPEAK	Cantonese	6-alternative (**17%** chance)	● Group mean: **40%**
Wei et al. (2000)	28 children	26 Nucleus22/24 2 Clarion	Cantonese	6-alternative (**17%** chance)	● Group mean: **66**, **70**, and **65%** after 6-, 12-, and 24-month implantation (36% before implantation)
Ciocca et al. (2000)	17 children, aged 4-9	6 SPEAK & 11 ACE	Cantonese	tonal pair (**50%** chance)	● Group mean: **50 - 61%** for each of tonal pairs, only 3 out of 8 pairs above chance and all involving the high-level tone. ● Only 2 out of 17 children performed above chance

Lee et al. (2002)	15 children, aged 6-11	Nucleus SPEAK	Cantonese	tonal pair (**50%** chance)	●	Group results: **53**, **68**, and **69%** for 3 tonal pairs, the latter 2 pairs above chance
Barry et al. (2002)	16 children, aged 4-11	Nucleus 7 SPEAK & 9 ACE	Cantonese	tonal pair (**50%** chance) change/no change paradigm	●	Group results: 11 out of 15 (SPEAK) and 10 out of 15 (ACE) tonal pairs are above chance (contrasts involving the high-level tones were discriminated better
Wu & Yang (2003)	16 children, aged 4-9	Nucleus ACE	Mandarin	4-alternative (**25%** chance)	●	Group mean: **73** and **79%** after 12- and 24-months implantation
Peng et al. (2004)	30 children, aged 6-12	11 Med-EL CIS 19 Nucleus SPEAK	Mandarin	tonal pair (**50%** chance)	● ●	Group mean: **73%** All 6 pairs were significant above chance (contrasts involving the high-falling tone were identified better)

Table 5.1 Summary of previous results on tone recognition by implant users of a tone language.

House/3M single-channel implant devices

In the House/3M single-channel device, speech is bandpassed through a 340-2700Hz filter and then used to modulate a 16 kHz carrier. The output signals preserve the periodicity information of the original speech. Therefore, some information about voice pitch can be elicited from the modulated signals (Dorman, 1993; Loizou, 1998). Studies investigating perception of prosody in English speakers reported that users of the House/3M device had a rather poor performance on a question/statement task, averaging about 60% correct (50% chance), though a small number of users achieved up to 75-80% correct (Edgerton *et al.*, 1983; Tyler *et al.*, 1985). Rosen *et al.* (1989) pointed out that the signals presented by the House/3M device might be too complicated to allow voice f0 to be perceived easily. They found performance on the same task could be improved by simplifying speech waveforms to better signal f0 information.

Two studies investigating tone perception in Cantonese speakers with the House/3M single-channel implant reported some improvement in tone perception after implantation (Tang *et al.*, 1990; Kwok *et al.*, 1991). Average performance was about 30-40% correct, compared to chance before implantation (17% chance). Some users did not perform significantly differently from chance. One subject in Kwok *et al.* (1991) achieved the best performance of 67% correct. These results demonstrated that users of the House/3M single-channel implant device were able to obtain some information about tonal contrasts, but the benefits were limited.

Multiple-channel implant devices with feature extraction strategies

Feature extraction strategies, including F0/F2, F0/F1/F2, and MPEAK (Multipeak), were used in the early versions of the Nucleus devices. In the feature extraction strategy, voice f0 is used to determine the stimulation rate during voiced speech (see more details in the Chapter 2, section 2.2 'speech coding strategy'). The explicit encoding of f0 is expected to give a clearer indication of voice pitch than that carried by temporal fluctuations of speech envelope in single-channel implants, or the latest multiple-channel implants with a CIS-like strategy.

A number of investigations of tonal contrasts were conducted in Mandarin speakers, and an average performance of around 60-70% correct (25% chance) was reported (Huang *et al.*, 1996, Liu *et al.*, 1997; Sun *et al.*, 1998). Some users achieved nearly perfect performance (over 90% correct), whereas some performed at just about chance level. A study by Xu *et al.* (1987), conducted in one subject only who was a native speaker of Mandarin and also familiar with English, reported 100% correct in recognising tonal contrasts in Mandarin and statement/question contrasts in English. The high level of performance by some implant users with the feature-extraction strategy indicates that information about f0 can be represented relatively successfully through variation in stimulation rate. However, the fact that not all of the users with such a strategy performed equally well suggests that there may be individual difference in the ability to use rate pitch information.

Multiple-channel implant devices based on filter-bank processing technology

In recent filter-bank approach, such as CIS (Continuous Interleaved Sampling), SPEAK (Spectral Peak), and ACE (Advanced Combination Encoders), speech signals

are analysed into a number of frequency channels, and the amplitude envelope of each channel is extracted and used to modulate a pulse carrier. Some information about voice pitch can be conveyed in such a strategy if the frequency of envelope-extraction filter includes the fundamental frequency of voice pitch, and the pulse carrier has a sufficiently high stimulation rate. Studies have shown it is necessary for the frequency of the pulse carrier to be 4 or 5 times higher than the f0 frequency range to sample it properly (e.g. Busby et al., 1993; Wilson, 1997). The CIS and ACE strategies both use a pulse carrier with a high stimulation rate, whereas the SPEAK strategy uses a pulse carrier with a relatively low one, 300 pps or even lower. Therefore, voice f0 is unlikely to be represented reliably in the SPEAK strategy. Even though the representation of f0 would be expected to be better in CIS and ACE, the information about voice f0 carried by the fluctuations of temporal envelopes is considered to be rather weak.

Green, Faulkner, and Rosen (2004) reported that, in Clarion users with the CIS strategy, performance on labelling the pitch movement of synthetic vowel glides was well above chance, but still limited. Even for stimuli with an f0 change up to one octave, these implant users performed only around 80% correct (chance 50%). A significantly better performance was achieved with a modified CIS processing, which had a simplified waveform matching the periodicity information of the input signals. However, the improvement was rather small. Further investigation of pitch perception using natural speech stimuli in Green, Faulkner, Rosen, and Macherey (2005) reported a mean score of about 70% correct in question/statement identification by users of the Clarion CIS devices. Again, a significantly better performance, around 5% increase, was achieved with the modified CIS processing.

A number of studies have investigated the perception of lexical tones in prelingually deafened children using an implant with filter-bank processing technology. Results from these children were even more inconclusive about whether these early deafened children were able to obtain sufficient information about tonal contrasts through their implant devices. Ciocca *et al.* (2002) and Lee *et al.* (2002) both reported that Cantonese children had great difficulty in recognizing tonal contrasts. Ciocca *et al.* (2002) investigated the perception of Cantonese tone in 17 early deafened children, aged between 4 and 9, with either SPEAK or ACE strategies. The six Cantonese tones were grouped into eight tonal pairs to examine listeners' sensitivity to pitch level (contrasts HL/ML, HL/LL, and ML/LL), the difference of final frequency (HR/LR, LR/LL, LF/LR, and LF/LL), and the difference of the initial frequency (HL/HR)[7]. Only two out of the 17 children performed above chance, indicating that the majority of these implanted children were not able to identify tonal contrasts reliably. Group results showed that only 3 out of 8 tonal pairs were recognised above chance, and none was over 61% correct (chance 50%). All three tonal pairs involved high-level tone (contrasts HL/ML, HL/LL, and HL/HR), the tone with a more distinct frequency range from others, suggesting that tonal contrasts might be recognised correctly only when two tones had a relatively large difference in f0. Lee *et al.* (2002) examined three tonal pairs (HL/HR, HL/LF, and HR/LF) in 15 implanted children, aged between 6 and 11, with SPEAK, and also reported a relatively poor performance. The three tones used in this study, HL, HR, and LF, were three early acquired tones in Cantonese (So & Dodd, 1995; Tse, 1978), and therefore

[7] There are six contrastive tones in Cantonese: high level (HL), high rising (HR), mid level (ML), low falling (LF), low rising (LR), and low level (LL). **Contrasts HL/ML, HL/LL, and ML/LL** differ in their pitch height; **contrasts HR/LR, LR/LL, LF/LR, and LF/LL** start at approximately the same frequency, but end at different frequencies. Except HR/LR with the same rising contour, the other 3 pairs have different pitch contours. **Contrast HL/HR** has different initial frequencies but a similar final frequency.

considered to be easier. All three tonal pairs were recognised correctly over 90% by 3-year-old normal-hearing children. For implanted children, two tonal pairs (HL/LF and HR/LF) were recognised around 70% correct (chance 50%). The tonal pair HL/HR, which involved the high-level tone, was not recognised above chance.

Studies by Barry *et al.* (2002) and Wei *et al.* (2000), also carried out in implanted children of Cantonese, have shown more positive results in perceiving tonal contrasts. Barry, Blamey, Martin, Lee, Tang, Yuen, and van Hasselt (2002) conducted a novel testing procedure to examine the ability to detect the change of the different tones in 16 implanted children using the Nucleus device with SPEAK or ACE strategy. The discrimination task was adapted from the change/no change paradigm. The syllable /wi/ with a particular tone was played at a fixed rate continuously as a "background" sound. The same syllable with a different tone then replaced the background one at different times and was played 3 times, before the background sound was played again. Children were asked to give a response when they heard the change in tone. The advantage of this task was that it did not require any cognitive skills or linguistic knowledge as the identification task used in most studies did. Therefore, it could be applied to children at a very young age or with little or no hearing experience. The syllable /wi/ with any tone does not correspond to any word in Cantonese. This allowed listeners to focus only on the change of f0 without mapping the sound to lexical words. The six Cantonese tones were arranged into 15 tonal pairs, all the combinations of the six tones, to examine the relative difficulty of discriminating the different tonal pair. Two-thirds of the tonal pairs were discriminated well above chance by implanted children. As in Ciocca (2002), those contrasts combining the high level tone were discriminated relatively well. No significant difference was found in performance between ACE and SPEAK users.

Their results indicated that implanted children were able to derive some information about tonal contrasts, though their performance was not as reliable as normal-hearing children. Wei, Wong, Hui, Au, Wong, Ho, Tsang, Kung, and Chung (2000) also reported that implanted children were able to perform relatively well in a tone identification task in Cantonese, achieving a mean score of about 65 % correct (17% chance).

In general, studies carried out in Mandarin-speaking children have often reported more positive results than those in Cantonese-speaking children. For instance, Peng, Tomblin, Cheung, Lin, and Wang (2004) examined tone perception in 30 implanted children speaking Mandarin, 19 with Nucleus SPEAK and 11 with Med-El CIS. All six tonal pairs, invoving all possible combinations of the four Mandarin tones[8], were recognised above chance, with overall average performance at 73% (chance 50%). All tonal pairs including the falling tone (tone 4) were recognised significantly better than other pairs. Six out of 30 children achieved around 90% correct, and all of them were SPEAK users. Wu & Yang (2003) also reported a relatively high performance in tone identification, a mean score of around 75% correct (chance 25%), by 16 children with the Nucleus devices using the ACE strategy.

The relatively higher performance in Mandarin-speaking children is likely due to the fact that Mandarin tones can be recognised to certain extent by other acoustic cues such as amplitude envelope and duration which can be transmitted relatively well in implants. Also, the overall intensity and creaky voice may also be potential cues in Mandarin. In contrast, Cantonese tones are almost exclusively distinguished by their

[8] The four Mandarin tones are level, rising, falling-rising, and falling tones, often referred as tones 1 to 4, respectively.

pitch patterns, even though there are also a few other acoustic cues associated with certain tones. For instance, the high-level tone is produced more intense by some speakers (Fok Chan, 1974), and the low-falling tone is often accompanied by creaky voice (Vance, 1977). However, none of these cues are consistently present to aid in the recognition of tonal contrasts by Cantonese listeners. Furthermore, there are more tones in Cantonese (six contrastive tones in Cantonese, and four in Mandarin). With more tones crowded in the low frequency range (Vance, 1977), it might be more difficult to distinguish Cantonese tones. Gandour (1983) has reported that pitch level was the most important perceptual dimension of tones, followed by pitch contour. Pitch contour played a more important role for Cantonese speakers than for Mandarin or Taiwanese speakers. However, results from Barry *et al.* (2002) suggested that implant devices with either SPEAK or ACE might not provide sufficient information for implanted listeners to perceive the change of pitch contour. For instance, implanted children performed relatively poorly on the contrast of tones 4 and 5, low-falling and low-rising tones, which normal-hearing children did not find it especially difficult. The low-rising and low-falling tones are very different perceptibly, and almost never confused by normal-hearing adults (Varley and So, 1995). The poor representation for the movement of f0 direction in implant devices might have a more negative effect in Cantonese than in Mandarin tones.

Possibly better representation for tonal contrats through implants

An explicit coding of f0

The information about f0 presented through an implant is determined by the speech coding strategy used for transforming speech signals into electrical stimuli.

Since f0 is the primary cue for tonal contrasts, presumably, those speech processing strategies with better encoding for voice f0 could be more beneficial than other strategies in terms of perceiving tonal information. For instance, voice f0 is encoded explicitly in the MPEAK strategy by the stimulation rate, whereas only rather weak information about voice f0 is provided by in SPEAK strategy through the fluctuations of temporal envelopes. Theoretically, implant users with the former strategy might be expected to perform better than those with the latter strategy. However, studies in implant users often did not show the benefit of explicit coding of voice f0.

Liu *et al.* (1997) examined the perception of Mandarin tones in 5 subjects who were experienced implant users with MPEAK and just about to change to SPEAK. Group performance showed no significant difference between the two strategies (mean scores were 64.7 and 65.9 for MPEAK and SPEAK, respectively). One subject had a significantly higher score with MPEAK, while another one performed significantly better with SPEAK. Another study by Jones *et al.* (1994) investigated whether adding an explicit coding of f0 information to the SMSP strategy (identical to the SPEAK in many aspects) could aid the identification of suprasegmental information in English. The modified SMSP strategy stimulated the most apical electrode with the rate of f0, and remaining electrodes at a constant rate. Five experienced SMSP users were asked to identify roving stress, rising/falling intonation, and question/statement contrasts using SMSP and the modified strategy. No significant improvement was found with the additional rate-encoded representation of f0. One possible explanation for the lack of effect with an explicit coding for f0 information might due to individual differences in the ability of using the rate pitch information. It is also possible that, as pointed out by Loizou (1998), the extraction of

voice f0 in early feature-extraction strategies might not always succeed in bringing out maximum effect of a clear indication of voice f0.

Two studies used acoustic simulation to investigate the effect of explicit f0 in tone language users, both reporting significant benefits for the encoding of explicit f0. Given that f0 changes contribute relatively little to speech intelligibility in English (Hillenbrand, 2003), Fearn (2001) applied a similar speech coding scheme as Jones *et al.*'s (1994) in normal-hearing listeners using tone languages. Native speakers of Cantonese and Thai, two in each, were asked to identify tonal contrasts of their languages. The perception of lexical tones improved significantly with the extra f0 coding through the stimulation rate. Lan *et al.* (2004) incorporated explicit f0 information into the CIS strategy, and applied it to native speakers of Mandarin using acoustic simulations. They also reported a significant improvement in recognition of lexical tone and running speech with the explicit coding for f0.

Higher stimulation rate

A higher stimulation rate is expected to give a better representation of f0, which may lead to better perception for lexical tones. For instance, strategies with higher stimulation rate such as ACE and CIS are expected to be more beneficial than SPEAK which has a considerably lower stimulation rate, in terms of transmitting temporal voice pitch information. In the study by Barry *et al.* (2002), which compared 16 children using either SPEAK or ACE speech processing strategy with the Nucleus devices for their ability to discriminate Cantonese tones, no significant difference was found between SPEAK and ACE users. Peng *et al.* (2004) also reported no significant difference in the identification of Mandarin tones between 19 Nucleus SPEAK and 11

Med-el CIS users. However, the lack of the effect for stimulation rate might be partially due to the use of between-subject comparisons, and, therefore, might be confounded by other factors such as subject variables associated with users of certain strategies. For instance, ACE users were generally younger in age and were implanted earlier.

Au (2003) examined the effect of stimulation rate on Cantonese tone recognition in the same group of subjects. Eleven implanted adults with the Med-el CIS strategy were asked to discriminate and identify Cantonese tones with stimulation at high, moderate, and low rates (1800, 800, and 400 pps, respectively). Group results showed that the higher the stimulation rate, the better the performance in both discrimination and identification tasks. However, there were considerable individual differences. While most subjects showed a higher performance with a higher stimulation rate, two subjects performed best with a low stimulation rate (400 pps).

Fu *et al.* (2004) also examined the effect of different speech coding strategies with various stimulation rates in nine Nucleus-24 users. Recognition of Mandarin tones as measured with SPEAK (250pps), ACE with three different rates (900, 1200, and 1800pps), and CIS with four different rates (1200, 1800, 2400, and 3600pps; the number of channels also varied from 12, 8, 6, to 4, respectively, due to the technical capacities of CIS). The results showed tone recognition performance with ACE and CIS was significantly higher than with SPEAK. There was no significant difference among various stimulation rates with ACE, nor CIS. Thus, relatively high stimulation rates, such as over 900pps in this study, appeared to be more beneficial than lower ones with SPEAK for tone recognition.

More frequency channels

Since more frequency channels provide greater frequency resolution in implant devices, it thus might help in recognising tonal contrasts. Results from both simulation studies and clinical studies have not yet confirmed if this is indeed the case. A simulation study by Fu *et al.* (1998) examined Mandarin tone perception with frequency channels varying from 1 to 4, reporting no significant difference on tone recognition scores for different numbers of channels. In contrast, Xu *et al.* (2002) and Kong & Zeng (2004), also using acoustic simulations, reported Mandarin tone recognition to improve significantly with better resolution in frequency.

Hsu *et al.* (2000) examined the effect of the number of channels on tone perception in 3 implant users with the SPEAK strategy (1 subject was compared with 7, 10, and 20 channels, and 2 were compared with 5, 7, and 14 channels). The number of channels showed no significant effect on the recognition of tone, compared to the strong effect on recognition of words and sentences. Liu *et al.* (2004) also examined the effect of channel number on tone recognition performance in 6 children with Nucleus-24 devices. They found a small but significant decrease in tone performance when the number of frequency channels was reduced by half (from 22 to 11 channels by eliminating even-numbered electrodes, average scores from 90 to 84% correct). Performance was further reduced, but still considered to be relatively high (58%), when only six apical electrodes were activated, though this reduction could be attributable to not enough time for subjects to adjust to their new maps (30min only).

5.2 Experiment III: Recognition of tonal contrasts by implanted children - Effect of amplitude envelope

Aims and experimental predictions

This experiment intended to investigate if amplitude envelope could be used by implanted listeners to assist the recognition of tonal contrasts. As discussed in the beginning of this chapter, amplitude contour could be a usable cue for tonal contrasts due to its similarity to the change of pitch direction in some Mandarin tones. To examine the effect of amplitude envelope on tonal identification, a set of speech syllables with natural amplitude changes or with neutralised amplitude contours (relatively flat ones) were presented to a group of implanted children to identify.

It is hypothesised that since implanted listeners could only perceive relatively weak pitch information from unresolved harmonics, they might make use of other available cues such as amplitude envelope. The results from Experiment I (Fig 3.4 in Chapter 3) clearly showed that, in the conditions without salient pitch information, the presence of the amplitude envelope cue did increase tone recognition performance significantly (for instance, better performance for stimuli NoiseAFD than NoiseFD, NoiseAF than NoiseF, and NoiseAD than NoiseD). The present study manipulated the presence of natural amplitude variations, and it is expected that the lack of natural amplitude variations would result in some degree of decrease on tone identification performance in these implanted listeners.

5.2.1 Stimuli and signal processing

Two types of syllable were used as stimuli: natural speech with its original pitch and amplitude contours, and processed speech with its original pitch contours but a relatively constant amplitude envelope. Speech was processed in MATLAB and the procedure used to remove amplitude variations in voiced speech is illustrated in Figure 5.1. The voiced segment of a syllable was identified by laryngograph signals that were recorded with the speech simultaneously. For each syllable, a simplified envelope was generated with the same duration but constant amplitude. The initial and final 20 ms of the simplified envelope had sinusoidal rises and falls to give a smooth change of amplitude at the beginning and the end of voiced speech. The original amplitude envelope was extracted by full-wave rectification and forward and backward filtering with a 20 Hz fourth-order elliptical lowpass filter. Then, the simplified envelope was divided by the original envelope, and this envelope was used to multiply the voiced speech segment. The modulated voiced speech was then scaled to the same rms level as the original voiced one, and placed back into the syllable. The new envelope extracted from the processed speech showed a relatively constant amplitude contour.

Pitch information available in different stimuli

The manipulation described above was to eliminate the potential amplitude cue, while the original pitch contour and duration remained unaltered. For processed speech syllables, the temporal pitch information, derived from unresolved harmonics, and duration were still preserved and could be used by implanted listeners for recognising tonal information. The acoustic cues to tonal contrasts in the processed

stimuli were similar to those available in stimuli NoiseFD in Chapter 3 (temporal f0 and duration), and the cues in the original speech stimuli were similar to those in stimuli NoiseAFD (amplitude envelope, temporal f0, and duration). The comparison of the tone recognition performance on processed and unprocessed speech was designed to determine if the natural variation in amplitude contours played some role in tonal recognition in implanted listeners.

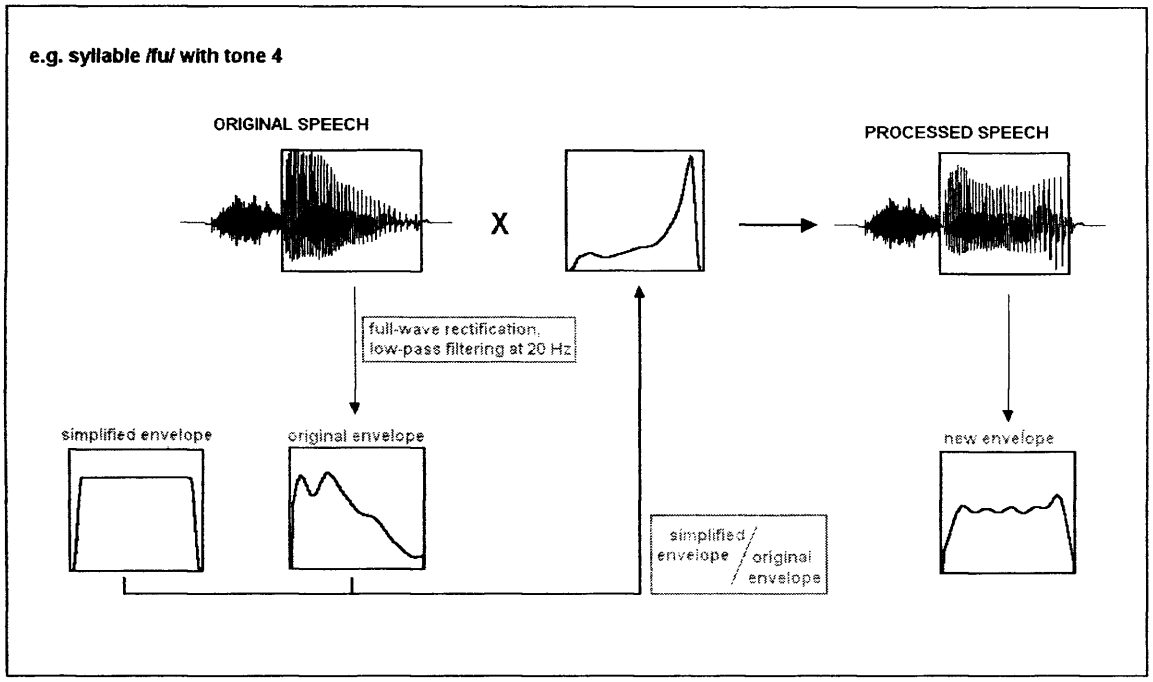

Figure 5.1 Schematic diagram for the procedure used to remove the amplitude variations in voiced speech.

There were 24 natural speech tokens, four syllables (/i/, /ba/, /fu/ and /tɕʰi/) with tones 1, 2, and 4 produced by one male and one female. Tone 3 was excluded as its realisation varies with speaker and context. The 24 natural and 24 processed speech were played 4 times each, resulting in a total of 192 stimuli. All stimuli were scaled to

give equal mean rms level for each tone and each speaker so that the overall intensity could be controlled but the natural variation in overall level still preserved.

5.2.2 Subjects

All children were native speakers of Mandarin recruited in Taiwan. Twenty-one prelingually deafened children, aged between 6 and 15, participated in the study. All of them were experienced implant users (at least 1-year of use), and were recommended as 'good' performers by therapists or schoolteachers. Sixteen were users of the Nucleus device, eight with the SPEAK and eight with the ACE strategy, and five were users of the Med-El devices with the CIS strategy. Details of the subjects are shown in Table 5.2.

Child	Implant System	Speech coding strategy	Age	Gender	Age at implantation	Duration of implantation
3	Nucleus	SPEAK	12y10m	F	4y8m	8y2m
4	Nucleus	SPEAK	12y6m	M	6y3m	6y3m
5	Nucleus	SPEAK	15y5m	F	8y2m	7y3m
8	Nucleus	SPEAK	15y9m	F	6y8m	9y1m
9	Nucleus	SPEAK	12y	F	5y4m	6y8m
10	Nucleus	SPEAK	11y10m	M	4y7m	7y3m
16	Nucleus	SPEAK	9y10m	F	3y5m	6y5m
18	Nucleus	SPEAK	9y9m	F	2y6m	7y3m
1	Nucleus	ACE	9y4m	M	4y	5y8m
13	Nucleus	ACE	9y7m	M	5y6m	4y1m
14	Nucleus	ACE	7y5m	F	3y6m	3y11m
15	Nucleus	ACE	6y7m	F	2y	4y7m
17	Nucleus	ACE	7y11m	F	5y6m	2y5m
19	Nucleus	ACE	6y7m	M	5y	1y7m
20	Nucleus	ACE	6y7m	M	2y11m	3y8m
21	Nucleus	ACE	6y2m	F	1y9m	4y5m

2	MED-EL	CIS	8y9m	F	3y6m	5y3m
6	MED-EL	CIS	12y11m	M	9y9m	3y2m
7	MED-EL	CIS	11y10m	F		
11	MED-EL	CIS	12y5m	F	6y1m	6y4m
12	MED-EL	CIS	8y10m	M	3y5m	5y5m

Table 5.2 General information for implanted children

5.2.3 Procedure

All children were tested individually. 11 children were tested in a therapy room at the Chi-Mei Hospital in Taiwan, and 10 children were tested in a quiet room in their home. The testing environments were generally quiet, with curtains and carpets which would absorb some reverberation. A tester was seated beside the child throughout testing to instruct and to help if necessary. Sounds were presented through a loudspeaker which was placed in front of the child at a distance of about 1 metre and at the height of around 1 metre above the ground. A graphical user interface (GUI) built in MATLAB was used to run this experiment. When a test session started, three Mandarin characters with the Chinese phonetic alphabet were shown on the screen first, and push buttons labelled with tones 1, 2, and 4 were below the corresponding characters. Then, a stimulus was played, and children were asked to make their identification response using a computer mouse to click on the screen. Three young children, unfamiliar with a computer mouse or preferring not to use it, made their responses by pointing to their answer on the screen, and the examiner clicked the mouse for them. Stimuli were presented in a randomised order. The intensity of stimuli varied randomly within a 3 dB range around the original level to eliminate possible cues derived from overall intensity of the stimuli. Before the testing started, children were presented all the Chinese characters, as well as the

Chinese phonetic alphabet symbols, used in the study to check if they had difficulty in recognising any words used for making their response. All children had learnt these words before. To familiarise children with the testing procedure, a training session using live voice was given first, followed by a short practice session which was exactly the same as the formal testing session. No stimuli used for familiarisation were used in the formal testing, and no feedback was ever given.

5.2.4 Results and discussion

1. The effect of amplitude envelope

Overall accuracy

Individual data are shown in Figure 5.2, which compares percentage correct tone recognition of speech with and without natural amplitude variation. As a group, average recognition scores were 70 and 65% for natural and processed speech, both scores well above chance (a binomial distribution reveals that scores of 40 or more out of 96, or 42%, are statistically different from 33% chance level). Most of these children obtained some information for tonal contrasts through their implants, even in the absence of the amplitude envelope cue. Around one-third of children achieved more than 80% correct for natural speech. In general, most children had a higher score for natural speech. This indicated that amplitude envelope was used by these implanted children in the recognition of tonal contrasts. Six of eight children with the ACE strategy reached a high level of performance compared to most implanted children, and their performance was less affected by the absence of amplitude envelope. The amplitude envelope seemed to have a greater effect on some children with the SPEAK strategy. Four of the eight children showed a decrease for processed

speech compared to natural speech (data points to the left of the diagonal). Overall, the effect of amplitude was not very large (65 vs. 70%, overall).

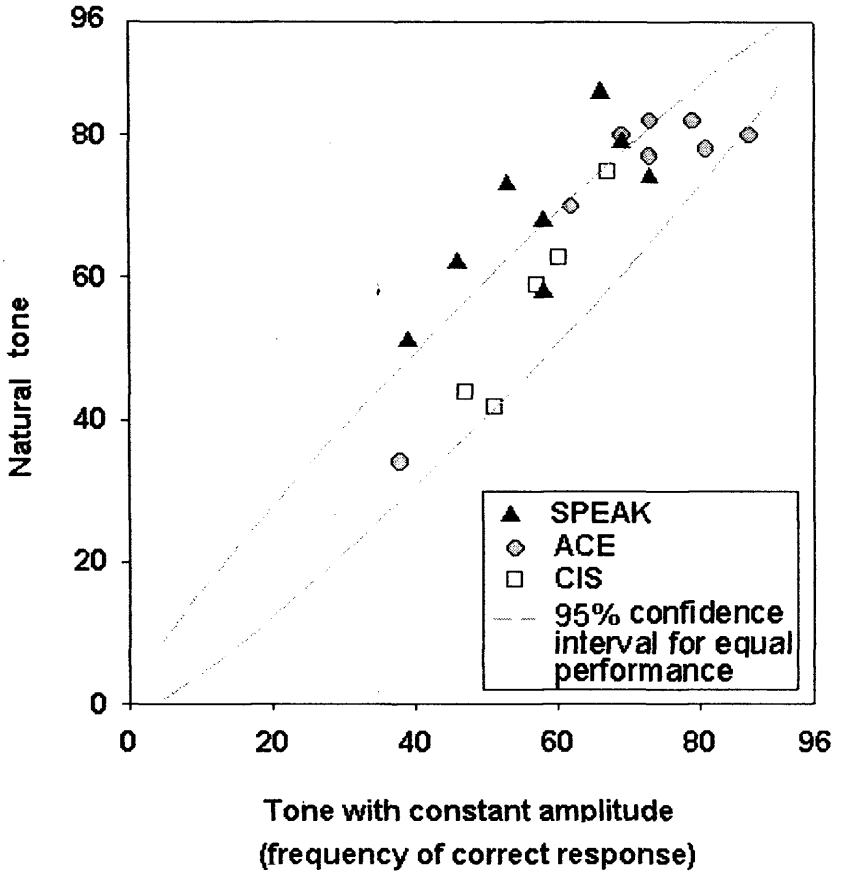

Figure 5.2 Scatterplot for individual performance on tone recognition for natural speech (y-axis) and processed speech (x-axis). The dashed curves represent 95% confidence level, using the binomial model, for equal performance for natural and processed speech. Data outside the curves represent a significant difference between speech with and without original amplitude envelope. Children with different speech coding strategies are marked by different symbols.

Performance for individual tones

Figure 5.3 shows the results for the three tones separately. For tones 1 and 2, most children did not show a significantly different performance between natural and processed speech. For those children whose data were outside the 95% confidence interval curves, amplitude variation appeared to have an inconsistent effect on tone

recognition. While some performed better for speech with the original amplitude envelope, others performed better for speech with a simplified amplitude envelope. For tone 4, the majority of children, regardless of the speech coding strategy, showed a significantly lower performance for speech without a natural amplitude envelope. Logistic regression was performed on the group results for the effect of amplitude envelope, and a significant difference was found for tone 4 ($p < 0.05$), but not for the other two tones. These results demonstrated that amplitude variation contributed significantly to tone 4, but not to tones 1 and 2.

The children with a significantly different performance on natural and processed speech for tone 1 (children 2, 3, 4, 13, 16, and 19; 3 with SPEAK, 2 with ACE, and 1 with CIS) mostly were *not* those children with significant differences for tone 2 (children 3, 9, 17, and 18; 3 with SPEAK and 1 with ACE), except child 3. Note that they were often SPEAK users.

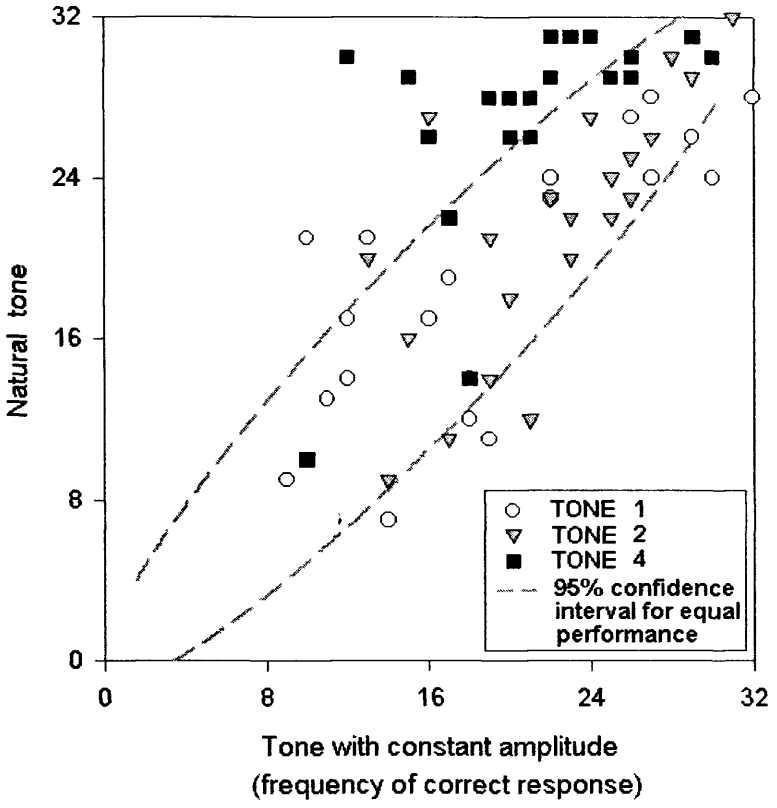

Figure 5.3 Scatterplot for frequency of correct response on natural and processed speech. Results for different tonal contrasts are marked by different symbols. The dashed curves represent 95% confidence level for equal performance for natural and processed speech. Data outside the curves represent a significant difference between speech with and without the original amplitude envelope.

2. Other acoustic cues to tonal contrasts: possible effect of duration

The above results showed some evidence for the utility of amplitude variation in tone recognition by implanted children. However, the effect did not seem to be as large as expected. These implanted children were able to recognise tonal contrasts correctly to some extent for speech without its original amplitude envelope. Another possible cue that might be used by implanted children was duration. Boxplots for the duration of voiced speech (simply called duration throughout this chapter) for tones 1, 2, and 4 are shown in Figure 5.4.

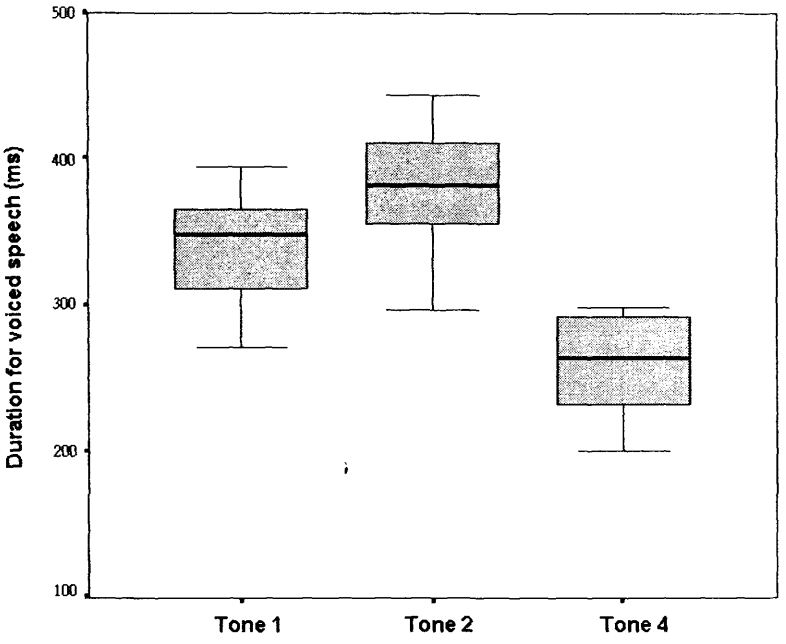

Figure 5.4 Boxplot of duration across different tones.

To investigate whether implanted children made their response based on the duration cue, the relationship between subject response and duration was examined. Figure 5.5 shows the frequency of subject response for the three tones as a function of duration, with a regression line for each response. The regression line for the tone 4 response is clearly distinguished from those for tones 1 and 2. The shorter a syllable, the more likely it is to be labelled as tone 4, and the longer a syllable, the more likely it is to be labelled tone 1, and especially tone 2. In labelling tone 4, there seemed to be a criterion duration of about 300ms. A syllable with duration above 300ms was rarely called tone 4. This suggested that perhaps implanted children used duration to identify tone, especially tone 4.

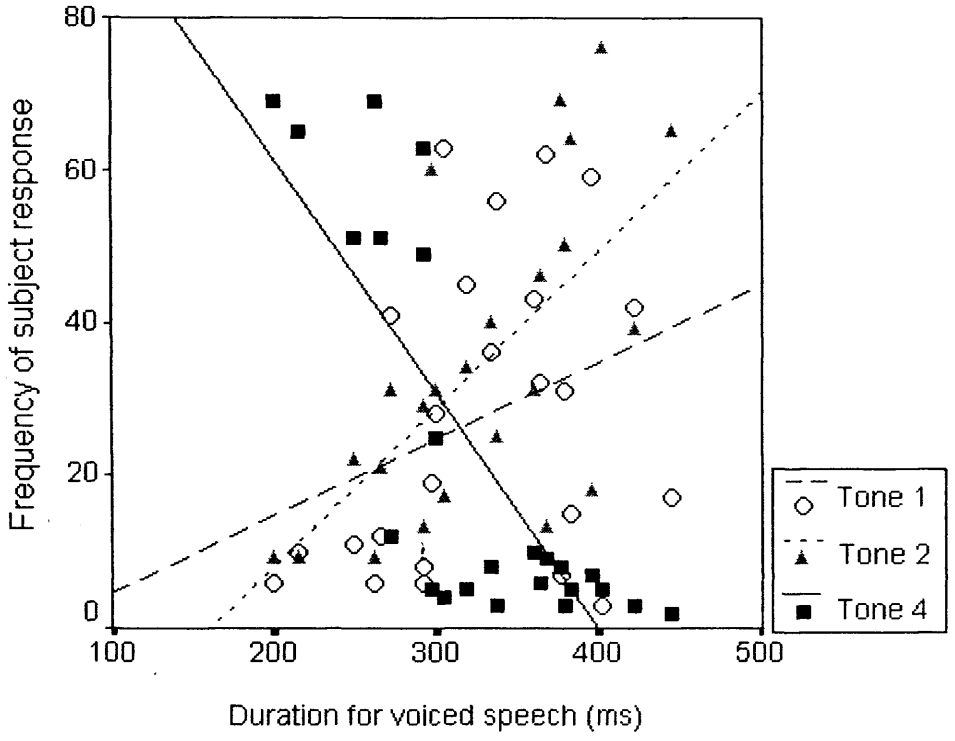

Figure 5.5 Frequency of subject response to each tone label as a function of duration for voiced speech. For instance, the syllable with voiced duration at 200 ms is labelled 69 times as tone 4, 9 times as tone 2, and 6 times as tone 1. A regression line is fit for each tone label. The Pearson correlation coefficients for tones 1, 2, and 4 are 0.33, 0.65, and –0.79. Correlations for tones 2 and 4 are statistically significant ($p<0.001$).

To further examine whether subject performance can be accounted for by the duration cue, a simple threshold model was used to generate confusion matrices to compare to the observed data. One model aimed for optimal performance (maximum percent correct) while the other aimed to best match the subject data (by minimising the Euclidean distance between the two matrices) (Table 5.3). Stimuli were first sorted by duration, and then classified by a set of two increasing cutoff durations. For example, a stimulus was classified as tone 4 when its duration was shorter than the first cutoff, and classified as tone 2 when the duration was longer than the second cutoff. Any stimulus with duration between the two cutoffs was classified as tone 1. This classification was based on the fact that tone 4 had the shortest mean duration

(259.3 ms) and tone 2 had the longest mean duration (379.5 ms) among the three tones for the stimuli used here (mean duration of tone 1 was 339.6 ms). However, note that different results might be found in different studies, and become more complex when tone 3 is included.

A technique of exhaustive search was used to find the best set of cutoff durations to classify stimuli. Optimal performance was found for cutoff durations of 300 and 370 ms, resulting in 83.3% correct. This result was better than the average performance from implanted children. Only three children (13, 17, and 19) achieved about the same or slightly better scores. The set of cutoff durations for the best match to the observed data was 293 and 365 ms. The prediction was generally close to the data from the implanted children, though not in all detail. For instance, in labelling tone 4, the subjects responded to tone 2 more than tone 1, while the model predicted tone 1 would be used more than tone 2. But all the other numbers and relationships were similar, suggesting subject performance on the processed stimuli with constant amplitude could be achieved purely using the duration cue.

A.

Stimulus Tone	*Subject performance* Response		
	1	2	4
1	**0.60**	0.32	0.08
2	0.25	**0.69**	0.06
4	0.13	0.21	**0.66**

B.

Stimulus Tone	*Model of optimal performance* Response		
	1	2	4
1	**0.75**	0.125	0.125
2	0.125	**0.75**	0.125
4	0	0	**1**

C.

Model of best match to observed data Response		
1	2	4
0.625	0.25	0.125
0.25	**0.75**	0
0.125	0	**0.875**

Table 5.3 Response matrices (in percentage) for A. Subject performance for stimuli without amplitude variation, B. model for optimal performance based on duration, and C. model matched best to subject performance.

5.3 Experiment IIIA: Recognition of tonal contrasts by implanted children - Effect of amplitude envelope and duration

Since the results from experiment III suggested that duration could be an important cue, here, duration was further controlled to investigate various cues that might operate in the perception of lexical tone. This sub-experiment was carried out in two implanted children who were users of the ACE strategy and were excellent performers in the previous experiment. Theoretically, some information about voice f0 is provided by implants with a high stimulation rate such as ACE and CIS. By eliminating amplitude envelope and duration cues, it was possible to examine whether implanted children were able to use temporal pitch information to identify tonal contrasts.

5.3.1 Stimuli and signal processing

Natural speech and three types of processed speech were used as stimuli. There were 48 natural speech tokens, four syllables (/i/, /ba/, /fu/ and /tɕ^{h}i/) with tones 1, 2, and 4 produced twice by one male and one female. All the processed speech preserved its original pitch contour but was manipulated to remove amplitude envelope and/or duration cues. In the ***AFD*** condition, natural speech retained its

original f0, amplitude envelope, and duration. In the ***AF*** condition, natural speech was processed in PRAAT using the PSOLA (Pitch-Synchronous Overlap and Add) method to give a fixed duration for the voiced speech (309 ms). The ***FD*** condition was the same as in the first experiment. Speech was processed to give a relatively constant amplitude envelope in its voiced segment using the procedure described before. These stimuli preserved original f0 contours and duration. The ***F*** condition was processed speech without natural variation in amplitude envelope and duration, but still preserved its original f0 contour. The F-stimuli were generated by further processing FD-stimuli in PRAAT to scale duration using PSOLA. Stimuli were scaled to give equal mean rms for each of the three tones and each of the speakers so as to control the overall intensity. The acoustic cues to tonal contrasts available for implanted listeners in different stimuli were summarised in Table 5.4.

Cues / Conditions	**Temporal pitch (F)**	**Amplitude envelope (A)**	**Duration (D)**
AFD	v	v	v
FD	v		v
AF	v	v	
F	v		

Table 5.4 Summary of different acoustic cues available in different conditions.

5.3.2 Subjects

Two children, who performed over 80% correct in the previous experiment, participated in this study (subject numbers 17 and 20). This sub-experiment was conducted one year later after the Experiment III. One was 8-years-old and the other

was 7-years-old when they participated this study. Both were users of the Nucleus device with the ACE strategy.

5.3.3 Procedure

The 192 stimuli (48 x 4 conditions) were played twice, in two separate randomised orders for a total of 384 stimuli. The testing procedure was the same as in the first experiment, except that stimulus was presented by direct connection between the laptop and the speech processor.

5.3.4 Results and discussions

Overall accuracy

Figure 5.6 shows the results from the two children. The overall percentages correct for stimuli with different acoustic cues are displayed on the left of each panel. For both children, performance on all four conditions was significantly above chance (42.7 %, calculated from the binomial distribution, which revealed that 41 or more out of 96 is statistically different from chance). These two children had a slightly different distribution for their data across conditions. For child 20, the performance on stimuli without either amplitude envelope or duration (conditions AF and FD) was lower than on natural speech (condition AFD), and further decreased on stimuli without both cues (condition F). Frequency of correct response was consistently reduced for the absence of amplitude envelope (AFD vs. FD, AF vs. F) and duration (AFD vs. AF and FD vs. F), indicating both amplitude envelope and duration were used by this child to aid recognition of tonal contrasts. For child 17, performance was less affected by the absence of amplitude envelope and duration. Duration showed some effect for

stimuli with relatively constant amplitude (FD vs. F). For stimuli without both amplitude envelope and duration cues (F), performance of the two children was both significantly above chance. They seemed able to make some use of voice f0 information, though only to a moderate extent. A logistic regression was performed for the effect of condition and listeners, and found that none of the main effects, nor the interaction, was significant. The results for each child were also analyzed separately, and there was still no significant effect. This is likely due to there not being enough data to reach any statistical significance.

Although the results from only two implanted children are not enough to draw any further conclusion, it does show that, at least, there is some evidence for the use of temporal pitch information in recognising tonal contrasts by prelingually deafened implanted children since they performed above chance in condition F. It is likely that amplitude envelope and duration cues may be used by some implant users, but their effects are unlikely to be large.

Performance for individual tones

Results for individual tone are showed on the right of each panel. Information transfer scores were computed from stimulus/response matrices to present an unbiased measure for each tone (for instance, a 2x2 matrix classifying stimuli as tone 1 vs. the other two tones, and responses in the same way). Across all conditions, tone 4 was recognised best among the three tones. For child 20, information transfer scores were consistently reduced when either or both of amplitude envelope and duration were neutralised, and this effect was especially clear for tones 2 and 4.

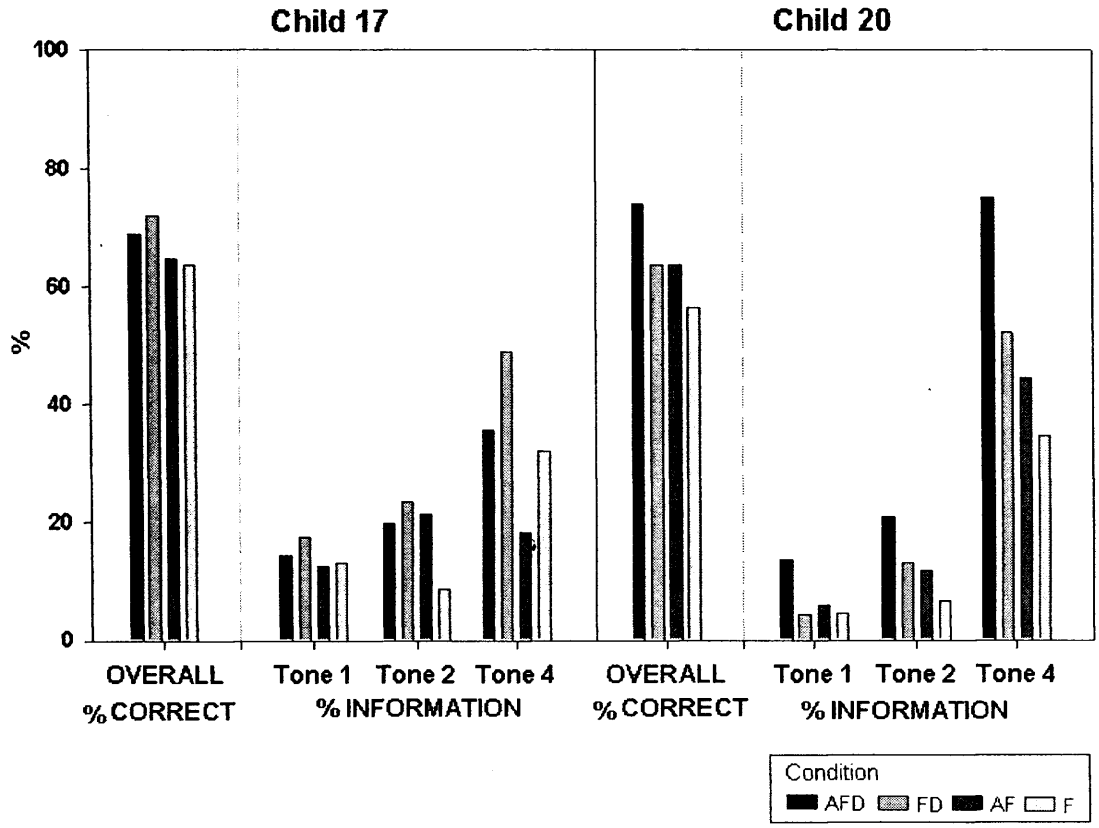

Figure 5.6 Results of further investigation on tone perception from two implanted children. The first set of bars represents percent correct for overall performance across the four conditions, whereas the others represent percent information transfer for each of the three tones.

5.4 Experiment IV: Effect of voice f0 in sentence recognition by implanted children

Aims and experimental predictions

In this experiment, the primary cue for tonal contrasts, f0, was manipulated directly to examine its effect on sentence recognition in implanted children. Results from the previous study in Experiment 2 using acoustic simulations in normal-hearing listeners demonstrated a significant contribution of voice f0 to sentence recognition when spectral information was degraded. The effect of f0 was strong across all age groups of listeners from 6 years to adult. Given the limited information about voice f0 provided in current implant devices, the question that interested us was how much benefit from f0, if any, was received by implanted children.

Unlike normal-hearing listeners, implant users were unable to resolve individual harmonics mainly due to the relatively gross frequency analysis in their implant devices. They might be able to extract some pitch information from the temporal structure of unresolved harmonics, though this temporal pitch information was considered to be rather weak (Faulkner et al., 2000). As in Experiment 2, sentences with natural pitch contours and with slightly falling ones were presented for identification, but here by implanted listeners. Since the pitch information evoked only from the temporal structure of unresolved harmonics is ambiguous and rather weak, it is therefore predicted that the effect of the presence/absence of natural f0 in implanted listeners would not be as strong as what had been observed in normal-hearing listeners shown in the results of Experiment II (chapter 4).

5.4.1 Stimuli and signal processing

Two types of sentence were used as stimuli: unprocessed sentences with natural f0 contours and processed sentences with slightly falling pitch contours. Processed sentences were re-synthesized in PRAAT using PSOLA. Examples of f0 contours of a sentence before and after processing are shown in Figure 5.7. The pitch values of the falling pitch contour were calculated according to the pitch range of the speaker. The initial frequency of the falling contour was a random value within one standard deviation around the mean f0 for each speaker (the range was 140 $\pm$ 8Hz for the male speaker and 240 $\pm$ 14Hz for the female speaker). The final frequency was 6% lower than the initial frequency. The transition was linear in frequency.

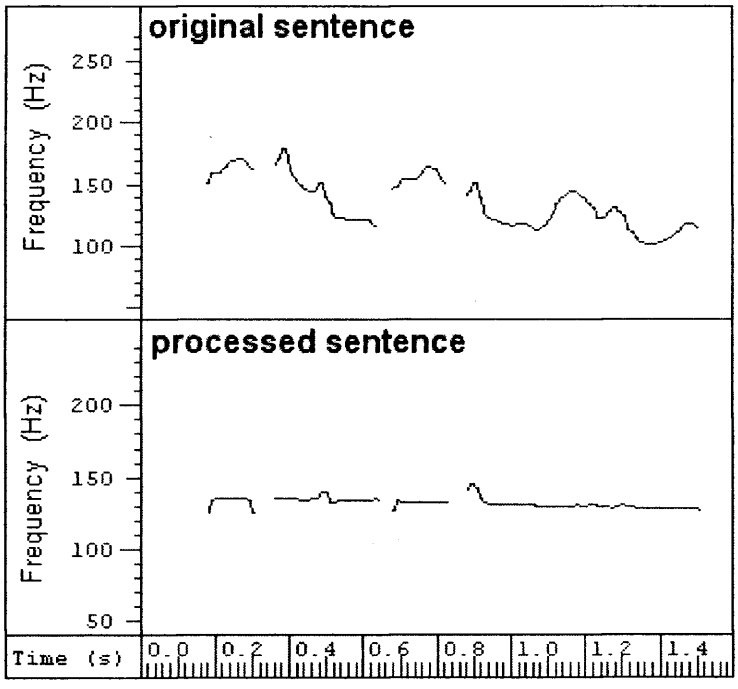

Figure 5.7 Examples of a sentence produced by a male speaker. The upper panel presents the natural f0 contour for the original sentence, and the lower panel presents the slightly falling f0 contour for the processed sentence.

Pitch information available in different stimuli

For unprocessed sentences which contained original f0 variations, implanted listeners might be able to derive some pitch information from the temporal periodicity of unresolved harmonics, and therefore some information about tonal contrasts. For processed sentences, implanted listeners might be also able to derive some information about pitch. However, the processed sentence had slightly falling pitch contours that were irrelevant to tonal contrasts.

5.4.2 Subjects

Subjects were the same as used in Experiment III, save for one (child 6) who did not participate.

5.4.3 Procedure

The data was collected at the same time as the Experiment III. The testing settings and procedures were generally the same as in Experiment III, save for some difference in recording subject responses. The examiner was seated beside the child throughout testing to instruct and to monitor the testing process.

For older children, a GUI built in MATLAB was provided to play the stimuli, and they were allowed to proceed at their own speed. The children were asked to repeat what they heard first and then write down on provided answer sheets. The reason for asking children to repeat back first was to minimise the possibility that they might forget some words, especially those in the end of sentences, once they started writing. If there were some words repeated back but which not been written down, the

examiner would ask the child to say the sentence again, but not to mention if they forgot anything. Only when words repeated at the second time were apparently different from what had been written down, the child would be reminded if he/she wrote down all the words. Only those words which had been written down were scored.

For younger children, the examiner played the stimuli when the child was ready. Children were asked to repeat what they heard and their response was recorded on audiotapes and scored later. Children were asked to repeat their answer again whenever their pronounciation was not clear enough. The tape recorder was placed on the table, at a distance about 30 cm from the children. All children were encouraged to write down or say as many words as possible for their answer, and to guess. A short practice session was given before the test session started, and no feedback was given.

5.4.4 Scoring method

The 'loose keyword' scoring method described in section 4.4.6 'scoring method' was used. For those recorded responses from young children, if the answer could still be understood clearly, key words would be scored as correct even though some tonal contrasts were not produced completely correctly. This was due to the fact that implanted children often had difficulty in producing tonal contrasts, but this sentence task aimed to examine the extent to which implanted children were able to understand, not produce, running speech.

5.4.5 Results and discussion

1. The effect of f0

Here, the recognition scores for sentences with and without natural f0 from implanted children were compared with the data from normal-hearing listeners using acoustic simulations of a cochlear implant in Experiment II. The simulation data were collected from four age groups of normal-hearing listeners (aged 6, 9, 12, and 20 with 10 subjects in each) using vocoder-like techniques with different number of frequency channels (2, 4, 8, and 16) for signal processing. The presence or absence of a natural f0 contour was created by using a f0-controlled pulse carrier or a pulse carrier with a slightly falling f0 contour during voiced speech. A noise carrier was always used for voiceless speech (Details about the signal processing were described in Chapter 4).

Figure 5.8 shows a clear difference for the effect of natural f0 in implanted children (presented by black square symbols) and in normal-hearing subjects listening to simulations of cochlear implants (presented by gray symbols). The simulation study demonstrated clearly that recognition scores were significantly higher for sentences with the natural f0 in normal-hearing listeners. However, results from implanted children showed no difference between natural and processed sentences. Unlike the data from simulations which were spread over to the left of the diagonal, data from implanted children were close the diagonal. Even the two children (17 and 20), who were able to make some use of f0 to recognise tone in isolated syllables, did not show much difference. Group results for sentences with and without natural f0 were about the same (both were around 46% correct). While the natural f0 enhanced the recognition of sentences in normal-hearing listeners using simulations, no significant

difference was found in implanted children between sentences with and without natural f0.

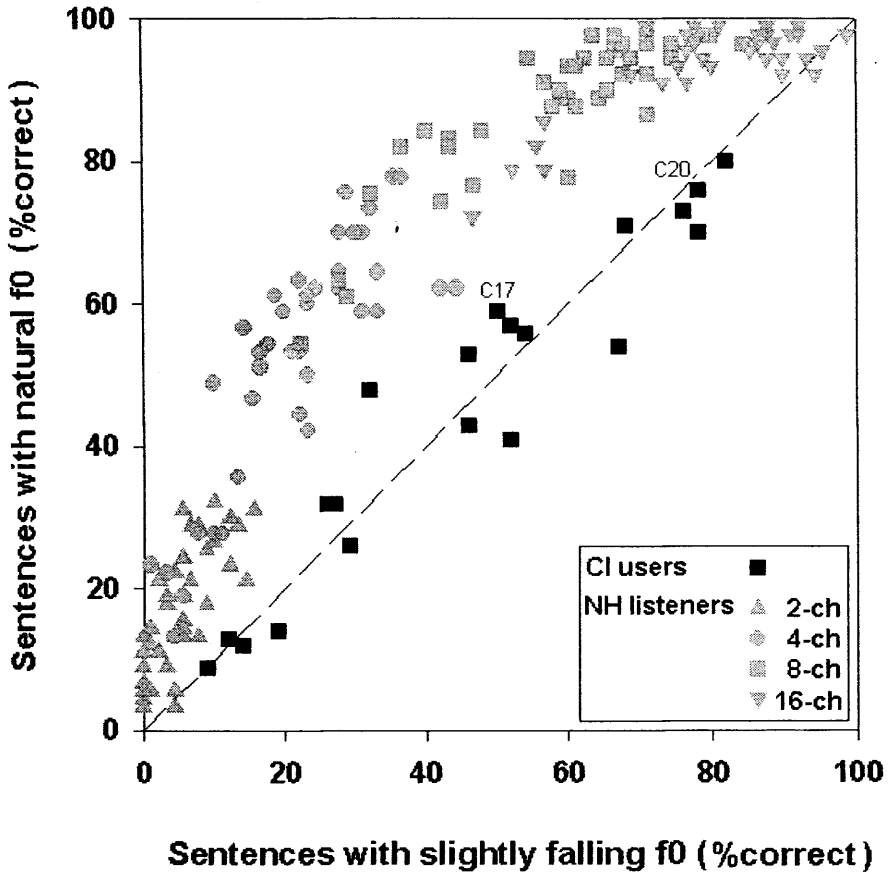

Figure 5.8 Scatterplot for performance on sentences with natural f0 (y-axis) and sentences with slightly falling f0 contours (x-axis). Results of implanted children are presented by black symbols, and those of normal-hearing subjects listening to acoustic simulation are presented in grey, with different symbols representing different numbers of channels in the simulations. Data to the left of the diagonal represent a higher score for sentences with natural f0 than with a slightly falling f0 contour. Results for children 17 and 20, who showed some evidence for using f0 in perceiving tone in the experiment IIIA, are marked.

2. Relationship between sentence recognition and tone recognition, age, age at implantation, and duration of implant use

Figure 5.9 shows a scatterplot of recognition scores for natural sentences and lexical tones by implanted children. In general, children who performed relatively well on sentence recognition also tended have better performance on tone recognition, and vice versa. A significantly positive correlation was found between sentence and tone recognition performance (The Pearson correlation coefficient is 0.67, $p<0.01$).

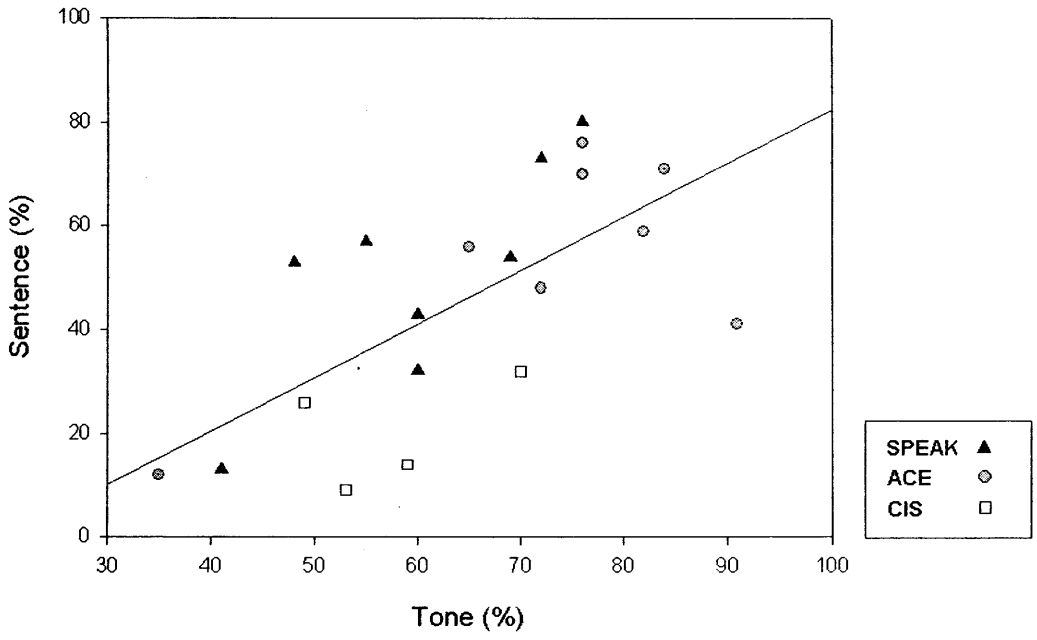

Figure 5.9 Scatterplot for percentage of correct recognition for natural sentences and tones by implanted children.

Figure 5.10 plots results of sentence recognition from the implanted children as a function of age, with reference lines from normal-hearing subjects listening to simulated sentences without natural f0. Overall, the performance of CI children was unrelated to their age (The Pearson correlation coefficient for sentence recognition

and age is -0.15). Note that some implanted children had a very impressive performance. For instance, three 6-year-old children performed at a level that was comparable to the best performance by normal-hearing children of the same age listening to the16-channel simulation. Figure 5.11 and 5.12 show sentence recognition performance as a function of the age at implantation and the duration of implant use. Again, sentence performance was unrelated to the age at implantation and the duration of implant use (The Pearson correlation coefficients are -0.17 and -0.06, repectively).

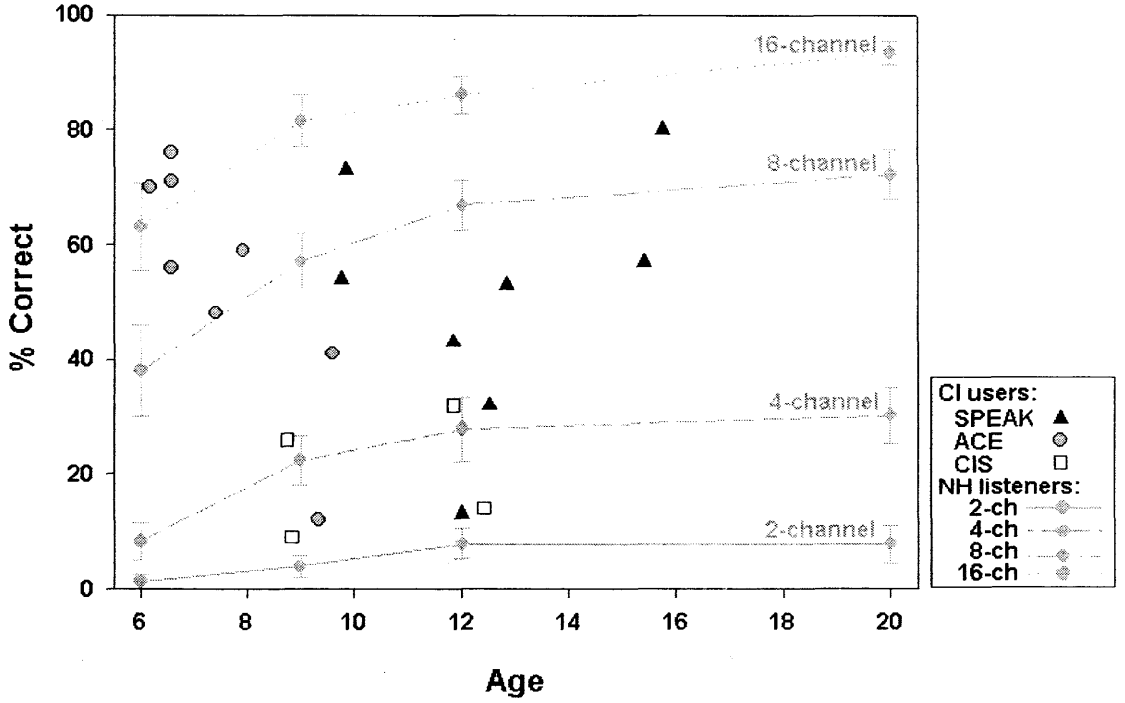

Figure 5.10 Percentage of correct recognition for natural sentences from implanted children, as a function of the age, in comparison to data from four age groups of normal-hearing subjects listening to simulated sentences without natural f0.

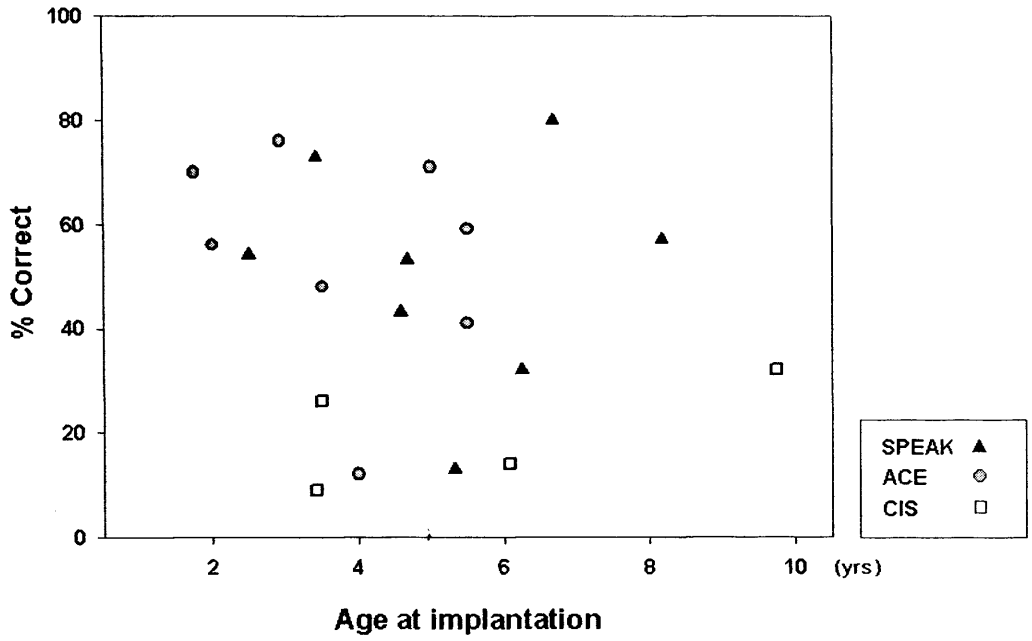

Figure 5.11 Percentage of correct recognition for natural sentences from implanted children, as a function of the age at implantation.

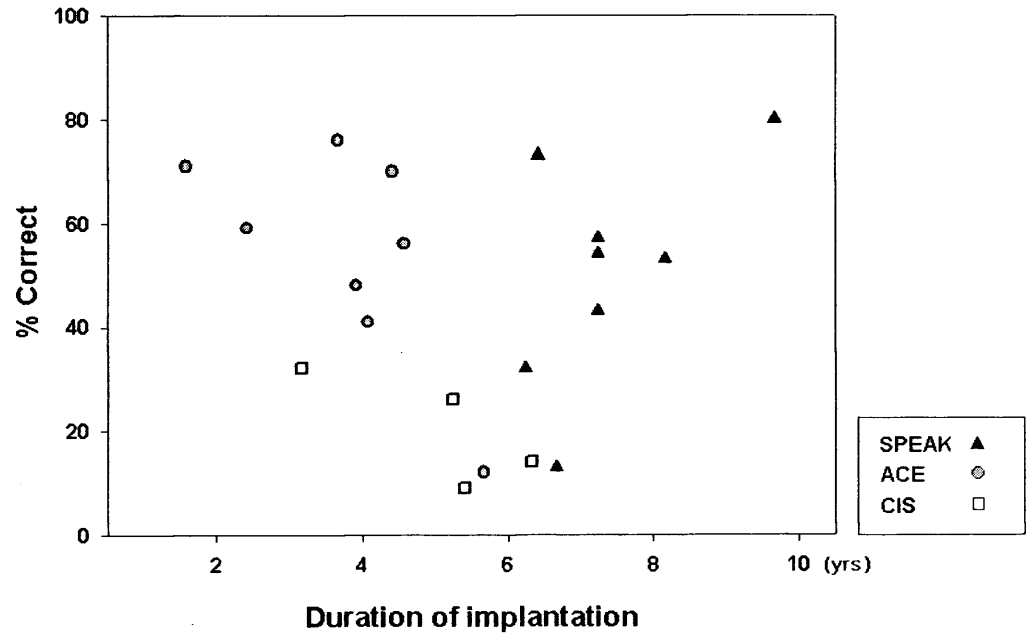

Figure 5.12 Percentage of correct recognition for natural sentences from implanted children, as a function of the duration of implant use.

5.5 General discussion

In this chapter, the use of secondary cues to tone recognition was examined in children using an implant. The results generally showed some effect for the use of amplitude envelope, duration, and temporal pitch information in recognising tonal contrasts. The effect of amplitude envelope was rather small. While simulation studies in normal-hearing listeners reported in Chapter 3 and other studies have demonstrated that amplitude envelope can make a substantial contribution to tone recognition, results from implanted children here have only shown a slight reduction on the overall performance when the natural variations of amplitude contour were neutralised. One possible explanation for the limited effect might be due to the nature of the speech tokens used in the current study. The effect of amplitude envelope on tonal contrasts is thought to be due to the similarity between pitch and amplitude contours, and the change of amplitude is, therefore, used by listeners to interpret the direction of pitch (Whalen & Xu, 1992). However, the similarity between pitch and amplitude contours can vary greatly across speakers (Fu & Zeng, 2000). Some speakers produced less effective amplitude cue than others. Also, the analysis in Chapter 3 showed that the amplitude contour of one tone was not always correlated the best to its own pitch contour (for instance, the amplitude contour of tone 1 for the male speaker was better correlated to the pitch contour of tone 4 than tone1; 0.69 and 0.38, respectively, for the correlation coefficients). Furthermore, tone 3 was not included in the current investigation in implanted children due to its various realizations (falling-rising, low-falling, and rising). Despite the *falling-rising* pattern which has been well known and used most frequently in previous studies, tone 3 appears more often to be the *low-falling* pattern, or sometimes the *rising* pattern due to the third tone sandhi rule in

Mandarin. Studies by Whalen & Xu (1992) and Fu & Zeng (2000) have reported that tones 3 and 4 were better recognised than others on the basis of amplitude envelope alone. The exclusion of tone 3 might also contribute to the relatively small effect of amplitude envelope in the current results.

Results from sentence recognition in implanted children have shown that the presence/absence of natural f0 variations made no difference in recognising sentences. This is likely attributable to the limitation of current cochlear implants in representing voice f0 information. As the results of the simulation study reported in Chapter 4 have demonstrated clearly the significant contribution of voice f0 to tone language users, it is most likely that implant users of a tone language will further benefit if better f0 information can be transmitted through implants. There have been a number of studies investigating how to improve pitch perception in implanted listeners, such as adding an extra coding of explicit voice f0, providing a more salient representation for temporal pitch in the speech envelope, or better place-coding (e.g. Jones, 1994; Geurts & Wouters, 2001; 2004; Green, Faulkner, & Rosen, 2004). However, the results so far are still not very encouraging. Before better f0 information can be made available to implanted listeners, it may bring some benefits for tone language users if tone perception can be improved by enhancing other acoustic cues to tonal contrasts. Although duration can be an important cue for tonal contrasts in isolated syllables, it may not be very useful in running speech. Amplitude envelope, on the other hand, may be a possible approach to improve tone perception in implant users. A study by Luo & Fu (2004) directly manipulated the similarity between f0 and amplitude contours in Mandarin tones using a noise-vocoder technique in 6 normal-hearing listeners. They reported a significant improvement in tone recognition with increasing similarity of pitch and amplitude contours (average performance increased by more

than 10 percentage points), suggesting that the modification of amplitude envelope to better resembling the f0 contour could be a possible approach to deliver better tonal information for implanted listeners. The studies reported in the current chapter have shown some evidence for the use of amplitude envelope in implanted listeners, and it is likely that implant users may further benefit from the amplitude modification to closely resemble the change of f0.

5.6 Summary

- Implanted children were able to recognise tone in isolated syllables to some extent, and some achieved a high level of performance. The absence of amplitude envelope caused some decrease in tone recognition performance, especially for implant users with SPEAK. However, the overall effect was generally small.
- Both amplitude envelope and duration could be used to aid in recognition of tonal contrasts by implanted children. Duration was particularly useful for recognition of tone 4.
- In contrast to the significant effect in normal-hearing subjects listening to acoustic simulations, the absence of natural f0 variations did not show any significant effect on sentence recognition in implanted listeners. This indicated that the representation of the information about voice f0 was extremely limited; therefore, the presence or absence of natural f0 made no difference in recognising running speech for implanted children.

Chapter 6

General discussion

The role of pitch and tonal information in tonal languages

For profoundly hearing-impaired people who receive little or no benefit from conventional hearing aids, cochlear implants provide an alternative in restoring some hearing. The advance of implant techniques in the latest 10 years has brought remarkable improvements in speech communication for implant users, though there are still some limitations in current devices. Since the development of cochlear implants has mainly been based on data from profoundly hearing-impaired English speakers, it is important to determine if this device provides sufficient information for users of all languages. Compared to the numbers of studies conducted in English, only a very limited number of studies have been carried out in users of other languages, especially focusing on the effectiveness in terms of different features of different languages.

This thesis investigated the use of cochlear implants in users of a tone language, from the acoustic features available for tonal contrasts to the effect of natural variations of voice f0 on high-level speech understanding. The results of tone recognition by normal-hearing subjects listening to simplified stimuli with different acoustic cues (Experiment I) demonstrated clearly the difference between the performance that could be achieved by a clear indication of voice f0 and that which was possible with only temporal cues (amplitude envelope, duration, and temporal pitch information). For stimuli in which f0 was conveyed by sawtooth carriers (SawAFD, SawAF, SawFD, and SawF), information from both resolved and

unresolved harmonics was available for determining pitch, and these cues evoked a clear pitch percept. Listeners therefore performed very well in recognising tonal contrasts. For stimuli with f0 carried by noise carriers (NoiseAFD, NoiseAF, NoiseFD, and NoiseF), pitch perception was based purely on the temporal fluctuations of modulated noise. Because temporal periodicity is a relatively weak cue to pitch, tone recognition performance based on the temporal pitch information was never comparable to that achieved with a clear pitch percept. The performance without a clear indication of f0 provided some insights about the difficulty that implant users might encounter in recognising this essential phonemic feature in their language. Even though there may be secondary cues to tone in Mandarin (such as amplitude envelope and duration) which aid in the recognition of tonal contrasts, the benefit is very limited.

The investigations on sentence perception examined how much speech understanding would be affected by the lack of natural pitch contours. Sentences with the original f0 contour of natural speech and with a slightly-falling neutralised f0 contour were presented to normal-hearing and implanted listeners for identification. Results from normal-hearing subjects listening to acoustic simulations of implant processing showed a strong effect of natural f0 information in sentence recognition (Experiment II), while results from the implanted children found no effect at all for the presence/absence of natural f0 (Experiment IV, see Figure 5.8). This indicated that, because of the gross frequency resolution in present devices, implant users seem unable to take advantage of the presence of natural voice pitch information when listening to running speech. For normal-hearing listeners, pitch perception can be derived from information about the frequencies of resolved harmonics and the periodicity of unresolved harmonics. When listening to vocoded sentences with

natural f0 contours (FxNx sentences), the clear pitch percept allowed listeners easily to perceive tonal contrasts, and therefore aided the recognition of speech. Normal-hearing listeners could also obtain a clear pitch percept for those sentences with neutralised slightly-falling f0 contours (VxNx sentences). However, the pitch information was irrelevant to the original tonal contrasts. For implanted listeners, on the other hand, only the periodicity information of unresolved harmonics could be used to determine pitch. Since periodicity information is a relatively weak cue to pitch, the presence/absence of natural f0 has little or no effect on recognising sentences. Although the results of tone identification experiments from implanted children (Experiment IIIA) suggested that implanted users might be able to make some use of voice f0 in isolated syllables for recognising tonal contrasts, this relatively weak pitch information might only be perceived in highly constrained situations, but hardly could be used in a more real-life running speech situation.

Lexical tone and speech perception in implanted children

Given the limited contributions that can be made by all the three temporal cues to tonal recognition, it is surprising that a few implant users were able to achieve a high level of performance on tone recognition. This is even more amazing for users with SPEAK, a speech coding strategy with a relatively slow stimulation rate in which the periodicity information is unlikely to be well represented by the temporal fluctuations of speech envelope. Some extremely good performers have also been reported in other studies. For instance, in the study by Peng *et al.* (2004), 6 out of 30 Mandarin speaking children reached about 90% correct on a tone identification task, and 4 of them were SPEAK users (chance 50%). In addition, in Cantonese speaking children, Barry *et al.* (2002) reported that one SPEAK user achieved 100% correct in

detecting the change of different tones for 14 out of 15 tonal pairs, and another SPEAK user also performed 100% correct on 12 out of 15 tonal pairs (average scores for both children were over 95%, chance 50%). While Mandarin tones might be cued by acoustic characteristics other than f0, Cantonese tones were almost exclusively signalled by their f0 patterns, and the overall intensity and duration of the stimuli used in Barry *et al.* (2002) were both controlled. Another possible cue, which might be used to recognise tonal contrasts and also available for SPEAK users, is the spectral movement across channels. This cue may be more salient for speech produced with a higher voice pitch such as a woman's or a child's voice. However, it is expected to be rather difficult to detect voice pitch by the spectral movement in real speech due to the obstruction of the formant movement of vowels (Green *et al.*, 2002).

Today, it is estimated that more than 100,000 people around the world have received cochlear implantation, and more than one-third are children. Implantation has now been applied to prelingually deafened children under 2 years. For children whose speech development mainly relies on electrical hearing from their implant devices, one essential issue is, with only limited information about tonal contrasts provided, how this would affect the process of spoken language acquisition in children learning a tone language. In normal-hearing children, the acquisition of lexical tone has been reported to be mastered at a very young age, well before the complete development of segmentals (e.g. Li & Thompson, 1977; Clumeck, 1980; So & Dodd, 1995). However, this is not the case for implanted children. Studies assessing the production of lexical tones in children using implants have reported that the development in tonal inventory was often slower than in vowel inventory (Barry, Blamey, Lee, & Cheung, 2000; Barry & Blamey, 2004; Xu *et al.*, 2004). Studies on tone perception have also shown no significant improvement over the period of

implant use, contrary to performance on vowels, consonants, and sentences which improve significantly over time for the majority of implanted children (Wei *et al.*, 2000; Wu & Yang, 2003). Although the exact relationship between tone perception and production of lexical tones has not yet been completely clear, investigation in implanted children generally suggests that good perception on tonal contrasts might be necessary for good performance on tone production (Barry *et al.*, 2004; Peng *et al.*, 2004). For instance, the results in the study by Peng *et al.* (2004) have shown that, although a high performance in recognising tonal contrasts did not always lead to a good performance on tone production, a low performance on tone perception never did. With insufficient information about voice f0 represented in current implant systems, most implanted children often show some difficulty in both recognising and producing tonal contrasts, and may subsequently have some effect on their speech understanding and intelligibility. Providing better information about voice f0 should further benefit implanted children, especially tone languages users, in speech communication, let alone music perception and understanding speech in noise as well. If implanted children can obtain better tonal information, even just with one speaker, one-to-one communication in a highly constrained environment, it could be make a significant difference to language development.

References

Arndt P., Staller S., Arcaroli J., Hines A. & Ebinger K. (1999) Within-subject comparison of advanced coding strategies in the Nucleus 24 cochlear implant. Cochlear Corporation Report (cited in Clark, G.: Cochlear implants - Fundamentals & Applications, 2003, New York:Springer).

Au D.K.K. (2003) Effects of stimulation rates on Cantonese lexical tone perception by cochlear implant users in Hong Kong. *ClinicalOtolaryngology and Allied Science* 28, 533-538.

Barry J.G., Blamey P.J., Lee K. & Cheung D. (2000) Differentiation in tone production in Cantonese-speaking hearing-impaired children, Proceedings of the 6th International Conference on Spoken Language Processing, 16-20 October. 1, 669-672. Beijing: China Military Friendship Publish.

Barry J.G., Blamey P.J., Martin L.F.A., Lee K.Y.S., Tang T., Ming Y.Y. & van Hasselt C.A. (2002) Tone discrimination in Cantonese-speaking children using a cochlear implant. *Clinical Linguistics and Phonetics* 16, 79-99.

Barry J.G. & Blamey P.J. (2004) The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese. *Journal of the Acoustical Society of America* 116, 1739-1748.

Battmer R.D., Zilberman Y., Haake P. & Lenarz T. (1999) SAS-CIS pilot comparison study in Europe. *Annals of Otology, Rhinology, and Laryngology* 177 (Suppl), 69-73.

Bench J. & Bamford J. (1979) *Speech-hearing tests and the spoken language of hearing-impaired children.* Academic Press, London.

Blamey P., Arndt P., Bergeron F., Bredberg G., Brimacombe J., Facer G., Larky J., Lindstrom B., Nedzelski J., Peterson A., Shipp D., Staller S. & Whitford L. (1996) Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants. *Audiology Neuro-Ololology* 1, 294-306.

Blicher D.L., Diehl R.L. & Cohen L.B. (1990) Effects of syllable duration on the perception of the Mandarin tone 2/tone 3 distinction: evidence of auditory enhancement. *Journal of Phonetics* 18, 37-49.

Boex C., Pelizzone M. & Montandon P. (1996) Speech recognition with a CIS strategy for the Ineraid multichannel cochlear implant. *American Journal of Otology* 17, 61-68.

Brill S.M., Gstottner W., Helms J., Ilberg C., Baumgartner W., Muller J. & Kiefer J. (1997) Optimization of channel number and stimulation rate for the fast continuous interleaved sampling strategy in the COMBI 40+. *American Journal of Otology* 18, S104-S106.

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

Busby P.A., Tong Y.C. & Clark G.M. (1993) The perception of temporal modulations by cochlear implant patients. *Journal of the Acoustical Society of America* 94, 124-131.

Busby P.A., Whitford L.A., Blamey P.J., Richardson L.M. & Clark G.M. (1994) Pitch perception for different modes of stimulation using the cochlear multiple-electrode prosthesis. *Journal of the Acoustical Society of America* 95, 2658-2669.

Busby P.A. & Clark G.M. (2000) Pitch estimation by early-deafened subjects using a multiple-electrode cochlear implant. *Journal of the Acoustical Society of America* 107, 547-558.

Chao Y.-R. (1948) *Mandarin Primer: A intensive course in spoken Chinese.* Harvard University Press, Cambridge, MA.

Chao Y.-R. (1968) *A Grammar of Spoken Chinese.* University of California: Berkeley & Angeles, Berkeley, CA.

Chen M.Y. (2000) *Tone sandhi - patterns across Chinese dialects.* Cambridge University Press.

Ciocca V., Francis A.L., Aisha R. & Wong L. (2002) The perception of Cantonese lexical tones by early-deafened cochlear implantee. *Journal of the Acoustical Society of America* 111, 2250-2256.

Clark G. (2003) *Cochlear implants: Fundamentals & Applications.* New York: Springer.

Clark G.M., Blamey P.J., Brown A.M., Busby P.A., Dowell R.C., Franz B.K.H., Pyman B.C., Shepherd R.K., Tong Y.C., Webb R.L., Hirshorn M.S., Kuzma J.A., Mecklenbury D.J., Money D.K., Patrick J.F. & Seligman P.M. (1987) The University of Melbourne - Nucleus multi-electrode cochlear implant. *Advances in Oto-Rhino-Laryngology* 38. Basel: Karger.

Clark G.M., Cowan R.S.C. & Dowell R.C. (1997) *Cochlear implantation for infants and children - advances*. Singular Publishing, San Diego.

Clumeck H. (1980) The acquisition of tone. In: *Child Phonology* (eds Yeni-komshian G.H., Kavanagh J.F. & Ferguson C.A.), pp. 257-275. Academic Press, New York.

Collett D. (2003) *Modelling binary data*, 2nd edn. Boca Raton, FL: CRC Press.

Dorman M.F. (1993) Speech perception by adults. In: *Cochlear Implants: Audiological Foundations* (ed Tyler R.S.), pp. 145-190. Singular Publishing, San Diego.

Dorman M.F. & Loizou P.C. (1997) Mechanisms of vowel recognition for Ineraid patients fit with continuous interleaved sampling processors. *Journal of the Acoustical Society of America* 102, 581-587.

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

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

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 and Hearing* 19, 162-166.

Dorman M.F., Loizou P.C., Fitzke J. & Tu Z. (1998a) The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels. *Journal of the Acoustical Society of America* 104, 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 and Hearing* 19, 481-484.

Dorman M.F. (2000) Speech perception by adults. In: *Cochlear implants* (eds Waltzman S.B. & Cohen N.L.), pp. 317-329. Thieme, New York.

Dorman M.F., Loizou P.C., Kemp L.L. & Kirk K.I. (2000) Word recognition by children listening to speech processed into a small number of channels: data from normal-hearing children and children with cochlear implants. *Ear and Hearing* 21, 596.

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. *Journal of Speech, Language, and Hearing Research* 45, 783-788.

Dorman M.F. & Ketten D. (2003) Adaptation by a cochlear-implant patient to upward shifts in the frequency representation of speech. *Ear and Hearing* 24, 457-460

Dowell R.C., Seligman P.M., Blamey P.J. & Clark G.M. (1987a) Evaluation of a two-formant speech-processing strategy for a multichannel cochlear prosthesis. *Annals of Otology, Rhinology and Laryngology* 96 (Suppl. 128), 132-134.

Dowell R.C., Seligman P.M., Blamey P.J. & Clark G.M. (1987b) Speech perception using a two-formant 22-electrode cochlear prosthesis in quiet and in noise. *Acta Oto-Laryngologica* 104, 439-446.

Dowell R.C., Dawson P.W., Dettman S.J., Shepherd R.K., Whitford L.A., Seligman P.M. & Clark G.M. (1991) Multichannel cochlear implantation in children: a summary of current work at the University of Melbourne. *The American Journal of Otology* 12 (Suppl), 137-143.

Eddington D.K. (1980) Speech discrimination in deaf subjects with cochlear implants. *Journal of the Acoustical Society of America* 68, 885-891.

Edgerton B., Prietto A. & Danhauer J.L. (1983) Cochlear implant patient performance on the MAC battery. *Otolaryngologic Clinics of North America* 16, 267-280.

Eisenberg L.S., Shannon R.V., Martinez A.S. & Wygonski J. (2000) Speech recognition with reduced spectral cues as a function of age. *Journal of the Acoustical Society of America* 107, 2704-2710.

Eisenberg L.S., Martinez A.S., Holowecky S.R. & Pogorelsky S. (2002) Recognition of lexically controlled words and sentences by children with normal hearing and children with cochlear implants. *Ear and Hearing* 23, 450-462.

Elliott L.L. (1979) Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *Journal of the Acoustical Society of America* 66, 651-653.

Faulkner A., Rosen S. & Smith C. (2000) Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech:

Implications of cochlear implants. *Journal of the Acoustical Society of America* 108, 1877-1887.

Faulkner A., Rosen S. & Wilkinson L. (2001) Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking trough a simulated cochlear implant speech processor. *Ear and Hearing* 22, 431-438.

Faulkner A., Rosen S. & Stanton D. (2003) Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth. *Journal of the Acoustical Society of America* 113, 1073-1080.

Fearn R.A. (2001) Music and pitch perception of cochlear implant recipients. PhD Thesis. University of New South Wales.

Fernald A. & Simon T. (1984) Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* 20, 104-113.

Fishman K.E., Shannon R.V. & Slattery W.H. (1997) Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant. *Journal of Speech, Language, and Hearing Research* 40, 1201-1205.

Fok Chan Y.Y.A. (1974) *A perceptual study of tones in Cantonese*. Centre of Asian Studies, University of Hong Kong, Hong Kong.

Fok Chan Y.Y.A. (1984) The teaching of tones to children with profound hearing impairment. *British Journal of Disorders of Communication* 19, 225-236.

Fon J. & Chiang W.-Y. (2000) What does Chao have to say about tones? - a case study of Taiwan Mandarin. *Journal of Chinese Linguistics* 27, 13-37.

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. *Journal of the Acoustical Society of America* 110, 1150-1163.

Frijns J.H.M., de Snoo S.L. & Schoonhoven R. (1995) Potential distributions and neural excitation patterns in a rotationally symmetric model of the electrically stimulated cochlear. *Hearing Research* 87, 170-186.

Frijns J.H.M., de Snoo S.L. & ten Kate J.H. (1996) Spatial selectivity in a rotationally symmetric model of the electrically stimulated cochlear. *Hearing Research* 95, 170-186.

Frijns J.H.M., Briaire J.J. & Grote J.J. (2001) The importance of human cochlear anatomy for the results of modiolus-hugging multichannel cochlear implants. *Otology and Neurotology* 22, 340-349.

Frijns J.H.M., Briaire J.J., de Laat J.A.P.M. & Grote J.J. (2002) Initial evaluation of the Clarion CII cochlear implant: speech perception and neural response imaging. *Ear and Hearing* 23, 184-197.

Frijns J.H.M., Klop W.M.C., Bonnet R.M. & Briaire J.J. (2003) Optimizing the number of electrodes with high-rate stimulation of the Clarion CII cochlear implant. *Acta Oto-Laryngologica* 123, 138-142.

Fryauf-Bertsch H., Tyler R.S., Kelsay D.M.R., Gantz B.J. & Woodworth G.G. (1997) Cochlear implant use by pre lingually deafened children: the influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research* 40, 183-199.

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

Fu Q.-J., Zeng F.-G., Shannon R.V. & Soli S.D. (1998) Importance of tonal envelope cues in Chinese speech recognition. *Journal of the Acoustical Society of America* 104, 505-510.

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

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

Fu Q.-J. & Zeng F.-G. (2000) Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing* 5, 45-57.

Fu Q.-J., Shannon R.V. & Galvin III J.J. (2002) Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant. *Journal of the Acoustical Society of America* 112, 1664-1674.

Fu Q.-J. & Galvin III J.J. (2003) The effects of short-term training for spectrally mismatched noise-band speech. *Journal of the Acoustical Society of America* 113, 1065-1072.

Fu Q.-J., Hsu C.-J. & Horng M.-J. (2004) Effects of speech processing strategy on Chinese tone recognition by Nucleus-24 cochlear implant users. *Ear and Hearing* 25, 501-508.

Gandour J. (1983) Tone perception in far eastern languages. *Journal of Phonetics* 11, 149-175.

Garding E., Kratochvil P., Svantesson J.O. & Zhang J. (1986) Tone 4 and tone 3 discrimination in Modern Standard Chinese. *Language and Speech* 29, 281-293.

Geurts L. & Wouters J. (2001) Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants. *Journal of the Acoustical Society of America* 109, 713-726.

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

Goldstein J.L. (1973) An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America* 54, 1496-1516.

Green T., Faulkner A. & Rosen S. (2002) Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants. *Journal of the Acoustical Society of America* 112, 2155-2164.

Green T., Faulkner A. & Rosen S. (2004) Enhancing temporal cues to voice pitch in continuous-interleaved-sampling cochlear implants. *Journal of the Acoustical Society of America* 116, 2298-2310.

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

Greenwood D.D. (1990) A cochlear frequency-position function for several species - 29 years later. *Journal of the Acoustical Society of America* 87, 2592-2605.

Gstoettner W.K., Adunka O., Franz P. & Hamzavi J.J. (2001) Perimodiolar electrodes in cochlear implant surgery. *Acta Oto-Laryngologica* 121, 216-219.

Hillenbrand J.M. (2003) Some effects of intonation contour on sentence intelligibility. *Journal of the Acoustical Society of America* 114, 2338.

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

Houtsma A.J.M. & Goldstein J.L. (1972) The central origin of the pitch of pure tones: evidence from musical interval recognition. *Journal of the Acoustical Society of America* 51, 520-529.

Howie J.M. (1976) *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press, Cambridge.

Hsu C.-J., Horng M.-J. & Fu Q.-J. (2000) Effects of the number of active electrodes on tone and speech perception by Nucleus 22 cochlear implant users with SPEAK strategy. *Advances in Oto-Rhino-Laryngology* 57, 257-259.

Huang T.-S., Wang N.-M. & Liu S.-Y. (1996) Nucleus 22-channel cochlear mini-system implantations in Mandarin-speaking patients. *The American Journal of Otology* 17, 46-52.

Jones P.A., McDermott H.J., Seligman P.M. & Millar J.B. (1994) Coding of voice source information in the Nucleus cochlear implant system. *Annals of Otology, Rhinology and Laryngology* 104, Supplement 166 - September 1995, 363-365.

Jusczyk P.W. (1997) *The discovery of spoken language*. The MIT Press.

Ketten D.R., Viemeister N.F., Skinner M.W., Gates G.A., Wang G. & Neely J.G. (1998) *In vivo* measures of cochlear length and insertion depth of Nucleus cochlear implant electrode arrays. *Annals of Otology, Rhinology and Laryngology* 107, 1-16.

Kiefer J., von Ilberg C., Rupprecht V., Huber-Egener J., Baumgartner W., Gstottner W. & Stephan K. (1999) Optimized speech understanding with the CIS-speech coding strategy in cochlear implants: The effect of variations in stimulus rate and numbers of channels. In: *Cochlear implants* (eds Waltzman S.B. & Cohen N.) pp. 339-340. Thieme Medical and Scientific, New York.

Kiefer J., Hohl S., Sturzebecher E., Pfennigdorff T. & Gstoettner 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, 32-42.

Kwong Y.-Y. & Zeng F.-G. (2004) Temporal and spectral cues in Mandarin. *Journal of the Acoustical Society of America* 115, 2545.

Kwok C.L., Wong C.M., So K.W., Yiu M.L., Lau C.C., Luk W.S. & Tang S.O. (1991) Speech and lexical-tone perception in Cantonese-speaking cochlear implant patients. *Australian Journal of Human Communication Disorders* 19, 77-90.

Lan N., Nie K., Gao S. & Zeng F.G. (2004) A novel speech processing strategy of cochlear implants incorporating tonal information. *IEEE Transactions on Biomedical Engineering* 51, 752-760.

Lawson D.T., Wilson B.S., Zerbi M. & Finley C.C. (1996) Speech processors for auditory prostheses: 22 electrode percutaneous study - results for the first five subjects. Third Quarterly Progress Report, NIH project N01-DC-5-2103, Neural Prosthesis Program, National Institudes of Health, Bethesda, MD.

Lee K.Y.S., van Hasselt C.A., Chiu S.N. & Cheung D.M.C. (2002) Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children. *International Journal of Pediatric Otorhinolaryngology* 63, 137-147.

Li C. & Thompson S. (1977) The acquisition of tone in Mandarin-speaking children. *Journal of Child Language* 4, 185-199.

Lin H.-B. & Repp B.H. (1989) Cues to the perception of Taiwanese tones. *Language and Speech* 32, 25-44.

Lin M.-C. (1988) The acoustic characteristics and perceptual cues of tones in Standard Chinese. *Chinese Yuwen* 204, 182-193.

Liu S.-Y., Huang T.-S. & Follent M. (1997) The field trial of the SPEAK versus MPEAK speech coding strategies in Mandarin Chinese. *Advances in Oto-Rhino-Laryngology* 52, 113-116.

Liu T.-C., Chen H.-P. & Lin H.-C. (2004) Effects of limiting the number of active electrodes on Mandarin tone perception in young children using cochlear implants. *Acta Oto-Laryngologica* 124, 1149-1154.

Loizou P.C. (1998) Mimicking the human ear: an overview of signal-processing strategies for converting sound into electrical signals in cochlear implants. *IEEE Signal Processing Magazine* September, 101-130.

Loizou P.C., Dorman M.F. & Tu Z. (1999) On the number of channels needed to understand speech. *Journal of the Acoustical Society of America* 106, 2097-2103.

Loizou P.C., Dorman M.F., Tu Z. & Fitzke J. (2000a) The recognition of sentence in noise by normal-hearing listeners using simulations of SPEAK-type cochlear implant processors. *Annals of Otology, Rhinology and Laryngology* 109 (12, Suppl. 185), 67-68.

Loizou P.C., Poroy O. & Dorman M.F. (2000b) The effect of parametric variations of cochlear implant processors on speech understanding. *Journal of the Acoustical Society of America* 108, 790-802.

Luo X. & Fu Q.-J. (2004) Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. *Journal of the Acoustical Society of America* 116, 3659-3667.

MacLeod A. & Summerfield Q. (1990) A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology* 24, 29-43.

McDermott H.J. & McKay C.M. (1994) Pitch ranking with non-simultaneous dual-electrode electrical stimulation of the cochlea. *Journal of the Acoustical Society of America* 96, 155-162.

Merzenich M.M., Rebscher S.J., Loeb G.E., Byers C.L. & Schindler R.A. (1984) The UCSF cochlear implant project: state of development. *Advances in Audiology* 2, 119-144.

Miller G.A. & Taylor W. (1948) The perception of repeated bursts of noise. *Journal of the Acoustical Society of America* 20, 171-182.

Moore B.C.J. (1996) Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear and Hearing* 17, 133-160.

Moore B.C.J. (2003) Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implant. *Otology and Neurotology* 24, 243-254.

Moore B.C.J. & Peter R.W. (1992) Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity. *Journal of the Acoustical Society of America* 91, 2881-2893.

Moore B.C.J. & Rosen S.M. (1979) Tune recognition with reduced pitch and interval information. *Quarterly Journal of Experimental Psychology* 31, 229-240.

Nelson D.A., Van Tasell D.J., Schroder A.C., Soli S. & Levine S. (1995) Electrode ranking of "place pitch" and speech recognition in electrical hearing . *Journal of the Acoustical Society of America* 98, 1987-1999.

O'Halpin R. (2001) Intonation issues in the speech of hearing impaired children: analysis, transcription and remediation. *Clinical Linguistics and Phonetics* 15, 529-550.

Osberger M.J. & Fisher L. (2000) New directions in speech processing: patient performance with simultaneous analog stimulation. *American Journal of Otology* 185 (suppl.), 70-73.

Patrick J.F. & Clark G.M. (1991) The Nucleus 22-channel cochlear implant system. *Ear and Hearing* 12 (4 Suppl), 3S-9S.

Pelizzone M., Cosendai G. & Tinembart J. (1999) Within-patient longitudinal speech reception measures with continuous interleaved sampling processors for Ineraid implanted subjects. *Ear and Hearing* 20, 228-237.

Peng S.-C., Tomblin J.B., Cheung H., Lin Y.-Y. & Wang L.-S. (2004) Perception and production of Mandarin tones in prelingually deaf children with cochlear implants. *Ear and Hearing* 25, 251-264.

Plomp R. (1967) Pitch of complex tones. *Journal of the Acoustical Society of America* 41, 1526-1533.

Ritsma R.J. (1967) Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America* 42, 191-198.

Rosen S., Walliker J., Brimacombe J.A. & Edgerton B.J. (1989) Prosodic and segmental aspects of speech perception with the House/3M single-channel implant. *Journal of Speech and Hearing Research* 32, 93-111.

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

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

Rossi M. (1978) Interactions of intensity glides and frequency glissandos. *Language and Speech* 21, 384-396.

Rubinstein J.T., Wilson B.S., Finley C.C. & Abbas P.J. (1999) Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation. *Hearing Research* 127, 108-118.

Rubinstein J.T. (2002) Paediatric cochlear implantation: prosthetic hearing and language developement. *Lancet* 360, 483-485.

Shannon R.V. (1983) Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics. *Hearing Research* 11, 157-189.

Shannon R.V. (1992) Temporal modulation transfer functions in patients with cochler implants. *Journal of the Acoustical Society of America* 91, 2156-2164.

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

Shannon R.V., Zeng F.-G. & Wygonski J. (1998) Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America* 104, 2467-2476.

Shannon R.V., Fu Q.-J. & Galvin III J. (2004) The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Oto-Laryngologica* Suppl 552, 50-54.

Schouten J.F., Ritsma R.J., & Cardozo B.L. (1962) Pitch of the residue. *Journal of the Acoustical Society of America* 34, 1418-1424.

Shen X.S. & Lin M. (1991) A perceptual study of Mandarin tones 2 and 3. *Language and Speech* 34, 145-156.

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

Shih C.-L. (1988) Tone and intonation in Mandarin. *Working Papers of Cornell Phonetics Laboratory* 3, 83-97.

Skinner M.W., Holden L.K., Holden T.A., Dowell R.C., Seligman P.M., Brimacombe J.A. & Beiter A.L. (1991) Performance of postlinguistically deaf adults with the Wearable Speech Processor (WSP III) and Mini Speech Processor (MSP) of the Nucleus multi-electrode cochlear implant. *Ear and Hearing* 12, 3-22

Snow C.E. & Ferguson C.A. (1977) *Talking to children: Language input and acquisition.* Cambridge University Press, Cambridge.

So L.K.H. & Dodd B.J. (1995) The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language* 22, 473-495.

Sun J.C., Skinner M.W., Liu S.-Y., Wang F.N.M., Huang T.S. & Lin T. (1998) Optimization of speech processor fitting strategies for Chinese-speaking cochlear implantees. *Laryngoscope* 108, 560-568.

Tang S.O., Luk W.S. & Lau C.C. (1990) Cochlear implant in Hong Kong Cantonese. *The American Journal of Otology* 11, 421-426.

Terhardt E. (1974) Pitch, consonance, and harmony. *Journal of the Acoustical Society of America* 55, 1061-1069.

Tseng C.-Y. (1990) *An acoustic phonetic study on tones in Mandarin Chinese.* Institute of History & Philology, Academia Sinica, Special Publications, No. 94, Taipei, Taiwan.

Tong M.C.F., Cheung D.M.C., Lee K.Y.S., Wong T.K.C., Leung E.K.S. & van Hasselt C.A. (2000) Perspectives in cochlear implantation in a tone language population. In: *Cochlear Implants* (eds Waltzman S.B. & Cohen N.L.), pp. 353-354. Thieme, New York.

Tong Y.C., Clark G.M., Blamey P.J., Busby P.A. & Dowell R.C. (1982) Psychophysical studies for tow multiple-channel cochlear implant patients. *Journal of the Acoustical Society of America* 71, 153-160.

Tong Y.C. & Clark G.M. (1985) Absolute identification of electric pulse rates and electrode positions by cochlear implant patients. *Journal of the Acoustical Society of America* 77, 1881-1888.

Tye-Murray N., Lowder M. & Tyler R.S. (1990) Comparison of the F0F2 and F0F1F2 processing strategies for the Cochlear Corporation cochlear implant. *Ear and Hearing* 11, 195-200.

Tykocinski M., Saunders E., Cohen L.T., Treaba C., Briggs R.J.S., Gibson P., Clark G.M. & Cowan R.S.C. (2001) The contour electrode array: safety study and initial patient trials of a new perimodiolar design. *Otology and Neurotology* 22, 33-41.

Tyler R.S., Gantz B.J., McCabe B.F., Lowder M.W., Otto S.R. & Preece J.P. (1985) Audiological results with two single channel cochlear implants. *Annals of Otology, Rhinology and Laryngology* 94, 133-139.

Tyler R.S., Teagle H.F.B., Kelsay D.M.R., Gantz B.J., Woodworth G.G. & Parkinson A.J. (2000) Speech perception by prelingually deaf children after six years of cochlear implant use: effects of age at implantation. *Annals of Otology, Rhinology and Laryngology* 109 (suppl 185), 82-84.

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

Van Tasell D.J. & Yanz J.L. (1987) Speech recognition threshold in noise: effects of hearing loss, frequency response, and speech materials. *Journal of Speech and Hearing Research* 30, 377-386.

Van Tasell D.J., Larsen S.Y. & Fabry D.A. (1988) Effects of an adaptive filter hearing aid on speech recognition in noise by hearing-impaired subjects. *Ear and Hearing* 9, 15-21.

Vance T.J. (1976) An experimental investigation of tone and intonation in Cantonese. *Phonetica* 33, 368-392.

Vance T.J. (1977) Tonal distinctions in Cantonese. *Phonetica* 34, 93-107.

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 and Hearing* 21, 608-624.

Verschuur C.A. (2005) Effect of stimulation rate on speech perception in adult users of the Med-El CIS speech processing strategy. *International Journal of Audiology* 44, 58-63.

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

Wei W.I., Wong R., Hui Y., Au D.K.K., Wong B.Y.K., Ho W.K., Tsang A., Kung P. & Chung E. (2000) Chinese tonal language rehabilitation following cochlear implantation in children. *Acta Oto-Laryngologica* 120, 218-221.

Whalen D.H. & Xu Y. (1992) Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49, 25-47.

Whitford L.A., Seligman P.M., Blamey P.J., McDermott H.J. & Patrick J.F. (1993) Comparison of current speech coding strategies. *Advances in Oto-Rhino-Laryngology* 48, 85-90.

Wilson B., Finley C., Farmer J., Lawson D., Weber B., Wolford R., Kenan P., White M., Merzenich M. & Schindler R. (1988) Comparative studies of speech proceessing strategies for cochlear implants. *Laryngoscope* 98, 1069-1077.

Wilson B.S., Finley C.C. & Lawson D.T. (1990) Models of the electrically stimulated ear. In: *Cochlear implants* (eds Miller J.M. & Spelman F.A.), pp. 339-376. Springer, New York.

Wilson B., Finley C., Lawson D., Wolford R., Eddington D. & Rabinowitz W. (1991) Better speech recognition with cochlear implants. *Nature (London)* 352, 236-238.

Wilson B. (1993) Signal processing. In: *Cochlear implants: Audiological Foundations* (ed Tyler R.S.), pp. 35-85. Singular Publishing, San Diego.

Wilson B. (1997) The future of cochlear implants. *British Journal of Audiology* 31, 205-225.

Wilson B., Finley C., Lawson D. & Zerbi M. (1998) Temporal representations with cochlear implants. *American Journal of Otology* 18 (suppl), S30-S34.

Wilson B.S. (2004) Engineering Design of Cochlear implants. In: *Cochlear Implants: auditory prostheses and electric hearing* (eds Fang-Gang Zeng, Arthur N.Popper & Richard R.Fay), pp. 14-52. Springer-Verlag, New York.

Wu J.L. & Yang H.M. (2003) Speech perception of Mandarin Chinese speaking young children after cochlear implant use: effect of age at implantation. *International Journal of Pediatric Otorhinolaryngology* 67, 247-253.

Xu L., Tsai Y. & Pfingst B.E. (2002) Features of simulation affecting tonal-speech perception: Implications for cochlear prostheses. *Journal of the Acoustical Society of America* 112, 247-258.

Xu L. & Pfingst B.E. (2003) Relative importance of temporal envelope and fine structure in lexical-tone perception (L). *Journal of the Acoustical Society of America* 114, 3024-3027.

Xu L., Li Y., Hao J., Chen X., Xue S.A. & Han D. (2004) Tone production in Mandarin-speaking children with cochlear Implants: a preliminary study. *Acta Oto-Laryngologica* 124, 363-367.

Xu S.A., Dowell R.C. & Clark G.M. (1987) Results for Chinese and English in a multichannel cochlear implant patient. *Annals of Otology, Rhinology, and Laryngology* 97 (suppl 128), 126-127.

Yip M. (1980) *The tonal phonology of Chinese*. PhD Dissertation, MIT. Published 1990, New York: Garland Publishing.

Yip M. (2002) *Tone*. Cambridge University Press, Cambridge.

Zee E. (1978) Duration and intensity as correlates of f0. *Journal of Phonetics* 6, 213-220.

Zeng F.-G. (2002) Temporal pitch in electric hearing. *Hearing Research* 174, 101-106.

Zeng F.-G., Nie K., Stickney G.S., Kong Y.-Y., Vongphoe M. & Bhargave A. (2005) Speech recognition with amplitude and frequency modulations. *PNAS* 102, 2293-2298.

Ziese M., Stutzel A., von Specht H., Begall K., Freigang B., Stroka S. & Nopp P. (2000) Speech understanding with the CIS and the n-of-m strategy in the MED-EL COMBI 40+ system. *ORL Journal for Oto-rhino-laryngology and its Related Specialties* 62, 321-329.

Zwolan T., Kileny P.R., Smith S., Mills D., Koch D. & Osberger M.J. (2001) Adult cochlear implant patient performance with evolving electrode technology. *Otology and Neurotology* 22, 844-849.

Appendix 1: Syllables for tone recognition

Syllable	Tone	Meaning	Chinese Character
/i/	1	one	一
	2	to move	移
	3	chair	椅
	4	meaning	義
/ba/	1	eight	八
	2	to pull	拔
	3	to hold	把
	4	father	爸
/fu/	1	husband	夫
	2	to support with hand	扶
	3	palace	府
	4	to pay	付
/ $tɕ^h$i/	1	seven	七
	2	to ride	騎
	3	to beg	乞
	4	gas	汽

Appendix 2: Sentence lists

Each list comprises 15 sentences, each with three key words (underlined), giving a total of 45 key words per list.

Female-spoken lists: f1~8

f1

	Mandarin	English
1.01	妹妹閉上眼睛	She closed her eyes.(she -> young sister)
1.02	姐姐在看書	She read her book.(she -> elder sister)
1.03	舊衣服很髒	The old clothes were dirty.
1.04	郵差送來一封信	The postman brings a letter.
1.05	牛肉漢堡很好吃	The cheese pie was good.->the beef hamburger was good.
1.06	女孩正在洗頭髮	The girl's washing her hair.
1.07	他們看著火車	They're watching the train
1.08	弟弟的錢掉了	He dropped his money.(he -> young brother)
1.09	郵局在學校附近	The post office was near.-> The post office was near the school.
1.10	太陽把雪融化了	The sun melted the snow
1.11	他們向火車揮手	They're waving at th train.
1.12	掃把放在牆角	The broom stood in the corner
1.13	火車準時到達	The train arrived on time.
1.14	機器蠻吵的	The machine was quite noisy
1.15	老先生很擔心	The old man worries

f2

2.01	姐姐拿起外套	She's taking her coat.(she->elder sister)
2.02	她寫信給哥哥	She writes to her brother
2.03	他們油漆著牆壁	They painted the wall.
2.04	警察找到了小狗	The policeman found a dog.
2.05	泥巴黏在鞋子上	The mud stuck on his shoe
2.06	小孩抓著玩具	The child grabs the toy
2.07	火柴掉到地板上	The match fell on the floor.
2.08	公車突然停住	The bus stopped suddenly
2.09	球打破了窗戶	The ball broke the window
2.10	新毛巾很乾淨	The new towel was clean.
2.11	嬰兒的奶嘴不見了	The baby lost his rattle.
2.12	足球比賽結束了	The football game's over.
2.13	阿姨蠻生氣的	The lady was quite cross.
2.14	小偷帶著一個梯子	The thief brought a ladder.
2.15	輪船慢慢的行駛	The yacht sailed past. -> the ship sailed slowly.

f3

3.01	他們騎著腳踏車	They're cycling along.-> they're riding bicycles.
3.02	牛奶放在門口	The milk was by the front door.-> the milk was put by the door
3.03	叔叔穿著長褲	Men wear long trousers.(men -> uncle)
3.04	女孩的洋娃娃不見了	The girl lost her doll
3.05	廚房的窗戶很乾淨	The kitchen window was clean.
3.06	小男孩睡著了	The small boy was asleep
3.07	工人用掃把掃地	The cleaner used a broom.
3.08	哥哥拿著棍子	He carried a stick.(he -> elder brother)
3.09	老鼠跑到桌子下面	A mouse ran down the hole->A mouse ran down the table
3.1	弟弟拿到了杯子	He reached for a cup.(he ->young brother)
3.11	葉子從樹上落下	The leaves dropped from the trees.
3.12	小狗舔著主人	The puppy licked his master.
3.13	媽媽攪拌著紅茶	The mother stirs the tea
3.14	他們偷了一些蘋果	They're stealing the apples.
3.15	蛋糕店開門了	The cake shop's opening.

f4

4.01	妹妹對著洋娃娃說話	She talked to her doll.(she -> young sister)
4.02	池塘的水很髒	The pond water's dirty
4.03	叔叔用鉛筆畫圖	The man drew with a pencil.(man ->uncle)
4.04	房子有個美麗的花園	The house had a nice garden
4.05	哥哥找到了弟弟	He found his brother.->the elder brother found his young brother)
4.06	他們帶了一些塑膠袋	They carry some shopping bags. -> they carry some plastic bags
4.07	他們跪在地上	They're kneeling down -> they knee on the ground
4.08	她正在等公車	She's waiting for her bus.
4.09	貓跳下了圍牆	A cat jumped off the fence.(fence -> wall)
4.1	媽媽縫著窗簾	Mother made some curtains.
4.11	油漆滴在地板上	The paint dripped on the ground
4.12	小孩吃了一些果醬	The child ate some jam.
4.13	新房子是空的	The new house was empty.
4.14	女孩和嬰兒一起玩	The girl plays with the baby.
4.15	廚房的鐘是錯的	The kitchen clock was wrong.

f5

5.01	毛巾掉到地板上	The towel droped on the floor.
5.02	姐姐背著相機	She brought her camera.(she -> elder sister)
5.03	聰明的女孩在讀書	The clever girls are reading
5.04	他們在喝茶	They're drinking tea.
5.05	他們完成了拼圖	They finished the jigsaw.
5.06	弟弟閉著眼睛	He closed his eyes.(he -> young brother)
5.07	男孩忘了帶書	The boy forgot his book
5.08	警察追著一輛車	The police chased the car
5.09	媽媽切著生日蛋糕	Mother cut the Christmas cake.-> mother cut the birthday cake.
5.1	哥哥帶著雨衣	He's bringing his raincoat.(he -> elder brother)
5.11	阿姨洗著襯衫	The lady washed the shirt.(lady ->aunt)
5.12	罐子放在架子上	The jug stood on the shelf
5.13	郵差按著門鈴	The postman leaned on the fence.-> The postman ring the doorbell.
5.14	植物長在牆上	The plant grows on the wall.
5.15	火柴盒是空的	The match boxes are empty

f6

6.01	弟弟的帽子不見了	He lost his hat.(he -> young brother)
6.02	外套放在椅子上	The coat lies on a chair
6.03	媽媽打開抽屜	Mother opens the drawer.
6.04	這件雨衣非常溼	The raincoat's very wet.
6.05	小嬰兒在睡覺	The little baby sleeps
6.06	姐姐梳著頭髮	She brushed her hair.(she -> elder sister)
6.07	牆上掛著一幅畫	The picture hung on the wall.
6.08	小狗跳上椅子	The dog jumped on the chair.
6.09	小孩向火車揮手	The children wave at the train.
6.1	前面的花園很漂亮	The front garden was pretty.
6.11	廚師烤著蛋糕	The baker iced the cake.->the cook baked the cake
6.12	哥哥割到手指	He cut his finger(he -> elder brother)
6.13	女孩削著鉛筆	The girl sharpened her pencil.
6.14	他們買了一些豆腐	They're shopping for cheese.-> they buy some tofu.
6.15	舊手套很髒	The old gloves are dirty

f7

7.01	姐姐聽著收音機	She's listening to the radio.(she -> elder sister)
7.02	他們全部站起來	They're standing up.-> They all stand up.
7.03	媽媽搖搖頭	The mother shook her head.
7.04	他們準備出門	They're going out.->they're ready to go out.
7.05	阿姨做了一個玩具	The lady's making a toy.(lady -> aunt)
7.06	小孩幫忙工友	The children help the milkman.(milkman -> workman)
7.07	妹妹在浴室唱歌	She sings in the bath. ->younger sister sings in the bathroom
7.08	裙子掛在衣櫥裏	The shirts hang in the cupborard.
7.09	廚師正在做蛋糕	The cook's making a cake
7.1	男孩夾到了手指	The boy hit his thumb.
7.11	水龍頭在水槽上面	The taps are above the sink.
7.12	烤箱太熱了	The oven's too hot.
7.13	弟弟用鏟子挖土	He dug with his spade.(he -> young brother)
7.14	公車停在商店前	The bus stopped at the shops.
7.15	果醬罐是滿的	The jam jar was full

f8

8.01	他們有著快樂的一天	They had a lovely day
8.02	爸爸在洗車	The husband cleaned the car.->father washed the car.
8.03	學校提早放學	The school finished early
8.04	小孩坐在樹下	The children sit under the tree.
8.05	弟弟爬到梯子上	He climbed the ladder.(he -> young brother)
8.06	小狗在籃子裏睡覺	The dog sleeps in a basket.
8.07	媽媽關上窗戶	The mother shut the window.
8.08	男孩有隻玩具恐龍	The boy had a toy dragon.
8.09	阿姨打掃房子	The woman tidied her house.(worman -> aunt)
8.1	檸檬長在樹上	Lemons grow on the trees
8.11	哥哥戴著領帶	He's wearing a tie. (he -> elder brother)
8.12	姐姐照著鏡子	She looked in her mirror.(she -> elder sister)
8.13	抹布蠻濕的	The teacloth's quite wet. (teacloth->cloth)
8.14	繩子太短了	The rope was too short.
8.15	紅蘋果在碗裏	The red apples were in a bowl.

Male-spoken lists: m1~8

m1

1.01	姐姐拿起錢包	She took her purse.
1.02	叔叔去銀行	The man goes to the bank. (man ->uncle)
1.03	天空非常的藍	The sky was very blue.
1.04	雨從天空落下	The rain came down. -> The rain came down from the sky
1.05	弟弟拿到了湯匙	He's reaching for his spoon.(he->young brother)
1.06	他們把傢俱搬走了	They moved the furniture.
1.07	他們買了門票	They bought some tikets.
1.08	小狗玩著一顆球	The puppy plays with a ball.
1.09	房子有九個房間	The house had nine rooms.
1.1	水果放在盒子裏	The fruit came in a box. -> the fruit was put in a box
1.11	排骨湯煮好了	The apple pie's cooking. ->the rib soup's cooked
1.12	女孩拿著鏡子	The girl held a mirror.
1.13	一些好心的人走過來	Some nice people are coming.
1.14	男孩從窗口摔下來	A boy fell from the window
1.15	玻璃碗破了	The glass bowl broke

m2

2.01	車子撞上牆壁	The car hit a wall
2.02	妹妹用杯子喝水	She drinks from her cup. (she -> young sister)
2.03	小孩走路回家	The children are walking home
2.04	他們吃完晚餐	They finished the dinner.
2.05	男孩上床睡覺	The boy got into bed.->the boy went to bed
2.06	哥哥穿著黃色襯衫	He wore his yellow shirt. -> elder brother wore his yellow shirt.
2.07	卡車爬上山坡	The lorry climbed the hill.
2.08	弟弟在地板上玩	He slid on the floor. ->young brother play on the floor
2.09	小狗追著小貓	The dog chased the cat.
2.1	他們拿了一些馬鈴著	They wanted some potatoes -> they take some potatoes
2.11	媽媽拿了一個鍋子	Mother fetches a saucepan
2.12	衣服被釘子勾住了	The shirt caught on a nail.(shirt->clothes)
2.13	湯非常熱	The fire was very hot. ->the soup was very hot
2.14	老鼠找到乳酪	The mouse found the cheese.
2.15	工人油漆著大門	The man painted the gate. ->The workman painted the door.

m3

3.01	妹妹用湯匙吃飯	She used her spoon. (she -> young sister)
3.02	媽媽在看報紙	The mother reads a paper.
3.03	書放在書架上	The book sits on the shelf.
3.04	洗澡水是溫的	The bath water was warm.
3.05	姐姐找到了錢包	She found her purse. (she -> elder sister)
3.06	火車停在火車站	The train stops tha the station.
3.07	貓咪坐在床上	A cat sits on the bed.
3.08	橘子蠻甜的	The orange was quite sweet
3.09	兩個農夫在說話	The two farmers are talking.
3.1	他們油漆著天花板	They painted the ceiling.
3.11	水壺非常燙	The kettle's quite hot
3.12	他們吃了檸檬果凍	They ate the lemon jelly
3.13	郵差關上大門	The postman shut the gate.
3.14	司機發動引擎	The driver starts the engine.
3.15	蘋果派是熱的	The apple pie was hot.

m4

4.01	草莓果醬很甜	The strawberry jam was sweet.
4.02	小孩丟下書包	The children dropped the bag.
4.03	他們有兩個空瓶子	They had two empty bottles.
4.04	哥哥在洗車	He's cleaning his car. (elder brother's washing his car)
4.05	水果放在地板上	The fruit lies on the ground
4.06	姐姐縫著紐扣	She's sewing on a button. (she -> elder sister)
4.07	阿姨穿著外套	The lady wore a coat.(lady -> aunt)
4.08	魚放在盤子裏	A fish lay on the plate
4.09	袋子掉到地板上	The bag bumps on the ground.
4.1	男孩從樓梯跌下來	The boy slipped on the stairs.
4.11	他們走過草地	They walked across the grass.
4.12	糖非常甜	Sugar's very sweet.
4.13	貓咪跳上桌子	The cat jumped onto the table.
4.14	外套掛在衣櫥裏	The coat hang in a cupboard.
4.15	弟弟吸著姆指	He's sucking his thumb.(he -> young brother)

m5

5.01	小孩喜歡草莓	Children like strawberries.
5.02	爸爸把禮物藏起來	Father's hiding the presents.
5.03	嬰兒打破了杯子	Baby broke his mug
5.04	她和妹妹在吵架	She argued with her sister.
5.05	牛奶放在瓶子裏	The milk came in a bottle. -> The milk was saved in a bottle.
5.06	叔叔在整理花園	The man dug his garden. -> uncle cleaned his garden
5.07	杯子放在盤子上	The cup stood on a saucer.
5.08	草莓冰淇淋是粉紅色的	The ice cream was pink. -> The strawberry ice cream is pink.
5.09	他們在吃果醬麵包	They had some jam pudding. -> they had some jam bread.
5.1	洗澡的毛巾是濕的	The bath towel was wet
5.11	車子發動著引擎	The car engine's running.
5.12	阿姨有件羊毛外套	The lady has a fur coat. (lady -> aunt) (fur -> wool)
5.13	植物掛在門上	The plant hangs above the door.
5.14	他們正在爬樹	They're climbing the tree
5.15	新鞋子很緊	The new shoes were tighty.

m6

6.01	阿姨坐在椅子上	The lady sat on her chair. (lady -> aunt)
6.02	弟弟玩著火車	He played with his train. (he -> young brother)
6.03	馬鈴薯長在地上	Potatoes grow in the ground.
6.04	廚師切著洋蔥	The cook cut some onions
6.05	地板看起來很乾淨	The floor looked clean
6.06	他們在公園裡玩	They're playing in the park.
6.07	這個袋子非常重	The bag was very heavy.
6.08	哥哥正在擦桌子	He's wiping the table. (he -> young brother)
6.09	綠色的番茄很小	The green tomatoes are small.
6.1	糖果店是空的	The sweet shop was empty.
6.11	房子有個美麗的花園	The house had a lovely garden.
6.12	她敲敲窗戶	She tapped at the window.
6.13	他們搬起箱子	They're lifting the box.
6.14	嬰兒抱著奶瓶	The baby wants his bottle. (wants -> holds)
6.15	貓咪玩著毛線	The cat played with some wool.

m7

7.01	妹妹的手受傷了	She hurt her hand. (she->young sister)
7.02	小孩喝了一些牛奶	The child drank some milk
7.03	叔叔和警察說話	A man told the police.
7.04	這雙鞋子非常髒	The shoes were very dirty
7.05	火車快速行駛	The train's moving fast
7.06	他們採了一些草莓	They picked some raspberries. (raspberries->strawberries)
7.07	阿姨的手臂受傷了	The lady hurt her arm.
7.08	毛巾掉到地毯上	The towel dripped on the carpet.
7.09	卡車載著水果	The lorry carried fruit.
7.1	這張桌子有三隻腳	The table has three legs
7.11	司機按著喇叭	The driver hooted his horn.
7.12	工友送來牛奶	The milkman carried the cream.-> a workman carried the milk
7.13	他們敲敲窗戶	They knocked on the window.
7.14	哥哥繫好鞋帶	He tied his shoelaces. (he -> elder brother)
7.15	男孩跑得很遠	The boy's running away

m8

8.01	有人拿走了錢	Somebody took the money
8.02	叔叔打開水龍頭	A man's turning the tap.
8.03	爸爸寫了一封信	The father writes a letter.
8.04	哥哥正在洗臉	He's washing his face (he -> elder brother)
8.05	姐姐站在窗戶邊	She stood near the window. (she -> elder sister)
8.06	他們看著時鐘	They're looking at the clock.
8.07	火柴放在架子上	The matches lie on the shelf
8.08	他們買了一些麵包	They're buying some bread.
8.09	窗簾太短了	The curtains were too short.
8.1	廚房的水槽是空的	The kitchen sink's empty
8.11	樹上的葉子掉光了	The tree lost its leaves.
8.12	小狗玩著樹枝	The dog played with a stick
8.13	公車提早開走了	The bus went early
8.14	弟弟撞到了頭	He hit his head. (he -> elder brother)
8.15	五個工人在工作	The five men are working (men -> workmen)