Exploitation of Effective Temporal Cues for Lexical Tone Recognition of Chinese

YUAN, Meng

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

March 2009

To my loving family

Acknowledgements

I owe enormous gratitude to my thesis advisor, Prof. Tan LEE. He guided me since I was a novice in the field of speech processing for hearing prosthesis and gave me insightful advices throughout this research. His help, understanding and patience played an important role in increasing my confidence and improving my technical writing and presentation skills. He also supported me to attend academic conferences.

I would like to express my deepest gratitude to my co-supervisor, Dr. Sigfrid D. SOLI, for his invaluable advices and suggestions. My discussions with him have been always constructive and enjoyable. He has always been a great resource for ideas and solutions, and his encouragement made me less frustrated when I was in the face of the difficulties for thesis exploration.

I also would like to sincerely thank Prof. Pak-Chung CHING, Prof. William Shi-Yuan WANG and Prof. Ken Ma for their precious suggestions. Thanks are due to all my colleagues and friends in DSP and speech technology groups. They have helped me in many different ways. I also would like to express my special regards to Dr. Kevin C. P. Yuen for his great help and close collaboration along the past four years since I started to work on hearing-related researches.

Finally, I am deeply grateful to my wife and parents for their love and continuous support.

Abstract of thesis entitled:

Exploitation of Effective Temporal Cues for Lexical Tone Recognition of Chinese

Submitted by **YUAN Meng**

for the degree of **Doctor of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong** in

March 2009.

Lexical tone plays an important role in tonal languages. Acoustically, pitch is determined by the periodicity of speech, which is measured as the fundamental frequency ($F0$) of acoustic signals. In each tonal language, there are a certain number of lexical tones that are described by distinctive pitch contours. Cantonese and Mandarin have four and six tones, respectively.

People with sensorineural hearing loss have difficulty in utilizing spectral information for speech recognition and rely heavily on temporal information. The temporal information of speech is divided into three parts, based on the rate of amplitude fluctuation: temporal envelope (below 50 Hz), periodicity (50 – 500 Hz), and fine structure (above 500 Hz).

The goals of this thesis are to investigate what are the effective temporal cues for lexical tone perception of Chinese and how to manipulate or enhance these cues for better performance of tone perception. We adopt the research method of acoustic simulation with normal-hearing subjects. A four-channel noise-excited vocoder is used to generate test stimuli for tone identification.

We compare the contributions of temporal envelope and periodicity components (TEPCs) from different frequency regions to tone recognition in Cantonese and Mandarin. It is observed that TEPCs from high-frequency region (1 - 4 kHz) are more important than those from low-frequency region ($<$ 1 kHz). In noise condition, tone recognition performance with temporal cues degrades and more spectral information is needed.

Previous studies show that hearing-impaired people have difficulties in perceiving tones, even though they are aided with cochlear implants (CIs). In this thesis, two approaches are investigated to improve Chinese tone recognition. In the first approach, TEPCs go through a process of non-linear expansion in order to increase the modulation depth of periodicity-related amplitude fluctuation. Results of listening tests show that TEPC expansion leads to a noticeable improvement on tone identification accuracy. In the second approach, the effectiveness of enhancing temporal periodicity cues in noise is investigated. Temporal periodicity cues are simplified into a sinusoidal wave with frequency equivalent to the $F0$ of speech. This leads to a consistent and significant improvement on tone identification performance at different noise levels. This part of research is expected to be helpful in designing CI processing strategy for effective speech perception of tonal languages.

摘要

聲調在有聲調語言中有十分重要的。從聲學角度講，音調 (pitch)是由聲音周期決定的， 而聲音周期通常是通過測量聲音信號的基頻得到的。對於每一種有聲調語言，不同聲調的數量是固定的，不同的聲調是靠音調的變化描述的。廣東話有六個聲調，而普通話則有四個不同的聲調。

對於有聽覺神經損傷的人來説，充分利用聲音的頻譜信息是困難的， 因此，他們會更多的依賴瞬時信息。根據不同的振幅變化速度，瞬時特性可以被分爲以下三部分：瞬時包絡 (temporal envelope) (50 Hz 以下)， 瞬時周期 (periodicity) (50-500 Hz),和瞬時精細結構 (~~fine~~ structure) (500 Hz 以上)。

本文的目的是要探討哪些瞬時特性是對中文聲調識別特別有效的， 並且，本文也將探討是否可以通過對以上有效的瞬時信號的處理來增強聲調識別能力。我們採用了一個聲學仿真的模型來做研究。以正常聽力人士為測試對象。一個四通道噪聲激勵的語音編解碼器被用來模擬真實人工耳蝸中的聲音處理過程。通過此編解碼器，我們可以得到語音測試資料來進行聲調辨識。

我們比較了來自不同頻段的瞬時包絡及周期特性 （TEPC）對廣東話及普通話聲調識別的貢獻。一致的結果是，高頻段的 TEPC 信息對於聲調的識別更重要。在噪聲情況下，利用瞬時特性來進行語音識別的能力明顯下降，更多的頻譜信息可以提供一定的幫助。

研究表明，對於聽力障礙人士來講，即使是在使用人工耳蝸的情況下，他們仍然無法聽到聲調。基於此，本文探討了兩种試圖增強中文聲調識別能力的方法。在第一种嘗試中，我們利用一種非綫性放大的方法對 TEPC 進行處理，以期增加與周期相關的調制深度。實驗結果顯示，這種對 TEPC 的放大方法可以使聲調辨別率明顯增加。在第二种嘗試中，我們試圖強化帶噪語音的瞬時週期性。我們用一個與語音基頻信號相關的正弦信號對原始的瞬時周期信號進行簡化。實驗結果表明，此种方法可有效提高不同噪聲程度下的聲調辨別能力。這部分研究的結果對於今後開發適用于人工耳蝸的針對有聲調語言的聲音處理策略有一定的作用。

Declaration

This is to certify that

- the thesis comprises only my original work towards the PhD except where indicated in the Preface below
- the thesis is less than 100,000 words in length

Preface

The work on the design of Cantonese disyllabic word list in Chapter 2 was carried out in collaboration with Kevin C. P. Yuen and Janet Pang. Kevin designed the structure of the disyllabic word list and Janet helped selecting the appropriate words. I was responsible for the processing of the recorded speech signals, including re-sampling, energy equalization, and stimuli preparation with the noise-excited vocoder.

Contents

List of Tables

List of Figures

Abbreviations

The following acronyms have been used throughout this thesis

CI	Cochlear Implant
CIS	Continuous Interleaved Sampling
ACE	Advanced Combination Encoding
TEPC	Temporal Envelope and Periodicity Component
TEC	Temporal Envelope Component
TPC	Temporal Periodicity Component
TFSC	Temporal Fine Structure Component
AM	Amplitude Modulation
RMS	Root Mean Square
AGC	Automatic Gain Control
NH	Normal Hearing
HI	Hearing Impaired
SNR	Signal-to-Noise Ratio
IIR	Infinite Impulse Response
AFC	Alternative-Forced-Choice
ANOVA	Analysis of Variance
HSD	Honest Significant Difference
MEM	Multi-channel Envelope Modulation
CHINT	Cantonese Hearing in Noise Test
HL	Hearing Level

DWT	Discrete Wavelet Transform
LPF	Low-pass Filter
HPF	High-pass Filter
BPF	Band-pass Filter
ATOPEX	Application for TOne Perception EXperiments
GUI	Graphical User Interface
CANDILET-N	Computerized CANtonese DIsyllabic LExical Tone Identification Test in Noise
BTE	Behind-The-Ear
ASR	Automatic Speech Recognition
dbHL	Decibels Hearing Level

Chapter 1

Introduction

Summary

This chapter provides the background and motivation of this thesis. We start with describing the mechanism of sound perception and discussion about hearing loss and hearing prostheses. Then we focus on the use of temporal cues in one of the commonly used type of hearing prostheses, namely cochlear implant. We review the previous studies on investigating the temporal information on speech perception of tonal languages. Accordingly, we establish the motivation of our research, which is exploring possible signal processing techniques for improving tone perception. The chapter concludes with an overview of the major research questions of this thesis.

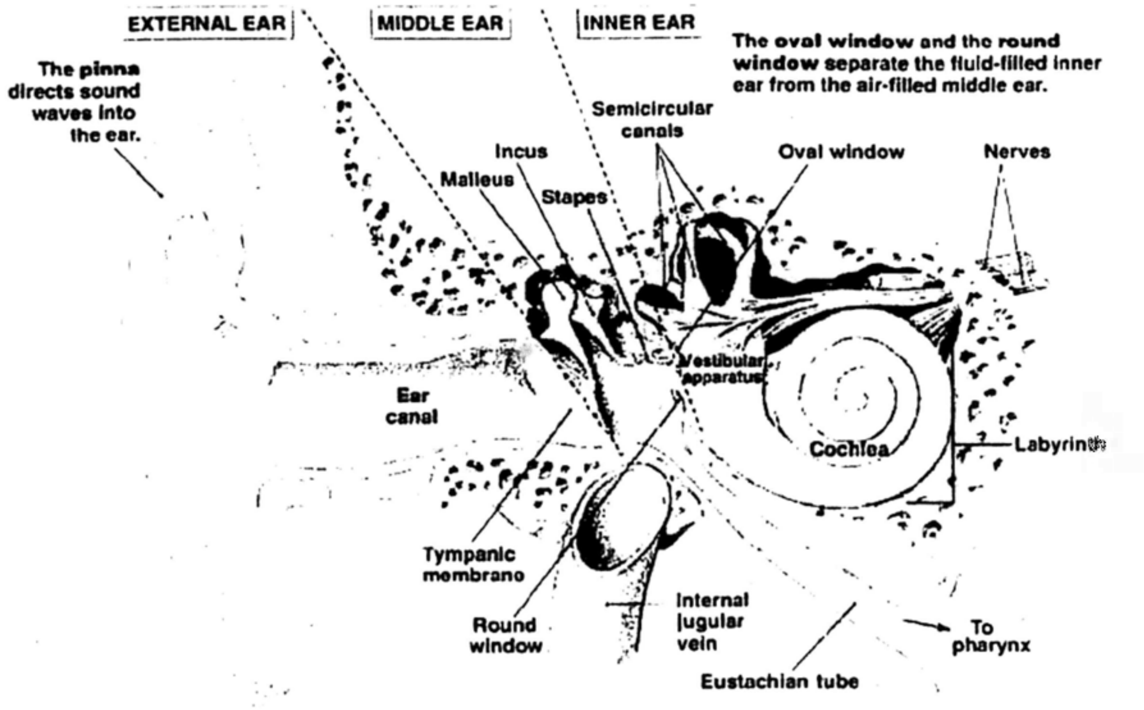
1.1 Hearing

1.1.1 Sound Perception

Sounds reaching the ears are made by mechanical vibration of air particles and perceived by the sense of hearing [Shambaugh, 1930]. Amplitude of the sound wave determines the loudness and its frequency determines the pitch. Perceivable sound is within a certain range of frequencies and intensities. A young person with normal hearing can hear sounds with frequencies between 20 Hz and 20,000 Hz [Cutnell, 1998]. The intensity range is not the same at different frequencies. For frequencies between 1 kHz and 6 kHz, which is the most sensitive range to human hearing, the sound is audible over a range of about 120 dB, which corresponds to the intensity range of about 1,000 billion to one.

Sounds travel through the air as pressure waves and are captured by the outer ear. The ear transforms this pressure wave into a neural code which is interpreted by our brain. As shown in Figure 1.1, our ear consists of three parts: the external/outer ear, the middle ear and the inner ear. The outer ear is composed of the pinna, the ear canal and the eardrum. The outer ear works like a horn and amplifies the sound wave by approximately 12 – 15 dB in the frequency region around 2.5 kHz [Moore, 1998]. The major function of the middle ear is to ensure the efficient transfer of sound energy from the air to the fluids in the cochlea and counter-balance the difference in impedance between the air in the outer ear and the fluid in the inner ear. The outer ear and the middle ear together behave essentially like a linear system for moderate sound levels [Moore et al., 1997].

The major function of the inner ear is to convert the sound into neural activities that are transmitted to the brain. In mammals, the auditory portion of the inner ear is a coiled structure, called 'cochlea'. The region nearest the oval window is the base of the cochlea; the other end, or top, is referred to as the apex. The cochlea is filled with incompressible fluids, and it has bony rigid walls. The principal elements for converting sounds into neural activities are



The pinna directs sound waves into the ear.

The oval window and the round window separate the fluid-filled inner ear from the air-filled middle ear.

Figure 1.1: Diagram of outer, middle and inner ear in human. Redrawn from [Denes and Pinson, 1973].

found on the basilar membrane (BM). Basilar membrane is a flexible structure that separates two liquid-filled tubes that run along the coil of the cochlea. The organ of Corti is located on the top of BM. It comprises the auditory sensory cells, or 'hair cells'. The hair cells are connected at their bases to the nerve fibers of the auditory nerve. There is an important mechanism of the cochlea called 'tonotopy'. Different frequencies are separated from high to low along the cochlea. Therefore, the cochlea works as a bank of filters. The center frequencies of the filters are spaced logarithmically from apex to base [Greenwood, 1990] and the bandwidths of the filters also increase with the distance along the BM from low to high frequency region [Glasberg and Moore, 1990]. The electrical activity induced in the hair cells stimulates the fibers of the auditory nerve. The auditory nerve reacts to the neurotransmitter by producing neural spikes (electrochemical pulses) that are sent to the brain along the auditory nerve. The higher the intensity of the sound, the more neural spikes being sent to the brain. The brain acts as a central auditory processor in interpreting complex sounds such as speech.

1.1.2 Hearing Loss and Hearing Prostheses

Hearing loss is regarded as one of the most common disorders in human beings. It is defined as a reduction in an individual's ability to hear sounds. Loss of hearing can affect a person of any age. Nevertheless it is accepted that hearing loss tends to occur gradually as a person gets older. In Hong Kong, about 165 infants were born with severe hearing impairment every year at an annual birth rate of 55,000 [Lam, 2003]. The prevalence of hearing loss for elder people aged above 65 years was 19 - 25 % in Hong Kong [Lee et al., 2002c; Ho and J.Woo, 1994] and 27.4% in United States [Adams and Hardy, 1989]. These people found it difficult to communicate with others in daily life and felt frustrated in the society.

There are three kinds of hearing loss, conductive, neural/ sensorineural, and the combination of these two. Conductive hearing loss is caused by the malfunction of the middle ear system. It can be caused by a number of factors. For example, fluids accumulated in the middle ear, which is caused by flu, ear infections, allergies and perforated eardrums, are possible factors. Sensorineural hearing loss is caused by the malfunction of the cochlea, or the auditory system. It can be caused by drugs that may cause injury to the hearing system, diseases, head trauma, aging and genetic syndromes.

Doctors and scientists have engaged in finding ways to restore normal hearing. In the early years, scientists tried to amplify sounds to make them audible for hearing impaired. The first hearing aid was a large horn-shaped device, which was developed two hundred years ago. It does not use electricity and simply amplifies the sound in acoustic domain.

The advent of electricity and the technology of transforming acoustical sound into electrical signals, e.g., telephone, rapidly changed the hearing aid technology. More powerful and smaller devices became possible for portable use. Digital hearing aids appeared in mid-1990s. User-defined programs could be implemented and adjusted to meet different needs of individual patients [Popelka and

Engebretson, 1983]. Advanced signal processing techniques have been adopted to reduce the effects of background noise, reverberation and acoustic feedback. And the performance of hearing aids has been substantially improved [Lim, 1983; Preves et al., 1986].

Along with the development of conventional hearing aids, other types of hearing prostheses were also invented. Cochlear implants (CI) are widely used nowadays. A CI device delivers electrical stimulations directly to auditory nerves in the cochlea. It is suitable for patients with moderate to profound hearing loss, who might not benefit from a hearing aid.

A CI consists of an internal part and an external part. The internal part containing a receiver and a series of electrodes is implanted inside the patient's head by surgical operation. The external part is worn by the patient. It consists of a microphone, a sound processor, and a coil transmitter. Figure 1.2 shows a body-worn CI system. The sound processor is located in the box worn by the patient. The acoustical signal is picked up by the microphone and sent to the sound processor via a wire. The sound processor converts the sound from analog into digital and calculates the current amplitudes based upon the incoming sound. The signal travels back to the headpiece that contains a coil transmitting coded radio-frequency signal across the skin. The implanted circuits serve as a decoder to decode the signals transmitted from the outlier coil, convert them into electrical currents, and send them to the cochlea via the wires. The electrode array at the end of the wire stimulates the auditory nerve and activates the central nervous system.

The sound processor is the "brain" of a CI system. It determines the sound features that are transmitted to the electrodes. Throughout the 40 years history of CI, many different speech processing strategies have been proposed [Wilson, 2000]. In most existing commercial products, the Continuous Interleaved Sampling (CIS) strategy has become a standard. It was first proposed by Wilson and his colleagues [Wilson et al., 1991]. Detailed information about CIS will be given in Chapter 3.

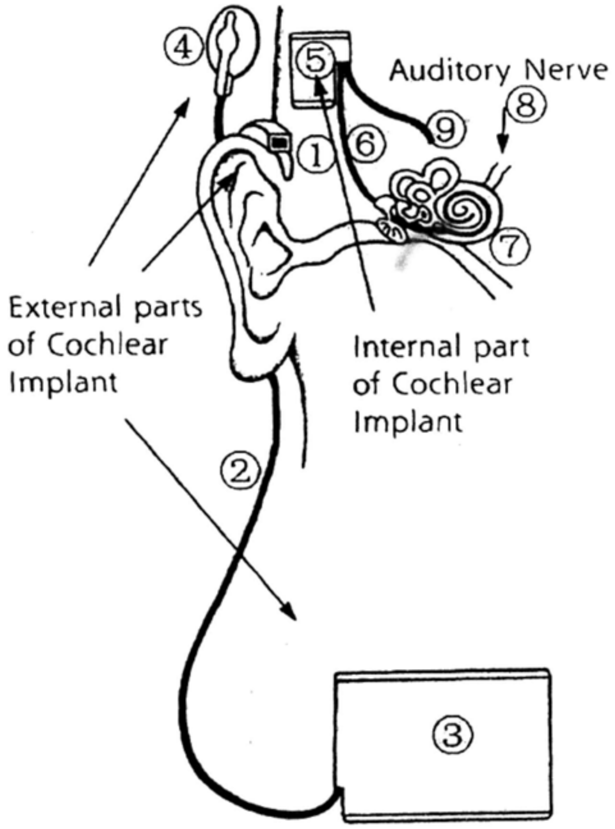


Figure 1.2: The body-worn CI system (Med-El Combi-40+). It contains: ① a microphone, ② a wire, ③ a speech processor, ④ a headpiece, ⑤ a magnet, ⑥ a wire to the implant, ⑦ an electrode array, and ⑨ another wire; ⑧ the auditory nerve.

1.2 Temporal Cues and Spectral Cues

1.2.1 Definitions

Speech signals can be described in time domain or frequency domain. Time-domain analysis concerns temporal amplitude variation of the signal, while frequency-domain analysis describes how signals change and how fast they change. In frequency domain, the signal spectrum can be decomposed into spectral envelope and spectral fine structure. Spectral envelope is defined as the general shape of the spectrum, a smooth curve that passes though the peaks of the spectrum [Hartmann, 1997]. For human speech, the spectral envelope represents the properties of the vocal tract, i.e., formant structure. The spectral fine structure, or the spectral details, refer to the detailed frequency components in the power spectrum. The spectral fine structure represents the vocal source information, which has a harmonic structure for voiced speech and noise nature for unvoiced speech. In time domain, Rosen proposed to divide the temporal information into three categories, depending on the rate of amplitude fluctuation: (1) temporal envelope (2–50 Hz) component (TEC); (2) temporal periodicity (50–500 Hz) component (TPC); and (3) temporal fine structure (500–10000 Hz) component (TFSC) [Rosen, 1992]. In other studies, the temporal cues may be roughly divided into two categories: the slow-varying cue – the temporal envelope and periodicity (2–500 Hz) component (TEPC); and contrarily, the fast-varying cue – the temporal fine structure (500–10000 Hz) component (TFSC) [Kong and Zeng, 2006].

TEPC and TFSC carry different linguistic contrast functions of speech. TEPCs are mostly responsible for carrying the contrasts in the prosodic domain including tempo, rhythm, syllabicity, stress and intonation, and contrasts for the manner of articulation and voicing in the segmental domain; whereas TFSCs are responsible for carrying contrasts in the segmental domain mainly on the place of articulation and voice quality. A detailed discussion on the framework for temporal information in speech can be found in [Rosen, 1992].

1.2.2 Extraction of Temporal Cues

There are two major methods of extracting a signal's temporal envelope. The first method, which has been widely applied in the CIS strategy, is the rectification and low-pass filtering method. In this method, the speech signal is first full-wave or half-wave rectified. Then a low-pass filter is used to limit the variation rate of the envelope. The typical cutoff frequency is set to 200 Hz or 400 Hz in CIS. Simulation studies demonstrated no significant effect of the envelope cutoff frequency on speech recognition in English by normal-hearing (NH) listeners [Shannon et al., 1995]. Conversely, changing the cutoff frequency from 50 to 500 Hz had a significant effect on Chinese tone recognition [Fu et al., 1998b; Xu et al., 2002]. This indicates that periodicity-related information are important to tone perception for tonal languages.

The other envelope detection method is based on the Hilbert transform [Hilbert, 1912]. Hilbert transform is a mathematical tool that represents a signal as the product of a slowly-varying envelope and a "carrier" signal, that contains fine structure of the waveform. Detailed mathematical presentation of Hilbert transform will be given in Appendix.

Psychophysical studies on CI users showed that the methods with full-wave rectification and Hilbert transform achieved significantly better speech recognition performance than with half-wave rectification on CI listeners [Nie et al., 2006]. In this thesis, we adopt the full-wave rectification and low-pass filtering method for easy implementation.

1.3 Temporal Cues for Speech Perception

The importance of temporal cues to speech recognition of non-tonal and tonal languages has been investigated extensively over the past twenty years. Smith and his colleagues constructed a set of acoustic stimuli, called "auditory chimeras" to investigate the relative importance of temporal envelope and temporal fine structure to speech recognition [Smith et al., 2002]. Each stimulus contains the envelope of one sound and the fine structure of another sound.

For example, a speech-speech chimera is synthesized such that it contains the information from one sentence in the envelope and the information from another sentence in the fine structure. Smith et al. found that temporal envelopes played an important role in speech-speech chimeras. The result is consistent with the observation on English-speaking CI users reported in Wilson et al. [1991]. In Wilson's study, CI users achieved high speech recognition performance using CIS with a limited number of frequency bands (4 to 6) where the temporal fine structure is absent. On the other hand, for melody-melody chimeras the subjects mainly perceived the melody based on the temporal fine structure. A follow-up study on the relative importance of temporal envelope and fine structure cues to lexical tone perception was carried out by means of the chimera approach [Xu and Pfingst, 2003]. It was found that when only limited spectral information is available the temporal fine structure information is more important to Mandarin tone classification than the temporal envelope information. This result indicated that the relative contributions of temporal envelope and fine structure for tone perception have a similar pattern to that for melody recognition. In Liu and Zeng [2006], speech-speech chimeras were synthesized. English speech intelligibility for NH listeners was measured over a wide range of signal-to-noise ratios (SNRs). The results showed that temporal envelope information contributed more in clean condition and at high signal-to-noise ratios, whereas temporal fine structure cues contributed more at low signal-to-noise ratios.

Shannon [2002] reviewed the speech perception research with CI users and NH subjects for English. They found that temporal cues up to 20 Hz fluctuation rate were useful for speech recognition. Temporal information above 20 Hz is related more to speech quality, and contributes little to speech recognition. Four frequency channels can produce good speech understanding, but more channels are required for difficult listening situations, e.g., in background noise or with competing speakers. Kong and Zeng systemically evaluated the contributions of temporal cues (envelope and fine structure) and spectral cues (envelope and fine structure) to Mandarin tone recognition. They found that NH listeners with acoustic CI simulation achieved nearly perfect tone recognition performance with either spectral or temporal fine structure in quiet, but only 70% – 80% correct with envelope cues. 32 channels were required to achieve a comparable performance to that obtained with the original stimuli, but only four bands were necessary if additional temporal fine structure was provided. When no temporal fine structure was available, the tone recognition accuracy could reach about 80% with only one band. The study concluded that tone recognition is a robust process that makes use of both spectral and temporal cues. Different from English, they found that the inclusion of temporal periodicity cues below 500 Hz produced significantly better tone recognition performance than the temporal envelope cues below 50 Hz. Xu and Pfingst [2008] summarized the relative contributions of temporal and spectral cues for phoneme recognition and lexical tone recognition. They showed that consonant and vowel recognition in quiet reached plateau at 8 and 12 channels, with low-pass cutoff frequencies of 16 Hz and 4 Hz, respectively. For Mandarin tone recognition, the performance was remarkably poorer than that for English phoneme recognition under comparable conditions. Higher temporal envelope cutoff frequency and more frequency bands were required [Fu et al., 1998b; Xu and Pfingst, 2008].

1.4 Outline of Thesis

1.4.1 Major Research Questions

This thesis is focussed on tone recognition in tonal languages, namely Cantonese and Mandarin. The research questions being addressed are two-fold: fundamental psychophysical principles and application-oriented signal processing issues. The psychophysical principles concern the contributions of temporal and spectral cues on tone perception. More specifically, our goal is to reveal (i) the importance of temporal cues to tone perception; (ii) the contribution of temporal cues from different frequency regions; (iii) the importance of temporal cues to different tonal languages, i.e. Cantonese and Mandarin; (iv) the effect of spectral details on tone perception in noise. A practical problem encountered by CI patients is that the tone-related information has not been sufficiently delivered in the state-of-the-art CIs. The other goal of this thesis is to explore possible signal processing techniques for improving tone perception. The design of these signal processing algorithms is based upon the findings of the psychophysical studies.

1.4.2 Chapter by Chapter Overview

This thesis is divided into chapters based upon different studies on the above stated research questions.

Chapter 2 discusses the fundamentals of pitch and tone perception. The pitch perception mechanisms explain how pitch is perceived in the human hearing system. The importance of lexical tone perception in tonal languages is discussed. The relation between pitch perception and tone perception is shown.

Chapter 3 describes the design of tone perception test. It involves the design of speech materials, the basic signal processing methodology, and the design of the research platform used for the test.

Chapter 4 investigates the effects of temporal and spectral cues on lexical tone identification. Fundamental psychophysical problems on the contribution of TEPCs from different frequency regions and the effect of spectral resolution on tone perception were investigated.

In Chapter 5, a non-linear temporal envelope expansion scheme was applied aiming to improve the tone perception in Cantonese. The enhancement method expands the modulation depth of the temporal envelope such that the periodic fluctuation which is related to $F0$ becomes more salient to detect. Experimental results showed that (i) expansion of TEPC from high-frequency band leads to noticeable improvement on tone identification accuracy; (ii) the effectiveness of TEPC expansion is more significant for female voice than male voice.

In Chapter 6 the effectiveness of enhancing temporal periodicity cues for Cantonese tone perception was investigated. The $F0$s of the voiced speech were explicitly encoded in the temporal envelope by a simplified periodic pattern, i.e. sinusoidal wave. The periodicity-related modulation depth in the temporal envelope was increased to make the periodicity more salient. Experimental results showed that the periodicity-enhanced speech processing strategy led to more accurate tone identification than the standard strategy, especially for noisy speech. Enhancing TEPC in high-frequency region (> 1 kHz) is more effective than in low-frequency region. The results are useful for the design of CI speech processing strategies to improve speech recognition of tonal languages.

Finally, Chapter 7 provides the overall conclusions of the thesis as well as some suggestions for further research.

□ **End of chapter.**

Chapter 2

Lexical Tone Perception

Summary

Tone perception is a very important part of speech perception, especially for tonal languages. In this chapter, we discuss the tone perception mechanism physiologically and linguistically. The chapter starts with an introduction to the pitch perception mechanisms. The relation between pitch and tone perception is also discussed. Tone patterns of Chinese dialects, i.e., Cantonese and Mandarin, are illustrated. Previous studies on investigating the tone perception on CI patients are reviewed.

2.1 Pitch

The American Standards Association defined pitch as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" [ASA, 1960]. An important aspect of this definition is the term 'sensation'. It implies that pitch is a subjective attribute of sound based on what is heard [Moore, 1998]. It also means that variation in pitch gives rise to a sense of melody [Plack and Oxenham, 2005]. The word 'pitch' does not refer to a *physical* attribute of a sound. The more recent American National Standards defined pitch, without referring to music, as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends primarily on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus" [ANSI, 1994]. This is a fairly broad definition, pointing out that "low" and "high" are to be associated with pitch/frequency. In this thesis, we followed the more recent ANSI definition to define pitch. An advantage of this definition is that pitch is not limited to be associated with music. It provides a more general view on pitch.

Pitch is an important attribute of sound. In fact, many sounds in our environment have acoustic waveforms that repeat over time. These sounds are often perceived as having a pitch. Musical instruments produce a pitch such that melodies and chords are generated. Vowel sounds in speech are "voiced" and can be associated with a pitch.

Pitch is known as the most relevant perceptual dimension in most forms of music. It is the basis for musical concepts such as key, consonance, chords, harmony, contour, etc. The melody of a song is carried by the variations in pitch of the musical notes. When these pitch variations are dropped, only the rhythm remains and the music becomes rather dull. Thus, pitch is crucial for music perception. Pitch is also important for speech communication. Although pitch is not essential for speech understanding of western languages, i.e. English speech can be understood when the pitch contour is changed, pitch does carry suprasegmental information. Pitch and pitch changes provide information about intonation (whether a sentence is a statement or a question), the emotion of the speaker, the identity of the speaker, or the dialect. Pitch can also be used to highlight a focussed word in a sentence or put stress on a specific part of a sentence. In tonal languages like Mandarin and Cantonese, pitch carries semantic information. In these languages, pitch pattern determines the meaning or the grammatical function of words. Words pronounced with different pitch contours have different meanings. The movement of pitch over the voiced portion of one or two syllables corresponds to a lexical tone.

Another important aspect of pitch is that it improves segregating sound sources by detecting the differences of the sources in fundamental frequency ($F0$). This allows NH subjects to focus attention on one voice and better understand speech in very noisy/competitive environments (e.g., cocktail party).

As pitch is a subjective attribute, it can not be directly measured. In order to investigate pitch-related subjects, we need subjective human reports about their sensations, which can possibly be biased or ambiguous. For example, the pitch of a sound is assessed by adjusting the frequency of the sinusoid until the perceived pitch of the sinusoid matches the perceived pitch of the sound in question. The frequency of the sinusoid gives a measure of pitch. Sometimes, a periodic complex sound, such as a pulse train, is used as a matching stimulus. In this case, the repetition rate of the pulse train gives a measure of pitch. Thus in acoustic terms, pitch corresponds to the frequency of a pure tone (a sinusoid), or to the $F0$ of a periodic complex tone. It is related to the physical repetition rate of the waveform of a sound.

2.2 Pitch Perception

Pitch perception is not a purely academic research topic. As more is known about how human auditory systems process pitch, we will be able to develop a variety of useful applications. For example, we may improve the performance of automatic speech recognition (ASR) systems such that the systems are more robust to interfering sounds. State-of-the-art ASR systems are unable to distinguish different acoustic sources, while humans do not have much difficulty. Understanding pitch perception mechanisms will also help us in designing new signal processing strategies for hearing prostheses to deliver the most relevant information which is available to the listeners.

It has been agreed that the perception of pitch involves both the place mechanism and the temporal mechanism. Figure 2.1 shows a simulation of the response on the basilar membrane to a complex tone [Moore, 1998]. The complex tone is a periodic pulse train at a rate of 200 Hz. In spectral domain, there are many equal-amplitude harmonics. In the figure, the waveforms represent the observations at those points corresponding to the frequency values on the left. The horizontal lines on the BM illustrate the critical bands. Frequencies from different bands can be resolved in human perception. It is seen that the lower harmonics are partly resolved on the BM. At a place where a low harmonic resides, the response is approximately a sinusoid at that harmonic frequency. For example, at the place with a frequency of 400 Hz, the response waveform is a 400 Hz sinusoid. For such resolved low harmonics, i.e., the 1st, 2nd, 3rd, and 4th harmonics at 200 Hz, 400 Hz, 600 Hz, and 800 Hz, the auditory system detect the harmonic frequencies based on both the place on the BM, and the inter-spikes intervals in neuron with frequencies close to individual harmonics. Given the resolved low harmonics, the auditory system uses a pattern recognizer to find the fundamental frequency that matches these harmonics. In this example, the fundamental frequency matching the above mentioned harmonics is 200 Hz. The perceived pitch corresponding to this fundamental frequency is thus the 'place pitch'. On the other hand, the pitch of a complex tone can also be perceived in the form of 'temporal pitch'. The temporal pitch is provided by the high unresolved harmonics. As shown in the figure, the waveforms at the places corresponding to the high harmonics from 1 kHz to 6.4 kHz are complex. However, they show a common repetition rate equal to the fundamental frequency. The neural impulses tend to be evoked by the major peaks in the waveform, i.e., the peaks close to the location of envelope maxima. The impulses are separated by time intervals corresponding to the period of the sound. In Figure 2.1 the input has a repetition rate of 200 Hz and the period is 5 ms. The time intervals between the nerve spikes would cluster around integer multiples of 5 ms, i.e. 5, 10, 15, 20 ... ms. The pitch can be detected based on these time intervals. This is the mechanism of temporal pitch perception.

While both the low harmonics and high harmonics contribute to pitch perception, experimental evidence suggested that the resolved low harmonics give a clearer pitch than the high harmonics [Moore et al., 1985]. A residue pitch can be heard when only unresolvable high harmonics are present.

For HI people with cochlear hearing loss, the auditory filters are broader than normal. This makes it more difficult to resolve individual harmonics in the low-frequency region as NH people do. For example, for the fundamental frequency ($F0$) of 200 Hz, the 4th and 5th harmonics can be resolved in a normal auditory system, but may not be resolved in an impaired ear where the auditory filters were, say, three times broader than normal. Because of the reduced frequency selectivity, HI people tend to depend more on unresolved harmonics to capture the pitch based on the temporal mechanism.

For HI patients who wear CI devices, pitch information may be obtained via either temporal or place cues (or both). The temporal cues include variations in the frequency (or rate) of the stimuli delivered to one or more electrodes, or the periodic fluctuations in the amplitude of the stimulus. Additionally, the place of stimulation in the cochlea may be used to convey pitch information. Electrodes near the basal area elicit a higher pitch than those near the apical area [Nelson et al., 1995].

For human pitch perception, the ability to detect changes in pitch (pitch discrimination) is an important aspect. The changes in pitch are reflected by the changes in frequency. The smallest detectable change in frequency is called the frequency difference limen. There have been two common ways of investigating the frequency discrimination. One measure involves the discrimination of two successive tones with slightly different frequencies. In each trial, the tones are presented in a random order, the listener is required to indicate whether the first or the second tone is higher in frequency. The frequency difference between

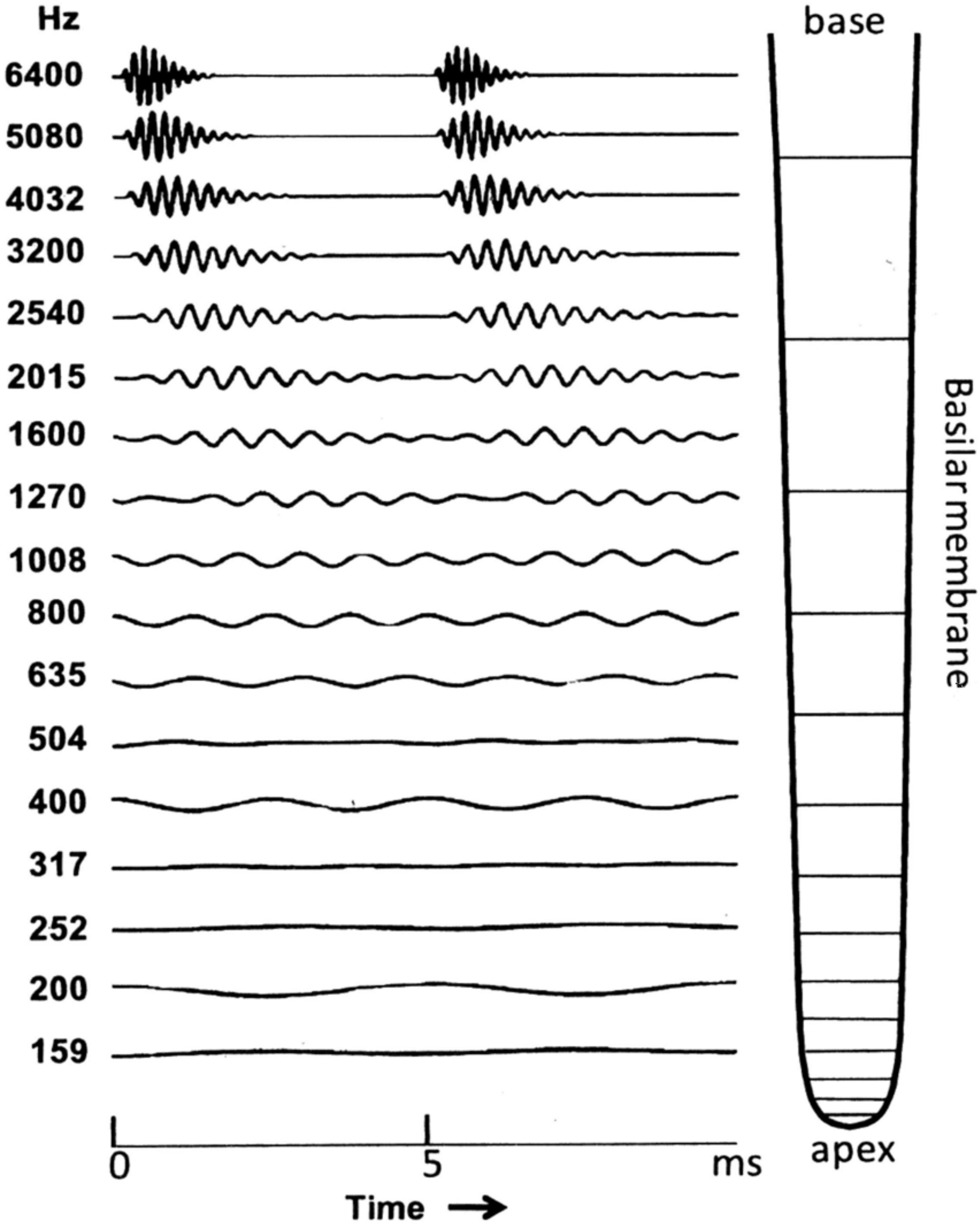


Figure 2.1: A simulation of the responses on the BM to periodic impulses of rate 200 pulses per second. The numbers on the left correspond to the frequencies which would maximally excite a given point on the BM. The waveforms that are plotted opposite to the numbers are the observations at those points, as a function of time. (The figure is adapted from [Moore, 1998])

the two tones is adjusted until the listener achieves a criterion of percentage correctness. This measure is called the DLF (difference limen for frequency). The second measure uses tones that are frequency modulated. Two tones are presented successively. One is modulated in frequency (between 2 and 20 Hz) and the other has a steady frequency. The order of the tones on each trial is randomized. The listener is required to indicate whether the first or the second tone is modulated. This measure is called the FMDL (frequency modulation detection limen).

2.3 Lexical Tones in Tonal Languages

Tone is described as the pitch contour over the voiced portion of one syllable. Different tones are characterized and identified by their distinctive pitch patterns [Francis et al., 2003]. Lexical tone plays an important role in tonal languages. A word carrying different tone patterns may have totally different meanings. For example, the Cantonese tone system comprises a set of six phonemic tones which form an additional but intrinsic part of Cantonese phonology. If the tone changes, the lexeme has a different meaning. For example, "愛國" /ngoi3 gwok3/ means 'patriot' but "外國" /ngoi6 gwok3/ means 'foreign country'. The only difference between the words is on the lexical tone of the first syllable (Tone 3 and Tone 6, respectively). Lexical tones carry a significant amount of linguistic information. This phenomenon happens only in tonal languages, but not in non-tonal languages.

2.3.1 Tonal Languages

By some estimates about 70% of the world's languages are tonal. They include languages spoken by huge populations, and in geographically diverse countries; Mandarin Chinese, Yoruba, and Swedish are all tonal. There are certain areas where almost all languages are tonal, such as sub-Saharan Africa, China, and Central America. A tonal language is defined as "a language with tone in which an indication of pitch enters into the lexical realization of at least some morphemes" [Hyman, 2001]. In addition to the consonants and vowels, a tonal language speaker uses an additional phonetic dimension for contrasting words by manipulating the pitch of the voice. Different pitch levels and movements can be used to represent different lexical and grammatical meanings of words [Bauer, 1997]. Such distinctive pitch patterns are called tones (or lexical tones). For both speech perception and production, tones are as important as the consonants and vowels.

In this thesis, we mainly focus on Cantonese and Mandarin which are the two major Chinese dialects spoken by over 800 million people around the world [Ethnologue, 2004a,b]. Cantonese speech is seen as a string of tonal syllables. Each syllable corresponds to some Chinese characters. Compared with English, Cantonese and Mandarin have a simpler and more restricted syllabic structure. All syllables have the structure of (consonant)- vowel-(consonant), where only the vowel nucleus is an obligatory element. If we consider only the phonemic composition of a syllable without tone, the syllable is referred to as a base syllable [Wang et al., 1995; Lee et al., 2002d; Qian et al., 2003]. Following the convention of Chinese phonology, each base syllable is divided into two parts, namely Initial and Final [Hashimoto, 1972]. The Initial (onset) includes what precedes the vowels while the Final includes the vowels (nucleus) and what follows it (coda). The tone resides on the voiced portion of a syllable. A base syllable carrying a tone becomes a tonal syllable, which specifies the complete pronunciation of a Chinese character. A character, however, can have multiple pronunciations, and a syllable typically can represent a number of different characters. For example, a Chinese character "和" has five different pronunciations in Mandarin. With different pronunciations, the character may have totally different meanings, i.e., "和" meaning "sum" or "and" in /he2/, "join in (the singing)" in /he4/, "mix" in /huo2/, "blend" in /huo4/, and "win in gambling" in /hu2/. Another example illustrating the phenomenon that a syllable represents a number of different characters is shown as following. The syllable /yi1/ in Mandarin may represents a large number of characters: "一" meaning "one", "衣" meaning "cloth", "依" meaning "according to", or "医"

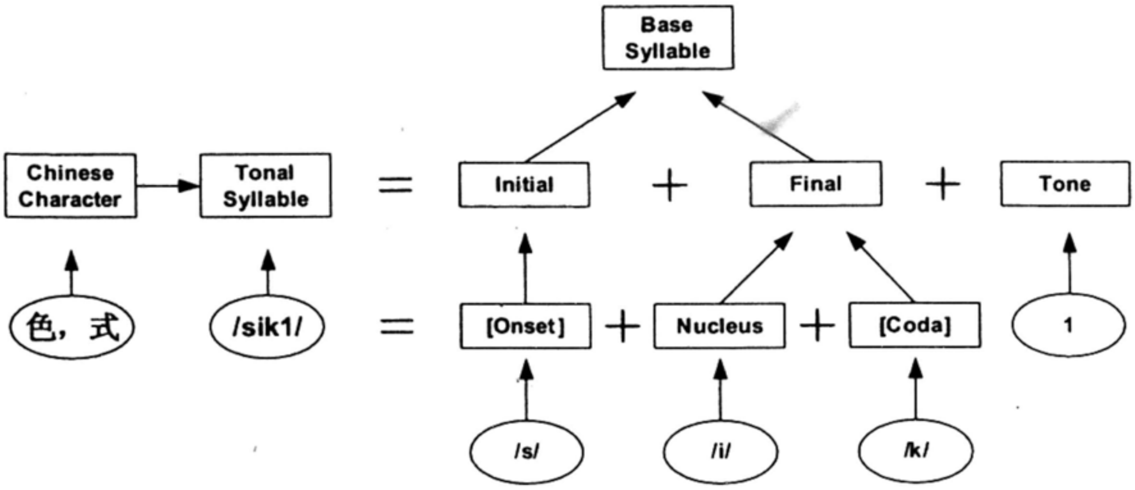


Figure 2.2: Structure of a Cantonese syllable.[] means optional.

meaning “medical”.

Figure 2.2 explains the composition of a Cantonese syllable with an example tonal syllable. The syllable /sik1/ may correspond to different characters, e.g., “色” meaning “color”, or “式” meaning “equation”. The tonal syllable is composed of two parts: the base syllable and the tone. In the base syllable, the Final is contains the nucleus or vowel. Table A.1 in Appendix gives the statistics on Cantonese syllables in comparison with those on Mandarin. Cantonese appears to have a richer inventory of syllable sounds than Mandarin.

There are 20 Initials and 53 Finals in Cantonese, in contrast to 21[1] Initials and 37 Finals[2] in Mandarin. Details of the Cantonese and Mandarin phonemes are listed in Appendix for reference.

2.3.2 Tone Patterns in Cantonese and Mandarin

Each tonal language has its own tone system. Cantonese is said to have nine tones that can be represented by their distinctive pitch patterns [Fok, 1974].

[1] The null Initial of /i/-start Finals is usually labeled as /j/, while the null Initial of /u/-start Finals is usually labeled as /w/, so it can be said there are 23 Initials in Mandarin.

[2] In Mandarin, the Final /ɹ/ is associated exclusively with Initials /ts/, /tsʻ/ and /s/, and /ʅ/ is associated exclusively with Initials /ʈʂ/, /ʈʂʻ/ and /ʂ/. In the Pinyin system, the above two Finals and the Final /i/ are labeled with the same symbol ‘i’, so there are only 35 Finals in the Pinyin system of Mandarin.

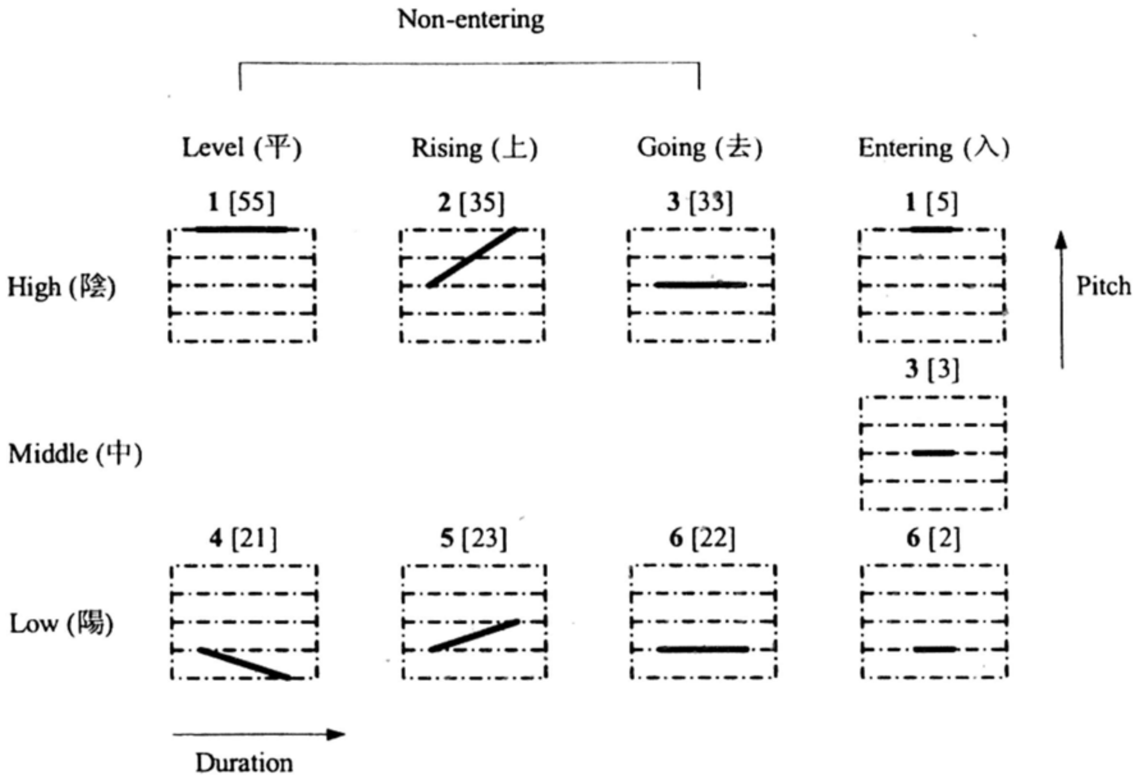


Figure 2.3: Schematic description of Cantonese tones and labeling schemes

Lexical tones can also be described impressionistically by using Chao's (1947) tone-letter notation system, in which "1" represents the lowest pitch level of a speaker's speech pitch range and "5" represents the highest pitch level. The Cantonese tone system is described in Figure 2.3 [LSHK, 1997]. It gives three different schemes of tone transcription. The Chinese labeling is from the historical tone category transcription [Fok, 1974]. The digits in [ ] are from Chao's system [Chao, 1947; Chen et al., 1997]. The bold digits are the tone labels from LSHK transcription [LSHK, 1997]. In this thesis, we adopt the LSHK labeling system. The six Cantonese tones are Tone 1 [55] (high level), Tone 2 [25] (high rising), Tone 3 [33] (middle level), Tone 4 [21] (low falling), Tone 5 [23] (low rising), and Tone 6 [22] (low level) [Bauer, 1997]. Examples of six Cantonese words with the same syllable [si] but different tones are /si1/ (詩,"poem"), /si2/ (史, "history"), /si3/ (試, "try"), /si4/ (時, "time"), /si5/ (市, "city"), /si6/ (氏, "surname").

In Cantonese, the so-called "entering" tones occur exclusively with "checked" syllables, i.e. syllables ending in an occlusive coda [p], [t] or [k].

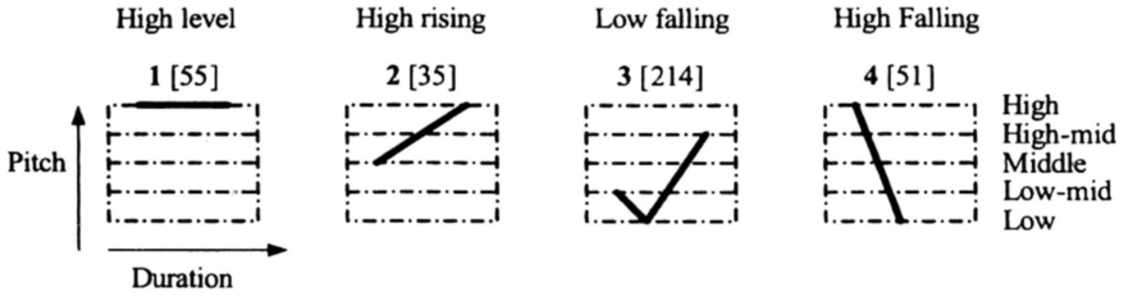


Figure 2.4: Schematic description of Mandarin tones and labeling schemes

They are contrastively shorter in duration than the "non-entering" tones. In terms of pitch level, each entering tone coincides roughly with a non-entering counterpart. Thus in LSHK scheme, only six distinctive tone categories are defined.

The tone system of Mandarin is simpler than that of Cantonese. There are four lexical tones in Mandarin. Each tone may be described as a contrastive pitch pattern. The four tones are shown in Figure 2.4. Same terminology as for Cantonese is used for Mandarin tone system. Different from Cantonese, the four tones in Mandarin are mainly distinguished in tone contour and duration. Tone 1 is high level; Tone 2 is high rising; Tone 3 is low falling; and Tone 4 is high falling. It is also found that Tone 4 has the shortest duration than the other three tones.

An interesting phenomena involving tones in Chinese dialects is called tone sandhi [Chen, 2000]. Tone sandhi refers to the change of tones when syllables are juxtaposed. To put it differently, a syllable has one of the tones in the language when it stands alone, but the same syllable may take on a different tone without a change in meaning when it is followed by another syllable. The most important tone sandhi rules in Mandarin involve the third tone. If a syllable has a weak stress or is unstressed, it loses its contrastive, relative pitch and therefore doesn't have one of the four tones described above. In such a case, the syllable is said to carry a neutral tone.

In general, tones are described by both pitch height and pitch contour. In Mandarin Chinese, the four tones are distinguished primarily by the pitch contour, while tones in Cantonese rely greatly on the pitch height to distinguish from each other. While fundamental frequency is the primary cue to discriminate lexical tones, other acoustic properties, such as syllable duration and amplitude contour, also convey tonal information [Liang, 1963; Whalen and Xu, 1992; Fu et al., 1998b; Fu and Zeng, 2000].

2.4 Tone Perception

Tone perception is part of speech perception. It plays an important role in speech communication with tonal languages. In a tonal language, there are a limited number of tones. Each of them has a distinctive pitch pattern. Tone perception is thus closely related to pitch perception, in that $F0$ is the primary acoustic cue.

The perception of pitch requires the detection of fundamental frequency of a complex tone with multiple harmonics or the detection of the frequency of a pure tone. Pitch perception experiments are commonly carried out in the form of pitch discrimination task. The listeners are asked to judge whether there exist a pitch difference between two pure tone stimuli with different $F0$s. These experiments reveal the ability of human to detect different pitches. In a more complex situation, continuous pitch movement over time is taken into consideration (from high to low or from low to high). The ability to detect the gradually changed pitch is investigated in these experiments.

Tone perception is different from pitch perception in that a perceptually detectable pitch difference does not necessarily causes a phonological tone contrast. Pitch movement (rising or falling) and pitch height/level (high, middle or low) together determine the lexical tones [3]. Tonal languages provide not only phonological but also linguistic features for understanding the importance of pitch cues in language perception. Different from pure tone and complex tone, lexical tone in tonal languages is a suprasegmental feature that is used to distinguish different words. The perception of lexical tones depends not only

[3] Hereafter, 'pitch movement', 'pitch contour' and 'tone contour' have the same meaning; 'pitch height', 'pitch level' and 'tone level' have the same meaning.

on pitch level, but also on the perception of other cues, such as pitch movement direction, vowel duration and amplitude contours, etc.

Tone perception is said to be a kind of categorical perception. When stimuli are perceived categorically, equivalent acoustic differences between two tokens may be treated differently, depending on whether the two tokens are heard as members of the same category or as members of different categories. Two members in the same category are less discriminable than two tokens from different categories, although the acoustic differences may be the same in the two cases. With experience on a language, listeners learn the location of specific category boundaries along various acoustic continua. By increasing discrimination accuracy across these boundaries and reducing it within boundaries, listeners improve their ability to hear two acoustically different members of one category as the same and, conversely, improve their ability to hear two acoustically similar members from different categories as different. It was shown that tone perception in Mandarin and Cantonese is categorical perception [Francis et al., 2003; Wang, 1976]. Francis et al. [2003] also reported that tonal category boundaries are determined by a combination of regions of natural auditory sensitivity and the influence of linguistic experience.

Acoustic features that are related with tone recognition have been studied extensively [Gandour, 1984; Liang, 1963; Lin, 1988; Whalen and Xu, 1992; Abramson, 1978; Xu and Pfingst, 2003; Kong and Zeng, 2006]. The first systematic study on acoustic cues for Mandarin tone perception was performed in [Liang, 1963]. In his study, perfect tone recognition (90% correct or above) was observed with a male voice by filtering out either ~~of the~~ first, second or the third harmonic, or any two of them from broad band speech signals. This revealed that the fundamental frequency is useful in tone perception, but it is not the only cue. Even with the absence of $F0$, tone perception is possible with the second and the third harmonics. The same study further demonstrated that robust tone recognition (64% correct) could be achieved on whispered speech, in which only temporal and spectral envelope cues were preserved, and spectral fine structure related to $F0$ and its harmonics are not available. Several follow-up studies found that the temporal cues such as vowel duration and amplitude contours also contributed to Mandarin tone recognition [Garding et al., 1986; Blicher et al., 1990; Whalen and Xu, 1992]. The contribution of these temporal cues was significant only when the $F0$ and its harmonics were not present [Lin, 1988].

2.5 Tone Perception in CI

Pitch perception in CI users is found to be far from satisfactory [Pijl, 1997; McDermott, 2004; Gfeller et al., 2006; Sucher and McDermott, 2007; Green et al., 2002; Qin and Oxenham, 2005; Stickney et al., 2004]. It caused poor discrimination of musical intervals [Gfeller et al., 2002, 2006] and lexical tones in tonal languages [Ciocca et al., 2002; Wong and Wong, 2004; Luo et al., 2008].

In a CI system, each frequency channel contains a signal which is the combination of several harmonics when a complex tone is presented. McDermott [2004] investigated the music perception by CI users and addressed two potential cues that are available for pitch perception through CIs. The first cue is place pitch which corresponds to the place of stimulation along the cochlea. Due to the limited number of frequency channels (electrodes) in current CI system and the broad bandwidth of the band-pass filters, individual harmonics of a complex sound can not be resolved as do the NH listeners [Geurts and Wouters, 2001]. This inevitably limits the use of place pitch cue for accurate pitch perception. The second cue for pitch perception comes from the use of rate pitch (corresponding to the periodicity cues in normal hearing). The rate pitch in CI systems is derived from the within-channel amplitude modulation of electrical pulse train. Therefore, it is expected that enhancing the coding of periodicity cues to $F0$ may provide better speech understanding in tonal languages and better music perception [Rubinstein, 2004; Vandali et al., 2005].

Some studies have investigated the importance of temporal and/or spectral cues for tone recognition in CI users. There are a number of factors that may affect the performance of tone perception in CI. It has been shown that CI listeners are capable of processing $F0$ cues which are presented in the speech sound to perform lexical tone recognition [Fu et al., 1998b]. Wei et al. [2004] tested five Nucleus-22 users for Mandarin tone recognition. The users showed a clear dependence of tone recognition on the number of channels. The performance improved from 30% correct to 70% correct with one channel and ten channels, respectively. Liu et al. [2004] tested tone recognition with part of the electrodes inactivated in six Mandarin-speaking children. The results indicated that the temporal envelope cues from the high-frequency region were more important for tone recognition than those from low-frequency region. Fu et al. [2004b] tested the effects of various stimulation rates on tone recognition with different signal processing strategies. Results showed that no significant difference in performance among the various stimulation rates. Contrarily, high pulse rates (900 Hz or above) were found to have negative effect on tone perception in other studies [Barry et al., 2002; Ciocca et al., 2002; Lee et al., 2002b; Wong and Wong, 2004]. A slower stimulation rate at 250 Hz also showed a significant decrease in tone recognition. It was not clear whether the results were due to the fact that the subjects were tested with the strategies he did not use in daily life. There was a tradeoff between the number of electrodes and the stimulation rate for tone recognition as for vowel recognition [Nie et al., 2006]. It was noted that the subjects showed similar tone recognition performance when fitted with the CIS strategy using 12, 8, 6, and 4 electrodes coupled with stimulation rates of 1200, 1800, 2400, and 3600 Hz, respectively. According to the sampling theorem, the stimulation rate must be at least two times higher than the cutoff frequency of the low-pass filter. McKay [McKay et al., 1994] suggested that it needs to be at least four times of the envelope cutoff frequency.

□ **End of chapter.**

Chapter 3

Design of Tone Perception Tests

Summary

This research is carried out with a series of psychoacoustic experiments on tone identification. In this chapter, we discuss about the design of these tests. The tests are special in that tone is not independent, it must be tested in a speech recognition task, such as a word recognition task. We designed different types of speech materials, including monosyllabic and disyllabic words, which are shown to be suitable for the tone identification tests. The speech signal processing is based on a multi-band noise-excited vocoder to simulate signal processing in CIs. A PC-based research platform is developed to control the stimulus presentation and record the subjects' responses.

3.1 Overview

Because pitch is a subjective sensation, assessing tone perception which is based on the perception of pitch is only possible through experiments with humans. This raises the need for an experimental setup that allows testing the subjects in a controlled way so that the results can be reproduced. The design of a tone perception test involves a number of considerations including the speech materials, the signal processing methods, the test procedure, the recruitment of subjects, and the test interface.

Tone can not exist without a particular syllable carrying it. The patterns of lexical tones are language-specific, e.g., Cantonese and Mandarin have totally different tone patterns. Researchers should determine which tonal language should be tested and what kind of speech elements should be used, e.g., vowel, word, or even sentence. Speech stimuli are generated from the speech materials by the use of certain signal processing methods. The signal processing is the 'heart' of a scientific experiment, which determines the information delivered to the subjects. The generated speech stimuli should be presented to the subjects in an appropriate way. It requires a well-controlled and scientifically logic procedure. Such a psychoacoustic setup needs a research platform to realize the requirements. It should preferably be versatile and user-friendly at the same time.

In the following chapter, every part described above will be discussed. Speech materials are monosyllabic and disyllabic words with full tone patterns in both Mandarin and Cantonese. A multi-channel noise-excited vocoder is used as a baseline signal processor. Tone identification task is performed. A research platform, referred to as ATOPEX (Application for TOne Perception EXperiments), is developed that allows the researcher to set up arbitrary auditory psychophysical experiments.

3.2 Speech Materials

In Cantonese and Mandarin, lexical tones are carried by syllables and can not be separately tested without words or sentences. Monosyllabic words have been commonly used as the carriers for tone perception test. In such tests, the segmental structures of the words were kept constant while the tones were different. In some of the studies, researchers applied a small set of syllables, e.g., 1 to 6 different syllables in the tone perception test [Ma et al., 2006; Luo et al., 2008; Khouw and Ciocca, 2007; Wong and Wong, 2004] while some of the researchers used a large set of syllables, e.g., over 20 different syllables [Liang, 1963; Kong and Zeng, 2006; Wei et al., 2004]. The various number of syllables used may depend on the goals of the tests, e.g., for children [Lee et al., 2002b; Liu et al., 2004; Wong and Wong, 2004]. It is also restricted because of the characteristics of the language. For example, there are only a few Cantonese base syllables that can be combined with the six tones to form meaningful words [Khouw and Ciocca, 2007; Ma et al., 2006]. Whereas in Mandarin, there are many such monosyllabic words [Wei et al., 2004; Kong and Zeng, 2006; Liang, 1963]. Khouw and Ciocca [2007] used four sets of monosyllabic Cantonese words with six words in each set. They explained that it was not always possible to use minimal contrasts in Cantonese. They therefore used words that differ in initial consonants in addition to the differences in tones. Due to the different limitation of the language characteristics, we adopted different criterions in designing the speech materials in Cantonese and Mandarin. In this thesis, monosyllabic and disyllabic words are used as speech materials for lexical tone test. Monosyllabic words provide clear comparisons in tones with the same base syllable. Disyllabic words carry more contextual information for linguistic comprehension of a language. Speech materials with more complex segmental structures, e.g., phrases and sentences, may also be considered for a tone perception experiment. However, they are not included in our study because the subjects may depend on the context information rather than the tone information for phrase or sentence recognition.

Table 3.1: The written characters of the three sets of lexical tones.

	Lexical tone number					
	1	2	3	4	5	6
/fu/	膚	虎	富	符	婦	負
/ji/	衣	椅	意	兒	耳	二
/wai/	威	委	餵	圍	偉	胃

3.2.1 Cantonese Speech Materials

Monosyllabic Speech Material

We used three sets of monosyllabic words, each containing different segmental structure with the following three base syllables: /fu/, /ji/ and /wai/. These syllables were chosen based on the considerations that there exist valid words for all of the six tones. These words occur frequently in daily communication. The Chinese characters of the three sets of words are listed in Table 3.1. Ideally, including more test sets would be helpful to avoid the bias towards the specific formant structures of syllables. As shown in Figure 3.1, the selected three monosyllables have different formant structures.

Digital recordings of the speech materials were prepared from two native Cantonese speakers (one male and one female). The recording was done in a sound-proof room. The Cool Edit Pro 2.0 software (Syntrillium Software Corporation, 2002) was used to perform recording with a sampling rate of 44,100 Hz and 16 bit resolution.

During recording, the speaker spoke a carrier phrase "這個詞是 _ _"(This word is _ _). There is a pause inserted between the carrier phrase and the target syllable. Five trials were done for each target syllable. One of them was manually selected and used in the study.

For each set of the six words, the durations of the syllables carrying those tones were equalized according to the mean duration of the six recorded syllables in the set. The Praat v.4.1.2.1 software [Boersma and Weenink, 1992 – 2003] was used to perform this equalization. The quality and naturalness of

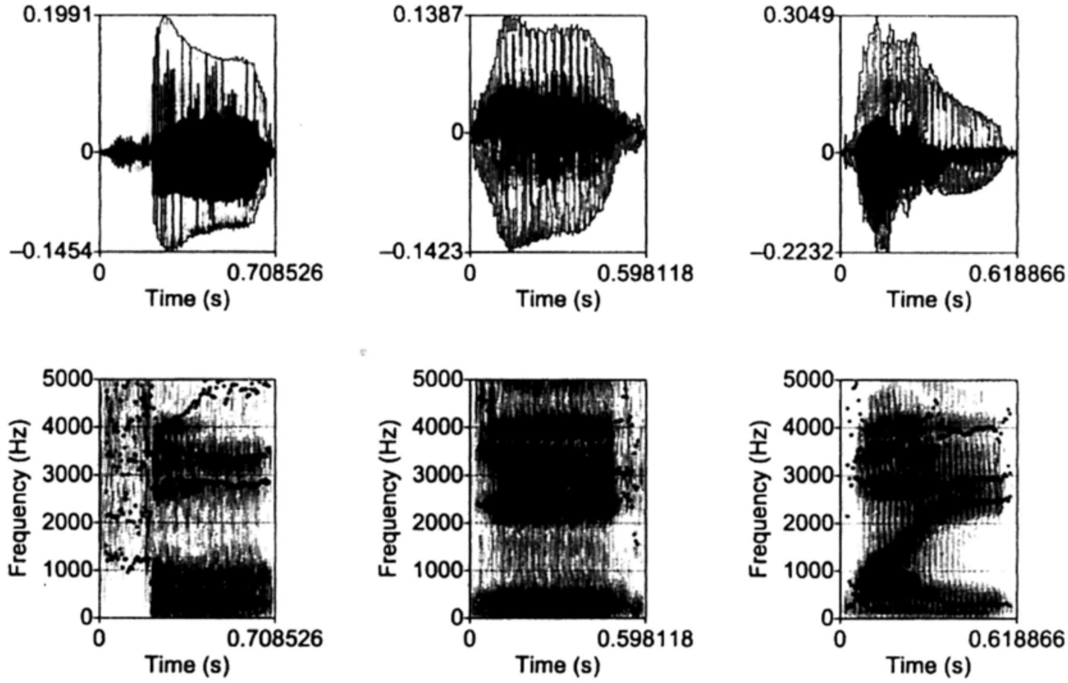


Figure 3.1: Formant structures of the three monosyllables, /fu/, /ji/ and /wai/, respectively. The upper plots are the waveforms in time domain. The corresponding lower plots are the spectrograms with formant structures (dotted curves).

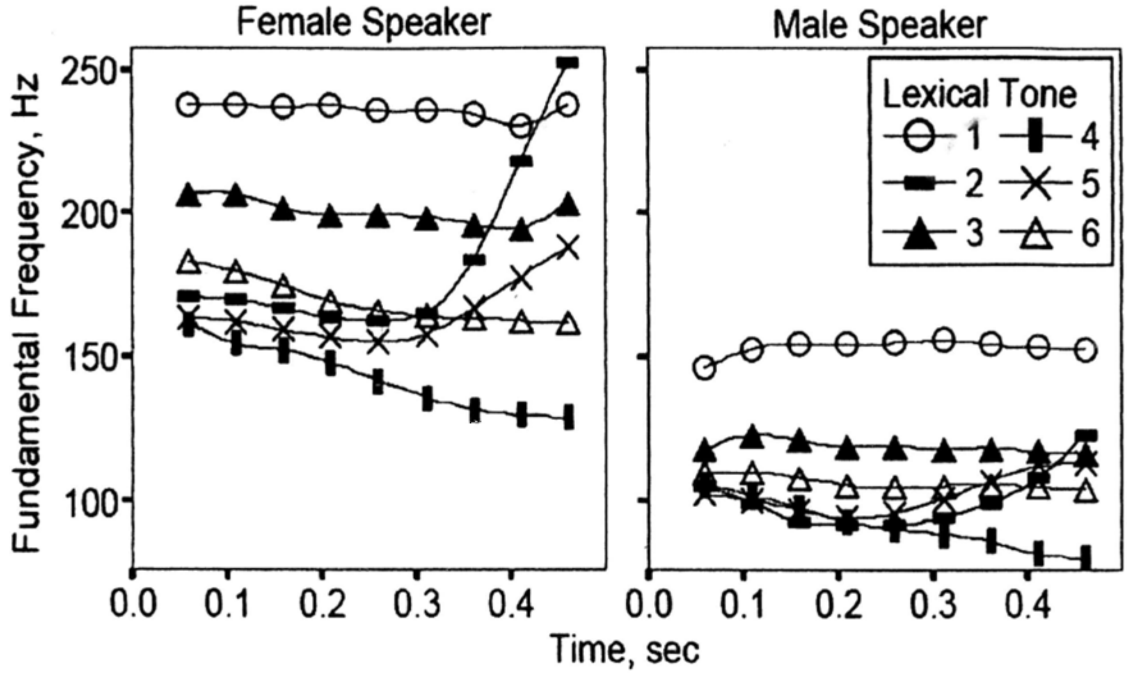


Figure 3.2: Average $F0$ changes over time.

these duration-equalized speech were affirmed by two native Cantonese speakers. Figure 3.2 gives the average $F0$ contours of the six tones over the three base syllables. Male speech and female speech are displayed separately.

In each experimental condition, there are a total of 36 presentations on each monosyllable. Each presentation consists of two lexical tone stimuli with the same syllable. The two stimuli were presented sequentially with a 500 ms silence interval in between. The 36 presentations exhausted all possible pairing combinations of the 6 lexical tones (6 lexical tones × 6 lexical tones). Therefore, there were a total of 72 lexical tone items presented. Each tone has 12 representatives. The 36 presentations were in random order. The subject's response from each presentation was entered into the test interface and stored for future analysis. The subject was allowed to listen to one repetition for the first five presentations. For the remaining 31 presentations no repetition was allowed. Unfilled response was not allowed and the subjects were requested to provide guesses for any uncertainties. Since the response format was a six-alternative force choice task, the chance level performance was 16.7%.

Disyllabic Speech Material

In Ciocca et al. [2002] and Wei et al. [2004], tone identification tests were carried out using a set of monosyllabic words that have the same segmental structure and carry different tones. Due to limited segmental variation in the test materials, the linguistic role of tone was not fully reflected in these tests. Moore and Jongman [1997] stated that the intrinsic acoustic properties ($F0$ level and contour) specific to each lexical tone might be sufficient for the correct identification of Mandarin tones that contrast in both $F0$ height and contour. However, the presence of three level tones (Tone 1, 3, and 6) and two rising tones (Tone 2 and 5) in Cantonese makes perception based only on intrinsic acoustic properties unlikely, as $F0$ height is a significant factor in differentiating tones of the same contour. For example, Fok [1974] reported confusion between Tone 3 and Tone 6 in their perceptual experiment in Cantonese with the targets presented in isolation; she proposed that the confusion is due to the absence of extrinsic context cues with these stimuli. Fok also reported that Tone 1 and Tone 4 were more resistant to confusion in the absence of extrinsic context when compared with other tones. This suggested that perception of the six tones might depend on the extrinsic context to different degrees as the intrinsic $F0$ changes of some tones (e.g., the falling contour of Tone 4 [21], as it is the only falling tone in Cantonese) may be of saliency to the listeners for correct tone identification. Therefore, another purpose of using disyllabic words was to provide extrinsic context cue in perception of all six tones in Cantonese.

Tone is used as a contrastive linguistic component for word identification. In Chinese, disyllabic words are much more commonly used than monosyllabic words [Chin, 1998], and therefore considered to be more appropriate for tone recognition test. Ideally, we need many sets of disyllabic words and each set of words are minimally contrasted by the tone of one of the syllables. Given the lexical constraints of Cantonese, it is difficult to find even one set of words that cover all of the six tones with the same segmental structures. Therefore, we decide to include only a pair of contrasting tones in each set of words. There are a total of 15 contrasting tone pairs. The contrasting tone may be on either the first or the second syllable of the disyllabic words. Thus, we need 30 sets of words.

Table 3.2 lists the 30 sets of words being used in our study. There are four words in each set. The left two words, denoted by MC_A and MC_B, carry the intended contrasting tones with the same segmental properties. For example, MC_A and MC_B in Set 1 are /ging1 lik6/ and /ging2 lik6/, respectively, which differ in the tones carried by the first syllables. The same tone contrast also occurs in Set 16, but on the second syllables in the words. However, if each test trial involves only two candidates words, the subjects may easily realize that the test is focused on one of the syllables. To minimize the learning effect, two additional words are included in each set, as shown in the right two columns of Table 3.2. These words are referred to as quasi-controlled tone contrasting words and denoted by QC_A and QC_B. They share the same tone contrast as that between MC_A and MC_B, but have slightly different segmental compositions. The whole set of 120 disyllabic words cover nearly 90% of the Cantonese phonemes.

Table 3.3 shows all possible outcomes of the test. T and $\overline{T}$ are used to represent the cases of correct and wrong identification of tone, respectively, while S and $\overline{S}$ refer to correct and wrong identification of segmental structures, respectively. For example, $(T, \overline{S})$ means that tone is correctly identified but the segmental identification is wrong. Although this study is focused primarily on tone recognition, the test results will also facilitate the investigation on the importance of temporal envelope and periodicity cues in conveying segmental information.

The speech materials were recorded from a female and a male native speaker of Hong Kong Cantonese. Recordings were made in a sound-treated booth with an Etymotic Research ER-11 microphone connected to a desktop computer with an external sound-card. The recorded signals were digitized with a sampling frequency of 44,100 Hz and 16-bit resolution. To maintain a consistent pitch level of the speakers' voice, all words were recorded with the same carrier sentence: "這個詞是 _ _"(This word is _ _). Five repetitions were recorded for each

Table 3.2: The list of Cantonese disyllabic words. Each row contains four candidate words for selection in one test trial.

Set	MC_A	MC_B	QC_A	QC_B
1	ging1 lik6 經歷	ging2 lik6 警力	gung1 lik6 功力	geng2 lik6 頸力
2	gei1 gin2 機件	gei3 gin2 寄件	gai1 gin2 雞件	gei3 cin2 寄錢
3	jau1 mei5 優美	jau4 mei5 柔美	au1 mei5 歐美	ngau4 mei5 牛尾
4	jau1 ji6 優異	jau5 ji6 有異	jat1 ji6 一二	jau5 si6 有事
5	dak1 ji3 得意	dak6 ji3 特意	hak1 ji3 刻意	dik6 ji3 敵意
6	gu2 jan4 古人	gu3 jan4 故人	gaa2 jan4 假人	go3 jan4 個人
7	jan2 jing4 隱形	jan4 jing4 人形	jan2 cing4 隱情	jan4 cing4 人情
8	waan2 gau3 玩夠	waan5 gau3 挽救	gaan2 gau3 揀夠	laan5 gau3 懶夠
9	gau2 paai4 狗牌	gau6 paai4 舊牌	zau2 paai4 酒牌	hau6 paai4 後排
10	paa3 gou1 怕高	paa4 gou1 爬高	gwaa3 gou1 掛高	ngaa4 gou1 牙膏
11	jau3 ji4 幼兒	jau5 ji4 友誼	jau3 si4 幼時	jau5 si4 有時
12	ngoi3 gwok3 愛國	ngoi6 gwok3 外國	ngoi3 gwo3 愛過	hoi6 gwok3 害國
13	mei4 miu6 微妙	mei5 miu6 美妙	kei4 miu6 奇妙	mei5 maau6 美貌
14	jyun4 ji3 原意	jyun6 ji3 願意	cyun4 ji3 傳意	gyun6 ji3 倦意
15	lou5 jan4 老人	lou6 jan4 路人	mou5 jan4 冇人	zou6 jan4 做人
16	daai6 baan1 大班	daai6 baan2 大阪	daai6 caan1 大餐	daai6 daan2 大蛋
17	daai6 ji1 大衣	daai6 ji3 大意	daai6 si1 大師	daai6 si3 大使
18	do1 jyu1 多於	do1 jyu4 多餘	do1 syu1 多書	do1 ji4 多疑
19	daai6 jyu1 大於	daai6 jyu5 大雨	daai6 zyu1 大豬	daai6 ji5 大耳
20	jau5 jik1 有益	jau5 jik6 有翼	jau5 sik1 有色	jau5 lik6 有力
21	jat1 bun2 一本	jat1 bun3 一半	jat1 wun2 一碗	jat1 gun3 一罐
22	jyu4 leon2 魚卵	jyu4 leon4 魚鱗	jyu4 ceon2 愚蠢	jyu4 seon4 魚唇
23	daai6 jyu2 大魚	daai6 jyu5 大雨	daai6 jyun2 大丸	daai6 jau5 大有
24	gaau1 doi2 膠袋	gaau1 doi6 交待	gaau1 daai2 膠帶	gaau1 ngoi6 郊外
25	daa2 ping3 打拼	daa2 ping4 打平	daa2 ting3 打聽	daa2 sing4 打成
26	kyut3 ji3 決意	kyut3 ji5 決議	kyut3 zi3 決志	syut3 ji5 雪耳
27	daai6 bou3 大埔	daai6 bou6 大步	daai6 ngou3 大澳	daai6 dou6 大盜
28	dou6 jau4 導遊	dou6 jau5 道友	ngou6 jau4 遨遊	deoi6 jau5 隊友
29	juk6 maa4 肉麻	juk6 maa6 辱罵	juk6 ngaa4 肉芽	juk6 baa6 欲罷
30	jyun4 mei5 完美	jyun4 mei6 原味	jyun4 lei5 原理	jyun4 bei6 完備

Table 3.3: Possible outcomes of the word identification tests. T and $\overline{T}$ denote correct and wrong identification of tone, respectively. S and $\overline{S}$ denote correct and wrong identification of segmental structures, respectively.

	MC_A	MC_B	QC_A	QC_B
MC_A	(T, S)	$(\overline{T}, S)$	$(T, \overline{S})$	$(\overline{T}, \overline{S})$
MC_B	$(\overline{T}, S)$	(T, S)	$(\overline{T}, \overline{S})$	$(T, \overline{S})$
QC_A	$(T, \overline{S})$	$(\overline{T}, \overline{S})$	(T, S)	$(\overline{T}, \overline{S})$
QC_B	$(\overline{T}, \overline{S})$	$(T, \overline{S})$	$(\overline{T}, \overline{S})$	(T, S)

test word. Another two native speakers assessed the quality and naturalness of these recordings and selected the best one for our experiments. The test words were manually excised from the carrier sentences.

3.2.2 Mandarin Speech Materials

The speech materials consist of two parts: monosyllabic words and disyllabic words.

Monosyllabic Speech Material

Different from Cantonese, it is much easier to find many sets of monosyllables with four tones meaningfully. There are 20 sets of monosyllabic words included in our study. Each set contains four words that share the same segmental structure but different tones. For example, one set of words was: "妈"(mother), "麻"(linen),"马"(horse), "骂"(curse), with the same consonant-vowel syllable, /ma/ in Tone 1 to 4, respectively. The selection criteria of these monosyllabic words includes: (1) being commonly used in daily communication; and (2) covering as many consonant-vowel combinations as possible. One male and one female speaker were asked to utter the 80 monosyllabic words 5 times and the acoustically optimal one was manually selected as the speech material. Table 3.4 shows the monosyllabic words designed for our Mandarin tone perception test.

Disyllabic Speech Material

The disyllabic words were recorded from the same speakers. The design of the Mandarin disyllabic words shares the same concept as that of the Cantonese disyllabic words. One pair of contrasting tones, e.g., “奴隶 ” /nu2 li4/ (slave) and “努力” /nu3 li4/ (struggle), is included in each test set. There are totally 6 contrasting pairs for the four tones in Mandarin. The contrasting tone may be on either the first or the second syllable of the disyllabic words while the segmental properties of the words are identical. For each tone pair, two sets of different words were created, e.g., Set 7 and Set 8 both contain the tone pair, Tone 2 and Tone 3, but has different segmental structures, i.e. /yang qi/ and /nu li/, respectively. Thus, we have 24 sets of words. Additionally, another two words were included in each set, which consist the same tone contrast as that between the two tone contrasting words, but have slightly different segmental compositions. Such words that correspond to the two tone contrasting words in the above example are “无力 ” /wu2 li4/ (disability) and “牡蛎 ” /mu3 li4/ (oyster). They are intended to minimize the learning effect and provide segmental variation of the syllables. The whole set of 96 disyllabic words covers nearly 90% of the Mandarin phonemes. The disyllabic words are shown in Table 3.5.

Recordings of both monosyllabic and disyllabic words were made in a sound-proof booth with a high-quality microphone and a desktop computer. All the speech signals were equalized to the same intensity and low-pass limited to 4 kHz.

3.3 Acoustic Simulation with Noise-excited Vocoder

Throughout this thesis, the signal processing strategies are based upon a noise-excited vocoder. It simulates the Continuous Interleaved Sampling (CIS) speech

Table 3.4: The list of Mandarin monosyllabic words. Each row contains four candidate words for selection in one test trial.

	Tone number			
	1	2	3	4
ma	妈	麻	马	骂
po	坡	婆	叵	破
ke	科	壳	可	课
du	都	读	赌	度
wei	微	围	委	卫
bao	包	薄	饱	暴
qiao	敲	桥	巧	翘
you	优	游	有	又
duo	多	夺	躲	堕
wan	弯	玩	晚	万
tong	通	同	统	痛
nian	蔫	年	碾	念
pin	拼	频	品	聘
liu	溜	流	柳	六
guo	锅	国	果	过
bai	掰	白	摆	拜
mi	眯	迷	米	密
tang	汤	唐	躺	烫
yun	晕	云	陨	运
ying	英	赢	影	硬

Table 3.5: The list of Mandarin disyllabic words. Each row contains four candidate words for selection in one test trial.

Set	MC_A	MC_B	QC_A	QC_B
1	ge1 duan4 割断	ge2 duan4 隔断	bo1 duan4 波段	zhe2 duan4 折断
2	bo1 wen2 波纹	bo2 wen2 博闻	zhong1 wen2 中文	de2 wen2 德文
3	tao1 qian2 掏钱	tao3 qian2 讨钱	chao1 qian2 超前	zhao3 qian2 找钱
4	biao1 ming2 标明	biao3 ming2 表明	gao1 ming2 高明	tiao3 ming2 挑明
5	tui1 chu1 推出	tui4 chu1 退出	tu1 chu1 突出	hui4 chu1 汇出
6	duan1 kou3 端口	duan4 kou3 断口	guan1 kou3 关口	chuan4 kou3 串口
7	yang2 qi4 洋气	yang3 qi4 氧气	fang2 qi4 房契	nuan3 qi4 暖气
8	nu2 li4 奴隶	nu3 li4 努力	wu2 li4 无力	mu3 li4 牡蛎
9	nan2 you3 男友	nan4 you3 难友	han2 you3 含有	zhan4 you3 战友
10	pan2 tao2 蟠桃	pan4 tao2 叛逃	qian2 tao2 潜逃	cuan4 tao2 窜逃
11	guan3 yong4 管用	guan4 yong4 惯用	xuan3 yong4 选用	lan4 yong4 滥用
12	ke3 guan1 可观	ke4 guan1 客观	jing3 guan1 景观	guo4 guan1 过关
13	dong4 ting1 动听	dong4 ting2 洞庭	dong4 xin1 动心	dong4 qing2 动情
14	min2 ge1 民歌	min2 ge2 民革	min2 jian1 民间	min2 guo2 民国
15	tong4 ku1 痛哭	tong4 ku3 痛苦	tong4 shi1 痛失	tong4 chu3 痛楚
16	tong2 ban1 同班	tong2 ban3 铜板	tong2 guan1 潼关	tong2 gan3 同感
17	kai1 tuo1 开脱	kai1 tuo4 开拓	kai1 guo1 开锅	kai1 kuo4 开阔
18	qing1 dan1 清单	qing1 dan4 清淡	qing1 shan1 青山	qing1 suan4 清算
19	mu4 tong2 牧童	mu4 tong3 木桶	mu4 peng2 木棚	mu4 ou3 木偶
20	wei2 du2 惟独	wei2 du3 围堵	wei2 nu2 为奴	wei2 bu3 围捕
21	mi4 tan2 密谈	mi4 tan4 密探	mi4 han2 密函	mi4 jian4 蜜饯
22	liu2 lian2 流连	liu2 lian4 留恋	liu2 yan2 留言	liu2 nian4 留念
23	zi4 li3 自理	zi4 li4 自立	zi4 ji3 自己	zi4 bi4 自闭
24	qing1 li3 清理	qing1 li4 倾力	qing1 xi3 清洗	qing1 xin4 轻信

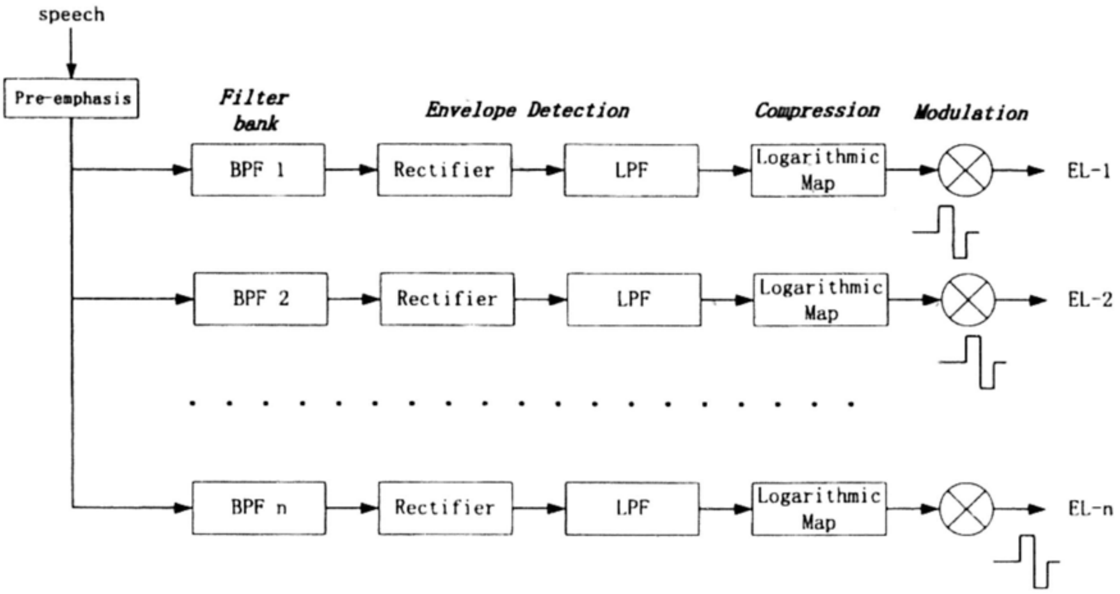


Figure 3.3: Flowchart of the CIS strategy process. BPF means band-pass filtering; LPF means low-pass filtering; and EL-n represents the n^{th} electrode.

processing strategy, which is commonly implemented in CI devices [Wilson et al., 1991; Wilson, 2000; Vandali et al., 2005].

CIS is based on a speech vocoder that could be traced back to 1940s [Dudly, 1939]. It was proposed by Wilson and his colleagues in 1991 [Wilson et al., 1991]. Figure 3.3 shows a flowchart of the signal processing in CIS.

In CIS the signal is first pre-emphasized. The pre-emphasis process is a high-pass filter which attenuates strong low-frequency components below 1.2 kHz at 6 dB/octave in speech that might mask the important high-frequency components. The signal then undergoes a filter bank which separates the signal into band-limited channels. The number of channels often corresponds to the number of active electrodes in the implant [Zeng, 2004]. Bandpass filtering simulates the physiology of the cochlea where different frequencies are detected at different positions on the BM. In each analysis band, the temporal envelope of the signal is extracted by the envelope detection function. This is done by full-wave/half-wave rectification and low-pass filtering. There are some other methods that can be used to detect the envelope, e.g., Hilbert transform. Detailed discussion has been presented in Section 1.2.2. A compression function is then applied to the envelope amplitude in each channel. The compression is an

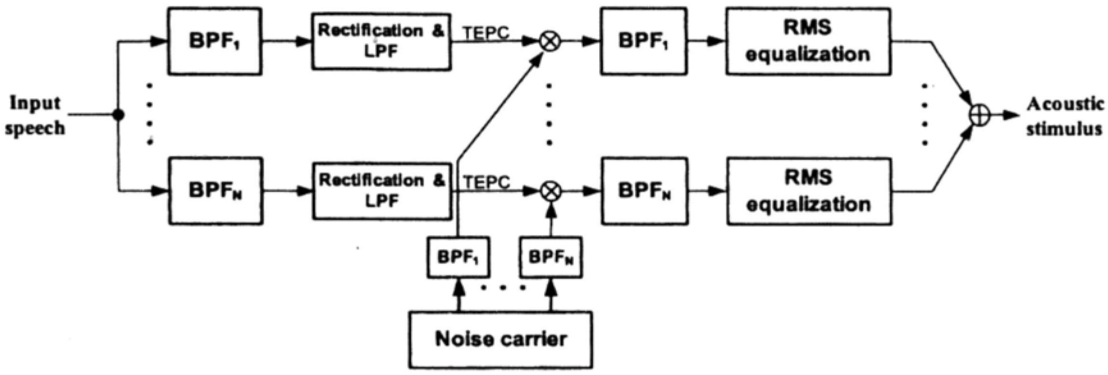


Figure 3.4: Flowchart of the noise-excited vocoder. BPF and LPF abbreviate band-pass filter and low-pass filter respectively.

essential component of the CIS processor. This function is necessary because the range of acoustic amplitudes in conversational speech is considerably larger than the CI user's dynamic range.

The key feature of CIS is that the electric pulses on different channels are not overlapped in time. They are interleaved to avoid the interaction among electrodes. If the electrodes are simultaneously stimulated, the interaction would be produced through vector summation of the electrical fields from the simultaneously stimulated electrodes. Such interaction would interfere the salience of the channel-related speech cues [Wilson et al., 1991].

Figure 3.4 depicts the standard implementation of the noise-excited vocoder [Shannon et al., 1995; Xu et al., 2002]. Similar to CIS, the input speech goes through a set of band-pass filters. TEPCs are extracted by the use of full-wave rectification and low-pass filtering. The major difference from CIS is that the TEPCs are used to amplitude-modulate noise carriers in the respective bands, instead of pulse trains in CIS. No compression is needed since NH listeners are tested. The modulated noise signals are band-passed again with the same analysis band to remove undesirable frequency components generated in the modulation process. At each band, the signal's intensity level is adjusted to match that of the original band-passed speech. Finally, acoustic stimuli for psychophysical experiments are generated by combining the modulated signals from each band.

If not particularly claimed, the speech materials undergo a four-channel noise-excited envelope vocoder, similar to the procedures described in many previous studies, e.g. [Dorman et al., 1997; Fu et al., 1998b; Shannon et al., 1995]. The frequency ranges of the four sub-bands were: 60 – 500 Hz, 500 -- 1000 Hz, 1000 – 2000 Hz, and 2000 – 4000 Hz. This band structure is slightly different from those used in some previous studies [Dorman et al., 1997; Fu et al., 1998b]. Shannon et al. [1998] compared various definitions of sub-bands and found that they were not critical for speech recognition. All band-pass filters were elliptical filters. The low-pass filter for TEPC extraction is an elliptical filter with cut-off frequency of 500 Hz. However, different parameter settings may be used depending on the particular experimental requirements. In our implementation of the noise-excited vocoder, the pre-emphasis procedure was not included. In pilot tests, it was found that pre-emphasis adversely affected speech recognition performance, especially for vowel recognition.

3.4 Research Platform

This section describes a PC-based research platform for controlling the stimulus presentation and collecting the subjects' responses. The platform enables researchers to set up auditory experiments in a user-friendly way. The stimuli and the graphical user interface for testing can be adapted for different tasks.

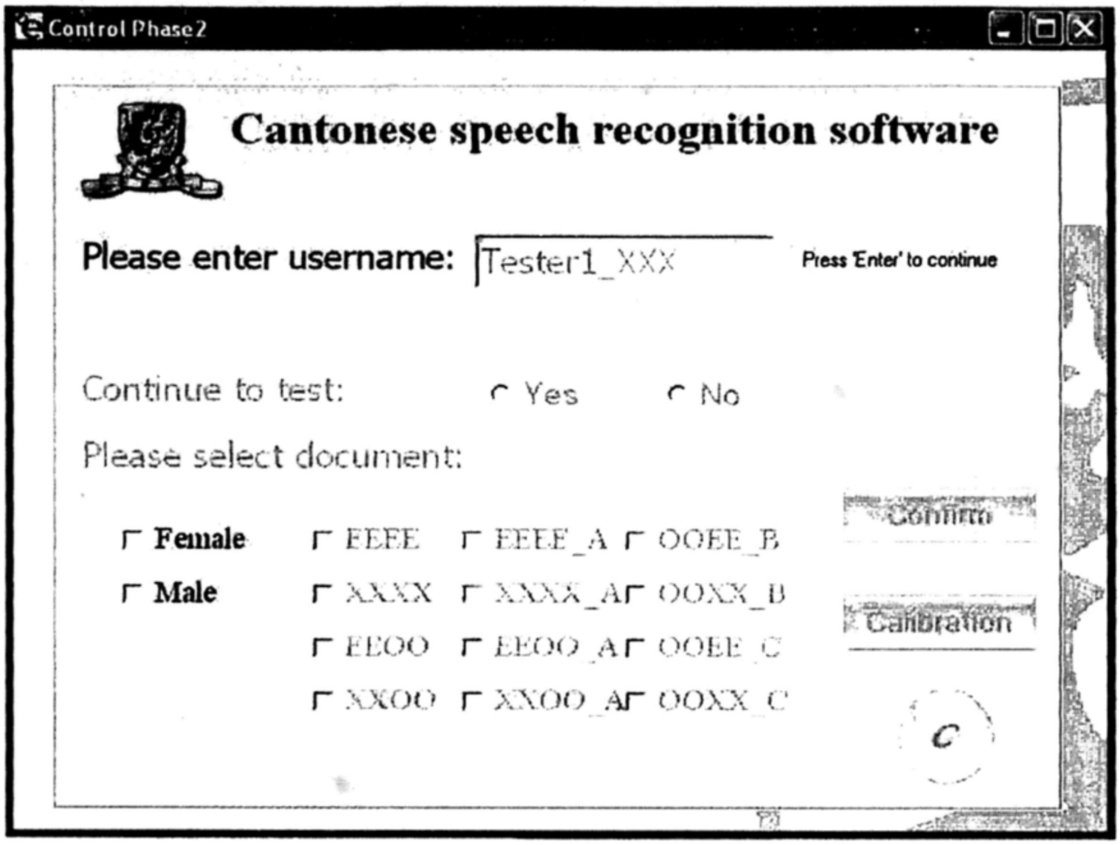
The use of PC-based software enables researchers to set up more complex test designs and to easily vary more parameters for a human controller. Moveover, the software driven experiments may provide high precision of stimulus presentation and the response collection. Due to the above advantages, psychophysical experiments are always controlled by software. Since the research targets are quite different from one to another, most of the software is specially requested and designed for one particular experiment. There are still some software packages appeared recently which can demonstrate both acoustical and electrical stimulation directly through CIs [Laneau et al., 2005; Fu, 2002; Vandali et al., 2005]. However, these software packages are not capable of handling all the specific requirements of a particular psychoacoustic test.

In our study, we developed a software package called ATOPEX (Application for TOne Perception EXperiments) which performs auditory experiments with acoustical stimuli for NH listeners. This software runs under the Windows operating system. The graphical user interface (GUI) can be easily set up for different tasks, including tone identification tests with monosyllabic and disyllabic words in Cantonese or Mandarin languages. The Cantonese disyllabic word set has been published as a standard Cantonese disyllabic tone test, called CANDILET-N (Computerized CANtonese DIsyllabic LExical Tone Identification Test in Noise).

3.4.1 Overview of the Software

For a particular study, the design of a psychoacoustic experiment can be divided into the following steps, based on the research hypothesis/expectation: (1) stimulus preparation, (2) determination of the experimental procedure, (3) subject selection, (4) running the real experiment, and (5) analysis of the collected response data.

For the first step, all stimuli are created offline with MATLAB program and are stored on hard disk in the computer. This allows the researcher to generate stimuli with very complex signal processing algorithms which may demand relatively large computation time. This also allows the researcher to verify the stimuli and their proposed processing ideas before running the experiment. ATOPEX was user-friendly and didn't require any programming skills for creating and setting up an experiment. Once the experimental procedures and parameter values are defined by the researcher and programmed in ATOPEX, the experiment controller is only required to follow the given description of the test procedure to run the experiment which is straightforward for understanding. Figure 3.5 shows an example of the control window for a Cantonese tone identification experiment with disyllabic words using ATOPEX. The experiment controller only needs to type in the subject's ID (username) and select the presentation order of the two speakers' speech stimuli which has been decided by the



Control Phase2

Cantonese speech recognition software

Please enter username: Tester1_XXX

Press 'Enter' to continue

Continue to test: ○ Yes ○ No

Please select document:

□ Female □ EEEE □ EEEE_A □ OOEE_B

□ Male □ XXXX □ XXXX_A □ OOXX_B

□ EEOO □ EEOO_A □ OOEE_C

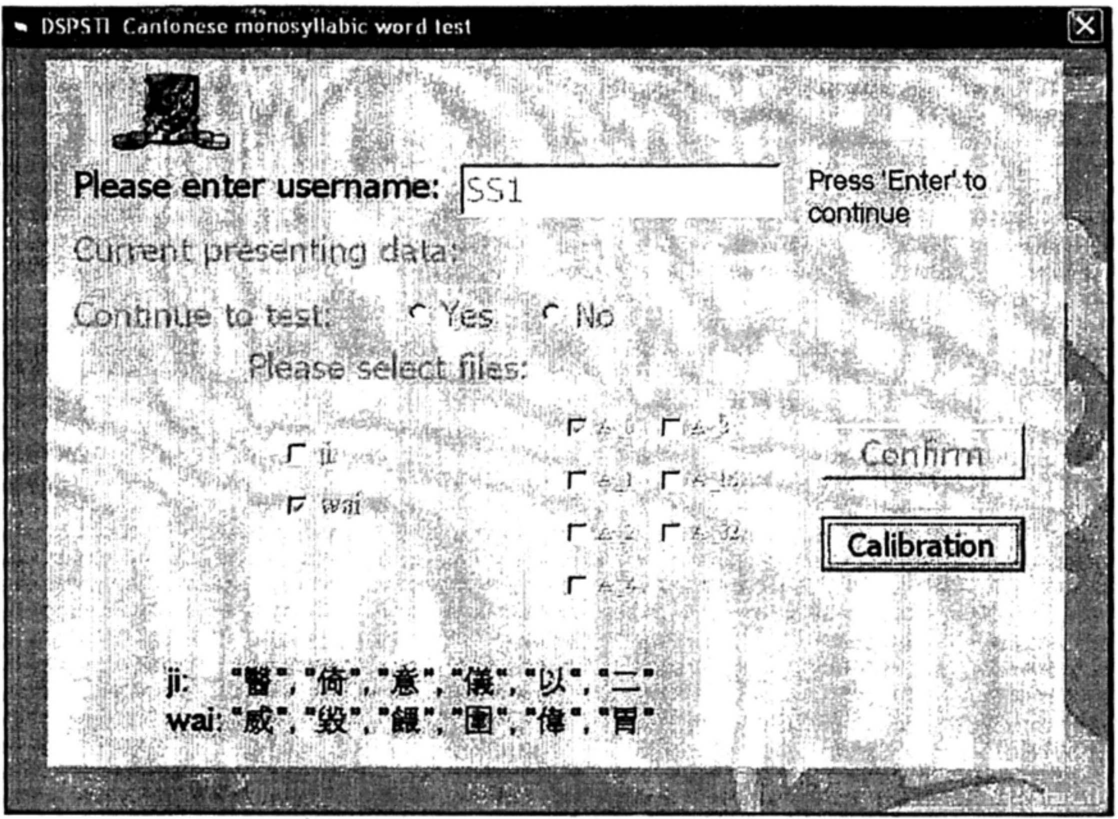
□ XXOO □ XXOO_A □ OOXX_C

Figure 3.5: Screenshot of an example of the control window for a Cantonese tone identification experiment using ATOPEX. The controller only needs to type in the username and select the presentation order of the two speakers' speech stimuli.

researcher. In this example, the presentation order of the processing conditions (e.g., 'EEEE', 'XXXX', and etc.) should be randomized which is automatically controlled by ATOPEX.

Figure 3.6 shows the first control window in the software for Cantonese monosyllabic words. First, the subject ID (username) was entered by the experiment controller. The presentation order for the two syllables (/ji/ and /wai/) and the seven test conditions were automatically randomized by the software for each subject. The items, A_0 to A_32 indicate the specific test conditions. Each time, the original unprocessed stimuli will be played first to make the subject familiarize with the stimuli and noise condition.

After the experiment controller loads in the required parameters and values in the control window, subjects can perform the experiment. Before the real test,



DSPSTI Cantonese monosyllabic word test

Please enter username: SS1

Press 'Enter' to continue

Current presenting data:

Continue to test: ○ Yes ○ No

Please select files:

□ ji

☑ wai

☑ [illegible] □ [illegible]

□ [illegible] □ [illegible]

□ [illegible] □ [illegible]

□ [illegible]

ji: "醫", "倚", "意", "儀", "以", "二"

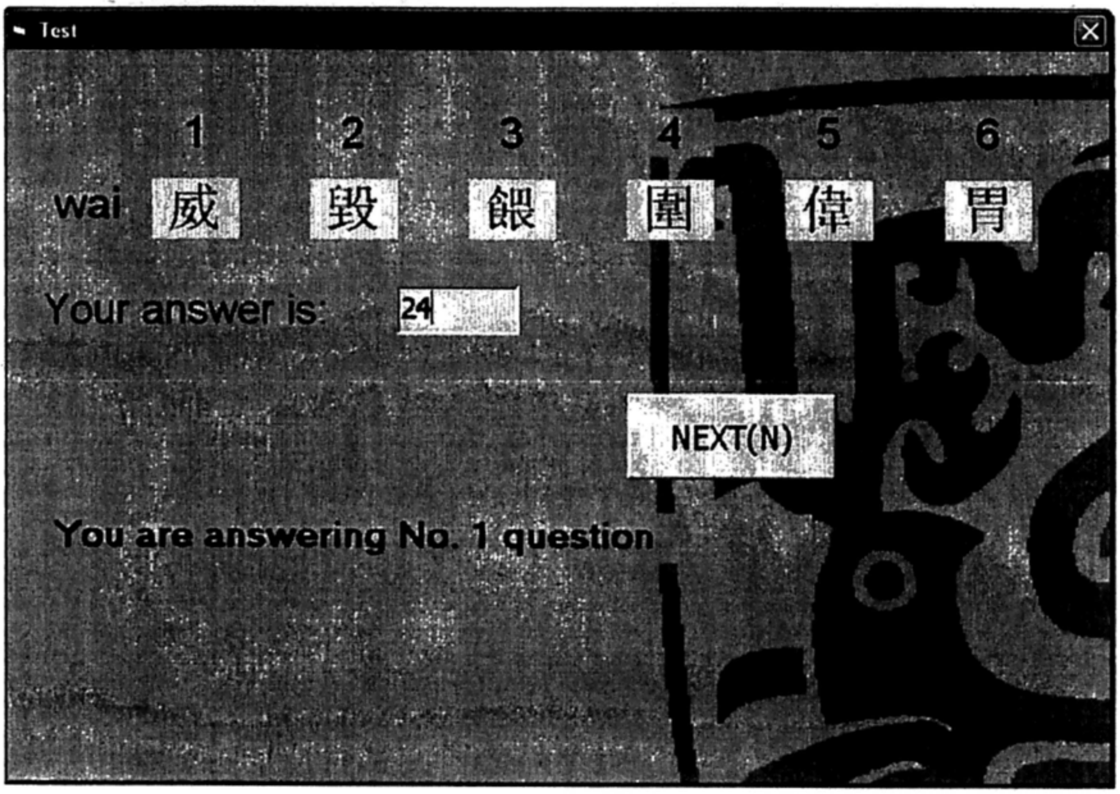
wai: "威", "毀", "餵", "圍", "偉", "胃"

Figure 3.6: Configuration window for tone test with monosyllabic words.

the experiment controller needs to press the 'Calibration' button to play some sounds produced by the speaker who whose speech sound was used for creating the real test stimuli. This process calibrates the frequency range of the tones presented by the speaker for the subject. When the subject gets familiarized with the tone range of the speaker, the real test begins. The subject sits in front of a graphical user interface and is presented with the sequence of trials. After each trial the subject enters his response. This can be done by clicking the possible answer using the PC mouse. Optional feedback can be provided after each response. The time interval between two trials can be fixed to a constant value or flexible to the response time of the subject. The display position of the four possible answers can be fixed or randomly assigned on the screen.

ATOPEX has different test windows for different tasks. For Cantonese monosyllabic word, the presentation and response window are shown in Figure 3.7. In this window, the monosyllable with six tones were presented in traditional Chinese characters. '1' to '6' represents the tone number. In each trial, the subject will hear two monosyllabic words sequentially. He/she is required to type the corresponding tone numbers in the text box besides the test 'Your answer is:'. Then the subject needs to click on the 'NEXT(N)' button using the mouse or press the N/n button on the keyboard to proceed to the next trial. A warning window will appear to alarm the subject to check his/her answer and input again if the subject only input one number or nothing, or the two inputs are not in the correct range (1 to 6). For Cantonese disyllabic word, the window is shown as in Figure 3.8. In this window, there are four squared buttons showing the traditional Chinese characters of one set of words in Table 3.2. The subject needs to click on one button after he/she heard the acoustic sound and then press N/n button on the keyboard to proceed. The corresponding disyllable and tone are illustrated besides the button in dotted boxes which are not shown to the subject during the test. Different from Cantonese, Mandarin monosyllabic word does not need contextual information for contrasting the tone. Only one word will be played in each trial. Figure 3.9 shows the presentation and response window for the tone identification test with Mandarin monosyllabic words. Since the concept of designing the Mandarin disyllabic word speech materials is the same as that for Cantonese, the interface for Mandarin disyllabic words is similar to that for Cantonese disyllabic words. The difference is that the type of Chinese characters presented in Mandarin tests is simplified Chinese, not traditional Chinese.

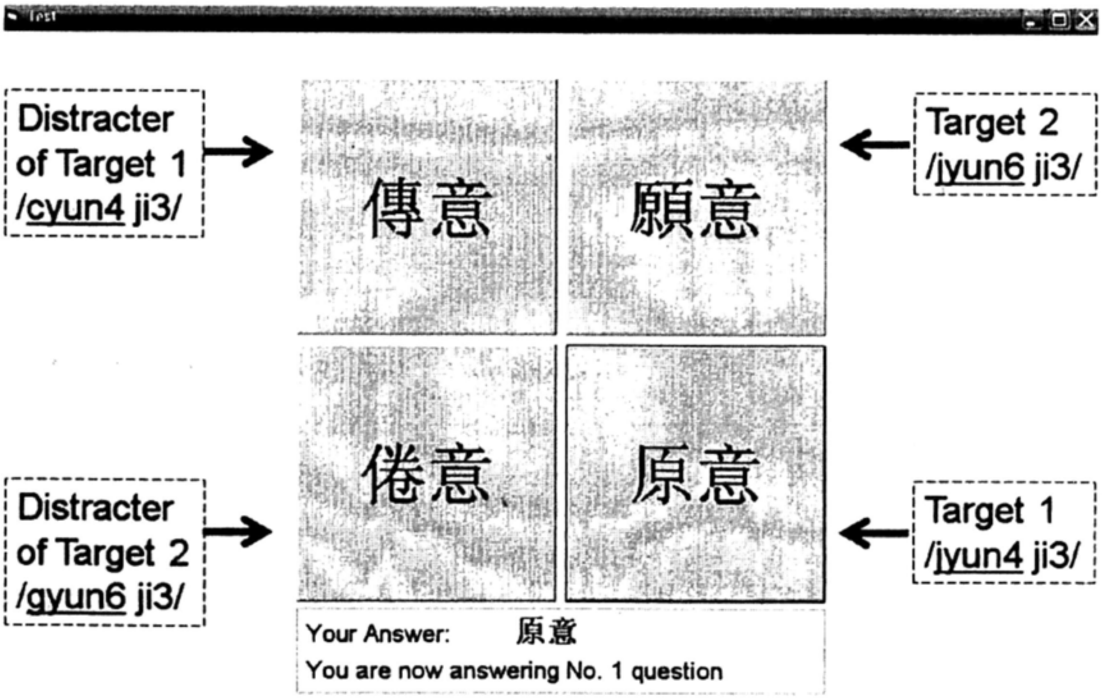
In order to permanently associate the experimental parameters with the obtained results, the program collects all responses and outputs them to a text file and names the text file with pre-defined format which should include the subject ID and test number information for individual subject. For the tone identification the program computes the correctness of each response in both tone and word recognition levels, and prints the results in the text file. These data can then easily be imported into statistical software packages for further analysis of the results.



Your answer is: 24

You are answering No. 1 question

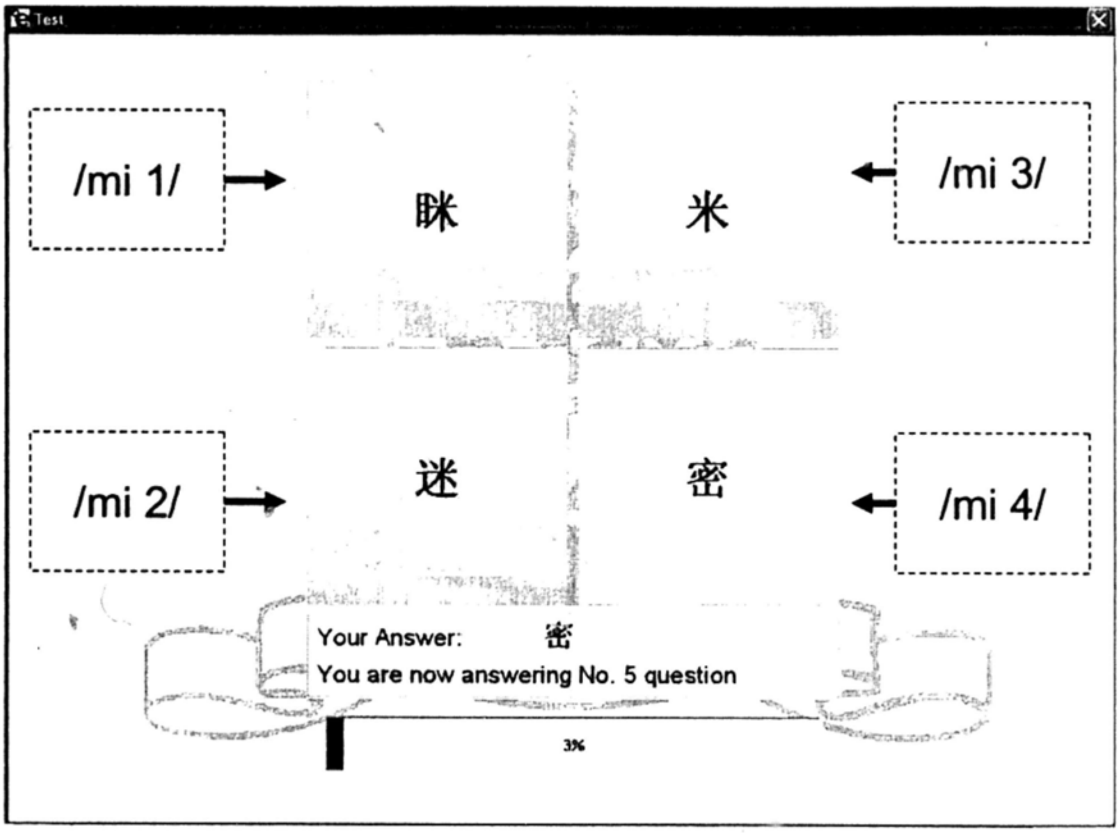
Figure 3.7: Presentation and response window for tone test with Cantonese monosyllabic words.


Test
Distracter of Target 1 /cyun4 ji3/
傳意
願意
Target 2 /jyun6 ji3/
Distracter of Target 2 /gyun6 ji3/
倦意
原意
Target 1 /jyun4 ji3/

Your Answer: 原意

You are now answering No. 1 question

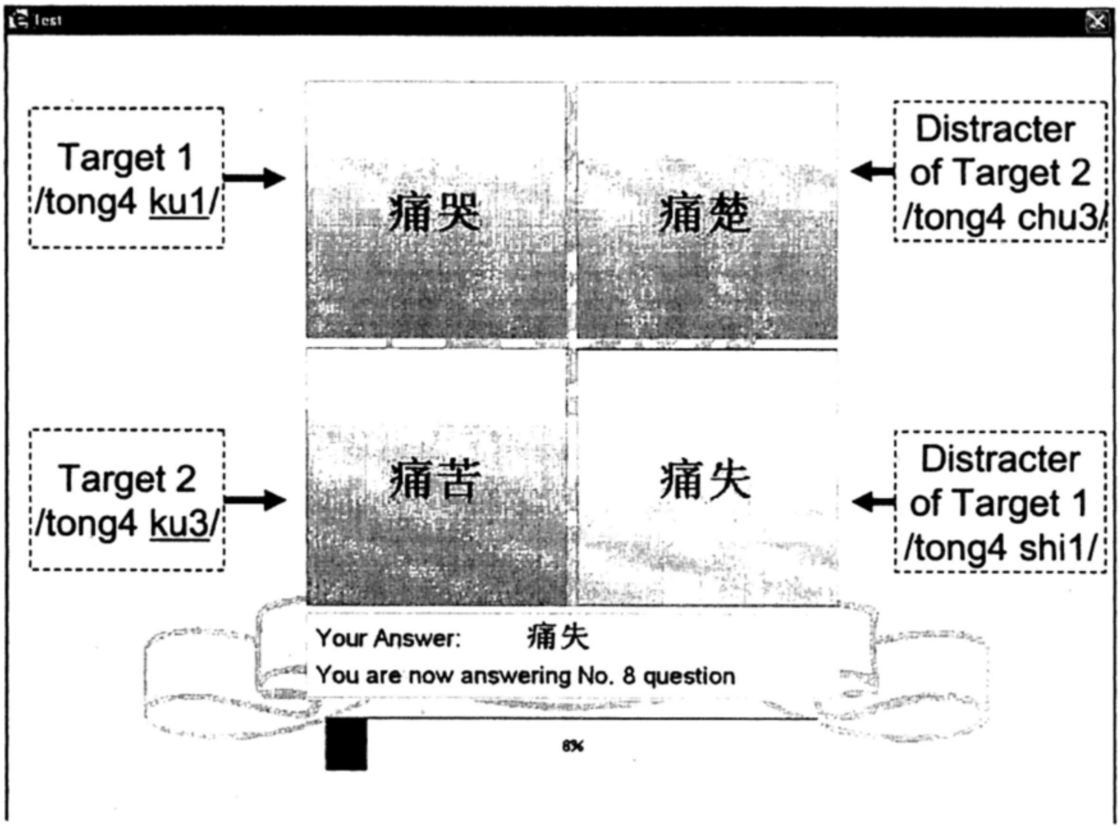
Figure 3.8: Presentation and response window for tone test with Cantonese disyllabic words.



Your Answer: 密

You are now answering No. 5 question

Figure 3.9: Presentation and response window for tone test with Mandarin monosyllabic words.



Target 1 /tong4 ku1/

Target 2 /tong4 ku3/

Distracter of Target 2 /tong4 chu3/

Distracter of Target 1 /tong4 shi1/

Your Answer: 痛失

You are now answering No. 8 question

Figure 3.10: Presentation and response window for tone test with Mandarin disyllabic words.

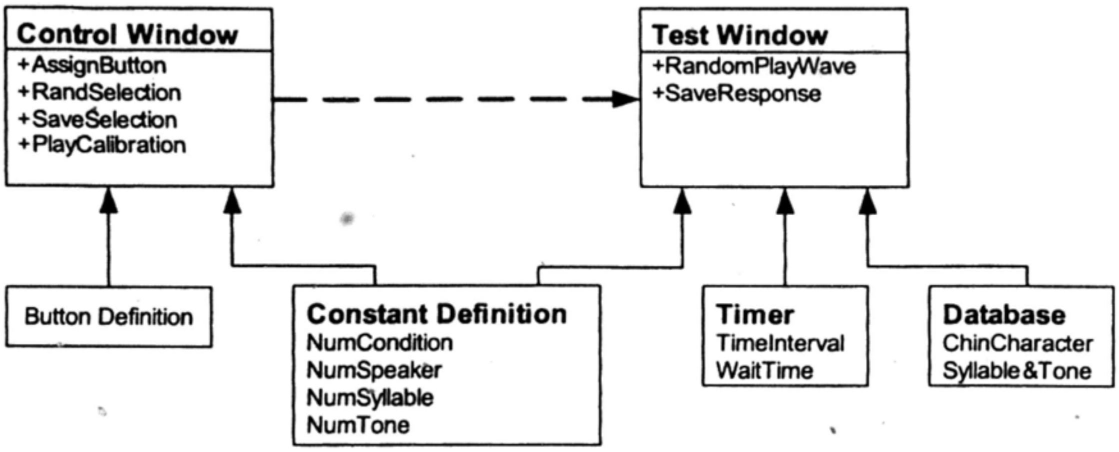


Figure 3.11: Internal structure of the ATOPEX source code.

3.4.2 Internal Structure

The program was designed in Microsoft Visual Basic Professional 6.0. The Visual Basic program is powerful of graphical user interface design and data management. Figure 3.11 shows the internal structure of the ATOPEX software source code. In the control window, the program first loads the button definition and constants, i.e., number of test conditions (NumCondition), number of speakers (NumSpeaker), number of syllables (NumSyllable), and number of tones (NumTone) for test. Then, the control window assigns the positions of the defined buttons and some textual blocks in the window. The researcher decides the type of parameter selection (automatically randomized, manually chosen, etc). The current and previous selected parameters are saved in a text file and prepared to be loaded in the test window source code. 'Calibrated' speech sounds produced by the same speaker who made the speech materials are played each time before the real test. The test window initially loads in the database which contains the test stimuli information, including the syllables and tones, together with the respective Chinese characters. In the test window, the program also loads in the timers which determines the time interval between two stimuli and the waiting time for the subject's response. The test stimuli in each trial are randomly played out via a loudspeaker/headphone and the subject's response will be automatically saved in another text file for further analysis.

3.5 Summary

This chapter first describes the design method of the speech test materials used in our study. Tone perception is not only another task of pitch discrimination. It is more related to linguistic aspects in tonal languages than in non-tonal languages. Therefore, speech material for lexical tone perception experiment should contain more segmental variations, e.g., using disyllabic words or sentences. To better control the contextual effects on tone perception, disyllabic words were used in our study. Due to the lexical constraints of Cantonese and Mandarin languages, it is hard to find a set of disyllabic words that have identical segmental structures but are minimally contrasted by the tone (six in Cantonese and four in Mandarin) carried by one syllable of the disyllabic word. Therefore, only a pair of contrasting tones were included in each set of words. Moreover, monosyllabic words were also implemented in our study similar to the conventional settings by other researchers.

In our study, a user-friendly research platform has been developed. It allows for psychoacoustic experiments simulating the process of CI devices. With different test stimuli, various experiments can be demonstrated including Cantonese and Mandarin tone identifications with disyllabic and monosyllabic words. It is flexible to adjust the presentation order depending on different experiment requirements in the software. However, it is still a specified software package which fits only the needs in our research.

□ **End of chapter.**

Chapter 4

Effects of Temporal and Spectral Cues on Lexical Tone Perception

Summary

Temporal cues play an important role in speech recognition, especially when the spectral cues are limited. Some studies showed that the contributions of temporal information from different frequency regions to speech recognition, i.e., consonant, vowel and sentence recognition are different. But very few studies have been done on tone recognition. The goal of this study is to evaluate the effect of frequency-specific TEPCs on tone recognition for tonal languages, i.e., Cantonese and Mandarin. We also aim to investigate the contribution of spectral resolution to tone perception in noise, by varying the number of frequency bands. Psychoacoustic experiments consistently show that: (i) TEPCs from high-frequency region are more important for tone perception; (ii) Cantonese tone perception accuracy will increase as the number of frequency bands increases (1 to 32 bands) at a signal-to-noise ratio of 10 dB. Four or more frequency bands are necessary to provide good tone perception.

4.1 Introduction

Previous chapters show that sounds are analyzed from different frequency regions in the inner ear according to the natural tonotopic organization of the cochlea. Despite the spectral approach to speech understanding, the temporal structure of speech is also important. A number of behavioral studies have shown that useful speech information is conveyed by the low-frequency amplitude fluctuations, i.e., the temporal envelope of speech, and that normal-hearing (NH) [Tasell et al., 1987, 1992; Shannon et al., 1995] and hearing-impaired (HI) listeners [Turner et al., 1995; Apoux et al., 2001] can use this information to recognize speech.

Recently, there have been a number of studies on the contribution of temporal envelopes extracted from different frequency regions to speech recognition. For instance, some researchers found that in quiet condition, different frequency regions contribute unequally for the recognition of consonant, vowel, and sentence [Shannon et al., 2001] while some researchers found the same contributions of the frequency regions for consonant recognition [Apoux and Bacon, 2004; Kasturi and Loizou, 2002]. Under noise conditions, it is found that the temporal information from high-frequency region is more important for consonant recognition [Apoux and Bacon, 2004]. Tone recognition, as a part of speech recognition, is important in the understanding of tonal languages. However, to our knowledge, the relative importance of the temporal cues from different frequency regions have not been investigated in tone perception tasks. In this chapter, we aim to investigate the effect of frequency-specific temporal cues to lexical tone identification of Chinese, i.e., Cantonese and Mandarin. Different from the recognition of non-tonal languages, temporal periodicity cues are found to be useful for providing tone-related information. Therefore, the temporal envelope and periodicity component (TEPC) is extracted from each frequency band.

The number of frequency channels controls the degree of the spectral details. The importance of spectral resolution for speech recognition has been investigated using the spectral smearing technique [Villchur, 1977; Baer and Moore, 1993, 1994; ter Keurs et al., 1992, 1993]. It is found in those studies that the speech recognition in quiet is hardly affected by spectral smearing. However, the speech recognition in noise was adversely affected, especially at low SNRs [Baer and Moore, 1993, 1994]. Shannon et al. [1995] used the vocoder technique and showed that four channels are sufficient to achieve good (i.e., $\geq 85\%$ correct) consonant, vowel, and sentence recognition of English in quiet. There is converged evidence that 4 – 16 channels are needed to achieve good speech recognition in English. The major factors are the difficulties of test tasks (e.g., vowel, consonant, or sentences) and listening conditions (e.g., quiet or noise) [Shannon et al., 1995, 2004; Dorman et al., 1997; Loizou, 1999; Zeng et al., 2005; Xu and Zheng, 2007]. The effect of spectral resolution on speech recognition is also investigated in tonal languages. Consistent with the observations in English, recognition performance keeps improving as the number of frequency bands increases for Mandrin phonemes and sentences [Fu et al., 1998b]. In a more recent work, Xu and Pfingst [2008] further showed that phoneme recognition in quiet reached a high level with 8 channels and more channels were needed under noise condition. For lexical tone recognition, larger number of channels is required [Xu et al., 2002; Xu and Pfingst, 2003; Luo and Fu, 2006; Lin et al., 2007]. In our study, the effect of spectral resolution on tone perception of Cantonese was evaluated in noise. Compared to Mandarin, Cantonese has one of the most complex tone systems in all languages. It is expected that more spectral cues would be required for the perception of Cantonese tones. The effect of noise is also investigated to show whether the temporal envelope cues are susceptible to the interference of background noise.

Two psychoacoustic experiments are reported in this chapter. In Experiment 1, test stimuli were created from the combinations of modulated noise carriers of different frequency bands. The test stimuli are monosyllabic and disyllabic words in Cantonese and Mandarin as described in Section 3.2. The purpose of this research is to investigate the contributions of TEPCs from different frequency regions to tone perception. In Experiment 2, the number of the frequency bands is varied from 1 to 32, providing different spectral resolution.

The speech materials are masked by background noise at a SNR of 0 dB. The purpose of this study is to evaluate the spectral resolution to tone perception in noise.

4.2 Experiment 1: TEPCs for Lexical Tone Identification

4.2.1 Experiment 1A: TEPCs for Cantonese Tone Identification

Materials and Methods

Subjects

Eighteen subjects (nine males and nine females) aged from 19 to 24 years participated in the tone identification test with monosyllabic words. Another ten subjects (five males and five females) attended the test with disyllabic words. All subjects had NH sensitivities with pure-tone air condition thresholds of 25dB HL or better in both ears at octave frequencies from 125 to 4000 Hz. All subjects are native Cantonese speakers with no reported history of ear diseases or hearing difficulties.

Speech Materials

Monosyllabic and disyllabic Cantonese words are used for lexical tone recognition in this test. The design of these speech materials was presented in Section 3.2.1.

Speech Processing and Stimuli

All of the speech materials were first low-pass filtered at 10 kHz. Their root mean square (RMS) intensity levels were equalized with A-weighting correction applied.

The test stimuli were generated with a four-channel noise-excited vocoder as described in Section 3.3. Different combinations of the four modulated noise bands were used to creat stimuli and the following experimental conditions are defined: (1) *ALL*, all four modulated noise bands are included; (2) *LOW*, with the two bands of 60 – 500 Hz and 500 – 1000 Hz; (3) *MID*, with the two bands of 500 – 1000 Hz and 1000 – 2000 Hz; (4) *HIGH*, with the two bands of 1 – 2 kHz and 2 – 4 kHz. Together with the original unprocessed signal (ORG), there were a total of five processing conditions.

Psychophysical Procedures

With Monosyllabic Words

For tone identification test with monosyllabic words, each subject participated in two test sessions, each lasting for 1.5 hour. In the first session, the subject was asked to fill in a research consent form followed by the air-conduction pure-tone audiometry assessment. Then all test stimuli from one of the two speakers were presented in the first session; the stimuli from the other speaker were presented in the second session. The test sequence of speakers was counter-balanced over all subjects.

Before the commencement of actual test, a test administrator introduced the six tones by presenting a response display containing the corresponding Chinese characters to the subject. The respective tone numbers (1 to 6) were shown next to the characters. The administrator spoke two words to the subject. The subject was asked to verbally repeat the words. The training procedure was repeated until the subject was familiar with the test requirement. This procedure is important for Cantonese speakers because they usually are not quite aware about the tone differences in their daily communication. This training procedure can make them pay more attention to the perception of tones, which is essentially important in our study.

The acoustic stimuli were presented via a loudspeaker. The subject was seated one meter away in front of the loudspeaker. A desktop computer with a graphic-user-interface (GUI) software was used to playback the test stimuli via the loudspeaker, which is connected with a GSI 10 audiometer. At the beginning, a noise signal with the same A-weighted RMS level as the monosyllable stimuli was used to calibrate a presentation level of 65 dBA at the subject's location. The calibration noise was generated from a white noise shaped by the average speech spectrum computed from all monosyllabic speech materials. All test sessions were conducted in a sound-proof room.

During each test session, the order for presenting the three blocks of base syllables was randomized. Within each block, the ORG condition was tested first, and then the other four conditions (ALL, LOW, MID and $HIGH$) were tested in randomized order. After finishing each test block, the administrator repeated the lexical tone training for the next base syllable.

With Disyllabic Words

During tone identification tests with disyllabic words, each subject was required to attend two sessions. Each session involved all test stimuli from one speaker. The presentation order of the two speakers was counter-balanced over all subjects. The NH subjects were seated in a sound-proof room and listened to the stimuli presented via a Paired E.A.R. Tone 3A Insert Earphone at 65 dBA. For each speaker, the unprocessed clean speech materials (ORG) were presented at the beginning to let the subject familiarize with the process and the test materials. Subsequently the four sets of processed stimuli (LOW, MID, $HIGH$ and ALL) were presented in randomized order. In each test set, the presentation order of the 120 words was randomized. No feedback was provided.

Results of Analysis

The test results were analyzed using the analysis of variance (ANOVA) method to investigate the main effects of different factors. *Post-hoc* tests were also applied to analyze within-factor differences.

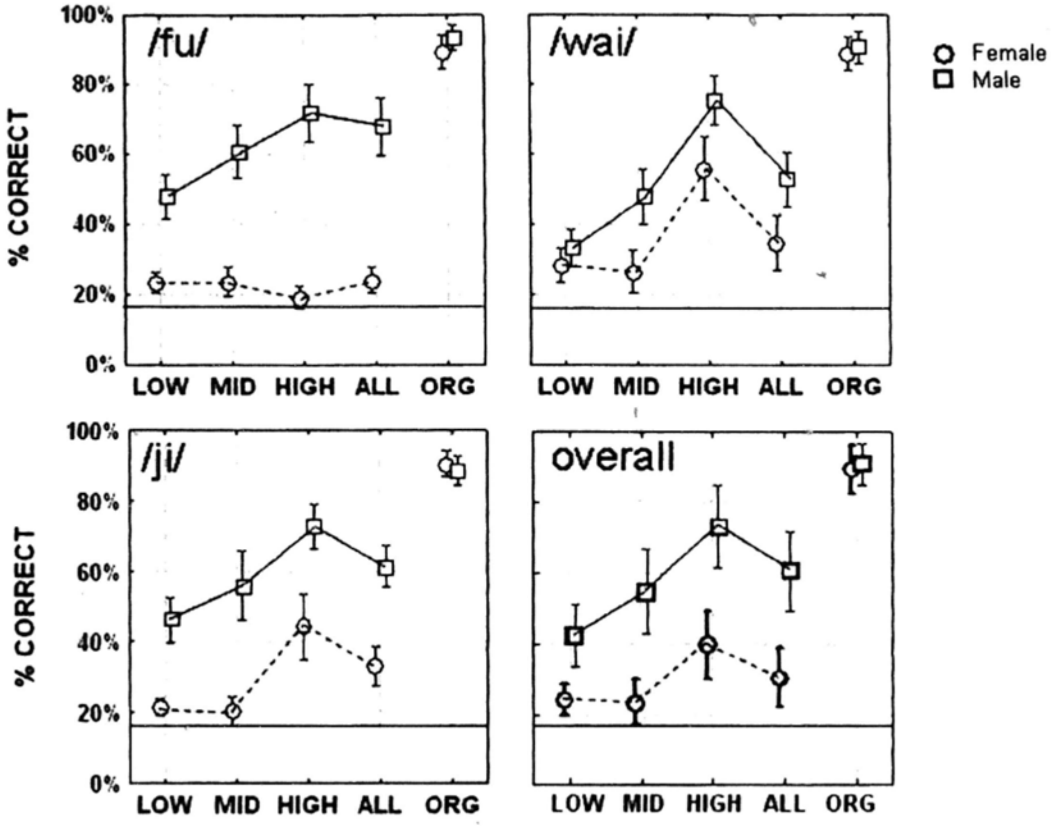


Figure 4.1: Tone identification accuracy of the five processing conditions with Cantonese monosyllabic words. The bottom right plot shows the overall results and the other three plots show the results breakdown from each phonetic composition. The error bars show the 95% confidence intervals of the means.

Results

Tests on Monosyllabic Words

Figure 4.1 gives the test results in terms of percentage correction of tone identification. The bottom right plot is the average result over the three base syllables. The results for individual base syllables are shown in the other plots. In each plot, the tone identification accuracies on male and female speech are shown separately. The error bars show the 95% confidence intervals of the means. The accuracy on male speech was better than that on female speech. The *LOW* processing condition always gave the worst performance. For natural unprocessed speech (*ORG*), the subjects could attain an identification accuracy of 90%.

A three-way repeated measures ANOVA was conducted to investigate the main effects from the factors of processing condition (*ALL*, *LOW*, *MID* and *HIGH*), base syllable (/fu/, /ji/ and /wai/), and speaker (male and female). The results revealed that there were significant main effects of processing condition [F (3,51) = 60.96, $p < 0.000001$] and speaker [F (1,17) = 128.62, $p < 0.000001$]. The factor of base syllable did not show a main effect [F (2,34) = 1.84, p = 0.17]. There were also significant effects for interactions of the three factors ($p < 0.005$).

Tukey *post-hoc* Honest Significant Different (HSD) test revealed that the performance on male speech was significantly better than on female speech ($p < 0.05$). The tone accuracy in *HIGH* condition was significantly higher than those in *LOW* and *MID*. There was no significant difference between *HIGH* and *ALL*. Looking into male and female speech separately, similar trends were observed except that *HIGH* was better than *ALL* for male speech ($p < 0.05$). For individual base syllable stimuli, the tone accuracy in *HIGH* condition was significantly higher than that in *LOW*. The only exception was noted on the base syllable /fu/ in female speech, in which the accuracies in all of the four conditions were very low (close to chance level). *HIGH* was significantly better than *ALL* for /ji/ and /wai/ for both male and female speech.

Tests on Disyllabic Words

Figure 4.2 shows test results with disyllabic words. The observations are similar to those on Figure 4.1. The *HIGH* condition was better than the other processing conditions and *LOW* was the worst. The accuracy on male voice was higher than those on female voice. Two-way ANOVA was used to analyze the main effects of the two factors: processing condition (*ALL*, *LOW*, *MID* and *HIGH*) and speaker (female and male). Both factors showed significant main effects. The interaction between them also showed significant main effect [F (4,36) = 541.9 , $p < 0.000001$]. This interaction reflected the fact that the difference in tone identification score between HIGH and other processing conditions was more obvious for female than for male speech. The high performance on male voice led to ceiling effect, which hindered further improvement on the *HIGH*

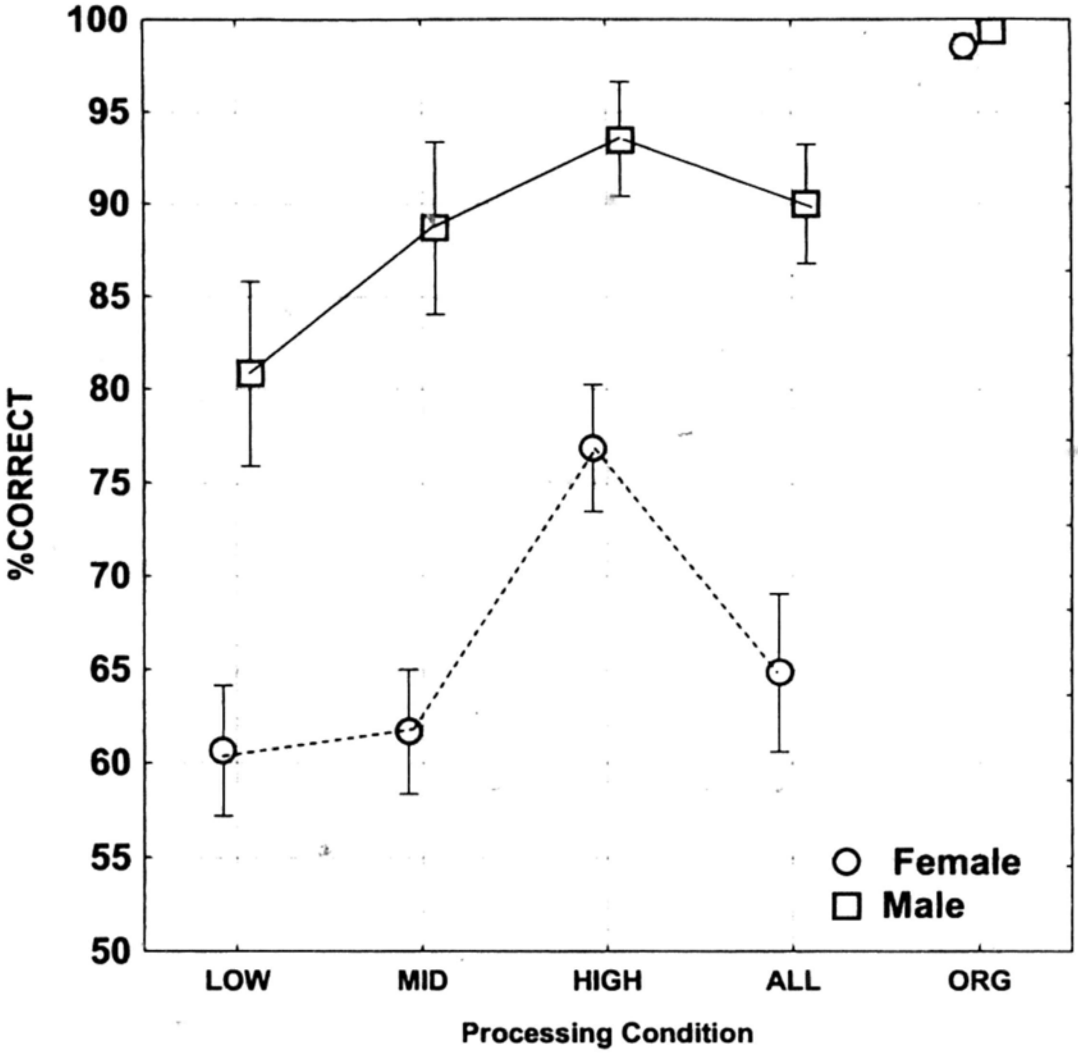


Figure 4.2: Tone identification accuracy of the five processing conditions with Cantonese disyllabic words. The error bars show the 95% confidence intervals of the means.

condition.

Tukey *post-hoc* HSD test showed that, the performance on male speech was significantly better than on female speech (p < 0.05). The tone accuracy in $HIGH$ condition was significantly higher than those in LOW and MID ($p <$ 0.05). For female speech, $HIGH$ was better than ALL ($p < 0.05$). A consistent trend was shown as following: $HIGH > ALL > MID > LOW$.

4.2.2 Experiment 1B: TEPCs for Mandarin Tone Identification

Materials and Methods

Subjects

Five male and five female subjects participated in the study. They are all native Mandarin speakers. Their ages are from 24 to 30. All subjects have normal hearing with pure-tone thresholds higher than 25 dB HL at octave frequencies from 125 to 4000 Hz in both ears.

Speech Materials

The speech materials consist of two parts: monosyllabic words and disyllabic words. Detailed information was given in Section 3.2.2.

Speech Processing and Stimuli

The signal processing method is exactly the same as the one used in the previous experiment on Cantonese. A four-channel noise-excited vocoder was implemented. Three processing conditions were generated with TEPC-modulated noise carriers: *LOW* (60 – 1000 Hz), *HIGH* (1 – 4 kHz) and *ALL* (60 – 4000 Hz), respectively. The *MID* condition (500 – 2000 Hz) was not included to reduce the duration of the test.

Psychophysical Procedure

The subjects were tested using the same equipments as those in Cantonese disyllabic word tests. Each subject was required to attend two sessions. Each session involved test stimuli from either monosyllabic or disyllabic word list, with the order counter-balanced over all subjects. The unprocessed clean speech materials were always presented at the beginning. Subsequently the three sets of processed stimuli were presented in randomized order. In each test set, the words were presented in randomized order without repetition. Tone identification tests were

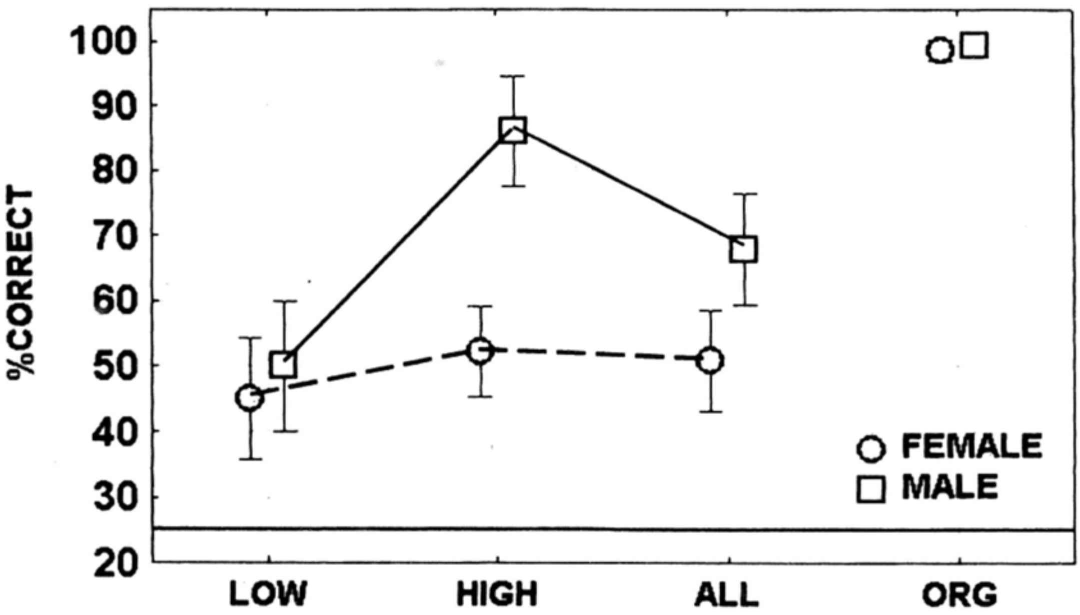


Figure 4.3: Tone identification accuracy with Mandarin monosyllabic words. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance levels are 25%.

conducted using a 4-alternative, forced-choice (4-AFC) procedure. No feedback was provided.

Results

Tests on Monosyllabic Words

Figure 4.3 shows the test results for monosyllabic words. The tone identification accuracies for male and female voices are plotted separately. The accuracies by the three processing conditions are compared. The original unprocessed speech condition (*ORG*) is included for reference. The tone accuracies in *HIGH* and *ALL* conditions were higher than those in *LOW*. *HIGH* was much better than *ALL* for male speech, while little difference between *HIGH* and *ALL* was observed for female speech. The accuracies on male speech were always higher than those on female speech under the same processing condition.

The results are analyzed using a two-way repeated-measures ANOVA with the factors of processing condition and speaker. The analysis showed significant main effects of both factors [F(3,27) = 114.12, $p < 0.0005$] and speaker [F(1,9) = 59.36, $p < 0.0005$]. A main effect on the interaction between the two factors was also significant [F(3,27) = 27.94, $p < 0.0005$], reflecting that the difference between $HIGH$ and ALL was more obvious for male speech than for female speech. Fisher LSD *post-hoc* tests showed that, for both female and male speech, the tone accuracies in $HIGH$ and ALL conditions were significantly higher than those in LOW ($p < 0.05$). For male speech, *post-hoc* comparison showed that the tone accuracy in $HIGH$ was significantly higher than ALL. For all processing conditions, the accuracies on male speech were higher than those on female speech. Significant difference between male and female speech was observed in $HIGH$ and ALL.

The average accuracies for individual tones are plotted in Figure 4.4. It was shown that the accuracies for Tone 1 and Tone 2 were lower than those for Tone 3 and Tone 4 in all processing conditions on female voice. For male voice, similar trends were observed except that the tone identification accuracies in $HIGH$ were similar across all tones.

Tests on Disyllabic Words

Figure 4.5 shows the test results with disyllabic words. Three similar trends were observed as in the monosyllabic word test. Additionally, the difference between tone accuracies in $HIGH$ and ALL and those in LOW was more noticeable for female speech.

ANOVAs revealed significant main effects of both processing condition [F(3,27) = 61.71, $p < 0.0005$] and speaker [F(1,9) = 13.69, $p < 0.005$]. The interaction also showed a significant main effect [F(3,27) = 4.44, $p < 0.05$]. LSD *post-hoc* tests showed that $HIGH$ and ALL were significantly better than LOW ($p < 0.05$) on both male and female speech. For male speech, $HIGH$ was significantly better than ALL. For all processing conditions, the accuracies on male speech were higher than those on female speech. However, only one significant difference between male and female speech was observed in $HIGH$ condition.

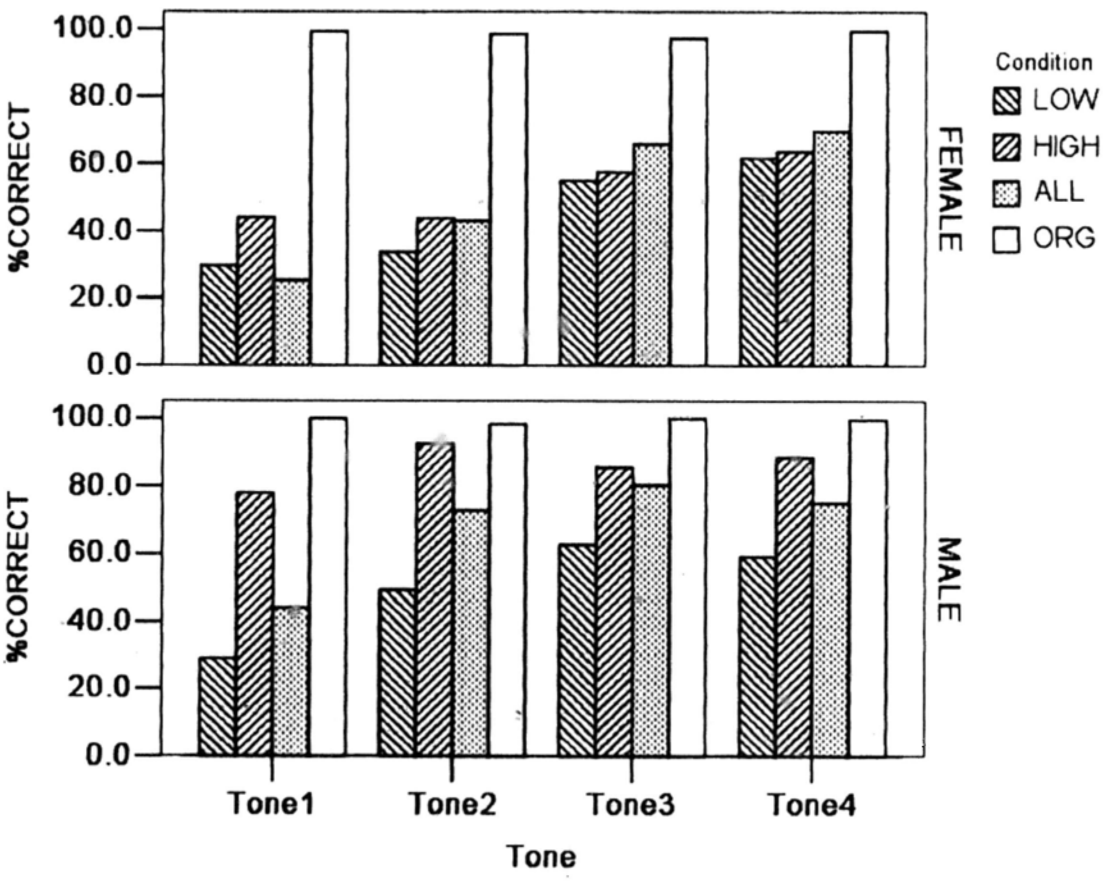


Figure 4.4: Percentage correct scores of the four tones with Mandarin monosyllabic words under different processing conditions.

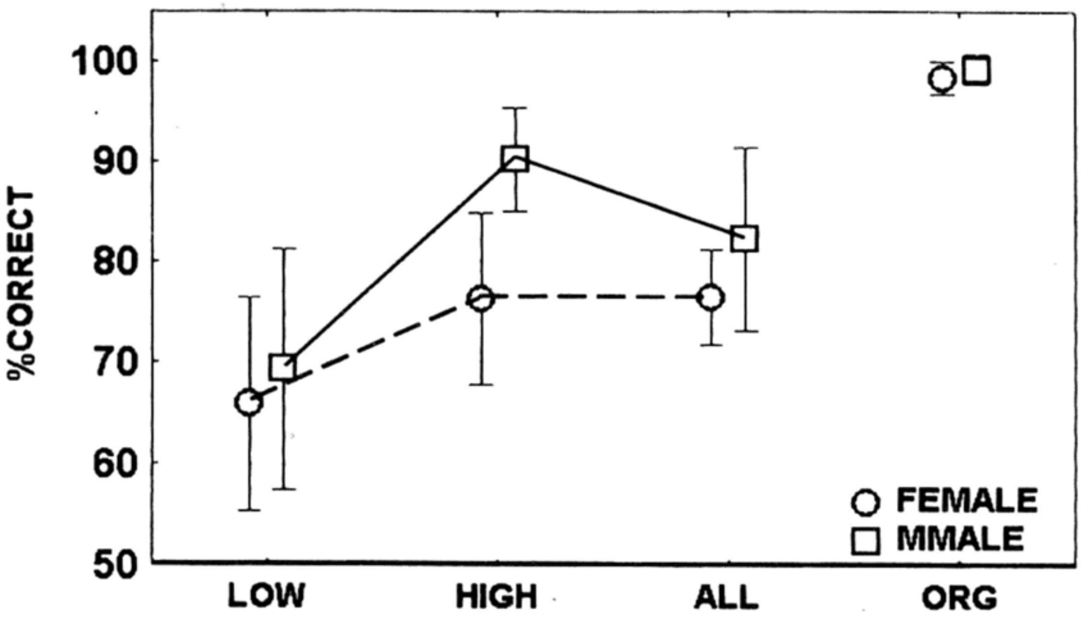


Figure 4.5: Tone identification accuracy with Mandarin disyllabic words. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance level is 50%.

4.2.3 Discussion

Our experiments showed some interesting results that are common in both Cantonese and Mandarin:

(1)Tone identification performance with TEPCs from high-frequency region is significantly better than that from low-frequency region. It is mainly related to two factors: the limitation in the peripheral filtering of normal ears, and spectrum overlapping between TEPC and the noise carrier in low-frequency region. More details will be discussed in the general discussion section in this chapter.

(2) Tone identification accuracies on male speech are significantly higher than those on female speech. It indicated that the relatively high $F0$ of modulation frequency in female speech might not be represented as well as the low $F0$ in male speech. Qin and Oxenham [2005] reported that pitch perception of TEPC was poorer with high $F0$ modulations than low $F0$ modulations. In their study, $F0$ difference limens for synthetic stimuli were worse for the ones

Table 4.1: Confusion matrix of Mandarin tone identification. The stimulus tones are presented in the y-axis. The response tones are presented in the x-axis.

Tone	1	2	3	4
1	67	14	5	14
2	11	75	6	8
3	2	10	82	6
4	9	4	4	83

with higher $F0$ (220Hz) than those with lower $F0$ (130Hz), especially in conditions where there were limited number of channels. Kohlrausch and Fassel [2000] also showed that modulation detection thresholds worsen with increasing modulation rate and suggested that the auditory system works like a low-pass filter in processing temporal information.

The Mandarin tone scores in all processing conditions were higher than those of Cantonese in the respective processing conditions, showing that Mandarin tone perception task is relatively easier than Cantonese tone perception. Cantonese tone system is complex which includes six tones. Tone level and tone contour are important for distinguishing the four level tones and two rising tones. Compared to Cantonese, Mandarin tone system contains four tones. Tone contour is the primary cue to distinguish different tones. Our experimental results also showed that tone identification performance was consistently much better for male voice than for female voice. The same results were observed in Cantonese tone identification tasks.

Table 4.1 and Table 4.2 showed the confusion matrices of tone identification of Mandarin and Cantonese monosyllabic words, respectively in the $HIGH$ condition. The total number of occurrences for each tone has been unified to 100 for ease of comparison. For Mandarin, different tones exhibit very different accuracy level. Tone 1 is the worst case. Tone 3 and 4 are the best recognized tones. Fu and Zeng investigated the contributions of temporal cues (amplitude, duration and periodicity) to Mandarin tone recognition [Fu and Zeng, 2000].

Table 4.2: Confusion matrix of Cantonese tone identification. The stimulus tones are presented in the y-axis. The response tones are presented in the x-axis.

Tone	1	2	3	4	5	6
1	85	2	6	3	2	2
2	2	78	5	4	9	2
3	6	4	75	4	4	7
4	3	5	6	78	3	5
5	2	19	7	5	63	4
6	3	5	14	8	8	62

Their results showed that the recognition accuracy for Tone 1 was the lowest among the four tones. They also found that syllable durations of Tone 3 and 4 were the longest and shortest respectively, and duration could be used as a cue to discriminate Tone 3 and 4 from other tones. In addition, they showed that the range of $F0$ variation in different tones were very different: Tone 4 had widest $F0$ range and Tone 1 had the narrowest range. The confusion matrix also showed that Tone 1 was easily confused with Tone 2 and Tone 4. For Cantonese, the confusion patterns can be categorized into two groups. Major confusions are seen in Table 4.2 between level tones that are close in pitch levels (Tone 1 and 3, Tone 3 and 4, Tone 3 and 6, and Tone 4 and 6), and between the two rising tones (Tone 2 and 5). The highest correct score was attained on Tone 1. This is a somewhat contradictory result to Mandarin since the perception of high-level tone in Mandarin was the worst.

4.3 Experiment 2: Number of Filter Banks to Cantonese Tone Identification

4.3.1 Materials and Methods

Subjects

Five male and five female Cantonese-speaking subjects participated in this study. Their ages are from 19 to 22. All of them have normal hearing with pure-tone thresholds of 25dB HL or better in both ears at octave frequencies from 125 to 4000 Hz.

Speech Materials

The speech materials contain monosyllabic and disyllabic words. The monosyllabic word stimuli are the two base syllables, /ji/ and /wai/. It was observed that these two syllables showed similar tone identification accuracy.

Noisy speech materials were generated by adding noise signals to clean speech at SNR of 0 dB. A noise signal was generated by shaping the spectrum of white noise according to the average spectrum of all test words spoken by the respective speaker. The speech materials were low-pass filtered with a cut-off frequency of 4 kHz. In this study, stimuli from clean speech materials were not tested in the consideration of long test period.

Experimental Conditions

To investigate the effect of number of frequency bands to tone perception, the original broadband speech materials were divided into 1, 2, 4, 8, 16, or 32 frequency bands using sixth-order elliptical bandpass filters. The cutoff frequencies of each band were determined by approximating equal cochlear distance for each band according to the Greenwood map [Greenwood, 1990]. In this map, the frequency-position function was defined as in Equation 4.1, where $A = 165.4$, $\alpha = 2.1$ and $k = 1$. The corner frequencies of these frequency banks are given in Appendix.

$$F = A(10^{\alpha x} - k) \tag{4.1}$$

Together with the original unprocessed noisy speech signal condition at 0 dB SNR (*ORG*), there were a total of seven experimental conditions: 1-band, 2-band, 4-band, 8-band, 16-band, 32-band, and *ORG*.

Psychophysical Procedure

The subjects attended the test in a sound-proof room. Each subject need to finish two sessions. One session contains the stimuli of monosyllabic words, while another session contains all the stimuli of disyllabic words. The order for these two sessions were balanced over all the subjects.

For both monosyllabic and disyllabic word tests, the experimental procedure is the same as in Section 4.2.1, except that the number of processing conditions was seven.

4.3.2 Results

Tests on Monosyllabic Words

Figure 4.6 shows the test results on monosyllabic words. The tone accuracies of the seven processing conditions are compared. The original unprocessed speech condition (*ORG*) is included for reference. It is clearly seen that the tone identification performance improves as the number of frequency bands increases.

The results were analyzed using a two-way repeated-measures ANOVA on two factors: spectral detail (number of frequency bands) and base syllable (/ji/ and /wai/). The analysis revealed significant main effects of both factors. Fisher LSD *post-hoc* tests showed that the tone scores attained from 1-band and 2-band conditions were significantly lower than the other conditions ($p < 0.05$). The 4-band, 8-band, and 16-band conditions showed similar tone scores (47.78%, 51.46% and 54.38%, respectively) ($p > 0.05$). Detailed spectral information,

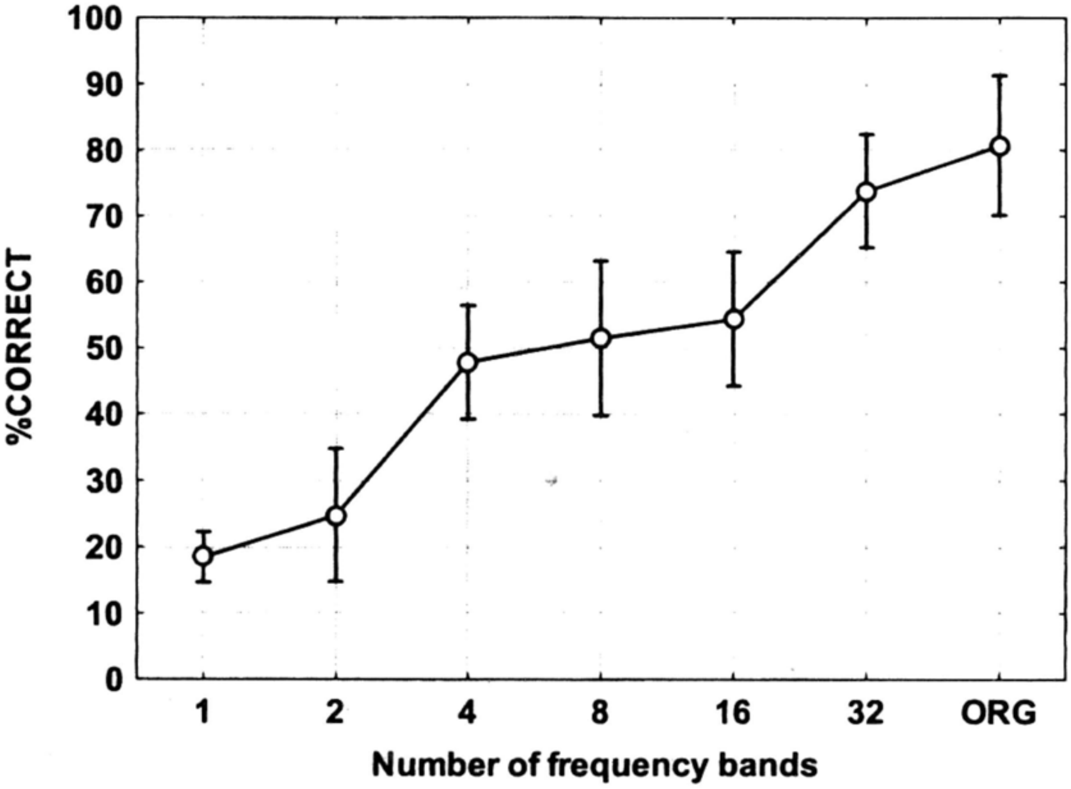


Figure 4.6: Tone identification accuracy with Cantonese monosyllabic words. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance level is 16.7%.

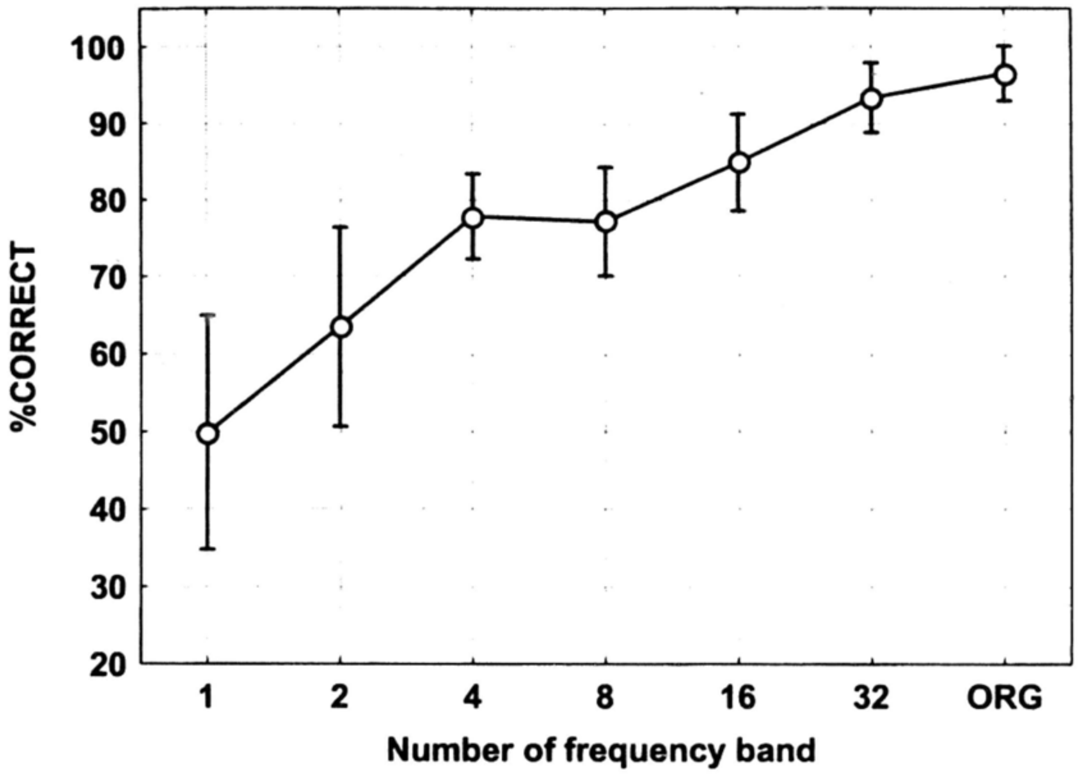


Figure 4.7: Tone identification accuracy with Cantonese disyllabic words. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance level is 50%.

with as many as 32 frequency bands, was required to produce tone recognition performance close to the original unprocessed speech (ORG). This agrees with previous study of Kong and Zeng [2006].

Tests on Disyllabic Words

Figure 4.7 shows the test results with disyllabic words. The trends are similar to those in monosyllabic words.

For disyllabic words, ANOVAs revealed significant main effects of spectral detail on the tone identification [F(6,54) = 40.89, $p < 0.00001$]. LSD *post-hoc* tests showed that the tone accuracy in the 1-band condition was significantly lower than the 2-band condition ($p < 0.05$). These two conditions performed significantly worse than the other conditions ($p < 0.05$). The *post-hoc* comparison showed no significant difference among 4-band, 8-band and 16-band conditions, as well as between 32-band and ORG ($p > 0.05$).

4.3.3 Discussion

This study investigated the effect of spectral resolution on tone perception in noise. The amount of spectral information wsa determined by the number of frequency bands. When there was no spectral detail available, i.e., 1-band condition, the subjects could achieve only 18.47% correctness of tone identification with monosyllabic word stimuli and 49.83% with disyllabic words. With increasing number of frequency bands, spectral details became richer and the tone identification performance improved. In our study, temporal fluctuations below 500 Hz were preserved in each frequency band. This covers the frequency range of $F0$ for tone perception purpose. However, the subjects could not perceive tone without spectral information (1-band condition). This indicated that the temporal periodicity cues can not transmit the tone information with the complete absence of spectral cues. As the number of channels increases, the temporal information in the envelope for each channel becomes restricted to narrower bands and the overall presentation of spectral-temporal information becomes more detailed. For example, Figure 4.8 and Figure 4.9 shows the waveforms and their spectrograms of the Cantonese syllable /ji/ in tone 1 to tone 6. The left most panel is the original unprocessed signal waveform. The following six panels were the stimuli waveforms with numbers of channels of 1, 2, 4, 8, 16 and 32. The spectrograms show that the $F0$ and its harmonics are absent in all tokens. With as few as 4 channels, the formants (i.e., F1, F2 and F3) can roughly be discerned. With 8 channels or more, the formants become clearer. Together with the temporal periodicity in each frequency band, the temporal-spectral patterns of the six tone tokens will be somewhat distinguishable. This observation is consistent with Xu and Pfingst [2008].

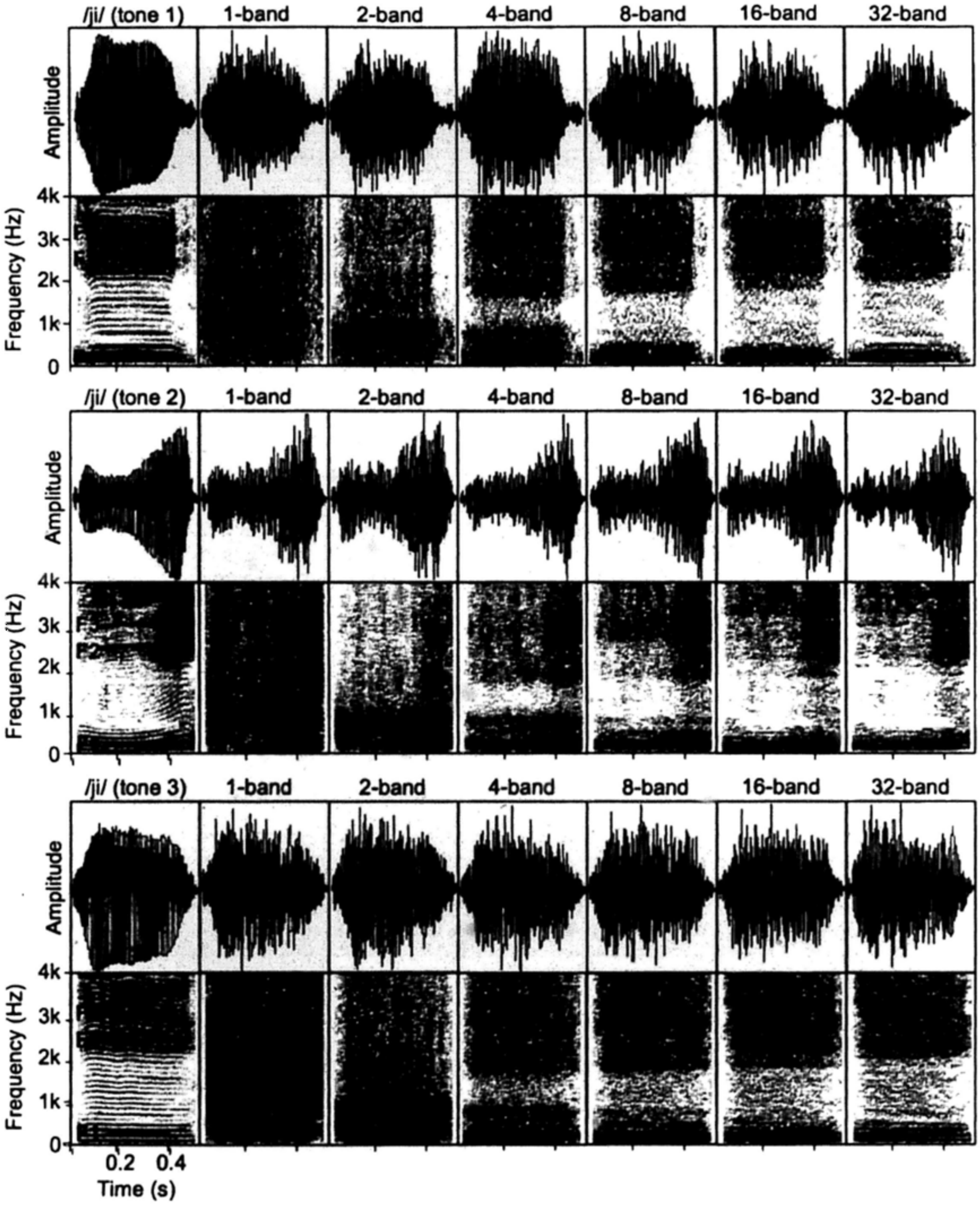


Figure 4.8: Time waveforms and the narrow-band spectrograms of vocoder processed Cantonese syllable /ji/ in tone 1–3 with numbers of channels of 1, 2, 4, 8, 16 and 32. The left most panel is the original unprocessed signal waveform. The three lines in this waveform indicate the three formants (F1, F2 and F3) extracted in the middle of the vowel of the original, unprocessed speech signal.

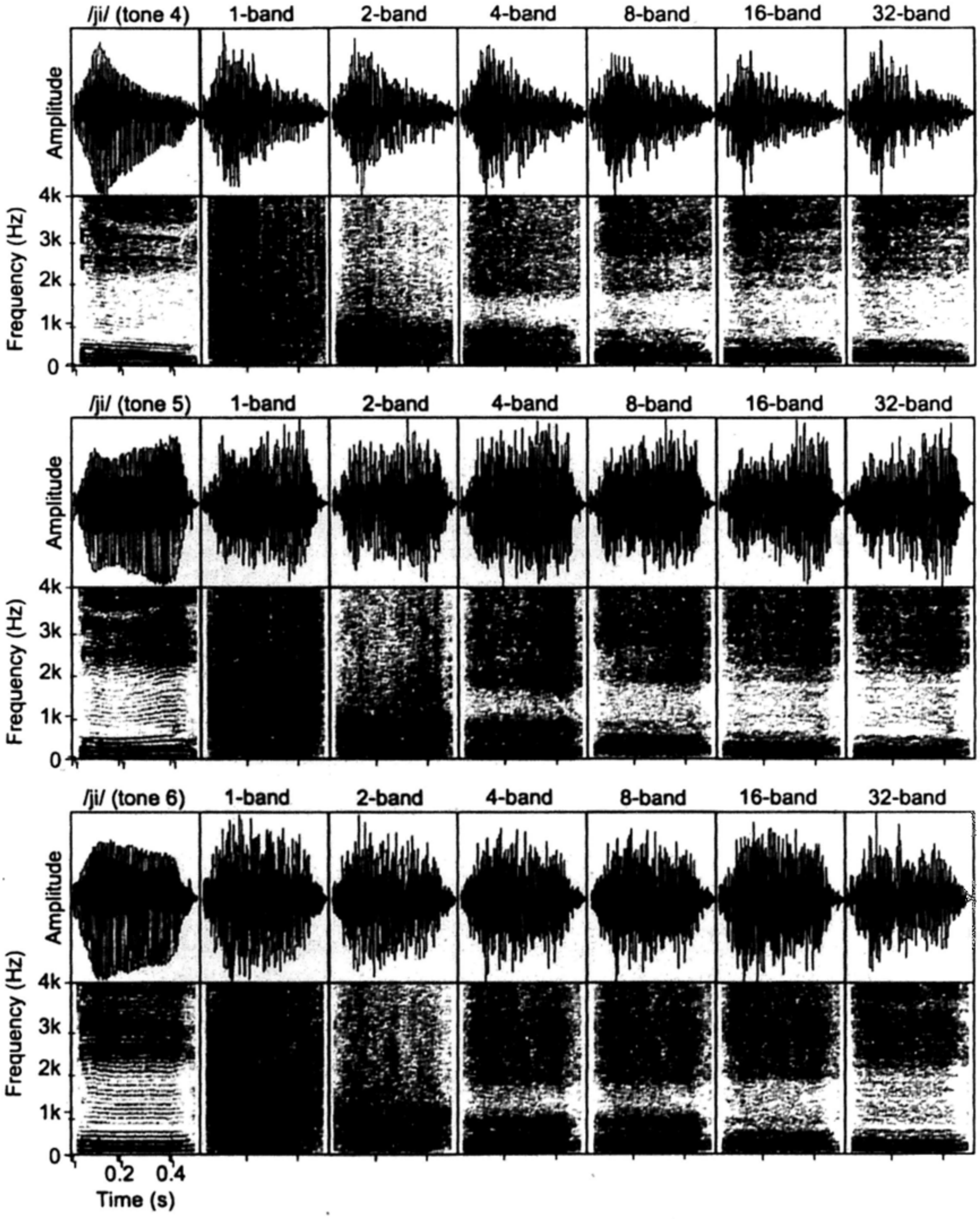


Figure 4.9: Time waveforms and the narrow-band spectrograms of vocoder processed Cantonese syllable /ji/ in tone 4–6 with numbers of channels of 1, 2, 4, 8, 16 and 32. The left most panel is the original unprocessed signal waveform. The three lines in this waveform indicate the three formants (F1, F2 and F3) extracted in the middle of the vowel of the original, unprocessed speech signal.

4.4 General Discussion

In state-of-the-art CI devices, temporal envelope cues are the primary speech information transmitted to the auditory nerves. Temporal cues that are delivered to CI patients are the relatively slow fluctuations in individual frequency channels. The availability of spectral information are mainly determined by the number and places of the electrodes (or frequency channels). In order to investigate effective temporal cues to speech recognition with flexibly-controlled parameters, a noise-excited vocoder is often used as an acoustic model for simulating the speech process in CI devices [Shannon et al., 1995; Laneau et al., 2006a]. In the vocoder, speech signal is split into a number of frequency sub-bands. In each sub-band, temporal envelope cue is extracted and used to modulate noise carrier within the same band. Acoustic output is the combination of the amplitude-modulated noise carriers. This output is presented to NH subjects in listening tests.

Temporal periodicity component (TPC) between 50 and 500 Hz was found to carry important information for tone perception when temporal fine-structure component (TFSC) is not available [Kong and Zeng, 2006; Fu et al., 1998b]. The test results from Experiment 1 agree with those from [Shannon et al., 2001; Apoux and Bacon, 2004]. That is, TEPCs in high-frequency region are critical for lexical tone recognition. There are two major factors that explain these different contributions. One factor is due to the limitation in the peripheral filtering of normal ears. Hanna [1992] showed that at low-frequency region the narrow bandwidth of peripheral filters limits the usefulness of the temporal pitch cues. In the noise-excited vocoder, the spectral region of the output signal is matched to the spectral region of the input signal. For the test stimuli in *LOW* condition, all signal energy is concentrated in the low-frequency region. For modulation rates above 100 Hz, modulation rate discrimination is limited by peripheral filtering. The basilar membrane starts oscillating at its characteristic frequency when excited with noise. In low frequency region, these oscillations can obscure the modulation of the noise band. Also, due to the limited bandwidth of the basilar membrane at low frequency region, the effective modulation depth of the F0-related modulations is reduced. Another major factor is due to spectrum overlapping between TEPC and the noise carrier in low-frequency region. The lowest band has a frequency range of 60 – 500 Hz, which overlaps with the range of fluctuation range of the extracted TEPC. The noise band may mask the TEPC and cause the poor performance of tone recognition. In the *MID* condition, the extracted TEPC does not have the same frequency range with noise carriers. However, the performance of tone recognition was significantly poorer than that in the *HIGH* condition. McKay et al. [1995] found that the carrier frequency in CI should be at least 4 times of $F0$. Otherwise pitch perception would be affected by the interference of the carrier. In our case, the carrier frequency should be at least 2000 Hz to avoid interference with the TEPC modulation frequency. In a related study [Yuen et al., 2007], a fixed high-frequency noise carrier ($>$ 1000 Hz) is used for TEPCs across all frequency bands. In this way, the effect of noise carrier's frequency is normalized. The test results showed that with the same noise carrier, the TEPC from high-frequency region was more useful than the other frequency regions. This indicated that periodicity information located in the high frequency TEPC is more important and salient for lexical tone perception. For NH people, pitch can be perceived from both the temporal cues and the spectral cues. As described in Section 2.2, temporal pitch is mainly derived from vibration of the unresolved harmonics in the high-frequency region. The temporal periodic fluctuations in these bands represent the fundamental frequency of the speech signal. Meanwhile, the low-order harmonics in the low-frequency region can be resolved and used for spectral pitch detection. However, in the noise-excited vocoder, the low-order harmonics carried in TEPCs from the low-frequency region are not resolvable due to the limited number of frequency bands (only two bands below 1000 Hz). Thus, NH subjects could not use spectral cues to detect pitch as they can do with original unprocessed speech. Their ability of using unresolved harmonics in low-frequency region is not as good as in high-frequency region. This is illustrated as in Figure 2.1. In low-frequency region, i.e., below 1000 Hz, the responses of basilar membrane correspond to the characteristic frequencies, i.e., the first four harmonics at 200 Hz, 400 Hz, 600 Hz and 800 Hz. The temporal periodicity is not salient compared to the high-frequency region. This reflects that the ability of using temporal envelope and periodicity cues to detect pitch in low-frequency region is not as good as that in high-frequency region.

Our observed results are consistent with previous studies. In Apoux and Bacon [2004], the relative importance of temporal information in broad spectral regions for consonant identification was assessed in NH listeners. The speech sounds were spectrally degraded using four-band noise-excited vocoder processing to make the listeners to use primarily temporal envelope cues. Their results showed that all bands contributed equally to consonant identification when presented in quiet. However, in noise condition, the listeners consistently placed relatively more weight upon the highest frequency band indicating that TEPC (< 500 Hz) in high frequency region (2.5 - 5 kHz) was more important for consonant identification. Lorenzi et al. [1999] suggested that the contribution of amplitude fluctuations in those high-frequency bands is more important to the intelligibility of the speech with reduced spectral information. They measured consonant identification under conditions of greatly reduced spectral information similar to the conditions tested in the present experiments. In addition, temporal modulations in the envelope of speech stimuli were preserved, degraded, or expanded. Their results showed a significant improvement in intelligibility when the temporal envelope from the high-frequency band was expanded. However, to our knowledge, there has been no investigation on the contributions of TEPCs from different frequency regions to tone perception. Our findings may have clinical implications for CI users who speak tonal languages. By emphasizing the TEPCs from high-frequency region, the lexical tone recognition may be improved such that overall performance of the speech recognition can be enhanced.

We also investigated the effect of the number of frequency bands to lexical tone recognition. With four or more frequency bands, over 75% correctness of tone identification can be achieved with disyllabic words in Cantonese. The tone recognition performance improves monotonically as the number of frequency bands increases. As many as 32 frequency bands, the performance became similar to the original unprocessed speech. Such observations were also found in other studies [Kong and Zeng, 2006; Fu et al., 2004a; Xu and Zheng, 2007; Xu et al., 2005; Green et al., 2002]. It indicates that the spectral resolution is important for both tonal language and non-tonal language recognition. The contribution of spectral cues was also investigated on CI users. Wei et al. [2004] evaluated the Mandarin tone recognition in CI listeners as a function of the number of electrodes. They found that implant listeners performed 57% correctness when listening with a 20-electrode map. The inability of cochlear-implant listeners to perceive lexical tones was also reported in Cantonese. Ciocca et al. [2002] studied a group of early-deafened Cantonese-speaking cochlear-implant children. Very few patients performed above chance in a tone identification task.

In Experiment 2, the speech signals were at a SNR of 0 dB. Compared to the test results in Experiment 1A, the tone identification accuracies at 0 dB SNR were 10% lower than in quiet. Friesen et al. [2001] found that a larger number of channels was required in noise conditions for CI users to achieve performance equivalent to that in quiet conditions. Similar results were obtained in Fu et al. [1998a]; Dorman et al. [1998]. Wei et al. [2004] tested Mandarin tone recognition with CI subjects and found that the performance could not be further improved with more than 7 electrodes. Similar observations were achieved for English and German speech perception [Fishman et al., 1997; Friesen et al., 2001; Garnham et al., 2002]. In the following chapters, we mainly focus on the effect of temporal cues to tone perception. The number of frequency bands is fixed at 4 for ease.

4.5 Conclusion

Lexical tone identification was measured in multi-channel noise-excited vocoder. The contribution of TEPCs from different frequency regions and the effect of number of frequency bands were evaluated on the tone identification in Cantonese and Mandarin. The results from this chapter can be summarized as follows:

1. Temporal envelope and periodicity components (TEPCs) play an important role for lexical tone identification in tonal languages, i.e. Cantonese and Mandarin when the spectral cues are limited. This is consistent with other studies showing similar observations on tonal languages.

2. TEPCs from high-frequency region (> 1000 Hz) contain more important information for tone perception in tonal languages. This performance agrees with the results obtained in other studies for consonant, vowel and sentence recognition.

3. The spectral resolution contributes to the tone perception, especially in noise. The tone identification monotonically improved as increasing the number of frequency bands. Relatively good recognition performance can be achieved with the number of frequency bands not less than 4. As the number of bands increased to 32, the recognition performance is close to the original unprocessed speech.

□ **End of chapter.**

Chapter 5

TEPC Expansion for Tone Perception

Summary

CI patients are found to have poor tone perception ability for tonal languages, e.g., Cantonese. It implicates that the commercially available signal processing strategies are not efficient in providing tone-related information to the subjects. This chapter describes a study on the effectiveness of expanding the TEPCs for Cantonese tone perception. The ultimate goal is to develop speech processing techniques that can improve speech perception of hearing prosthesis users. Psychophysical listening tests on Cantonese tone identification are carried out with expanded and unexpanded TEPCs. Based on the conclusion obtained in the previous chapter that TEPCs from high-frequency region are more important for tone recognition, the expansion is applied to the TEPCs from high-frequency region. The experimental results show that: (i) expansion of TEPC leads to noticeable improvement on tone identification accuracy; (ii) the effectiveness of TEPC expansion is more significant for female voice than male voice.

5.1 Introduction

A number of studies have shown that temporal envelope - the slow-varying amplitude fluctuations of speech sounds - plays an important role in speech perception when spectral cues are not available in NH listeners [Shannon et al., 1995; Smith et al., 2002] and listeners with cochlear damage [Apoux et al., 2001; Turner et al., 1995].

This finding has led to the idea that speech recognition may be improved or restored by enhancing (i.e., expanding) these temporal envelope cues, or in other words, by artificially increasing the modulation depth of the speech envelope. A number of studies have investigated the effects of temporal envelope expansion on speech recognition in quiet and noise conditions. Different implementations of the expansion scheme have been considered (power-law, envelope thresholding, compression-expansion, etc.). However, these studies gave conflicting results. Clarkson and Bahgat [1991] showed a very small benefit from expansion (+6%) in NH listeners on word recognition. Fu and Shannon [1999]; Lorenzi et al. [1999]; Apoux et al. [2001] reported greater improvements in phoneme recognition in both NH and HI listeners. Contrarily, no effects or even detrimental effects of envelope amplitude expansion in both NH and HI listeners were shown in Freyman and Nerbonne [1996]; van Buuren et al. [1999]; Apoux et al. [2004].

Part of the inconsistent results of the above studies may arise from the use of different expansion schemes, speech stimuli (phonemes, syllables, and sentences), and psychoacoustic paradigms (measurement of speech reception thresholds, reaction times, etc.). A critical concern is about the different frequency range of envelope fluctuations to expand. The studies that showed more positive results used higher cut-off frequencies for the envelope, such as 160 Hz [Fu and Shannon, 1999] or 500 Hz [Apoux et al., 2001; Lorenzi et al., 1999].

The goal of the present study was to further investigate the effects of temporal envelope expansion on tone perception. To our knowledge, no one has investigated such a research topic. According to our previous studies, the temporal amplitude fluctuations below 500 Hz from high-frequency bands are expanded.

Cantonese tone identification task is carried out. We use a noise-excited vocoder which extracts TEPCs from different sub-bands of the input speech signal and re-synthesizes the output signal by modulating random noise with these TEPCs [Fu et al., 1998b; Xu et al., 2002]. In our work, the sub-band TEPCs undergo a nonlinear expansion process such that their periodicity is strengthened.

5.2 Materials and Methods

5.2.1 Subjects

Five male and five female subjects aged from 19 to 24 years participated in the perceptual experiments. All of them are normal-hearing people with pure-tone air condition thresholds of 25 dB HL or better in both ears at octave frequencies from 125 to 4000 Hz. All subjects are native Cantonese speakers with no reported history of ear diseases or hearing difficulties.

5.2.2 Speech Materials

The Cantonese syllables /ji/ and /wai/ were used as the base syllables.

5.2.3 Speech Processing and Stimuli

In this study, the periodicity-enhancement method is compared with the standard CIS strategy as shown in Section 3.3.

The TEPC expansion is a non-linear amplification process that attempts to make the periodicity more salient. First of all, the envelope of TEPC is determined. It describes the variation that is slower than the pitch periodicity. This is done by applying discrete wavelet transform (DWT) analysis on the TEPC. A $7th$-order Daubechies wavelet is applied. The DWT analysis consists two processing steps, decomposition and reconstruction. In the decomposition step, the coarser coefficient A_7 and the detail coefficients D_1 to D_7 are derived. In the reconstruction step, only the coarser coefficient A_7 is retained, from which the TEPC envelope can be constructed. The TEPC envelope is used as a threshold for the expansion. That is, a sample on the TEPC is amplified only if the sample value is higher than the threshold at the respective time instant, i.e.,

$$p[k] = \begin{cases} \alpha \cdot (x[k] - e[k]) + e[k] & if \quad x[k] \geq e[k] \\ x[k] \quad if \quad x[k] < e[k] \end{cases} \tag{5.1}$$

where $x[k]$ denotes the TEPC, $e[k]$ is the expansion threshold, α is the expansion factor that determines the degree of expansion, and $p[k]$ is the expanded TEPC.

We use a synthesized speech segment to demonstrate the effect of TEPC expansion. Figure 5.1 shows the TEPC extracted from the second sub-band (500 – 1000 Hz) of a synthesized segment of vowel 'a' with the $F0$ value of 110 Hz. The upper pane shows the original TEPC and the lower one is the expanded TEPC (amplitude-normalized) with $\alpha = 100$. The measured period of this speech segment is equal to exactly $1/F0$. After expansion, the modulation depth of the periodic signal is increased.

Figure 5.2 shows the frequency spectra of the TEPC signals. Before expansion, the TEPC contains the $F0$ and the first harmonic. The magnitude of the first harmonic component is 25 dB lower than that of the fundamental frequency component. After expansion, the difference is reduced to about 6 dB and a number of higher harmonics are introduced.

In our previous work, it was found that TEPC from the two higher bands (1 – 4 kHz) leads to significantly higher accuracy of Cantonese lexical tone identification than that of the two lower bands (60 – 1000 Hz). Accordingly we limit the current study to the comparison between a single high-frequency band (denoted by HIGH) and a single low-frequency band (denoted by LOW), with 1 kHz being the boundary. Speech stimuli were created for the following experimental conditions:

- **HIGH_STD** - Only the high-frequency unexpanded TEPC is presented. No signal is included for the low-frequency band.
- **HIGH_EXP** - Similar to HIGH_STD, except that the TEPC is expanded.

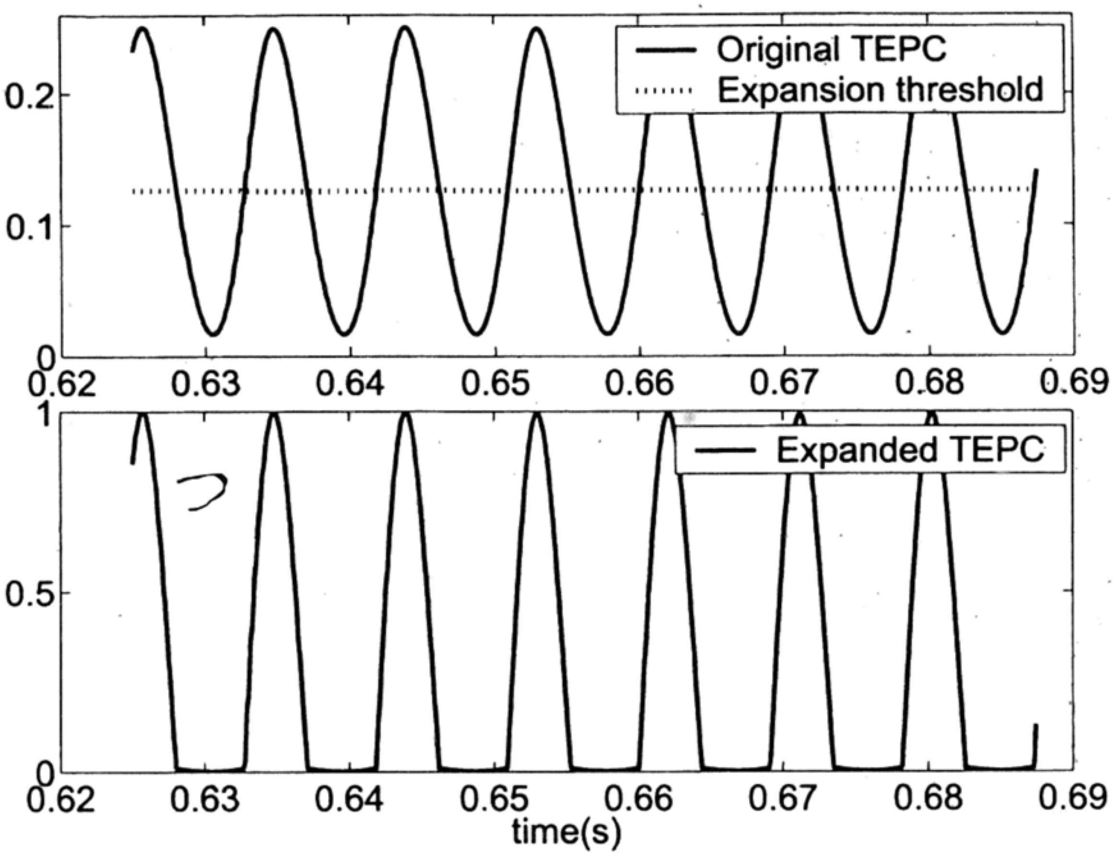


Figure 5.1: TEPC expansion effect for synthesized vowel /a/ with $F_0 = 110$ Hz in the 2^{nd} sub-band.

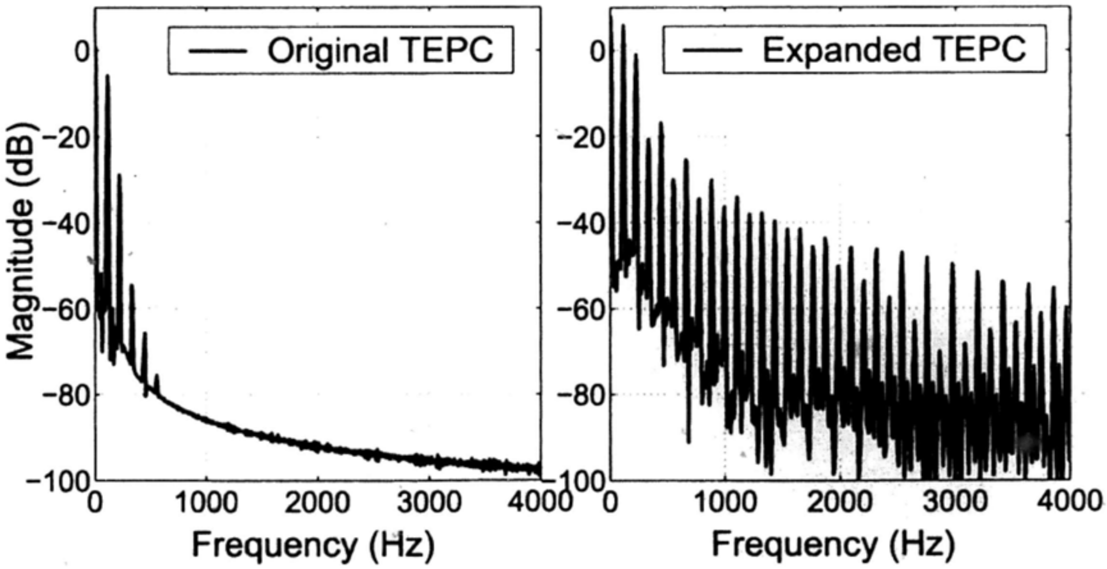


Figure 5.2: Spectrum of TEPC with the expansion effect (left pane: original TEPC, right pane: expanded TEPC).

- **ALL_STD** - Both low-frequency and high-frequency TEPCs are included and they are not expanded.

- **ALL_EXP** - Similar to ALL_STD, except the high-frequency TEPC is expanded.

In addition to the above conditions, the original speech was also used as a reference condition. The expansion factor α is fixed to 100 throughout the test based on a pilot experiment.

Each time a subject is presented with a pair of stimuli that carry the same base syllable. As there are six different tones in Cantonese, there are 36 presentations prepared for each of /ji/ and /wai/. In total, we had 72 syllable-pair stimuli under each experimental condition.

5.2.4 Psychophysical Procedure

A high-quality loudspeaker was used to present the signals at a presentation level of 65 dBA. The psychophysical procedure for this experiment is the same as the one described in Section 4.2.1. But this test contains different processing conditions as mentioned in the previous section.

5.3 Results

Figure 5.3 shows the tone identification accuracy averaged over all subjects under different conditions. The dash line indicates the chance-level correctness. A repeated ANOVA was performed to investigate the main effects of three factors: speaker (male, female), base syllable (/ji/, /wai/) and processing condition (HIGH_STD, HIGH_EXP, ALL_STD, ALL_EXP). The results revealed that all factors significantly affect the recognition performance: speaker ($p < .001$), base syllable ($p < .05$) and processing condition ($p < .0001$). Noticeable performance differences were also seen in the interaction term of speaker and processing condition ($p < .0001$). This reflected that the contributions of the processing conditions vary between different speakers. Special care should be taken when

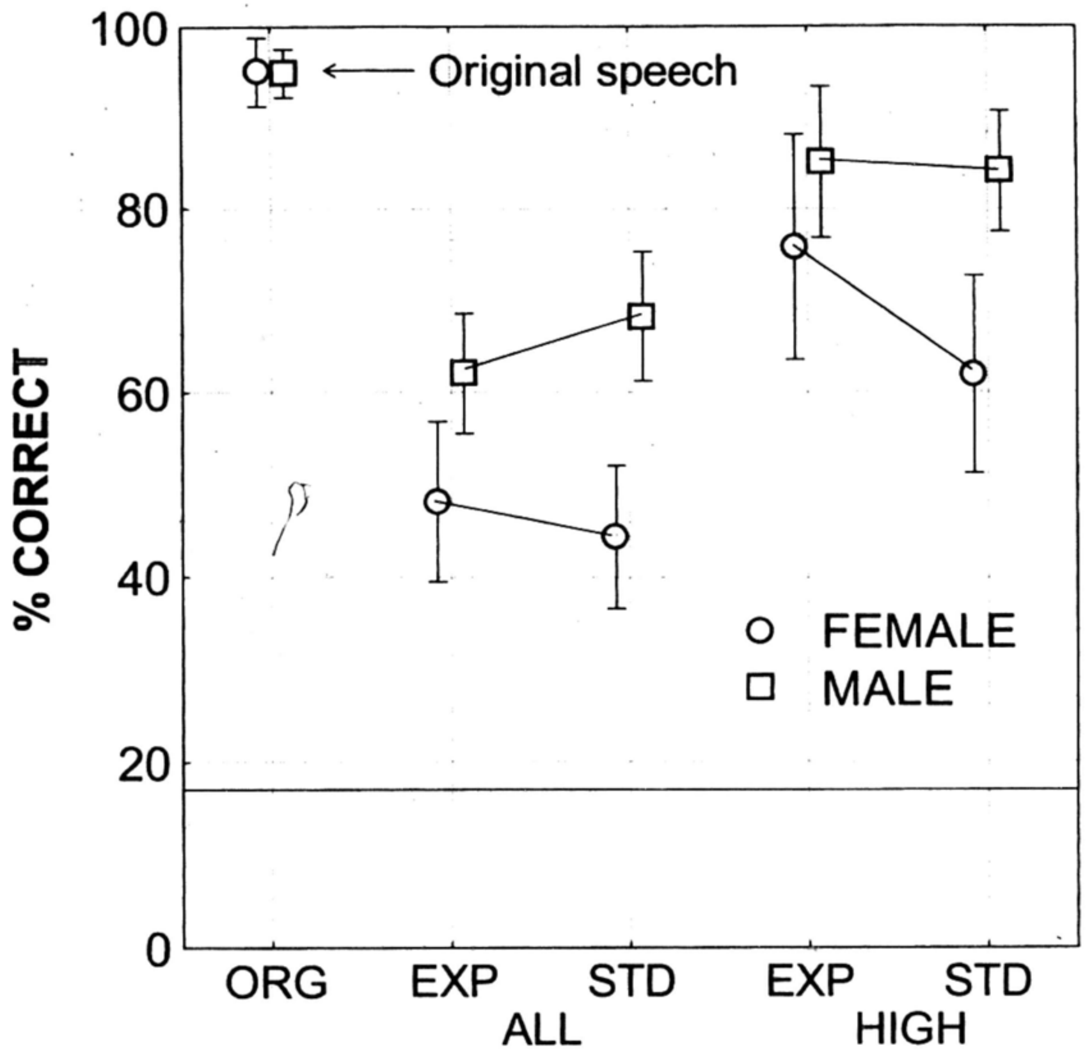


Figure 5.3: Tone identification accuracy with Cantonese monosyllabic words under different processing conditions.

interpreting the pairwise difference for the processing conditions.

Pairwise comparisons were performed by the *post-hoc* Fisher least-square-difference (LSD) Tests ($p < .05$). The demission of the low-frequency band gives significantly better performance than retaining it. The performance of tone recognition significantly improved by the expansion of high-frequency TEPC for female speech but no significant change was observed for male speech ($p = 0.73$), due to the ceiling effect of the high scores for male speech. With the presence of the low-frequency band, there is no significant difference between the expanded and unexpanded conditions for female speech ($p = 0.21$).

5.4 Discussion and Conclusion

In the cases that only the high-frequency TEPC is presented (conditions HIGH_STD and HIGH_EXP), the accuracy of tone identification is significantly better than the cases that both low- and high-frequency TEPCs are presented (conditions ALL_STD and ALL_EXP). This is consistent with our previous findings.

When only TEPC from high-frequency region is presented, the expansion leads a noticeable improvement on tone identification performance for female speech. The improvement is not significant for male speech. It is noted that, without TEPC expansion, the tone identification accuracy for male speech is much higher than that for female speech (85% vs. 60%). There is relatively less room to demonstrate the effectiveness of TEPC expansion for male speech.

□ **End of chapter.**

Chapter 6

Improved Tone Perception with $F0$-modulated TEPCs

Summary

Temporal periodicity cues are found to be important to tone recognition. Better representation of the tone-related information may improve the tone recognition performance. This chapter investigated the effectiveness of enhancing these cues for tone recognition in noise. The periodicity cues between 20 and 500 Hz in the TEPCs were simplified into a sinusoidal wave of the same fundamental frequency. Tone identification experiments were carried out using Cantonese disyllabic words. Results showed that the use of periodicity-enhanced TEPCs led to consistent improvement of tone identification performance. The improvement was more significant at low SNRs than for clean speech, and more noticeable for female speech than male speech. The analysis of error distributions showed that the periodicity enhancement method reduced the number of tone identification errors and influence to the recognition of segmental structures was subtle.

6.1 Introduction

In realistic hearing environments, periodicity cues can be easily deteriorated by background noise. To make pitch information more salient against noise, it was suggested to increase the modulation depth of the periodicity cues [Lorenzi et al., 1999; Vandali et al., 2005]. By using a simple power-n expansion of temporal envelope, Lorenzi et al. [1999] reported a small but consistent performance improvement of speech recognition in noise. We proposed a similar approach using non-linear envelope expansion which has been shown in Chapter 5. There are many other approaches that focused on improving the tone perception by manipulating the temporal envelope cues. Vandali et al. [2005] evaluated the pitch ranking performance of CI users with different speech processing strategies. In addition to the standard strategies commonly used in commercial devices, a few experimental algorithms were developed to encode $F0$-related periodicity information in the stimulus signal. Across all activated electrodes, the modulation depths of $F0$-related fluctuation were increased in a synchronous manner. The experimental results showed that these new strategies could improve pitch perception of CI recipients. In Lan et al. [2004] and Luo and Fu [2004a], acoustic simulations were carried out with $F0$-controlled sinusoidal or pulse-train carriers modulated by sub-band temporal envelopes. They demonstrated better performance of Mandarin tone recognition than systems with fixed-frequency carrier or noise carrier. In Green et al. [2004], the complex periodicity cue extracted from original speech was simplified as a sinusoidal or a sawtooth wave at the same fundamental frequency, and then combined with intact envelope cue. Improvements on pitch perception were observed in both acoustic simulations with NH subjects and listening tests with CI users.

There have been relatively fewer works on Cantonese speech recognition with temporal information than on English and Mandarin. Perceptual studies on lexical tones and intonation of Cantonese speech were reported in Ma et al. [2005, 2006]. It was shown that tonal context played an important role in Cantonese tone perception. Six-tone identification accuracy varied from 98.2% for tones in natural sentences to 78.8% for those presented in isolation. Lee et al.

[2002a] showed a big gap of Cantonese tone identification ability between NH and CI children (92% vs. 64%). Au [2003] tested Cantonese tone identification of a group of post-lingually deafened Cantonese-speaking CI users. The average accuracy was about 69% and great individual differences were observed. These results indicated that existing CI systems are not effective in delivering tone-related information.

The present study extends our previous work in two aspects. First, the effect of noise was investigated by using test stimuli at different SNRs. Second, in order to improve tone recognition in noise, the effectiveness of enhancing temporal periodicity cues was studied. We adopted the processing algorithm described in Green et al. [2004]. A slowly-varying temporal envelope component (TEC) was extracted by full-wave rectification and low-pass filtering at 20 Hz. In each sub-band, the TEC was multiplied with a constant-amplitude sinusoidal wave that follows the $F0$ trajectory of the original speech. This produced a modified TEPC, in which the temporal periodicity component (TPC) was simplified. Tone identification experiments by NH subjects were carried out with test stimuli generated by the standard CIS strategy and the periodicity-enhanced one. We expected that: (1) Cantonese tone identification accuracy would be improved by using periodicity-enhanced TEPCs; (2) periodicity enhancement on TEPCs from high-frequency region (1 – 4 kHz) would be particularly effective to improve tone identification accuracy; (3) the effect of periodicity enhancement would be more prominent for noisy speech than for clean speech.

6.2 Materials and Methods

6.2.1 Subjects

Five male and five female subjects participated in this experiment. Their ages ranged from 20 to 23 years. All of them are native Cantonese speakers with normal hearing. Their pure-tone thresholds were better than 20 dB HL at octave frequencies from 125 to 4000 Hz in both ears.

6.2.2 Speech Materials

In this study, the Cantonese disyllabic words were used as the speech materials. The full set of disyllabic words is same as the one described in Section 4.2.1.

Several sets of noisy speech materials were generated by adding noise signals to clean speech at different SNRs. For each of the speakers, a noise signal was generated by shaping the spectrum of white noise to follow the average spectrum of all test words spoken by this speaker. The noise signal was then added to the clean utterances at the SNRs of 0 dB, 10 dB and 20 dB. The SNR was controlled by fixing the Root-Mean-Square (RMS) intensity level of speech signals at -25 dB and varying the noise level. Both clean and noisy speech materials were low-pass limited to 4 kHz.

6.2.3 Speech Processing and Stimuli

In this study, the standard noise-excited vocoder was implemented. Detailed description of the vocoder has been shown in Section 3.3. Figure 6.1 explains the modified speech processing strategy that incorporates explicit $F0$-related periodicity information. It differs from the standard strategy in that the TEPC used to modulate the noise carrier is not directly derived from the input speech. At each band, a slowly-varying TEC is extracted with a low cut-off frequency of 20 Hz. The TECs are then multiplied with a sinusoidal wave that follows the $F0$ variation of the original clean speech. In other words, the complex periodicity cues are replaced by a simplified periodicity pattern [Green et al., 2004]. Given these modified sub-band TEPCs, the subsequent steps of acoustic stimuli generation are the same as the standard strategy.

In our study, a four-channel vocoder was used. The parameters of the vocoder, e.g., corner frequencies of the analysis bands, low-pass cut-off frequency for envelope extraction, the parameter for determining the filters, are the same as described in Section 3.3.

The $F0$ trajectories of all test stimuli were pre-computed from full-band clean speech signals. This was done with the pitch estimation algorithm imple-

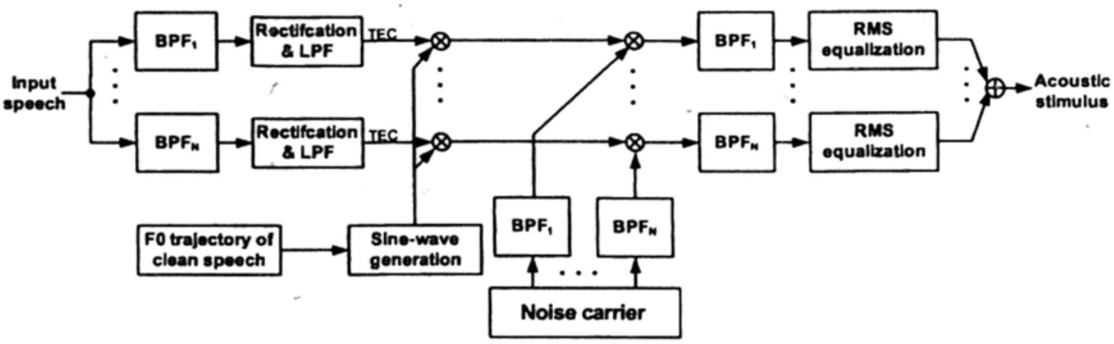


Figure 6.1: The modified speech processing strategy with enhanced periodicity cues.

mented in the PRAAT software [1]. The $F0$ values were manually checked and errors were corrected.

To investigate the contributions of envelope and periodicity information from different frequency regions, several sets of test stimuli were generated with different combinations of sub-bands. Based on our findings in Chapter 4, three different sub-band combinations were considered:

LOW : 60 Hz – 1 kHz (including the two low-frequency bands)

HIGH : 1 – 4 kHz (including the two high-frequency bands)

ALL : 60 Hz – 4 kHz (including all of the four bands)

With the three frequency regions and the two processing strategies, there were a total of six different test conditions as shown in Table 6.1. The standard and modified strategies are abbreviated as *STD* and *MOD*, respectively. For clean speech materials and noisy speech at 10 dB SNR, all of the six conditions were tested. For noisy speech at 0 dB and 20 dB SNR, only the *HIGH* conditions were tested. Including the unprocessed natural speech, there were 17 sets of test stimuli for each speaker.

[1] PRAAT 5.0.20 Copyright ©1992 – 2008 by Paul Boersma and David Weenink (www.praat.org)

Table 6.1: The 16 sets of processed test stimuli used in this study. Each row represents a specific way of generating TEPC. The four columns represent different noise conditions.

	$CLEAN$	$20dB$	$10dB$	$0dB$	
ALL_{STD}	√		√		Original TEPCs from 60 Hz – 4 kHz
ALL_{MOD}	√		√		Modified TEPCs from 60 Hz – 4 kHz
LOW_{STD}	√		√		Original TEPCs from 60 Hz – 1 kHz
LOW_{MOD}	√		√		Modified TEPCs from 60 Hz – 1 kHz
$HIGH_{STD}$	√	√	√	√	Original TEPCs from 1 – 4 kHz
$HIGH_{MOD}$	√	√	√	√	Modified TEPCs from 1 – 4 kHz

6.2.4 Psychophysical Procedure

The equipments included a laptop computer with a high-quality external audio interface (TASCAM US-122). Acoustic stimuli were presented to the subject via a Paired E.A.R. Tone 3A Insert Earphone (50 ohm). A computer software with graphical user interface was developed to control the presentation of test stimuli and collect responses from subjects as shown in Section 3.4.

Each subject was required to attend two test sessions on different days. Each session involved all test stimuli from one of the speakers. The presentation order of the two speakers was balanced over all subjects. In each test session, the unprocessed clean speech was presented at the beginning so that the subjects could familiarize themselves with the process and materials. Subsequently the 16 sets of processed stimuli were presented in randomized order.

Each set of stimuli included the 120 disyllabic words as described in Section 3.2.1. They were presented in randomized order without repetition. A four-alternative forced-choice (4-AFC) procedure was adopted. The four choices were displayed in the form of Chinese characters and the display positions were randomly assigned. After the presentation of a stimulus item, the subject was asked to select by mouse clicking the word that he/she had heard. The time for responding to each test item was fixed to 5 seconds. The subjects were encouraged to make a guess if they were not sure about the correct answer. If there was no response after 5 seconds, the test item would be regarded as "incorrectly recognized" and the system proceeded to the next item automatically. No feedback was given to the subjects. The 120 test items were presented in continuation without pausing midway. The duration for an entire test session was about 3 hours and the subjects were instructed to take a 5-minute break every one hour.

6.2.5 Method of Result Analysis

For each set of test stimuli in Table 6.1, percentage correctness of tone identification and word identification were evaluated over all subjects. The tone score counted all responses with correct tone identification, i.e., the recognized word carries the same tone as the presented word. This includes the items (T, S) and $(T, \overline{S})$ in the confusion matrix of Table 3.3. The chance level for tone identification is 50%. The word score counted the answers that exactly matched the presented words, i.e., the item (T, S). The chance level for word identification is 25%.

Tone scores and word scores were compared among different test conditions. Two primary factors that affect test results are frequency region (*LOW, HIGH, ALL*) and processing strategy (*STD, MOD*). Statistical analysis and comparison were performed using ANOVA and post-hoc test techniques. We also looked into the effect of noise level and the difference between male and female voices.

As seen in Table 3.3, there were three different types of errors: $(T, \overline{S})$, $(\overline{T}, S)$ and $(\overline{T}, \overline{S})$. The percentage distributions of these errors were analyzed to reveal the effects of the above condition factors on perception of lexical tones and segmental structures.

6.3 Results

6.3.1 Contributions from Different Frequency Regions

Tone Identification

Figure 6.2 shows the test results on tone identification for clean speech and noisy speech at 10 dB SNR. The results for male and female speech are displayed separately. The tone scores attained with the six processing conditions in Table 6.1 are compared side by side. It was consistently observed that the tone scores in the *HIGH* condition were better than *LOW*. The scores in *ALL* condition were between *HIGH* and *LOW* in most cases. The second observation was that *MOD* strategy produced better performances than *STD*, especially for female speech. The improvement was more noticeable for noisy speech than for clean speech. The third observation was that the scores for male speech were significantly higher than those for female speech.

The results of tone identification were analyzed using a two-way repeated-measures ANOVA with factors of processing strategy and frequency region. The ANOVA was done separately for different noise conditions (*CLEAN* and *SNR10dB*) and different speakers (*MALE* and *FEMALE*). The analyses revealed significant main effects of both factors ($p < 0.05$). There was an exception for *CLEAN-MALE*, where the main effect of processing strategy was not significant [F(1,9) = 0.03, $p = 0.87$]. This might be related to the high level of performances attained with the standard strategy (85% – 96% accuracy). The ceiling effect hindered further improvement. A significant main effect was found on the two-way interaction between processing strategy and frequency region for *SNR10dB-FEMALE* [F(2,18) = 5.18, $p = 0.017$], reflecting that the effect of MOD was more obvious for *HIGH* than for *LOW* and *ALL*, as seen from Figure 6.2.

Tukey HSD *post-hoc* tests confirmed that, for female speech, the scores were very different among different frequency region conditions, with *HIGH* being the highest and *LOW* the lowest ($p < 0.05$). The same trend was observed for male speech but without significant difference ($p > 0.05$). *Post-hoc* comparison also showed that, under the *HIGH* condition, the tone score with *MOD* was significantly higher than that with *STD*. For clean male speech, the modified processing strategy did not improve tone identification performance. This again might be due to high performance level attained with the standard strategy (~ 95%). The results also showed significant difference between the tone scores for the two speakers ($p < 0.05$).

Word Identification

Figure 6.3 shows the test results of word identification under different test conditions. The performances under the *HIGH* condition were consistently higher than *LOW*. Being different from tone identification results, *ALL* was better than *HIGH* in most cases. The *MOD* strategy improved the accuracy over *STD*, especially for female speech in noise condition. The word scores for male speech were higher than those for female speech under the same condition.

Two-way ANOVAs were carried out to analyze the effect of frequency region and processing strategy. Significant main effects from both factors were found for different noise levels and different speakers ($p < 0.05$). Similar to tone identification results, the main effect of processing strategy for *CLEAN-MALE* was not significant [F(1,9) = 0.08, $p = 0.78$]. Two-way interaction between processing strategy and frequency region for *SNR10dB-FEMALE* showed a significant main effect. Tukey HSD *post-hoc* comparison revealed no significant difference between *STD* and *MOD* except for *SNR10dB-FEMALE* ($p < 0.05$), indicating that the periodicity enhancement method may not be beneficial to word recognition. The word scores of *HIGH* and *ALL* were significantly higher than those of *LOW* in most cases ($p < 0.05$). For both processing strategies, there was no significant difference between *HIGH* and *ALL* ($p > 0.05$), except for *SNR10dB-MALE*. Overall speaking, the periodicity-enhanced processing method did not show any negative effect on word recognition but did show a positive effect on tone identification, especially for noisy speech.

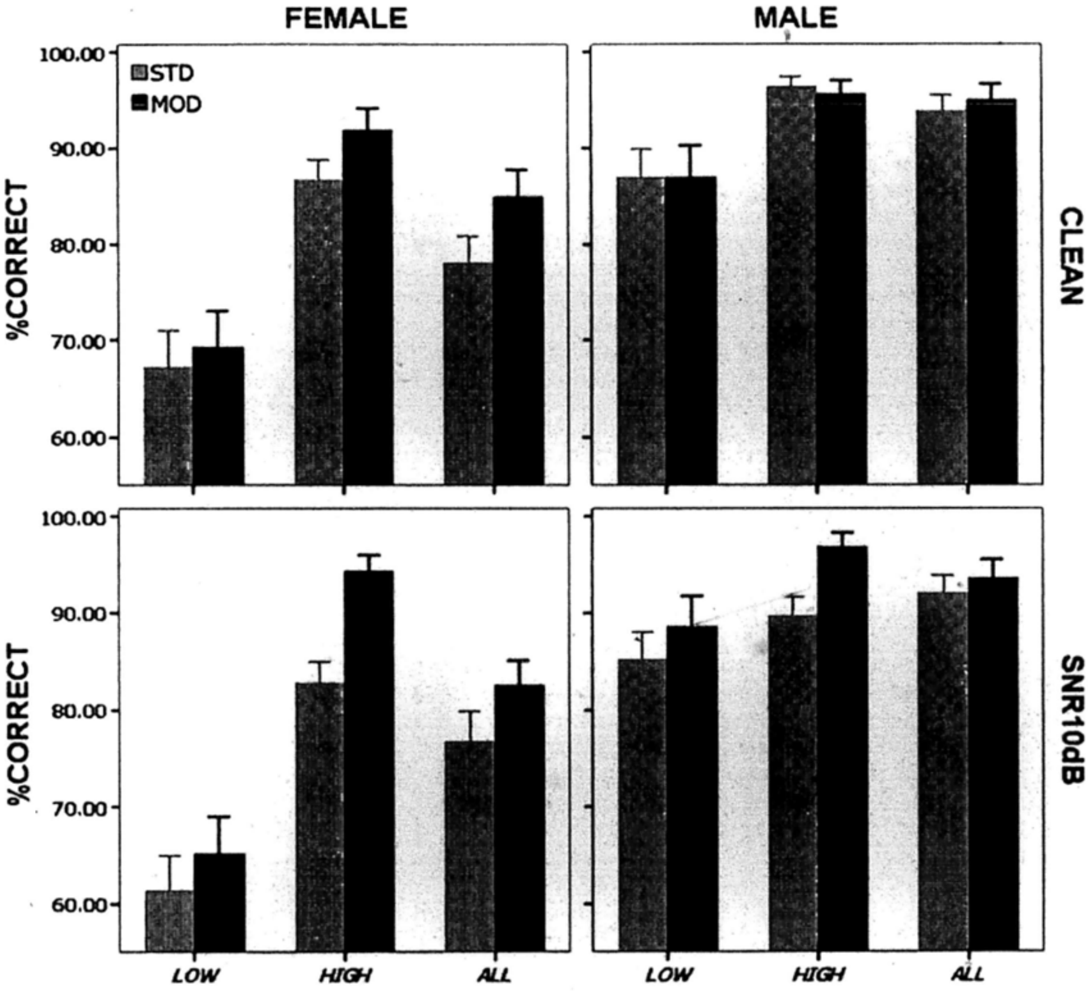


Figure 6.2: Tone identification accuracy. The results for clean speech and noisy speech (10 dB SNR), and for female and male voices are shown in separate panes. Each pane contains the tone scores attained with the six processing conditions. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance level is 50%.

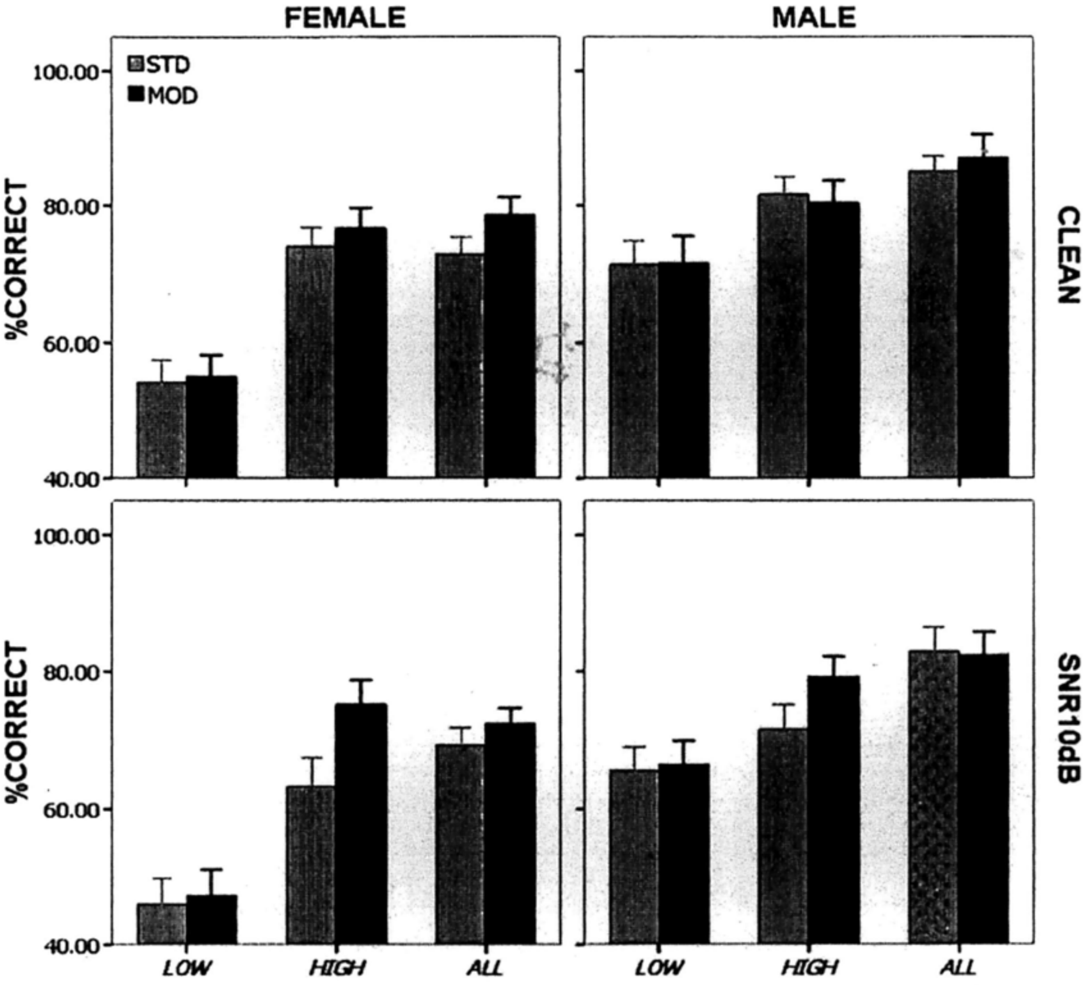


Figure 6.3: Word identification accuracy. The results for clean speech and noisy speech (10 dB SNR), and for female and male voices are shown in separate panes. Each pane contains the tone scores attained with the six processing conditions. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance level is 25%.

6.3.2 Effect of Noise

Figure 6.4 shows the test results as a function of SNR: clean, 20 dB, 10 dB and 0 dB under the *HIGH* condition. The scores for unprocessed clean speech are also given for reference. In general, the performances of both tone identification and word identification declined as the SNR decreased. *MOD* showed a consistently higher performance level than *STD*. For tone identification, *MOD* maintained an accuracy of about 90% across all noise conditions.

Two-way ANOVA was used to analyze the effects of noise level and processing strategy. Both factors and their interaction were shown to have significant main effects on tone identification and word identification ($p < 0.05$). The two-way interaction between processing strategy and noise level reflects the result that the performance difference between *STD* and *MOD* depends on the noise level. The effectiveness of *MOD* was more noticeable at low SNR than at high SNR. Tukey HSD *post-hoc* tests indicated that the tone scores and word scores of *MOD* were significantly higher than those of *STD* for *SNR10dB* and *SNR0dB* ($p < 0.05$). *Post-hoc* tests also showed no significant difference between the scores of *CLEAN* and *SNR20dB* with *STD* ($p > 0.05$). However, there were significant differences among *CLEAN*, *SNR10dB* and *SNR0dB* ($p < 0.05$). With *MOD*, there was no significant difference among all of the noise levels for male voice ($p > 0.05$). For female voice, only the scores at *SNR0dB* were significantly different from other noise levels ($p < 0.05$).

6.3.3 Analysis of Error Distributions

Figure 6.4 shows that, under the *HIGH* condition, the use of periodicity-enhanced TEPCs leads to improvement on both tone identification and word identification accuracies, especially at SNR of 10 dB or below. A word identification error may be caused by tone error, segmental error or both. Figure 6.5 shows the percentage distributions of different types of errors in the test results. It is noted that tone errors, which include $(\overline{T}, S)$ and $(\overline{T}, \overline{S})$, were substantially reduced by the periodicity-enhanced processing strategy, especially at low SNR.

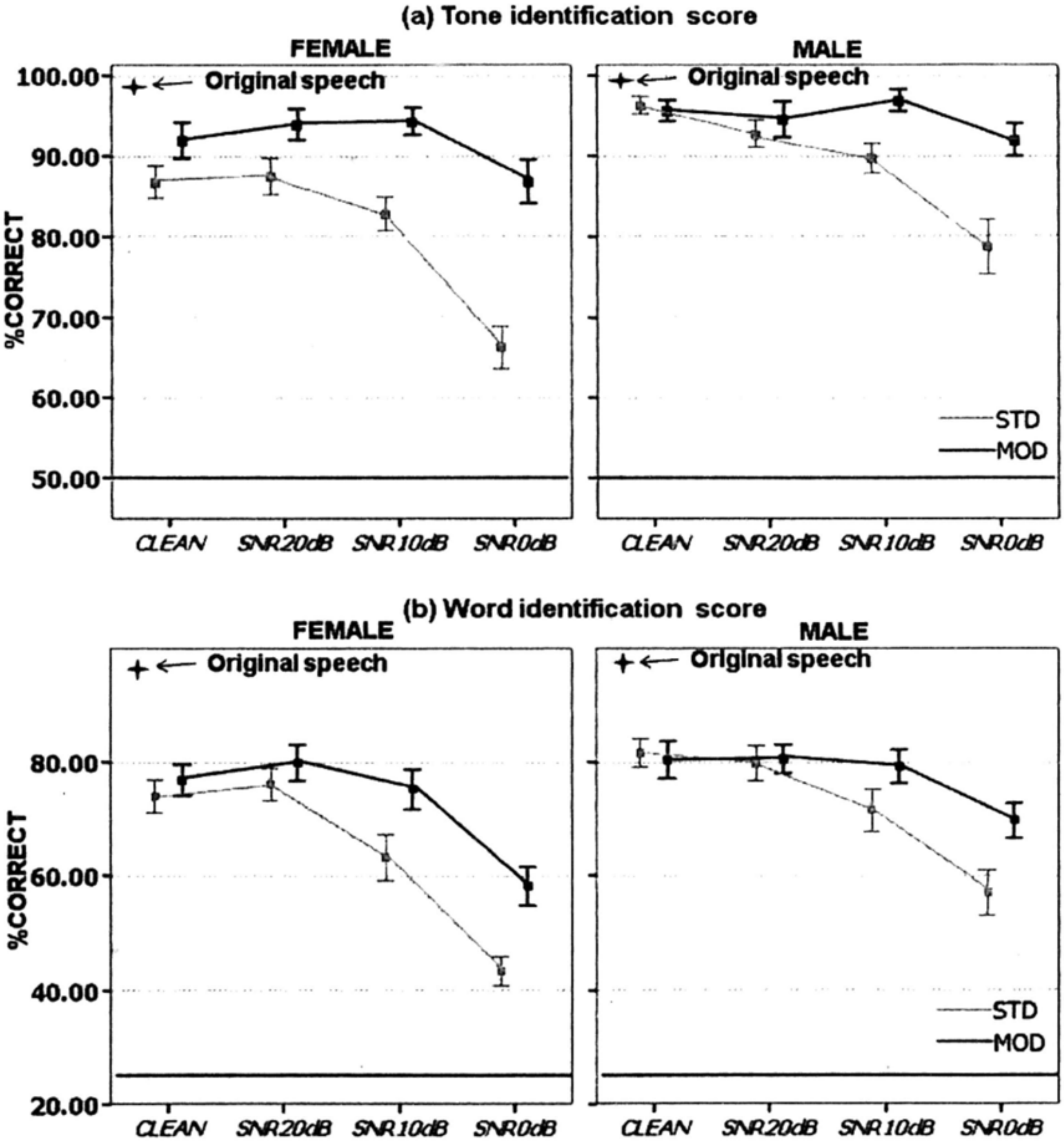


Figure 6.4: Percentage correctness of (a) tone identification and (b) word identification, as a function of SNR. The error bars indicate the 95% confidence intervals of the mean scores over all subjects.

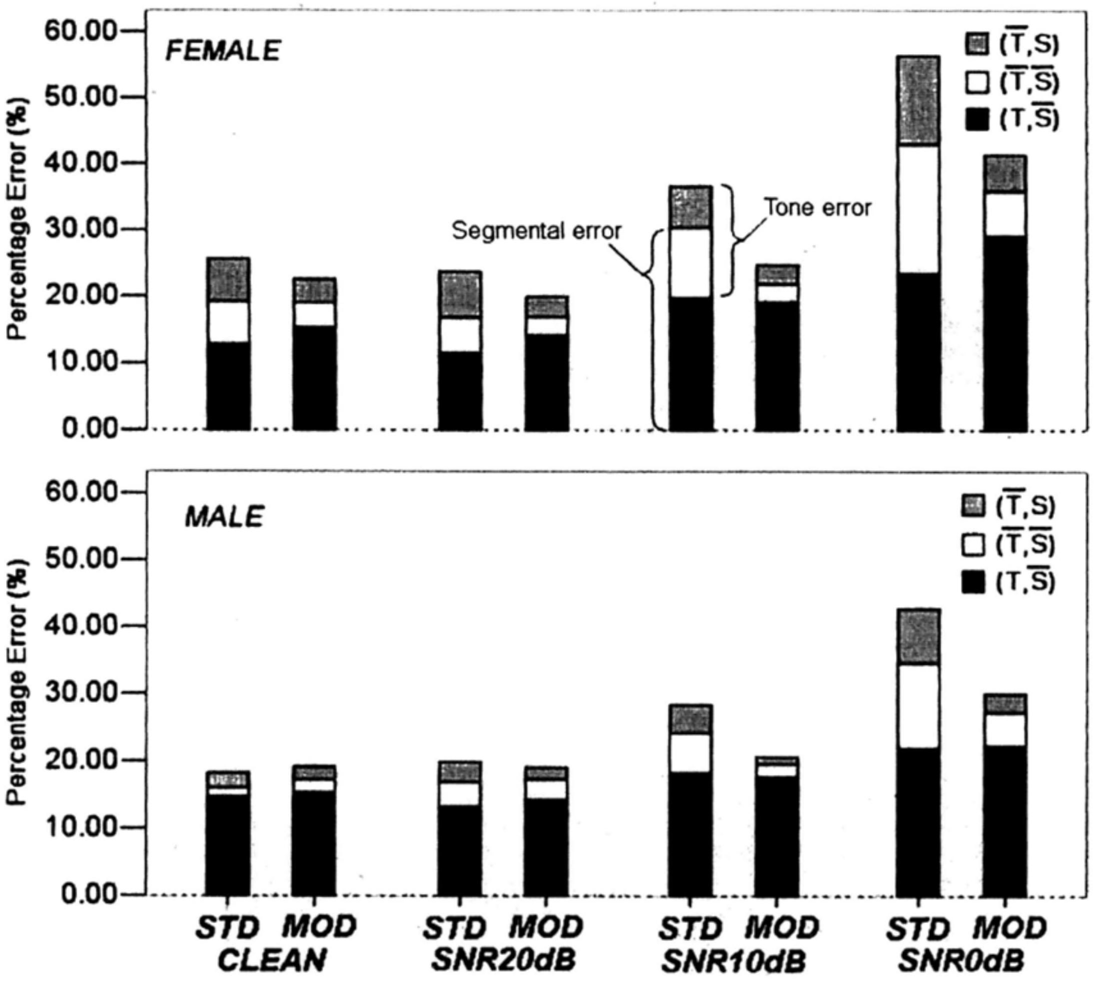


Figure 6.5: Comparison of percentage error distributions between the standard and the modified processing strategies at different SNRs. The percentage error is computed as the ratio of the number of the respective type of errors over the total number of test items for all subjects.

Meanwhile, the number of the $(T, \overline{S})$ errors increased because the improved tone identification made some of the $(\overline{T}, \overline{S})$ errors become $(T, \overline{S})$. The total number of segmental errors, which include $(T, \overline{S})$ and $(\overline{T}, \overline{S})$, decreased at low SNR and remained intact at high SNR. In other words, although the modified processing strategy removes the periodicity details of the original speech, it doesn't seem to affect the delivery of segmental information.

Figure 6.6 compares the distributions of different types of errors among the *LOW*, *HIGH* and *ALL* conditions. The number of the $(T, \overline{S})$ errors was similar between *LOW* and *HIGH*, but considerably smaller in *ALL*. This indicates

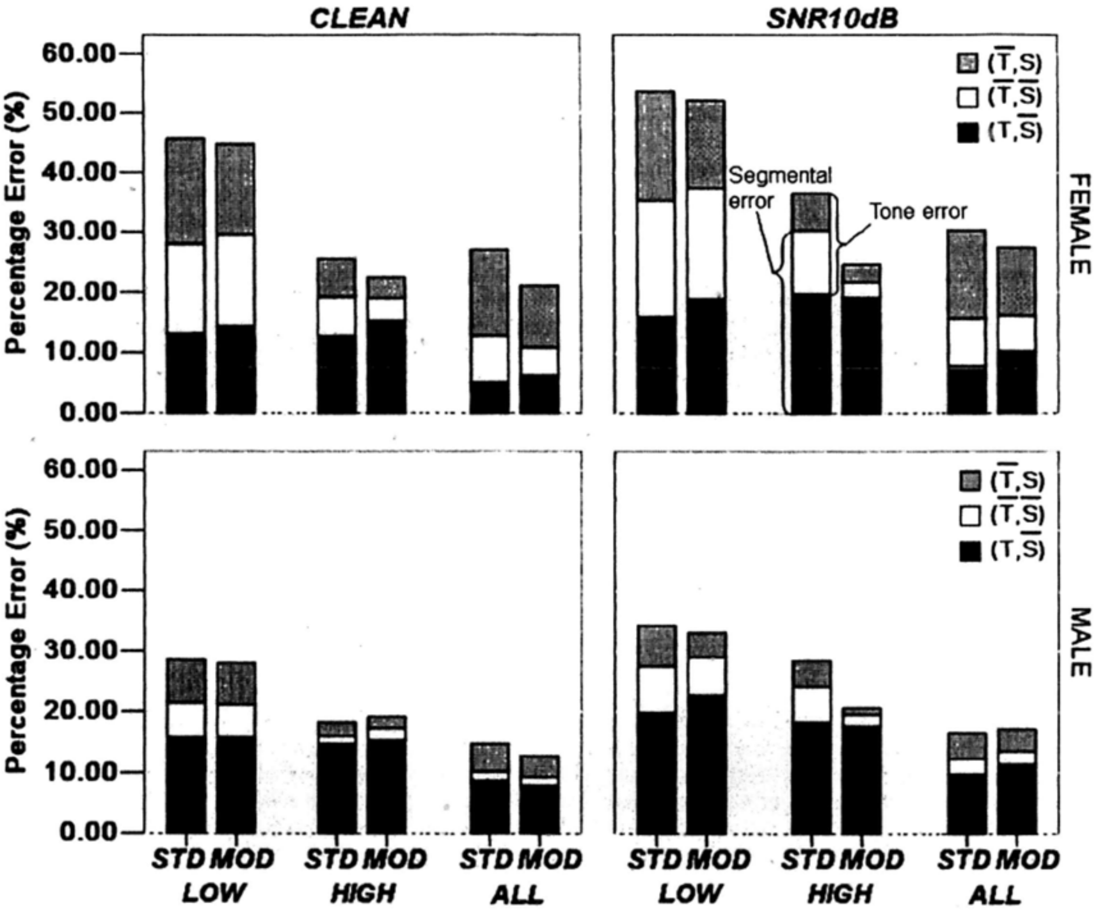


Figure 6.6: Comparison of percentage error distributions between the standard and the modified processing strategies under different frequency region conditions.

that TEPCs from all frequency bands contain useful information for identifying segmental cues. On the other hand, there were much fewer $(\overline{T}, S)$ errors in *HIGH* than in *LOW* and *ALL*. The use of TEPCs from low-frequency region seems to negatively affect tone recognition. The number of the $(\overline{T}, \overline{S})$ errors in *LOW* was much greater than in *HIGH* and *ALL*. Apparently, the superiority of *HIGH* to *LOW* in word identification was due to the improved tone identification.

6.4 Discussion and Conclusion

6.4.1 Effectiveness of Periodicity Enhancement

Our experimental results showed that Cantonese tone identification in quiet could reach a high performance level without using any fine structure cues. When speech is masked by noise, the performance deteriorated drastically. With the restricted spectral resolution, tone perception of CI users relies largely on temporal periodicity cues. The presence of noise leads to many spurious temporal peaks, which contaminate the representation of periodicity in the extracted TEPCs and hence adversely affect tone perception.

In this study, we hypothesized that tone perception could be improved with better representation of temporal periodicity and investigated the effect of enhancing temporal periodicity on Cantonese tone identification. We used a speech processing strategy modified from the conventional multi-channel noise-excited vocoder. Temporal periodicity was made more salient in three different ways. First, a simple periodicity pattern was used to replace the complex periodicity cues in the original speech. Second, the periodicity-related modulation depth was increased. Third, the F0-related periodicity pattern is synchronized across channels which maximizes the $F0$-related peaks in the output stimuli. As an example, Figure 6.7 compares the original TEPC and the periodicity-enhanced one of a Cantonese syllable. The modified TEPC shows a relatively simple $F0$ modulation pattern, i.e., in each pitch cycle, only one primary peak is retained and all secondary peaks are removed. At the same time, the modulation depth is increased to 100%.

This method of periodicity enhancement was first proposed and experimented by Green et al. [2004]. The test stimuli were synthesized English diphthong segments with gliding $F0$. The subjects were asked to distinguish between "rising" and "falling" pitch contours. By using $F0$-related sinusoidal or sawtooth waveforms to replace complex periodicity cues in speech signals, pitch discrimination capabilities of both NH subjects and CI recipients were improved noticeably. In our study, the benefit of using simplified periodicity cues was evaluated in a linguistic task for a specific language. The experimental results not only confirmed that tone perception had been improved but also showed that speech recognition could be improved in both quiet and noisy conditions. The tone scores attained by the modified strategy at SNRs of 20 dB and 10 dB were very close to that for clean speech. Even for very noisy speech of 0 dB SNR, the tone score maintained a very high level. In human sound perception, the peaks of a temporal envelope stimulus are translated into neural impulses. The intervals between successive impulses correspond approximately to the period of the sound or its integer multiples [Moore, 1998]. The complex periodicity cues, especially those extracted from noisy speech, may contain many pitch-irrelevant fluctuations such that the true $F0$ can not be clearly represented in the neural firing pattern. It is believed that a simplified periodicity pattern provides a better representation of $F0$ in the neural firing pattern [Green et al., 2004].

Temporal periodicity enhancement by increasing the modulation depth in TEPC has been widely studied [McKay et al., 1995; Lorenzi et al., 1999; Geurts and Wouters, 2001; Vandali et al., 2005]. McKay et al. [1995] found that, if the modulation depth was too small, the perceived pitch would correspond to the frequency of the pulse train carrier instead of the $F0$-related modulating frequency. Geurts and Wouters [2001] used sinusoidally amplitude-modulated (SAM) pulse trains. Their results showed that pitch discrimination performance of CI users degraded when modulation depth was decreased. In Lorenzi et al. [1999], noticeable performance improvement on speech recognition in noise was achieved with power-n expansion of TEPC. In the present study, the modulation depth of the periodicity-enhanced TEPCs was set to be 100%. The test results confirmed the effectiveness of increasing modulation depth for tone recognition.

6.4.2 Practical Implications

Ciocca et al. [2002] investigated Cantonese tone perception of a group of early-deafened CI users. They found that the children had great difficulty in extracting pitch information from temporal cues. There is a need to improve existing CI speech processing strategies for better pitch perception and tone recogni-

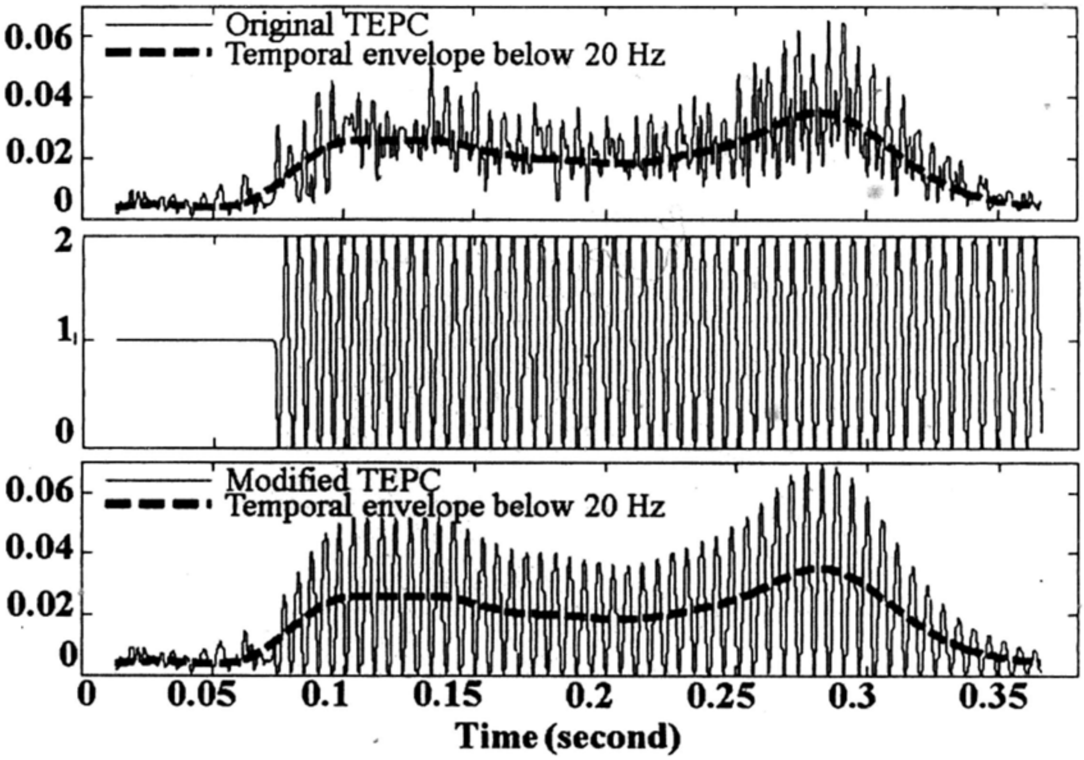


Figure 6.7: Comparison of the original TEPC and the periodicity-enhanced one. The speech segment contains a Cantonese syllable with additive noise at 10 dB SNR. The TEPCs are extracted from the sub-band of 2 – 4 kHz.

tion. The results of our acoustic simulation study indicate a possible approach to enhancing tone-related periodicity cues in CI devices for Cantonese-speaking users. It must be noted that acoustic simulations are the basis of further CI tests to some extent although real CI test results may not agree with the simulation results due to their differences [Laneau et al., 2006a].

For the implementation of this method, one major practical problem is real-time $F0$ estimation in realistic acoustic environments. In noise-suppressed speech, both time-domain periodicity and spectral-domain harmonic structure may be distorted. Algorithms for pitch detection can be broadly classified into three categories: algorithms using time domain properties, algorithms using frequency domain properties, and algorithms using both time and frequency properties. In time domain, the algorithms operate directly on speech waveform based on the measurements of peak and valley, zero-crossings and autocorrelation. The two most commonly used methods, i.e., autocorrelation function (ACF) and average magnitude difference function (AMDF), were jointly used for robust pitch estimation in Shimamura and Kobayashi [2001]. In frequency domain, simple measurements can be made on the frequency spectrum or a nonlinearly transformed version of it, as in the cepstral method [Noll, 1967], to estimate the pitch period of the signal by detecting the fundamental frequency or interval between successive harmonic components. Kunieda et al. [2000] used the autocorrelation of log spectrum to detect pitch harmonics in the presence of noise. Lahat et al. [1987] developed a spectral autocorrelation method of $F0$ extraction for noise-corrupted speech. Hybrid methods exploit both time-domain and frequency-domain approaches [Ahmadi and Spanias, 1999; Markel, 1972]. A pitch extraction algorithm in noise was recently proposed based on temporal and spectral representations [Shahnaz et al., 2008]. In state-of-the-art speech coding systems, pitch estimation has been implemented mostly based on ACF and AMDF algorithms. Oh and Un [1984] compared several pitch detection algorithms in noisy speech and found that the pitch detection accuracy could reach about 88% for clean speech and 85% for noisy speech at +5 dB SNR. Although many attempts have been made, most existing algorithms can handle only a limited range of noise conditions [Shahnaz et al., 2005, 2007; Kinjo and Funaki, 2006; Krini and Schmidt, 2007]. It is not yet possible to find a robust algorithm that can deal with adverse conditions as well as humans do. In our study, the $F0$ trajectory of the speech signal is extracted from the original clean speech. It optimally represents the $F0$ of input speech. Practically, the $F0$ trajectory will not be so accurate because they are extracted from noisy speech signals. It is also very difficult to perform accurate real-time estimation of $F0$ at affordable computation [Choi, 1997]. However, it is noted that precise $F0$ contours may not be necessary as far as tone perception is concerned. Ciocca et al. [2002] found that Cantonese tone perception relies mainly on the relative pitch levels. In Li and Lee [2007], it was shown that Cantonese tone contours could be approximated by simple linear movements without noticeable perceptual difference. Thus the $F0$ estimation problem may be alleviated by using coarsely predicted tone contours. We hypothesized that good tone perception performance can be achieved by providing approximated tone contours in the temporal periodic fluctuations. A supplementary experiment was carried out to investigate the effect of approximated tone contours to tone perception.

6.5 Supplementary Test on Cantonese Tone Identification with Approximated $F0$ Contour

6.5.1 Introduction

In the present study, we investigated the effect of approximated tone contour to lexical tone perception in temporal periodicity cues. This involves a decomposition of the envelope into two separate components. One consists of a slow-varying information which presents the dynamic changes of the spectral shape that are crucial for speech, while the second presents the $F0$-related periodicity in the form of a simplified synthesized waveform. The simplified waveform was generated in two different ways. First, the frequency of simplified waveform exactly follows the $F0$s of the tone contour of the original speech [Green et al., 2004]. Second, another waveform was synthesized, whose frequency generally shows the trend of pitch movement of the tones. In this approach each tone contour is approximated as a concatenation of line segments that describe the syllable-wide trend of $F0$ movement [Li and Lee, 2007]. The modified TEPCs were fed into the noise-excited vocoder to generate acoustic stimuli. Cantonese tone identification experiments by NH subjects were carried out with test stimuli generated by the standard CIS simulation and the two periodicity-enhanced ones in noisy speech. We hypothesized that (i) Cantonese tone perception would be improved with the use of periodicity-enhanced TEPCs; (ii) the perceptual test performance with the approximated tone contour method would be comparable to that with the exact tone contour approach. This implies that it is unnecessary to provide exact $F0$ estimation of voiced speech for Cantonese tone perception. It may indicate the possibility to improve the tone perception ability of CI users in practical environment using tone contours which are not very accurate but shows the trend of the tone contours.

6.5.2 Materials and Methods

Subjects

Five male and five female native Cantonese-speaking listeners participated in this study, aging from 19 to 21. All of them were normal hearing and had pure-tone thresholds better than 25 dB HL at octave frequencies from 125 to 4000 Hz in both ears.

Speech Materials

Cantonese disyllables were used in this study. These speech materials have been described in Chapter 3. The clean speech materials were corrupted by additive noises at a SNR of 10 dB. For each of the speakers, the noise signal was generated by shaping the spectrum of white noise to follow the average spectrum of all test words spoken by this speaker. All the noisy speech signals were equalized to the same intensity and low-pass limited to 4 kHz.

Speech Processing and Stimuli

In this study, a standard noise-excited vocoder was implemented which is the same as the one depicted in Chapter 3. We also adopted the modified speech processing strategy as shown in Section 6.2.3. The difference is the deriving method of the $F0$ trajectory from clean speech.

Tone Contour Estimation and Approximation

The $F0$ trajectories of all test stimuli were pre-computed from full-band clean speech signals. This was done with the pitch estimation algorithm implemented in the PRAAT software [2]. The $F0$ values were manually checked and errors were corrected. In our study, this estimation approach was supposed to provide the natural tone contour of the original speech.

Since $F0$s of the natural tone contour are computed frame by frame, it demands a heavy computation which is not suitable for practical implementation. On the other hand, the exact $F0$ estimation is not stable for noisy speech. In order to simplify the $F0$ estimation and provide perceptually acceptable tone information, the tone contour was approximated. Different approximation strategies are used for level tones and rising tones, as shown in Figure 6.8. They are designed on the basis of both the phonological descriptions in Figure 2.3 and acoustic observations from natural speech. For level tones, a single linear movement is used (i.e., Tone 1, Tone3, Tone 4 and Tone 6). For rising tones, two linear movements are used (i.e., Tone 2 and Tone 5). The linearly approximated tone contours were used to generate the sinusoidal wave which replaced the original complex temporal periodicity cues in each sub-band.

In total, there were three different processing strategies tested in this study. The first one is the standard noise-excited vocoder with original TEPCs below 500 Hz in each frequency bands (denoted as STD); the second one is the

[2] PRAAT 5.0.20 Copyright ©1992 - 2008 by Paul Boersma and David Weenink (www.praat.org)

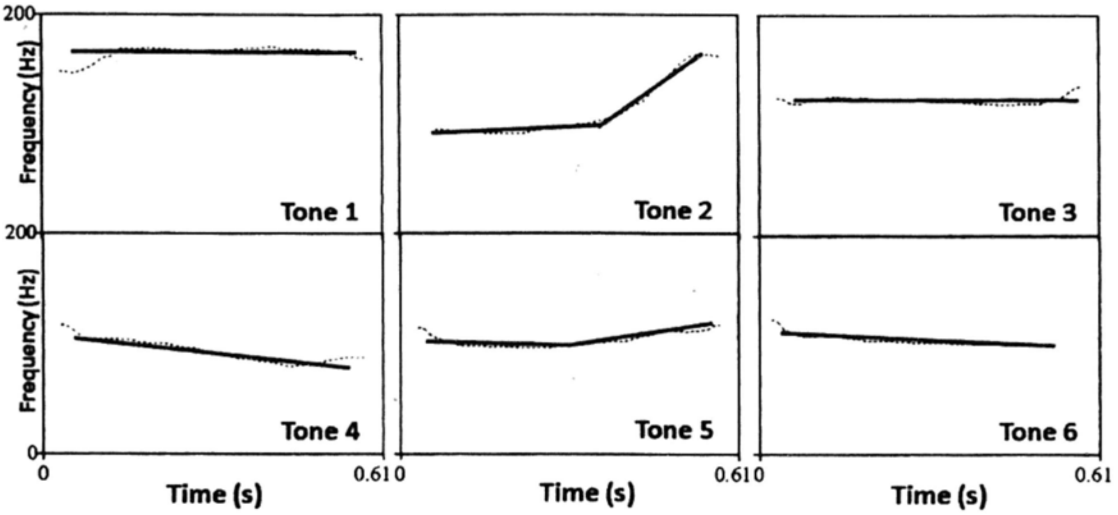


Figure 6.8: Approximations of isolated tone contours

periodicity-enhanced vocoder with original TEC below 20 Hz and simplified $F0$ trajectory following the natural tone contour (denoted as $MOD1$); the last is the periodicity-enhanced one with original TEC below 20 Hz and simplified $F0$ trajectory with approximated tone contour (denoted as $MOD2$). For all the processing strategies, the stimuli only contain the outputs from the higher two frequency bands (> 1000 Hz) while the lower two bands were discarded. From our previous observations, we found that the TEPCs from the higher two bands were more important from tone perception.

Psychophysical Procedure

The NH subjects were seated in a sound-treated booth and listened to the stimuli presented via a Paired E.A.R. Tone 3A Insert Earphone at 65 dBA. The test procedure is the same as the one in Section 6.2.4, except that more processing conditions were included which corresponded to the modified method with approximated $F0$ trajectory.

6.5.3 Results

Figure 6.9 shows the percentage correctness of tone identification for noisy speech at 10 dB SNR. The results for male and female speech are displayed separately. The tone scores of the three processing conditions are compared.

The scores of the original unprocessed speech condition were included for reference. Three apparent trends were observed from the tone scores. First, the tone scores under *MOD1* and *MOD2* conditions were better than those under the *STD* condition. Second, comparable tone scores were observed between the two periodicity-enhanced strategies. Third, the scores for male voice were always higher than those for female voice under the same condition.

The test results were analyzed using a two-way repeated-measures ANOVA with the factors of processing strategy and speaker ($p < 0.05$). The analyses revealed significant main effects of both factors. A significant main effect on the two-way interaction between processing strategy and speaker was also found, reflecting that the performance difference of STD and the two modified strategies was more salient for female speech than for male speech. This is because of the high level of performance for male speech which hindered further improvements by the modified strategies. Tukey HSD *post-hoc* tests showed that, for female speech, the tone scores attained from the two periodicity-enhanced methods were significantly higher than those from STD ($p < 0.05$). The same trend was observed for male speech but without significant difference between $MOD2$ and STD ($p = 0.09$). *Post-hoc* comparison also showed that the tone scores between *MOD1* and *MOD2* were not significantly different for both male and female speech ($p > 0.05$). For all processing strategies, the scores from male speech were higher than those from female speech. However, only one significant difference was observed for STD condition.

6.5.4 Discussion and Conclusion

In our periodicity-enhancement methods, the complex temporal periodicity in TEPC was replaced by a simplified sinusoidal wave whose frequency was equal to the exact $F0$ value of the original speech or followed the trend of $F0$ movement. At the same time, the modulation depth of the temporal periodicity was increased to 100%. Psychoacoustical experiments showed that the scores of periodicity enhancement were above 90%, significantly higher than the original TEPC. The $MOD1$ method was similar to Green et al. [2004]. In their study,

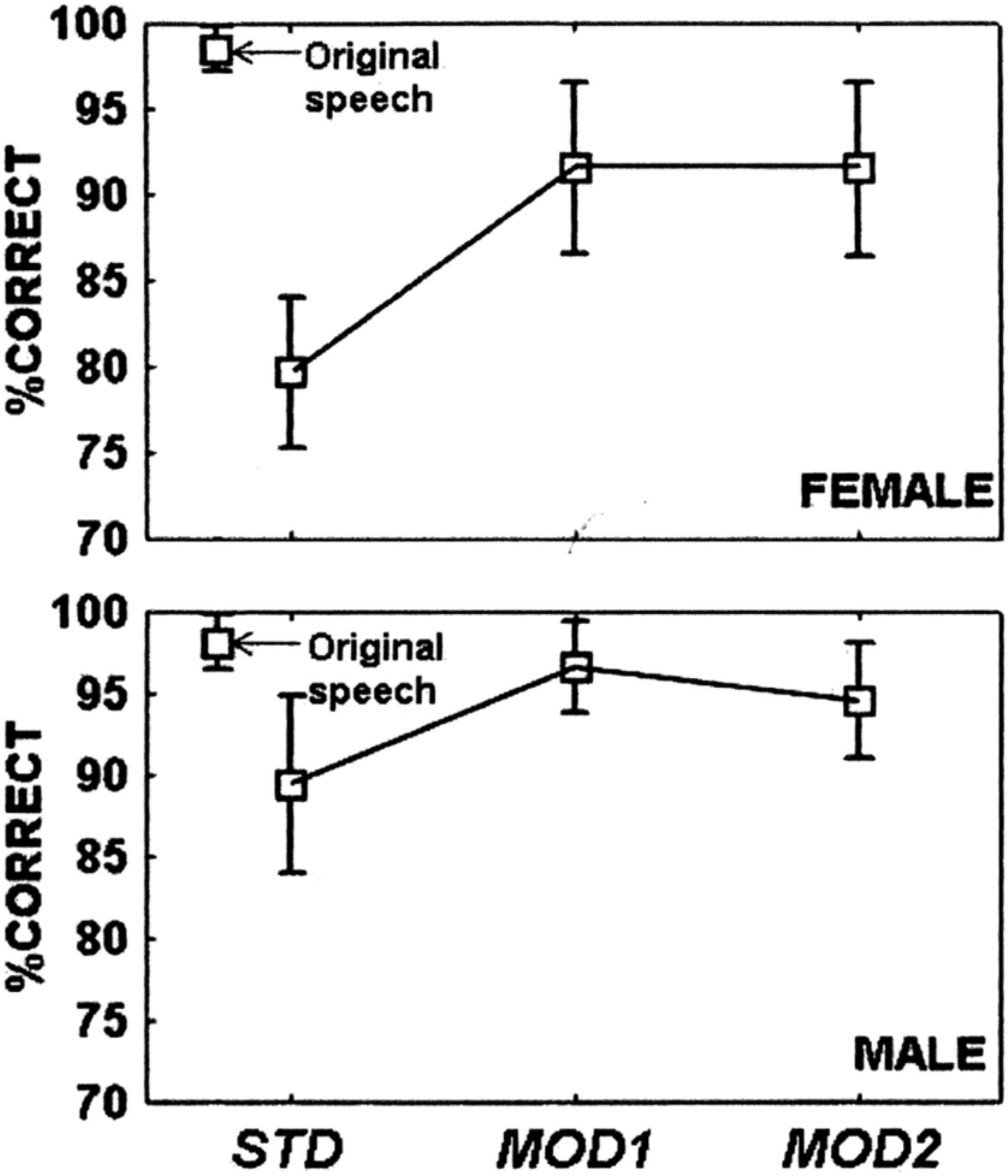


Figure 6.9: Percentage correctness of tone identification. The results for female and male voices in 10 dB SNR are shown in separate panes. Each pane contains the tone scores attained with the three processing conditions. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance level is 50%.

the test stimuli were synthesized English diphthong segments with gliding $F0$. By using $F0$-related sinusoidal waveforms to replace complex periodicity cues in speech signals, pitch discrimination capabilities of both NH subjects and CI recipients were improved noticeably. In human sound perception, the peaks of a temporal envelope stimulus are translated into neural impulses. The intervals between successive impulses correspond approximately to the period of the sound or its integer multiples [Moore, 1998]. The complex periodicity cues, especially those extracted from noisy speech, may contain many pitch-irrelevant fluctuations such that the true $F0$ can not be clearly represented in the neural firing pattern. It is believed that a simplified periodicity pattern provides a better representation of $F0$ in the neural firing pattern [Green et al., 2004].

Studies by Green et al. [2004, 2005] suggested that enhancing the temporal fluctuations of spectrally degraded speech may produce small improvements in $F0$ processing, but certain kinds of enhancements may result in reduced transmission of other speech features. In their modified processing scheme, the standard 400 Hz smoothed amplitude envelope was replaced by the product of a slow rate envelope and simplified $F0$-related modulation (sinusoidal or sawtooth). Both in acoustic simulations and in implant users, the ability to use intonation information to identify sentences as question or statement was significantly better with modified processing. However, while there was no difference in vowel recognition in the acoustic simulation, implant users performed worse with modified processing both in vowel recognition and in formant frequency discrimination. It appears that, while enhancing pitch perception, modified processing harmed the transmission of spectral information. In Laneau et al. [2006b], a new sound processing scheme (F0mod) was designed to optimize pitch perception, and its performance for music and pitch perception was compared in four different experiments to that of the current clinically used sound processing scheme (ACE) in six Nucleus CI24 subjects. In the F0mod scheme, slowly varying channel envelopes are explicitly modulated sinusoidally at the fundamental frequency of the input signal, with 100% modulation depth and in phase across channels to maximize temporal envelope pitch cues. With the new F0mod scheme, music perception was found to be improved significantly with respect to the current most often used sound processing strategy, ACE, for Nucleus recipients. These results indicate that explicit $F0$ modulation of the channel envelopes improves music perception in CI subjects. Another interesting work was done by Luo and Fu [2004b]. They found that enhancing the co-varying amplitude envelope information improved tone recognition by NH listeners.

In addition to the benefits from explicit coding of $F0$ in the channel envelopes, Vandali et al. [2005] developed an experimental strategy, multichannel envelope modulation (MEM), that implicitly enhances the coding of $F0$-periodicity cues in the incoming signal. The strategy extracts the low-frequency (80-400 Hz) envelope of the broadband signal, which contains $F0$ periodicity information, and uses it to modulate the envelope of the band-pass filtered channel signals derived from the ACE strategy. As a result of this processing, $F0$ periodicity information in the envelope of the broadband signal is presented coincidentally in time across all stimulation channels. Furthermore, the modulation depth of the $F0$ periodicity information is expanded in the stimulus envelope so as to enhance its perception. Results for CI users using this strategy in pitch ranking tests were significantly better than those using the ACE and CIS strategies. In addition, no degradation in speech perception in quiet and noise was observed using the MEM strategy compared to the ACE and CIS strategies. Wong et al. [2008] followed the MEM strategy and tested the CI users who are Cantonese-speaking postlingually deafened adults. Speech intelligibility in speech-spectrum shaped noise was measured using the Cantonese hearing in noise test (CHINT). However, MEM did not demonstrate any advantages for speech recognition in noise. Subjects preferred ACE for daily listening situations, and a few preferred MEM in noise.

In our second periodicity-enhancement strategy, which is similar to $MOD1$, the complex temporal periodicity was simplified by a sinusoidal wave at 100% modulation depth. The frequency of the sine wave corresponded to the linear approximation of the pitch movement of the original tone contour. We proposed this strategy based on the following hypotheses. First, tone perception does not rely on the exact $F0$ values at particular time instants, but depends on the trend of pitch movement. Second, linear approximation of the tone contours is sufficient to achieve competitive Cantonese tone perception using temporal periodicity cues. In Li and Lee [2007], a perceptual study on approximated Cantonese tone contours was conducted. Cantonese monosyllabic words, disyllabic words and sentences were tested in their study. The $F0$ contour was extracted and modified from broad-band speech. New speech materials were synthesized from the original speech and the modified $F0$ contour by PRAAT software. They concluded that perception of tone contours relies mainly on the major trend of pitch movement and linear approximation of tone contours was adequate to describe the pitch movement. In the current study, approximated $F0$ contours were incorporated into TEPCs. Our experimental results showed that Cantonese tone identification performance with periodicity-enhanced TEPCs was significantly higher than that with original temporal envelope cues below 500 Hz. It showed that the use of linear approximation of tone contour in temporal periodicity-enhancement could attain competitive perceptual performance as that with exact tone contour. This indicated that exact $F0$ estimation is not a critical requirement for improving tone perception in CI with pitch enhancement strategy.

□ **End of chapter.**

Chapter 7

Conclusions and Future Research

7.1 Conclusions

Temporal cues play an important role in human speech recognition, especially when the use of spectral cues is limited. For people who suffered from severe to profound hearing loss, their abilities of utilizing spectral details of sounds are adversely affected. These people rely more on the use of slowly-varying temporal cues for speech perception. Currently cochlear implants (CI) employ advanced signal processing techniques to deliver temporal cues to hearing impaired people by electrically stimulating the auditory nerves. The devices are of importance to recover part of the hearing. In quiet environment, speech recognition performance can reach a relatively high level for some of the patients. However, there is still a big gap on hearing ability between patients with hearing prostheses and people with normal hearing. CI patients who speak tonal language find it difficult to perceive tones, which are critical to the understanding of tonal languages. This problem reflects the fact that existing speech processing strategies in CI devices are not effective in delivering tone-related information. This triggered our research to investigate useful temporal cues and the possibility of enhancing them for better tone perception.

In this thesis, we focussed on the study of temporal cues for tone perception of Cantonese and Mandarin and investigate the ways to improve tone perception. The following questions were addressed:

- Are temporal envelope and periodicity components (TEPCs) important to Cantonese and Mandarin tone perception?
- Do the TEPCs from different frequency regions have different importance for tone perception?
- Is it possible to improve tone perception by manipulating the TEPCs? What are the possible signal processing procedures to make tone-related information more salient?

A multi-channel noise-excited vocoder is used. The temporal cues from a small number of channels are extracted and evaluated. Acoustic signals are obtained as the outputs from the vocoder. Psychoacoustic hearing tests were carried out with NH subjects listening to the acoustic stimuli. A number of factors were investigated in our study. The major conclusions are given below.

TEPCs from Different Frequency Regions

We investigated the contributions of TEPCs extracted from different frequency regions to tone perception. Test stimuli were created from the combinations of TEPC-modulated noise carriers from different frequency bands. The experiments showed a consistent observation that TEPCs obtained from high-frequency region (1 – 4 kHz) are more important than those obtained from low (60 – 1000 Hz) and middle (500 – 2000 Hz) frequency regions. This is explained in different perspectives: (1) In low and middle frequency regions, the frequency range of TEPC ($< 500Hz$) overlaps with or is close to that of the noise carriers. This causes a severe distortion on TEPC such that tone-related periodicity is destroyed; (2) In the low-frequency region, the peripheral auditory filters limit the perceived modulation depth, and consequently the sensitivity to temporal pitch. The choice of the carrier frequency thus requires careful consideration. A noise-band vocoder with high center frequencies for the re-synthesis filters (e.g., mimicking a shallow insertion depth of the simulated electrode array) can overcome the limitations of the peripheral auditory filters.

Our study confirmed the importance of TEPC to tone perception, which was demonstrated in many previous studies. As shown in Chapter 4, the highest tone identification accuracy was 92% in Figure 4.2.

Signal Processing Methods

Tone perception relies on the perception of temporal pitch when spectral information is not available. In noisy condition, whether the temporal periodicity could be kept salient is most critical to perception of tone. There are different ways of enhancing temporal periodicity cues, which include increasing the modulation depth of TEPCs, synchronizing periodicity cues across channels and simplifying temporal periodicity cue.

In Chapter 5, a non-linear expansion method on TEPCs was described. Its aim is to expand the modulation depth of the TEPCs such that periodicity becomes more prominent. In line with previous studies, our results showed that TEPC expansion leads a noticeable improvement on tone identification performance, especially for female voice, when TEPCs from high-frequency regions are presented.

In Chapter 6 we investigated the effectiveness of enhancing periodicity cues for Cantonese tone recognition in noise. The temporal periodicity cues between 20 and 500 Hz were simplified into a sinusoidal wave with the pitch contour following the original speech. The simplified temporal periodicity was applied in a synchronized manner across all channels. The results showed that the use of periodicity-enhanced TEPCs led to a consistent improvement of tone identification performance under different test conditions. The periodicity-enhanced method was very effective in low SNRs. For example, it could increase tone identification accuracy for female voice by 20% at 0 dB SNR. We analyzed the distribution of errors to reveal the effects of the periodicity-enhanced method on lexical tone perception and segmental recognition. It was shown that tone identification errors can be effectively reduced when SNR was 10 dB or below.

Meanwhile, the segmental information was not affected although the modified processing method removes some periodicity details in the original speech.

Speaker Gender Effects

Test results showed that tone identification accuracy on male voice was significantly higher than that on female voice. This suggests that the higher $F0$ modulation frequency in TEPC of female speech could not be represented as well as that of male speech. In TEPC below 500 Hz, there are more harmonics included in a sound with lower $F0$s than in a sound with higher $F0$s. For example, for a sound with $F0$ of 120 Hz, there are four harmonics contained in TEPC. For a sound with $F0$ of 200 Hz, there are only two harmonics in TEPC. More harmonics lead to deeper modulation such that the F0-related periodicity cues become more salient for perception. As a result, tone information can be perceived more easily.

Noise Effects

Additive background noise adversely affects the perception of tonal information. Cantonese tone identification accuracy could reach about 92% in clean speech condition, while the accuracy dropped to 72% at a SNR of 0 dB in a four-channel vocoder. More frequency bands are needed to achieve better performance as the noise levels increase. This means that more electrodes should be provided to the CI listeners when they are exposed to noisy environments.

Number of Frequency Bands

Increasing the number of frequency bands will make more spectral information available for perception. Compared to temporal cues, spectral cues are more robust to noise. With the increase of frequency bands, tone perception is improved as well as phoneme perception. With 32 bands, Cantonese tone identification accuracy reach the same level as for original unprocessed speech.

Speech Materials

Different types of speech materials, i.e., monosyllabic and disyllabic words, were designed for tone identification tests. The use of disyllabic words provides more segmental variations and reveals the importance of tone perception as part of speech perception. Consistent results have been obtained by the use of our speech materials, confirming the appropriateness of the speech materials. The design would be useful in other similar tasks.

7.2 Summary of Contributions

The major contributions from this thesis are summarized below:

- This thesis is focused on Cantonese, one of the major Chinese dialects. Cantonese has one of the most complicated tone systems among all languages in the world. However, research works on Cantonese tone perception with temporal cues have been quite limited, compared to Mandarin. This work is a valuable first effort on this important language.

- This is the first study on evaluating the contributions of different frequency regions to speech recognition of tonal languages, especially Cantonese. Our study shows that TEPCs from high-frequency region are more important for tone perception than those from low-frequency region. This leads to an awareness of choosing appropriate carrier frequencies in a vocoder. Low-frequency region should carry relatively lower weights to reduce the negative effect of low-frequency carriers to the temporal cues.

- Tone perception can be improved by increasing the modulation depth in TEPCs and providing cleaner and synchronized temporal periodicity cues relating to voice fundamental frequency. Our processing method with the $F0$-modulated TEPC shows an possibility to enhance tone perception performance in CI users by providing a simplified temporal periodicity cue with tone-related information to the listeners.

- The design of tone perception tests is original and of great value for future work in this area. Different from other studies, the speech materials were capable of providing segmental variation and facilitating the study of the segmental recognition performance together with the study of tone perception.

7.3 Future Work

The research platform, ATOPEX, as described in Chapter 2, allows us to setup and run the experiments described in this thesis. Although ATOPEX is very easy to use, it has a number of limitations. Firstly, the extensibility of the software needs to be improved. Secondly, the current platform can support tone identification tasks only with a static test procedure. An adaptive test procedure is expected with adaptive closed-set identification and speech in noise testing with variable SNR. Another limitation is that the present ATOPEX doesn't record the response time of the subjects. It might also be interesting to record the response time because the response time is often related to the subjects' performance.

In this thesis, we investigated different methods of improving tone perception by manipulating temporal envelope and periodicity information. Monosyllabic and disyllabic words were used as the test materials. Although the proposed speech materials have made a significant step towards investigating speech perception of tonal languages, it would be desirable to evaluate speech recognition performance with wider segmental variations (i.e. vowel and consonant recognition) and super-segmental cues (i.e. sentence recognition). Moreover, it would be important to test HI subjects with the $F0$-enhanced electrical stimuli.

In current CI systems, useful speech information is delivered to the patients by the envelope cues within each channel. Despite the progress in the design and performance of CI systems, the patients do not hear as well as normal hearing listeners. Particularly in some adverse situations, e.g., with competitive talkers or low SNRs. Temporal fine structure cues have been found to be more robust against noise than the temporal envelope cues. However, the current CI systems are not adequate in providing speech temporal fine structure information to the patients. Therefore, a possible direction for enhancing the speech recognition performance would be effective inclusion of temporal fine structure in CI systems.

□ End of chapter.

Appendix A

Tables and Equations

Tables of Statistics on Cantonese and Mandarin

Table A.1: Statistics on Cantonese and Mandarin syllables.

	Cantonese (LSHK, 1997)	Mandarin (CCDICT, 2000)
Total number of base syllables	625	420
Total number of tonal syllables	1,761	1,471
Average number of tones per base syllable	2.8	3.5
Average number of base syllable pronunciations per character	1.1	1.6
Average number of tonal syllable pronunciations per character	1.2	2
Average number of homophonous characters per base syllable	17	31
Average number of homophonous characters per tonal syllable	6	8

Table A.2: The 19 Cantonese Initials (labeled in the Jyut-Ping scheme).

LSHK symbols	Manner of Articulation	Place of Articulation
[b]	Plosive, unaspirated	Labial
[d]	Plosive, unaspirated	Alveolar
[g]	Plosive, unaspirated	Velar
[p]	Plosive, aspirated	Labial
[t]	Plosive, aspirated	Alveolar
[k]	Plosive, aspirated	Velar
[gw]	Plosive, unaspirated, lip-rounded	Velar, labial
[kw]	Plosive, aspirated, lip-rounded	Velar, labial
[z]	Affricate, unaspirated	Alveolar
[c]	Affricate, aspirated	Alveolar
[s]	Fricative	Alveolar
[f]	Fricative	Dental-labial
[h]	Fricative	Vocal
[j]	Glide	Alveolar
[w]	Glide	Labial
[l]	Liquid	Lateral
[m]	Nasal	Labial
[n]	Nasal	Alveolar
[ng]	Nasal	Velar

Table A.3: The 21 Mandarin Initials (labeled in the Pinyin scheme).

Pinyin symbols	Manner of Articulation	Place of Articulation
[b]	Plosive, unaspirated	Labial
[d]	Plosive, unaspirated	Alveolar
[g]	Plosive, unaspirated	Velar
[p]	Plosive, aspirated	Labial
[t]	Plosive, aspirated	Alveolar
[k]	Plosive, aspirated	Velar
[gw]	Plosive, unaspirated, lip-rounded	Velar, labial
[kw]	Plosive, aspirated, lip-rounded	Velar, labial
[z]	Affricate, unaspirated	Alveolar
[c]	Affricate, aspirated	Alveolar
[s]	Fricative	Alveolar
[f]	Fricative	Dental-labial
[h]	Fricative	Vocal
[j]	Glide	Alveolar
[w]	Glide	Labial
[l]	Liquid	Lateral
[m]	Nasal	Labial
[n]	Nasal	Alveolar
[zh]	Affricate, unaspirated	Retroflex
[ch]	Affricate, aspirated	Retroflex
[sh]	Fricative	Retroflex
[j]	Affricate, unaspirated	Palatal
[q]	Affricate, aspirated	Palatal
[x]	Fricative	Palatal
[r]	Approximant	Retroflex

Table A.4: The 53 Cantonese Finals (labeled in the Jyut-Ping scheme).

NUCLEUS	CODA								
	Null	[i]	[u]	[p]	[t]	[k]	[m]	[n]	[ng]
[aa]	[aa]	[aai]	[aau]	[aap]	[aat]	[aak]	[aam]	[aan]	[aang]
[a]		[ai]	[au]	[ap]	[at]	[ak]	[am]	[an]	[ang]
[e]	[e]	[ei]				[ek]			[eng]
[i]		[i]	[iu]	[ip]	[it]	[ik]	[im]	[in]	[ing]
[o]	[o]	[oi]	[ou]		[ot]	[ok]		[on]	[ong]
[u]	[u]	[ui]			[ut]	[uk]		[un]	[ung]
[yu]	[yu]				[yut]			[yun]	
[oe]	[oe]	[eoi]			[eot]	[oek]		[eon]	[oeng]
							[m]		[ng]

Table A.5: The 37 Mandarin Finals (labeled in the Pinyin scheme).

NUCLEUS	CODA											
	Null	[e]	[a]	[ei]	[ai]	[ou]	[ao]	[en]	[an]	[eng]	[ang]	[er]
[i]	[i]	[ie]	[ia]			[iu]	[iao]	[in]	[ian]	[ing]	[iang]	
[u]	[u]	(u)o	[ua]	[ui]	[uai]			[un]	[uan]	[ong]	[uang]	
[ü]	[ü]	[üe]						[ün]	[üan]	[iong]		

Hilbert Transform for Temporal Cues Extraction

In order to obtain the envelope from the filtered output from one channel, a analytic signal is generated by the following equation:

$$s(t) = s_r(t) + is_i(t) \tag{A.1}$$

where $s_r(t)$ is the filter output in one channel, $s_i(t)$ is the Hilbert transform of $s_r(t)$ at time t, and i is the imaginary number (*i.e.*, square root of -1).

The Hilbert envelope is the magnitude of the analytic signal [Ville, 1948]:

$$a(t) = |s(t)| = \sqrt{s_r^2 + s_i^2} \tag{A.2}$$

The Hilbert fine structure is $\cos\phi(t)$, where $\phi(t)$ is the called the instantaneous phase of the analytic signal:

$$\phi(t) = \arctan\frac{s_i(t)}{s_r(t)} \tag{A.3}$$

And the derivative of $\phi(t)$ produces the instantaneous frequency of the signal, which is time varying:

$$f = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{A.4}$$

Specially, the original band-passed signal can be recovered as:

$$s_r(t) = a(t)\cos\phi(t) \tag{A.5}$$

In practice, the Hilbert transform is often combined with the band-pass filtering process by the use of complex filters of which real and imaginary parts are in quadrature, such as in a Fast Fourier Transform (FFT).

Table of Cut-off Frequencies of Band-pass Filters (in Hz)

Number of Frequency Bands					
1-band	2-band	4-band	8-band	16-band	32-band
60	60	60	60	60	60
4000	804	302	159	105	82
	4000	804	302	159	105
		1844	508	224	131
		4000	804	302	159
			1230	395	190
			1844	508	224
			2727	642	261
			4000	804	302
				997	347
				1230	395
				1509	449
				1844	508
				2245	572
				2727	642
				3306	719
				4000	804
					896
					997
					1180
					1230
					1363
					1509
					1669
					1844
					2035
					2245
					2475
					2727
					3004
					3306
					3637
					4000

□ **End of chapter.**

Bibliography

Acoustical terminology si, 1-1960, 1960.

A. S. Abramson. Static and dynamic acoustic cues in distinctive tone. *Lang. Speech*, 21:319 – 325, 1978.

P. F. Adams and A. M. Hardy. Current estimates from the national health interview survey: United states, 1988. *Vital Health Stat. 10.*, 10(173):Hyattsville, Md.: National Center for Health Statistics, 1989.

S. Ahmadi and A. S. Spanias. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. *IEEE Trans. Speech Audio Proces.*, 7:333–338, 1999.

ANSI. American national standard acoustical terminology, 1994.

F. Apoux and S. Bacon. Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *J. Acoust. Soc. Am.*, 116:1671–1680, 2004.

F. Apoux, O. Crouzet, and C. Lorenzi. Temporal envelope expansion of speech in noise for normal-hearing and hearing-impaired listeners: effects on identification performance and response times. *Hear. Res.*, 153(1-2):123–131, Mar 2001.

F. Apoux, N. Tribut, X. Debruille, and C. Lorenzi. Identification of envelope-expanded sentences in normal-hearing and hearing-impaired listeners. *Hear. Res.*, 189:13–24, 2004.

D. K. K. Au. Effects of stimulation rates on Cantonese lexical tone perception by cochlear implant users in hong kong. *Clin. Otolaryngol.*, 28:533–538, 2003.

T. Baer and B. Moore. Effects of spectral smearing on the intelligibility of sentences in noise. *J. Acoust. Soc. Am.*, 94:1229 – 1241, 1993.

T. Baer and B. Moore. Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *J. Acoust. Soc. Am.*, 95:2277 – 2280, 1994.

J. G. Barry, P. J. Blamey, L. F. A. Martin, K. Y. S. Lee, T. Tang, Y. Y. Ming, and C. A. .V. Hasselt. Tone discrimination in cantonese-speaking children using a cochlear implant. *Clin Linguist Phon*, 16(2):79–99, Mar 2002.

R. S. Bauer. *Modern Cantonese Phonology*, volume 103, chapter 2, page 109. New York: Mouton de Gruyter, 1997.

D. Blicher, R. Diehl, and L. Cohen. Effects of syllable duration on the recognition of the Mandarin tone 2/tone 3 distinction: evidence for auditory enhancement. *J. Phonetics*, 18:37 – 49, 1990.

P. Boersma and D. Weenink. Praat doing phonetics by computer. Technical report, Version 4.1.2., 1992 – 2003.

Y. R. Chao. *Cantonese Primer.* Cambridge: Cambridge University Press, 1947.

C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen. New methods in continuous mandarin speech recognition. In *Proc. Eurospeech*, pages 1543–1546, 1997.

M. Y. Chen. *Tone Sandhi: Patterns Across Chinese Dialects.* Cambridge University Press, 2000.

A. C. Chin. *Quantitative and Computational Studies on the Chinese Language.* Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong, 1998.

A. Choi. Real-time fundamental frequency estimation by least-square fitting. *IEEE Trans. Speech Audio Process.*, 5(2):201–205, 1997.

V. Ciocca, A. L. Francis, R. Aisha, and L. Wong. The perception of Cantonese lexical tones by early-deafened cochlear implantees. *J. Acoust. Soc. Am.*, 111: 2250–2256, 2002.

P. M. Clarkson and S. F. Bahgat. Envelope expansion methods for speech enhancement. *J. Acoust. Soc. Am.*, 89(3):1378–1382, 1991. doi: 10.1121/1.400538. URL `http://link.aip.org/link/?JAS/89/1378/1`.

J. D. Cutnell. *Physics*, chapter 16, page 466. New York: Wiley, 1998.

P. B. Denes and E. N. Pinson. *The speech chain: the physics and biology of spoken language.* Garden City, N.Y.: Anchor Press, 1973.

M. F. Dorman, P. C. Loizou, and D. Rainey. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J. Acoust. Soc. Am.*, 102(4):2403–2411, Oct 1997.

M. F. Dorman, P. C. Loizou, and D. Rainey. The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels. *J. Acoust. Soc. Am.*, 104:3583–3585, 1998.

H. Dudly. *The vocoder.* Bell Labs Record, 1939.

Ethnologue. Chinese, Mandarin: A language of china. URL $http : //www.ethnologue.com/show_language.asp?code = CHN$. On-line entry, 2004a.

Ethnologue. Chinese, yue: A language of china. URL $http : //www.ethnologue.com/show_language.asp?code = yue$. On-line entry, 2004b.

K. Fishman, R. V. Shannon, and W. H. Slattery. Speech recognition as a function of the number of electrodes used in the speak cochlear implant speech processor. *J. Speech Lang. Hear. Res.*, 40:1201 – 1215, 1997.

A. C. Y. Y. Fok. A perceptual study of tones in Cantonese. In *Hong Kong: Centre of Asian Studies*. University of Hong Kong, 1974.

A. L. Francis, V. Ciocca, and B. K. C. Ng. On the (non)categorical perception of lexical tones. *Percept. Psychophys.*, 65(7):1029–1044, Oct 2003.

R. L. Freyman and G. P. Nerbonne. Consonant confusions in amplitude-expanded speech. *J. Speech Hear. Res.*, 39(6):1124–1137, Dec 1996.

L. Friesen, R. Shannon, D. Baskent, and X. Wang. Speech recognition in noise as a function of spectral channels: comparison of acoustic hearing and cochlear implants. *J. Acoust. Soc. Am.*, 110:1150–1163, 2001.

Q.-J. Fu. Temporal processing and speech recognition in cochlear implant users. *Neuroreport*, 13(13):1635–1639, Sep 2002.

Q.-J. Fu and R. V. Shannon. Recognition of spectrally degraded speech in noise with nonlinear amplitude mapping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 369–372, 15–19 March 1999. doi: 10.1109/ICASSP.1999.758139.

Q. J. Fu and F.-G. Zeng. Identification of temporal envelope cues in Chinese tone recognition. *Asia Pac. J. Speech, Lang. Hearing*, 5:45–57, 2000.

Q. J. Fu, R. V. Shannon, and X. Wang. Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *J. Acoust. Soc. Am.*, 104:3586–3596, 1998a.

Q. J. Fu, F.-G. Zeng, R. V. Shannon, and S. D. Soli. Importance of tonal envelope cues in Chinese speech recognition. *J. Acoust. Soc. Am.*, 104(1): 505–510, 1998b.

Q.-J. Fu, S. Chinchilla, and J. J. Galvin. The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users. *J. Assoc. Res. Otolaryngol.*, 5 (3):253–260, Sep 2004a. doi: 10.1007/s10162-004-4046-1. URL `http://dx.doi.org/10.1007/s10162-004-4046-1`.

Q. J. Fu, C. J. Hsu, and M. J. Horng. Effects of speech processing strategy on Chinese tone recognition by nucleus-24 cochlear implant users. *Ear Hear.*, 25 (5):501–508, Oct 2004b.

J. Gandour. Tone dissimilarity judgments by Chinese listeners. *J. Chin. Linguist.*, 12:235 – 260, 1984.

E. Garding, P. Kratochvil, J. O. Svantesson, and J. Zhang. Tone 4 and tone 3 discrimination in modern standard Chinese. *Lang. Speech*, 29:281 – 283, 1986.

C. Garnham, M. O'Driscoll, R. Ramsden, and S. Saeed. Speech understanding in noise with a med-el combi 40+ cochlear implant using reduced channel sets. *Ear Hear.*, 23:540–552, 2002.

L. Geurts and J. Wouters. Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants. *J. Acoust. Soc. Am.*, 109(2):713–726, 2001.

K. Gfeller, S. Witt, M. Adamek, M. Mehr, J. Rogers, J. Stordahl, and S. Ringgenberg. Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. *J. Am. Acad. Audiol.*, 13 (3):132–145, Mar 2002.

K. E. Gfeller, C. Olszewski, C. Turner, B. Gantz, and J. Oleson. Music perception with cochlear implants and residual hearing. *Audiol. Neuro-Otol.*, 11 Suppl 1:12–15, 2006. doi: 10.1159/000095608. URL `http://dx.doi.org/10.1159/000095608`.

B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47(1-2):103–138, Aug 1990.

T. Green, A. Faulkner, and S. Rosen. Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants. *J. Acoust. Soc. Am.*, 112(5):2155–2164, 2002.

T. Green, A. Faulkner, and S. Rosen. Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *J. Acoust. Soc. Am.*, 116(4):2298–2310, 2004.

T. Green, A. Faulkner, S. Rosen, and O. Macherey. Enhancement of temporal periodicity cues in cochlear implants: effects on prosodic perception and vowel identification. *J. Acoust. Soc. Am.*, 118(1):375–385, Jul 2005.

D. D. Greenwood. A cochlear frequency-position function for several species-29 years later. *J. Acoust. Soc. Am.*, 87(6):2592–2605, 1990.

T. Hanna. Discrimination and identification of modulation rate using a noise carrier. *J. Acoust. Soc. Am.*, 91:2122–2128, 1992.

W. M. Hartmann. *Signals, Sound and Sensation (Modern Acoustics and Signal Processing Series).* Springer-Verlag, 1997.

O.-K. Y. Hashimoto. *Phonology of Cantonese.* Cambridge University Press, 1972.

D. Hilbert. *Foundations of the general theory of linear integral calculus.* Teubner, Leipzig, 1912.

S. Ho and J.Woo. *Social and health profile of the Hong Kong old-old population.* Chinese University of Hong Kong, 1994.

L. Hyman. Tone systems. In M. Haspelmath, E. Konig, W. Oesterreicher, and W. Raible, editors, *Language typology and language universals: An international Handbook*, volume 2, pages 1367–1380. Berlin & New York: Walter de Gruyter, 2001.

K. Kasturi and P. C. Loizou. The intelligibility of speech with 'holes' in the spectrum. *J. Acoust. Soc. Am.*, 112(3):1102–1111, 2002.

E. Khouw and V. Ciocca. Perceptual correlates of Cantonese tones. *J. Phonetics*, 35:104 – 117, 2007.

T. Kinjo and K. Funaki. F0 estimation of noisy speech based on complex speech analysis. In *Proc. Digital Signal Processing Workshop, 12th - Signal Processing Education Workshop*, pages 434–437, 2006. doi: 10.1109/DSPWS.2006.265462.

A. Kohlrausch and R. Fassel. The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *J. Acoust. Soc. Am.*, 108:723 – 734, 2000.

Y.-Y. Kong and F.-G. Zeng. Temporal and spectral cues in Mandarin tone recognition. *J. Acoust. Soc. Am.*, 120(5):2830–2840, 2006.

M. Krini and G. Schmidt. Spectral refinement and its application to fundamental frequency estimation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pages 251–254, 2007. doi: 10.1109/ASPAA.2007.4393018.

N. Kunieda, T. Shimamura, and J. Suzuki. Pitch extraction by using autocorrelation function on the log spectrum. *Electronics and Communications in Japan*, 83:90–98, 2000.

M. Lahat, R. J. Niederjohn, and D. A. Krubsack. A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Trans. Acoust. Speech Signal Proces.*, ASSP-35:741–750, 1987.

B. C. Lam. Newborn hearing screening. In *Brainchild*, volume 3, pages 11–13, 2003.

N. Lan, K. B. Nie, S. K. Gao, and F.-G. Zeng. A novel speech processing strategy incorporating tonal information for cochlear implants. *IEEE Trans. Biomed. Eng.*, 51(5):752–760, May 2004.

J. Laneau, B. Boets, M. Moonen, A. van Wieringen, and J. Wouters. A flexible auditory research platform using acoustic or electric stimuli for adults and young children. *J. Neurosci. Methods*, 142(1):

131–136, Mar 2005. doi: 10.1016/j.jneumeth.2004.08.015. URL http://dx.doi.org/10.1016/j.jneumeth.2004.08.015.

J. Laneau, M. Moonen, and J. Wouters. Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants. *J. Acoust. Soc. Am.*, 119(1):491–506, Jan 2006a.

J. Laneau, J. Wouters, and M. Moonen. Improved music perception with explicit pitch coding in cochlear implants. *Audiol. Neuro-Otol.*, 11(1):38–52, 2006b.

K. Y. S. Lee, S. N. Chiu, and C. A. van Hasselt. Tone perception ability of Cantonese-speaking children. *Lang. Speech*, 45(Pt 4):387–406, Dec 2002a.

K. Y. S. Lee, C. A. van Hasselt, S. N. Chiu, and D. M. C. Cheung. Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children. *Int. J. Pediatr. Otorhinolaryngol.*, 63(2):137–147, Apr 2002b.

R. S. Y. Lee, K. S. Ho, K. L. Chua, K. Y. Lee, N. S. C. Wu, and W. M. Chan. Elderly health centres - the first year experience. *The Hong Kong Practitioner*, 24:530–539, 2002c.

T. Lee, W. K. Lo, P. C. Ching, and H. Meng. Spoken language resources for cantonese speech processing. *Speech Commun.*, 36(3):327–342, 2002d. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/S0167-6393(00)00101-1.

Y. J. Li and T. Lee. Perceptual equivalence of approximated Cantonese tone contours. In *Proc. ISCA Interspeech*, pages 2677–2680, 2007.

Z. A. Liang. Auditory perceptual cues in Mandarin tones. *Acta Phys. Sin.*, 26: 85 – 91, 1963.

J. S. Lim. *Speech Enhancement.* Prentice-Hall, Englewood Cliffs,NJ, 1983.

M. C. Lin. The acoustic characteristics and perceptual cues of tones in standard Chinese. *Chin. Yuwen*, 204:182–193, 1988.

Y.-S. Lin, F.-P. Lee, I.-S. Huang, and S.-C. Peng. Continuous improvement in mandarin lexical tone perception as the number of channels increased: a simulation study of cochlear implant. *Acta Oto-Laryngol.*, 127(5):505–514, May 2007. doi: 10.1080/00016480600951434. URL `http://dx.doi.org/10.1080/00016480600951434`.

S. Liu and F.-G. Zeng. Temporal properties in clear speech perception. *J. Acoust. Soc. Am.*, 120(1):424–432, Jul 2006.

T. C. Liu, H. P. Chen, and H. C. Lin. Effects of limiting the number of active electrodes on mandarin tone perception in young children using cochlear implants. *Acta Oto-Laryngol.*, 124(10):1149–1154, Dec 2004.

P. C. Loizou. Introduction to cochlear implants. *IEEE Eng. Med. Biol.*, 18(1): 32–42, 1999.

C. Lorenzi, F. Berthommier, F. Apoux, and N. Bacri. Effects of envelope expansion on speech recognition. *Hear. Res.*, 136:131–138, 1999.

LSHK. *Hong Kong Jyut Ping Characters Table.* Linguistic Society of Hong Kong Press, 1997.

X. Luo and Q.-J. Fu. Importance of pitch and periodicity to Chinese-speaking cochlear implant patients. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages iv–1–4, 2004a.

X. Luo and Q. J. Fu. Enhancing Chinese tone recognition by manipulating amplitude envelope: implications for cochlear implants. *J. Acoust. Soc. Am.*, 116(6):3659–3667, Dec 2004b.

X. Luo and Q.-J. Fu. Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations. *J. Acoust. Soc. Am.*, 120(4):2260–2266, Oct 2006.

X. Luo, Q.-J. Fu, C.-G. Wei, and K.-L. Cao. Speech recognition and temporal amplitude modulation processing by Mandarin-speaking cochlear implant users. *Ear Hear.*, 29(6), Sep 2008. doi: 10.1097/AUD.0b013e3181888f61. URL `http://dx.doi.org/10.1097/AUD.0b013e3181888f61`.

J. K.-Y. Ma, V. Ciocca, and T. Whitehill. Contextual effect on perception of lexical tones in Cantonese. In *Proc. Eurospeech*, pages 401–404, 2005.

J. K.-Y. Ma, V. Ciocca, and T. Whitehill. Effect of intonation on Cantonese lexical tones. *J. Acoust. Soc. Am.*, 102(6):3978–3987, 2006.

J. D. Markel. The sift algorithm for fundamental frequency estimation. *IEEE Trans. Audio Electroacoust*, AU-20:367–377, 1972.

H. J. McDermott. Music perception with cochlear implants: A review. *Trends Amp.*, 8:49–82, 2004.

C. M. McKay, H. J. McDermott, and G. M. Clark. Pitch percepts associated with amplitude-modulated current pulse trains in cochlear implantees. *J. Acoust. Soc. Am.*, 96:2664–2673, 1994.

C. M. McKay, H. J. McDermott, and G. M. Clark. Pitch matching of amplitude-modulated current pulse trains by cochlear implantees: the effect of modulation depth. *J. Acoust. Soc. Am.*, 97(3):1777–1785, Mar 1995.

B. Moore, B. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc.*, 45:224 – 240, 1997.

B. C. J. Moore. *Cochlear hearing loss.* Whurr Publishers Ltd, London, 1998.

B. C. J. Moore, B. R. Glasberg, and R. W. Peters. Relative dominance of individual partials in determining the pitch of complex tones. *J. Acoust. Soc. Am.*, 77(5):1853–1860, 1985. doi: 10.1121/1.391936. URL `http://link.aip.org/link/?JAS/77/1853/1`.

C. B. Moore and A. Jongman. Speaker normalization in the perception of Mandarin Chinese tones. *J. Acoust. Soc. Am.*, 102(3):1864–1877, Sep 1997.

D. A. Nelson, D. J. van Tasell, A. C. Schroder, S. Soli, and S. Levine. Electrode ranking of "place pitch" and speech recognition in electrical hearing. *J. Acoust. Soc. Am.*, 98:1987 – 1999, 1995.

K. Nie, A. Barco, and F.-G. Zeng. Spectral and temporal cues in cochlear implant speech perception. *Ear Hear.*, 27(2):208–217, Apr 2006.

A. M. Noll. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, 41:293–309, 1967.

K. Oh and C. Un. A performance comparison of pitch extraction algorithms for noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 85–88, Mar 1984.

S. Pijl. Labeling of musical interval size by cochlear implant patients and normally hearing subjects. *Ear Hear.*, 18(5):364–372, Oct 1997.

C. Plack and A. Oxenham. *The psychophysics of Pitch*, chapter 2, pages 7 – 55. Springer-Verlag, 2005.

G. R. Popelka and A. M. Engebretson. A computer-based system for hearing aid assessment. *Hear. Instrum.*, 34:6–9, 1983.

D. A. Preves, J. A. Sigelman, and P. R. LeMay. A feedback stabilizing circuit for hearing aids. *Hear. Instrum.*, 37(4):35–36, 1986.

Y. Qian, T. Lee, and Y. Li. Overlapped di-tone modeling for tone recognition in continuous Cantonese speech. In *Proc. Eurospeech*, pages 1845–1848, 2003.

M. K. Qin and A. J. Oxenham. Effects of envelope-vocoder processing on f0 discrimination and concurrent-vowel identification. *Ear Hear.*, 26(5):451–460, Oct 2005.

S. Rosen. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. London, Ser. B*, 336:367–373, 1992.

J. T. Rubinstein. How cochlear implants encode speech. *Curr. Opin. Otolaryngol. Head Neck Surg.*, 12(5):444–448, Oct 2004.

C. Shahnaz, W. P. Zhu, and M. O. Ahmad. Robust pitch estimation at very low snr exploiting time and frequency domain cues. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, number 389-392, 2005.

C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A robust pitch estimation algorithm in noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, number 1073-1076, 2007.

C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A pitch extraction algorithm in noise based on temporal and spectral representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, number 4477-4480, 2008.

G. E. Shambaugh. The theory of sound perception. *J. Acoust. Soc. Am.*, 1: 295–300, 1930.

R. V. Shannon. The relative importance of amplitude, temporal, and spectral cues for cochlear implant processor design. *Am. J. Audiol.*, 11(2):124–127, Dec 2002.

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

R. V. Shannon, F.-G. Zeng, and J. Wygonski. Speech recognition with altered spectral distribution of envelope cues. *J. Acoust. Soc. Am.*, 104(4):2467–2476, 1998.

R. V. Shannon, J. J. Galvin, and D. Baskent. Holes in hearing. *J. Assoc. Res. Otolaryngol.*, 3(2):185–199, Jun 2001.

R. V. Shannon, Q.-J. Fu, and J. Galvin. The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngol. Suppl.*, 552(552):50–54, May 2004.

T. Shimamura and H. Kobayashi. Weighted autocorrelation for pitch extraction of noisy speech. *IEEE Trans. Speech Audio Process.*, 9:727 – 730, 2001.

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

G. S. Stickney, F.-G. Zeng, R. Litovsky, and P. Assmann. Cochlear implant speech recognition with speech maskers. *J. Acoust. Soc. Am.*, 116(2):1081–1091, Aug 2004.

C. M. Sucher and H. J. McDermott. Pitch ranking of complex tones by normally hearing subjects and cochlear implant users. *Hear. Res.*, 230(1-2):80–87, Aug 2007. doi: 10.1016/j.heares.2007.05.002. URL `http://dx.doi.org/10.1016/j.heares.2007.05.002`.

D. J. Tasell, D. G. Greenfield, J. J. Logemann, and D. A. Nelson. Temporal cues for consonant recognition: training, talker generalization and use in evaluation of cochlear implants. *J. Acoust. Soc. Am.*, 92(3):1247–1257, 1992.

D. J. V. Tasell, S. D. Soli, V. M. Kirby, and G. P. Widin. Speech waveform envelope cues for consonant recognition. *J. Acoust. Soc. Am.*, 82(4):1152–1161, Oct 1987.

M. ter Keurs, J. M. Festen, and R. Plomp. Effect of spectral envelope smearing on speech reception i. *J. Acoust. Soc. Am.*, 91(5):2872–2880, May 1992.

M. ter Keurs, J. M. Festen, and R. Plomp. Effect of spectral envelope smearing on speech reception ii. *J. Acoust. Soc. Am.*, 93(3):1547–1552, Mar 1993.

C. W. Turner, P. E. Souza, and L. N. Forget. Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 97(4):2568–2576, Apr 1995.

R. A. van Buuren, J. M. Festen, and T. Houtgast. Compression and expansion of the temporal envelope: evaluation of speech intelligibility and sound quality. *J. Acoust. Soc. Am.*, 105(5):2903–2913, 1999.

A. E. Vandali, C. Sucher, D. J. Tsang, C. M. McKay, J. W. D. Chew, and H. J. McDermott. Pitch ranking ability of CI recipients: a comparison of sound processing strategies. *J. Acoust. Soc. Am.*, 117(5):3126–3138, 2005.

E. Villchur. Electronic models to simulate the effect of sensory distortions on speech perception by the deaf. *J. Acoust. Soc. Am.*, 62(3):665–674, Sep 1977.

J. Ville. Theorie et applications de la notion de signal analytique. *Cables Transmission*, 2:61–74, 1948.

H.-M. Wang, J.-L. Shen, Y.-J. Yang, C.-Y. Tseng, and L.-S. Lee. Complete recognition of continuous mandarin speech for chinese language with very large vocabulary but limited training data. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 61–64, 9–12 May 1995. doi: 10.1109/ICASSP.1995.479273.

W. S.-Y. Wang. *Language change.* New York: New York Academy of Sciences, 1976. Annals of the New York Academy of Sciences, Vol. 280, pp. 61-72.

C. G. Wei, K. L. Cao, and F.-G. Zeng. Mandarin tone recognition in cochlear-implant subjects. *Hear. Res.*, 197:87–95, 2004.

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

B. S. Wilson. Strategies for representing speech information with cochlear implants. In J. K. Niparko, editor, *Cochlear implants-principles and practices*, chapter 7, pages 129–170. Lippincott Willianms and Wilkins, 2000.

B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz. Better speech recognition with cochlear implants. *Nature*, 352:236–238, 1991.

A. O. C. Wong and L. L. N. Wong. Tone perception of Cantonese-speaking prelingually hearing-impaired children with cochlear implants. *Otolaryngol. Head Neck Surg.*, 130(6):751–758, Jun 2004. doi: 10.1016/j.otohns.2003.09.037. URL `http://dx.doi.org/10.1016/j.otohns.2003.09.037`.

L. L. N. Wong, A. E. Vandali, V. Ciocca, B. Luk, V. W. K. Ip, B. Murray, H. C. Yu, and I. Chung. New cochlear implant coding strategy for tonal language speakers. *Int. J. Audiol.*, 47(6):337–347, Jun 2008.

L. Xu and B. E. Pfingst. Relative importance of temporal envelope and fine structure in lexical-tone perception. *J. Acoust. Soc. Am.*, 114(6):3024–3027, 2003.

L. Xu and B. E. Pfingst. Spectral and temporal cues for speech recognition: implications for auditory prostheses. *Hear. Res.*, 242 (1-2):132–140, Aug 2008. doi: 10.1016/j.heares.2007.12.010. URL `http://dx.doi.org/10.1016/j.heares.2007.12.010`.

L. Xu and Y. Zheng. Spectral and temporal cues for phoneme recognition in noise. *J. Acoust. Soc. Am.*, 122(3):1758, Sep 2007. doi: 10.1121/1.2767000. URL `http://dx.doi.org/10.1121/1.2767000`.

L. Xu, Y. Tsai, and B. E. Pfingst. Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses. *J. Acoust. Soc. Am.*, 112(1): 247–258, 2002.

L. Xu, C. S. Thompson, and B. E. Pfingst. Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.*, 117(5):3255 – 3267, 2005.

K. C. P. Yuen, M. Yuan, T. Lee, S. Soli, M. C. F. Tong, and C. A. van Hasselt. Cantonese lexical tone recognition from frequency-specific temporal envelope and periodicity components in the same versus different noise band carriers. In *Asia Pacific Symposium on Cochlear Implant and Related Sciences (APSCI2007)*, 2007.

F.-G. Zeng. Trends in cochlear implants. *Trends Amp.*, 8:1–34, 2004.

F.-G. Zeng, Y.-Y. Kong, H. J. Michalewski, and A. Starr. Perceptual consequences of disrupted auditory nerve activity. *J. Neurophys-*

iol., 93(6):3050–3063, Jun 2005. doi: 10.1152/jn.00985.2004. URL `http://dx.doi.org/10.1152/jn.00985.2004`.