A Perceptual Study on Linearly Approximated F0 Contours in Cantonese Speech

LI, Yujia

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

The Chinese University of Hong Kong
January 2011

To my family and friends

Acknowledgments

First of all I would like to express my deepest gratitude to my thesis supervisor Prof. Tan Lee for his patience and guidance, without which I could never have made this to the end. Prof. Lee has been a great advisor and mentor, providing me enormous support and encouragement, especially during the moment when I was in difficulties of thesis exploration.

Sincere thanks to Prof. Chiu-Yu Tseng, Prof. Helen Meng, Prof. William Shi-Yuan Wang, Prof. Pak-Chung Ching and Dr. Frank Soong, for their constructive comments that improve the quality of this thesis.

Thanks are due to all my colleagues in DSP and speech technology group for their valuable collaboration, assistance and providing me the most enjoyable working environment.

Special gratitude to my friends Mr. Alex Wei Zhang, Ms. Heather Ting, Ma, Ms. Natalie Lai Tsang, Mr. Meng Yuan, Ms. Yvonne Siu-Wa Lee and many others for their kind friendship and inspiration.

Finally, I am deeply grateful to my family for their love, trust and continuous support.

Abstract of thesis entitled:

A Perceptual Study on Linearly Approximated F0 Contours in Cantonese Speech

Submitted by **LI Yujia**
for the degree of Doctor **of Philosophy**
in **Electronic Engineering**
at **The Chinese University of Hong Kong**
in **January 2011**.

F0 variation in speech is known to carry abundant information, both linguistic and paralinguistic. Its impact on speech communication is thus widely concerned. F0 variation in speech, being a major super-segmental acoustic feature, has received a lot of attention, particularly from the perspectives of production-acoustics and perception-acoustics. However, it is noted that perception-acoustic knowledge of F0 variation in association with speech naturalness is quite limited. This is especially the case in the studies of tonal languages, in which most efforts are made on acoustic cues related to tone identification.

F0 contours measured from human speech (observed contours) generally vary to a considerable extent. This research attempts to investigate perception-critical variations in these highly varying contours. In particular, F0 contours in Cantonese speech are concerned. Cantonese is a major Chinese dialect that is known of being rich in tones. Psychoacoustic findings suggest that human perception has limitations in perceiving pitch movements. This means that not all of the variations in the observed contours are perceivable. A major problem addressed in this study is to find the simplest acoustic representation of an observed F0 contour that is adequate to attain comparable perception with the natural speech.

Approximation of F0 contours in Cantonese speech is investigated. Multiple approximations are examined and evaluated. The modified speech utterances that carry the approximated contours at syllable, word and sentence levels are perceptually examined with reference to natural speech. It is found that linear approximation can adequately describe all perception-sensitive F0 variations in Cantonese speech. Each tone contour can be represented by one or two linear movements, and the transition between co-articulated tones can be represented by one linear movement.

The feasibility of using linear approximation greatly simplifies the way to understand and interpret F0 variations in speech processing, by the means of learning the properties of linear movements. Three steps of analysis are carried out on the generated linear approximations. The first one examines the movement slopes in the approximated F0 contours of isolated syllables, in comparison with the perceptual thresholds found in the psychoacoustic studies. The second analysis is performed over a set of linearly approximated F0 contours of polysyllabic Cantonese words. The determining attributes of these linear movements, i.e., movement slopes, movement heights and time locations of turning points are analyzed statistically. The last analysis concerns the evaluation of modified F0 contours. Objective evaluations are compared with perceptual evaluation. These analyses provide knowledge which can improve our understanding on how F0 variations are processed in speech path.

To explore the potentials of linear approximation in research of speech prosody, a perception-oriented framework of automatic approximation is developed, so as to replace the manual process in the feasibility study. The framework aims to make the process of deriving approximations standardized, consistent and efficient. It is formulated based on the experiences from manual approximations and is also implemented with other perceptual findings. The initial test on polysyllabic words gives promising results.

摘要

語音信號中的基頻 (F0) 變化傳遞豐富信息。因此，它在語音交互中的作用一直受到廣泛關注。基頻變化作爲主要的超音段聲學特徵，已經得到深入研究。這些研究的視角多是基於聲學特徵與發聲機制之間的聯係，或是聲學特徵與感知機制之間的聯係。盡管如此，我們對於聽者感受語音自然度的這個過程中基頻變化與感知機制之間的關係依然所知甚少。這個問題在對聲調語言的研究中尤其明顯，因爲大部分以感知機制爲主的研究更關注影響聲調識別的基頻變化。

從説話人語音信號中測量得到的基頻曲綫 (observed F0 contour) 變化十分複雜。本研究致力於從這些多變曲綫中捕捉對於人類感知系統真正起作用的那些變化。我們集中研究粵語語音中的基頻曲綫。粵語是中國的主要方言之一。聲調變化尤其豐富。感知聲學研究發現人類感知系統對於感知音調變化是有限度的。這意味著測量曲綫中的變化並不都是可感知的。因此，本論文主要要解決的一個問題就是為測量曲綫尋找一個最簡單的表征。同時要保證這種簡化的基頻表征不會對語音信號的感知產生任何負面影響。

我們以近似 (approximate) 測量曲綫為方法，為上述問題尋找答案。我們研究了不同的近似方式 (approximation)。這些近似方式被應用于字，詞，句層面的測量曲綫。依此而重新合成的語音會與説話人語音進行感知上的比較。研究發現綫性近似已足夠描述測量曲綫中所有對感知系統重要的變化。每一個粵語聲調曲綫可以用一或兩個綫性變化 (linear movement) 來描述。聲調與聲調之間的過渡曲綫可以用一個綫性變化來描述。

綫性近似讓我們在理解和解釋基頻變化方面更加容易。本論文對於實驗中所產生的綫性近似進行了分析研究。首先，我們分析了分離單音節的近似曲綫中綫性變化的斜率。關注其與感知的關係，並與感知聲學研究中發現的感知域值進行了比較。第二個分析基於幾百個多字詞的近似基頻曲綫。我們對這些近似曲綫中的綫性變化進行了分析。特別對它們的變化斜率，高度，以及銜接點的時間位置進行了詳細分析。最後一個分析關注的問題是對基頻曲綫的主觀評估與客觀評

估。客觀評估的結果與感知實驗的結果進行了對照比較。這些分析結果增強了我們對於語音交互系統處理基頻變化的認識。

為有助於綫性近似在語音韻律研究中發揮積極作用，我們設計了一個以感知為導向可自動近似測量曲綫的算法，以取代在可行性研究中一直採用的人工近似法。該算法旨在使近似過程標準化，更有效率性和一致性。算法設計上參考了人工近似法的經驗，也同時應用了其他感知學研究成果。我們于多字詞的測量曲綫上對這個算法進行了初步檢驗，結果理想。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech, as a fundamental and natural means of communication, is produced by a speaker and perceived by a listener. Acoustic signal is the actual physical medium via which speech information is delivered during the process. With the progress on research of acoustics and the ever-increasing power of computers, acoustic signals now become easily measurable and controllable by signal processing methods (Boersma and Weenink, 2010). This conversely allows studies about the relations between human speech production and acoustics, between human speech perception and acoustics, and between human speech production and perception, by means of modifying the acoustic signals. Findings of such studies will greatly improve our understanding on how human process speech.

Fundamental frequency (F0) is an important acoustic feature which is most related to the perceived pitch in sound signals. It is particularly significant in human speech. As an important component of speech prosody, the temporal variation of F0 is a major acoustic manifestation of super-segmental features like lexically defined tone patterns, accents and stresses. The F0 contours also carry paralinguistic information like intonation, focus and speaker's emotional state. These super-segmental features are found to determine the perceived speech naturalness (Fujiaski and Hirose, 1984; Kochanski and Shih, 2003; Li, Lee and Qian, 2004; Vaissiere, 2004; Fujisaki et al, 2005; Jongman et al, 2006).

Cantonese is a major Chinese dialect. It is a tonal language and is well known for its richness in tones. In this thesis, F0 contours in Cantonese speech are studied. The investigation focuses on perception-sensitive variations in the highly complicated surface contours. It is argued that simplified contours hold many advantages in processing and understanding F0 data in speech.

1.1. Perception and Production

Speech is transmitted via a path from human production to acoustics and from acoustics to human perception. Being the two ends in the path, perception and production depend on each other, as they are connected to the same acoustic signal. It is believed that, for most linguistic and paralinguistic events in speech, there exist underlying acoustic targets for the production system to reach, and for the perception system to recognize. Training effects between production and perception have been found to play a role in the process of speech communication (Wang et al, 1999; Wang, Jongman and Sereno, 2003). Such effects are expected to be more important in the process of paralinguistic events due to the lack of clear linguistic definitions on them.

Despite their close relation, speech production and speech perception are accomplished by different physical systems of human beings (Denes and Pinson, 1998). Their interactions with acoustics appear to be very different. For example, in tone production, it was found that the physical limitations of human articulators restrict the changing rate of F0 in the acoustic signal (Xu and Sun, 2002); in tone perception, the same F0 contour can be perceived as contrastively different tones, if the contexts are different (Xu, 1994). The change of perceived speaking rate does not mean pro-rata duration variation of speech units over the entire utterance. Duration change on stressed syllables was found to play a more important role (Pasdeloup, Espesser and Faraj, 2006).

It is believed that the non-trivial relation between perception and production should be treated seriously. There are several considerations. First, many variations in natural speech are just inevitable, due to the physical constraints of human production system. These variations do not serve to satisfy human perception. In (Xu, 2004), it was shown that the complex tone contour in natural human speech might not be speaker intended and it may not be perceptually desirable either, but rather articulately unavoidable. In (Janse, 2004), an investigation on speeded-up speech revealed that naturally-fast speech is perceptually less preferred and less intelligible than a linearly speeded-up version. The author argued that the natural prosodic patterns in fast speech is not aimed at helping the listeners but because the speakers could not speed up in any other way. Second, human perception might not be able to detect every minute variation in human produced acoustics. Previous research suggested that perception has limited resolution in differentiating acoustics (t' Hart, Collier and Cohen, 1990). Hence, the natural acoustic realization of an underlying target can be simplified and this would not create perceptual difference. Third, simplified acoustics might be perceptually more preferred than the original one. A view of selective perception states "*the biological resources available for perceptual processing of the speech signal are limited and must therefore be allocated to various tasks in a restricted manner*" (House, 2004). This view is experimentally supported by the work of (House, 1990), which revealed that pitch sensitivity decreases when the complexity of speech signals increases. Therefore, it is considered that simplification of the acoustic realization of a target can reduce the perceptual workload on processing it, and consequently enhance the perceptual capability on processing other tasks.

Based on these considerations, this study focuses on the perception-acoustic relation and investigates F0 contours in speech. Better understanding about perception of F0 contour and its use in real applications are expected.

1.2. Perception of F0 Variation

1.2.1. Pitch and F0

Pitch was defined as "*that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale*" [American Standards Association, 1960]. In other words, pitch variations are closely correlated with the sense of music melody. Pitch is a subjective attribute, which is perceived but not measured physically. Nevertheless, perceived pitch is related to the rate of waveform repetition in the acoustic sound. If the sound is a pure tone with sinusoidal waveform, pitch corresponds to the frequency of the waveform. If the sound contains a complex tone or has a quasi-periodic waveform, pitch is related to the fundamental frequency (F0) of the waveform, as illustrated in Figure 1.1. Quantitative description of the pitch of a sound generally refers to finding the frequency of a pure tone (sine wave) that has the same subjectively perceived pitch (Moore, 2004). Psychoacoustic studies found that pitch is approximately a linear function of F0 below 500 Hz. Above this level, perceived pitch corresponds more to a logarithmic scale of F0 (Jongman et al, 2006).

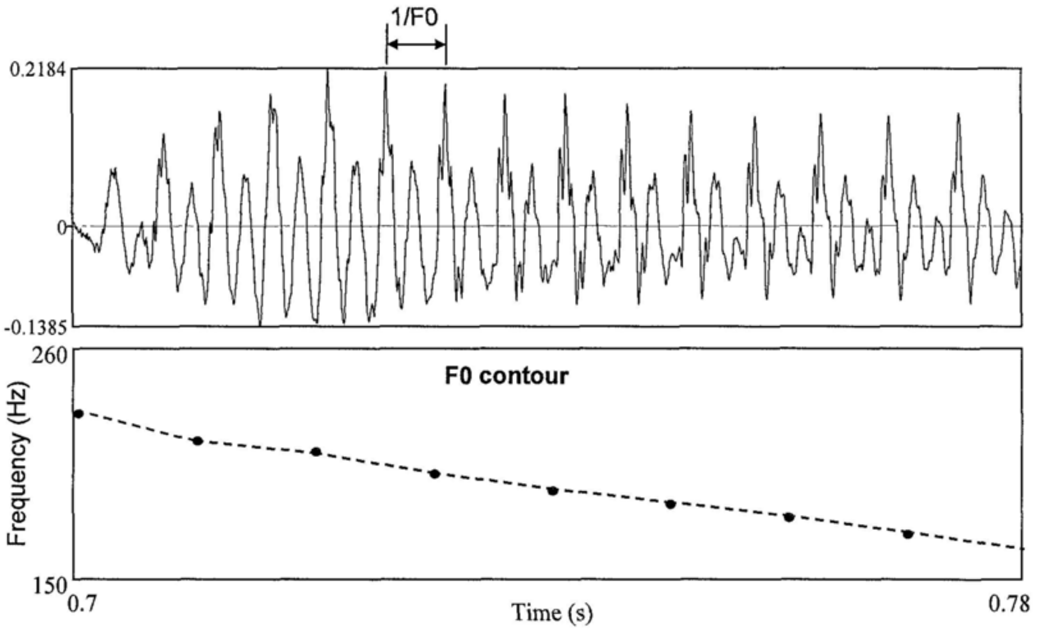


FIG. 1.1: F0 and F0 contour measured from a segment of sound waveform. The upper figure illustrates the measurement of F0 from a quasi-periodic waveform. The lower figure shows the F0 contour that is time-aligned with the waveform.

The values of F0 over a sound segment might be constant or varying. Typical computational algorithms for F0 extraction make estimations from the waveform with a time resolution of 10 ms (also known as the frame size), i.e., one F0 value is calculated every 10 ms. Along the time axis, we obtain an F0 contour aligned with the waveform, as exemplified in Figure 1.1. Perception of F0 contour is a complex process, not a simple mapping from F0 to pitch. For a contour with constant F0 value, most likely only one pitch level is perceived. If the contour is formed by time-varying F0 points, multiple pitch movements might be perceived, depending on the magnitude of the change and also the detection ability of human perception system. In the next section, the perceptual thresholds of detecting F0 variations will be discussed, in order to provide the psychoacoustic knowledge on how an F0 contour will be perceptually processed.

1.2.2. Thresholds of perceiving F0 variations

The range of F0 that can lead to perception of pitch is from 40 Hz to 4,000 Hz. For the frequencies beyond 4,000 Hz, the accuracy with which pitch can be judged decreases sharply (Henning, 1966). Pitch perception is influenced by the duration of a sound as well. Studies found that, to produce a reliable pitch perception, a stimulus needs to be longer than 30 ms (Cardozo and Ritsma, 1965; ‘t Hart, Collier and Cohen, 1990). On the other hand, if the sound duration is too long, human memory ability would pose a problem. Within the general limitations, the task of absolute pitch, i.e., telling the pitch of one tone presented in isolation, is found to be very difficult for human subjects (‘t Hart, Collier and Cohen, 1990). Therefore, most psychoacoustic studies focused on investigating the perceptual abilities on differentiating F0 differences. In psychoacoustic studies, the stimuli are generally steady ones, e.g. pure tones or complex tones.

1.2.2.1 The differential threshold of pitch

The differential threshold of pitch concerns the just-noticeable frequency difference (j.n.d.) between two successively presented (pure) tones (‘t Hart, Collier and Cohen, 1990). This threshold is influenced by both the reference frequency[1] and the duration of the stimuli. If taking the j.n.d. relative to the reference frequency as sensitivity, it is found that at reference frequency lower than 1,000 Hz, the sensitivity to frequency

[1] In comparison of the two presented tones, generally frequency of one tone is fixed, and frequency of the other tone is adjusted. The fixed frequency here is referred as the reference frequency in determination of the differential threshold of pitch.

difference becomes lower. For instance, when the reference frequency is 1,000 Hz, $\Delta f / f$ is found to be 0.54×10^{-3}; while this value is 1.36×10^{-3} when the reference frequency is 125 Hz (Nordmark, 1968). Furthermore, the precision of pitch sensation decreases as the stimulus duration decreases. For example, at a reference frequency of 1,000 Hz, $\Delta f / f$ of a 30-ms stimulus is more than double of that when the stimuli duration is 60 ms or longer (Cardozo and Ritsma, 1965).

The differential threshold was also examined with speech signals. The outcomes are discrepantly from 0.3%~0.5% (Flanagan and Saslow, 1958) to 4%~5% (Issachenko and Schädlich, 1970; Rossi and Chafcouloff, 1972) of a frequency region representative of male speech (140 Hz to 200 Hz).

1.2.2.2 The differential threshold of pitch distance

The differential threshold of pitch distance concerns the conditions in which one interval of pitch distance is perceived larger or smaller than the other interval ('t Hart, Collier and Cohen, 1990). Measurement of the threshold is a rather difficult task. There were large intra-subjects or inter-subjects variations of the results (Stevens and Volkman, 1940; Plomp, Wagenaar and Mimpen, 1973). In some conditions, musically trained subjects could occasionally hear differences below one semitone[2]. In speech signal, the threshold was measured by ('t Hart, 1981) using untrained subjects. The outcomes suggested that only pitch differences over three semitones can play a part in communicative situations.

[2] Definition and calculation of semitone is given in Appendix 1.

1.2.2.3 The absolute threshold of pitch change (glissando threshold)[3]

The absolute threshold of pitch change concerns the pitch variation over a given interval of time, i.e., how rapidly F0 should change in order to evoke a sensation of pitch change ('t Hart, Collier and Cohen, 1990). In the measurement of this threshold, it was argued whether the sensitivity to frequency changing rate or merely the sensitivity to frequency difference is measured. Some psychoacoustic experiments examining this threshold indicated that the total frequency change is independent of duration, supporting the hypothesis that the frequency difference is the important perceptual factor (Pollack, 1968). However, other experiments exhibited higher sensitivity with the longer-sweep tones (Sergeant and Harris, 1962). In summary, psychoacoustic studies found different relations between F0 changing rate and stimuli duration or between F0 changing rate and initial frequency (Shower and Biddulph, 1931; Sergeant and Harris, 1962; Pollack, 1968; Schouten, 1985). The threshold in speech-like signals was measured by (Rossi, 1971, 1978) and (Klatt, 1973). Being described in terms of speed of F0 change, the outcomes from these experiments could not agree with each other well.

The experimental data (Sergeant and Harris, 1962; Pollack, 1968; Rossi, 1971, 1978; Klatt, 1973; Schouten, 1985) were later investigated by the researchers of the Institute for Perception Research (IPO) ('t Hart, Collier and Cohen, 1990). The following equation is found to fit most of the data,

[3] Glissando is a term used in music. It represents the pitch gliding, i.e. rising or falling.

$$G_{thr} = 0.16 / T^2 \tag{1.1}$$

in which G_{thr} is the speed of frequency change at threshold in semitone/second, and T is the stimulus duration in second, varied from 20 ms to 5 seconds. IPO further examined the threshold in speech signals using synthetic vowel segments as stimuli. The experimental data could be well described by Equation 1.1 as well. The glissando threshold is rather consistent in psychoacoustic and "psychophonetic"[4] perception.

1.2.2.4 The differential threshold of pitch change (differential glissando threshold)

The differential threshold of pitch change concerns how different the slopes of two successively presented F0 glides should be in order to be just audibly distinguishable ('t Hart, Collier and Cohen, 1990). This threshold was initially measured by (Pollack, 1968) and (Nabelek and Hirsh, 1969) using sine waves. It was later investigated for speech by (Klatt, 1973) using synthetic vowels, and by IPO ('t Hart, Collier and Cohen, 1990) using VCVCV[5] sequences. Durations of all the stimuli are less than 1 second. The findings from psychoacoustic and psychophonetic studies were found to be consistent. Given g_1, the slope of one F0 glide, and g_2, the slope of the other F0 glide, most of the experimental data gave a threshold around $g_1 / g_2 = 2$, irrespective of the slopes of individual glides. It was concluded that "*in actual speech two just*

[4] The so called "psychophonetic" is used to differentiate from psychoacoustic studies when the tested stimuli use speech segments as carriers.

[5] V represents a vowel and C represents a consonant.

distinguishable rates of change of F0 will differ by a factor of at least two" ('t Hart, Collier and Cohen, 1990: 35).

1.3. Perception of F0 Contours in Speech

F0 is measurable for the voiced speech signals that have quasi-periodic waveforms. The value of F0 changes continuously over a speech segment. As a result, a temporally varying contour can be observed, excluding the regions in which the speech signals are unvoiced. An example of the F0 contour measured from an English utterance is shown in Figure 1.2.

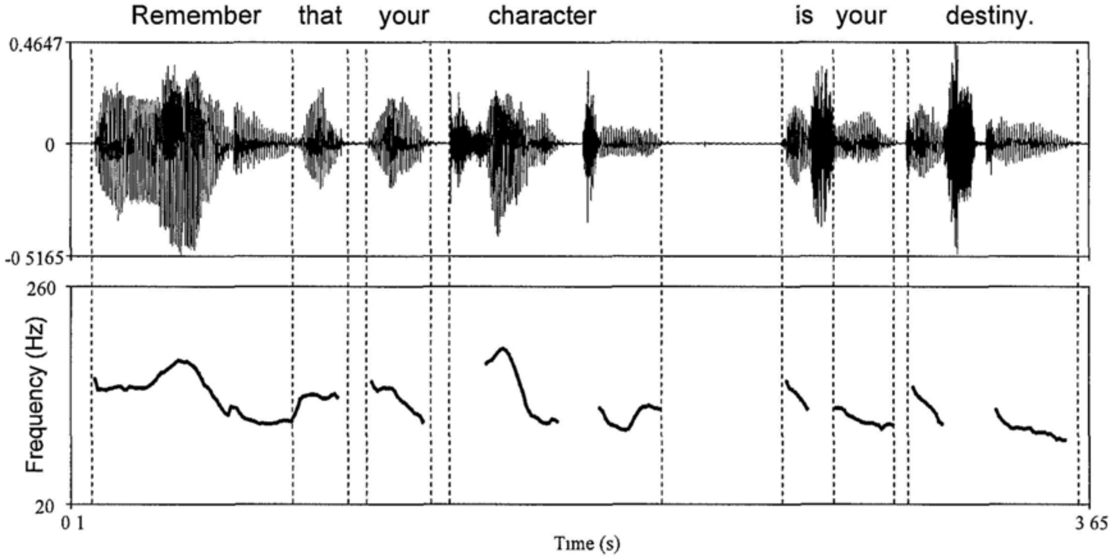


FIG. 1.2: The waveform (upper figure) and the F0 contour (lower figure) of an English utterance. Word boundaries are indicated by the dashed lines.

Perception of F0 variation in speech should be highly correlated with psychoacoustic perception of F0 variation, but the two are certainly not identical. Psychoacoustic research is set out to explore the extreme performance of human perception system. It is reasonable to expect that such utmost performance is not necessarily to be reached in daily life. Compared with the steady stimuli used in the psychoacoustic research, F0 variation in speech is different in the following three aspects. First, F0 variation in speech forms a going-on stream. Second, F0 contours carry major communicative functions. Third, it is perceptually processed in conjunction with other dimensions of variations. Hence, perception of F0 variation in speech is an important and interesting research topic.

1.3.1. Perceptual segmentation of F0 contour

F0 perception is a part of speech perception. F0 variations are always processed together with other speech information, such as segmental information. In real communication, the stream of speech goes on continuously. Due to the limited memory, speech must be processed piece by piece, and inevitably, segmentation is needed at an early stage of speech perception. In speech perception, syllable[6] is considered to be the basic unit (Ryalls 1996). F0 contour of a stream of speech, as a co-occurring physical property, is expected to be perceived on a syllable-size basis too. This hypothesis was experimentally proved by (House, 1990). His study revealed how the spectral and amplitude variations in speech jointly create a perceptual segmentation of pitch sequence into syllable-sized chunks. The perceived F0 variations in speech thus become locally correlated with syllables. Then the segmented F0 contours will be

[6] A syllable is a unit of organization for a sequence of speech sounds. A syllable is typically made up of a syllable nucleus (most often a vowel) with optional initial and final margins (typically, consonants). For example, the word “paper” is composed of two syllables: “pa” and “per”.

further perceptually interpreted into higher-level linguistic codes or prosodic properties.

1.3.2. Perceiving F0 variations with communicative functions

F0 contours in speech carry abundant information that is related to the prosodic properties of speech. Some of the properties are linguistically defined in the lexicon of a language, such as accent in Japanese, stress in English and tone in Chinese. F0 contours carrying such functions are locally aligned with syllables and approach some underlying acoustic targets that can be identified by perception. The other properties are connected with communicative functions that are linguistically irrelevant. They are referred as paralinguistic properties, for examples, intonation and speakers emotions. F0 contours conveying these functions are linked with speech units that are much larger than syllables, e.g., phrases, or sentences. Perception of paralinguistic functions is based on the perception of syllable-sized F0 contours and involves a higher-level abstraction. Local pitch movements are memorized and integrated into a meaningful interpretation in the process (Pierrehumbert, 1979; Thorsen, 1980; Wales and Taylor, 1987; House 1990). For the above communicative functions, F0 variation is an important but not the only cue for perception. Other prosodic parameters like duration and intensity often play a role.

As giving absolute pitch is very difficult, perceptual process of F0-related targets is often carried out in a relative sense, i.e., comparing F0 variations within a time interval. For local targets, such as lexical tones, the F0 variations of neighboring syllables might be referred. If the target is global, such as intonation, the comparison becomes more complex with the help of memory. Such comparison is particularly inevitable for the properties that have no clear linguistic definitions.

Compared with psychoacoustic perception, the perceptual sensitivity to F0 variations in speech might be decreased. This was evidenced in the studies of measuring different thresholds on detecting F0 variations (Section 1.2.2). When speech stimuli were used, the measured threshold was mostly higher than that derived from pure-tone or complex-tone stimuli. In speech perception, perception of F0 variations is not the only isolated goal. According to the view of selective perception (House, 1990), in the case of processing multiple tasks, the biological resources available for processing individual tasks decrease as the number of tasks increases. This implies a decreased perceptual sensitivity on F0 variation when it is a sub-task of speech perception. Communicative function of F0 variation is another factor that might cause decreased sensitivity. It was found that "*the pitch changes that are linguistically relevant are much larger than the limits of pitch discrimination measured psychophysically using steady stimuli*" (Moore, 1997: 557). This is supported by a universal tendency in tonal languages that only a limit number of pitch levels are used in their tone systems, despite that pitch discrimination is known to be very acute (Connell, 2000). In the studies of Mandarin tones, it was argued that, in order to differentiate between the high level and high rising tones, listeners had learned to suppress a natural auditory sensitivity to the difference between rising and "truly" level pitch contour, so as to facilitate the acquisition of a similarly heightened sensitivity to differences between rapidly rising and flat pitch glides. Thus, listeners were best served by shifting the location of an intrinsic sensitivity to the distinction between rising and level pitch glides (Wang, 1976; Stagray and Downs, 1993; Francis, Ciocca and Ng, 2003).

F0 contours are also processed together with other speech information. It is an important research question how the time alignment of F0 variations is in relation to segmental and syllable boundaries, and hence to realize the communicative functions. Perceptual studies found that in languages, a sequence of F0 movements could be created and perceptually associated with different categories of a communicative function, by changing the timing of these movements (House, 1997, 2004). For example, in Swedish, by shifting two F0 peaks along an utterance, four different dialect types were perceived (Bruce, 1983). Shifting F0 peaks was also found to affect the perception of prosodic phrase categories (Gårding and Eriksson, 1991), statement or question intonations (D'Imperio and House, 1997), and even the meaning of a word (Kohler, 1987). Shifting F0 rising or falling movements similarly leads to the perception of different categories (Bruce, 1977; House, 1990; Hermes, 1996). Many of the perceptual experiments seem to suggest that in the shifting, a timing difference of 50 ms is the threshold to change the perceptual interpretation of the whole contour.

1.3.3. Tone perception

In tonal languages, tone is a part of lexicon to determine the meaning of a spoken word and it is realized by aligning a distinctive pitch pattern over a syllable (Pike, 1948; Clark and Yallop, 1990). A slight majority of the languages in the world are tonal. There are many in East Asia, such as Chinese, Vietnamese and Thai. The tone system of a language consists of a limited number of tone patterns, usually no more than five (Maddieson, 1978). A tone pattern is defined by either a pitch level (level tone) or pitch movements (contour tone). Mandarin tones as shown in Figure 1.3 are considered to represent a contour tone system, in which each tone has a distinct shape of contour. On the other hand, most Bantu language[7] have register tone systems, where the distinguishing feature is the relative difference in pitch levels, e.g., high, mid, or low. Generally speaking, contour tones are more easily identified than level tones, since many acoustic cues other than pitch height can be referred (Khouw and Ciocca, 2007).

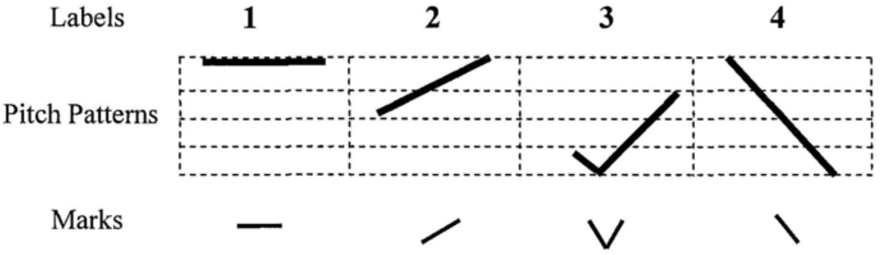


FIG. 1.3: Pitch patterns of four Mandarin tones.

Tone perception is a primary task of perceiving F0 variations in the speech of tonal languages. It is also a highly concerned issue in speech perception, as linguistic and paralinguistic knowledge are both carried by tonal F0 contours. Research on tone perception could improve our understanding about human perceptual process of super-segmental and prosodic features in speech.

Studies on tone perception mainly go into two branches. One branch cares about the acoustic cues that are used in human perception to differentiate or identify tones. In particular, Mandarin, Cantonese, Thai have received most attention. The acoustic cues are language-dependent and, to a large extent, are determined by how a tone system is defined. In the identification of Mandarin tones, Mandarin-speaking listeners seem to attach more importance to pitch contour than pitch level (Gandour, 1984; Massaro, Cohen and Tseng, 1985; Wang et al, 1999; Jongman et al, 2006). The turning point at

[7] Bantu is a group of languages of Africa.

which a contour changes from falling to rising as well as the F0 range from the onset of the tone to this turning point were found to be important cues (Gårding et al, 1986; Blicher, Diehl and Cohen, 1990; Shen and Lin, 1991; Shen, Lin and Yan, 1993; Moore and Jongman, 1997). In addition, duration has also been shown to affect Mandarin tone identification (Blicher, Diehl and Cohen, 1990). On the other hand, in tone perception of Cantonese or Thai where the tone system contains several level tones, the relative pitch height clearly plays an important role, and the turning point is not as critical as in Mandarin tones (Fok, 1974; Abramson, 1975; Vance, 1977; Gandour, 1981, 1983).

Another focus of tone perception research is on the phenomenon of categorical perception, in which equivalent acoustic differences between two stimuli are treated differently, depending on whether they are heard as members of the same category or from different categories. It aims to explain how listeners cope with the many-to-one mapping between acoustic patterns and phonological categories (Francis, Ciocca and Ng, 2003). Perception of consonants has been found to be categorical (Liberman et al, 1957; Liberman et al, 1961), while that for vowels are not (Fry et al, 1962; Stevens et al, 1969; Abramson, 1976). In recent years, there has been similar interest on studying categorical tone contrasts. A consistent conclusion is that the perception of the contrasts between level tones seems not to be categorical (Abramson, 1979; Francis, Ciocca and Ng, 2003), and that between contour tones tends to be categorical (Wang, 1976; Francis, Ciocca and Ng, 2003).

F0 variations in the speech of tonal languages carry rich information other than tones. Tone perception is also affected by the other functions of F0 variations. In continuous speech, perception of a tone is greatly influenced by its neighboring tones. The relative F0 levels of abutting tones were found to provide most contributive information (Xu, 1994; Ma, Ciocca and Whitehill, 2005; Zheng et al, 2006; Chen and Kuo, 2007). Intonation causes changes in F0 contour and therefore affects tone perception. This is especially evident on the tone at the final position of an intonation phrase. Level tones were more likely to be affected than contour tones, and question-type intonation had greater effect than statement-type (Fok, 1974; Connell, Hogan and Rozsypal, 1983; Ma, Ciocca and Whitehill, 2005). Moreover, the identification of tones depends on the knowledge about the talker's pitch range (Moore and Jongman, 1997; Wong and Diehl, 2003). Furthermore, the linguistic experience of a listener was also found to play a role in tone perception (Gandour, 1983; Leather, 1987; Stagray and Downs, 1993; Lee, Vakoch and Wurm, 1996; Burnham and Francis, 1997).

1.4. Approximation and Stylization of F0 Contours

In the preceding sections, F0 perception in respect to the perception-acoustic relation was extensively discussed, while speech production was not the concern. Nevertheless, inclusion of naturally produced acoustic signal in perceptual studies is capable of establishing good connections between perception and production.

The F0 contours measured from human speech (henceforth referred to as *observed contours*) vary to a considerable extent (Pike, 1948). Given the relation between perception and production as discussed in Section 1.1, some of the variations in human-generated acoustic signals are production-intended in order to achieve communicative functions, while some are just unavoidable, given the physical constraints of human speech production. For example, human articulators restrict the maximum rate of F0 change, and consequently, the transition between a pair of pitch targets at contrasting levels, e.g., high-low or low-high, must be gradual (Xu and Sun, 2002; Xu, 2004). On the other hand, how these variations can be accessed and interpreted in speech communication is determined by the perceptual ability on detecting F0 variations and the mapping of perceived variations to particular communicative functions. The perceptual limitations in discriminating different kinds of F0 variations have been introduced in Section 1.2.2. Among these limitations, glissando threshold and differential glissando threshold might be particularly important for perceiving F0 contours in speech. The existence of these thresholds suggests that not all the variations in the observed F0 contours are necessary for perception. It is possible to remove unnecessary variations and simplify the observed contours, with the requirement of maintaining perceptual quality. For this purpose, perception-based approximation or stylization of F0 contours has been an important research problem and is the direction of the present study.

1.4.1. Motivations

The motivations of studying approximation of F0 contours are multi-fold (t' Hart, Collier and Cohen, 1990). By using a simplified representation, the amount of acoustic data becomes more manageable and thus easier to be processed and interpreted. This will benefit corpus-based studies, which have become the one of the major approaches in spoken language research. On the other hand, since the process of contour approximation is based on perceptual tolerance, it will help improving our understanding about how F0 contours are processed in human perception and to identify the perception-critical F0 variations. Moreover, it is helpful to establish a relation between pitch variations and discrete linguistic events, which is the primary goal of research on speech prosody (Kochansk and Shih, 2003; Fujisaki et al, 2005; Prom-on, Xu and Thipakorn, 2009). Simple representation of F0 contour is also beneficial to prosody modeling in text-to-speech systems, although issues on system development are not addressed in this thesis.

1.4.2. Approximation of F0 contours in non-tonal languages

Stylization of intonation contours in European spoken languages was investigated in relatively good depth. In the IPO method (t' Hart, Collier and Cohen, 1990), pitch contours of Dutch utterances were investigated and each syllable-sized contour was simplified as a straight-line segment. The simplification process was performed by human subjects and equal perception to original speech was assured. The stylization was later applied on building an intonation generation model for Dutch. In another method by (d'Alessandro and Mertens, 1995), syllable-sized pitch contours in French utterances were simplified based on perceptual thresholds on detecting pitch movements. The simplified contours were evaluated with re-synthesized speech. It was reported that the perceptual quality of re-synthesized speech could approach natural speech.

1.4.3. Approximation of F0 contours in tonal languages

Compared with non-tonal languages, F0 contours in tonal languages are associated with additional communicative functions. Perception of them becomes more complicated due to the involvement of the important task — identification of tones.

However, the actual realizations of a lexical tone in natural speech often vary greatly and the observed F0 contours usually deviate a lot from the phonologically defined pitch pattern (Xu, 1997, 2004; Fujisaki et al, 2003; Li, Lee and Qian 2004). In tonal languages, many studies have been carried out on acoustic features that are related to tone identification (Fok, 1974; Vance, 1977; Gandour, 1981, 1983, 1984; Abramson, 1997; Francis, Ciocca and Ng, 2003; Ma, Ciocca and Whitehill, 2005; Jongman et al, 2006; Zheng et al, 2006; Khouw and Ciocca, 2007). Contrastively, very few efforts have been put on approximation of F0 contours from the perspective of speech perception. It remains unclear how the perceptually related F0 variations are described such that an equal perception can be maintained. Our understanding on the F0 variations that are critical to perceived speech naturalness is relatively limited.

1.5. Approximation of F0 Contours in Cantonese Speech

This study concerns the approximation of F0 contours in Cantonese speech. Cantonese is a tonal language, in which both level and contour tone contrasts are found.

1.5.1. Methodology

We attempt to find the simplest acoustic representation of F0 contours that can attain perceptual equivalence with the natural speech. Different approximations of the observed F0 contours are proposed and investigated via perceptual comparison. The natural speech is modified to follow the approximated contours, and perceptual tests are carried out to compare the modified speech with the natural one. When the human subjects cannot perceive any difference between them, the approximations are believed to carry information without loss, i.e., all the important pitch variations for perception of natural speech as well as tone identification are captured.

1.5.2. Linear approximation

For the selection of appropriate approximations for an observed F0 contour, our underlying hypothesis is that perception is only sensitive to the major trends of F0 variations. We argue that these trends can be captured by using a linear approximation of the contour. This is motivated from different perspectives. Phonologically, Cantonese tones[8] are defined by simple movements, e.g., level, falling, rising, among a limited number of pitch heights (Chao, 1947; Hashimoto, 1972; Bauer and Benedict, 1997; LSHK, 1997). These patterns can be considered as the underlying targets for speech production to reach, and for speech perception to identify. Previous studies found that most of the useful cues for Cantonese tone identification are related with linear F0 variations (Fok, 1974; Vance, 1977; Gandour, 1981, 1983; Khouw and Ciocca, 2007). For non-tonal languages, linear approximation has been proved to be appropriate for representing intonation (t' Hart, Collier and Cohen, 1990; d'Alessandro and Mertens, 1995). Our work examines whether linear approximation of F0 contour is adequate to represent the perceptually sensitive F0 variations in the speech of tonal languages, so as to improve our understanding on the perceptual processing of F0 variations in speech.

[8] More details of Cantonese tones are given in Chapter 2.

In our study, F0 contours are investigated on a syllable-size basis. Each observed F0 contour of the Cantonese tones is approximated as a concatenation of straight-line segments that describe the major trends of F0 variations over a syllable. The transitions between co-articulated tones are also approximated by line segments. The approximations are examined in a series of perceptual tests.

1.5.3. Manual and automatic approximation

The approximation of an observed F0 contour can be performed either manually or automatically. At the initial stage, when our knowledge about approximation is very limited, e.g., in our case, only "linear" is expected, manual modification is considered to be a good starting strategy to examine the hypothesis and to verify the suitable approximations. The manual approximation is performed in an interactive procedure in which an observed contour is approximated by a trial and reducing-error approach, such that the perceptual difference between the natural speech and the modified speech is minimized, at least for the human performer. In this way, the "best" approximation for each observed contour can be obtained easily.

Manual approximation is not reproduceable and hence not appropriate for large-scale scientific research. Its generalization ability to new data is weak. For practical applications, method of automatic approximation is needed to make the procedures standardized and efficient. Automatic approximation initially tries to interpret manual process systematically. Nevertheless, we gradually find this is not an easy task. Though the course of manual approximation seems easy and quick, its nature is rather complex. Human listeners are greatly directed by the perceived difference and meanwhile refer to multiple dimensions of knowledge with or without consciousness, e.g., visual observations and linguistic knowledge. With limited understanding on human perception of F0 contours, interpreting the course is difficult. Therefore, the task of automatic procedures is changed from imitating manual course to implementing perceptual knowledge on acoustic data. Manual approximations sketch out a basic structure of the linear approximation for the F0 contours of each tone category, e.g. the number of required linear movements. The structure is closely correlated with the phonological definition of the tone. Furthermore, perceptual knowledge is applied on the observed F0 contours to help specify each linear movement in terms of its location and slope. The integration of linguistic knowledge and perceptual model (d'Alessandro, Rosset and Rossi, 1998) differentiates the present study from previously proposed methods (d'Alessandro and Mertens, 1995).

1.5.4. Indications

Linear approximation provides us a basis to understand and interpret human perception of F0 contours in tonal speech, and the interpretation is not limited to tone perception any more. In addition, it facilitates the association between the limits of pitch perception and the perception of F0 in speech.

1.6. Thesis Outline

This thesis investigates the issue of linear approximation for F0 contours in Cantonese speech, mainly from the speech perception perspective, and also associating with speech production. It is organized as follows.

Chapter 2 gives an introduction to the language of Cantonese, focusing on its tones. This chapter aims to provide the essential background knowledge for the whole study.

Chapter 3 presents a basic study for linear approximations. It investigates how the F0 contours at different levels, i.e., syllable, word and sentence can be represented by linear movements, such that perception of the speech utterances is not affected.

In Chapter 4, based on the findings in the basic study, the generalization ability of linear approximation is further examined on a large set of polysyllabic words. Given amount of perceptually confirmed linear approximations, statistical analysis is carried out. The analysis findings are interesting and can improve our understanding on how the F0 variations function in speech communication.

Chapter 5 describes a framework of automatic approximation. This framework is perception-oriented. It incorporates not only our perceptual findings but also other perceptual knowledge. Both objective and subjective evaluations suggest that the proposed framework can generate good approximations for the observed F0 contours.

Lastly, conclusions and suggestions for future work are given in Chapter 6 and Chapter 7 respectively.

Chapter 2

Cantonese Tones

Delivering tone information is the primary task of F0 variations in the speech of tonal languages. Cantonese is rich in tones and its tone system is particularly interesting to both linguists and speech perception researchers. This chapter aims to provide essential background knowledge of Cantonese tones. The chapter starts with a brief introduction to the language of Cantonese and then describes the tone system from three aspects: phonological definitions, acoustical realizations and perception.

2.1. The Cantonese Dialect

Cantonese is a major Chinese dialect widely spoken in Southern China (Bauer and Benedict, 1997). It is also popular among overseas Chinese communities in Southeast Asia, North America, Australia and Europe. The dialect expanded farther and faster than any other Chinese dialects especially in the 20th century, mainly due to the cultural and media liberty available in Hong Kong. Today, its speaker population is up to 55.5 million in around 20 countries, ranking the third among all Chinese dialects and the twenty-first among the languages in the world (Lewis, 2009).

Cantonese plays a significant role in both sociological and cultural life. It is the predominant language in Hong Kong and Macau. It is also the only variety of Chinese other than Mandarin that is being used in official contexts. Cantonese preserves many ancient pronunciations. This makes it invaluable for the investigations on Chinese culture.

Cantonese, like Mandarin, is a monosyllabic and tonal language. Written Cantonese is composed of a sequence of Chinese characters. A Chinese character is the smallest meaningful unit in the language and is pronounced as a syllable carrying a specific lexical tone. As illustrated in Figure 2.1, each syllable consists of an *Initial* and a *Final*. The *Initial* is generally a consonant or being absent. The *Final* typically consists of a vowel nucleus and an optional consonant coda. A Cantonese utterance is seen as a sequence of continuously pronounced syllables.

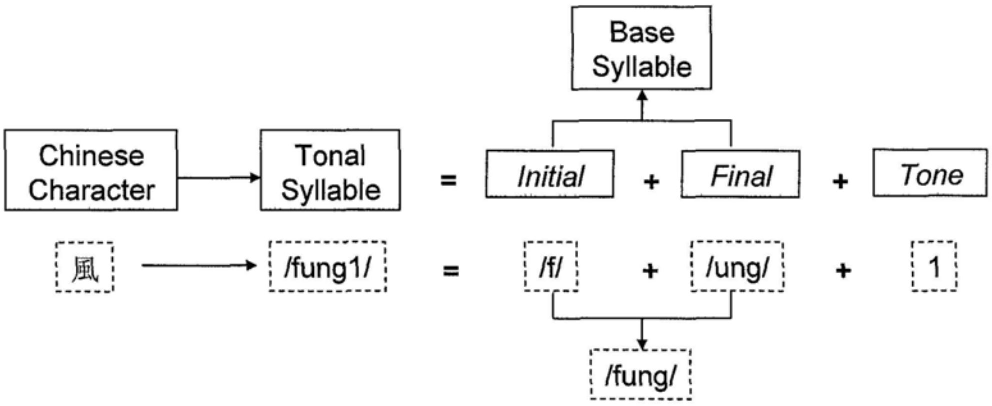


FIG. 2.1: Phonological structure of a Cantonese syllable.

In this study, Cantonese spoken in Hong Kong is investigated. The Jyut Ping system (LSHK, 1997) is adopted for Cantonese romanization.

2.2. Tones of Cantonese

2.2.1. Tone system

Cantonese has a relatively complicated tone system. In conventional Chinese phonology, Cantonese is said to have nine tones described by different pitch patterns as shown in Fig. 2.2. Such an abundant tone system is rather rare among tonal languages in the world (Maddieson, 1978; Connell, 2000).

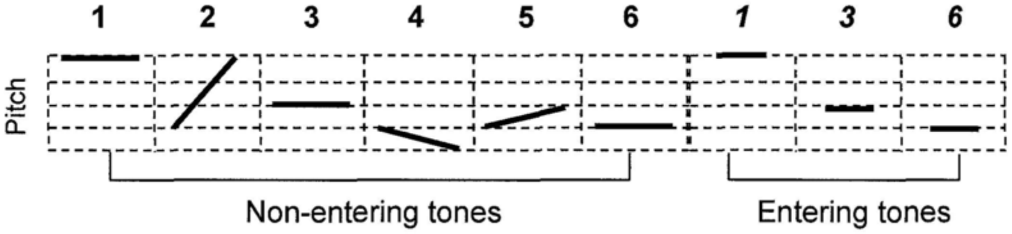


FIG. 2.2: Pitch patterns of Cantonese tones. According to the Jyut Ping transcription system (LSHK, 1997), the six tones are labeled by the numerals 1 to 6. Entering tones occur with checked syllables and are just shorter in duration with their non-entering counterparts.

"Entering tone" is a historically defined tonal category that was used as a cover term for tones that co-occur with "checked" syllables, i.e., syllables ending with an occlusive coda such as /p/, /t/ or /k/. Entering tones are contrastively shorter in duration but coincide with a non-entering counterpart in terms of pitch level (Lee and Qian, 2007). Many linguistic researchers have suggested treating the three entering tones as abbreviated versions of their non-entering counterparts (Chao, 1947; Hashimoto, 1972; Vance, 1976; Bauer and Benedict, 1997). In the Jyut Ping system (LSHK, 1997), only six distinctive tone categories, which are labeled by the numerals 1 to 6 (Figure 2.2), are defined. In our study, we follow the Jyut Ping system and use a six-tone system for labeling.

In terms of pitch height, the six tones are distributed in two ranges: high (Tone 1, 2, 3) and low (Tone 4, 5, 6). In terms of the shape of pitch contour, they can be classified as level tones (Tone 1, 3, 6), falling tone (Tone 4), and rising tones (Tone 2, 5). Unlike the contour tones in Mandarin which contrast each other by different shapes, Cantonese tone system is close to a register system, discrimination among tones much relying on pitch heights in relation to each other (Clark and Yallop, 1990). In Cantonese, the number of level tones is up to three. This is considered to approach the limit of level tones that can be used in a tone system (Maddieson, 1978; Connell, 2000).

The pitch patterns as shown in Figure 2.2 are the phonological descriptions of the tones. They can be considered as the underlying targets for speech production to reach, and for speech perception to identify.

2.2.2. Acoustical realization of produced tones

Acoustically, tone is represented by the F0 variation across the voiced portion of a syllable. In a Cantonese syllable, the *Final* can be regarded as voiced while the *Initial* is either voiced or unvoiced. In the case of isolated syllables, the acoustical realization of a lexical tone tends to resemble its phonological pattern. The produced F0 contours of an entering tone and its non-entering counterpart are often very similar (Lau, 2000).

The phonological patterns as well as the isolated tone contours are considered to be the idealistic representations of tones. In real speech, the actual realizations vary greatly with many factors, such as speakers, linguistic context and speaking rate (Pike, 1948). Being an example, the observed F0 contour of a Cantonese sentence utterance is given in Figure 2.3. The tone identity of each syllable is marked by the numerals and the phonological pitch pattern of each tone is shown above its observed contour. Obviously, tone contours in continuous speech deviate a lot from their canonical patterns. The same tone can be realized quite differently, such as the five occurrences of Tone 3 in the example utterance.

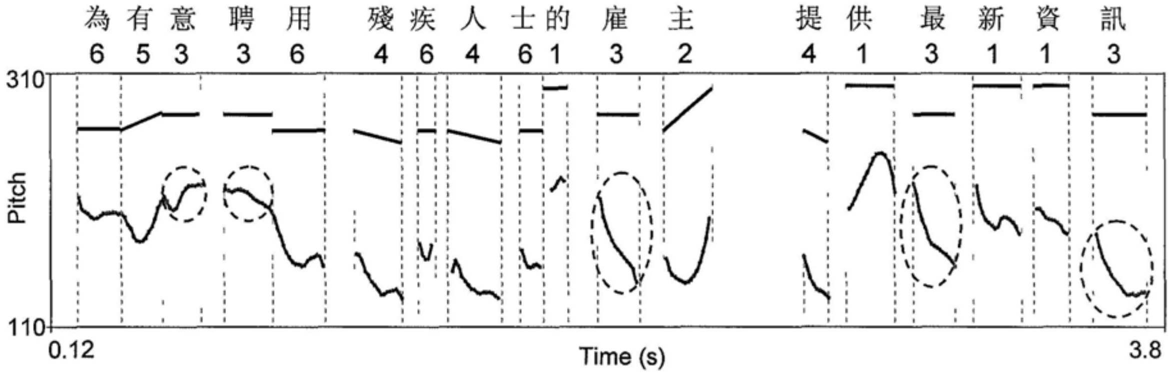


FIG. 2.3: Observed F0 contour of a Cantonese sentence utterance. Vertical dashed lines indicate syllable boundaries. The produced tone contours vary greatly from phonological pitch patterns. The same tone identity is realized very differently in different syllables.

Acoustical analysis with large-scale speech corpuses revealed that the F0 ranges of different Cantonese tones overlap each other substantially in both isolated and continuous speech (Li, 2003; Qian, Lee and Soong, 2007). Considering the nature of Cantonese tones, relation between neighboring tones in continuous speech is established (Li, 2003). Such relation has been explored effectively in tone normalization (Li, Lee and Qian, 2004), speaker modeling (Li, 2006) and tone recognition (Qian, Lee and Soong, 2007).

Co-articulation between successively produced tones is unavoidable due to continuous muscle movements in human speech production system (Ohala and Ewan, 1973; Sundberg, 1979; Stevens, 1998; Xu and Sun, 2002; Kochanski and Shih, 2003). In Cantonese, co-articulation leads to a carry-over effect and an anticipatory effect on the initial and the later portions of a tone respectively (Li, Lee and Qian, 2002), as that is observed in many other languages (Han and Kim, 1974; Gandour, Potisuk and Dechongkit, 1994; Xu, 1997). Explicit manifestation can be found at syllable boundaries. In Figure 2.3, it is clearly shown that tone contours compromise with each other to make a smooth transition, especially when the abutting tones have contrastive pitch levels.

Long-term downtrend of F0 contour within a single phrase or across phrases is observed in Cantonese (Li, Lee and Qian, 2004). It could be related with both the physiological constraints and the realization of intonation (Kochanski and Shih, 2003). It is commonly agreed that the surface F0 contour of an utterance is basically the superposed product of local tone contours and one or several long-term intonation contours (Fujiaski and Hirose, 1984; Dutoit, 1997; Holm and Bailly, 2000; Dong and Lua, 2002; Kochanski and shih, 2003; Li, Lee and Qian, 2004; Ni and Kawai, 2006).

2.2.3. Perception of Cantonese tones

Tone is always processed in together with segmental information (Culter and Chen, 1997). Hence, speech carriers are generally used when investigating the properties of tones. In speech perception, a syllabic decomposition occurs at the early stage (Ryalls, 1996). In tonal languages, linguistically, only tone-aligned syllables can define the smallest meaningful units completely; acoustically, tone is found to be fully in phase with syllable due to the physical constraints of human production (Xu and Sun, 2002; Xu, 2004; Olsberg, Xu and Green, 2007; Xu, 2008). Concerning F0 variations in relation with perceived speech naturalness, we believe that tone contour is also the essential unit. Therefore, our investigation will start with the F0 variations carried by tone contours.

Perception of isolated Cantonese tones is generally investigated with the task of tone identification, for the purpose of finding acoustic features that determine or influence human identification and discrimination of tones. Because of the overlapped pitch range among different tones and the existence of multiple level tones, identification of isolated Cantonese tones is not an easy task as that for Mandarin. The correctness percentage varies in the range of 60% ~ 90%. Contour tones are identified more accurately than level tones (Khouw and Ciocca, 2007; Yuen et al, 2007). Many studies suggested that the relative F0 level, the direction of F0 change, and the magnitude of F0 change are the most crucial perceptual correlates for tone identification (Fok, 1974; Vance, 1977; Gandour, 1981, 1983). In addition, it was found that F0 change over the later portion of a tone is much important for separating tones in both production and perception, as they often preserve the most representative canonical tonal patterns (Khouw and Ciocca, 2007). Temporal location of turning point was found to be acoustic cues for identification of Mandarin tones, while this might not be true for Cantonese tones. The temporal positions of turning points in the F0 contours of the two rising tones are often similar (Khouw and Ciocca, 2007).

In continuous speech, tone identification depends much on the context. This is particularly true for level tones. (Ma, Ciocca and Whitehill, 2005) found that listeners place more emphasis on the immediate context preceding the target and the context information contributes mainly on identification of F0 level, while the intrinsic acoustic properties of the tone help in identifying the F0 contour. Similar finding was also reported by (Zheng et al, 2006) that categorizing tones not only depends on the absolute F0 value of the target syllable but also on the F0 difference between the target syllable and the adjacent syllables.

For both isolated and continuous speech, the identification of tones is processed with the knowledge of the talker's pitch range (Moore and Jongman, 1997; Wong and Diehl, 2003). Intonation perception is said to be formed by a process of memorizing the stored tonal movements during the ongoing speech and then retrieving them after being integrated into a meaningful linguistic whole (House, 1990).

While acoustic features related to tone identification have been investigated extensively, how the tone contours give influence to perceived speech naturalness has not received much attention. In our previous study on Cantonese text-to-speech synthesis (Li, Lee and Qian, 2004), it was suggested that the naturalness of synthesized speech can be improved, given properly modeled tone and tone transition contours. We believe that tone contours are not only important for tone identification, but contribute much to speech naturalness. Meanwhile, it should be noted that due to the lack of clear definitions on the underlying targets related to speech naturalness, perceived speech naturalness could be highly connected with a training effect from human production. In studies of Cantonese tone contours, in addition to the acoustic features that are related to tone identification, the essential ones for influencing speech naturalness should be concerned as well. Nevertheless, this problem was not addressed in previous perceptual studies. This study tries to give answers to the question. Several perceptual investigations on the simplified F0 contours in Cantonese speech are carried out and they will be introduced one by one in the next several chapters.

Chapter 3

Perception of Linearly Approximated F0 Contours – Basic Study

In this chapter, we attempt to find the simplest acoustic representations for the observed F0 contours. Perceptual equivalence between the modified speech and the natural speech is required during the process. Different approximations of the observed F0 contours in Cantonese speech are suggested and investigated through perceptual comparison. Natural speech utterances are modified to follow the approximated contours. Perceptual tests are carried out to compare the modified speech against natural speech. When human subjects cannot perceive any difference between them, the approximations are believed to carry all the important pitch variations for speech perception. It is found that perception-sensitive F0 variations in Cantonese speech can be adequately described by the linear approximation of the observed F0 contours. Each tone contour can be represented by one or two linear movements, and the transition between a pair of co-articulated tones can be represented by one linear movement. Moreover, slopes of these linear movements are found to be closely related to perception. A lifted perceptual tolerance on F0 variations in speech is noticed.

In the following sections, three perceptual experiments on the approximated F0 contours will be described. They are carried out with speech materials at syllable, word and sentence levels respectively.

3.1. Experiment 1: Perception of Approximated Tone Contours of Isolated Syllables

Syllables spoken in isolation are the basic independent and intact units, on which the required pitch targets can be approached without contextual variation. Therefore we consider the F0 contours of isolated syllables as carrying the canonical tone patterns. This experiment investigates linear approximations of these canonical patterns and compares the observed F0 contours with their linear approximations perceptually.

3.1.1. Approximation

Based on the phonological descriptions and the acoustic observations, tone contours of isolated Cantonese syllables were approximated by one or two straight-line segments. According to the phonological descriptions, three of the six tones are contour tones (Tone 2, Tone 4 and Tone 5) and the other three are level tones (Tone 1, Tone 3 and Tone 6). However, acoustic observations suggested that level tones and Tone 4 commonly hold a major trend of falling, but to different degrees. For simplicity, we divided the six tones into two groups, namely rising and non-rising. The rising-tone group consists of Tone 2 and Tone 5, and the non-rising tone group consists of the other four tones. Figure 3.1 shows the approximation examples for the two groups. For the non-rising tones, two approximation patterns were examined: level (denoted by **La**) and falling (denoted by **Fa**). **La** is expected to be adequate for representing level tones, while **Fa** would describe the observed contours more precisely. For the rising tones, three different approximation patterns, namely rising (**Ra**), level-rising (**LRa**) and falling-rising (**FRa**), were investigated. They resemble the observed contours to different extents. As a result, there are a total of 14 different approximation patterns for the six tones, i.e., 2 patterns ×4 non-rising tones + 3 patterns ×2 rising tones.

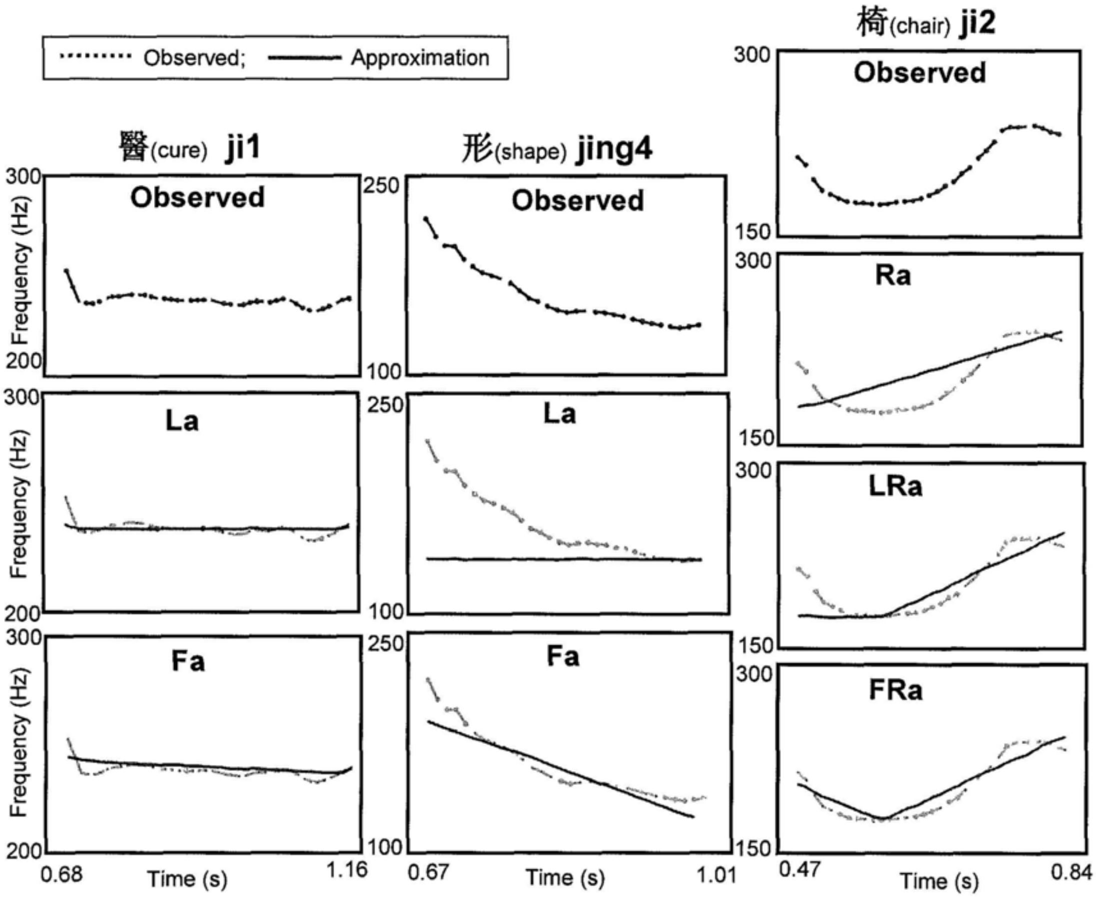


FIG. 3.1: Examples of linear approximations for tone contours carried by isolated syllables. (a) and (b) illustrate level (**La**) and falling (**Fa**) approximation patterns for non-rising tones; (c) depicts rising (**Ra**), level-rising (**LRa**) and falling-rising (**FRa**) approximation patterns for rising tones.

3.1.2. Method

3.1.2.1 Subjects

Ten subjects participated in the listening test. They include five female and five male students at the Chinese University of Hong Kong. Their ages are from 20 to 30 years. All of them are native Cantonese speakers with normal hearing and none of them has professional knowledge about speech technology or linguistics. The subjects were paid for their workload.

3.1.2.2 Stimuli

The speech carriers are three Cantonese syllables /wai/, /ji/ and /jing/. Experiments in Lee (2001) showed that the initial fragment of a tone contour corresponding to the initial consonant is important for recognizing Mandarin tones. (Xu, 2004, 2008) also found that Mandarin tone and syllable are totally in phase, from the perspective of speech production. Therefore, we choose syllables with voiced initial consonants such that the F0 contour covers the duration of the entire syllable. Each syllable can be associated with six different tones. As a result, there are 18 tonal syllables that are represented by different Chinese characters as shown in Table 3.1.

Table 3.1. List of carrier syllables used in Experiment 1. Three base syllables associating with 6 tones give 18 carriers. All the characters are transcribed according to LSHK (1997).

	Base syllable		
Tone	**wai**	**ji**	**jing**
1	威 (stateliness)	醫 (cure)	英(Britain)
	wai1	ji1	jing1
2	委 (committee)	椅 (chair)	影 (shadow)
	wai2	ji2	jing2
3	畏 (fear)	意 (meaning)	應 (respond)
	wai3	ji3	jing3
4	維 (dimension)	怡 (joyful)	形(shape)
	wai4	ji4	jing4
5	偉 (great)	耳 (ear)	郢 (Ying capital)[N]
	wai5	ji5	jing5
6	惠 (benefit)	二 (two)	認 (recognize)
	wai6	ji6	jing6

[N]: the capital name of Chu State in ancient China.

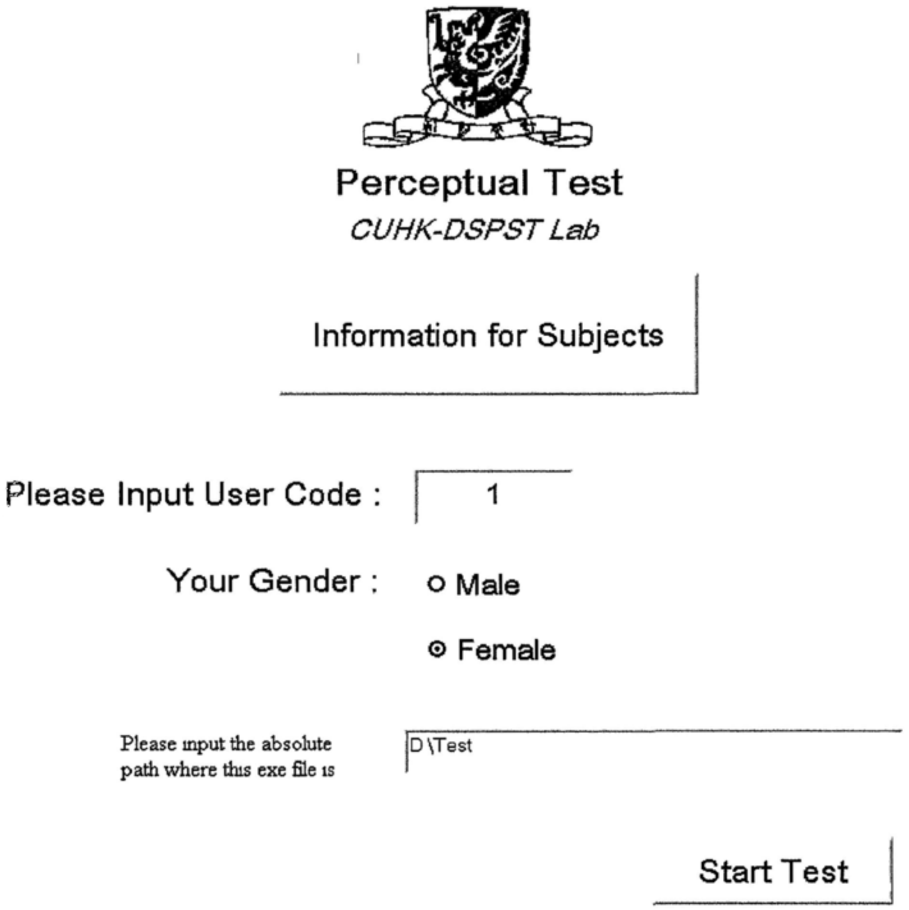
Natural utterances of these tonal syllables were recorded from a female speaker (Speaker I). The recording was done using a high-quality close-talking microphone (Shure SM10A) in a quiet room. The recorded signals were digitized at a sampling frequency of 44,100 Hz. Each syllable was recorded three times, and the one with the best sound quality was used in the experiment.

The F0 contour of each natural utterance was computed. The approximations for each observed contour were manually created by the author of this thesis through an iterative procedure using the Praat software (Boersma and Weenink, 2010). The initial trial was made to follow the contour trajectory. Then the re-synthesized speech with the initial approximation was assessed and compared perceptually with the natural utterance, and adjustments on the existing approximation were made accordingly. This iterative process of re-synthesis and adjustment continued till the perceptual difference could not be further reduced. The re-synthesized speech in which F0 variations follow the final resulted approximation is referred to as the *modified speech*.

For each syllable carrier, there were 14 approximations for the 6 observed tone contours. Accordingly, there were totally 42 approximated contours to be evaluated for the three syllable carriers.

3.1.2.3 Procedure

Pair comparison was adopted as the testing method. Each pair of stimuli contained a modified speech utterance and a natural one. The stimuli were presented to the subjects through an interactive computer interface. The working procedure of the interface is explained through Figure 3.2 to Figure 3.4. Figure 3.2 shows the first page of the interface, which serves to collect the information of the subject, such as the subject code and the gender. By clicking the button "Information for Subjects", the window as shown in Figure 3.3 pops up to display the test instructions. All subjects are required to read this instruction page before starting the test; otherwise a warning message would pop up when the "Start Test" button is clicked. Afterwards, the interface will go forward to the testing page as shown in Figure 3.4. It deals with one test trial.

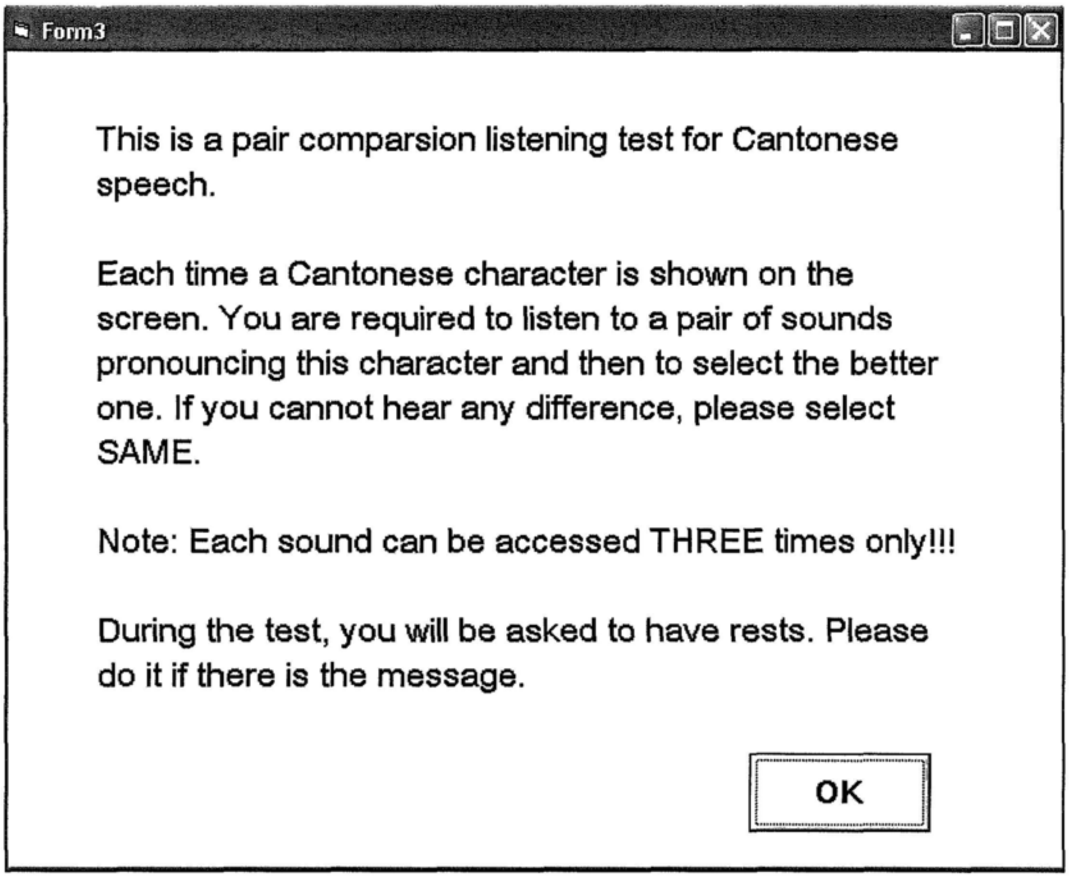


Please Input User Code : 1

Your Gender : ○ Male ◉ Female

Please input the absolute path where this exe file is: D\Test

FIG. 3.2: The first page of the interactive test interface — collecting subject information.



This is a pair comparsion listening test for Cantonese speech.

Each time a Cantonese character is shown on the screen. You are required to listen to a pair of sounds pronouncing this character and then to select the better one. If you cannot hear any difference, please select SAME.

Note: Each sound can be accessed THREE times only!!!

During the test, you will be asked to have rests. Please do it if there is the message.

FIG. 3.3: The information page of the interactive test interface — showing test instructions.

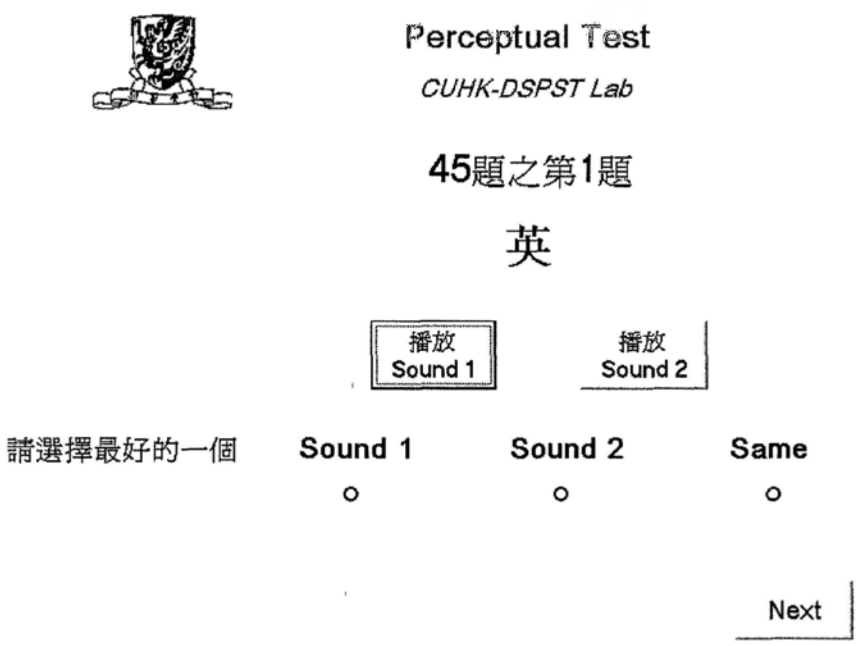


FIG. 3.4: The testing page of the interactive test interface — dealing with a test trial.

In each test trial, a Chinese character that corresponds to the spoken syllable is shown on the computer screen, and the subject can listen to the two stimuli by mouse-clicking the corresponding sound buttons, which are labeled as *sound 1* and *sound 2*. Each stimulus is allowed to be accessed at most three times. The subject is asked to select the preferred stimulus in terms of overall impression, or to rate them as the same by clicking one of the three response buttons on the screen, which are labeled as *sound 1, sound 2, same*, respectively. After a decision has been made, the response is recorded and the subject can proceed to the next pair of stimuli by clicking the button "Next". When the button "Next" is clicked, the interface automatically checks whether the two stimuli have been listened and whether a response has been recorded. If not, a reminder message will pop up for the missed action.

The subject's responses to the first three pairs of stimuli in each test are not counted. These pairs are intended to help subjects be familiar with the process. The subjects are not informed of this arrangement. The formal test goes through a full list of the 42 test pairs, which are presented in a randomized order. Given a test pair, two stimuli are also randomly associated with the two displayed sound buttons. The test lasts for about 15 minutes.

3.1.3. Results and analysis

The test results are shown as in Figure 3.5. Each vertical bar in the figure represents the 30 collected responses for one specific approximation pattern. The distributions of preference votes to "natural speech", "same" and "modified speech" are illustrated with different colors and textures. The test result on each approximation pattern is interpreted based on its preference distribution. If more than 50% of the votes are for "modified speech", we consider that the modified speech could attain *better perception* than the natural speech, and the respective approximation pattern is rated as *excellent*; if not less than 90% of the votes are for "same", the modified speech is considered to attain *equal perception* to the natural speech, and the approximation pattern is rated as *good*; if the votes for "natural speech" have a percentage of 10% ~ 50%, the modified speech is considered to attain *comparable perception* with the natural speech, and the respective approximation is said to be *appropriate*; if more than 50% of the votes are for "natural speech", the modified speech is considered to have *worse perception* than the natural speech, and the approximation is rated as *bad*. Furthermore, a statistical analysis of the test results was carried out by paired t-test[9] between different approximation patterns of a particular lexical tone category. The test results are given in Table 3.2.

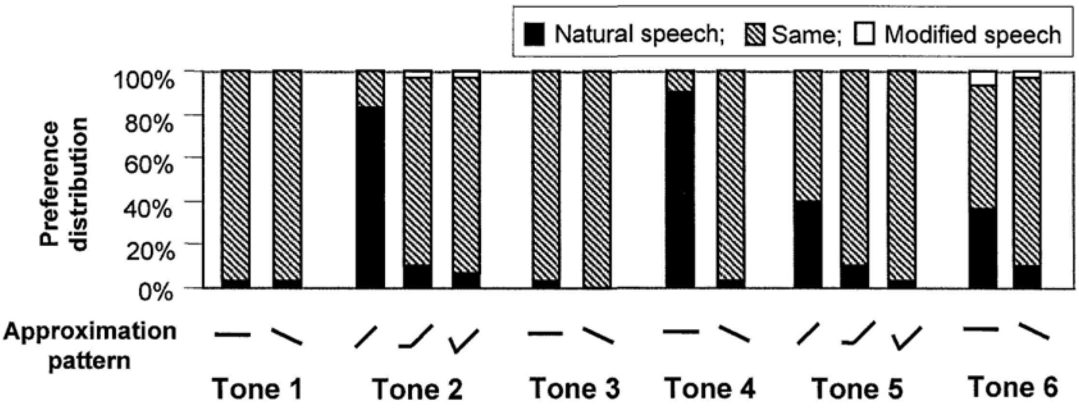


FIG. 3.5: Perceptual test results on the approximated tone contours carried by isolated syllables. Each vertical bar represents the distribution of 30 collected responses for a particular approximation pattern. The preference votes to "natural speech", "same" and "modified speech" are illustrated with different colors and textures.

[9] A brief introduction to the basic concepts of paired t-test and ANOVA is given in Appendix 2.

Table 3.2. Results of paired t-test between different approximation patterns of a particular lexical tone category.

Paired t-test (significant level: $p < 0.05$)									
Tone		**1**	**2**		**3**	**4**	**5**		**6**
	Approx.	**La**	**Ra**	**LRa**	**La**	**La**	**Ra**	**LRa**	**La**
1	**Fa**	same							
2	**LRa**		< 0.001						
	FRa		< 0.001	0.59					
3	**Fa**				0.34				
4	**Fa**					< 0.001			
5	**LRa**						0.03		
	FRa						0.007	0.17	
6	**Fa**								0.037

For Tone 1 and Tone 3, the modified speech with approximation patterns **La** and **Fa** are both perceived to be equal to the natural speech ($p \geq 0.34$). Whilst for Tone 4 and Tone 6, the results on **La** and **Fa** are significantly different ($p < 0.04$). **Fa** is found to be more appropriate than **La**. Evidently **La** is not applicable to Tone 4 ($p < 0.001$), but it is appropriate for Tone 6, though less preferred than **Fa** ($p = 0.037$). This implies that Tone 4 is a falling tone, and Tone 6 is more like a level tone. For the rising-tone group, the modified speech with approximation pattern **FRa** attains equal perception to the natural speech. **LRa** is slightly worse than **FRa** ($p \geq 0.17$). Ra, which contains only a single movement of pitch rising, is rated as a bad approximation for

Tone 2 ($p < 0.001$). For Tone 5, Ra is found to be appropriate, but it is much worse than **LRa** and **FRa** ($p \leq 0.03$). In summary, to reach comparable perception, three kinds of linear approximation patterns are needed. For high-level tones (Tone 1 and Tone 3), **La** is adequate; for low-level tones (Tone 4 and Tone 6), **Fa** is more appropriate; for rising tones (Tone 2 and Tone 5), **FRa** should be used. More generally speaking, a non-rising tone can be approximated by a single linear movement, while a rising tone needs two linear movements.

As mentioned in Section 3.1.2.3, a subject was allowed to access each stimulus 1 to 3 times. The actual number of access times reflects the degree of difficulty for the subjects to make decisions, which depends on the perceptual difference between of the two stimuli in a pair. The greater the difference, the easier the decision, and the smaller the number of access times. Statistical analysis was carried out on the number of access times. The results are shown in Table 3.3. The overall mean is 1.68. For the best perceived approximations, i.e., **Fa** for non-rising tones and **FRa** for rising tones, the average number of access times is 1.72, indicating reduced difference between the modified speech and the natural speech and the hesitation of the subjects in making judgments. Particularly, in such cases, if the preference response was on "natural speech" or "modified speech", the average number of access times is 2 and 2.13 respectively. This increase of difficulty in making decisions reflects that the modified speech is indeed very close to the natural one.

Table 3.3. Statistics of the number of access times in Experiment 1. The number of access times is defined as, to make a decision, how many times a pair of speech stimuli was listened. The upper limit is 3.

Overall mean: 1.68		
Average for the best approximations: 1.72		
For the best approximations, preference-dependent average		
to "same"	to "natural speech"	to "modified speech"
1.7	2	2.13

With the example approximations in Figure 3.1, the relation between acoustic variations and the corresponding test results can be better interpreted. The observed contours of the syllables carrying high-level tones (Tone 1 and Tone 3) typically show a trend of slight declining. Such declining seems not to be perceptually critical since the test results show that these contours can be adequately represented by **La**. For the low-level tones (Tone 4 and Tone 6), the observed contours exhibit a significant trend of falling. The modified speech with **La** was found to sound very differently from the natural speech. In the process of manual approximation, there was an interesting finding for Tone 4. Even if **La** was determined according to the lowest value in the observed contour, the modified speech was still perceived to have a higher pitch than the natural one.

The observed contours of rising tones (Tone 2 and Tone 5) clearly show a valley shape. **LRa** and **FRa** approach the valley contours to different extents and therefore the modified speech attained equal or comparable perception to the natural speech. Although the rising portion of the contours is important, it seems that the falling part at the beginning also plays a role. From the viewpoint of relative pitch perception (Bachem, 1937; Ritsma, 1965; t' Hart, Collier and Cohen, 1990), it is possible that the falling section in **FRa** can strengthen the perception of the succeeding rising section and make it sound more "rising". **Ra** cannot capture the valley trajectory and deviates greatly from the observed contour. This leads to noticeable difference in perception. Nevertheless, it is noted that **Ra** is a more appropriate approximation for Tone 5 than for Tone 2, particularly in terms of approximating the rising sections. This explains why the test results on **Ra** for Tone 5 are much better than those for Tone 2.

The above results revealed that F0 variations in the observed tone contours can be described by linear F0 movements. Being the defining attribute of a linear variation, movement slope plays a critical role in perception of an F0 contour, as suggested by the psychoacoustic studies (t' Hart, Collier and Cohen, 1990). The movement slopes in the linear approximations were examined in conjunction with the perceptual test results. Figure 3.6 gives the detailed test results for individual carrier syllables. The movement slopes in each tested approximation are given in Table 3.4.

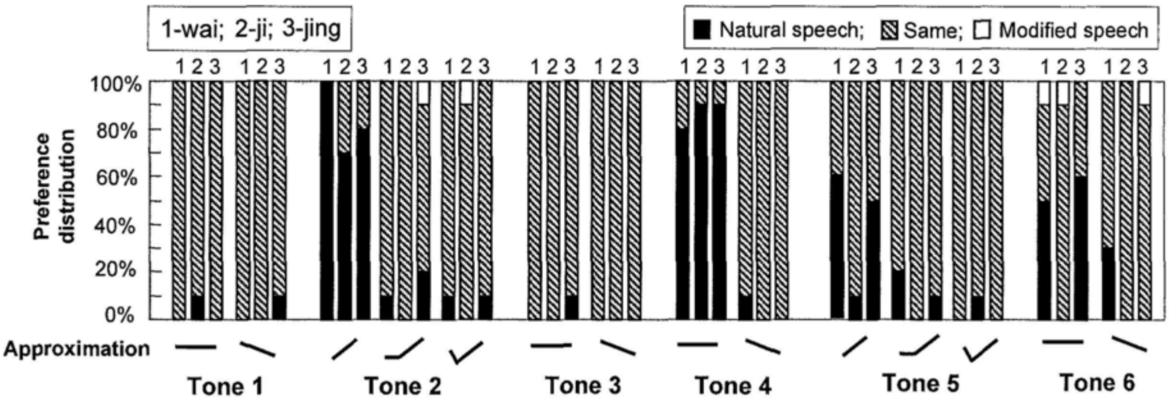


FIG. 3.6: Perceptual test results on the approximations carried by individual carrier syllables. The marker 1/2/3 above each bar refers to the respective base syllable.

Table 3.4. The measured slopes (in Hz/s) of the linear movements in the approximated tone contours carried by isolated syllables.

Tone	Approximation	Syllable		
		wai	ji	jing
1	**La**	-1	0	-4
	Fa	-17	-11	-74
2	**Ra**	300	178	206
	LRa	-1 [f]	-2 [f]	-4 [f]
		580 [s]	297 [s]	338 [s]
	FRa	-125 [f]	-249 [f]	-290 [f]
		597 [s]	283 [s]	365 [s]
3	**La**	0	-4	-4
	Fa	-44	-53	-72
4	**La**	-1	-2	-3
	Fa	-120	-166	-223
5	**Ra**	141	58	125
	LRa	-1 [f]	-2 [f]	4 [f]
		212 [s]	103 [s]	234 [s]
	FRa	-135 [f]	-269 [f]	-177 [f]
		189 [s]	98 [s]	231 [s]
6	**La**	7	-7	-13
	Fa	-129	-63	-53

[f] indicates the slope of the first movement in rising tones;
[s] indicates the slope of the second movement in rising tones.

For each approximation pattern of a specific tone, the three carriers might exhibit some differences. Such perceptual differences were not due to the effect of carriers ($p > 0.07$ by one-way ANOVA). They might be related to the discrepancies in the movement slopes. For example, the test results on **Ra** of Tone 2 – syllable 1 show significant individual difference from the other two syllables. In Table 3.4, this difference corresponds to a very different movement slope (300 Hz/s versus 178 Hz/s and 206 Hz/s). Similar relation is also found for **Ra** of Tone 5, and **Fa** of Tone 6. These observations provide evidences for the correlation between movement slope and contour perception. They also suggest that in Cantonese different contour tones might have different ranges of movement slope.

Perception of these linear approximations enables the comparison between the limits of pitch perception and the perception of F0 variations in speech, by examining glissando and differential glissando thresholds. In this experiment, the estimated value of glissando threshold is 1.17 ST/second for the average tone duration of 0.37 second, or 14 Hz/second for this speaker with a reference F0 level of about 200 Hz[10]. The perceptual test results are further examined based on a comparison of the movement slopes with the estimated glissando threshold and the differential glissando threshold. The following observations were made:

1) If the movement slopes of approximated contours are close to the glissando threshold, e.g., **Fa**(s) of Tone 1, or even two to three times higher than the threshold, e.g. **Fa**(s) of Tone 3, the falling movements could not be perceived.

[10] Calculation from semitone distance to F0 distance is given in Appendix 1.

Their test results are similar to those on **La**. This may indicate that the perceptual tolerance on F0 variation is increased in speech.

2) **Fa**(s) of Tone 6 and Tone 3 have similar slopes. However, the falling movements in Tone 6 could be perceived and those in Tone 3 could not be perceived. A possible explanation is that Tone 6 has a lower F0 level than Tone 3, and hence leads to a lower threshold (in Hz/s) in detecting pitch movements.

3) If the movement slopes are much higher than the glissando threshold, e.g., the approximations for Tone 2 and Tone 5, and **Fa**(s) of Tone 4, the movements could be perceived.

4) In most cases, the slope ratios between the rising movements of Tone 2 and Tone 5 and between the falling movements of Tone 4 and Tone 6 are in the range of 2 to 3. This is in agreement with the differential glissando threshold. For a purpose to be correctly identified, these movements should be at lease perceptually distinguishable from each other. This indicates that the differential glissando threshold, which was established originally from successively presented stimuli, is also applicable in differentiating separately presented F0 movements. Then the perception of the exceptional case, **Fa**(s) of Tone 4 and Tone 6 carried by syllable 1, can be explained. These two approximations were less preferred as compared to the same approximation pattern for the other two syllables, possibly because their slope ratio (≈ 1) does not satisfy the differential glissando threshold and therefore they cannot be distinguished.

The results of this experiment suggest that the observed F0 contours of isolated Cantonese tones can be described by linear movements, which do not affect the perception of tone contours. In this way linear F0 movements are associated with speech naturalness in tonal languages, in addition to their well understood functions in tone identification. Slopes of the linear movements are important for perception. It is found that the perceptual tolerance on F0 variations is lifted in speech signal as compared with those measured in psychoacoustic studies.

3.2. Experiment 2: Perception of Approximated F0 Contours of Words

This experiment investigated approximated F0 contours of Cantonese disyllabic words. As exhibited in Figure 3.7, in natural speech, co-articulation commonly occurs between neighboring tones (Han and Kim, 1974; Gandour, Potisuk and Dechongkit, 1994; Xu 1994; Li, Lee and Qian, 2002). As a consequence, the F0 contour of a naturally spoken polysyllabic word is not a simple concatenation of the tone contours of isolated syllables. A transition period is introduced between neighboring tones. This is particularly evident when the two tones are at very different pitch levels. The transition was found to be contributive to tone identification (Lee, 2001) and might be more influential in the perception of speech naturalness. It was reported that perception is highly sensitive to abrupt F0 changes (t' Hart, Collier and Cohen, 1990). A "smooth" transition is generally a condition of naturalness in speech.

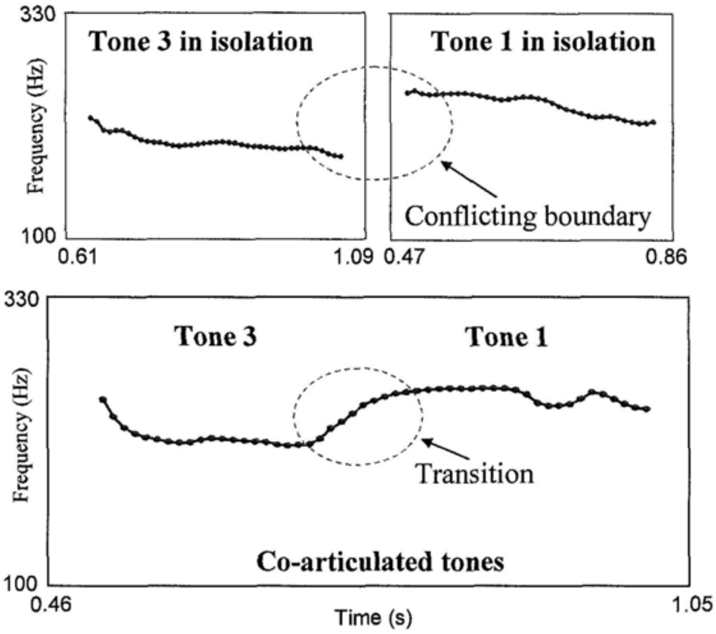


FIG. 3.7: Tone contours in isolation (upper figure) and in continuous speech (lower figure). F0 contour of a naturally spoken word is not a simple concatenation of the tone contours of isolated syllables. A transition period is introduced between neighboring tones.

3.2.1. Approximation

The way of approximating word-level F0 contours is illustrated as in Figure 3.8, using an example of co-articulated level tones and an example of co-articulated rising tones. The observed contours from natural speech show smooth transitions between the connected tones. A transition region was manually identified, with the turning points as the major indicators and the time-aligned speech waveform as reference. The identified transition contour was approximated by one linear movement, and the approximation of tone contours followed our findings in Experiment 1. That is, a non-rising tone was represented by a single movement and a rising tone was described by two movements.

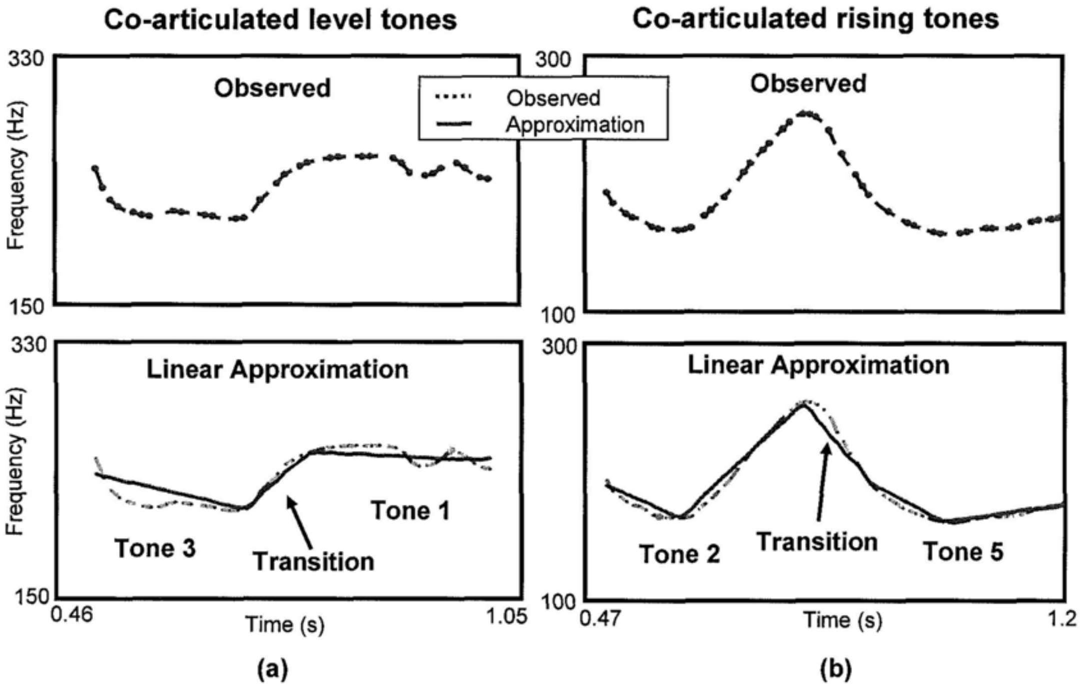


FIG. 3.8: Examples of approximations for the F0 contours of disyllabic words. (a) and (b) illustrate the approximations for co-articulated level tones and co-articulated rising tones respectively. Each non-rising tone is represented by a single movement and each rising tone is described by two movements. Tone transition is approximated by one linear movement.

3.2.2. Method

3.2.2.1 Subjects

Ten male and ten female subjects participated in the listening tests. None of them was involved in Experiment 1. The subject recruitment followed the same requirements as in Experiment 1. The subjects were evenly divided into two groups with balanced gender. Each group of subjects took part in only one of the two sub-tests, which evaluated two different groups of words (see Section 3.2.2.2).

3.2.2.2 Stimuli

The speech materials consist of 44 disyllabic-word carriers as listed in Table 3.5. These words carry only non-entering tones, and the two tones have conflicting pitch levels at the juncture, e.g., Tone 4 followed by Tone 1. The transition between such tone pairs would be acoustically prominent and their influence to perception is most interesting to us. The 44 words were divided into two groups. Group 1 consists of 16 words carrying only non-rising tones. Group 2 has 28 words and each of them contains at least one rising tone. The two groups of words were tested separately in two sub-tests. Natural speech utterances were recorded from the same speaker and under the same recording condition as in Experiment 1. For each recorded utterance, manual approximation was performed and a modified speech utterance was accordingly generated.

Table 3.5. List of disyllabic word carriers used in Experiment 2. Each carrier has a tone combination conflict at the juncture. In Group 1, the carriers contain only non-rising tones; in Group 2, each carrier contains at least one rising tone. All the characters are transcribed according to LSHK (1997).

Group	Tone	Word	Word	Tone	Word	Word
1	1-3	標誌 (sign) biu1-zi3	東亞 (East Asia) dung1-ngaa3	3-1	報章 (newspaper) bou3-zoeng1	對於 (for) deoi3-jyu1
	1-4	醫療 (medicine) ji1-liu4	青年 (youth) cing1-nin4	4-1	澄清 (clarify) cing4-cing1	荃灣(Tsuen Wan)[N] cyun4-waan1
	1-6	幫助 (help) bong1-zo6	操練 (practice) cou1-lin6	6-1	大專 (junior college) daai6-zyun1	訂單 (order form) deng6-daan1
	3-4	帶來 (bring) daai3-loi4	價錢 (price) gaa3-cin4	4-3	行政 (administration) hang4-zing3	球賽 (game) kau4-coi3
2	1-2	醫院 (hospital) ji1-jyun2	污水 (sewage) wu1-seoi2	2-2	首位 (primacy) sau2-wai2	院長 (dean) jyun2-zoeng2
	3-2	澳門 (Macao) ngou3-mun2	雇主 (employer) gu3-zyu2	5-2	女人 (woman) neoi5-jan2	永久 (forever) wing5-gao2
	2-3	估計 (estimation) gu2-gai3	位數 (digit) wai2-sou3	2-4	會員 (member) wui2-jyun4	廣場 (square) gwong2-coeng4
	2-5	擁有 (possession) jung2-jau5	股市 (stock market) gu2-si5	2-6	演藝 (show) jin2-ngai6	穩健 (firm) wan2-gin6
	1-5	英語 (English) jing1-jyu5	高企 (keep high) gou1-kei5	3-5	建議 (suggestion) gin3-ji5	對上 (match) deoi3-soeng5
	5-5	女友 (girlfriend) neoi5-jau5	尾市 (end of market) mei5-si5	5-1	每天 (everyday) mui5-tin1	已經 (already) ji5-ging1
	5-4	往來 (intercourse) wong5-loi4	舞臺 (stage) mou5-toi4	5-6	理論 (theory) lei5-leon6	馬上 (immediately) maa5-soeng6

[N]: a district name of Hong Kong.

3.2.2.3 Procedure

The test procedures were similar to those in Experiment 1. Each stimuli pair contains a modified utterance and a natural one. The stimuli were presented in randomized order to the subjects through the same computer interface as in Experiment 1. At the beginning of each test, three pairs of stimuli were used for training and their results were not counted. Each sub-test lasted for about 15 minutes.

3.2.3. Results and analysis

Test results are shown as in Figure 3.9. Each vertical bar accounts for the 10 responses collected for a particular word. All of the modified utterances attain at least comparable perception to the natural ones, despite some individual differences on different words. Among the total 440 responses, 69.9% voted for "same", 13.4% for "modified speech", and 16.8% for "natural speech". For nine words in Group 1, the modified utterances attained equal perception to the natural ones. In Group 2, the superiority of the modified speech is even more noticeable.

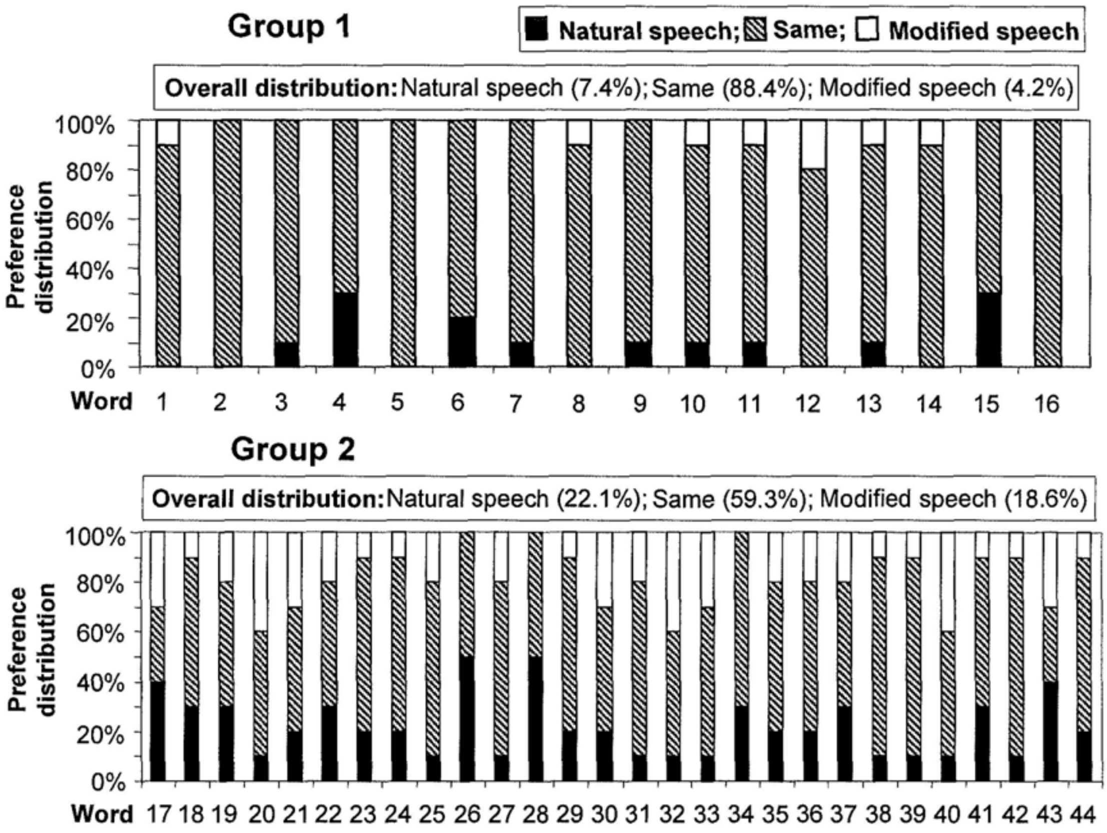


FIG. 3.9: Perceptual test results on the approximated F0 contours of disyllabic words. Each vertical bar corresponds to a particular word. The distribution of preference votes to "natural speech", "same" and "modified speech" is illustrated by different colors and textures. In Group 1, the carrier words contain only non-rising tones; in Group 2, each carrier word contains at least one rising tone.

Figure 3.10 shows the worst (Word 15) and the best (Word 12) perceived approximations among the words in Group 1. It is found that the approximated contour of Word 15 is actually closer to the observed contour than that of Word 12. This indicates that an approximation more closely resembling the observed contour does not necessarily lead to better perception.

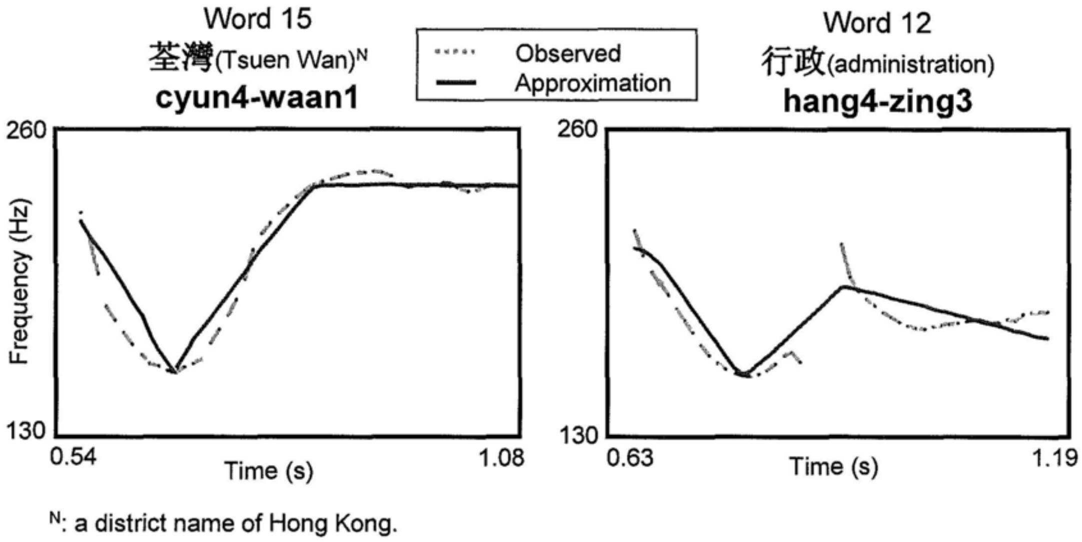


FIG. 3.10: Approximations of F0 contours of two example words in Group 1. The left one was perceived as the worst within the group, and the right was perceived as the best.

The average number of access times in Experiment 2 was 2.26 (the upper limit is 3). This indicates that the subjects had difficulties in perceiving the difference between a pair of stimuli. The subjects also verbally expressed that they could not tell the difference for most of the pairs.

The slopes of transition movements in the approximated contours were measured and the statistics is shown as in Figure 3.11. The transition slopes for words in Group 2 are averagely smaller than those in Group 1. An explanation is that in Group 2, if the second syllable in the word carries a rising tone, the boundary conflict would be weakened by the beginning falling movement in the rising tone. For the transition contour, the estimated value of glissando threshold is 6.9 ST/s (with average transition duration of 152 ms) or 98 Hz/s. Most of the transition slopes are much greater than the threshold and they vary over a wide range. However, we believe that perception is not sensitive to such rapidly changed transitions, unless their slopes are too high to exceed the perceptual tolerance and thus cause an unnatural "jump" (t' Hart, Collier and Cohen, 1990). In our approximations, there are cases that the transition slopes are over 800 Hz/s. These occurred on both voiced and unvoiced transitions, while perception seems not to be influenced. It suggests that the perceptual tolerance on the slopes of tone transitions should be higher than 800 Hz/s.

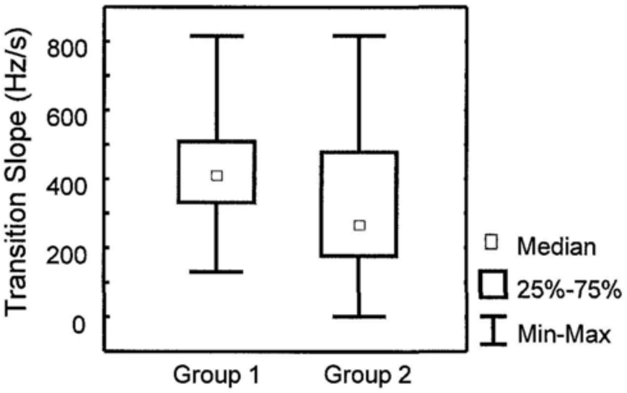


FIG. 3.11: Statistics of the transition slopes (absolute values in Hz/second) in the approximated F0 contours of disyllabic words.

3.3. Experiment 3: Perception of Approximated F0 Contours of Sentences

This experiment investigated linear approximations for F0 contours at sentence-level. The approximation procedures were the same as in Experiment 2.

3.3.1. Method

3.3.1.1 Subjects

Eight male and eight female subjects participated in the tests. All of them are new subjects. They were divided into two gender-balanced groups, for two sub-tests in which two kinds of approximations were studied. The recruitment of subjects followed the same requirement as in previous tests.

3.3.1.2 Stimuli

The speech materials consist of 5 naturally spoken sentences of Cantonese as shown in Table 3.6. Each sentence contains about 20 syllables. The natural utterances were taken from the *CUProsody* corpus (Li, 2003). They were recorded by a female speaker (denoted as Speaker II in this thesis) and were digitized at a sampling frequency of 16,000 Hz. For each natural utterance, two different approximations were manually created from its observed F0 contour. **FRa** was used to represent rising tones in one of the approximations and **LRa** was used in another. Figure 3.12 gives an example of these two types of approximations. The F0 levels of both natural and modified speech were down-shifted by 10 Hz, so as to compensate for the distortion of voice quality caused by time-domain F0 modification (Wong and Diehl, 2003).

Table 3.6. List of sentence carriers used in Experiment 3. Each carrier contains about 20 characters. All the characters are transcribed according to (LSHK, 1997). The natural speech of the sentence carriers are taken from the *CUProsody* corpus.

No.	ID in corpus	Sentence
1	4	我從小就很愛魔術，魔術就如我的女朋友，但我從沒想過可用這些伎倆救人一命。 ngo5 cung4-siu2 zau6 han2 ngoi3 mo1-seot6, mo1-seot6 zau6 jyu4 ngo5 dik1 neoi5-pang4-jau5, daan6 ngo5 cung4-mut6 soeng2-gwo3 ho2 jung6 ze5-se1 gei6-loeng5 gau3-jan4-jat1-ming6
2	21	一個開陽的露台將不再是置業者的奢想。 jat1-go3 hoi1-joeng4 dik1 lou6-toi4 zoeng1 bat1-zoi3 si6 zi3-jip6-ze2 dik1 ce1-soeng2
3	33	九廣鐵路紅磡車站外，昨天飄揚起一片綠色。 gau2-gwong2-tit3-lou6 hung4-ham3 ce1-zaam6 ngoi6, zok3-tin1 piu1-joeng4-hei2 jat1-pin3 luk6-sik1
4	61	中國入世會使到世貿組織真正成爲全球的多邊貿易組織。 zung1-gwok3 jap6-sai3-wui6 si2-dou3 sai3-mau6-zou2-zik1 zan1-zing3 sing4-wai4 cyun4-kau4 dik1-do1 bin1 mau6-jik6 zou2-zik1
5	151	警方今日呼籲市民，在非緊急的情況下切勿致電九九九緊急服務中心。 ging2-fong1 gam1-jat6 fu1-jyu6 si5-man4, zoi6 fei1-gan2-gap1 dik1 cing4-fong3-haa6 cit3-mat6 zi3-din6 gau2-gau2-gau2 gan2-gap1 fuk6-mou6 zung1-sam1

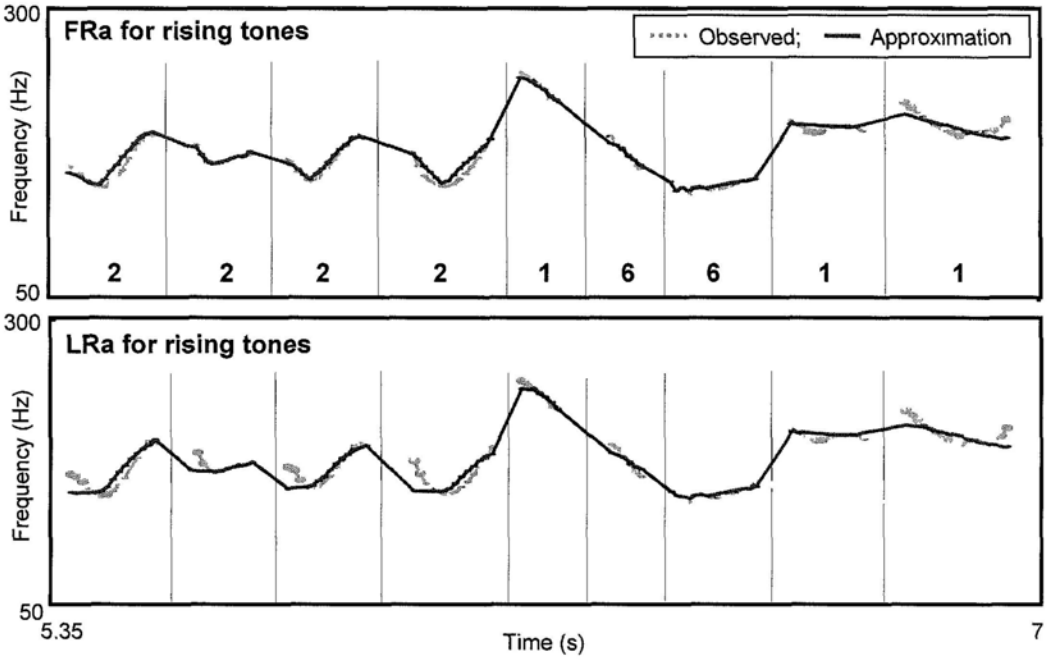


FIG. 3.12: The two different types of approximations for sentence-level F0 contour. The shown F0 contour is a part of Sentence 5 in the test materials. The vertical dashed lines are the syllable boundaries and the numerals refer to the respective tones. In the upper figure, **FRa** is used for approximating rising tones, and in the lower figure, **LRa** is used.

3.3.1.3 Procedure

The two types of approximations were tested in sub-test I and sub-test II respectively, following the similar procedures as in previous tests. At the beginning of each test, one pair of the stimuli that carry a sentence different from those in Table 3.6 was used for training and the result was not counted.

3.3.2. Results and analysis

Results of the two sub-tests are shown as in Figure 3.13. In each sub-test, 8 responses were collected for a particular sentence, and its preference distribution is illustrated by a vertical bar in the figure. In sub-test I, all of the modified utterances were perceived to be at least comparable to the natural ones. The modified speech of Sentence 2 and Sentence 4 even attained better perception than the natural speech. In terms of the distribution of the total responses, modified speech also received the most preferences (45%). In sub-test II, the use of a less accurate approximation for rising tones led to a noticeable perceptual difference and hence a decrease in the preferences to the modified speech (half of that in sub-test I).

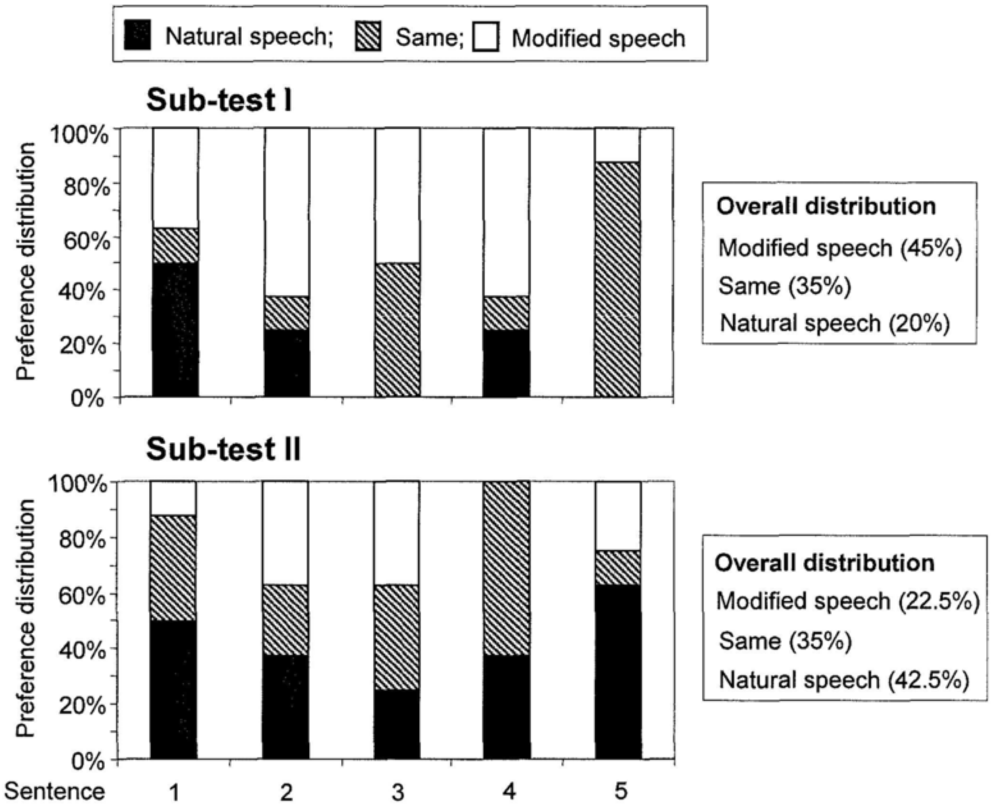


FIG. 3.13: Perceptual test results on the approximated F0 contours carried by 5 sentences. In sub-test I, **FRa** was used for rising tones; in sub-test II, **LRa** was used.

Paired t-test showed that the difference between the results of the two sub-tests is not statistically significant ($p > 0.1$). Nevertheless, we believe that **FRa** is more appropriate than **LRa** in representing rising tones. For example, for Sentence 5, the approximated contour using **LRa** failed to attain comparable perception, while equal perception was reached by using **FRa**. This sentence contains four successive syllables carrying Tone 2 (see Figure 3.12), making the deficiency of **LRa** very prominent.

3.4. Discussion

3.4.1. Linear approximation

In this study, linear approximations of the observed F0 contours in Cantonese speech were examined. It is found that these approximations adequately represent perception-relevant F0 variations such that comparable perception can be attained. With this finding, linear F0 movements are connected with speech naturalness in tonal languages, in addition to their demonstrated roles in tone identification. While linear approximation was studied in previous research, our findings suggest that such approximations are greatly constrained by the pitch patterns of lexical tones. The approximations developed for intonation, e.g., the IPO method (t' Hart, Collier and Cohen, 1990) assuming that one movement is sufficient for one syllable, and the automatic approximation applying the glissando and the differential glissando thresholds (d'Alessandro and Mertens, 1995), are either under-approximating or over-approximating for tone contours. Nevertheless, these evidences from both tonal and non-tonal languages support a general applicability of linearly approximated F0 contours in attaining comparable perception, although language-specific constraints may be needed in determining the approximations.

Being a defining attribute of linear movement, movement slope is closely correlated with the perception of F0 variations. This investigation made efforts to examine the movement slopes in the approximated contours. Compared with the findings of psychoacoustic experiments, our study revealed a possible lifting of perceptual tolerance on F0 variations in speech. The reason might be that speech is a complex signal and delivers multiple communicative functions simultaneously, therefore "*speech is conveyed using robust cues that do not severely tax the discrimination abilities of the auditory system*" (Moore, 1997: 557; Connell, 2000). Associating with the tone function, this finding is also consistent with the statement that "*the pitch changes that are linguistically relevant are much larger than the limits of pitch discrimination measured psychophysically*" (Moore, 1997: 557). Meanwhile, the linguistic experience of listeners may also play a role in such kind of tolerance lifting (Lee, Vakoch and Wurm, 1996; Burnham and Francis, 1997).

In previous studies there has been an argument on whether the endpoint or the changing slope of the F0 contours is used by the listeners to differentiate the contour tones (Francis, Ciocca and Ng, 2003). Our analysis of tonal movement slopes in Experiment 1 supports that the changing slope is highly correlated with the perception of contour tones. Based on this, our analysis additionally suggests that the movement slope of a particular Cantonese contour tone might need to change in a certain range. The range for "natural" contour should be more restrictive than the one for "identifiable" contour. For studies related to speech naturalness, investigations on perceptual tolerance that allows variants of an underlying target to change are interesting and meaningful.

Comparable perception of the linearly approximated F0 contours leads to another interesting finding that has not been evidently proved before. That is, a sharp or "non-smooth" F0 change around a turning point does not cause any side-effect for perception of the F0 contour; while in most previous works on generating F0 contours for the speech with high naturalness, it has been assumed that smooth change of F0 around turning points is needed (Kochanski and Shih, 2003; Li, Lee and Qian, 2004; Fujisaki et al, 2005; Ni and Kawai, 2006; Prom-on, Xu and Thipakorn, 2009).

Turning points are commonly considered important in perception of F0 contours. Previous studies suggested that the temporal position of the turning point in an F0 contour is an acoustic cue for identification of Mandarin tones (Shen and Lin, 1991; Shen, Lin and Yan, 1993; Jongman et al, 2006), while this might not be true for Cantonese (Francis, Ciocca and Ng, 2003; Khouw and Ciocca, 2007). We believe that in Cantonese turning points are more closely related with the naturalness of perceived speech than with tone identification / differentiation. In linear approximations, the inclusion or exclusion of a turning point implies a change in the number of constituent movements in the contour; and time shifting of a turning point would affect the perception of the two movements connected by it, and furthermore modify the linguistic interpretation of the perceived movement sequence (House, 2004; Kohler, 2005). In this investigation, turning points play a role in the approximation of rising tones as well as tone transitions. Their temporal positions have not been investigated in depth. Our understanding about the perceptual functions of turning points in Cantonese F0 contours is still very limited.

The current results may not suffice to conclude that the linear approximations implemented in our study are the best approximations for the respective F0 contours. Nevertheless, it is the simplest one that we have been able to find. Simplified representations will be useful in general to the research on speech prosody. By approximation, perception-irrelevant elements are removed, and prosody-related perceptual cues can be more easily identified. The use of linear approximation helps to confine the study of perception-critical F0 variations to a limited number of factors. Moreover, our results will be useful in a number of practical applications, for examples, language learning, text-to-speech synthesis, speech recognition, speech enhancement and hearing aid. In an effort on improving Cantonese speech perception of cochlear implant users (Yuan, 2009), linear approximation has been considered as a good solution for efficient prediction of F0 contours.

A limitation of this work is that the approximations were obtained manually. As a basic and pilot study, manual process is acceptable. But for practical usages, such a process will be subject to the problems of low efficiency and inconsistency. We believe that this limitation can be overcome. As we can see from the manual process, a proper approximation for an observed contour is not unique, due to the existence of perceptual tolerance. Consequently, automatic approximation is possible if the perceptual knowledge can be properly applied. In Chapter 5, we will introduce our continuous efforts on the development of a framework of automatic approximation. The preliminary results on approximating word-level F0 contours are very promising.

3.4.2. Perceptual preference on simplified representations

In our perceptual tests, there are a number of cases that the approximated contours were more preferred than the natural one. In the sentence-level experiment, the modified utterances with approximated contours even obtained a much higher overall preferences than the natural ones. These results indicate a possibility of perceptual preference on simplified acoustic representations than the naturally produced contours, though they are still not sufficient to support a general conclusion along this direction. As discussed in Section 1.1, such a preference can theoretically get supports from the view of selective perception (House, 2004). A simplified F0 contour can potentially reduce the perceptual efforts on processing it and consequently enhances the perception of other tasks.

3.5. Conclusions

In this chapter, we described a perceptual study on the linearly approximated F0 contours in Cantonese speech, at syllable, word and sentence levels. The approximated contours were found to adequately represent perception-sensitive F0 variations in Cantonese speech. It was consistently observed that the modified speech utterances carrying linearly approximated contours could attain comparable or equal perception to the natural speech. Hence, linear F0 movements are associated with the speech naturalness in tonal languages, in addition to their essential functions in tone identification. Each non-rising tone in Cantonese can be approximated by one linear movement and each rising tone needs to be represented by two movements. For the transition between two neighboring tones, one linear movement is appropriate and sufficient. The approximated F0 contours were examined in a series of perceptual tests, with speech materials from different speakers. The speech carriers were designed to have different lengths and rich segmental variations. The test results lead to general conclusions for Cantonese F0 contours. We also analyzed the slopes of the linear movements and compared them with previous psychoacoustic findings. We concluded that the contour slope is highly correlated with perception. It is also found that the thresholds of perceiving pitch movements in speech might be higher than those found in psychoacoustic studies.

Publications

Li Y.-J. and Lee Tan (2007). "Perceptual equivalence of approximated Cantonese tone contours," *Proc. Interspeech 2007*, pp. 2677-2780.

Li Y.-J. and Lee Tan (2008). "A perceptual study of approximated Cantonese tone contours," *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP) 2008*.

Li Y.-J. and Lee Tan (2010). "A perceptual study on linearly approximated F0 contours in Cantonese speech," submitted journal paper.

Chapter 4

Perception and Analysis of Approximated F0 Contours of Polysyllabic Words

In the previous chapter, linear approximations of the F0 contours in Cantonese speech were preliminarily examined. For focusing on F0 variations, the used speech carriers were carefully selected. Although the stimuli cover materials at different levels, namely syllable, word and sentence, the lexical and segmental variations are relatively limited. In this chapter, the generalization ability of linear approximation is further investigated with a large set of polysyllabic Cantonese words. Perceptual results clearly validate the effectiveness of the linear approximations of F0 contours.

Subsequently statistical analysis of the amount of generated linear approximations is carried out. The properties of linear F0 movements in Cantonese speech are learned in-depth through the examination of each determining attribute. This also allows an attempt of quantitatively interpreting F0 variations of tones and tone transitions. Thus, our understanding about linear F0 variations in relation to communicative functions is expected to be greatly improved.

Lastly, the issue of objective evaluations and perceptual performance is addressed. Two objective evaluations of the modified F0 contours, root mean square (RMS) error and contour correlation are compared with the true perceptual performance. It is found that neither of these objective measurements gives good prediction on perceived speech naturalness.

4.1. Perception of Approximated F0 Contours of Polysyllabic Words

F0 contours of Cantonese polysyllabic words are manually approximated using the same method as in the basic study. The accordingly generated utterances are examined through a large-scale perceptual test.

4.1.1. Method

4.1.1.1 Subjects

Twenty-six new subjects (fifteen female and eleven male) participated in the listening test. They are undergraduate or postgraduate students at the Chinese University of Hong Kong, aging from 20 to 30. All of them are native Cantonese speakers with normal hearing. None of them has professional knowledge about speech technology or linguistics. The subjects were paid for their workload.

4.1.1.2 Stimuli

335 Cantonese polysyllabic words were used as speech carriers. The words are 4 to 6 syllables long and carry only non-rising tones. The words are listed in Appendix 3.

The natural utterances of the words were obtained from the speech corpus *CUWord* (Lo, Lee and Ching, 1998). CUWord is a corpus containing a large set of Cantonese words recorded from multiple native speakers. The recording was carried out using a high-quality close-talking microphone in a quiet room. The recorded signals were digitized at a sampling frequency of 16,000 Hz. The utterances from a female speaker (denoted as Speaker III in this thesis) were used in this study. The speaker was selected based on the subjective judgment of the speech quality and naturalness.

The F0 contour of each natural utterance is manually approximated as described in the basic study. The F0 levels of both the observed and the approximated contours were down-shifted by 10 Hz, so as to compensate for the distortion of voice quality caused by time-domain F0 modification (Wong and Diehl, 2003).

4.1.1.3 Procedure

The procedures of perceptual tests are similar to those described in the preceding chapters. The modified utterances and the natural utterances were compared in pair. Apart from the 335 word pairs, the test stimuli additionally include about 10% "control" pairs (very different stimuli in a pair or identical stimuli in a pair), which serve to monitor whether a subject performs the comparison task in a reasonable manner.

In each test, a subject was required to go through half of the full set of test data, about 190 pairs of unrepeated words (10% of them are control pairs). All the stimuli pairs were presented to the subjects through the same interactive computer interface as in the previous tests. The presentation order of word pairs and the order of stimuli within a pair are randomized. The first three pairs of stimuli in each test are training pairs and their test results are not counted. The subjects were not informed of this arrangement. Each test lasted for about 40 minutes, including two mandatory 5-minute breaks. After each test, the subject was asked to give comments for an overall impression of the heard pairs.

4.1.2. Results and analysis

The results of six subjects (five female and one male) were discarded, due to their unreasonable responses to the control data. The total number of effective responses is 3350. Among them, 76.8% are for "same", 8.5% for "modified speech" and 14.7% for "natural speech" (Table 4.1).

Table 4.1. Perceptual test results on the approximated F0 contours of polysyllabic words — preference distribution of the total 3350 effective responses.

to "same"	to "modified speech"	to "natural speech"
76.8%	8.5%	14.7%

10 responses were collected for each word pair. If not more than half of them vote "natural speech", the modified speech is said to attain *comparable perception* with the natural speech. About 97.3% of the modified utterances attain comparable perception. According to the obtained votes on "same" and "modified speech" for each word pair, a distribution of the 335 word pairs is calculated. The result is shown in Figure 4.1. The perceptual test results on individual word pairs are given in Appendix 3. Comments from all the subjects indicated that the two stimuli in most pairs are hard to be differentiated from each other.

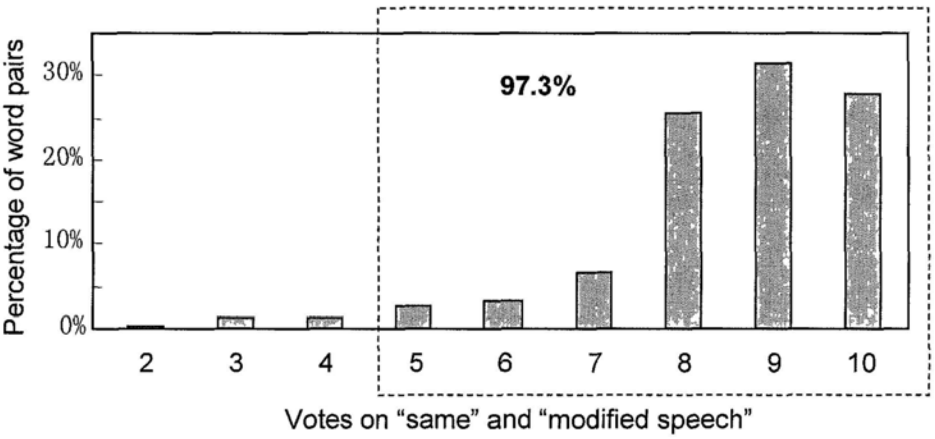


FIG. 4.1: Perceptual test results on the approximated F0 contours of polysyllabic words — distribution of the 335 word pairs according to the obtained votes on "same" and "modified speech" for each word pair (upper limit is 10).

The test results in this investigation confirm that linear approximation of F0 contours is applicable to different lexical and segmental variations, and also to the speech of different speakers. This leads to a generalizable conclusion that F0 contours in Cantonese speech can be approximated by linear movements, such that perception of the speech is not affected.

4.2. Analysis of Linear F0 Movements

The amount of generated linear approximations in the previous experiment, as having been perceptually confirmed, can be considered as a kind of perceptual data. These approximations can be looked on as connections between human produced and perceived F0 variations. On the one hand, they describe the major trends of human generated F0 variations; and on the other hand, they are greatly associated with perception. Analysis of them allows us to examine how the communicative functions of F0 variations, such as different tones, are processed, from a special viewpoint.

Given linear approximations, the properties of an F0 contour can be described by the attributes of its linear approximation. Being the major attributes, movement slopes, movement heights and time locations of turning points in the generated approximations were analyzed.

4.2.1. Slopes of linear movements

Table 4.2 and Figure 4.2 give the statistics of the movement slopes in the approximated contours, for different tone and tone transition categories. As expected, most slopes of tone movements are negative (falling movement). For non-entering tones, the movement slopes of Tone 1 approximations appear to be much smaller than others, averagely around -5 Hz/s (almost level movement), and with many cases of positive values (rising movement). Movement slopes of Tone 3 and Tone 6 are similar. Movement slopes of Tone 4 are large, more than two times of those for Tone 3 and Tone 6, and meanwhile vary in a large dynamic range. A very interesting finding is that the movement slopes of the four investigated non-entering tones clearly fall into three non-overlapped ranges, with those of Tone 3 and Tone 6 in a similar range. This finding indicates that movement slope could be at least acoustically a rather reliable feature for differentiating these tones. Given the importance of movement slopes in perceiving Cantonese tone contours (Khouw and Ciocca, 2007; Li and Lee, 2008), the separate ranges of slope values among different tones should be perception desired as well. Meanwhile, this finding is quite in line with our earlier discussion in Section 3.4.1 that the movement slope of a particular Cantonese tone might vary within a certain range. The relative relation among the movement slopes of different non-entering tones in this analysis is rather consistent with that observed in the basic study, which was on the approximations of isolated non-entering tone contours (refer to Section 3.1.3). The glissando threshold for these non-entering tone movements carried by polysyllabic words is estimated to be 50 Hz/s, given the movement duration of 200 ms and the reference F0 level of 200 Hz. Except Tone 1, the movement slopes of non-entering tones are much higher than the threshold. Whether these movements can be perceived is not for sure due to the lifted perceptual tolerance on F0 variations in speech (refer to Section 3.1.3).

Table 4.2. Analysis of movement slopes — statistics of the movement slopes in the approximated F0 contours of polysyllabic words, for different tone and tone transition categories.

Group	**Tone/transition**	**No. of cases**	**Mean slope** (Hz/s)
Non-entering tones	Tone 1	304	-5
	Tone 3	197	-180
	Tone 4	313	-429
	Tone 6	200	-173
Entering tones	Tone 1	97	-246
	Tone 3	92	-445
	Tone 6	142	-522
Transitions	Between non-entering tones	717	-309
	Before entering tones	265	-289
	After entering tones	244	-197

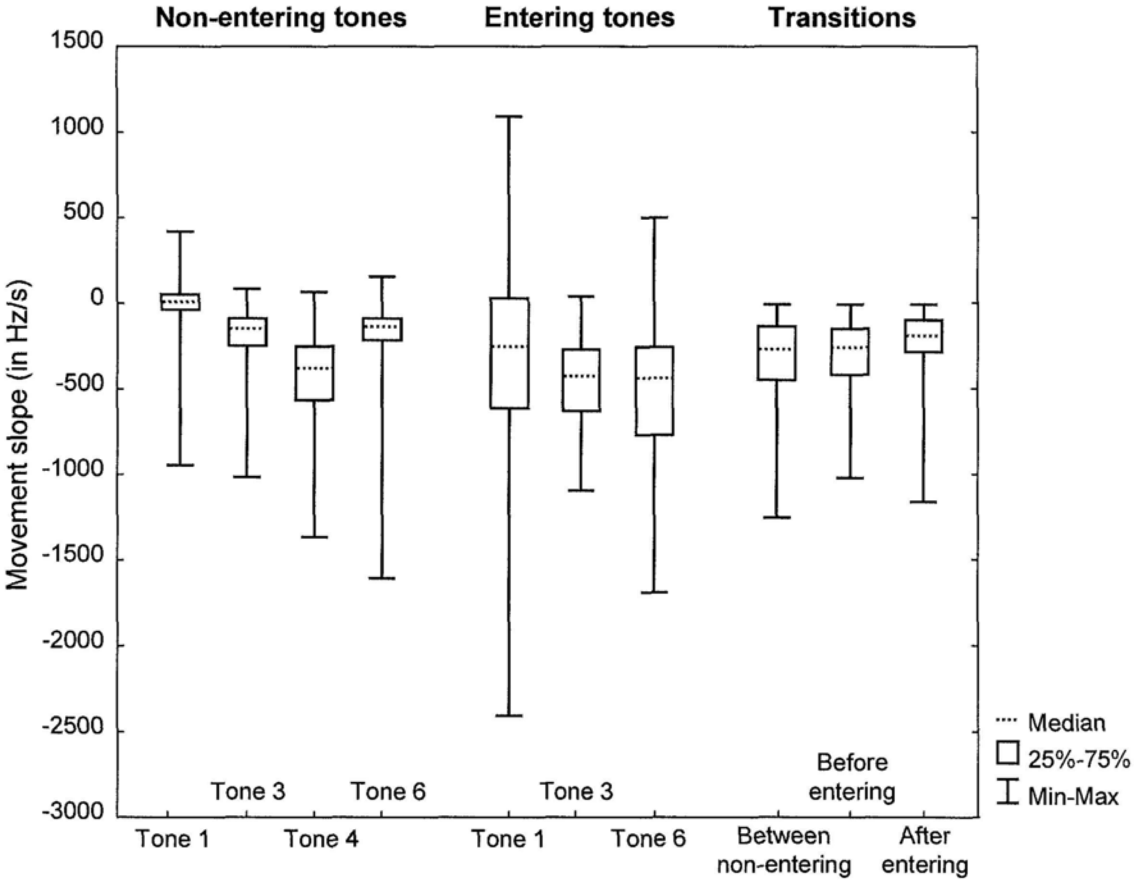


FIG. 4.2: Analysis of movement slopes — box plot of the movement slopes in the approximated F0 contours of polysyllabic words, for different tone and tone transition categories.

Movement slopes of entering tone approximations vary greatly. They are generally much larger than those of their non-entering counterparts. At the same time, these slopes are much higher than the estimated glissando threshold (about 100 Hz/s, given the movement duration of 150 ms and the reference F0 level of 200 Hz). It is noted that entering Tone 3 and Tone 6 each has a non-overlapped slope range with their non-entering counterparts. Nevertheless, for entering tones themselves, slope ranges are severely overlapped. Consequently, movement slope might not be a good feature for differentiating entering tones.

For the ease of comparison, the movement slopes of tone transitions are all converted into negative values. Both the cases of "between non-entering tones" and "before entering tones" have an average slope about 300 Hz/s, which are comparable with the finding in the basic study. Unexpectedly, the case of "after entering tones" has a lower value of average slope of 197 Hz/s. High movement slopes were initially supposed to appear in such cases, as stop coda would introduce in an unvoiced transition after each entering tone and accordingly an F0 reset of the succeeding tone. At this stage, we cannot yet find a reasonable explanation for this unexpected observation.

4.2.2. Heights of tone movements

Movement height is another important attribute of linear movements. It is also generally considered as important perceptual cue for identification of different Cantonese tones (Fok, 1974; Gandour, 1981, 1983; Vance, 1977; Khouw and Ciocca, 2007). Table 4.3 and Figure 4.3 give the statistics of the movement heights in the approximated contours, for different tones.

Table 4.3. Analysis of heights of tone movements — statistics of the movement heights in the approximated F0 contours of polysyllabic words, for different tones.

Group	**Tone**	**No. of cases**	**Mean height** (Hz)
Non-entering tones	Tone 1	310	274
	Tone 3	199	215
	Tone 4	315	179
	Tone 6	202	198
Entering tones	Tone 1	97	273
	Tone 3	92	218
	Tone 6	144	201

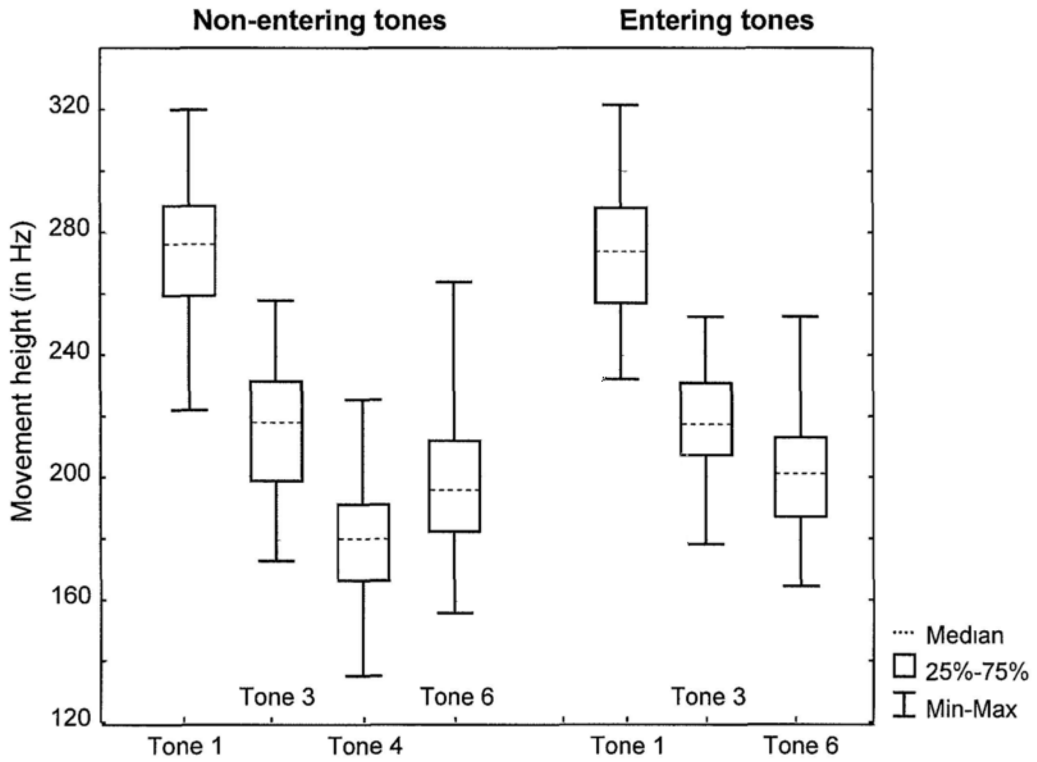


FIG. 4.3: Analysis of heights of tone movements — box plot of the movement heights in the approximated F0 contours of polysyllabic words, for different tones.

The movement heights of Tone 1, Tone 3, Tone 6 and Tone 4 are reasonably ordered from high to low, consistent with their phonological definitions. The movement height of an entering tone is similar to its non-entering counterpart. Movement heights of all the investigated tone categories vary with a comparable standard deviation.

Heights of non-entering Tone 3 and Tone 6 are expected to be discriminative, since their movement slopes have been found similar. However, the statistics do not give strong evidences on this hypothesis. Then how these two tones are perceptually differentiated in continuous speech becomes a very interesting question. On the other hand, the heights of entering tone contours are much more differentiable than their movement slopes. Hence, the contour heights are considered to be important acoustic cues and possible perceptual cues for differentiating entering tones.

It should be noted that being perceptual or acoustic cue, absolute height of a tone movement is not as reliable as its slope, since the height is easily affected by long-term downtrend or other short-term communicative functions like focus or emphasis. Our statistics is derived from the tone movements in polysyllabic words in which downtrend effect is weak and no other special communicative function is embedded.

4.2.3. Time locations of turning points

In the perceptual data, all the tone contours were approximated by a single movement. Hence, a turning point occurs to connect a transition movement with a tone movement, as that shown in Figure 4.4. In this analysis, we investigated the time locations of turning points with reference to the syllable boundaries. The syllable alignments were obtained by HMM force alignment. The location of a turning point is associated with the location where a transition starts or ends. The ratio of the pre-tone transition duration $d_{pre-tone}$ (as shown in Figure 4.4) relative to the syllable duration d_{syl} was calculated, i.e.,

$$r_{pre-tone} = d_{pre-tone} / d_{syl} \tag{4.1a}$$

Similarly, the ratio of the post-tone transition duration $d_{post-tone}$ in relation to the syllable duration d_{syl} was calculated as,

$$r_{post-tone} = d_{post-tone} / d_{syl} \tag{4.1b}$$

A negative value of the ratio means that the transition part starts or ends outside the syllable region. Examination of these ratios helps to learn how a transition extends within a syllable.

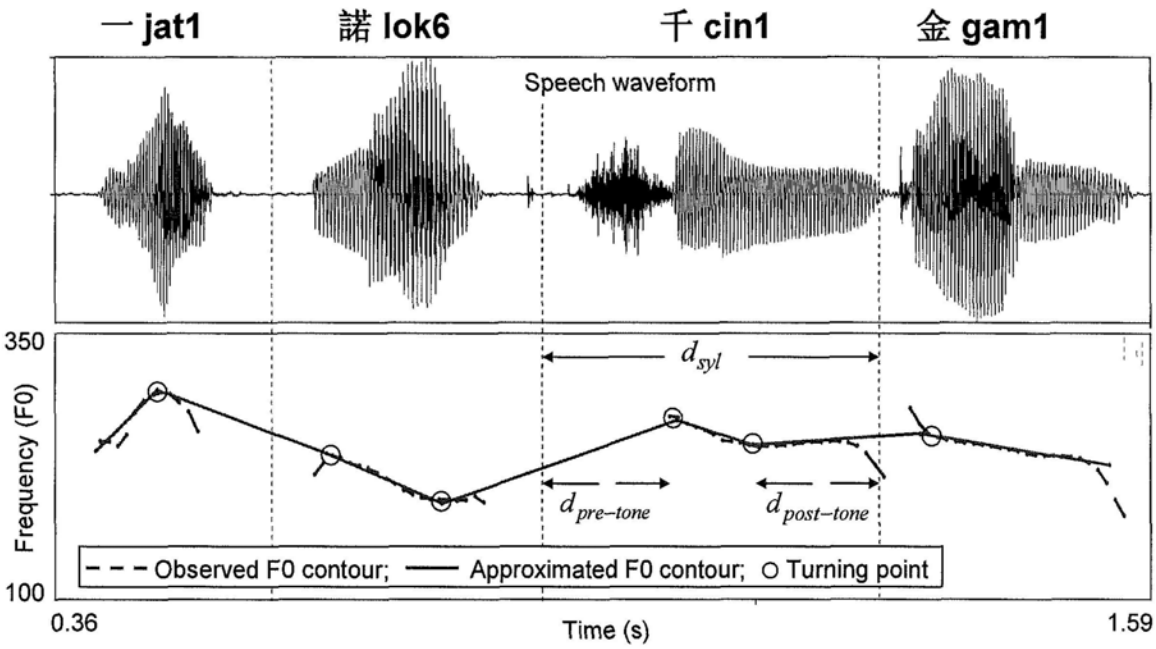


FIG. 4.4: Time locations of turning points in the linearly approximated F0 contour of a polysyllabic word. The dashed lines indicate syllable boundaries.

Statistics of the two ratios are given in Table 4.4 and Figure 4.5. Entering tones and non-entering tones were investigated separately. For non-entering tones, it appears that a pre-tone transition often extends to one third of a syllable region; while a post-tone transition generally starts at the last one sixth of a syllable region. The larger pre-tone ratios should come from the fact that most transitions need to go across a consonant region. For entering tones, pre-tone ratios are similar to the case of non-entering tones. Nevertheless, their post-tone ratios are much higher. This may be related with the unvoiced regions at the end of the syllables, which are introduced in by the stop codas. In summary, pre-tone transitions extend to around one-third of syllable regions; post-tone transitions start at a time much earlier in the syllables carrying entering tones than in the syllables carrying non-entering tones. Most ratios are smaller than 0.5, indicating that a transition rarely extends after or starts before the half of a syllable region. The ratios are all positive, implying that all the transitions stretch into the both connected syllables.

Table 4.4. Analysis of time locations of turning points — statistics of the ratios of pre-tone transition duration relative to syllable duration and the ratios of post-tone transition duration relative to syllable duration.

Group	Ratio type	No. of cases	Mean ratio
Non-entering tones	$r_{pre-tone}$	757	0.34
	$r_{post-tone}$	779	0.15
Entering tones	$r_{pre-tone}$	267	0.32
	$r_{post-tone}$	245	0.38

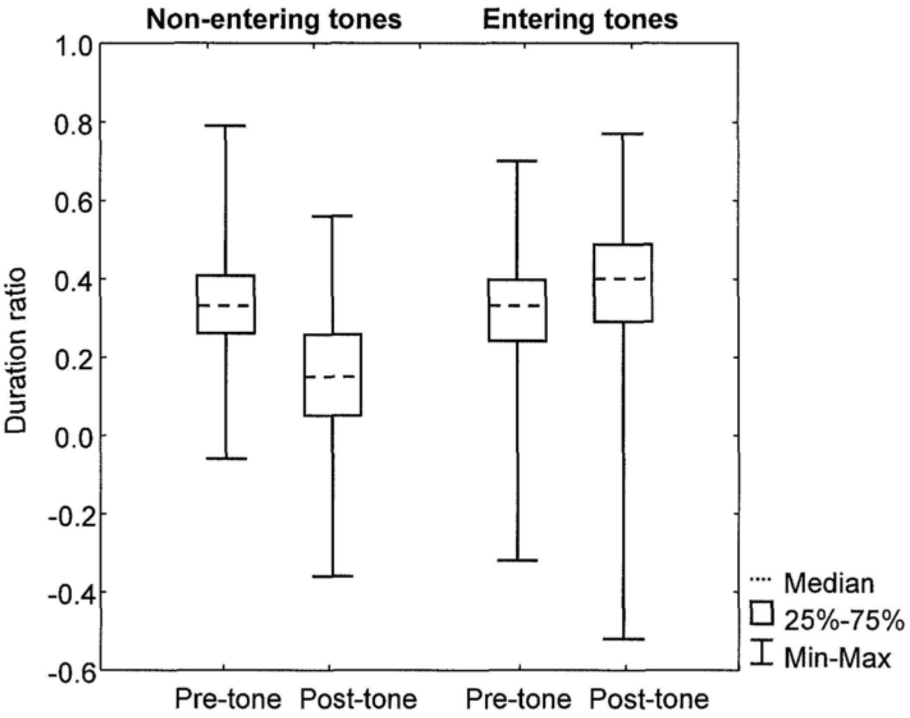


FIG. 4.5: Analysis of time location of turning points — box plot of the duration ratios for the four different cases.

4.3. Objective Evaluations and Perceptual Performance

For the development of speech synthesis systems, the evaluation of predicted F0 contours is an interesting and difficult problem. The root mean square (RMS) error (Kochanski and Shih, 2003; Fujisaki et al, 2005; Ni and Kawai, 2006; Prom-on, Xu and Thipakorn, 2009) and the correlation (Black and Hunt, 1996; Dusterhoff, Black and Taylor, 1999; Hu and Loizou, 2008) between the predicted and the observed contours have been commonly used as objective evaluation for the performance. A small RMS error or a high degree of correlation is assumed to imply good perceptual quality.

However, our analysis suggests that these performance indices are not good indicators as far as speech perception is concerned.

4.3.1. RMS error and perceptual performance

Consider an observed F0 contour $F0_o(n)$ and its linear approximation $F0_a(n)$ (n is the index of discrete time in the unit of 10 ms), both with length N, the RMS error between the two contours is calculated as,

$$error = \sqrt{(\sum_{n=1}^{N}(F0_o(n) - F0_a(n))^2)/N} \tag{4.2}$$

For each of the 335 tested contour pairs (observed vs. approximated), the RMS error was computed. The average RMS error is shown to be 10 Hz. Let the RMS error and the perceptual test result (for each stimuli pair, the votes on "natural speech") be two random variables X and Y. The relevance between them can be measured by the following correlation coefficient,

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \tag{4.3}$$

where $\rho_{X,Y}$, COV, E, μ, σ denote the normalized correlation coefficient, covariance, expectation, mean and standard deviation respectively. Our data gives a normalized correlation coefficient of 0.1, indicating that X and Y are uncorrelated. A scatter plot of the two variables is shown as in Figure 4.6. There are many inconsistent cases. For examples, an approximated contour with 6 Hz RMS error obtains only 20% preference votes, and the one with 36 Hz RMS error gives equal perception.

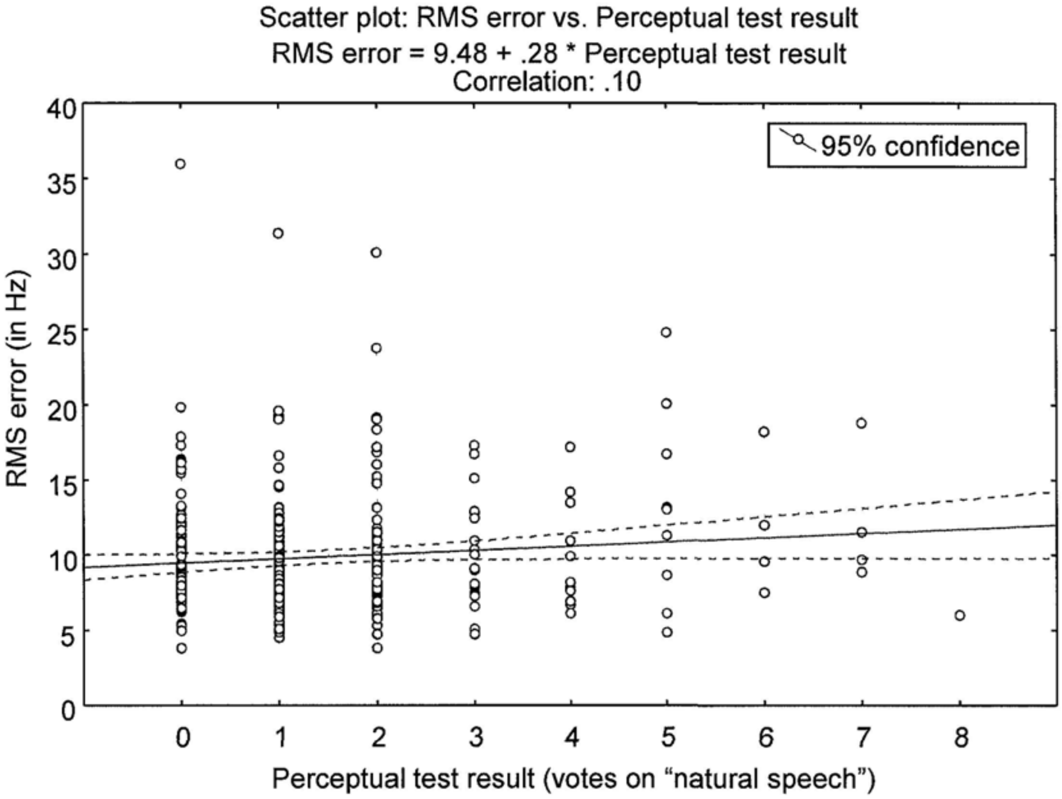


FIG. 4.6: RMS error and perceptual test result — scatter plot of the two variables.

4.3.2. Contour correlation and perceptual performance

For each of the test words, the correlation between the approximated and the observed F0 contours is also measured. For an observed contour $F0_o(n)$ and its approximation $F0_a(n)$, their correlation is calculated as,

$$r = \frac{1}{N}\sum_{n=1}^{N}\frac{(F0_o(n)-\mu(F0_o))(F0_a(n)-\mu(F0_a))}{\sigma(F0_o)\sigma(F0_a)} \tag{4.4}$$

where r, μ, σ denote normalized correlation coefficient, mean and standard variation respectively. The average correlation is 0.95. This implies a high similarity between the observed F0 contours and their approximations.

To evaluate the relevance between the two variables "contour correlation" and "perceptual test result", their correlation is calculated according to Equation 4.3. The result gives a normalized correlation coefficient of -0.09 and suggests a weak relation between them. Scatter plot of the two variables is shown in Figure 4.7. Inconsistency of the two variables is obvious.

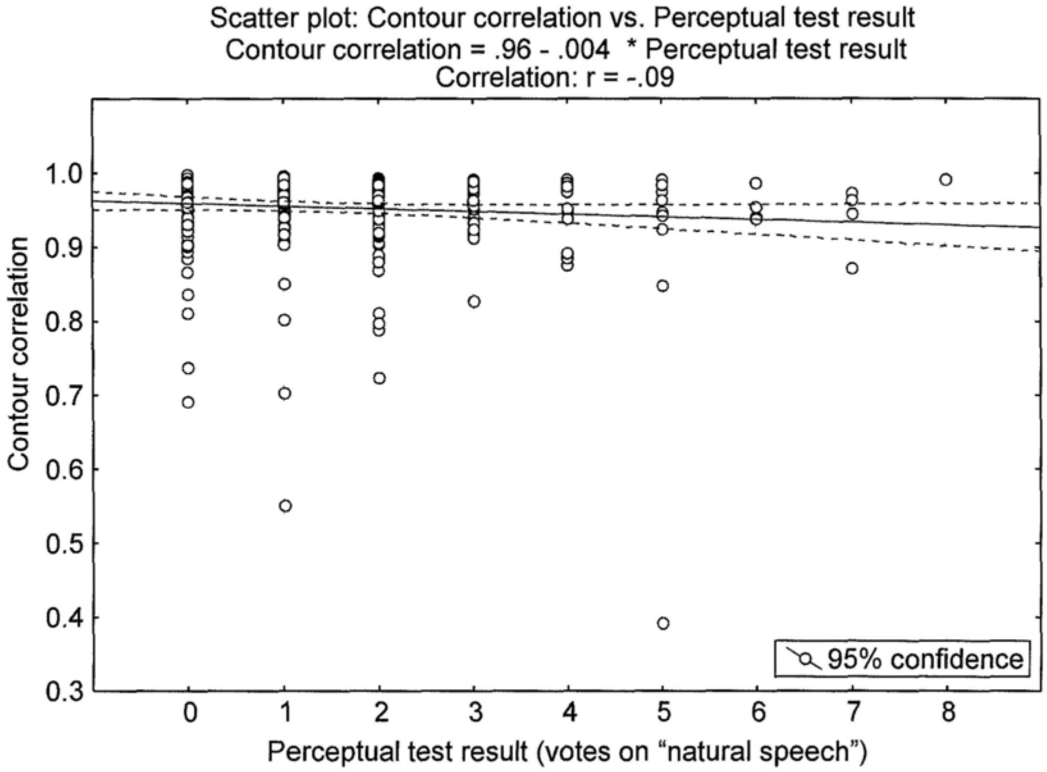


FIG. 4.7: Contour correlation and perceptual test result — scatter plot of the two variables.

4.3.3. RMS error and contour correlation

Being the two carried-out objective evaluations, RMS error and contour correlation were compared as well. Figure 4.8 shows the scatter plot of the two variables. Correlation between them is 0.59 (normalized correlation coefficient). This indicates that the two objective evaluations are different to a certain extent. In our case, we consider moving trends of a contour is important for perception, hence contour correlation would be a more reliable one between the two objective evaluations. In addition, in the examination of the cases with low contour correlation or high RMS error, we found that the two evaluations are very sensitive to the outlier points in the contours.

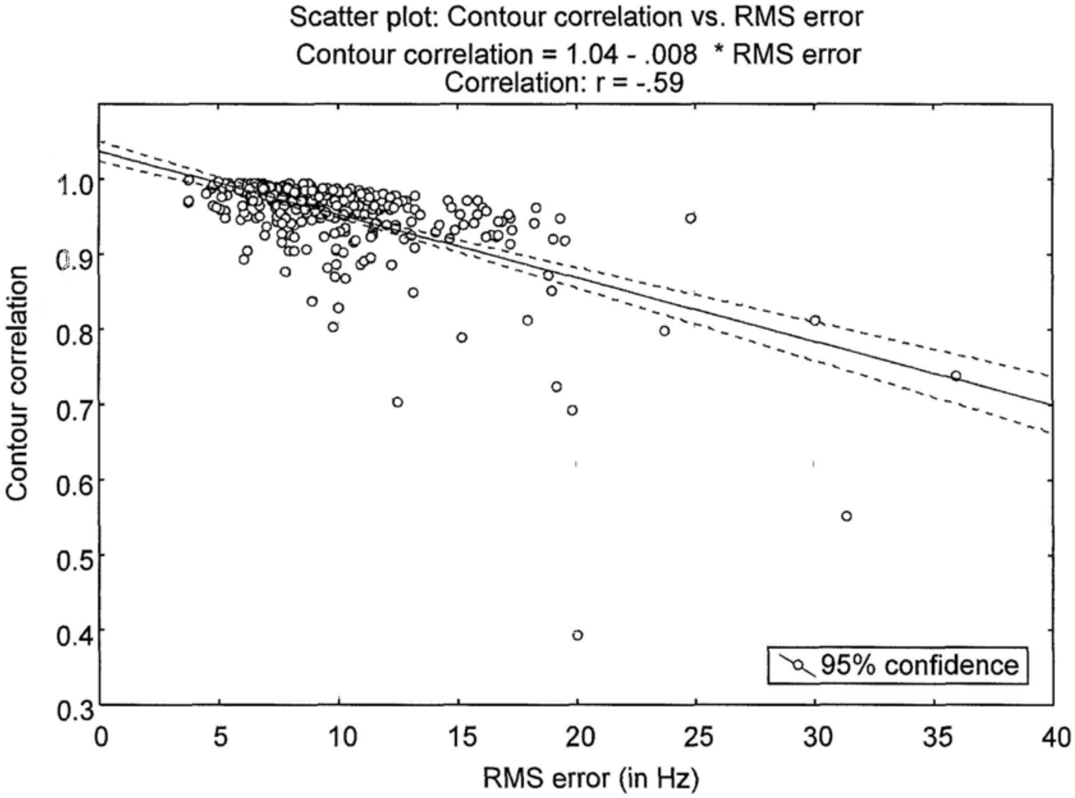


FIG. 4.8: Contour correlation and RMS error — scatter plot of the two variables.

4.3.4. Indications from objective evaluations

Examination of the objective evaluations with the perceptual performance clearly shows the inconsistency between speech perception and production. Nevertheless, such inconsistency would be very useful for investigations on perception-sensitive acoustic features. For example, if an approximated contour has a large RMS error but good perceptual performance, it must have successfully retained the most desirable perceptual information. Oppositely, if a contour has a small RMS error but poor perceptual performance, some perception-sensitive variations must have been removed by the approximation process.

4.4. Discussion and Conclusions

Based on our findings in the basic study, linear approximation was further examined on the F0 contours of hundreds of polysyllabic words. It is confirmed that regarding to maintaining the comparable perceptual quality, F0 contours in Cantonese speech can be represented by their linear approximations. This is a general observation, not affected by segmental variations, lexical variations, contextual variations, length of utterances or speaker differences.

The hundreds of comparably perceived linear approximations are considered as a set of perceptual data. The data set provides an opportunity to learn the properties of linear approximations in a statistical sense. All the determining attributes of approximations, i.e., movement slopes, movement heights and time locations of turning points were examined through the acoustic analysis of the approximated F0 contours. Although the perceptual experiment concerns perceived speech naturalness, the analysis shows interesting findings that can be associated with tone identification. Attributes of linear F0 movement though have been extensively examined in relation to tone identification, in this investigation they are accessed for the first time with a pre-condition that the investigated F0 contours have ensured perceptual quality in continuous speech.

The analysis results of movement slopes are quite consistent with those obtained from isolated syllables. Moreover, movement slopes of four investigated non-entering tones are found to fall into three non-overlapped ranges, indicating that movement slope is possibly a good feature for differentiating these tones. Movement heights keep stable between an entering tone and its non-entering counterpart. Meanwhile, movement heights are rather acoustically differentiable among different tones. Since both the movement slopes and the movement heights of Tone 3 and Tone 6 are severely overlapped in their changing ranges, the two tones are predicted to be not as differentiable from each other as from other tones. A tone identification test on the linearly approximated F0 contours is expected. Thus, the function of linear approximations in tone identification could be better interpreted. Then we can have a more in-depth and systematic understanding about perceiving F0 contours in tonal languages. In the examination of the time locations of turning points, the turning points are interpreted as how a transition stretches into or away from a syllable region. It is found that most transitions stretch up to one third but rarely exceed half of a syllable region. Post-tone transitions start much earlier in the syllables carrying entering tones (around last two fifth) than in the syllables carrying non-entering tones (around last one sixth).

Perceptual examination of amount of modified F0 contours also facilitates a comparison of objective evaluations with perceptual performance. The conventional measurements of RMS error and correlation between modified and non-modified F0 contours are found to be not reliable predictors for perceptual performance. How to objectively describe the perceptual difference of F0 contours is still a difficult problem. In terms of the two objective measurements themselves, they are indeed different.

Between the two, contour correlation is believed to be the better one, as far as the perceived difference in terms of speech naturalness is concerned.

In this chapter, the relation between human production and perception is touched again. The analysis of the attributes of linear F0 movements reveals the intent of human production to ease human perception. The acoustic variations belong to different linguistic categories are generated as uniquely as possible. On the other hand, the examination of objective evaluations with perceptual performance reveals the discrepancy that always exists between the two different systems.

Publications

Li Y.-J. and Lee Tan (2008). "A perceptual study of approximated Cantonese tone contours," *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP) 2008*.

Li Y.-J. and Lee Tan (2010). "Perception and analysis of linearly approximated F0 contours in Cantonese speech," *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP) 2010*, pp. 435-439.

Chapter 5

Perception-based Automatic Approximation of F0 Contours in Cantonese Speech

In the preceding chapters, it has been confirmed that F0 variations in Cantonese speech can be adequately represented by linear approximations of the observed contours, in the sense that comparable perception with the natural speech is maintained. A limitation in our investigations is that the approximated contours were derived manually. As a feasibility study, a manual process is not unreasonable. But for practical usages, such a process will be subject to the problems of low efficiency and inconsistency. Hence the application of the method would be very limited. A method of automatic approximation is needed to make the procedures standardized, consistent, efficient and reproduceable.

In this chapter, a framework of automatic linear approximation of the F0 contours in Cantonese speech will be described. The framework is developed based on the knowledge learned from the perceptual studies. The approximation process is carried out in three steps: contour smoothing, locating turning points and determining F0 values at the turning points. Perceptual evaluation is performed on the re-synthesized speech of hundreds of polysyllabic words. The results show that the proposed framework generates good approximations for observed F0 contours in a fully automatic manner. In most cases, the re-synthesized speech can reach comparable perception with the natural speech. The automatically approximated F0 contours attain similar preference level to the manually determined approximations.

5.1. Formulation of Automatic Approximation

Initially we try to design the process of automatic approximation by exactly following the procedures of manual approximation. However, this is found to be very difficult in reality. Manual approximation is complex in that the human listener makes reference to the knowledge in multiple dimensions, including the observed F0 contour, the speech waveform, the linguistic experiences and the perceptual feedbacks. To understand how these dimensions of knowledge contribute to the process is not an easy task by itself. Our efforts are then made on capturing perception-related knowledge from the manual procedures and other perceptual studies, and applying them properly on the design of the automatic process.

The automatic approximation process is formulated as the determination of a sequence of connected linear movements which capture the major trends of the observed contour, as illustrated in Figure 5.1. The number of movements can be estimated based on the number of syllables and the tone identities of individual syllables. The approximation would then be well fixed with properly arranged turning points and their F0 values. In manual approximations, turning points were found to reside at the instants when the movement trend changes, while the slope of an approximated movement depended on not only the observed contours but also the perceptual assessment.

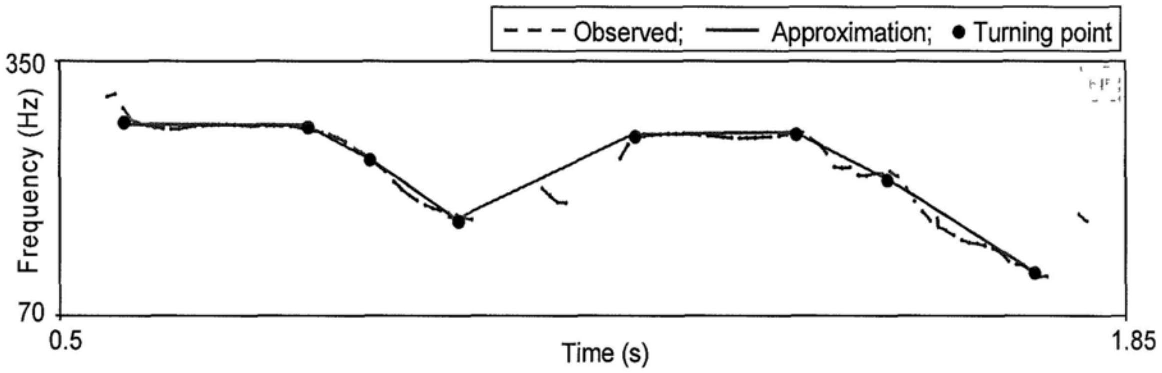


FIG. 5.1: Important elements in linear approximation of an observed F0 contour — connected linear movements, the number of movements, locations of turning points and F0 values at the turning points.

The procedures of automatic approximation aim mainly to solve two problems: (1) locating turning points on the contours; and (2) determining F0 values at the turning points using perception model. A contour smoothing process is needed at the beginning, so as to reduce the redundancy and noise in the raw data. The proposed framework is illustrated as in Figure 5.2. Each step will be described in detail. Evaluation of each step and the whole framework are carried out on the speech corpus of polysyllabic words that has been described in Chapter 4 (refer to Section 4.1.1.2 and Appendix 3).

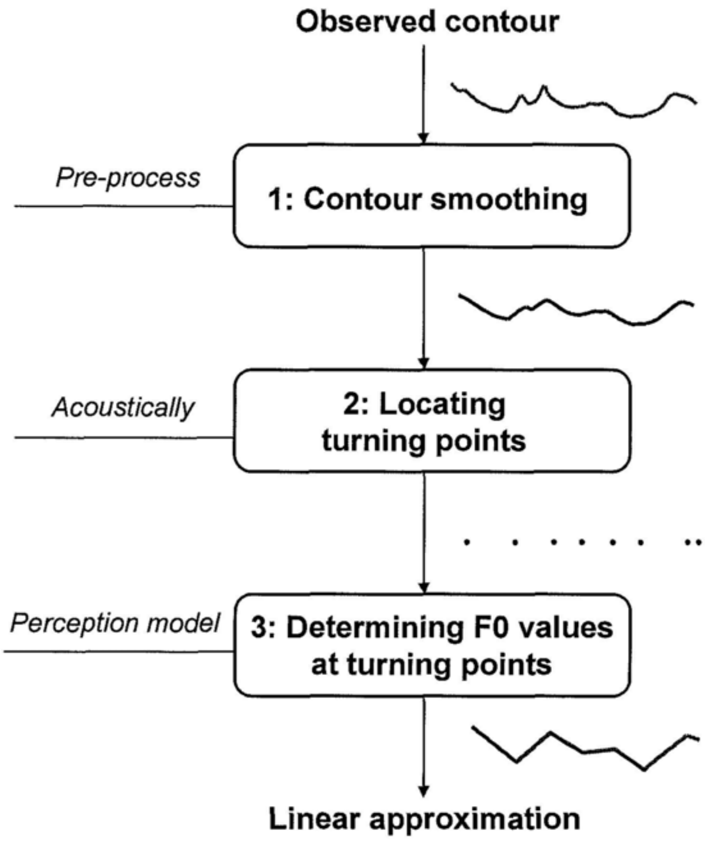


FIG. 5.2: Flow chart of the framework of automatic approximation.

5.2. Procedures of Automatic Approximation

5.2.1. Step 1: Contour smoothing

The observed F0 contours are extracted directly from the signal waveforms. They often contain many details that are not important to perception, but would cause problems in the contour approximation. Examples are the tiny variations as shown in Figure 5.3. These variations might be caused by muscle activities in human speech production system or be resulted from the errors made by the F0 estimation algorithm. In the search of perception-related F0 movements, such variations become a kind of “noise”.

Contour smoothing is therefore performed as a pre-processing step to eliminate these noises and to make the major moving trends of an F0 contour clearly identifiable.

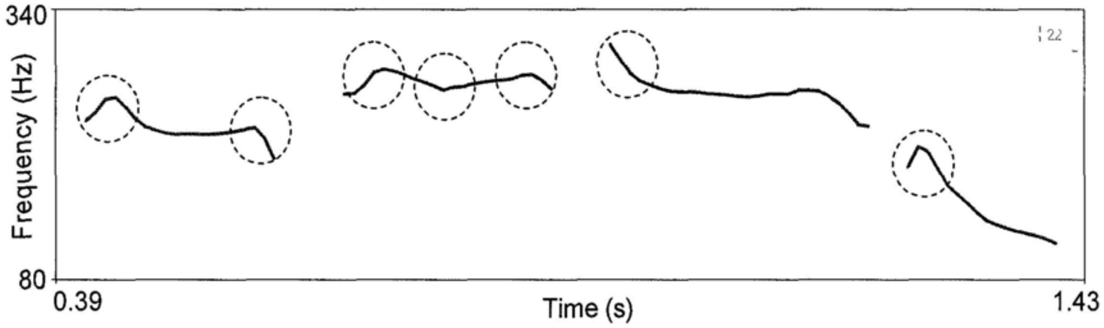


FIG. 5.3: Tiny variations (marked by dashed circles) in an observed F0 contour. These variations are not important to perception but may cause problems in automatic approximation.

Based on the experience from the manual approximations, the objectives of contour smoothing are set to be:

1) making the major moving trends of F0 contours prominent;

2) deleting irrelevant outlier points;

3) correcting wrongly estimated F0 values.

The extent of smoothing needs to be properly controlled as either over-smoothing or under-smoothing would be harmful to the subsequent steps of approximation. Two trade-offs are considered. One is between the standing-out of movement trends and retaining the contour variation; and the other one is between deleting outliers and keeping enough F0 points to represent a movement.

5.2.1.1 Method

Let the observed F0 contour of an utterance be denoted by $F0(n)$, where n is the time index in the unit of 10 ms. $F0(n)$ is considered as being composed of one or several *continuous segments*. A continuous segment (CS) is defined as a period of contour in which any two neighboring F0 points have a time distance smaller than 0.02 s. For example, there are four CS(s) in the utterance contour that is shown in Figure 5.3. Smoothing is performed within each CS.

First of all, tiny local fluctuations are flattened by a moving-window averaging. The window has a length of 3 (frames) and a shift step size of 1 (frame). As a result, the major trends of F0 change become more prominent.

Subsequently irrelevant outlier points on the contours are removed. A CS shorter than 0.04s is regarded as outliers. Some irrelevant points at the beginning or the end of a CS are also considered as outliers. These irrelevant points are typically located within the consonant-vowel transition regions. These points are identified by comparing with the calculated average F0 difference (AFD) between the neighboring points in a CS. AFD is given a minimum value of 3 Hz, in the case that the calculated value is smaller than 3 Hz. If the F0 difference between a CS beginning point and its neighbor is larger than AFD, the beginning point will be deleted. The outlier points at the end of a CS are identified and removed in a similar way.

Lastly, correction of wrongly estimated F0 values within a CS is performed. A wrongly estimated F0 point is detected when it has an F0 difference greater than AFD with one or both of its neighboring points. Then the value of this point is amended according to the values of its two neighbors. Deletion of outlier points and F0 correction are done iteratively for 4 times.

Figure 5.4 and Figure 5.5 give two examples of the smoothed contours. It is shown that the smoothing process well achieves all the objectives.

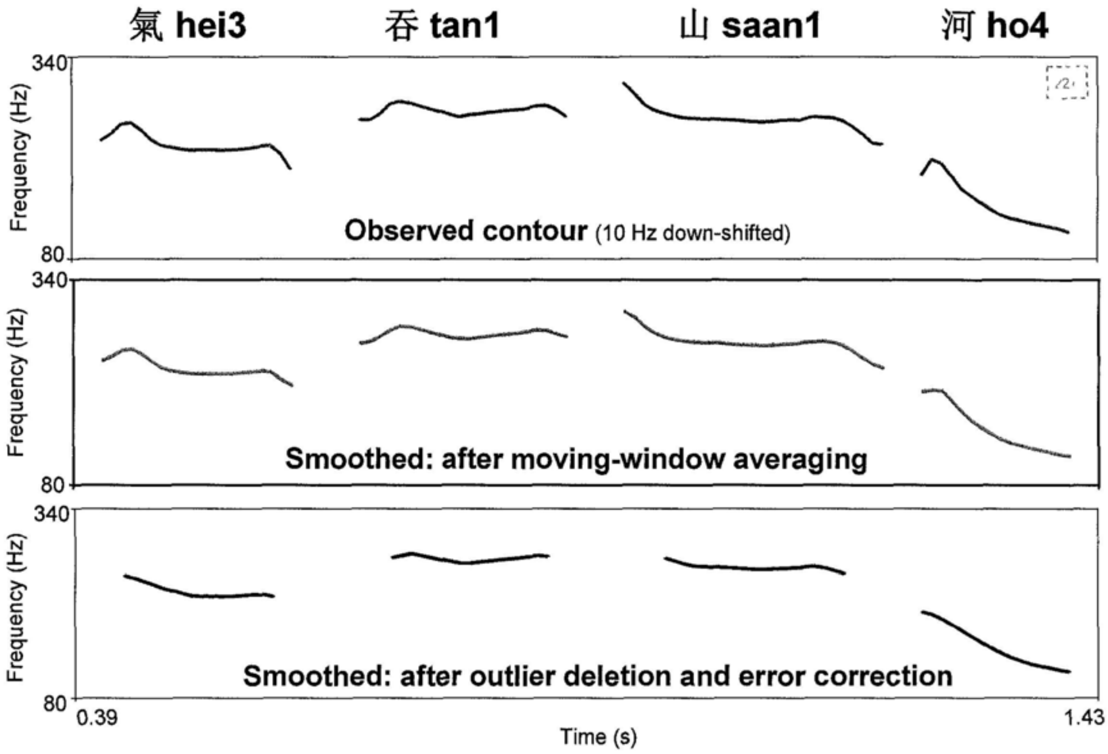


FIG. 5.4: An example of the smoothed contour.

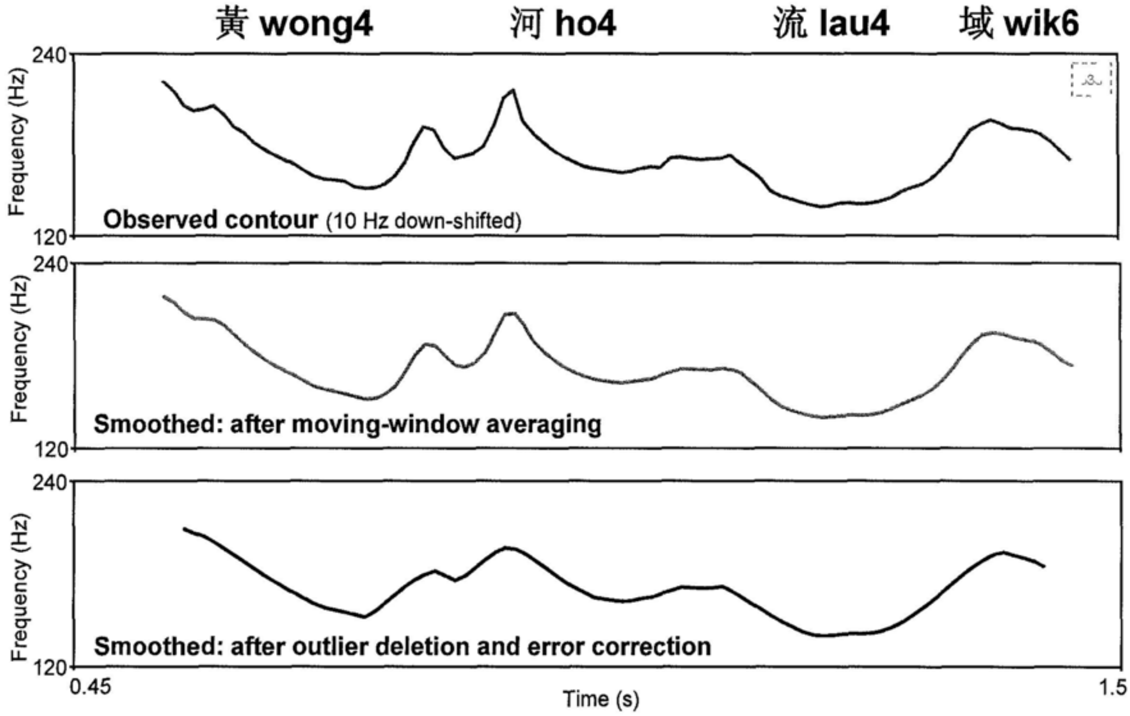


FIG. 5.5: An example of the smoothed contour.

5.2.1.2 Evaluations

To ensure that smoothing step would not cause serious error propagation, objective and perceptual evaluations were carried out using the evaluation data.

5.2.1.2.1 Objective evaluation

Among all of the F0 points in the observed F0 contours of 335 utterances, about 11.5% were identified as outliers and were removed. As shown in Table 5.1, after contour smoothing, the F0 differences between the neighboring F0 points are largely reduced, particularly obvious for its standard deviation. The reduction is contributed by all of the procedures implemented in the smoothing. The RMS error between the smoothed and the observed F0 contours[11] is 3.1 Hz on average.

Table 5.1. Comparison of F0 difference between the neighboring F0 points in the observed contours and the smoothed contours.

	Observed contours	**Smoothed contours**	**Reduction rate**
Mean	3.6 Hz	2.4 Hz	33.3%
Std.	5.44 Hz	2.16 Hz	60.3%

5.2.1.2.2 Perceptual evaluation

For all of the evaluated utterances, the smoothed contours were subjectively assessed in terms of perceptual distortion. The test method and procedure are similar with previous perceptual tests, except that the test was done in a smaller scale.

Six subjects, including three male and three female, participated in the listening test. None of them was involved in other perceptual tests of this study. They were recruited with the same requirements as in other tests.

For each utterance, the modified speech was the re-synthesized natural speech using the smoothed F0 contour. The F0 levels of both the observed and the smoothed

[11] Measurement is not carried out for removed outliers.

contours were down-shifted by 10 Hz. In the listening test, the modified speech was compared with the natural speech and the subject was asked to choose the better one.

Each subject went through half of the evaluation data. As a result, there were 3 responses collected for each word pair. According to the votes on "same" and "modified speech" among the 3 responses, a distribution of the 335 word pairs was calculated and is shown as in Figure 5.6. If no more than 1 of the 3 votes was given to "natural speech", the modified speech is said to attain comparable perception with the natural speech. As seen from Figure 5.6, 96% of the modified utterances could attain comparable perception. The preference distribution of all the collected responses is given in Table 5.2. The results confirm that the smoothing process does not introduce any side-effect on perception and the modified speech can attain comparable perception to the natural speech.

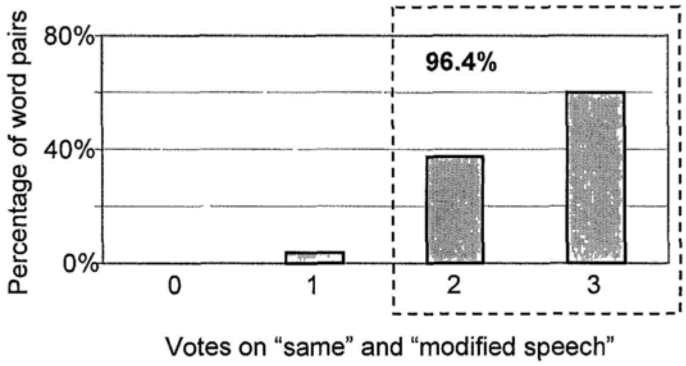


FIG. 5.6: Perceptual evaluation results on the smoothed contours — distribution of the 335 word pairs according to the obtained votes on "same" and "modified speech" for each word pair (upper limit is 3).

Table 5.2. Perceptual evaluation results on the smoothed contours — preference distribution of the total 1005 collected responses.

to "same"	to "modified speech"	to "natural speech"
72.7%	12.5%	14.8%

5.2.2. Step 2: Locating turning points

5.2.2.1 Method

The basic idea of locating turning points is to search N_t global contour peaks and valleys under a set of constraints, where N_t is the number of required turning points. N_t should be double that of the required linear movements; while the number of movements can be fixed, given the number of syllables and the tone identities.

5.2.2.1.1 Basic searching algorithm

This basic searching algorithm is applied to find a single peak/valley over a contour. The process is explained by Figure 5.7. Consider an F0 contour $F0(n)$, which is composed of N points (t_n, f_n), where $n \in [1, N]$. To find the global valley V, a reference line L is firstly defined to pass through the two contour terminals (t_1, f_1) and (t_N, f_N),

$$L:\quad at + bf + c = 0 \tag{5.1a}$$

Then the vertical distance (F0 offset) from each contour point (t_n, f_n) to L is computed by,

$$D_n = \frac{|at_n + bf_n + c|}{\sqrt{a^2 + b^2}} \tag{5.1b}$$

The global valley V is the point with the largest vertical distance to L,

$$V = \arg\max_n (D_n) \tag{5.1c}$$

In this algorithm, under a reference line, only one peak/valley will be found[12].

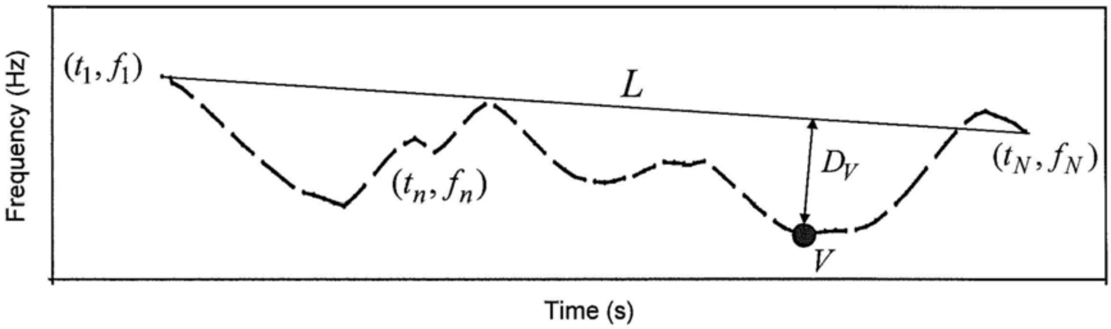


FIG. 5.7: Explanation of the basic searching algorithm — searching a single peak/valley over an F0 contour.

[12] For a shown F0 contour, scale division of time axis is given by every 0.01 s and that for frequency is given by every 1 Hz. To represent the observed distances, in calculation, time and frequency have to be in the same order of magnitude. Therefore, values of t in the equations from (5.1a) to (5.1c) are 100 times of the true values.

5.2.2.1.2 Global search

Global search is performed to find a certain number of global peaks and valleys over an F0 contour, by applying the basic searching algorithm. These peaks and valleys are taken as the candidates for turning points. Explanation of the global search is given in Figure 5.8. The search is formulated as an iterative process to find a new candidate between any two neighboring existing candidates. In the initial round of search, the two terminals (t_1, f_1) and (t_N, f_N) of the entire contour are used as the initial candidates. They are denoted as the beginning candidate C_B and the end candidate C_E respectively. The initial search defines a reference line L_1 passing through C_B and C_E. By applying the basic searching algorithm, the candidate C_1 is found. The next round of search checks the contour segments from C_B to C_1 and from C_1 to C_E, and locates the candidates C_2 and C_3. Subsequently a new round of search is performed from C_B to C_E, examining each contour segment between a pair of neighboring candidates.

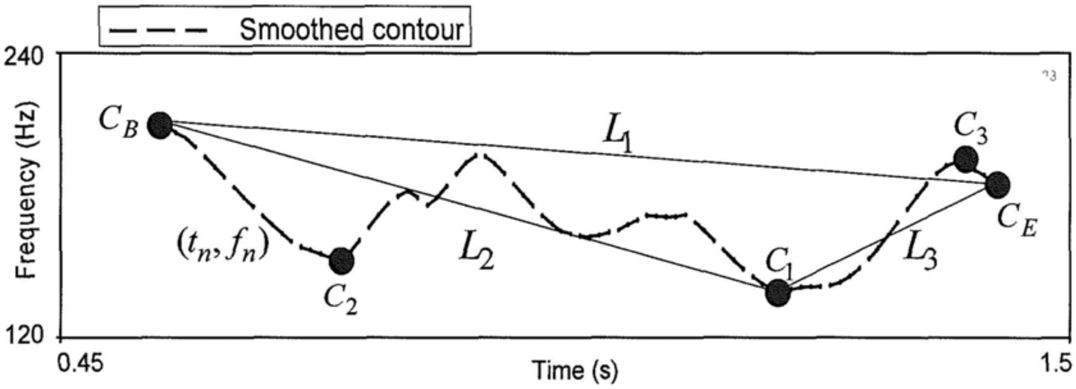


FIG. 5.8: Explanation of the global search — searching candidates of turning points on the smoothed F0 contour of an utterance.

In the above process, if a candidate satisfies a distance constraint and a time constraint, it will be regarded as a located turning point (LT). The distance constraint states that the distance from the candidate to the corresponding reference line must be greater than 8. The time constraint imposes a minimum time interval between the candidate and its neighboring candidates, which is computed as,

$$(t_N - t_1)/((N_t - 1) \times 3) \qquad (5.2)$$

This minimum time interval will be given a value of 0.03 s, when the result of Equation 5.2 is smaller than 0.03 s. C_B or C_E will be replaced if there is a candidate satisfying the distance constraint and meanwhile having ≤ 0.04 s time difference with it.

The search continues iteratively until the number of located turning points N_{LT} is greater than $N_t + 2$, or the total number of search rounds is greater than 8. After the global search, there would be over-found turning points ($N_{LT} > N_t$).

5.2.2.1.3 Refinement in continuous segments

After the global search, the LT(s) in each CS are further refined. The boundaries of a CS usually correspond to the boundaries of one or several tones. This enables an estimation of the syllable number in the CS (N_{CS_syl}) and accordingly leads to an estimated number of turning points in the CS (N_{CS_LT}). The estimation of N_{CS_syl} and N_{CS_LT} is explained by the pseudo code given in Table 5.3.

Table 5.3. Pseudo code for the estimation of N_{CS_syl} and N_{CS_LT}.

```
Define $N_{syl}$ - the syllable number in the utterance.
Define $D_{CS}$ - duration of a CS.
Define $D_{syl}$ - the estimated syllable duration within the utterance.
$D_{syl} = (\sum D_{CS}) / N_{syl}$.
$N_{CS\_syl}$ = integer rounding of $(D_{CS} / D_{syl})$.
$N_{CS\_LT} = 2 \times N_{CS\_syl}$.
```

The refinement process adjusts the number of LT(s) in the CS to N_{CS_LT} via deletion and insertion. Firstly, LT(s) at the beginning and the end of the CS are refined according to a more stringent time constraint which requires a longer than $D_{syl} / 5$ time difference between such an LT and its neighboring LT (in the same CS). The ones that do not satisfy the constraint are deleted. Then if the number of LT(s) is still larger than N_{CS_LT}, the LT(s) that are closest to the reference line are deleted. If the number of LT(s) is smaller than N_{CS_LT}, new LT would be searched between the pair of neighboring LT(s) that are most distant from each other. The refinement guarantees an even number of LT(s) in each CS and tries to make N_{LT} close to N_t.

After the refinement, N_{LT} will be greatly reduced compared with that obtained in the global search, but may be still slightly larger than N_t by an even number, such as two or four over-found turning points. Since the turning points are located globally, ensuring the perceptual performance by using the exact number of N_t located turning points is a difficult task. Considering perceptual performance as the most important, process of locating turning points allows slight over-approximation (with extra movements), while at the same time tries to make N_{LT} equal to N_t . Serious over-approximation has to be avoided; otherwise the approximation will be far away from its original intention of keeping only perception-necessary F0 variations.

5.2.2.2 Evaluation

The method of locating turning points was tested with the evaluation data. The turning points obtained in the manual approximations (Chapter 4) are used as the reference. Objective evaluation was performed by calculating the matching rate between the automatically located turning points and the manually marked ones. If an automatically located turning point is away from a manually marked one within 35 ms, it is considered a case of successful matching. For the first and the last located turning points in a CS, the time error tolerance is 45 ms. Previous perceptual studies have suggested that in the shifting of an F0 peak or valley in a contour, a timing difference of 50 ms is the threshold to change the perceptual interpretation of the contour (House, 2004).

Matching results are given as in Table 5.4. Each located turning point was matched only once. The results confirm that the turning points can be reliably located, with a matching rate of 87.5%. Figure 5.9 and 5.10 give the two examples that represent the best and the worst cases respectively. In the two figures, each turning point is seen as a dark dot. Manually marked turning points are shown on the manual approximations of the observed F0 contours. Automatically located turning points are shown below the smoothed contours. If an automatically located turning point matches the manually marked one, a "√ " sign is put under it.

Table 5.4. Result summary of the objective evaluation on the automatically located turning points (measurement of the matching rate between automatically located turning points and the manually marked ones).

No. of needed turning points (N_t)	2718 (8 1/utterance)
No. of located turning points (N_{LT})	2752 ($N_{LT} = 1.01\, N_t$) (8 2/utterance)
Matching rate	87.5%
Rate of full matched utterances (Matching rate=100%)	32.5% (109/335)
Rate of low matched utterances (Matching rate$\leq$60%)	2% (4/335)

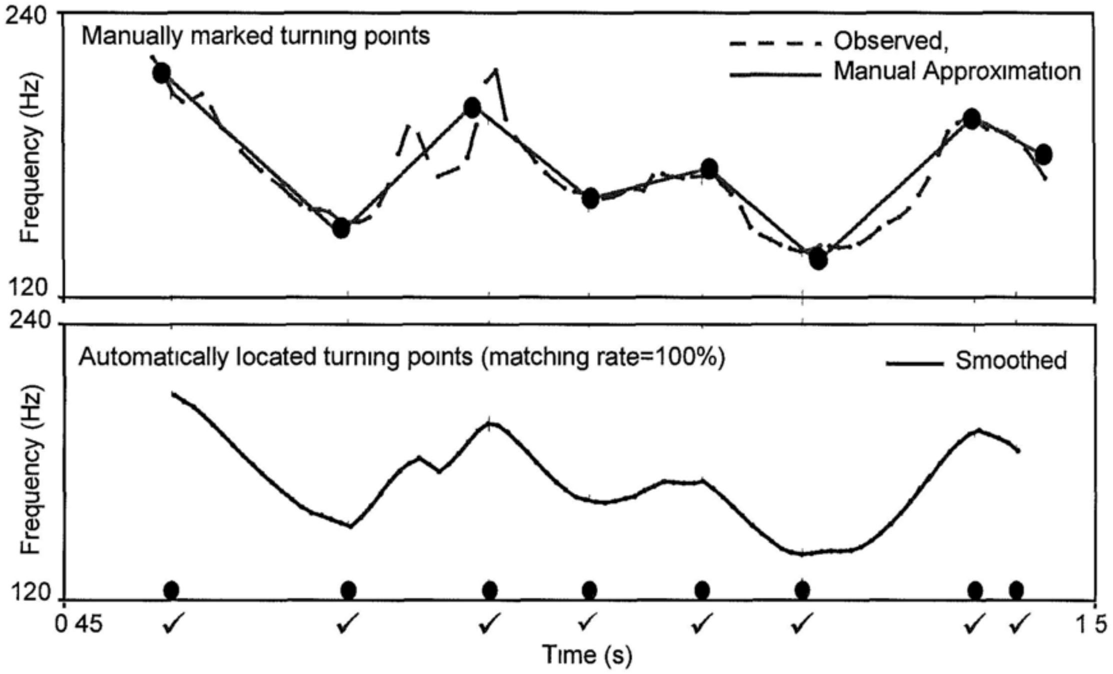


FIG. 5.9: An example of the cases with the best matching rate between automatically located turning points and manually marked ones.

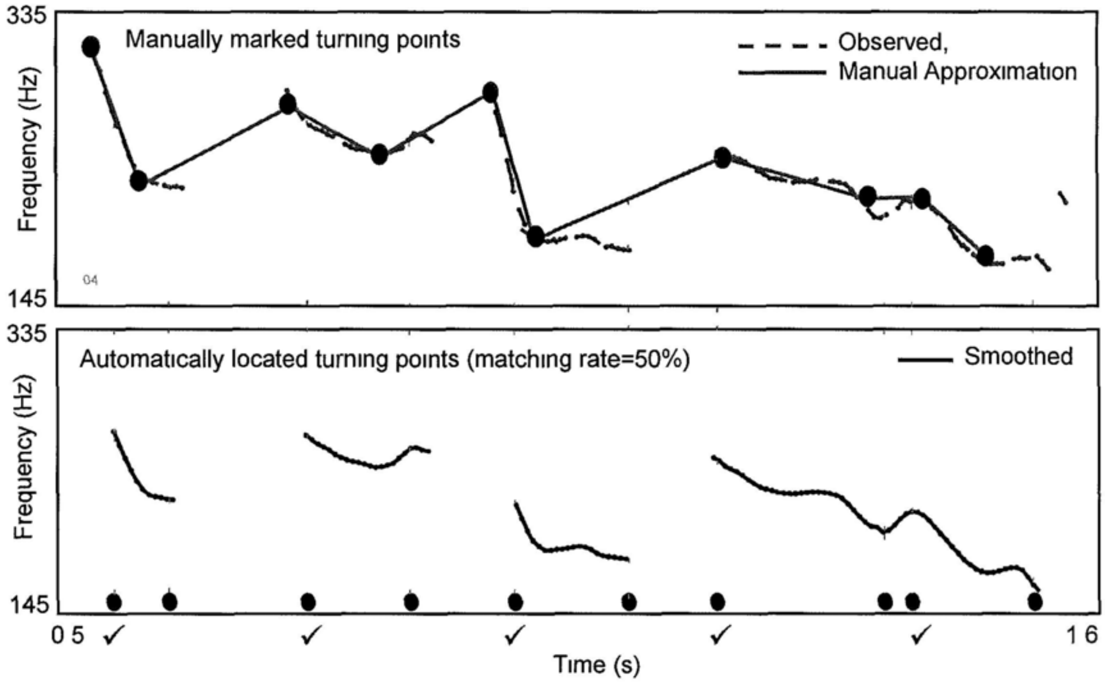


FIG. 5.10: An example of the cases with the worst matching rate between automatically located turning points and manually marked ones.

Matching rate with manually marked turning points is only a preliminary evaluation of this step. More reliable evaluation can be obtained by a perceptual examination on the final resulted approximations. Since the linear approximation of an observed contour leading to a comparable perception is not unique, manually marked turning points can not be guaranteed to be the best ones for approximations and they should be only considered as references. It is expected that an automatic approximation with a high matching rate of located turning points can be a good approximation. Nevertheless, an automatic approximation with a low matching rate is not necessarily associated with a bad approximation.

5.2.3. Step 3: Determining F0 values at turning points

5.2.3.1 Method

In this step, the question being asked is how to transform an observed contour segment between two turning points into a linear contour such that they are perceived to be similar. As shown in Figure 5.11, this similarity can be interpreted as that the two contours have the same perceived beginning pitches and the same perceived end pitches.

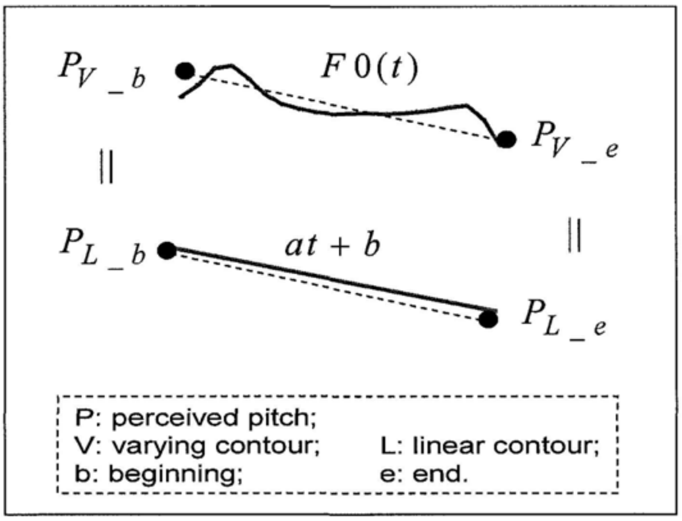


FIG. 5.11: Explanation of the similarity in perceiving two different contours — the two contours have the same perceived beginning pitches and the same perceived end pitches.

Perceived pitch of a segment of F0 contour was investigated in (d'Alessandro and Castellengo, 1994; d'Alessandro, Rosset and Rossi, 1998). Given an F0 contour

$F0(t)$ between t_1 and t_2, the perceived pitch level can be predicted using the weighted time average (WTA) model[13],

$$p = \frac{\int_{t_1}^{t_2} (e^{\alpha(\tau - t_2)} + \beta) F0(\tau) d\tau}{\int_{t_1}^{t_2} (e^{\alpha(\tau - t_2)} + \beta) d\tau} \tag{5.3}$$

This model can be understood based on a time averaging function,

$$p = \frac{\int_{t_1}^{t_2} F0(\tau) d\tau}{\int_{t_1}^{t_2} d\tau} \tag{5.4}$$

The WTA model raises an exponential window in the averaging period. α is the factor of the exponent in the window. It controls the weight to the final part of the contour. β is the height of the window. It controls the weight given to the whole contour.

In the case that the contour is perceived as a changing movement, WTA model is simplified with $\beta = 0$ as,

$$p = \frac{\int_{t_1}^{t_2} e^{\alpha(\tau - t_2)} F0(\tau) d\tau}{\int_{t_1}^{t_2} e^{\alpha(\tau - t_2)} d\tau} \tag{5.5}$$

The simplified version is used to predict the perceived beginning pitch and end pitch of the contour, by changing the interval of integration to $[0, \delta d]$ and $[d - \delta d, d]$ respectively, where d is the contour duration and δ is the integration interval as a

[13] Development of the WTA model is given in Appendix 4.

proportion of d. The values of α and δ were optimized in (d'Alessandro, Rosset and Rossi, 1998) for four different cases: perceived beginning and end pitches of a rising movement (denoted as R_b and R_e respectively), and perceived beginning and end pitches of a falling movement (denoted as F_b and F_e respectively). The optimal parameters were obtained by matching a large number of subjectively determined beginning and end pitch values for isolated tone contours. Table 5.5 shows the optimal parameter values and the corresponding matching rate between predicted and subjectively determined pitch values. We will adopt these values directly in the following procedures.

Table 5.5. Optimized parameters of WTA model for isolated tones, by matching amount of subjectively determined beginning and end pitches (d'Alessandro, Rosset and Rossi, 1998).

	α	δ	Matching rate (%)
R_b	-7.5	0.82	100
R_e	24.7	0.26	85
F_b	-19.4	0.56	90
F_e	13.5	0.75	100

In our problem, given an observed F0 contour $F0(t)$ ($t \in (0,d)$) between a pair of turning points, and the estimation of its changing trend, i.e. rising (when $F0(d) - F0(0) \geq 0$) or falling (when $F0(d) - F0(0) < 0$), the perceived beginning and end pitch values can be predicted by the WTA model. The desired linear approximation is expressed as $(at+b)$. Its perceived beginning and end pitch values can be also described by the WTA model, with two unknown variables a and b. To be perceived similarly, the two contours are considered to have the same perceived beginning pitch value and the same perceived end pitch value. Accordingly, we can set up the following linear equations with two unknown variables,

$$\begin{cases} p_b = \dfrac{\int_0^{\delta d} e^{\alpha(\tau-\delta d)} F0(\tau)d\tau}{\int_0^{\delta d} e^{\alpha(\tau-\delta d)} d\tau} = \dfrac{\int_0^{\delta d} e^{\alpha(\tau-\delta d)} (a\tau+b)d\tau}{\int_0^{\delta d} e^{\alpha(\tau-\delta d)} d\tau} \\ p_e = \dfrac{\int_{d-\delta d}^{d} e^{\alpha(\tau-d)} F0(\tau)d\tau}{\int_{d-\delta d}^{d} e^{\alpha(\tau-d)} d\tau} = \dfrac{\int_{d-\delta d}^{d} e^{\alpha(\tau-d)} (a\tau+b)d\tau}{\int_{d-\delta d}^{d} e^{\alpha(\tau-d)} d\tau} \end{cases} \tag{5.6}$$

In these equations a and b are solvable, and hence a linear F0 contour can be obtained. The first and the last points on this linear contour are where the two turning points should be. If the slope of a resulted linear movement is smaller than the glissando threshold (t' Hart, Collier and Cohen, 1990), the movement is further simplified as a level movement with the height equal to the median of the linear movement.

In continuous speech, given the case that all the tone identities are non-rising and all the turning points are ideally located, an F0 contour would be segmented into interlaced tone contours and transition contours, as shown in Fig. 5.12. Our previous study (in Chapter 3) suggested that perception is more sensitive to tone contours than tone transitions. Therefore, every two turning points derive their F0 values by using the in between tone contour. For examples, F0 values of the first two turning points in Figure 5.12 are predicted from the first tone contour; F0 values of the third and the fourth turning points are derived from the second tone contour. As a result, F0 values of all the turning points can be determined. The approximated contour is obtained by line-connection of these turning points.

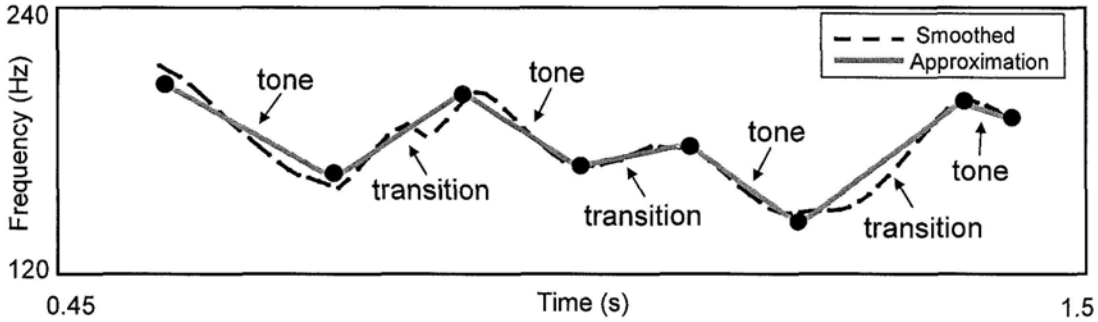


FIG. 5.12: In continuous speech, determining F0 values of every two turning points by using the in between tone contours.

5.2.3.2 Evaluation

An objective evaluation was carried out. It compared the corresponding movements in two kinds of approximations. One is manual approximation and the other one is partial automatic approximation using manually marked turning points and automatically determined F0 values. The average F0 difference at the corresponding turning points is 4.2 Hz. This suggests that at turning points a certain difference exists between predicted and manually determined F0 values. On the other hand, measurement of the slope ratios between the corresponding movements gives an average ratio of 1.4, lower than 2 that is the perceptual threshold to differentiate two movements (t' Hart, Collier and Cohen, 1990). It indicates that there would be no perceptual difference between the corresponding movements in the two kinds of approximations.

Figure 5.13 gives a comparison of the manual approximation and the partial automatic approximation. In the two sub-figures, locations of turning points are derived from manually marked turning points. In the lower figure, F0 values at turning points are automatically determined and they are quite close to the values in the manual approximation.

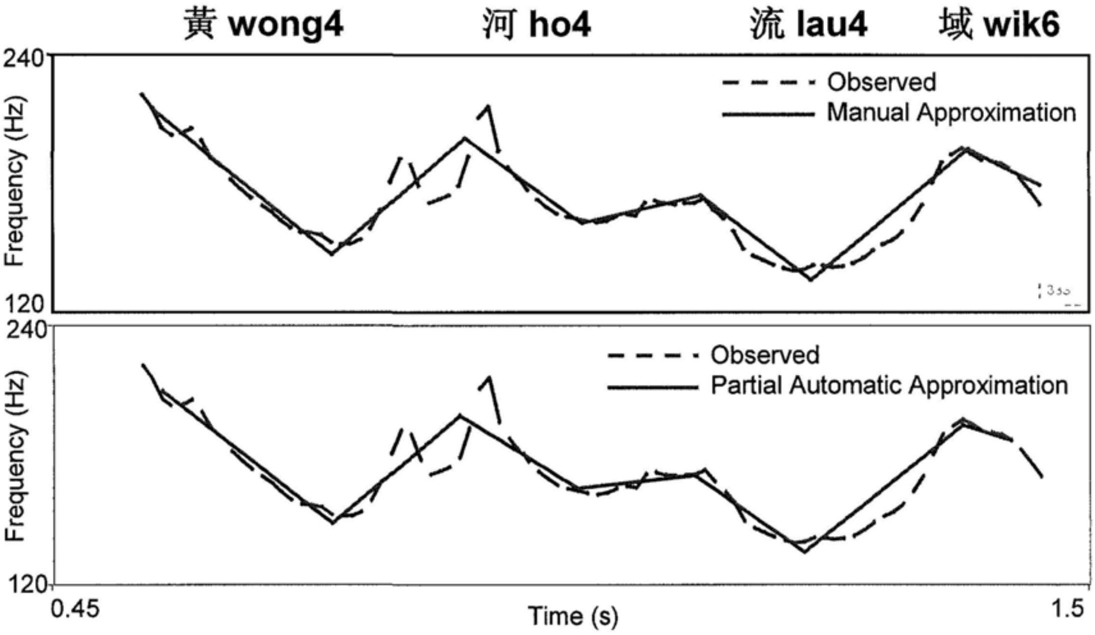


FIG. 5.13: Comparison of the manual approximation and the partial automatic approximation. The partial automatic approximation is derived from manually marked turning points and automatically determined F0 values.

5.3. Perceptual Evaluation of Automatic Approximation

The proposed framework of automatic approximation was perceptually examined using the evaluation data. The test procedures are similar to those in the previous tests.

5.3.1. Method

Twenty new and gender-balanced subjects participated in the perceptual test. Subject recruitment followed the same requirements as in previous tests.

The F0 contour of each natural utterance in the evaluation data was automatically approximated and a modified speech utterance was generated to follow the approximated contour. The F0 levels of both the observed and the approximated contours were down-shifted by 10 Hz.

In the test, the modified speech and the natural speech were presented in pair to the subjects through the same interactive computer interface as used in the previous tests. The subjects were asked to compare the two utterances and to select the preferred one in terms of overall impression. Each of the 20 subjects went through half of the words in the evaluation data.

5.3.2. Results

10 responses were collected for each pair of utterances. If not more than half of them vote "natural speech", the modified speech is said to attain *comparable perception* with the natural speech. About 93% of the modified utterances attain comparable perception. A distribution of word pairs according to the total votes on "same" and "modified speech" for each pair is given in Figure 5.14.

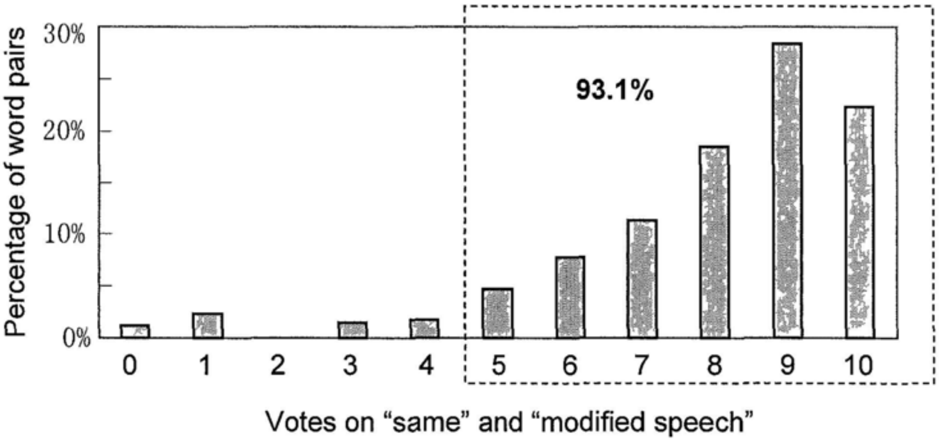


FIG. 5.14: Perceptual evaluation results on automatic approximation — distribution of the 335 word pairs according to the obtained votes on "same" and "modified speech" for each word pair (upper limit is 10).

Figure 5.15 and Figure 5.16 give two examples that represent, respectively, the best and the worst cases of the automatically approximated F0 contours in the evaluation data, in terms of perceptual test results. The examples suggest that the step of determining F0 values generally would not create serious errors, while the error of located turning points may greatly influence the perception of the approximated contours. Another indication is that, in some cases, automatic approximation may be better than manual approximation.

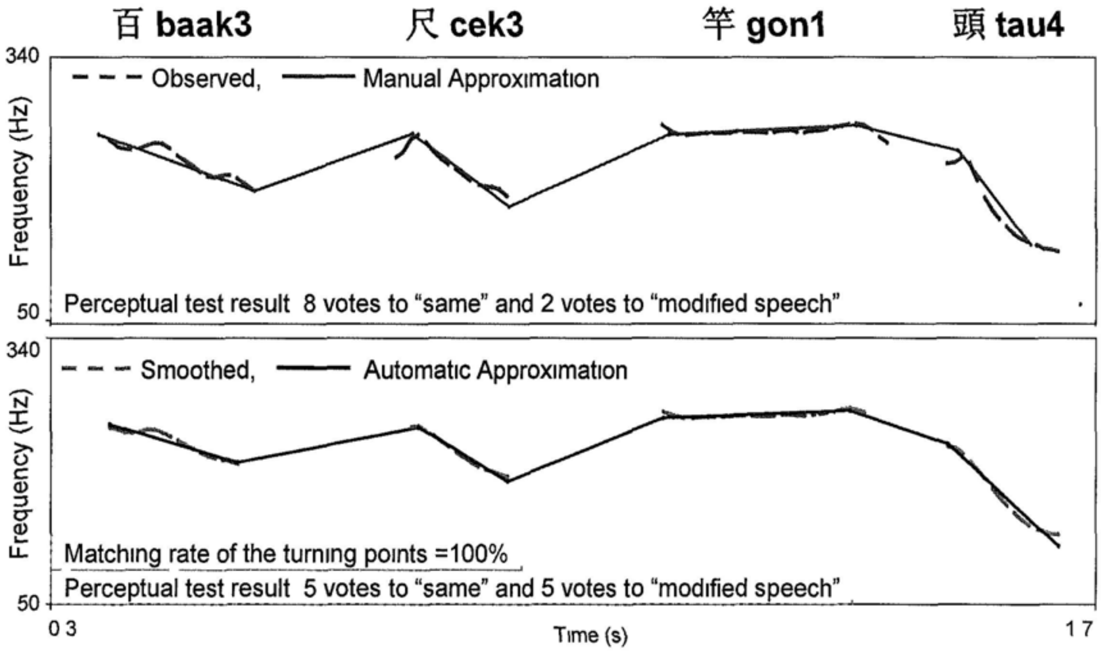


FIG. 5.15: An example of the best cases of automatically approximated F0 contours in the evaluation data, in terms of perceptual test results. "Matching rate of the turning points" indicates the matching rate between the automatically located turning points and the manually marked ones.

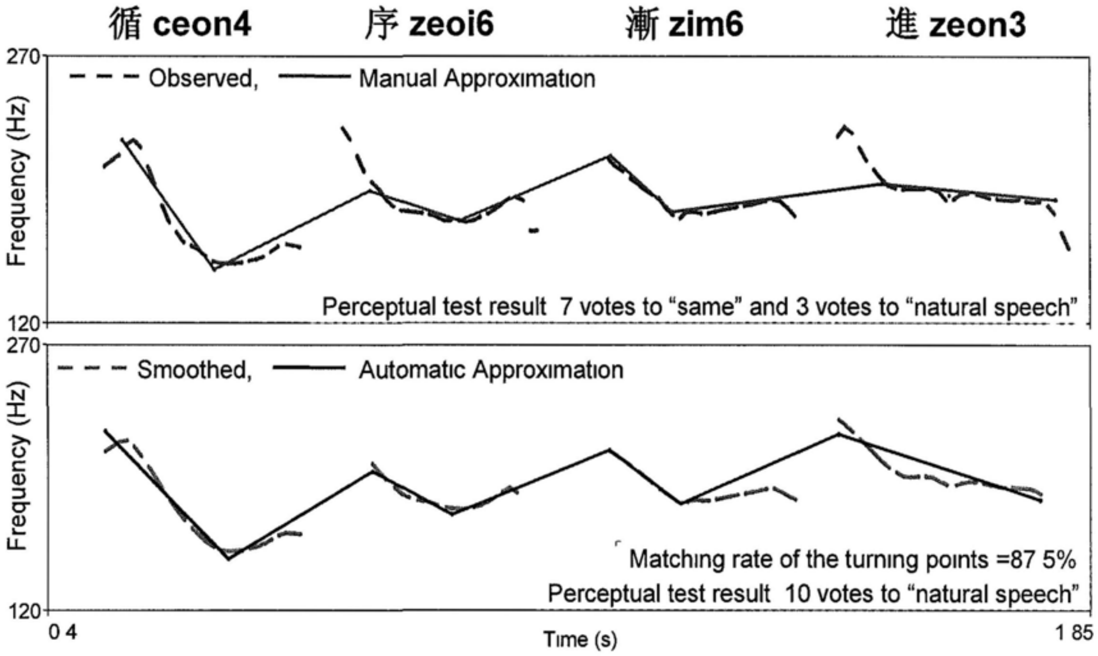


FIG. 5.16: An example of the worst cases of automatically approximated F0 contours in the evaluation data, in terms of perceptual test results. "Matching rate of the turning points" indicates the matching rate between the automatically located turning points and the manually marked ones.

Table 5.6 gives a comparison of the perceptual test results between automatic approximation and manual approximation. Not surprisingly, the degree of preferences on the automatic approximations is lower than that on the manual approximations, but the difference is not significant. We believe that the proposed process can generate good approximations for the observed F0 contours in terms of reaching comparable perception with natural speech.

Table 5.6. Comparison of the perceptual test results on the automatically approximated and the manually approximated F0 contours.

Approx.	Comparably perceived word pairs	Distribution of total 3350 responses		
		to "same"	to "modified speech"	to "natural speech"
Automatic	93.1%	71.9%	7.2%	20.9%
Manual	97.3%	76.8%	8.5%	14.7%

5.4. Discussion

The proposed process is efficient with low computational complexity. All of the procedures operate on the observed F0 contours directly. The only required prior knowledge is the number of syllables and the tone identity of each syllable, which are available from a syllable-level transcription of the utterance. In particular, since the turning points are globally located, prior knowledge about syllable boundaries is not necessary, unlike in other approximation algorithm for intonation contours (d'Alessandro and Mertens, 1995) or in our manual approximation.

Among the three steps, locating turning points incorporates most of the language-specific constraints and hence it is the most difficult and important. Although the information about syllable boundaries is not used in this step, we believe that such information is useful. It is expected that the process could be easier and the performance could be improved, when syllable boundaries are given. In our study, the evaluation data contain only non-rising tones. For approximating F0 contours of rising tones, adjustments have to be made on the step of locating turning points, so as to include extra constraints. Step of determining F0 values utilizes most perceptual knowledge that has been systematically described in previous psychoacoustic studies (t' Hart, Collier and Cohen, 1990; d'Alessandro, Rosset and Rossi, 1998). In this thesis, the incorporation of the knowledge to derive a linear movement is considered as an innovative effort. The step of contour smoothing is very important in the whole framework. The process of locating turning points is indeed very sensitive to noise and the determination of F0 values is influenced by noise as well. If the result of smoothing is unsatisfied, it would be very difficult for the framework to reach the current performance.

In this study, the framework is evaluated using word-level utterances. Extension of the framework to sentence-level utterances should be easy, since procedures of processing word-level and sentence-level utterances are similar in nature, as indicated from the manual approximations.

Though the approximation framework is developed with Cantonese speech data, it can be generalized to other languages provided that their perception-necessary F0 variations can be also adequately represented by linear approximations. The steps of contour smoothing and determining F0 values would be rather language-independent and the step of locating turning points would involve most language-specific constraints.

The framework is desired to be associated with a graphic user interface so that the result of each step can be visualized and adjustable. Such an interface can greatly increase the potentials of the linear approximations and the automatic approximations in practical usages.

5.5. Conclusions

In this chapter, a perception-oriented framework of automatic approximation of the F0 contours in continuous Cantonese speech has been developed. It aims at establishing a simplified representation of the highly varying acoustic observations without causing noticeable perceptual differences. The approximation process is designed and implemented based on our basic study on the manually approximated F0 contours. It consists of three steps: contour smoothing, locating turning points and determining F0 values at turning points. Each of the steps was examined individually. The complete framework was evaluated through a perceptual test, in which the modified utterances following automatically approximated F0 contours were compared with the natural speech. The results show that the modified speech can reach comparable perception with the natural speech. The automatically approximated F0 contours can attain similar preference level to the manually determined approximations.

Publication

Li Y.-J. and Lee Tan (2010). "Perception-based automatic approximation of F0 contours in Cantonese speech," *Proc. Interspeech 2010*, pp. 1425-1428.

Chapter 6

Conclusions

Speech production and speech perception are the most important component processes in speech communication. Their relation is complex and is difficult to be interpreted in a systematic way. Nevertheless, such knowledge is essential in theoretical research on speech communication and also for application development of speech technologies. This study is motivated by our strong desire to have a better understanding on the production-perception relation. Being considered as the connection between the two processes, acoustic signal is controlled and investigated. F0 variation is particularly concerned in this thesis.

F0 in speech carries both linguistic and paralinguistic information. Therefore, its impacts on speech communication are most interesting to people in this area. Despite extensive efforts on studying F0-related features, we found that the knowledge about F0 variation in relation to the perception of speech naturalness is quite limited. Furthermore, the problem of how speech production and speech perception are connected with F0 has not been investigated to a good depth.

F0 contours measured from human speech (the observed contours) vary to a considerable extent. We aim to identify perceptually critical variations from the varying F0 contours. A key problem addressed in this thesis is to search for the simplest acoustic representation of F0 contours in speech, with the requirement that the perceived speech quality is not affected. In particular, F0 contours in Cantonese speech are focused.

6.1. Linear Approximation

To tackle the above problem, approximations of the observed F0 contours in Cantonese speech were investigated. The modified speech utterances following approximated contours were compared with the natural utterances in a series of perceptual tests. With extensive and consistent test results, we argue that linear approximations of the observed F0 contours can adequately describe perception-sensitive F0 variations in Cantonese speech. The linear representations are dictated largely by the target pitch patterns of lexical tones. Specifically, the F0 contour of a rising tone requires two linear movements and that of a non-rising tone can be represented by a single movement. For the transition contour between a pair of co-articulated tones, one linear movement is appropriate and sufficient. These conclusions are drawn from the approximated contours at different levels, i.e., isolated syllables, words and sentences; from the speech carriers with rich segmental and contextual variations; and from the speech utterances of different speakers. Comparable perception between the modified speech and the natural speech were consistently observed. We consider these conclusions generally applicable to the F0 contours in Cantonese speech.

Our finding relates linear F0 movements with natural speech in tonal languages. Meanwhile, the fact that linear F0 movements are adequate for maintaining perceptual quality of speech provides additional evidences to support previous studies on tone-identification related acoustic cues, all of which were on the basis of linear F0 movements. In other words, all the communicative functions of F0 variations in speech can be hereafter consistently interpreted by the properties of linear movements.

6.2. Analysis of Linear Approximations

We further analyzed a large amount of generated and perceptually confirmed linear approximations. These approximations, on the one hand, resemble human produced acoustic signals; and on the other hand, are highly related with human perception. Analyses of them can reveal the connections between speech production and speech perception. Three analyses were carried out. They are summarized as below.

1) In Chapter 3, we examined the movement slopes in the approximated F0 contours of isolated syllables, in connection with perceptual test results. Results lead to a high correlation between contour slopes and perception of Cantonese speech, in terms of speech naturalness. The analysis also suggested a possible lifted perceptual tolerance on F0 variations in speech, as compared with those found in the psychoacoustic studies.

2) In Chapter 4, we performed a statistical analysis for hundreds of linear approximations of polysyllabic words. Each of the determining attributes of linear movements, i.e., movement slope, movement height and time location of a turning point was examined individually. Tonal movement slopes and heights among different tone categories were found to change in quite independent ranges. The intent of human production to ease human perception is clearly demonstrated. The investigations on the F0 changing rate (movement slopes) and the position of tone transition in relation to the syllable position (time location of turning points), that

are facilitated by the linear approximations, are particularly informative. These two features were rarely examined on the human generated F0 contours in previous studies, due to the difficulty in precisely measuring them.

3) In Chapter 4, we carried out a comparison of the objective evaluations of the modified F0 contours with their perceptual evaluation, with the help of perceptual test results on a large number of approximated contours. Two commonly-used objective measurements, RMS error and correlation between modified and non-modified F0 contours were examined. We found that neither of them is really reliable at predicting perceptual performance. Nevertheless, the discrepancy is considered to be useful in search of perception-critical F0 variations.

In this study, we established a set of linearly approximated F0 contours of polysyllabic words with guaranteed perceptual performance. The purpose is to learn F0 variations that are related to both human production and perception. Our current analyses on this data set are limited. More in-depth explorations are expected.

6.3. Framework of Automatic Approximation

Linear approximations were manually generated at the first stage of this study. In Chapter 5, an automatic process to generate approximations for the F0 contours in continuous Cantonese speech in a standardized, consistent and efficient way was investigated. A perception-oriented framework of automatic approximation was formulated based on the process of manual approximations and the implementation of other perceptual findings. Initial test on the polysyllabic words gave promising results. The generated approximations were capable of reaching comparable perception with the natural speech. The automatically approximated contours could attain a similar preference level to those manually obtained. The framework holds many advantages. It works on observed F0 contours directly with little prior knowledge as well as low computational cost. It is also extendable to other languages.

6.4. Testing Interface

Perceptual test plays an important role in this study. All the tests were conducted with our self-developed interactive computer interface. We consider this interface as additional contribution of this study. It is suitable for other similar perceptual tests which are based on pair comparison testing.

6.5. Summary

In this thesis, we dealt with a basic question in speech communication: what is the simplest acoustic representation of F0 variation, from speech perception points of view? To our knowledge, this is the first effort on this problem in tonal languages. The finding of linear approximations greatly simplifies the way of studying F0 variations in association with perceptual process, interpreting F0 variations with linguistic events and applying F0 variations in speech applications. We learned the properties of linear F0 movements in Cantonese speech through analyses. Results are very informative. The application potential of linear approximation is supported by the developed automatic approximation framework. With the help of this framework, we expect linear approximation to be beneficial to the research of speech prosody in general.

Chapter 7

Future Work

This study was motivated by our lack of understanding about the relation between speech production and speech perception. It addressed a basic problem in speech communication, and the results help improve our understanding on the perception of F0 in speech. As a basic study, our work is expected to be useful in general to the research of speech prosody. In this chapter, some suggestions on the follow-up works are made.

7.1. Linear Approximation

7.1.1. Analysis of linear approximations of rising tones

In Chapter 3, a preliminary study on the linear approximations for Cantonese rising tones is described. It is found that each rising tone can be described by two linear movements. The study also suggests that the movement slopes are important in perception of Cantonese rising tones. However, the used speech carriers only involve a small set of isolated syllables which are not adequate for stating a general conclusion. On the other hand, the current data size is too small for a statistical learning of the linear approximations for rising tones. A similar investigation as that carried out in Chapter 4 for non-rising tones is expected.

7.1.2. Evaluation of linearly approximated F0 contours using synthetic speech as carriers

In Chapter 3 and Chapter 4, the perceptual tests show some cases that the modified speech is more preferred than the natural speech. Such cases are supported by a view of selective perception, but still cannot be experimentally proved. Whether simplified F0 contours are more perceptually preferred is a very interesting problem. A positive answer would change the conventional way to model F0 variations for perception-targeted applications.

In our perceptual tests, all the modified utterances that carry linearly approximated F0 contours were compared with the natural speech. To compensate for the distortion of voice quality caused by F0 modification, the F0 levels of both the observed and the approximated contours were down-shifted by 10 Hz. Nevertheless, the distortion in the modified speech is still more severe than that in the natural speech. It would be more reasonable to use synthesized speech as carriers in the perceptual comparison of the two kinds of F0 contours, for a better balance of the influence from speech quality distortion.

7.1.3. Extension to other tonal languages

We believe that the comparable perception between observed F0 contours and their linear approximations is physically due to the perceptual limitations on detecting F0 variations. Hence, the capability of linearly approximated F0 contours should not be limited to a single language. With great possibility, this phenomenon is language-independent. Just as that has been found in several European languages, linearly approximated syllable-sized F0 contours do give comparable perception as well. Our work on Cantonese additionally found such linear approximations are greatly constrained by the lexical tone patterns. This will be particularly directive for similar studies on other tonal languages. In addition, the methodology in this study is also referable to other similar studies.

For the next language to be studied, Mandarin would be particularly interesting. Mandarin has a contour tone system which is much different from Cantonese. Moreover, studies on Mandarin are quite desired due to lots of practical demands.

7.2. Automatic Approximation

7.2.1. Syllable boundary as prior knowledge

In Chapter 5, it is noted that the performance of an approximation might be greatly decreased if there is wrongly located turning point(s). Hence, correctly locating turning points is very important in automatic approximation.

In our automatic process, turning points are globally located and the information of syllable boundaries is not referred. Indeed, the information of syllable boundaries is helpful for the process. It helps to fix the number of turning points for the individual contour regions that belong to different syllables. It also provides good references for examining whether the turning points are reasonably located. The cost of marking syllable boundaries is relatively low. Given well-trained acoustic models, syllable boundaries can be easily obtained by HMM force alignment. We believe prior knowledge of syllable boundary can lead to better located turning points and accordingly improved performance of automatic approximation.

7.2.2. Graphic user interface

In our point of view, automatic approximation is a basic step to the applicable issue of linear approximations. Subsequently, we hope that the framework can be more controllable with the help of a user-friendly graphic interface. The interface is expected to be capable of providing visualized results for each of the three steps in the automatic approximation. Allowable free adjustments on the visualized results are desired as additional function. Real time speech re-synthesis to follow any modified F0 contour must be embedded as a basic function. We hope such an interface can empower the developed automatic approximation as an analysis tool, to associate linearly approximated F0 contours with perception.

7.2.3. Extension studies

Our automatic approximation is developed based on the utterances of polysyllabic words that carry non-rising tones. As discussed in Chapter 5, automatic approximation of the F0 contours of rising tones would be some different, since two linear movements are needed. Adjustments have to be made on the step of locating turning points to include extra intra-tone turning points for rising tones.

The framework is currently examined on word-level utterances. Extension to sentence-level utterances is necessary. As indicated from the manual approximations, procedures of processing word-level and sentence-level utterances are similar in nature.

Consequently, we expect that in the extension, there would be no substantial modification on the algorithm, and the framework performance on the sentence-level utterances would be at a similar level as that obtained from the word-level utterances.

7.3. Perceptual Studies on Cantonese Tones

7.3.1. Identification test on the linearly approximated tone contours

In our study, there is an assumption that when human subjects cannot perceive any difference between an observed F0 contour and its approximation, the approximation carries all the important F0 variations for speech naturalness perception and tone identification. Nevertheless, for a better interpretation of linear F0 movements in function of tone identification, a tone identification test on the linearly approximated F0 contours is still necessary. Such a test will make our understanding on perceiving F0 contours in tonal languages more comprehensive and systematic.

7.3.2. Perceptual tolerance on the F0 variations with communicative functions

In Chapter 3, the analysis of movement slopes with perception suggests that the movement slope of a particular Cantonese contour tone might need to change in a certain range. In Chapter 4, the examination of a large number of linear approximations also shows that the linear movements in a particular tone category are much differentiable from those in other tone categories, in terms of the changing range of each attribute of the linear movement. In view of these findings, we consider for the studies related to speech naturalness, investigations on the perceptual tolerance that allows variants of an underlying target to change are very interesting and meaningful. Such perceptual tolerance on tone identification has been learned a lot. Based on our understanding on the perception of speech naturalness, we predict that the perceptual tolerance on the F0 variations for perceiving "natural" contours should be more restrictive than the one for perceiving "identifiable" contours.

7.3.3. Psychoacoustic perception of the linearly approximated F0 contours

Psychoacoustic perception of F0 variations here indicates that the F0 contours are presented to subjects without speech carriers, but in the form of simple signals like sinusoidal signal, pulses or hum.

To learn the effects of complex speech signals on the perception of F0 variations, we are quite interested in the perceptual difference between the observed F0 contours and their approximations, in condition that the speech carriers are removed. Firstly, we concern whether the comparably perceived contour pairs in our study are differentiable in such a condition. Secondly, we are eager to know which one between the two kinds of F0 contours, observed and linearly approximated, is better in terms of "pure" tone identification.

Appendix 1

Semitone and F0 Distance

Octave

In music, an *octave* is the interval between two sounds one of which has twice the frequency of the other. For example, if a sound has a frequency of 400 Hz, its one octave above sound should have a frequency of 800 Hz. The human ear tends to hear the both sounds which have an octave interval as being essentially "the same", due to closely related harmonics.

Semitone

An octave is divided into 12 semitones. A semitone which is the smallest musical interval is also called a half step or a half tone. Semitone distance is often used to represent the perceptual distance between two sounds.

Semitone and F0 Distance

A distance D_s, in semitones, between any two frequencies f_1 and f_2 is calculated by the following formula,

$$D_s = 12 \times \log_2 \frac{f_1}{f_2} \tag{A1.1}$$

Reversely, given a semitone distance D_s and a reference frequency f_r, the frequency difference (to the reference frequency) D_f can be estimated by the following formula,

$$D_f = 2^{D_s/12} \times f_r - f_r \tag{A1.2}$$

Appendix 2

ANOVA

Basic Concepts of ANOVA

Analysis of variance (ANOVA) provides statistical means to examine the significance level of difference among groups of data. In the testing, two kinds of variables are involved, categorical and quantitative. The main question is: do the means of the quantitative variables depend on which group (categorical variable) the individual is in? Using the data given in Table A2.1 as example, the simplest ANOVA tests two hypothesis:

H_0 : The means of all the groups (treatment here) are equal;

H_1 : Not the means of all the groups are equal.

Table A2.1. Data example for ANOVA analysis.

Subject: 25 patients with blisters.

Measurement: given a treatment, number of the days till the blisters heal.

Treatment	Measurements	Mean of measurements
A	5, 6, 6, 7, 7, 8, 9, 10	7.25
B	7, 7, 8, 9, 9, 10, 10, 11	8.88
C	7, 9, 9, 10, 10, 10, 11, 12, 13	10.11

The test measures *between-group* variation by calculating the differences between each group mean and the overall mean, and then averaging the differences,

$$MSG = \frac{\sum_{i}(\overline{X_i} - \overline{X})^2}{DFG} \tag{A2.1}$$

where MSG, $\overline{X_i}$, $\overline{X}$, DFG denotes the between-group variation, the data mean of group i, the mean of entire data set, and the number of calculated group differences respectively.

Meanwhile the test also measures *within-group* variation by calculating the difference between each data value and the mean of its group, and then averaging the differences,

$$MSE = \frac{\sum_{i}\sum_{j}(x_{ij} - \overline{X_i})^2}{DFE} \tag{A2.2}$$

where MSE, x_{ij}, $\overline{X_i}$, DFE denotes the within-group variation, data j in group i, the data mean of group i, and the number of calculated within group differences respectively.

The ANOVA F-statistic is an indicator of the difference level among the groups. It is a ratio of the between-group difference (MSG) relative to the within-group difference (MSE),

$$F(DFG, DFE) = \frac{Between\ group}{Within\ group} = \frac{MSG}{MSE} \tag{A2.3}$$

A large F is the evidence against H_0, since it indicates that there is more difference between groups than within groups.

Whether the difference among groups of data is significant is eventually determined by p-value. In statistical significance testing, the p-value is defined as the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The lower the p-value, the less likely the result is if the null hypothesis is true, and consequently the *more* "significant" the result is, in the sense of statistical significance. One often accepts the alternative hypothesis, (i.e. rejects a null hypothesis) if the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true. In ANOVA, p-value comes from $F(DFG, DFE)$. Higher F leads to lower p-value.

Several Types of Analysis in ANOVA

One-way ANOVA

One-way ANOVA works similarly with the basic concepts of ANOVA. It deals with one independent variable (categorical variable) and one dependent variable (quantitative variable). In other words, it is used to test for the difference among two or more independent groups (categorical variable). In the case that only two groups are compared, one-way ANOVA is equivalent with t-test.

Multi-factor ANOVA

Multi-factor ANOVA is used when effects of more than one independent variables (categorical variables) is studied. The most commonly used type is two-way ANOVA, where there are two independent variables. For example, in the data given in Table A2.2, there are two independent variables, "experimental group" and "gender". With two-way ANOVA, the interaction between the two independent variables can be measured. In a three-way ANOVA analysis, three independent variables (e.g. A, B, C) leads to three first-order interactions (AB, AC, BC) and one second-order interaction (ABC). As the number of independent variables (categorical) increases, the number of interactions increases and the interpretation of the model becomes more difficult.

Table A2.2. Data example for two-way ANOVA analysis.

Gender	**Experimental group**		
	Experimental group 1	**Experimental group 2**	**Experimental group 3**
Male	2	6	5
	3	7	7
	1	5	9
Female	4	8	9
	5	9	11
	3	7	7

Repeated measures design

Repeated measures design is suitable for the studies in which the same measures are collected multiple times for each subject but under different conditions. For instance, a group of subjects are asked to take a performance test four times — once under each of the four levels of noise distraction. Using the level of noise distraction as the factor of repeated measures (also called within-subject factor), the effects of the different noise levels on the subjects can be measured. In this example, if the variable used as the factor of repeated measures (e.g. the level of noise distraction) only has two categorical values (e.g., two levels of noise distraction), repeated measures would be similar as a paired (or dependent samples) t-test.

Appendix 3

Polysyllabic Words – Word List and Perceptual Test Results

CUWord Corpus

All the words and the natural speech utterances in the data set of polysyllabic words were selected from the speech corpus *CUWord*. CUWord is a multiple-speaker corpus and contains the utterances of 2527 Cantonese words. These words have lengths from one to seven syllables. All the utterances were recorded as prompt speech in a quiet room and were down-sampled to 16,000 Hz in a post-process. 28 native Cantonese speakers (13 male and 15 female) participated in the recording and each of them spoke all the words in the word list.

List of Selected Polysyllabic Words

In our study, we selected 335 words which carry only non-rising tones and the natural utterances were selected from the speech of a female speaker (CW02F). The observed F0 contours of these utterances were manually approximated in Chapter 4 and were automatically approximated in Chapter 5. The word list and the perceptual test results of the two kinds of approximations are given in Table A3.1.

Table A3.1. Word list of polysyllabic words and the perceptual test results on the manually and the automatically approximated F0 contours of these words.

ID: Word ID in *CUWord*.

s: Number of responses to "same" (total responses for a word is 10);

m: Number of responses to "modified speech" (total responses for a word is 10);

Matching rate: the matching rate between the automatically located turning points and the manually marked ones.

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
1	8	一諾千金 jat1-lok6-cin1-gam1	9	0	100	9	0
2	23	大肆擴張 daai6-si3-kwong3-zoeng1	8	0	87.5	7	2
3	24	大獲全勝 daai6-wok6-cyun4-sing3	8	2	87.5	9	1
4	28	大權在握 daai6-kyun4-zoi6-ngak1	7	2	75	8	1
5	33	不琢不成器 bat1-doek3-bat1-sing4-hei3	9	0	90	10	0
6	34	不進則退 bat1-zeon3-zak1-teoi3	8	0	87.5	4	1
7	40	天氣酷熱 tin1-hei3-hou6-jit6	8	0	75	5	2
8	41	天崩地裂 tin1-bang1-dei6-lit6	10	0	87.5	8	1
9	50	支撐大局 zi1-caang1-daai6-guk6	7	2	87.5	7	0
10	59	凹凸不平 lap1-dat6-bat1-ping4	9	0	87.5	10	0
11	68	叱吒風雲 cik1-caak1-fung1-wan4	9	0	87.5	9	1
12	72	平價出售 peng4-gaa3-ceot1-sau6	8	1	100	5	1
13	75	未卜先知 mei6-buk1-sin1-zi1	6	2	100	10	0
14	76	末期癌症 mut6-kei4-ngaam4-zing3	9	0	100	8	0
15	77	生物化學系 sang1-mat6-faa3-hok6-hai6	7	1	90	6	1
16	81	甲型肝炎 gaap3-jing4-gon1-jim4	4	2	75	10	0
17	86	全部囊括 cyun4-bou6-nong4-kut3	9	1	100	8	1
18	100	百尺竿頭 baak3-cek3-gon1-tau4	8	2	100	5	5

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
19	106	衣著樸素 ji1-zoek3-pok3-sou3	10	0	100	8	0
20	114	吸血殭屍 kap1-hyut3-goeng1-si1	7	1	100	10	0
21	118	夾心階層 gaap3-sam1-gaai1-cang4	8	2	75	3	0
22	121	抗日戰爭 kong3-jat6-zin3-zang1	9	1	100	7	2
23	128	沙嗲牛肉 saa3-de1-au4-juk6	9	0	87.5	7	2
24	135	迂迴曲折 jyu1-wui4-kuk1-zit3	8	2	87.5	6	3
25	144	奇門遁甲 kei4-mun4-deon6-gaap3	9	0	87.5	9	0
26	150	房屋津貼 fong4-nguk1-zeon1-tip3	8	1	100	9	1
27	156	放棄原則 fong3-hei3-jyun4-zak1	7	3	100	9	0
28	157	放輕腳步 fong3-heng1-goek3-bou6	7	1	100	6	1
29	159	明察秋毫 ming4-caat3-cau1-hou4	7	2	87.5	7	0
30	168	爭權奪利 zang1-kyun4-dyut6-lei6	9	0	100	7	1
31	173	花旗參燉雞 faa1-kei4-sam1-dan6-gai1	7	1	80	7	1
32	176	長命百歲 coeng4-meng6-baak3-seoi3	8	0	87.5	7	1
33	197	活潑動人 wut6-put3-dung6-jan4	10	0	100	10	0
34	199	皇道吉日 wong4-dou6-gat1-jat6	10	0	100	6	0
35	209	軍閥割據 gwan1-fat6-got3-geoi3	8	1	87.5	7	0
36	210	飛黃騰達 fei1-wong4-tang4-daat6	7	1	87.5	7	1
37	211	食飯行街 sik6-faan6-haang4-gaai1	7	0	75	9	0
38	217	唉聲嘆氣 aai1-sing1-taan3-hei3	10	0	100	6	1
39	222	旁敲側擊 pong4-haau1-zak1-gik1	8	2	87.5	8	0
40	227	氣吞山河 hei3-tan1-saan1-ho4	8	0	62.5	10	0
41	233	索然無味 sok3-jin4-mou4-mei6	9	1	75	9	1
42	239	荊軻刺秦王 ging1-ngo1-ci3-ceon4-wong4	4	0	80	8	1
43	240	迷惑眾生 mai4-waak6-zung3-sang1	7	0	100	10	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
44	243	高級文憑 gou1-kap1-man4-pang4	7	1	100	3	0
45	244	高深莫測 gou1-sam1-mok6-cak1	9	0	87.5	8	0
46	249	停下歇腳 ting4-haa6-hit3-goek3	9	1	87.5	7	1
47	261	域多利兵房 wik6-do1-lei6-bing1-fong4	3	0	90	4	0
48	263	執行任務 zap1-hang4-jam6-mou6	6	1	87.5	8	1
49	264	執行命令 zap1-hang4-ming6-ling6	9	1	87.5	9	1
50	269	強勁節拍 koeng4-ging6-zit3-paak3	9	1	75	6	1
51	275	掙扎求存 zang1-zaat3-kau4-cyun4	7	2	100	7	1
52	276	望而卻步 mong6-ji4-koek3-bou6	10	0	87.5	7	1
53	283	略佔優勢 loek6-zim3-jau1-sai3	10	0	87.5	8	1
54	288	脫胎換骨 tyut3-toi1-wun6-gwat1	10	0	87.5	7	1
55	299	尋尋覓覓 cam4-cam4-mik6-mik6	9	1	87.5	8	1
56	310	晴天霹靂 cing4-tin1-pik1-lik1	6	1	100	8	0
57	321	詞鋒銳利 ci4-fung1-jeoi6-lei6	8	0	75	10	0
58	325	超額認購 ciu1-ngaak6-jing6-kau3	7	1	87.5	1	0
59	327	量度儀器 loeng4-dok6-ji4-hei3	8	0	100	8	1
60	333	黃河流域 wong4-ho4-lau4-wik6	10	0	100	6	2
61	346	滅絕人性 mit6-zyut6-jan4-sing3	10	0	87.5	9	0
62	350	萬壽無疆 maan6-sau6-mou4-goeng1	6	1	75	10	0
63	355	誇誇其談 kwaa1-kwaa1-kei4-taam4	10	0	75	9	1
64	357	資源勘察 zi1-jyun4-ham3-caat3	7	0	75	4	3
65	364	過埠新娘 gwo3-fau6-san1-noeng4	10	0	75	8	1
66	368	零的突破 ling4-dik1-dat6-po3	8	1	100	5	0
67	369	預防霍亂 jyu6-fong4-fok3-lyun6	10	0	100	9	0
68	381	精心策劃 zing1-sam1-caak3-waak6	8	0	87.5	9	1

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
69	396	寬宏大量 fun1-wang4-daai6-loeng6	7	2	87.5	7	0
70	409	確真確實 kok3-zan1-kok3-sat6	7	1	100	10	0
71	411	窮鄉僻壤 kung4-hoeng1-pik1-joeng6	8	1	87.5	9	0
72	427	憑空捏造 pang4-hung1-lip6-zou6	8	1	87.5	6	0
73	433	橫街窄巷 waang4-gaai1-zaak3-hong6	3	1	100	1	0
74	455	優勝劣敗 jau1-sing3-lyut3-baai6	8	0	87.5	6	0
75	461	燦爛奪目 caan3-laan6-dyut6-muk6	9	0	100	9	0
76	463	獲得支持 wok6-dak1-zi1-ci4	8	1	100	8	1
77	473	舊曲新詞 gau6-kuk1-san1-ci4	10	0	87.5	10	0
78	487	證據確鑿 zing3-geoi3-kok3-zok6	7	2	100	8	2
79	492	懸崖峭壁 jyun4-ngaai4-ciu3-bik3	7	1	100	9	0
80	505	轟轟烈烈 gwang1-gwang1-lit6-lit6	9	1	75	5	1
81	513	鬱鬱而終 wat1-wat1-ji4-zung1	9	1	100	10	0
82	527	一團和氣 jat1-tyun4-wo4-hei3	9	0	87.5	8	1
83	528	一噸鋼鐵 jat1-deon1-gong3-tit3	8	0	75	10	0
84	530	一疊鈔票 jat1-dip6-caau1-piu3	10	0	100	7	0
85	533	七月十四 cat1-jyut6-sap6-sei3	8	1	100	4	0
86	537	人傑地靈 jan4-git6-dei6-ling4	7	1	100	8	1
87	557	不吐不快 bat1-tou3-bat1-faai3	6	4	100	8	1
88	561	不屑一看 bat1-sit3-jat1-hon3	9	1	100	10	0
89	573	心肺功能 sam1-fai3-gung1-nang4	6	0	87.5	7	1
90	574	心曠神怡 sam1-kwong3-san4-ji4	7	3	75	5	1
91	590	四腳爬爬 sei3-goek3-paa4-paa4	10	0	75	8	1
92	592	外牆剝落 ngoi6-coeng4-mok1-lok6	6	2	87.5	8	0
93	605	先進設備 sin1-zeon3-cit3-bei6	8	2	87.5	8	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
94	606	再接再厲 zoi3-zip3-zoi3-lai6	6	3	75	5	1
95	617	行俠仗義 hang4-haap6-zoeng3-ji6	8	1	75	4	1
96	629	劫富濟貧 gip3-fu3-zai3-pan4	9	0	87.5	8	1
97	630	劫數難逃 gip3-sou3-naan4-tou4	10	0	87.5	7	0
98	633	含情脈脈 ham4-cing4-mak6-mak6	7	1	75	8	1
99	634	含羞答答 ham4-sau1-daap3-daap3	6	3	100	8	2
100	645	沙漠梟雄 saa1-mok6-hiu1-hung4	9	0	75	0	0
101	651	狂蜂浪蝶 kong4-fung1-long6-dip6	7	0	75	4	1
102	659	依法沒收 ji1-faat3-mut6-sau1	8	0	100	9	0
103	661	來勢凶凶 loi4-sai3-hung1-hung1	9	1	75	8	0
104	665	叔伯兄弟 suk1-baak3-hing1-dai6	6	2	100	9	0
105	669	姍姍來遲 saan1-saan1-loi4-ci4	8	1	100	6	0
106	677	抽籤決定 cau1-cim1-kyut3-ding6	7	1	87.5	5	0
107	687	迎春接福 jing4-ceon1-zip3-fuk1	8	1	87.5	6	0
108	694	非常含蓄 fei1-soeng4-ham4-cuk1	5	0	87.5	8	2
109	698	垂頭喪氣 seoi4-tau4-song3-hei3	8	2	100	9	0
110	706	後悔莫及 hau6-fui3-mok6-kap6	9	1	100	10	0
111	720	英雄豪傑 jing1-hung4-hou4-git6	9	0	87.5	10	0
112	728	音樂噴泉 jam1-ngok6-pan3-cyun4	7	2	75	8	1
113	731	香滑雪糕 hoeng1-waat6-syut3-gou1	7	2	75	7	0
114	734	兼容並蓄 gim1-jung4-bing3-cuk1	7	2	87.5	8	0
115	742	座無虛設 zo6-mou4-heoi1-cit3	6	2	100	10	0
116	751	校慶聚餐 haau6-hing3-zeoi6-caan1	8	1	87.5	7	0
117	752	核心成員 hat6-sam1-sing4-jyun4	9	1	75	10	0
118	756	殷勤服務 jan1-kan4-fuk6-mou6	8	0	87.5	10	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
119	757	特別快車 dak6-bit6-faai3-ce1	8	1	100	9	0
120	782	售價昂貴 sau3-gaa3-ngong4-gwai3	8	1	75	8	1
121	784	國共內戰 gok3-gung6-noi6-zin3	7	2	87.5	9	1
122	790	從旁協助 cung4-pong4-hip3-zo6	7	1	100	7	1
123	817	設備齊全 cit3-bei6-cai4-cyun4	9	0	100	6	1
124	821	貧賤不能移 pan4-zin6-bat1-nang4-ji4	6	3	80	6	3
125	822	速遞公司 cuk1-dai6-gung1-si1	10	0	100	9	0
126	823	速戰速決 cuk1-zin3-cuk1-kyut3	7	2	100	7	0
127	832	喪權辱國 song3-kyun4-juk6-gok3	7	3	87.5	5	3
128	834	尋求協助 cam4-kau4-hip3-zo6	5	0	87.5	6	0
129	839	朝氣勃勃 ziu1-hei3-but6-but6	7	0	75	9	0
130	841	發掘潛力 faat3-gwat6-cim4-lik6	6	2	75	8	1
131	851	肅貪倡廉 suk1-taam1-coeng1-lim4	9	0	87.5	9	0
132	871	亂七八糟 lyun6-cat1-baat3-zou1	8	0	87.5	4	1
133	872	傾國傾城 king1-gok3-king1-sing4	9	1	87.5	10	0
134	884	新屋裝修 san1-nguk1-zong1-sau1	7	1	87.5	6	2
135	896	滔滔不絕 tou1-tou1-bat1-zyut6	10	0	87.5	9	0
136	905	腳踏實地 goek3-daap6-sat6-dei6	7	1	100	8	1
137	913	態度傲慢 taai3-dou6-ngou6-maan6	9	0	100	10	0
138	914	態度親切 taai3-dou6-can1-cit3	9	0	87.5	6	1
139	919	漆黑一片 cat1-hak1-jat1-pin3	9	0	100	9	0
140	921	瘋狂購物 fung1-kwong4-kau3-mat6	6	2	100	8	0
141	925	維持秩序 wai4-ci4-dit6-zeoi6	6	0	100	8	0
142	926	貌若潘安 maau6-joek6-pun1-ngon1	9	0	87.5	8	1
143	927	輕鬆自如 hing1-sung1-zi6-jyu4	8	1	75	10	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
144	932	銀行透支 ngan4-hong4-tau3-zi1	7	1	87.5	7	3
145	947	層出不窮 cang4-ceot1-bat1-kung4	8	0	87.5	7	0
146	957	熱切盼望 jit6-cit3-paan3-mong6	8	0	100	7	3
147	967	趣味盎然 ceoi3-mei6-ong3-jin4	9	0	87.5	6	3
148	984	頭頭碰著黑 tau4-tau4-pung3-zoek3-hak1	9	0	100	9	1
149	987	餐廳侍應 caan1-teng1-si6-jing3	10	0	87.5	10	0
150	989	默默無聞 mak6-mak6-mou4-man4	8	1	87.5	10	0
151	990	黔驢技窮 kim4-lou4-gei6-kung4	7	1	62.5	4	1
152	992	應接不暇 jing3-zip3-bat1-haa4	7	1	87.5	6	1
153	998	斷斷續續 dyun6-dyun6-zuk6-zuk6	9	1	100	2	7
154	1004	難兄難弟 naan4-hing1-naan4-dai6	7	2	100	6	3
155	1022	凈賺一萬 zing6-zaan6-jat1-maan6	9	1	75	0	0
156	1038	人丁單薄 jan4-ding1-daan1-bok6	7	1	100	7	0
157	1040	三妻四妾 saam1-cai1-sei3-cip3	8	0	87.5	4	0
158	1044	千奇百怪 cin1-kei4-baak3-gwaai3	7	1	87.5	9	1
159	1050	中西合璧 zung1-sai1-hap6-bik1	8	0	100	4	0
160	1058	分裂國家 fan1-lit6-gok3-gaa1	8	2	87.5	6	2
161	1062	天下太平 tin1-haa6-taai3-ping4	8	2	87.5	9	1
162	1063	天生麗質 tin1-saang1-lai6-zat1	8	1	87.5	8	0
163	1066	太空穿梭機 taai3-hung1-cyun1-so1-gei1	7	1	90	8	1
164	1067	太陽落山 taai3-joeng4-lok6-saan1	8	1	100	7	1
165	1068	心花怒放 sam1-faa1-nou6-fong3	9	0	100	10	0
166	1069	心急如焚 sam1-gap1-jyu4-fan4	7	2	87.5	0	1
167	1075	月缺月圓 jyut6-kyut3-jyut6-jyun4	10	0	100	1	0
168	1078	欠債還錢 him3-zaai3-waan4-cin4	6	1	75	6	1

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
169	1083	世界杯出線權 sai3-gaai3-bui1-ceot1-sin3-kyun4	8	2	66.7	9	0
170	1086	加官晉爵 gaa1-gun1-zeon3-zoeng3	9	0	87.5	8	0
171	1088	加強訓練 gaa1-koeng4-fan3-lin6	9	0	100	9	0
172	1111	光怪陸離 gwong1-gwaai3-luk6-lei4	7	1	87.5	5	0
173	1113	全身濕透 cyun4-san1-sap1-tau3	5	0	87.5	9	0
174	1121	地鐵上蓋物業 dei6-tit3-soeng6-goi3-mat6-jip6	7	2	75	8	1
175	1125	如坐針氈 jyu4-zo6-zam1-zin1	7	3	87.5	8	1
176	1132	百變梅艷芳 baak3-bin3-mui4-jim6-fong1	5	0	80	8	0
177	1137	血債血償 hyut3-zaai3-hyut3-soeng4	9	0	75	9	1
178	1154	含冤待雪 ham4-jyun1-doi6-syut3	7	2	87.5	7	1
179	1159	完璧歸趙 jyun4-bik1-gwai1-ziu6	8	0	87.5	10	0
180	1166	技術高超 gei6-seot6-gou1-ciu1	8	2	87.5	7	1
181	1176	沉著應戰 cam4-zoek3-jing3-zin3	8	0	75	9	0
182	1183	身兼數職 san1-gim1-sou3-zik1	8	0	100	6	0
183	1207	拖拖拉拉 to1-to1-laai1-laai1	6	1	50	6	2
184	1219	欣欣向榮 jan1-jan1-hoeng3-wing4	8	1	75	8	2
185	1235	金碧輝煌 gam1-bik1-fai1-wong4	9	1	100	7	1
186	1240	亭亭玉立 ting4-ting4-juk6-lap6	9	1	87.5	8	0
187	1253	城寨居民 sing4-zaai6-geoi1-man4	6	2	50	8	0
188	1256	建立威信 gin3-lap6-wai1-seon3	8	0	100	8	1
189	1258	怒髮衝冠 nou6-faat3-cung1-gun1	6	1	87.5	9	0
190	1260	急凍雞翼 gap1-dung3-gai1-jik6	7	1	100	6	0
191	1268	挑燈夜讀 tiu1-dang1-je6-duk6	10	0	87.5	9	1
192	1271	殃及池魚 joeng1-kap6-ci4-jyu4	8	0	87.5	4	2
193	1273	洪福齊天 hung4-fuk1-cai4-tin1	7	1	100	5	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
194	1276	炸醬撈麵 zaa3-zoeng3-lou1-min6	7	3	75	2	3
195	1288	訂立契約 ding6-lap6-kai3-joek3	8	2	87.5	10	0
196	1309	家賊難防 gaa1-caak6-naan4-fong4	7	1	100	9	1
197	1314	恩愛夫妻 jan1-ngoi3-fu1-cai1	8	0	87.5	7	2
198	1316	拳拳到肉 kyun4-kyun4-dou3-juk6	6	0	75	4	2
199	1347	貢獻良多 gung3-hin3-loeng4-do1	8	1	75	9	0
200	1356	飢寒交迫 gei1-hon4-gaau1-bik1	9	0	100	7	0
201	1361	高溫消毒 gou1-wan1-siu1-duk6	8	0	87.5	8	0
202	1379	埠際足球賽 fau6-zai3-zuk1-kau4-coi3	5	2	90	8	1
203	1390	推卸責任 teoi1-se3-zaak3-jam6	9	1	75	9	0
204	1414	率先進入 seot1-sin1-zeon3-jap6	8	1	100	7	1
205	1423	細心鑑別 sai3-sam1-gaam3-bit6	9	1	75	7	0
206	1424	終身監禁 zung1-san1-gaam1-gam3	5	1	100	3	0
207	1426	船堅炮利 syun4-gin1-paau3-lei6	7	1	75	7	1
208	1436	通貨膨脹 tung1-fo3-paang4-zoeng3	7	1	100	7	1
209	1442	雀巢咖啡 zoek3-caau4-gaa3-fe1	6	0	87.5	8	1
210	1456	富貴榮華 fu3-gwai3-wing4-waa4	7	1	87.5	5	0
211	1459	循序漸進 ceon4-zeoi6-zim6-zeon3	7	0	87.5	0	0
212	1464	欺人太甚 hei1-jan4-taai3-sam6	9	1	87.5	5	0
213	1471	無怨無悔 mou4-jyun3-mou4-fui3	9	0	87.5	9	1
214	1488	華麗吊燈 waa4-lai6-diu3-dang1	10	0	87.5	9	0
215	1496	越級挑戰 jyut6-kap1-tiu1-zin3	9	0	87.5	7	0
216	1508	債臺高築 zaai3-toi4-gou1-zuk1	7	1	100	4	2
217	1513	敬業樂業 ging3-jip6-ngaau6-jip6	5	0	87.5	8	0
218	1519	照價賠償 ziu3-gaa3-pui4-soeng4	9	1	75	4	5

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
219	1522	萬象更新 maan6-zoeng6-gang1-san3	7	1	87.5	9	0
220	1539	道聽途說 dou6-ting3-tou4-syut3	8	0	87.5	9	1
221	1546	僥倖勝出 hiu1-hang6-sing3-ceot1	7	1	87.5	7	1
222	1555	慢慢浸淫 maan6-maan6-zam3-jam4	6	2	75	8	0
223	1569	碧血丹心 bik1-hyut3-daan1-sam1	9	1	100	9	0
224	1570	碧波蕩漾 bik1-bo1-dong6-joeng6	5	2	87.5	5	4
225	1573	精乖伶俐 zeng1-gwaai1-ling4-lei6	9	0	75	10	0
226	1584	誤信讒言 m6-seon3-caam4-jin4	7	1	75	7	1
227	1593	嘻笑怒罵 hei1-siu3-nou6-maa6	9	1	87.5	8	2
228	1594	墮入陷阱 do6-jap6-ham6-zing6	8	2	75	7	1
229	1607	磋砣歲月 co1-to4-seoi3-jyut6	7	0	87.5	9	0
230	1621	鴉雀無聲 aa1-zoek3-mou4-sing1	9	1	100	9	1
231	1624	戰術成功 zin3-seot6-sing4-gung1	7	2	75	7	0
232	1643	龍蛇混雜 lung4-se4-wan6-zaap6	8	2	100	8	1
233	1650	濫竽充數 laam6-jyu4-cung1-sou3	9	0	100	10	0
234	1654	繁榮昌盛 faan4-wing4-coeng1-sing6	9	1	87.5	4	2
235	1658	薄利多銷 bok6-lei6-do1-siu1	3	0	100	7	0
236	1666	蟬聯冠軍 sim4-lyun4-gun3-gwan1	6	2	87.5	7	0
237	1667	豐儉由人 fung1-gim6-jau4-jan4	8	2	87.5	4	0
238	1671	懷恨在心 waai4-han6-zoi6-sam1	7	2	75	8	0
239	1674	爆炸現場 baau3-zaa3-jin6-coeng4	10	0	75	8	0
240	1684	競爭激烈 ging6-zang1-gik1-lit6	7	1	87.5	8	2
241	1704	鶴立雞群 hok6-lap6-gai1-kwan4	10	0	100	9	0
242	1707	權力鬥爭 kyun4-lik6-dau3-zang1	9	1	62.5	7	0
243	1753	出入平安 ceot1-jap6-ping4-ngon1	7	1	75	8	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
244	1766	平鋪直敘 ping4-pou1-zik6-zeoi6	4	1	100	8	0
245	1770	用力撬開 jung6-lik6-giu6-hoi1	9	0	87.5	9	0
246	1777	名校學生 ming4-haau6-hok6-sang1	9	0	100	9	0
247	1781	地位超然 dei6-wai6-ciu1-jin4	2	0	75	0	0
248	1796	作奸犯科 zok3-gaan1-faan6-fo1	9	0	100	8	1
249	1797	呆若木雞 ngoi4-joek6-muk6-gai1	9	0	100	10	0
250	1798	含苞待放 ham4-baau1-doi6-fong3	7	1	100	1	0
251	1829	明文規定 ming4-man4-kwai1-ding6	8	0	87.5	1	0
252	1840	前路崎嶇 cin4-lou6-kei1-keoi1	6	3	100	9	0
253	1849	流行音樂 lau4-hang4-jam1-ngok6	8	1	75	8	0
254	1855	看破紅塵 hon3-po3-hung4-can4	9	0	75	7	2
255	1856	研究昆蟲 jin4-gau3-kwan1-cung4	6	0	75	8	1
256	1862	重振雄風 cung4-zan3-hung4-fung1	3	0	87.5	7	0
257	1867	倒掛金鉤 dou3-gwaa3-gam1-au1	6	4	100	6	1
258	1876	浩浩蕩蕩 hou6-hou6-dong6-dong6	9	0	87.5	5	4
259	1891	參加宴會 caam1-gaa1-jin3-wui6	8	2	100	9	0
260	1895	唯肖唯妙 wai4-ciu3-wai4-miu6	6	2	75	8	1
261	1897	婆婆媽媽 po4-po4-maa1-maa1	7	0	75	9	1
262	1917	貨真價實 fo3-zan1-gaa3-sat6	6	2	87.5	5	0
263	1924	逢年過節 fung4-nin4-go3-zit3	9	0	100	9	1
264	1925	陰差陽錯 jam1-caa1-joeng4-co3	8	0	62.5	6	2
265	1937	登堂入室 dang1-tong4-jap6-sat1	9	1	100	10	0
266	1940	筋疲力盡 gan1-pei4-lik6-zeon6	8	0	100	6	0
267	1944	陽光燦爛 joeng4-gwong1-caan3-laan6	7	1	75	5	1
268	1946	順應潮流 seon6-jing3-ciu4-lau4	3	1	62.5	4	3

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
269	1983	禍從天降 wo6-cung4-tin1-gong3	7	1	75	7	2
270	1989	豪邁奔放 hou4-maai6-ban1-fong3	8	1	75	4	2
271	2006	衝鋒陷陣 cung1-fung1-ham6-zan6	8	0	75	5	1
272	2009	賢良淑德 jin4-loeng4-suk6-dak1	8	1	100	9	1
273	2025	繁殖後代 faan4-zik6-hau6-doi6	9	0	87.5	3	0
274	2042	難得糊塗 naan4-dak1-wu4-tou4	5	1	100	5	0
275	2078	多謝光臨 do1-ze6-gwong1-lam4	8	2	87.5	10	0
276	2134	搬搬抬抬 bun1-bun1-toi4-toi4	9	0	87.5	9	1
277	2190	三更半夜 saam1-gaang1-bun3-je6	9	1	75	10	0
278	2194	驚鴻一瞥 ging1-hung4-jat1-pit3	8	1	75	10	0
279	2202	全新創作 cyun4-san1-cong3-zok3	8	2	75	3	1
280	2204	收拾殘局 sau1-sap6-caan4-guk6	7	3	100	7	0
281	2209	春夏秋冬 ceon1-haa6-cau1-dung1	6	1	100	9	0
282	2216	梅開二度 mui4-hoi1-ji6-dou6	7	0	100	9	0
283	2217	堪稱佳作 ham1-cing1-gaai1-zok3	6	2	87.5	7	0
284	2218	無疾而終 mou4-zat6-ji4-zung1	7	3	100	2	8
285	2219	開壇作法 hoi1-taan4-zok3-faat3	9	1	87.5	10	0
286	2220	微服出巡 mei4-fuk6-ceot1-ceon4	9	1	100	5	4
287	2233	大汗淋漓 daai6-hon6-lam4-lei4	7	1	75	1	0
288	2239	危害健康 ngai4-hoi6-gin6-hong1	9	1	87.5	8	2
289	2246	到處留情 dou3-cyu3-lau4-cing4	8	1	87.5	8	1
290	2248	和氣生財 wo4-hei3-sang1-coi4	5	0	75	7	2
291	2249	安居思危 ngon1-geoi1-si1-ngai4	8	1	87.5	8	1
292	2250	林蔭大道 lam4-jam3-daai6-dou6	8	0	100	8	1
293	2253	後羿射箭 hau6-ngai6-se6-zin3	6	0	75	8	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
294	2262	疏財仗義 so1-coi4-zoeng6-ji6	8	1	87.5	6	3
295	2269	植樹造林 zik6-syu6-zou6-lam4	9	0	75	3	0
296	2274	新陳代謝 san1-can4-doi6-ze6	8	0	75	7	1
297	2291	大江南北 daai6-gong1-naam4-bak1	8	2	75	5	1
298	2294	破敵之計 po3-dik6-zi1-gai3	2	2	87.5	6	0
299	2295	恨之入骨 han6-zi1-jap6-gwat1	9	0	75	5	0
300	2298	知慳識儉 zi1-haan1-sik1-gim6	7	1	87.5	8	1
301	2300	禽獸不如 kam4-sau3-bat1-jyu4	7	1	75	9	0
302	2301	人浮於事 jan4-fau4-jyu1-si6	8	2	87.5	6	0
303	2304	習慣成自然 zaap6-gwaan3-sing4-zi6-jin4	6	1	50	4	1
304	2305	一竅不通 jat1-hiu3-bat1-tung1	8	1	100	9	0
305	2306	熱帶地區 jit6-daai3-dei6-keoi1	9	0	100	6	0
306	2309	疾惡如仇 zat6-ngok3-jyu4-sau4	9	0	87.5	10	0
307	2314	爭拗不斷 zang1-ngaau3-bat1-dyun6	9	1	100	6	2
308	2315	配襯衣物 pui3-can3-ji1-mat6	8	2	62.5	4	3
309	2316	飄忽不定 piu1-fat1-bat1-ding6	9	0	75	9	0
310	2318	東亞病夫 dung1-ngaa3-beng6-fu1	8	1	87.5	6	0
311	2322	特殊待遇 dak6-syu4-doi6-jyu6	7	0	87.5	7	1
312	2325	專心學業 zyun1-sam1-hok6-jip6	5	1	87.5	7	1
313	2328	眾目睽睽 zung3-muk6-kwai4-kwai4	7	0	75	7	1
314	2330	六國大封相 luk6-gwok3-daai6-fung1-soeng3	8	1	100	5	2
315	2334	特別行政區 dak6-bit6-hang4-zing3-keoi1	4	1	90	1	0
316	2346	中華民國 zung1-waa4-man4-gok3	6	3	75	9	0
317	2353	天旋地轉 tin1-syun4-dei6-zyun3	9	0	100	8	0
318	2359	去向未明 heoi3-hoeng3-mei6-ming4	7	2	50	10	0

No.	ID	Word	Results of Perceptual Test				
			Manual Approx.		Automatic Approx.		
			s	m	Matching rate (%)	s	m
319	2367	民族尊嚴 man4-zuk6-zyun1-jim4	9	1	100	10	0
320	2374	再三退讓 zoi3-saam1-teoi3-joeng6	3	0	62.5	5	2
321	2376	名門望族 ming4-mun4-mong6-zuk6	8	2	100	9	0
322	2396	足球初賽 zuk1-kau4-co1-coi3	9	0	87.5	7	1
323	2403	物盡其用 mat6-zeon6-kei4-jung6	9	0	87.5	7	1
324	2417	英國殖民地 jing1-gok3-zik6-man4-dei6	9	0	90	9	1
325	2426	財運亨通 coi4-wan6-hang1-tung1	6	3	62.5	9	1
326	2446	街頭霸王 gaai1-tau4-baa3-wong4	8	2	75	8	0
327	2450	開門見山 hoi1-mun4-gin3-saan1	8	0	87.5	9	0
328	2455	搖搖欲墜 jiu4-jiu4-juk6-zeoi6	5	0	87.5	6	1
329	2456	搖頭歎息 jiu4-tau4-taan3-sik1	9	0	100	10	0
330	2462	跪地求饒 gwai6-dei6-kau4-jiu4	9	1	100	6	2
331	2466	雍容華貴 jung1-jung4-waa4-gwai3	8	2	62.5	10	0
332	2515	無無聊聊 mou4-mou4-liu4-liu4	7	1	100	6	1
333	2519	經濟學家 ging1-zai3-hok6-gaa1	8	0	87.5	8	0
334	2523	意態娉婷 ji3-taai3-ping1-ting4	7	0	87.5	9	0
335	2524	南京大屠殺 naam4-ging1-daai6-tou4-saat3	6	0	90	5	1

Appendix 4

Weighted Time Average Model

For Perceived Pitch Level

In the perceptual studies of perceived pitch of an F0 contour, for the cases that the contours change with a small extent and only a constant pitch is perceived, the experimental data suggested that in the pitch judgment, F0 contours were time averaged and the final portion of the contour had a larger weight. Weighted time average (WTA) model is a quantitative model describing such a process (d'Alessandro and Castellengo, 1994). Let $p(t)$ denote the pitch perceived at time t, f the time-varying F0 function beginning at time 0, and let α, β be two constants. WTA model is given as,

$$p(t) = \frac{\int_0^t (e^{\alpha(\tau - t)} + \beta) f(\tau) d\tau}{\int_0^t (e^{\alpha(\tau - t)} + \beta) d\tau} \tag{A4.1}$$

This model raises an exponential memory function for the data window, so that events in the past contribute exponentially less to the average. An exponential memory function is not sufficient, because only the recent past is taken into account. Thus, a time average function of the entire contour is combined. The meaning of Equation (A4.1) is twofold: the constant β accounts for the time averaging, and the constant α accounts for the weighting of the past. Figure A4.1 is a simulation of the weight function $e^{\alpha(\tau-t)}$, when $t = 300\,ms$. A positive α indicates that a higher perceptual weight is assigned to the most recent information (end). A negative α indicates that a higher perceptual weight is assigned to the most distant information (beginning). The magnitude of α represents the amount of damping introduced in averaging. A larger magnitude means that a higher perceptual weight is assigned to the information close to the extremity matched. A small magnitude indicates that all the parts of the time interval used for averaging have almost the same weight.

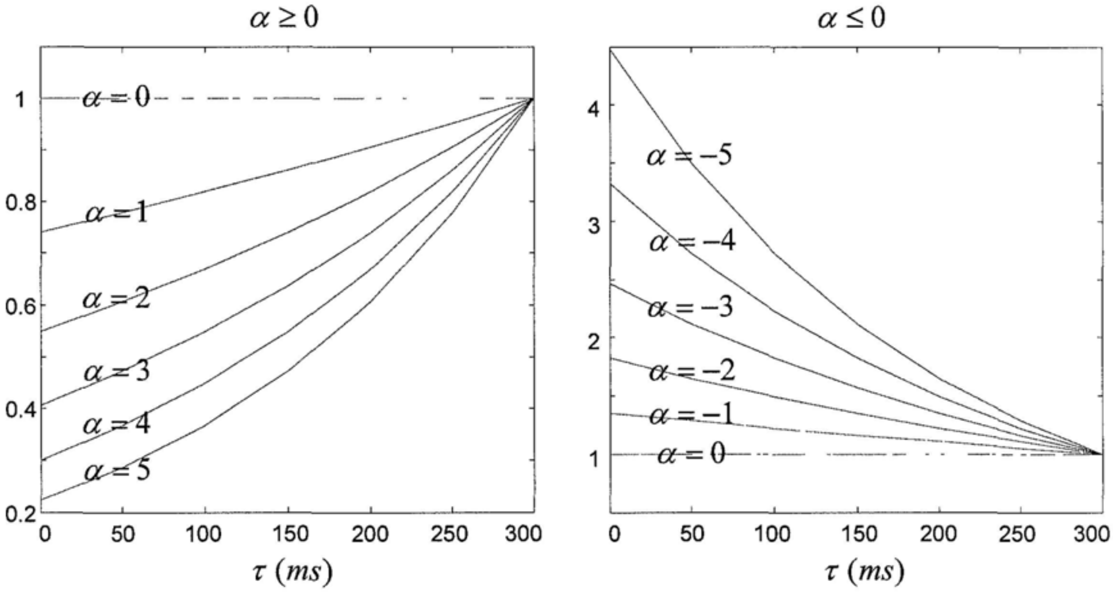


FIG. A4.1: A simulation of the weight function given by the exponential component in the WTA model.

α and β were optimized by matching the model responses and the experimental data. The optimal parameters are $\alpha = 22$ and $\beta = 0.2$, indicating a larger weight at the end portions of the contour and a 20% long-term time average.

For Perceived Pitch Movement

In the case that an F0 contour changes with a large extent and is perceived as a movement, the extrema of the movement are perceived as two independent auditory events. In such cases, pitch judgment of one extremity of the movement will not take into account the distant occurrences. It does mean that the amount of constant time average represented by β in this kind of pitch judgment is reduced. Therefore, when this model is applied on the prediction of the extremity pitches of movements, it is given by Equation (A4.2) with $\beta = 0$ as,

$$p(t) = \frac{\int_0^t (e^{\alpha(\tau - t)}) f(\tau) d\tau}{\int_0^t (e^{\alpha(\tau - t)}) d\tau} \tag{A4.2}$$

In predicting the perceived beginning and end pitch values of a movement, it is reasonable to assume that time plays a role in the averaging process, so more attention must be paid to the end in case of judging the end pitch, and to the beginning in case of judging the beginning pitch. Let d be the contour duration and δ represent the duration of the averaging interval expressed as a proportion of the contour duration. Then the time limits of integration are $[0, \delta d]$ for a judgment at the beginning, and $[d - \delta d, d]$ for a judgment at the end. The WTA model is computed as,

$$p(\alpha, \delta) = \begin{cases} p(t) = \dfrac{\int_0^{\delta d} (e^{\alpha(\tau - t)}) f(\tau) d\tau}{\int_0^{\delta d} (e^{\alpha(\tau - t)}) d\tau}, & \textit{judgment at the beginning} \\ p(t) = \dfrac{\int_{d-\delta d}^{d} (e^{\alpha(\tau - t)}) f(\tau) d\tau}{\int_{d-\delta d}^{d} (e^{\alpha(\tau - t)}) d\tau}, & \textit{judgment at the end} \end{cases} \tag{A4.3}$$

The values of α and δ were optimized in (d'Alessandro, Rosset and Rossi, 1998) for four different cases: perceived beginning and end pitches of a rising movement, and perceived beginning and end pitches of a falling movement. The optimized parameters were obtained by matching amount of subjectively determined beginning and end pitch values of isolated tone contours. The optimized parameter values are given in Table 5.5. The optimized parameters suggest that a higher weight is given to the contour parts close to the end for the judgment at the end, and a higher weight is given to the contour parts close to the beginning for the judgment at the beginning. The averaging interval is longer for the judgment of the lower extremity pitch than for the judgment of the higher extremity pitch.

Bibliography

Abramson, A.S. (1975). “The tones of central Thai: some perceptual experiments,” *Studies in Thai Linguistics* (Eds: Harris, J.G. and Chamberlain, J.; Bangkok: Central Institute of English Language).

Abramson, A.S. (1976). “Thai tones as a reference system,” *Tai linguistics in honor of Fang-Kuei Li* (Eds: Gething, W., Harris, G.. and Kullavanijaya, P.; Bangkok: Chulalongkorn University Press).

Abramson, A.S. (1979). “The noncategorical perception of tone categories in Thai,” *Frontiers of speech communication research* (Eds: Lindblom, B. and Öhman, S.; London: Academic Press).

Abramson, A.S. (1997). “The Thai tonal space,” *Southeast Asian Linguistic Studies in Honour of Vichin Panupong* (Bangkok: Chulalongkorn University Press).

d'Alessandro, C. and Castellengo, M. (1994). “The pitch of short-duration vibrato tones,” J. Acoust. Soc. Am. 95(3), 1617-1630.

d'Alessandro, C. and Mertens, P. (1995). “Automatic pitch contour stylization using a model of tonal perception,” Computer Speech and Language 9, 257-288.

d'Alessandro, C., Rosset, S. and Rossi J.-P. (1998). “The pitch of short-duration fundamental frequency glissandos,” J. Acoust. Soc. Am. 104(4), 2339-2348.

Bachem, A. (1937). “Various types of absolute pitch,” J. Acoust. Soc. Am. 9, 146-151.

Bauer, R.S. and Benedict, P.K. (1997). *Modern Cantonese Phonology* (New York: Mouton de Gruyter).

Black, A. and Hunt, A. (1996). “Generating F0 contours from ToBI labels using linear regression,” *Proc. International Conference on Spoken Language Processing 1996.*

Blicher, D.L., Diehl, R. and Cohen, L.B. (1990). “Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement,” J. Phon. 18, 37-49.

Boersma, P. and Weenink, D. (2010). *Praat: Doing Phonetics by Computer*, www.praat.org (date last viewed 03/31/10).

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective*, (Lund: Gleerup).

Bruce, G. (1983). "Accentuation and timing in Swedish," Folia Linguistica 17, 221-238.

Burnham, D. and Francis, E. (1997). "The role of linguistic experience in the perception of Thai tones," *Southeast Asian Linguistic Studies in Honour of Vichin Panupong* (Bangkok: Chulalongkorn University Press).

Cardozo, B.L. and Ritsma, R.J. (1965). "Short-time characteristics of periodicity pitch," *Proc. The Fifth International Congress on Acoustics* (Eds: Commins, D.E.), paper B37.

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

Clark, J. and Yallop, C. (1990). *An Introduction to Phonetic and Phonology* (Cambridge, MA: Basil Blackwell, Inc.).

Chen, S.-H. and Kuo, C.-C. (2007). "Perceptual relevance of pitch contours of Mandarin tones and its efficacy in prosody generation of speech synthesis," *Proc. Interspeech 2007*, pp. 2669-2672.

Connell, B. (2000). "The perception of lexical tone in Mambila," Language and Speech 43(2), 163-182.

Connell, B.A., Hogan, J.T., and Rozsypal, A.J. (1983). "Experimental evidence of interaction between tone and intonation in Mandarin Chinese," J. Phon. 11, 337-351.

Cutler, A. and Chen, H.-C. (1997). "Lexical tone in Cantonese spoken-word processing," Percept. Psychophys. 59(2), 165-179.

Denes, P.B. and Pinson, E.N. (1998). *The Speech Chain: The Physics and Biology of Spoken Language* (New York: W.H. Freeman and Company).

Dong, M. and Lua, K.T. (2002). "Pitch contour model for Chinese text-to-speech using CART and statistical method," *Proc. International Conference on Spoken Language Processing 2002*, pp. 2405-2408.

Dusterhoff, K.E., Black, A.W. and Taylor, P.A. (1999). "Using decision trees within the tilt intonation model to predict F0 contours," *Proc. of Eurospeech 1999.*

Dutiot, T. (1997). *An Introduction to Text-to-Speech Synthesis* (Kluwer Academic Publishers).

Flanagan, J.L. and Saslow, M.G. (1958). "Pitch discrimination for synthetic vowels," J. Acoust. Soc. Am. 30, 435-442.

Fok, C.Y.Y. (1974). *A Perceptual Study of Tones in Cantonese (Occasional Papers and Monographs No. 18)* (Hong Kong: University of Hong Kong, Centre of Asian Studies).

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

Fry, D.B., Abramson, A.S., Eimas, P.D., and Liberman, A.M. (1962). "Identification and discrimination of synthetic vowels," Language and Speech 5, 171-189.

Fujiaski, H. and Hirose, K. (1984). "Analysis of voice fundamental frequency contours of declarative sentences of Japanese," J. Acoust. Soc. Japan (E) 5(4), 233-242.

Fujisaki, H., Ohno, S. and Luksaneeyanawin, S. (2003). "Analysis and synthesis of F0 contours of Thai utterances based on the command-response model," *Proc. International Congress of Phonetic Sciences 2003*.

Fujisaki, H., Wang, C., Ohno, S., and Gu, W. (2005). "Analysis and synthesis of fundamental frequency contours of standard Chinese using the command–response model," Speech Communication 47, 59-70.

Gandour, J. (1981). "Perceptual dimensions of tone: evidence from Cantonese," J. Chin. linguist 9, 20-36.

Gandour, J. (1983). "Tone perception in far eastern languages," J. phon. 11, 149-175.

Gandour, J. (1984). "Tone dissimilarity judgments by Chinese listeners," J. Chin. linguist 12, 235-261.

Gandour, J., Potisuk, S., and Dechongkit, S. (1994). "Tonal coarticulation in Thai", J. phon. 22, 477-492.

Gårding, E. and Eriksson, L. (1991). "On the perception of prosodic phrase patterns," Working Papers, Department of Linguistics, Lund University 38, 45-70.

Gårding, E., Kratochvil, P., Svantesson, J.-O., and Zhang, J. (1986). "Tone 4 and Tone 3 discrimination in modern standard Chinese," Language and Speech 29, 281-293.

Han, M.S., and Kim, K. (1974). "Phonetic variation of Vietnamese tones in dysyllabic utterances," J. phon. 2, 223-232.

t' Hart, J. (1981). "Differential sensitivity to pitch distance, particularly in speech," J. Acoust. Soc. Am. 6, 811-821.

t' Hart, J., Collier, R. and Cohen, A. (1990). *A Perceptual Study of Intonation: Experimental-phonetic Approach to Speech Melody* (Cambridge: Cambridge University Press).

Hashimoto, O.-K.Y. (1972). *Studies in Yue Dialects 1: Phonology of Cantonese* (Cambridge: Cambridge University Press).

Henning, G.B. (1966). "Frequency discrimination of random-amplitude tones," J. Acoust. Soc. Am. 39, 336-339.

Hermes, D. (1996). "Timing of pitch movements and accentuation of syllables," *Proc. International Conference on Spoken Language Processing 1996*, pp. 1197-1200.

Holm, B. and Bailly, G. (2000). "Generating prosody by superposing multi-parametric overlapping contours," *Proc. International Conference on Spoken Language Processing 2000*, pp. 203-206.

House, D. (1990). *Tonal Perception in Speech* (Lund: Lund University Press).

House, D. (1997). "Perceptual thresholds and tonal categories," *Proc. Fonetic*, pp. 179-182.

House, D. (2004). "Pitch and alignment in the perception of tone and intonation: pragmatic signals and biological codes," *Proc. International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages*, pp. 93-96.

Hu, Y. and Loizou, P.C. (2008). "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio Speech Language Process. 16 (1), 229–238.

D'Imperio, M. and House, D. (1997). "Perception of questions and statements in Neqpolitan Italian," *Proc. Eurospeech 1997*.

Issachenko, A.V. and Schädlich, H.-J. (1970). *A Model of Standard German Intonation* (The Hague/Paris: Mouton).

Janse, E. (2004). "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," Speech Communication 42(2), 155-173.

Jongman, A., Wang, Y., Moore, C.B. and Sereno, J. (2006). "Perception and production of Mandarin Chinese tone," *Handbook of East Asian Psycholinguistics* (Vol. 1: Chinese)

(Eds.: Li, P., Tan, L.H., Bates, E. and Tzeng, O.J.L.; Cambridge: Cambridge University Press).

Khouw, E. and Ciocca, V. (2007). "Perceptual correlates of Cantonese tones," J. phon. 35, 104-117.

Kochanski, G.P. and Shih, C. (2003). "Prosody modeling with soft templates," Speech Communication 39, 311-352.

Klatt, D.H. (1973). "Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception," J. Acoust. Soc. Am. 53, 8-16.

Kohler, J.K. (1987). "Categorical pitch perception," *Proc. The Eleventh International Congress of Phonetic Sciences*, pp. 331-333.

Kohler, J.K. (2005). "Timing and communicative functions of pitch contours," Phonetica 62, 88-105.

Lau, W. (2000). *Attributes and Extraction of Tone Information for Continuous Cantonese Speech Recognition*, M. Phil. Thesis. Department of Electronic Engineering, The Chinese University of Hong Kong.

Leather, J. (1987). "F0 pattern inference in the perceptual acquisition of second language tone," *Sound Patterns in Second Language Acquisition* (Eds: James, A. and Leather, J.; Dordrecht: Foris), pp. 59–81.

Lee, C.-Y. (2001). *Lexical Tone in Spoken Word Recognition: A View from Mandarin Chinese*, Ph.D. Dissertation, Brown University.

Lee, Tan and Qian, Y. (2007). "Tone modeling for speech recognition," *Advances in Chinese Spoken Language Processing* (Eds.: Lee, C.-H., Li H., Lee, L.S., Wang, R. H. and Huo, Q.; Singapore: World Scientific Publishing).

Lee, Y.-S., Vakoch, D.A. and Wurm, L.H. (1996). "Tone perception in Cantonese and Mandarin: a cross-linguistic comparison," J. Psycholinguistic Res. 25(5), 527-542.

Lewis, M.P. (Eds.) (2009). *ETHNOLOGUE: Languages of the World (16th Edition)*, http://www.ethnologue.com (date last viewed 06/01/10).

Li, Y.-J. (2003). *Prosody Analysis and Modeling for Cantonese Text-to-Speech*, M. Phil. Thesis, Electronic Engineering Department, The Chinese University of Hong Kong.

Li, Y.-J. (2006). "Tone ratios combined with F0 register in Cantonese as speaker-dependent characteristic," *Proc. SpeechProsody 2006.*

Li, Y.-J. and Lee Tan (2008). "A perceptual study of approximated Cantonese tone contours," *Proc. International Symposium on Chinese Spoken Language Processing 2008.*

Li, Y.-J., Lee, Tan and Qian, Y. (2002), "Acoustical F0 analysis of continuous Cantonese speech," *Proc. International Symposium on Chinese Spoken Language Processing 2002.*

Li, Y.-J., Lee, Tan and Qian, Y. (2004). "Analysis and modeling of F0 contours for Cantonese text-to-speech," ACM Trans. Asian Language Information Processing 3(3), 169-180.

Liberman, A.M., Harris, K.S., Eimas, P.D., Lisker, L., and Bastian, J. (1961). "An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance," Language and Speech 4, 175-195.

Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957). "The discrimination of speech sounds within and across phoneme boundaries," J. exp. psychol. 54, 358-368.

Liberman, A.M., Harris, K.S., Kinney, J.A., and Lane, H. (1961). "The discrimination of relative onset time of the components of certain speech and nonspeech patterns," J. exp. psychol. 61, 379-388.

Linguistic Society of Hong Kong (LSHK) (1997). *Hong Kong Jyut Ping Characters Table* (Hong Kong: Linguistic Society of Hong Kong Press).

Lo, W.K., Lee Tan and Ching, P.C. (1998). "Development of Cantonese spoken language corpora for speech applications," *Proc. International Symposium on Chinese Spoken Language Processing 1998*, pp. 102-107.

Ma, J. K-Y., Ciocca, V., and Whitehill, T. (2005). "Contextual effect on perception of lexical tones in Cantonese," *Proc. Eurospeech 2005*, pp. 401-404.

Maddieson, I. (1978). "The frequency of tones," UCLA Working Papers in Phonetics 41, 43-52.

Massaro, D.W., Cohen, M.M., and Tseng, C.C. (1985). "The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese", J. Chin. linguist. 13, 267-289.

Moore, B.C.J. (1997). "Aspects of auditory processing related to speech perception," *The Handbook of Phonetic Sciences* (Eds.: Hardcastle, W.J. and Laver, J.; UK, Oxford: Blackwell).

Moore, B.C.J. (2004). *An Introduction to the Psychology of Hearing (Fifth edition)* (Elsevier Academic Press).

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

Nabelek, I. and Hirsh, I.J. (1969). “On the discrimination of frequency transitions,” J. Acoust. Soc. Am. 45, 1510-1519.

Ni, J. and Kawai, H. (2006). “Constrained tone transformation technique for separation and combination of Mandarin tone and intonation,” J. Acoust. Soc. Am. 119(3), 1764-1782.

Nordmark, J.O. (1968). “Mechanisms of frequency discrimination,” J. Acoust. Soc. Am. 44, 1533-1540.

Ohala, J.J. and Ewan, W.G. (1973). “Speed of pitch change,” J. Acoust. Soc. Am. 53, 345(A).

Olsberg, M., Xu, Y. and Green, J. (2007). “Dependence of tone perception on syllable perception,” *Proc. Interspeech 2007*, pp. 2649-2652.

Pasdeloup, V., Espesser, R. and Faraj, M. (2006). “Rate sensitivity of syllable in French: a perceptual illusion?” *Proc. SpeechProsody 2006.*

Pierrehumbert, J. (1979). “The perception of fundamental frequency declination,” J. Acoust. Soc. Am. 66, 363-369.

Pike, K.L. (1948). *Tone Languages* (Ann Arbor: University of Michigan Press).

Plomp, R. Wagenaar, W.A. and Mimpen, A.M. (1973). “Musical interval recognition with simultaneous tones,” Acustica, 29, 101-109.

Pollack, I. (1968). “Detection of rate of change of auditory frequency,” J. exp. psycho. 77, 535-541.

Prom-on, S., Xu, Y., and Thipakorn, B. (2009). “Modeling tone and intonation in Mandarin and English as a process of target approximation,” J. Acoust. Soc. Am. 125, 405-424.

Qian, Y., Lee Tan and Soong, Frank K. (2007). “Tone recognition in continuous Cantonese speech using supratone models,” J. Acoust. Soc. Am. 121, 2936-2945.

Ritsma, R.J. (1965). “Pitch discrimination and frequency discrimination,” *Proc. The Fifth International Congress on Acoustics*, pp. B22.

Rossi, M. (1971). “Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole,” Phonetica 23, 1-33.

Rossi, M. (1978). “La perception des glissandos descendants dans les contours prosodiques (The perception of falling glissandos in prosodic contours),” Phonetica 35, 11-40.

Rossi, M. and Chafcouloff, M. (1972). “Recherche sur le seuil differential de frequence fondamentale dans la parole,” Travaux de l’Institut de Phonetique d’Aix (1), 179-185.

Ryalls, J. (1996). *A Basic Introduction to Speech Perception* (San Diego, Calif. : Singular Pub. Group).

Schouten, H.E.M. (1985). “Identification and discrimination of sweep tones,” Percept. Psychophys. 37, 369-376.

Sergeant, R.L. and Harris, J.D. (1962). “Sensitivity to unidirectional frequency modulation,” J. Acoust. Soc. Am. 34, 1625-1628..

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

Shen, X., Lin, M. and Yan, J. (1993). “F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3,” J. Acoust. Soc. Am. 93, 2241-2243.

Shower, E.G. and Biddulph, R. (1931). “Differential pitch sensitivity of the ear,” J. Acoust. Soc. Am. 3, 275-287.

Sundberg, J. (1979). “Maximum speed of pitch changes in singers and untrained subjects,” J. phon. 7, 71-79.

Stagray, J.R. and Downs, D. (1993). “Differential sensitivity for frequency among speakers of a tone and a non-tone language,” J. Chin. linguist. 21, 144-163.

Stevens, K. (1998). *Acoustic Phonetics* (Cambridge, MA: The MIT Press).

Stevens, K.N., Liberman, A.M., Studdert-Kennedy, M., and Öhman, S.E.G. (1969). “Crosslanguage study of vowel perception,” Language and Speech 12, 1-23.

Stevens, S.S. and Volkman, J. (1940). “The relation of pitch to frequency: a revised scale,” American Journal of Psychology 53, 329-353.

Thorsen, N.G. (1980). “A study of the perception of sentence intonation – evidence from Danish,” J. Acoust. Soc. Am. 67, 1014-1030.

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

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

Vaissiere, J. (2004). "Perception of intonation," *Handbook of Speech Perception* (Eds: Pisoni, D.B. and Remez, R. E.; Oxford: Blackwell).

Wales, R. and Taylor, S. (1987). "Intonation cues to questions and statements: how are they perceived?" Language and Speech 30, 199-210.

Wang, W.S.-Y. (1976). "Language change," *Origins and Evolution of Language and Speech* (Annuals of the New York Academy of Sciences, 280, 61-72) (Eds: Harnad, S.R., Steklis, H.D. and Lancaster, J.; New York: New York Academy of Sciences).

Wang, Y., Jongman, A. and Sereno, J.A. (2003). "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," J. Acoust. Soc. Am. 113(2), 1033-1043.

Wang, Y., Spence, M.M., Jongman, A. and Sereno, J.A. (1999). "Training American listeners to perceive Mandarin tones," J. Acoust. Soc. Am. 106, 3649-3658.

Wong, P.C.M. and Diehl, R.L. (2003). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," Journal of Speech, Language, and Hearing Research 46, 413-421.

Xu, Y. (1994). "Production and perception of coarticulated tones," J. Acoust. Soc. Am. 95(4), 2240-2253.

Xu, Y. (1997). "Contextual tonal variations in Mandarin," J. phon. 25(1), 61-83.

Xu, Y. (2004). "Understanding tone from the perspective of production and perception," Language and Linguistics 5(4), 757-797.

Xu, Y., (2008). "Timing and coordination in tone and intonation -- An articulatory-functional perspective," Lingua 119, 906-927.

Xu, Y., and Sun, X. (2002). "Maximum speed of pitch change and how it may relate to speech," J. Acoust. Soc. Am. 111, 1399-1413.

Yuan, M. (2009). *Exploitation of Effective Temporal Cues for Lexical Tone Recognition of Chinese*, Ph.D. Dissertation, The Chinese University of Hong Kong.

Yuen, K.C.P., Yuan, M., Lee Tan, Soli, S., Tong, M.C.F. and van Hasselt, C.A. (2007). "Frequency-specific temporal envelope and periodicity components for lexical tone identification in Cantonese," Ear & Hearing 28, 107s-113s.

Zheng, H., Peng, G., Tsang, P. W-M., & Wang, W. S.-Y. (2006). "Perception of Cantonese level tones influenced by context position," *Proc. SpeechProsody 2006.*