SCUOLA NORMALE SUPERIORE DI PISA

DOCTORAL DISSERTATION

A STUDY ON FORM AND FUNCTION OF PROSODY BASED ON ACOUSTICS, INTERPRETATION, AND MODELLING

-WITH EVIDENCE FROM THE ANALYSIS BY SYNTHESIS OF MANDARIN SPEECH PROSODY

BY

ZHI NA

PISA, ITALY

Nov 2012

Abstract

An analysis-by-synthesis study on Mandarin speech prosody is conducted in the present dissertation. The features of Mandarin speech prosody are discussed by focusing on two salient aspects: the function of prosody and the form of prosody. The study attempts to find a plausible way in which the two aspects can be mapped onto each other through the functional analysis of prosody and the multi-level formal representation. The form of Mandarin speech prosody is a complex F0 picture due to the simultaneous uses of pitch contours by both lexical tones and sentential intonation. The phenomenon of tone sandhi in speech context triggers more puzzling issues when researchers are confronted with the acoustic form of Mandarin prosody. The functional use of prosody in Mandarin speech concerns: at the lexical level for word identity (Tone1, Tone2, Tone3, Tone4, and Tone0); at the sentential level for prominence marking (sentence accents) and the indication of prosodic boundaries (intonation boundary tones). In the present study, the analysis of prosodic function at the two levels provides a basic framework in coding the surface melodic form of Mandarin prosody, which consists of pitch contours in tonal units and boundary tones at the beginning and end of intonation unit. For the formal representation of Mandarin speech prosody, the surface F0 contour of each utterance is coded into a sequence of INTSINT symbols, and subject to the Prozed tool for speech synthesis. It is shown that the synthesized stimuli derived from the symbolic coding can closely follow the melodic features and correctly express the prosodic function of the original Mandarin utterances. The present study employs acoustic data, symbolic coding, and speech synthesis for the derivative mapping between prosodic function and form, which aims to interpret the complex prosodic phenomenon, and provide an insight for the annotation and analysis of Mandarin speech prosody.

Acknowledgements

There are no words to express my gratitude to my advisor, Prof. Pier Marco Bertinetto, whose support has been crucial during all these years in Italy. Without his constant and patient support and guidance, very little of this thesis could have been written. His support is not limited to this thesis, however, as he constantly provides me with very detailed comments and suggestions about submitting papers and attendance to many international conferences. His careful and strict attitude in pursuing for precision in academics has greatly influenced me. Moreover, he has introduced me to the linguistics academia, and provided crucial help in getting me sponsored as visiting student in France.

I would like to sincerely thank Prof. Daniel Hirst of CNRS Laboratoire Parole et Langage of Aix-Marseille Universit éin France, a very kind and patient professor in guiding me with many ideas in the prosodic research. He often encourages me to explore new proposals, and use evaluation results to improve the attempts. His tutorials in conferences and summer schools have provided me with a lot of inspirational ideas in my research. I also express my sincere gratitude for the friendship with Prof. Hirst and his wife, Mrs. Yvette Hirst. They are very nice people who have taken good care of me during my several times' study in France. Their encouragement has provided me with great confidence in completing the dissertation.

I am also very grateful to Prof. Giovanna Marotta of Universit àdi Pisa, whose course on *Phonetics and Phonology* has provided me with a wide range of essential knowledge for the prosodic study. She has great interest in Chinese language and has encouraged me a lot in my research study and completion of the dissertation.

My gratitude also goes to Prof. Alessandro Lenci, Prof. Francesco Del Punta, Prof. Lavinia Merlini Barbaresi, and the committee members of my PhD dissertation, such as Prof. Claudio Ciociola, Dr. Chiara Celata, Dr. Luca D'Onghia and Dr. Shu Shi.

I would like to thank the staff in the laboratorio di Linguistica, especially Chiara Bertini, with her hard work in data analysis, we have collaborated with Prof. Bertinetto for three international conference papers. I am also very grateful to Chiara Celata, Irene Ricci, Valentina Bambini, and the PhD students: Anna Lentovskaya, Anna Alexandrova, Clementine Pacmogda, Danica Pusic, Emanuela Sanfelici, Emanuele Saiu, Giovanna Lenoci, Marta Ghio, Michelangelo Falco, Luca Ciucci, Luca Pesini, Simona Di Paola, Rosalba Nodari, etc. I would also like to extend my gratitude to the administrative staff of Scuola Normale Superiore di Pisa, especially, Ambra Vettori, Elisabetta Terzuoli, Emanuela Brustolon, Mario Landucci, and Luca De Francesco, who have all provided me with a lot of help during my four years' study and living in Italy.

My acknowledgment also extends to my Italian and international friends, Annamaria Peri, Chiara Tramontani, Chiara Viola, Federico Camelo, Orso Maria Piavento, Silvia Milano, Celine Delooze, Chen Qi, Dalazi Moustafa, Daniela Caceres Jilabert, Gao Ji, José Sandí Morales, Laura Lahaix, Lu Lingyan, Lu Xinyang, Matthias Lepucier, Maxime Chapuis, Nika Shamugia, Rafael Gaune Corradi, Roxana Blaga, Sevgi Dogan, Sunghye Baek, Takashi Araya, Xie Mingguang, Yohana Lévêque, etc., with whom I have spent much great time in knowing, appreciating and enjoying Italian culture and other international cultures.

I would like to express my gratitude to a kind Chinese family in Pisa: Aunt Yanjun, Uncle Chunzeng, and Sister Feifei who have offered me tremendous help and support during my four years' abroad life. Their kindness makes me feel quite warm and touched, especially in each Chinese Lunar New year that I have spent with them in the past four years.

Last but not least, my deep appreciation goes to my beloved family members in China, my mother, Wang Yajun, my father, Xuan Zhaorui, my husband, Gao Runpeng, and my cousin, Wang Ying. They have provided me with enormous love and support. Their encouragement is always a great impetus for my completion of the long-journey writing of the dissertation.

Publications

- Zhi, N., Hirst, D. and Bertinetto, P. M. 2010. Automatic analysis of the intonation of a tone language -applying the Momel algorithm to spontaneous Standard Chinese (Beijing). *Proc. 11th Interspeech Conference*, Makuhari, Japan.
- Zhi, N., Bertinetto, P. M. and Bertini, C. 2011. The speech rhythm of Beijing Chinese, in the framework of CCI. *Proc. 17th International Congress of Phonetic Sciences,* Hong Kong, China, 2316-2319.
- De Looze, C., Zhi, N., Auran, C., Cho, H., Herment, S. and Nesterenko, I. 2012. Providing linguists with better tools: Daniel Hirst's contribution to prosodic annotation. *Proc.6th International Conference of Speech Prosody*, Shanghai, China, 615-618.
- Bertinetto, P. M., Ricci, I. and Zhi, N. 2010. Le nasali sorde dell' ayoreo: prime prospezioni. *Atti del 60 Convegno Nazionale AISV*, Napoli, Italia.
- Bertini, C., Bertinetto, P. M. and Zhi, N. 2011. Chinese and Italian speech rhythm, normalization and the CCI algorithm. *Proc. 12th Interspeech Conference,* Florence, Italy, 1853-1856.
- Bertinetto, P. M., Bertini, C. and Zhi, N. 2012. A comparision of accentual features between Chinese and Italian speech. *Proc.6th International Conference of Speech Prosody*, Shanghai, China, 520-523.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background of prosodic study of natural languages

Prosody refers to the speech characteristics beyond the text level, thus it is often considered as the "suprasegmental" aspect of spoken language. In everyday speech, prosody together with the lexical-syntactic information of spoken text contributes to the interpretation of speech, which is illustrated by an equation in Hirst (2011a) as, *Speech = Text + Prosody.* In social interactions, prosody performs salient functions for the need of communicative exchange between speech participants.

According to the analyses of a variety of natural discourse data in the book, *The Music of Everyday Speech* (Wennerstrom 2001), the way in which prosody contributes to the processing and understanding of spoken language could be found in various aspects. For instance, prosody indicates the coherence of speech by means of pause, length, and other intonational cues at boundaries of speech units. Such function of indicating the finality/non-finality of speech is widely used in the turn-taking between participants of the conversational interaction; Prosody also contributes to the illocutionary force of speech by means of certain intonation patterns, which can be observed from the speech productions of lawyers in the court room, where intentional speech acts are achieved with particular intonational cues. Prosody can also provide evidence in understanding the subtle features of human social behaviour with the indication of "tone concord" which refers to the synchronized matching of speakers' pitch range, key level and rhythmic structure in a conversation, in particular, speech participants adopt the similar pitch level in matching with each other at one’s completion of the turn and the other’s turn-taking. It has been observed that such agreement in “tone concord” can be achieved by conversants in a harmonious context of interaction, whereas in less supportive situations, there occurs the “concord breaking” conveyed by discordant responses from conversants, i.e., using high key level competitively as a power struggle, or breaking down the regularity of rhythmic intervals in the interactive speech exchange; Prosody is also an essential non-verbal cue in expressing the emotions and attitudes of speakers. It is found that story-tellers often make full use of prosody to avoid monotonous speech by means of variating pitch, loudness, length and speech rate, etc. to achieve the purpose of enriching the narrative production with strong emotional colour in attracting the attention of listeners; Due to the importance of prosody in spoken language, it is proposed by Wennerstrom (2001) that the prosodic acquisition should be highlighted in the materials of second-language teaching.

Marotta (2008) highlighted the sociolinguistic status of prosody, as she held that prosodic elements such as intonation serves as an important socio-phonetic cue, with which listeners can perceive and distinguish the different varieties of the same language, as found in Italian (Marotta & Sardelli 2003, 2007; Marotta *et al*. 2004). For the phonological status of prosody, she proposed to draw a boundary between the language core grammar level and the pragmatic level, i.e., the sociolinguistic use.

Cresti (1995) discussed the discourse function of prosody, and proposed “language as an act”, in which the intonation of an utterance serves the function of a “systematic marker of informational unit, whose fundamental principle is establishment of illocution”. From her observation of an Italian spoken corpus, she summarized that the (topic)-comment pattern can be regarded as the basic pattern in Italian oral speech. Topic is optional as it mainly carries given information, while comment is privileged and sufficient in performing the main function of speech act, as it usually corresponds to the predicate or a whole sentence, conveying new information.

Prosody, as a compositional term, has a broad definition, and researchers may have different understandings on the subject. For some researchers, prosody reflects speakers' emotion and attitude. According to Pike (1945: 10), the distinctiveness of intonational meaning is not defined by the grammatical sentence type, but rather by the attitude of the speaker at the time when the utterance is given. Bolinger (1989:1) also emphasized the paralinguistic function of intonation and defines intonation for its function in reflecting speakers' inner states, and as a "nonarbitrary, sound symbolic system with intimate ties to facial expression and bodily gesture".

Hirst & Di Cristo (1998) held that prosody conveys semantic meanings and interpersonal functions together with the lexical and syntactic components of an utterance. Ladd (1996) summarized the linguistic and paralinguistic facts of intonation, and regarded intonation as "the use of suprasegmental phonetic features to convey 'postlexical' or sentence-level pragmatic meanings in a linguistically structured way". In his understanding, intonation has three main compulsory features, namely, (a) "suprasegmental"; (b) "'postlexical' or sentence-level pragmatic meanings"; and (c) "linguistically structured".

From the above review on the literature of prosody, one can note that the prosodic study of natural languages can extend to a very broad scale, due to the fact that prosody is the general non-verbal aspect of oral speech, which could encompass a variety of linguistic and paralinguistic phenomena. Hirst (2001) claims that a considerable number of factors contribute to a language's prosodic features, and they could be universal, language specific, dialectal, individual, syntactic, phonological, semantic, pragmatic, discursive, attitudinal, emotional, and the list is obviously not complete.

In the present study, I shall focus the discussions of prosody onto its linguistic functional aspect. The dissertation consists of six chapters. Chapter 1 introduces the background of speech prosody, with discussions from both the functional aspect and formal aspect of prosody. Moreover, the chapter reviews the previous relevant studies on Mandarin speech prosody, and presents the objective of the present study by using speech synthesis technology in evaluating the proposal of form-function mapping. Chapter 2 is a review on the important related literature on speech prosody, including the classical studies on intonation of non-tonal languages by the British, the Dutch and the American school. The review aims to provide an overview of background information of prosodic study. Chapter 3 presents the details of the two Mandarin speech corpora employed in the present study: a spontaneous dialogue corpus and a read speech corpus. Discussions of the labeling methodology of the two corpora and the selection criteria of speech data are presented. In Chapter 4, detailed discussions are given on three important functional components of Mandarin prosody: lexical tones, intonation, and accents. The three factors closely interact in contributing to the prosodic form of Mandarin speech, and together play a salient role in the linguistic functional aspect. Chapter 5 discusses the results of speech synthesis derived from the symbolic representation of prosodic form with the INTSINT annotation system. The principle consists in finding a plausible way in which the physical facts of prosodic form can be related to the functional aspect of prosody. The proposal is tested through the implementation of the Prozed tool, with the predicted symbolic coding of 60 Mandarin utterances synthesized into stimuli, and compared to the original utterances on the melodic features. It is shown that satisfactory synthesized results can be derived: the stimuli can closely follow the surface contour movement of the original ones. Chapter 6 summarizes the mutual interrelations among the three factors in Chinese speech prosody, and concluded that the phonetic makeup of speech could be correlated to the functional components of prosody.

1.1.1 Function of prosody

According to Hirst (2005), many of the linguistic functions of prosody are nearly universal, as in all languages prosody is an essential part for the word identity at the lexical level (via tone, accent and quantity); above that, prosody can highlight the key information by marking out certain words from the background for expressing prominence; moreover, prosody can indicate the finality and non-finality of speech unit with boundary tones. The linguistic function of prosody in natural language can be summarized according to its contribution at two distinct levels, as illustrated in Figure 1.1 from Hirst & Di Cristo (1998: 4). Such distinction of prosodic function at lexical and non-lexical levels can provide a basic framework for the comparative description of prosodic systems across languages.

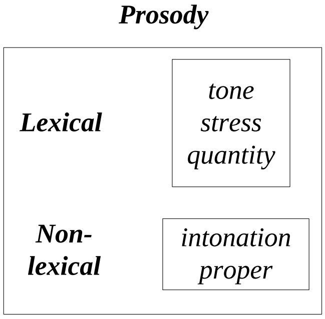


Figure 1.1: Prosodic function at the lexical level and non-lexical level

1.1.1.1 Prosodic function at the lexical level

For the prosodic contribution at the lexical level, there are language-specific ways in the use of prosodic parameters for the distinction of lexical identity. In a tone language, like Mandarin Chinese, the pitch contrast on each syllabic unit contributes to the lexical distinctive function. There are four distinct tones in Mandarin: the first tone (Tone 1), the second tone (Tone 2), the third tone (Tone 3), and the fourth tone (Tone 4). The four lexical tones in Mandarin can be marked in written form with an iconic diacritic above the nucleus of the associated syllable, i.e., zhī, zhí, zhǐ, zhì. Therefore, the identical syllable, when associated with four different tones, leads to four completely different word morphemes, as seen in the following Table 1.1.

Syllable	zhī	zh í	zhǐ	zh ì
Tone	Tone 1	Tone 2	Tone 3	Tone 4

Morpheme	Noun. *"knowledge"*; Verb. *"to know"*	Adj.*"straight"*	Noun.*"paper"*	Adj.*"intelligent"*

Table 1.1: Identical syllable with four different tones

For the tonal representation, Chao (1930) proposed a numerical notation system of five scaling points, 1 to 5 corresponding respectively to low, half-low, medium, half-high and high position within a speaker's normal pitch range. According to native speaker's perception, the four lexical tones can be represented with a succession of numerals marking respectively the significant points of the pitch contour of each tone. Accordingly, Tone 1 is transcribed as [55], Tone 2 as [35], Tone 3 as [214], and Tone 4 as [51], indicating the high-level pitch form of T1, the high-rising form of T2, the low-falling-rising form of T3, and the high falling form of T4. The idealized form of the four tonal shapes and heights (in relative values) could be represented in Figure 1.2:

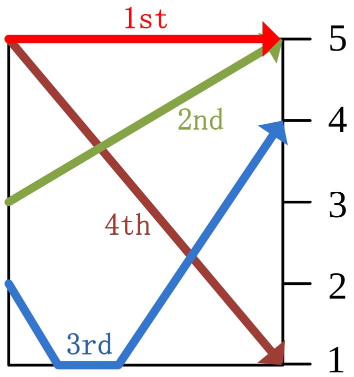


Figure 1.2: The four lexical tones in Mandarin

In most cases, each monosyllabic morpheme of Chinese aligns with one of the above citational tones. There are also morphemes which carry no tones, or so called 'neutral' tones. According to Chao (1968), such neutral tones occur on two types of syllables. First, there are inherently toneless syllables, which are usually grammatical morphemes, such as particles. Second, there are tonal syllables whose canonical tonal features are intentionally neutralized by speakers, especially, when such lexical morphemes are located on the second syllables in disyllabic words, such as the kinship term *ma1 ma0* (‘mother’), where the second repetitive syllable *ma* (with citational Tone1) is actually neutralized. Besides the above two types of toneless syllables, neutral tone also occurs in certain lexical words, where the tone of the second syllable is neutralized, and such lexical words usually form a minimal pair with the identical phonotactic ones without tonal neutralization, such as *dong1 xi0* (NOM ‘thing’) vs. “dong1 xi1” (NOM/ADJ ‘east-west’). These neutral-tone syllables are phonetically analogous to unstressed syllables in English, with features such as vowel centralization, short duration and weak intensity. According to Yip (2002), all the neutral-tone syllables are regarded as default unstressed syllables in contrast with the full-tone syllables in Chinese speech.

In a non-tonal language, such paradigmatic opposition of tonal pitch does not contribute to the lexical identity, for instance in English or Italian. The two languages employ quantity and accent, instead of tone, in making the distinctions of lexical identity. There are also languages in which prosody is not required to perform any function at the lexical level. In Hirst (2004), it was mentioned that modern standard French does not distinguish lexical identity according to either accent, or tone, or quantity; thus, the perceptual prosodic contrast in French fluent speech is not derived from the underlying lexical level, but rather from the surface sentential level. The discussion of prosodic typology at the lexical level can be seen in the study of Hyman (2009), in which it was proposed that word-prosody should be conducted in a property-driven approach and the author argued against the use of the term “pitch-accent” for a language like Japanese, which has an intermediate property of accentual tone between English (a stress language) and Mandarin (a tone language).

1.1.1.2 Prosodic function at the non-lexical level

For the prosodic contribution at the non-lexical level, there exist a large quantity of discussion in the literature on the language intonation systems. Such kind of literature has especially flourished in English language, inspired by the pedagogical purpose of

English-as-a-second-language (ESL) acquisition. It can be seen in many earlier works, such as Palmer (1922), Armstrong & Ward (1926), Kingdon (1958), Cystal (1969), Halliday (1967, 1970), O'Connor & Arnold (1973) and etc., which formed the well-developed approach of the *British school* in the study of intonational functions and forms.

Despite the fact that intonation may present particular forms in different languages, it is generally agreed that the linguistic functional role of intonation in expressing prominence, called the *weighting function* of intonation by Gårding (1989), and the function of marking connective and demarcative features in speech, called *grouping function* of intonation (ibid), are nearly universal in all languages.

In this study on Mandarin prosody, the nature of abstract prosodic properties, both at the functional and formal level will be explored. The discussion on prosodic function is focused on the lexical distinctive function of tones and the weighting and grouping function of intonation in connected speech. The functional analysis of prosody will pave the way for the interpretation and modelling of the acoustic data of prosodic form. The plausibility of the form-function mapping proposal is evaluated through the Momel-INTSINT system (Hirst & Espesser 1993, Hirst and Dicristo 1998, Hirst 2005). The aim is to examine the underlying and potential factors, which contribute to the surface phenomena, by means of combining acoustics with interpretation, and evaluating proposals with the synthesis result.

In the following section, I shall discuss the physical events related to the form of prosody.

1.1.2 Form of prosody

The form of prosody mainly concerns the salient sound events perceived by listeners, which are indicated by parameters such as pitch and length, which will be discussed in details in the present section, as the two serve as important correlates of prosody in contributing to the linguistic functions in speech.

1.1.2.1 Pitch

Pitch is acknowledged as the primary parameter in listeners' perception of speech melodic form, which generally consists of intonation and also lexical tones in certain languages. Previous studies on intonation forms, such as the British School, greatly depend on listeners' impression of the “rising” or “falling” tendency of pitch contour movement especially at the sentence end, which are believed to be associated with the expression of modality, in particular, statement is stylized with a falling pitch contour, while question is expressed with a rising pitch pattern. Pitch, as the primary acoustic parameter in the perception of intonation, was also discussed in Hirst & Di Cristo (1998: 4).

In a tone language like Mandarin, speech melody is a combined carrier of pitch information at the lexical level as well as the sentential level. The pitch movement of the component syllables in each utterance contributes to the basic melodic form of a Mandarin utterance, while the intonation pattern at the sentential level interact with the lexical pitch contours, which can be in a way of either modifying (broadening or narrowing) the pitch range of the local lexical contour, or adding intonation boundary tones at the initial and final boundaries of the utterance, or manipulating or even changing the citational pitch movement of lexical tones in a competitive way. These two phonological features are “phonetically intertwined in the tempo and pitch contour of an utterance” (Beckman 1986: 28).

Due to such multiple uses of pitch by both lexical tones and intonation, the formation of speech melody in Mandarin, and the way in which tones and intonation interact with each other is still a debated issue, which have been discussed in many earlier studies, as seen in the section §1.2.2.

The physical measurement of pitch is fundamental frequency (F0) which corresponds to the vibration rate of the vocal cord within the larynx, where the circular tenoid muscles control the open and closeness of the vocal cord in quick succession of the airflow. The F0 contour in speech can be modeled with a synthesis system, such as Momel (MOdelling

MELody) algorithm, which was first proposed by Hirst & Espesser (1993). According to the basic principle of the Momel modelling, the raw F0 contour of an intonation unit is analyzed as a product of two components: a **macroprosodic** component and a **microprosodic** component. The former corresponds to the global intonation pattern of an utterance, which is the essential component in contributing to the linguistic function of the contour; while the latter refers to the local effect of pitch from the individual segments, i.e., voiced/unvoiced, obstruent, sonorant, vocalic and etc., which manifest as deviations from the marcroprosodic curve. The Momel algorithm takes the information of raw F0 contour as input, and provides a modelling output, which is a smooth and continuous contour quadratically interpolated by the "target points" detected by the algorithm. The major phonetic events on the raw F0 contour can be closely followed by the stylized modelling of Momel algorithm.

As the target points of Momel modelling are derived solely from the fundamental frequency curve, without access to the information of neither speech production nor speech perception, it has a 'theory-neutral' nature, which allows it to be employed by different approaches in deriving the melodic representation of speech, for instance, it has been used as the first step with the Fujisaki model (Mixdorff 2000), the ToBI system for English (Wightman & Campbell 1994, Maghbouleh 1998), the K-ToBI system for Korean (Cho 2009, Cho & Rauzy 2008), and the INTSINT coding system for a variety of languages: English (Auran, 2004), French (Nicolas 1995, Bertrand 1999, Portes 2004), Italian (Giordano 2005), Catalan (Estruch 2000), Brazilian Portuguese (Fernandez-cruz 2000), Venezuelan Spanish (Mora Gallardo 1996), Russian (Nesterenko 2006), Arabic (Najim 1995), IsiZulu (Louw & Barnard 2004), and Korean (Kim *et al*. 2008, Cho 2009), as well as the intonation of L2 speakers, in particular, the French learners of English (Tortel 2008), and English learners of French, with the comparison between native speakers and learners (Herment *et al*. 2012).

An attempt of applying the Momel algorithm in the automatic modelling of a tone language, such as Mandarin, was also conducted, and the evaluation of the resynthesized 100 spontaneous utterances was reported in the study of Zhi *et al*. (2010). According to the study, 100 utterances from a spontaneous Mandarin speech corpus were synthesized with the values from the automatic coding and quadratic spline interpolation with the Momel system. The synthetic output of each utterance was evaluated according to the judgment of the first author (a native speaker) on the perception of lexical tones and intonation, as distinguished into three categories (Zhi *et al*. 2010):

a) lexical tone error - one or more of the lexical items in the utterance was perceived as being pronounced with the wrong lexical tone.

b) intonation error - the resynthesised utterance was perceived as being produced with the correct lexical tones but with a perceptibly different intonational meaning.

c) correct tone and intonation - even if the utterance did not sound exactly the same as the original, there was no perceptible lexical or intonational difference of meaning.

According to the evaluation with the perceptual judgment, it was found that the majority of errors in synthesized utterances was lexical tone errors, and needed to undergo manual corrections by means of adding or deleting targets points, or readjusting the locations of existing target points.

Table 1.2 shows the number of syllables which were manually corrected from the total corpus. A, B, C and D represent the four speakers of the study: A and B are female speakers, C and D male speakers. T1, T2, T3, T4 and T0 in the left columns refer to the tone types in Mandarin speech. The numbers in the table (outside the bracket) refer to the number of syllables of the given tone type produced by the given speaker, while the numbers within brackets indicate the number of syllables that required manual correction, due to error in automatic pitch detection.

	A	**B**	**C**	**D**
T1	63 (12)	73 (4)	47 (3)	23 (0)
T2	83 (12)	86 (13)	59 (10)	36 (8)
T3	52 (14)	62 (8)	41 (9)	41 (9)
T4	107 (22)	144 (19)	85 (7)	63 (10)
T0	57 (7)	53 (6)	35 (2)	23 (0)

Table 1.2: Number of syllables for each tone and for each speaker and the number of syllables that required manual correction (in brackets)

The evaluation result shows the particularly challenging task in the automatic modelling of Mandarin utterances due to the intertwined use of pitch by both lexical tones and intonation contour in the raw F0 data. However, attempt to improve the automatic modelling system of a complex tone language is expected to be far more fruitful than with languages with no syllabic tones.

In the present study, I further explore the melodic form of Mandarin prosody by integrating the Momel modelling algorithm with the INTSINT coding system (Hirst & Di Cristo 1998, Hirst *et al*. 2000) for the modelling of Mandarin speech. By deriving the prosodic form from the abstract symbols manually annotated at the surface phonological level representing prosodic functions, it is aimed to obtain the synthesized data which can closely captures the pitch features of the original utterance at the physical level, and contains all the necessary information, both lexical and sentential, for expressing the functions of Mandarin prosody.

1.1.2.2 Length

Another important feature of speech prosody is indicated by length, as “speaking is a time-dependent activity” (Fox 2000: 12). The different organization of timing events in a language gives the impression of its particular mode of speech rhythm. In classical studies on rhythmic features of speech, two basic rhythmic tendencies were distinguished as, the ‘machine-gun’ rhythm and ‘morse-code’ rhythm (Lloyd James 1940), which were categorized by Pike (1945) respectively as, “syllable-timed” and “stress-timed” languages. Such distinction is made on the basis of certain assumed temporal regularity in speech, in particular, syllable-timed language tendentially have the uniform syllable durations; while stress-timed language retain equal intervals between adjacent stressed syllables. The distinction of such two rhythmic types is characterized by the isochronic tendency of perceived timing events in different domains, i.e., syllable vs. inter-stress interval.

However, such assumed isochronicity is merely based on impressionistic judgments, which can be easily falsified with the objective measurement of the length of speech sounds. Therefore, such distinction of extreme temporal equality, indicated by the categorization of syllable-timed and stress-timed features were suggested to be abandoned by some linguists, such as Dasher & Bolinger (1982), Bertinetto (1989), and Levelt (1989), who proposed to consider alternative notions, and turn the study focus to the underlying phonological structure which governs the surface phonetic behavior.

In the study of Bertinetto (1989: 108-109), a number of factors which may contribute to the isochronic phonetic features, such as the phonological structure and the segmental makeup of the language, which are quoted as follows:

a) vowel reduction vs. full articulation in unstressed syllables;

b) relative uncertainty vs. certainty in syllable counting, at least in some cases;

c) tempo acceleration obtained (mainly) through compression of unstressed syllables vs. proportional compression;

d) complex syllable structure, with relatively uncertain syllable boundaries, vs. simple structure and well-defined boundaries;

e) tendency of stress to attract segmental material in order to build up heavy syllables vs. no such tendency;

f) relative flexibility in stress placement (cf. the so-called «rhythm rule») vs. comparatively stronger rigidity of prominence;

g) relative density of secondary stresses, with the corresponding tendency towards short ISI (interstress intervals), and (conversely) relative tolerance for large discrepancies in the extent of the ISI. This feature seems to oppose languages like English or German on the one side, to languages like Italian or Spanish on the other.

Bertinetto (1989) proposed the use of terms such as “compensatory” and “controlling”, indicating a continuum of “how languages diverge in terms of the coupling vocalic and consonantal gestures”. To examine the length variations of segments in speech, an algorithmic model, Control/Compensation index (CCI), was proposed in studying the rhythmic tendencies of natural languages, with details in Bertinetto & Bertini (2008, 2010).

The CCI model in its full realization (Bertinetto & Bertini 2010) is based on a two-level conception: level-I (phonotactics) and level-II (phrasal). The model takes into account both the speech durational behavior and the degree of phonotactic complexity. At level I, the following formula is proposed:

$$CCI = \frac{1}{m-1}\sum_{k=1}^{m-1}\left|\frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}}\right| \qquad \textbf{Formula 1.1}$$

In Formula 1.1, *m* stands for number of intervals (vocalic or consonantal, as separately considered), *d* stands for duration (in ms), and *n* represents number of segments within the relevant interval.

CCI is a modification of the PVI (pairwise variability index) model proposed by Grabe & Low (2002), in which the PVI was expressed in the following mathematic formula:

$$rPVI = \frac{1}{m-1}\sum_{k=1}^{m-1}\left|d_k - d_{k+1}\right| \qquad \textbf{Formula 1.2}$$

According to the comparison of the two formulas, it can be seen that CCI considers the number of segments composing each vocalic and consonantal interval. In particular, it takes into account whether a consonantal interval contains a single C, or a geminate, or a C cluster, and the same applies to the V intervals (with a single V, a long V, or a V sequence).

According to the formula, in ideal situations, a perfectly controlling language falls on or above the bisecting line, due to the identical C and V durational fluctuations or at least the stronger stability in the V intervals; while a compensating language should have substantial V-reduction, fluctuating more in the V than in the C component.

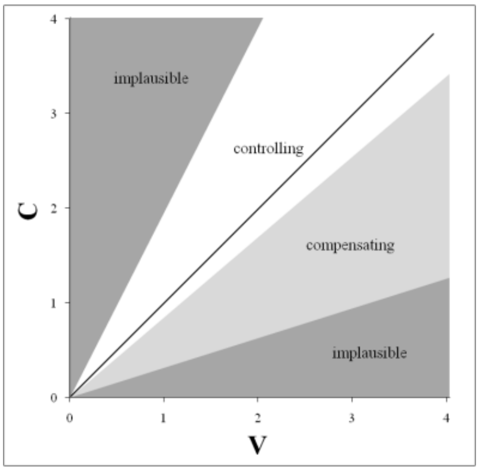


Figure 1.3: The major rhythmic tendencies in languages according to CCI framework

The CCI measure has been applied in the modelling of the level-I rhythmic features of Italian (Bertinetto & Bertini 2008, 2010), Mandarin (Zhi *et al*. 2011), and a comparative study on the rhythmic features between the two languages (Bertini *et al*. 2011, Zhi *et al*. 2011). The CCI model has also been applied in Brazilian Portuguese and German as seen in Bertinetto *et al*. (2012), where the degree of control/compensation tendencies of the two languages were discussed and also compared to the rhythmic features of Italian and Mandarin. The paper presented further inspections on the CCI construction, by discussing its conception in yielding the approximation of different languages' rhythmic tendencies.

In the discussion of the timing factors which contribute to the perceived rhythm of a language, one consideration is on the length variations of the constituent segments within the speech unit, another factor which deserves attention is the timing behavior of sentential accents. The Level-II of the CCI model examines the rhythmic feature at the sentence level, considering the coupling of two oscillators: the sentence-accent oscillator and the syllable-peak oscillator. By adopting suggestions in O'Dell & Nieminen (1999), the oscillators' coupling is expressed by the following mathematical formula, where *I* stands for duration of the inter-accentual intervals, *n* for number of syllable-peaks, ω_1 and ω_2 for the velocity of the two oscillators:

$$I(n) = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2} n$$

Formula 1.3

The formula relates the duration of the inter-accentual interval to the number of syllable-peaks comprised in it. A study of the role of sentence accents on the rhythm of Mandarin and Pisa Italian was conducted in Bertinetto *et al*. (2012).

According to a series of comparative studies between Mandarin and Italian on the rhythmic features with the CCI model, it is found that both languages exhibit a "controlling" rhythmic tendency, in particular according to the computations at Level I, as seen in the following Figure 1.4 from Bertinetto *et al*. (2012), where SBC represents the spontaneous Beijing Chinese, RPI for read Pisa Italian, and SPI for spontaneous Pisa Italian.

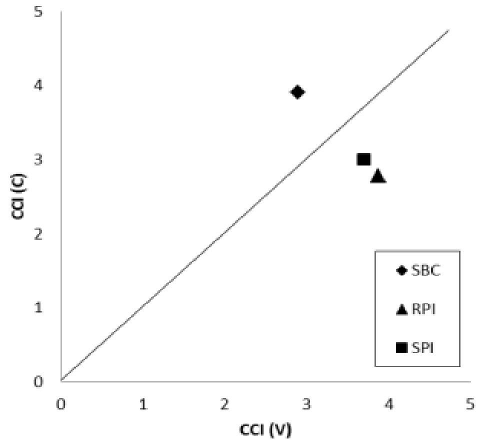


Figure 1.4: The rhythmic behavior of spontanous Beijing Chinese (SBC) in comparison with that of read Pisa Italian (RPI) and spontaneous Pisa Italian (SPI)

For the study of rhythmic behaviors at Level II, Mandarin, as compared with Italian, does not present salient sentence accents, its controlling rhythmic behavior thus can mostly be attributed to the language phonotactic organization at Level I, while the rhythmic accents at Level II plays a marginal role. In Italian, the sentence accents at Level II are part of prosodic competence, which together with the phonotactic structure at Level I contribute to the controlling feature of the language.

Speech rate is known to exert a crucial role in the rhythmic behavior of natural languages, as often noted in the specialized literature. In the study of Zhi *et al.* (2011) and Bertinetto *et al.* (2012), the factor of speech rate on the rhythmic tendencies of Beijing Chinese and Pisa Italian were checked at Level-I and Level-II, respectively. By means of dividing the speech productions of each language into three tempo-groups: slow, medium and fast, Zhi *et al.* (2011) found that the acceleration of speech rate exerts a tendentially linear effect on both V and C intervals of the two languages, as predicted for the rhythmic tendencies of controlling languages at Level-I. The study of Bertinetto *et al.* (2012) showed that the Level-II rhythmic organization of Pisa Italian although oriented towards the controlling behavior, appears to be variegated. The actual result is dependent on the speech rate, as intersecting the style factor (read vs. spontaneous). The study considered that: “the effect of speed increase is not uniform”, and “depending on the speed the mean dimension of the inter-accentual intervals allows varying degrees of internal syllabic flexibility” (Bertinetto *et al*. 2012). As for Beijing Chinese, the situation is fairly stable, the phonotactic components at Level-I do most of the rhythmic job, with the controlling behavior emerged, while the sentence accents at Level-II play a weak role, which leads to the stability of the syllabic oscillator. The relative SA-deafness of the Chinese speakers was reflected on the low inter-subjective convergence in the SA identification. More details of the two studies can be found in section §4.3.1.

An interesting study on the influence of speech rate on the acoustic and perceptual correlates of rhythm can be seen in Dellwo (2010). In his study, an experimental study on rate analysis was conducted with five languages: English, French, German, Italian and Czech. The result of perception experiments in the study revealed that speech rate is an important cue used by listeners in perceiving the different rhythmic categories, and it was observed that languages with a simpler syllable structure have higher speech rates than languages with more complex syllable structures. Dellwo proposed in his study that speech rate has an equal value as well as the durational variability of C- and V- intervals in distinguishing rhythmic classes, thus the effect of speech rate should be highlighted in the study of rhythmic modeling of different languages.

In the rhythmic study of timing in speech, the principle for the measurement of length is a debated issue, due to the fact that “speech is not readily segmentable … and the boundaries between sounds are seldom clearcut” (Fox 2000: 13). In connected speech, due to the existence of various sound changes, such as segmental deletions, insertions and shifts, the consistency in the annotation of sound phonemes and the acoustic measurement of segmental durations is essential. The details of manual and automatic labeling of Mandarin speech data will be discussed in the section §3.2.

1.2 Mandarin prosody: issues and previous approaches

1.2.1 Introduction of Mandarin

China is a country which has the largest population in the world, and which makes the Chinese language spoken by the largest group of people on the earth. The spoken language of China contains many regional varieties, which can be categorized into seven groups: the Mandarin group (the northern dialect), the Wu group (Shanghai and Zhejiang province), the Yue group (Cantonese), the Min group (Hainan, Fujian and Anhui provinces), the Hakka group (south Jiangxi and west Fujian provinces), the Xiang group (Hunan province) and the Gan group (Jiangxi and east Hunan provinces), with strong regional identities.

Such varieties of the Chinese language can be quite diversified, or even mutually unintelligible among citizens from different geographical regions. Therefore, there exist debates among linguists over the categorization of the language varieties, as whether which shall be regarded as the dialects of the same language, or rather separate languages. For most Chinese linguists and sociologists, the term *fangyan* (“dialect”) is preferentially used in marking such distinctions of the language spoken in different regions, due to the fact that despite of the variable speaking ways, these groups share the common use of written characters.

Among all the spoken variety, Mandarin is the official language of the mainland China, as decided by the National Language Unification Commission in 1932, and adopted till nowadays as the standard language used for education and national television broadcasts since the foundation of the People’s Republic of China in 1949. Mandarin is also called as *standard Chinese* or *Putonghua* (“common speech”), and it is based on the dialect spoken in the national capital, Beijing. As Mandarin is the common language shared by most speakers of the country, it has aroused the greatest interest by researchers in China and abroad. In the present discussion, the prosodic features of Mandarin are the central theme of the study, in particular, three salient components of Mandarin prosody: tone, intonation and rhythmic accent.

On the functional aspects of Mandarin prosody, the contributions of lexical tone and intonation are the most discussed topic. The functional use of sentence intonation in tonal languages is believed to be as free as those in non-tonal languages. However, the intonation study in a tonal language is much more complicated due to the binary use of pitch by both lexical tones and intonation.

In Chinese, the interactive relation between lexical tones and the “purely intonational use” (Peng *et al*. 2005: 231) of pitch prosody make the whole F0 picture extricate, and is a big interest of prosodic study.

1.2.2 Earlier studies on the interaction between tone and intonation

In the historical literature, the superimposed relation of Chinese lexical tones and intonation is described with a metaphor by Chao (1968: 28) as, “small ripples riding on top of large waves”, where lexical tones are regarded as “small ripples” in an ocean, which add on the “large waves” of the overall intonation melody, resulting in “an algebraic sum” of wave forms.

According to Chao (1933), there exist two possible ways in which tones and intonation interact on the pitch melodic contour, namely, *simultaneous addition* and *successive addition*. For the case of simultaneous addition, intonation is superimposed on lexical tones, which may lead to the overall raising or lowering of pitch level, and also the expansion and compression of pitch range. To be specific, when the direction of a lexical tone coincides with the overall tendency of the intonation pattern, its pitch value is enhanced, while in conflictive situations is the pitch value is decreased.

In terms of successive addition, Chao (1933) assumed that a terminal rise or fall, as an important intonational cue in expressing the interrogative or declarative form of an utterance, adds to the end of an utterance in a successive pattern. In this proposal, the pitch contour of a sentence is dominated by the composing lexical tones, while intonation just presents itself as a terminal rising or falling tone at the end of an utterance, leaving the lexical tones intact from the influence of intonation.

The interaction between tones and intonation in Mandarin was also discussed in Gårding (1984, 1987), within the "grid model" framework, which was originally proposed in Bruce & Gårding (1978) for the intonation studies of Swedish varieties. The model was later applied to explore the universal intonation features across languages. In the grid model, intonation pattern is "the global melody of a phrase or sentence to which the local pitch movements are subordinated" (Gårding, 1993: 36).

In the study of Chinese prosody with the grid model, the local pitch contour is specified by lexical tones of the constituent syllables of a sentence, and the local pitch range is a "grid", which consists of two parallel straight lines, a top line and a base line, which are defined by fitting most values of local peaks and valleys on F0 contour between the two lines. The grid reflects the global property of sentence intonation pattern. The direction of grid movement, such as the overall rising or falling, indicates the intonation of different sentence types. The turning points associated with lexical tones correspond to the peaks and valleys on the local contour, and a pivot often corresponds to the syntactic boundary in speech. According to Gårding (1984), a Mandarin statement usually presents a falling grid, while an interrogative sentence is often associated with a rising grid. The distance between the top line and base line of the grid is closely related to the sentence focus, as a grid becomes broadened under focus, and gets compressed at the post-focal region. In some extreme cases, a grid can become compressed into one straight line.

The grid model provides a schematic representation of the relationship between lexical tones and intonation in Mandarin. To be specific, the global property of intonation, represented by the range and direction of grid, influence the local and subordinated features of lexical tones. An overall review of the grid representation from Gårding can be represented as follows:

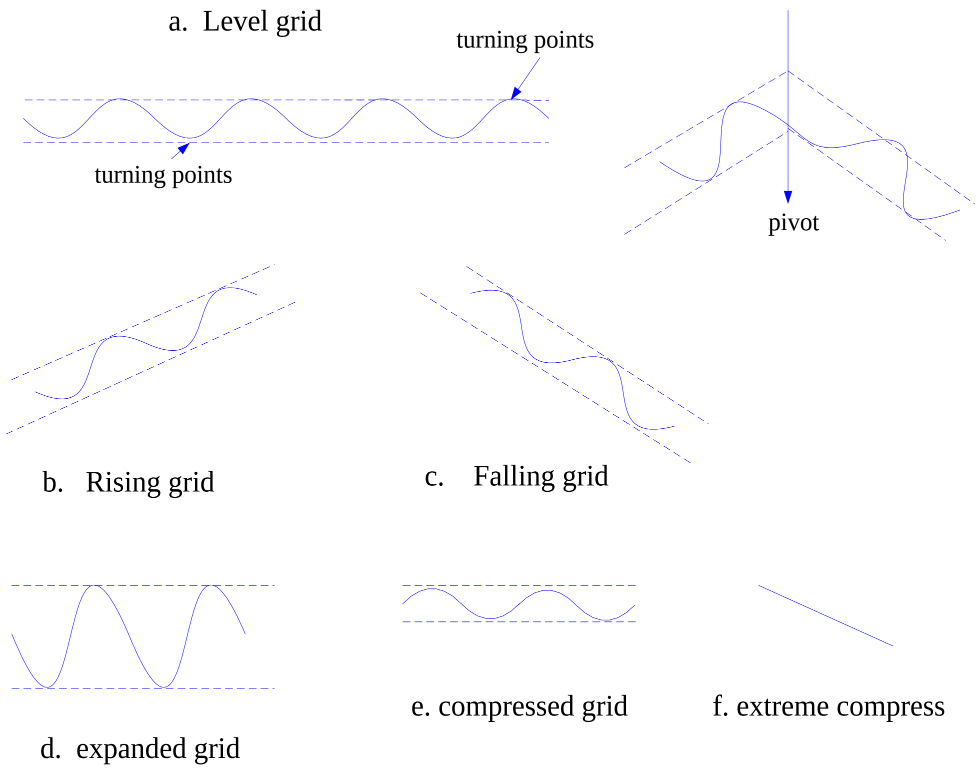


Figure 1.5: A sketch of the grid representation from Gårding (1983)

According to Gårding's study on Mandarin prosody, the interrelationship between tones and intonation on the melodic form of an utterance is represented by a grid model, where intonation associated with certain sentence type (interrogative sentences with a rising pattern, declarative sentences with a falling pattern) is reflected by the grid direction, and the intonational prominence is reflected by the expansion of grid, which becomes compressed at the post-focal portion; the local contour movement which fit themselves between the top line and bottom line of the grid are the lexical tones.

The components involved in the melodic form of Chinese prosody are also suggested in Shen's study (1990), according to which (ibid: 67-68), the following components contribute to the output of Mandarin speech prosody:

a) lexical tones at the basic morphemic level;

b) normal stress denoting possible lexical stress at the boundary of word or phrase level;

c) tune at the sentence level expressing sentence modality;

d) emphatic focus;

e) emotive intonation at the sentence level superimposed on the overall pitch movement of the utterance.

Besides the interactive relationship of lexical tones and intonation in Mandarin speech prosody, another big challenge in analyzing the melodic form of connected speech is the "unpredicted" and divergent lexical tonal phenomena. The varied forms of tones are called "tone sandhi", which are found when they are located in the tonal contexts, especially when the tone bearing syllables are juxtaposed with each other in connected speech. For instance, an underlying rising form of a Tone 2 may actually present a falling configuration, and some syllables even lose their pitch features partially or even entirely. The deviational degrees of tones are sensitive to factors, such as speech rate, speech style, the functional load of the word, the frequency of the word occurrence, the pragmatic context, etc.

However, there exist an interesting and paradoxical phenomenon, as observed in Tseng's study (1981) of spontaneous speech data: despite the fact that "a substantial portion of a sentence might be produced with partial acoustic information", listeners have no difficultly in understanding the meaning of such sentence, as the native subjects can often supply their knowledge of the language beyond the phonetic level in decoding the acoustic signals.

Yip (2002) also held that, for successful communication, tones must be perceived by listeners as linguistic objects rather than musical humming pitch. However, as the citational features of tones are often modified by various factors, how the speakers perceive the correct tones despite their distorted phonetic output is still a debated issue.

It needs to be reminded that the distinctive function of lexical tone is important in efficient mutual exchange; however, its canonical shape should not be overestimated. In real speech communications, native speakers do not have to only rely on the acoustic cues of F0 parameters for tonal recognition. The present study will show that the speech context, either grammatical or extralinguistic, is a salient factor in causing tone sandhi, as it provides a lot of anticipated information for the purpose of communication.

The speech context influence on tone sandhi can be accounted for in terms of speakers' balance between ease of production and ease of perception. To be specific, due to the speech production mechanism, speakers would use the "minimum production efforts" to reach the "minimization of perceptual confusion" (Boersma 1998). In some cases, the canonical form of tones can become redundant codings of information in speech exchange. According to Xu (2011), there exist both prosodic codes and syntactic codes for speech information transmission, and "if a function is already syntactically coded, there is no need to also encode it prosodically, and vice versa", although he admits that redundant coding can still be found in speech.

Tonal sandhi in connected speech, according to Chen (2004: preface), can extend to cover a large variety of related phenomena, such as "the allotonic variations, intonation effects, and the morphologically or syntactically conditioned tone changes". In the present study, the analysis on the phenomenon of tonal variation and its interaction with intonation on the melodic contour shall be explored by discussing a number of issues, such as the functional load of lexical prosody, the phonological rules of tone process, the potential domain of tonal co-articulation, the distinction between obligatory tone sandhi and optional tone sandhi, and the representation of prosodic functions and prosodic forms at the sentential level, and the plausibility in mapping the two through several intermediate phases.

1.2.3 Previous studies on rhythmic accent

The previous studies on the accentual features of Mandarin are much fewer as compared to those conducted on tones and intonation. Mandarin is not a "stress-language" like English or Italian, where accent plays a functional role for lexical contrast at the local prosodic level. Except for a few minimal word pairs which employ accent for lexical distinction, Mandarin does not require any stress pattern at the word level. However, at the sentential level, Mandarin is like any other language, where each syllable of an utterance is not uttered with the equal accentual status. There does exist the alternation of stressed syllables and unstressed ones in the perception of Mandarin speech, which can be accounted for with the words of Levelt (1989: 297) as follows:

> Speaking is a rhythmic process. A speaker organizes his utterance in patterns of stressed and unstressed syllables, and can assign various degrees of stress or accent to different syllables.

An early analysis of Chinese accents can be seen in the study of Chao (1968: 35), in which, accents were distinguished into three degrees, contrastive accent, normal accent and weak accent. Syllables with contrastive accent correspond to the ones marked out at the sentential level for the expression of emphatic focus, and often present features such as exaggerated high F0 value, great intensity and long duration; syllables with normal accent in an utterance are those which carry the perceptible tonal features, but do not convey contrastive intonation focus; weak-accented syllables refer to those with short duration, little intensity, and also centralized vowel features. Such syllables are produced with little articulatory effort and with neutralized tone features, which often correspond to grammatical suffixes and particles which are default toneless syllables, and the syllables which lose their original tone features when situated at the non-initial position of lexical word or phrase.

In the above distinctions of three accentual categories, the feature of accent is closely related to the tonal manifestation, in particular, weak-accented syllables are characterized by neutral tones, which are in contrast to the “full-tone” features of syllables with normal accent and contrastive accent. Such analysis of regarding “neutral tone” feature as a salient phonetic cue in the categorization of weak-accented syllable is also accepted in Yip’s study (2002), where it was suggested that neutral-tone syllables in Mandarin are all unstressed syllables, with features such as, short duration, weak intensity and vowel centralization, which are contrastive to those features of accented syllables often aligned with full tones. Beckman (1986: 104) also stated that atonic Mandarin syllables are analogous to the ones with weak accentual status in English, which has reduced vowel features and short duration. However, the idea of relating accent with tone is not endorsed by some linguists, such as Xu & Wang (2005): they argue that Mandarin is not a language with lexical contrast by accent, the neutral tone should be generally regarded as a tone feature rather than an unaccented phenomenon. Hyman (2009) also provide his distinction between stress and tone, stating that:

> Stress is a structural property in which syllables are metrically hierarchized as relative strong vs. weak, while tone is a feature property referring to contrastive relative pitch.

However, in actual data analysis, it is not easy to make a systematic distinction between tone feature and accent in Mandarin, as the acoustic manifestation of the two prosodic features are closely tied with each other. In the study of Kratochv ĺ (1968: 154), it was stated that accent is a multidimentional complex, and each of its respective dimension can be shared by other speech features. It is specified that the acoustic relationship between accent and tone is particularly acute in the case of Beijing Mandarin, and perhaps in all tone languages. He (1986:154) commented as follows:

> The whole complex of dimensions relevant to stress, namely, fundamental frequency, intensity level, and duration, is also relevant to the feature known as the tone.

Fox (2000) also stated that pitch is a constituent element in the phonetic manifestation of accent, and accent is a “cover-term” for a number of phonetic bases. The indication of accented sound is related with speaker’s increased effort, as the more muscular energy is spent from the producer, the more perceptual salience can be received by the listener. In Mandarin, it is normal that the syllable with more prominent accentual status is related with more canonical tone feature.

Duanmu (2007) analyzed the accentual features in Mandarin speech from the functional aspect, and he proposed an *Information-stress principle*, in which he assumed that the accentual distribution is closely related to the information load of words. It is thus fairly well understood that pronouns are unaccented, when they are either predictable or have already been mentioned in the earlier text; similarly, functional words are also unaccented as they do not carry much information. Sentence accent is often assigned with great flexibility, due to the variable information load of words, depending on the particular context. Duanmu believed that such information-stress principle is a universal rule across languages that high degrees of accent are given to words which convey the important information, while words with low degrees carry less important information.

Similar ideas are proposed in the study of Kochanski *et al*. (2003), according to him, in most cases, the syllable strength has correlation with their “semantic weight”. Speakers contribute intentionally more strength to salient information, while devote little energy to unimportant messages. In Mandarin speech, the greater the articulatory strength attached to a syllable, the more canonical feature its lexical tone shows on the global melodic contour, which indicates that a speaker is careful to produce a precise prosody by increasing the energy cost. Content words such as nouns and verbs are normally in prominent status, thus their aligned tonal features are often realized in canonical forms. While grammatical words such as prepositions and functional particles often show reduced features in the pitch shape of lexical tones.

In the studies of Kochanski & Shih (2001, 2003), a Soft-template Mark-up Language

(Stem-ML) model is proposed, which is defined on a physically motivated assumption originally raised by Ohala (1992), that speech is a compromised result of energy-saving effort and communicative clarity. When speakers produce utterances, they tend to reduce their muscular expenditure to a minimal extent, while at the same time they have to ensure that the speech is understandable with a low rate of communication errors. The prosody of a language is believed to be generated through a balance of two factors, energy expenditure and communicative demands, represented by an equation: effort + error. The preplanned physical effort in speech production guarantees a continuous pitch trajectory with smooth transitions between adjacent accents, as muscles cannot change their positions discontinuously. In Mandarin, the tones of syllables with greater strength dominate the general frame of a global F0 contour, while the tones of weaker syllables, where effort is minimized by the speaker, usually adjust themselves to accommodate to the global F0 pattern. Weak tones retain their original pitch features only when they are compatible with the global intonation tendency, while in contradictory context or in fast running speech, weak tones sacrifice their features partially or completely and present as interpolation between adjacent strong tones. The intonation contour is the result of "coordinative strategy", with strong tones playing the decisive role. As the actual F0 realization of a lexical tone is to a large extent related to the strength degree of its associated syllable, an idealized contour pattern of an utterance can be predicted based on the above strength assumption of syllables.

In the modelling of Mandarin utterances with Stem-ML, a number of tags are placed on syllables in an utterance with adjustable parameters, such as a strength parameter attached on each syllable. Those adjustable parameters can specify the position and stretching scale (shape) of a tag. The accented shape of a tag is a template, which is "soft" as it could adapt to the pitch tendency of neighboring accents. Speakers' muscular tensions determine the deviational degrees of tonal templates. The F0 value of each syllable can be derived by the current and nearby tonal templates and the strength parameters. The Stem-ML model believes that there are alternating metrical patterns in words, and speakers marks the hierarchical relations of words, phrases, and sentences in speech with the articulation strength.

In the above review of earlier studies, the close correlation between accent and tone are discussed. Syllables which carry "contrastive accent" and "normal accent" present "hyper-articulated" tone features; comparatively, the less important syllables with weak accent are either default neutral-tone syllables such as grammatical particles, or syllables with "hypo-articulated" or even neutralized tone features, as those situated in non-initial word position.

However, the phonetic nature of the accentual phenomenon is still an open question. Fox (2000: 143-144) summarized three facts of the confusing picture found in a systematic study of accents: (1) Accent has elusive and inconsistent phonetic correlates, as it is itself a "cover-term" encompassing a set of phonetic properties. Therefore, no adequate definition of accent as a phonological phenomenon can be based on phonetic criteria; (2) Accent cannot be approached systematically on a paradigmatic basis, as the occurrence of an accentual contrast in different domains may result in separate classifications of the phenomenon. Thus, there are no systems of mutually substitutable items for the accentual contrasts in different places; (3) Accent is multi-layered, and the different levels of accentual contrast could be hierarchically organized. In some languages, such as English, the nuclear accent at the intonational level is superimposed on the basic accentual pattern. Fox proposed that the study of accent should focus on its functional aspect, instead of the phonetic manifestation of surface accentual phenomenon.

According to Beckman (1986), accent has an organizational role in speech. The accentual feature and its correlation to the speech hierarchical structure have been widely discussed in studies of Metrical phonology (Liberman & Prince 1977, Hogg & McCully 1987, Goldsmith 1990). In the present dissertation, the "organizational" role of accent at the sentential level in Mandarin shall be further discussed.

At the sentential level, accent plays an important role in defining the hierarchical prosodic structure in speech and contributing to intonational prominence. In section §4.3.3 the rhythmic unit formed by the accented syllable and the following unaccented ones will be shown to be a potential domain where tonal coarticulation occurs. The idea of locating a certain domain where tone sandhi takes place was also discussed in previous studies, such as Chen (2004), where the unit was termed as *Minimal Rhythmic Unit*. However, no consensus has been reached on the specific rules defining the potential domain of tone sandhi, although it was generally believed that the prosodic and syntactic structure of speech plays a salient role.

In the present study, a perception experiment is conducted among native subjects in the identification of accents at the sentential level in Mandarin speech, and the organizational function of sentence accents is explored by analyzing the hierarchical structure of Mandarin utterances. The rhythmic unit chunked by sentential accents is discussed as the basic domain of tone sandhi process.

1.2.4 Recent studies: from data to modelling

The tendency of recent prosodic studies shows a practice from speech data to modelling. There are a large number of works devoted to the development of automatic tools for prosodic modelling, such as the Momel-INTSINT system (Hirst & Espesser 1993, Hirst *et al*. 2000, Hirst 2001, 2007), the PENTATrainer model (Xu 2005, Prom-on *et al*. 2009, Xu & Prom-on 2012), the Fujisaki model (1983, 1988), the Stem-ML model (Kochanski & Shih 2001, 2003), etc.

In the following subsections, two prosodic models, the PENTATrainer model and the Momel-INTSINT modelling system are reviewed in details.

1.2.4.1 The PENTATrainer system

According to Xu & Prom-on (2012),

> The significance of prosody research is not only the contribution to basic knowledge in speech science but also the advancement in speech

technology, particular speech synthesis.

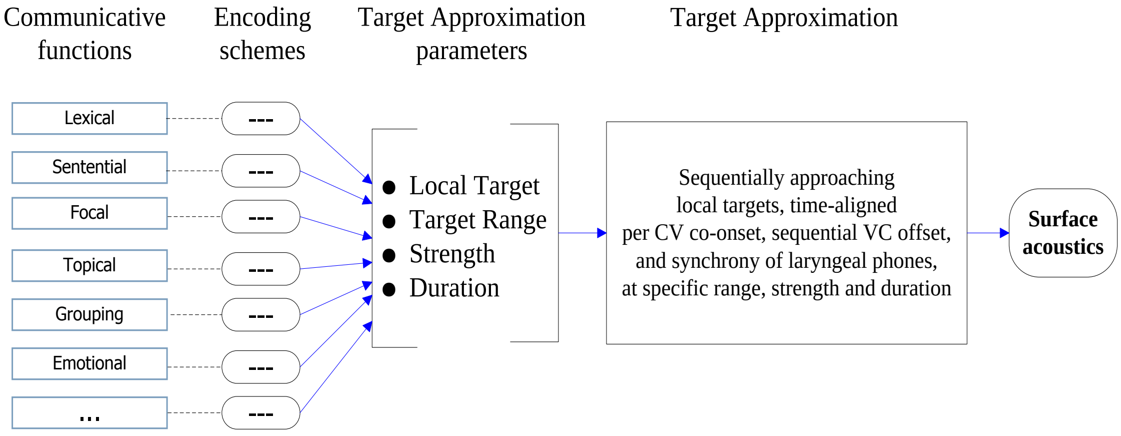
The Parallel Encoding Target Approximation (PENTA) proposal (Xu & Wang 2001, Xu 2005, 2011) was originally designed for the study of Mandarin speech prosody. In such an approach, human's articulatory mechanism and language's communicative function are two salient factors considered in exploring the speech features, as according to Xu's words (2007):

> Successful speech modelling can be achieved only if communicative functions and biophysical mechanisms are treated as the core rather than margins of speech.

He believes that message conveying is the essential speech function, as the message (either linguistic or paralinguistic) is deliberately expressed by a speaker to convey intentional meanings. These communicative functions are linked to the acoustic signal in speech through encoding schemes that directly control articulation.

a. Basic principle

The PENTA model is based on the assumption that the multiple layers of communicative functions are the driving force of prosody in speech, and are directly linked to the articulatory mechanism specified by encoding schemes. The surface F0 contour is a result of sequential tonal-target approximation, and is filled with deviational features due to people's physical constraint in articulation. A schematic sketch of the PENTA model from Xu (2009) is quoted as follows:


Communicative functions
Encoding schemes
Target Approximation parameters
Target Approximation
Lexical
Sentential
Focal
Topical
Grouping
Emotional
...

● Local Target
● Target Range
● Strength
● Duration
Sequentially approaching local targets, time-aligned per CV co-onset, sequential VC offset, and synchrony of laryngeal phones, at specific range, strength and duration
Surface acoustics

Figure 1.6: The proposal of parallel encoding of communicative functions and the parameters of articulatory dynamics in realizing the surface acoustics

In the above scheme, the various melodic features of speech are defined in the form of their communicative functions, such as lexical, sentential, focal, topical, grouping and emotional, etc. These multiple communicative functions are independent of each other and parallel in relations. Each function is realized though distinct encoding schemes, which specify the distinctive values of articulatory parameters, such as local pitch target, pitch range, prosodic strength and duration. Those articulatory parameters are all controllable elements, therefore, the local pitch target could be [high], [low], [rising], [falling] or [mid]; pitch range could be specified in height as [high], [low] or [mid], and in span as [wide], [narrow] or [normal]; strength could be [strong], [weak] or [normal]; duration could be [long], [short] or [normal] (Xu 2005).

The *target approximation* proposal is the basic principle of the PENTA model. It regards the surface F0 contour as a result of a goal-motivated process towards an underlying pitch target aligned with each syllable in speech. A sketch of the Target Approximation (TA) model from Xu (2009) is schemed as follows:

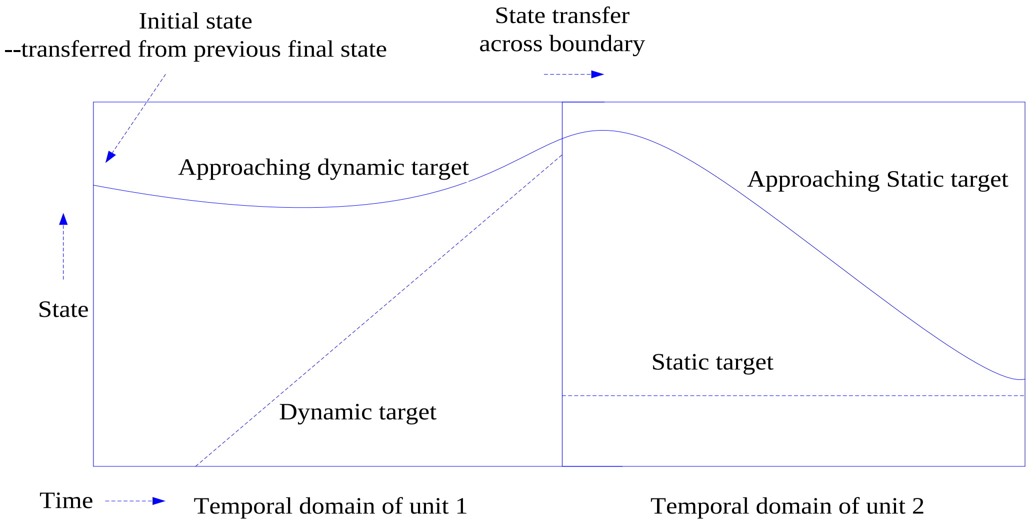

Initial state
--transferred from previous final state
State transfer
across boundary
Approaching dynamic target
Approaching Static target
State
Static target
Dynamic target
Time
Temporal domain of unit 1
Temporal domain of unit 2

Figure 1.7: Syllable-synchronized target approximation process

It is assumed that there is an underlying pitch target (represented by the dashed lines in the scheme) synchronously aligned with each syllable. Each pitch target could either be a static or a dynamic one and specified by two parameters, pitch height and contour slope. The surface F0 contour of a syllable is the result of a dynamic process approaching towards an underlying pitch target. Such approximation process starts at the syllable onset regardless of whether the initial consonant is voiced or not, and ends at the syllable offset. This synchronous movement of a tone with its attached syllable is accounted for by Xu (2005) with a proposal of coordinated biological movement, that is, based on a deep-rooted articulatory mechanism: one movement always becomes coordinated with its neighboring environment whenever pertinent. As pitch movement is often implemented with fast speed in running speech, the phase relation between pitch and syllable can only be manifested in full synchrony.

The F0 contour at the final portion of a syllable is mostly close to its underlying pitch target in terms of both pitch height and contour slope. At the syllable boundary, due to the physical inertia of articulatory states, the final velocity of the preceding syllable cannot stop immediately, thus it manifests as carryover effect by becoming the initial state of a following syllable, where a new cycle of target-approximation process begins. Within the duration of the following syllable, the F0 contour moves away from the preceding state and at the same time toward the current underlying tone. Such transition on F0 contour at the syllable boundary depends on humans' articulatory constraints which can finely account for the phenomena that, in some cases, the maximum or minimum F0 value of a syllable can be realized in the initial consonantal part of the following syllable.

Based on the above target-approximation assumption, a global F0 contour frame could be derived from the sequential target-approaching movement of each composing syllable in utterance. Such assumption of PENTA model is believed to be a universal feature in intonation generation, but the encoding schemes of specific communicative functions are language specific.

In Chinese, the lexical tonal features of each syllable are regarded as its underlying pitch target for asymptotic process, and the canonical shape of a target tone can be reached most approximately at the final portion of the syllable. Xu (1999) observes that the four tones do not manifest their canonical contours until arriving at the final portion of the aligned syllable. Specifically, T1 (H tone) shows its citation level form near the end of a syllable; T2 (R tone) shows the canonical rising contour from the middle portion of vowel(s); T3 (L tone) shows falling movement from the initial onset of a syllable, but the important lowest pitch point is not reached until the end of a syllable; in cases of T4 (F tone), when preceded by a H or R tone, it often begins the falling movement after the initial consonantal part of the syllable, while when preceded by a L tone or F tone, the falling movement starts at a much later time. Thus, Xu concludes that the latter portion of a syllable is the most significant in correct tonal perception, while the influence from tonal surroundings, the common carryover effect which is mainly manifested in the early part of a syllable, does not quite influence the correct tonal identity.

With the proposal of PENTA model, where the sequential syllable-bound asymptotical process is a basic mechanism for the generation of intonation contour, Xu (2005) argues against the Autosegmental Metrical approach (Pierrehumbert 1980; Beckman & Pierrehumbert 1986), where F0 contour is formed through interpolation between prominent pitch accents, while the contour of unaccented syllables are manifested as transitions between preceding and following accented syllables. Xu concludes that the differences in F0 contour between strong and weak syllables mainly result from differences in pitch targets and target-approximation speed. Accented syllables usually have longer duration, more strength and higher speed in articulation movement, which contribute to the complete realization of a canonical pitch target, while unaccented syllables, in the contrast, usually present partial or deviational tonal features, showing interpolation-like f0 forms.

The PENTAL model is quite different from the earlier mentioned Stem-ML model (Kochanski & Shih 2001, 2003), as it proposes that the surface F0 contour is a result of a sequential process of syllable-bound target approximation, where the left-to-right carryover effect is dominant in tonal interaction from surroundings, while in the Stem-ML model, the bidirectional smoothing mechanism with both carryover and anticipatory influence from neighboring tones is a basic principle for the formation of a continuous F0 contour. For instance, when a syllable with a falling tone is uttered with great strength, its preceding tone will increase the pitch level in advance for preparation.

The modelling program of PENTATrainer is implemented as a Praat script, with which a resynthesized F0 contour can be generated through an automatic process for extracting the pitch-target parameters based on the target approximation scheme, while another process of exhaustively optimizing the annotated parameters determines the ones which can be readily used for synthesis. The PENTATrainer system has been applied in the automatic prosodic modelling of Mandarin (Prom-on *et al*. 2009, 2011), English (Prom-on *et al*, 2009) and Thai (Prom-on & Xu 2012), etc.

b. Problems of Xu's approach

According to Xu (1997), due to the dominant carryover effect, the offset value of an early tone becomes the starting state of the following one, and the carryover influence could proceed to the following syllable to a different extent. However, as admitted by Xu (1999), there are also abundant cases in which a local peak or valley occurs within one syllable, indicating that the syllable-bound target approximation has completed within the aligned duration. Then according to the habitual behavior of articulatory mechanism, the remaining time of the syllable should be used for the preparation of a following tone, thus presenting the anticipatory effect on the following syllable. Such phenomena are absent in Xu's studies as it seems to contradict Xu's proposal that the canonical shape of an underlying contour is presented at the final portion of the aligned syllable due to the dominant carryover influence and the target approaching movement.

Moreover, in a tonal language like Chinese, the underlying pitch target of each syllable can correspond to its lexical tone. However, in non-tonal languages, like English, Xu (2005) also claims that the presence of a target is obligatory in each syllable, that is, an unstressed syllable carries a [mid] pitch target, and a stressed one has either a dynamic or static target. The targets in English are not constant like those in Chinese, and vary according to the positions of the aligned syllables within a word or a sentence, and also dependent on factors such as the position of sentence focus and the sentence type. Although Xu believes that the syllable-bound target approximation is a universal feature in contour generation, the assignment of compulsory underlying pitch targets in a non-tonal language does not seem quite convincing, as also seen in D'Imperio's comment (2006: preface):

> In intonational languages not each tone needs to be associated to a syllable, and each syllable does not need a tonal specification in non-tonal languages…The sparsity is potential problem for model such as the one proposed by Xu and colleagues, since it might be that synchronization between the laryngeal and supralaryngeal systems, and the pressure to achieve the kind of phase-locking proposed by this model, might be more true of languages such as Mandarin than for languages such as English, French, or Italian.

1.2.4.2 The MOMEL-INTSINT system

Hirst (2005) proposed that a central theme in the research of speech prosody concerns the way in which prosody contributes to meaning. He argues for the application of analysis-by-synthesis paradigm in testing different models of phonological structure.

According to Hirst & Espesser (1993), an F0 curve is a superposition of two components, namely, a *microprosodic* component which corresponds to the local effect of pitch from the individual segments, such as vowels and consonants, and the other component is a *macroprosodic* component which represents the global intonation pattern of an utterance. For the automatic modelling and symbolic coding of intonation patterns, there are two important algorithmic systems employed, MOMEL algorithm (Hirst & Espesser 1993; Hirst *et al*. 2000; Hirst 2007) and INTSINT coding system (Hirst 2001, 2007).

With MOMEL automatic algorithm, the macroprosoic component of F0 pattern is modeled, resulting in a continuous and smooth pitch contour. The algorithm detects the pitch movement on the raw contour, and reduces the countless pitch data to a limited number of target points, which are then interpolated in sequence with a quadratic spline function, with the transition between target points being smooth and continuous. The resulted new contour is the modeled output of the Momel algorithm, which can capture the significant phonetic properties on the raw F0 contour of the original utterance. A sample of the modelling output from Momel algorithm can be seen in the following Figure 1.8 quoted from Zhi *et al*. (2010):

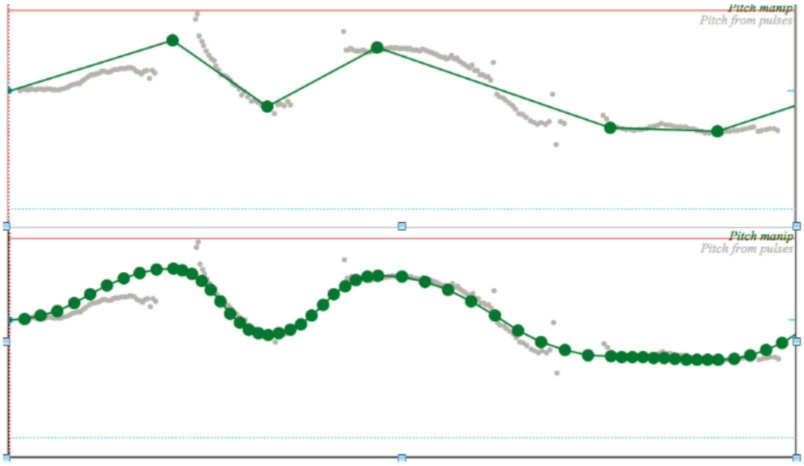


Figure 1.8: The F0-contour modelling with Momel algorithm

When a satisfactory model of F0 contour is obtained, the relationship between the target points and more abstract higher-level representation can be described by the coding system, INTSINT (INternational Transcription System for INTonation), which is presented in Hirst & Di Cristo (1998), where the intonation patterns of nine languages have been described with the coding alphabet, such as British English, Spanish, European Portuguese, Brazilian Portuguese, French, Romanian, Russian, Moroccan Arabic and Japanese.

INTSINT, as an annotation system, was developed to provide a narrow transcription of the pitch movement contour with a sequence of abstract symbols. In the book of Hirst & Di Cristo (1998), the following symbols were employed:

Top: ⇑ Bottom: ⇓ Mid: [Higher: ↑ Lower: ↓ Same: ⟶

Downstep: > Upstep: < Initial pitch: [Final pitch:] Resetting: [[

An intonation unit delimited by square brackets [ ] is defined as a local domain for transcription. The starting pitch range of a following unit is specified by symbols, “ < ” or “ > ” in round brackets, to be specific, (>) or (<) means the following phrase starts with an upstepped pitch range or a downstepped range, respectively. Therefore, with the transcription alphabet of

INTSINT system, the pitch contour of two intonation units can be represented with the abstract symbols as follows:

(⇑) [⇑ ↓ ↑ ↓ ⇑] (>) [⇑ ↓↑ ⇓]

T L H L T (D) T L H B

The use of the above transcription symbols aims to be applied in the intonation transcription across languages like the International Phonetic Alphabet does in describing vowels and consonants cross-linguistically. INTSINT focuses on describing the surface F0 features and provides "equivalent of a narrow phonetic transcription" (Hirst & Di Cristo 1998: 14) of an utterance pitch contour. It can be applied directly in transcribing the melodic form of a target language without much previous knowledge of its phonological features.

In the recent development of INTSINT (Hirst 2000, 2007, 2011a), Hirst proposed to use the low case letters, such as *t(op), m(iddle), b(ottom), h(igher), s(ame), l(ower), u(pstepped) or d(ownstepped)* instead of the capital letters, *T, M, B…*, which were employed in previous studies in representing the target points.

The INTSINT system employs an alphabet of 8 symbols for annotating the surface pitch movement, with each tone labeled relative to the immediately preceding one. The relative interpretation of the tonal segments is illustrated in the following figure (Hirst 2007):

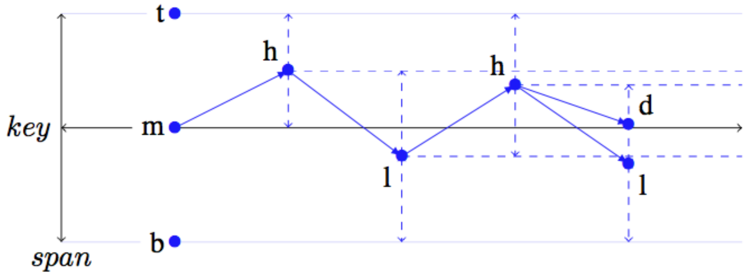


Figure 1.9: Relations among tonal segments of INTSINT system

In Figure 1.9, the points indicated by *t*, *m*, and *b* are **absolute tones**, which correspond to the relevant position in speakers' pitch range (defined with two parameters, key and span); the tones *h*, *l*, and *s* are **relative tones**, defined with respect to the preceding tonal targets; the symbols of *u* and *d* refer to the **iterative relative tones**, which are also defined relatively to the preceding tones, but generally with smaller upward or downward pitch changes.

The symbolic representation based on the relative movement of important pitch targets on the contour can be automatically converted to a series of target values, which could be applied as input for MOMEL algorithm to generate a stylized contour with a quadratic spline function. Such procedures can be applied automatically for the cyclic evaluation of the data for steady improvement in intonation synthesis. The following scheme (Hirst 2001) shows this cyclic procedure,

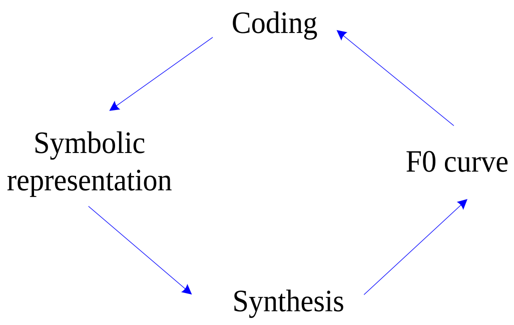


Figure 1.10: A reversible paradigm of F0 contour analysis with symbolic representation and re-synthesis evaluation

As both MOMEL and INTSINT algorithms only employ the important pitch movement of an intonation contour, it seems that some information of the original data are lost, but such loss can be neglected as it does not miss the qualitative values for a stylized contour. The automatically generated model is a stylized F0 contour deriving from interpolation of the pitch target points corresponding to the macroprosodic component of F0 contour, which are also the linguistically significant points in speech.

The INTSINT labeling model and MOMEL algorithm can be used together for automatic intonation synthesis. The data derived from one can be re-evaluated by the other, so the

INTSINT symbolic labels can be used as direct input for MOMEL stylized algorithm, and the stylized contour from MOMEL algorithm can be recoded into symbolic representation with INTSINT inventory. Therefore, this intonation system is also called an automatic MOMEL-based INTSINT transcription system. Such implementation could be seen in the following window quoted from Kim *et al*. (2008):

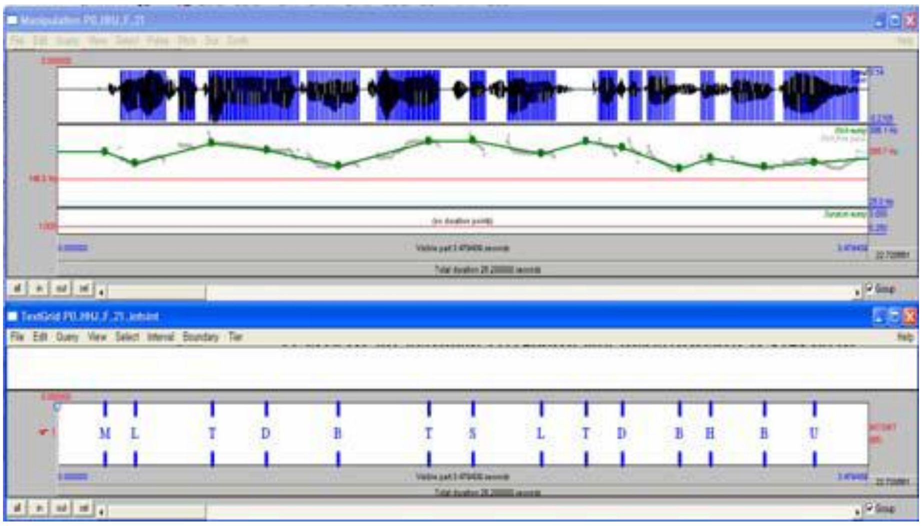


Figure 1.11: An implementation window of MOMEL modelling and INTSINT coding

Recently, The MOMEL-INTSINT algorithms have been implemented in a speech tool, *ProZed* (Hirst & Auran 2005, Hirst 2011a, 2012), which is designed for linguists to manipulate the prosody of utterances, either through the "immediate interactive assessment of the prosody determined by the annotation", or via the generated "synthesized stimuli for more formal perceptual experiments" (Hirst 2012). The Prozed tool is implemented as a plugin to Praat program (Boersma & Weenink 2012), and is still under development for the speech re-synthesis step in the analysis-by-synthesis paradigm.

In the present study, the analysis and modelling of the prosodic form in Mandarin speech was carried out. It is believed that the melodic property of Mandarin speech can be studied in the same way as those of non-tonal languages. The symbolic representation of speech melody provides an access in reducing the observable complexity from a large quantity of data to a more simplified model, which only retains the necessary information for expressing the functional contrasts in speech. The proposed mapping rules between representation of prosodic function and representation of prosodic form is evaluated with the re-synthesis stimuli generated by the Prozed tool. It will be shown in Chapter 5 that satisfactory re-synthesis of original utterances can be obtained with MOMEL algorithm from the manually annotated symbols of INTSINT at the surface phonological level.

1.3 Objective of the present study

1.3.1 The data of the study

Many prosodic studies preferred to use artificial utterances instead of utterances from daily-speech context, as it is believed that the former could serve the research purpose, avoiding any complex or disturbing contextual factor of spontaneous discourse. However, it is well known that speech prosody is quite context dependent, thus the complexity of real speech context may contribute to the explanation of many prosodic features which can never be expected from the artificial data.

The debate on whether to use real or experimental data is always on, as can be found in works such as Xu (2010) and Wennerstrom (2001) among others. As speech studies may need different types of data depending on the specific research purposes, one cannot apriori decide which kinds of data have more advantages.

The aim of this dissertation is to discuss the functions and forms of Chinese prosody in real spoken speech. Therefore, the authentic speech data selected from both read corpus and spontaneous corpus will be used. The word “authentic” was widely adopted in English language teaching literature, such as in Bernard (2006), Clarke (1990), Peacock (1997), Senior (2005), and Wallace (1992), as it is often argued that second-language learners should be provided in language class with authentic materials rather than the artificial ones, due to the real-life nature of the former. In the present study, the authentic speech data includes both the spontaneous daily dialogues, and the read speech with the spoken texts based on real-life themes.

The two corpora employed in the study are the *Chinese Spontaneous Conversation Corpus* (Li *et al*. 2001, Li 2002) and the *Chinese Multext* corpus (Komatsu 2009), respectively. Utterances in the two corpora have been labeled at the segmental level, as discussed in Chapter 3.

1.3.2 A plausible mapping between form and function

In *The Philosophy of Grammar* (1925), Jespersen defined three terms, *form*, *function* and *notion* in distinguishing language components and their usage; *form* denotes the sound actually spoken; *function* refers to the grammatical category denoted by the form; *notion* is the meaning expressed by the function.

In the study of speech prosody as reviewed in sections §1.1.1 and §1.1.2, the function of prosody contributes in various ways to the interpretation of speech meanings, while the form of prosody manifests how the acoustic events are perceived by listeners. In the process of analyzing and modelling the prosody of natural languages, the matching between form and function is still a poorly understood issue. An important reason is that there is no consensus on the representation of prosodic form and the representation of prosodic function.

In the study of Hirst (2005), he stated that a systematic distinction shall be made between the two levels of representation. Such proposal is different from the ToBI system (Silverman *et al*. 1992), in which pitch movement are annotated by symbols such as L* and H%, with L/H corresponds to a low/high tone, representing the prosodic form, while the symbols * and % refer to the functional aspect of prosody *accented* and *boundary*, respectively. With the employment of symbols L* and H%, ToBI system conflates the formal and functional aspects of prosody in the representation.

To provide a potential solution to the mapping between the function and form of speech prosody, Hirst (2000, 2005) postulates a multi-level organization for the form-function interface, and encourages linguists to define mapping rules between the representation of function and the representation of form through analysis by synthesis paradigm with speech modelling technology. According to him, both form and function of prosody exist in all languages, but what is different is the specific way that a language establishes for the mapping between the two levels.

The present prosodic study of Mandarin Chinese follows Hirst's idea in mapping the acoustic signals with linguistic functions by distinguishing a number of intermediate levels, and exploring their interactive relations. The basic principles of Hirst's mapping proposal are summarized in the following graph:

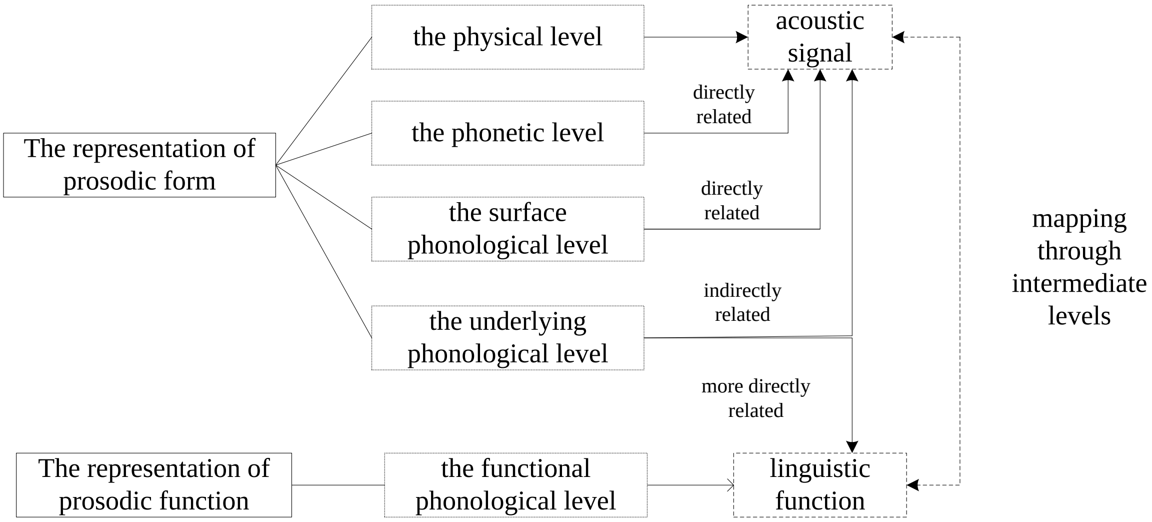

the physical level
acoustic signal
directly related
the phonetic level
The representation of prosodic form
the surface phonological level
directly related
mapping through intermediate levels
indirectly related
the underlying phonological level
more directly related
The representation of prosodic function
the functional phonological level
linguistic function

Figure 1.12: Multi-level representation of prosodic form and the form-function mapping proposal

In Figure 1.12, the theoretical status of each level is interpretable with respect to the adjacent levels in the multi-level organization for the form-function interface.

In Mandarin speech, the function of prosody contributes to word identity at the lexical level as well as intonation at the overall sentential level. Accordingly, the representation of prosodic function in Mandarin speech includes the lexical tones aligned with each syllable, such as Tone 1 (HH), Tone 2 (LH), Tone 3 (LLH), Tone 4 (HL) and Tone 0 (neutral tone), and also the sentential intonation, with its “grouping function” annotated with [±terminal] tones at the speech unit boundaries, and the “weighting function” annotated with [±accented] above the relevant syllables in marking the prominent status. The representation of prosodic function is directly related to the author’s linguistic interpretation of speech. The functional annotation of the prosody of a neutral Mandarin utterance could be presented as follows:

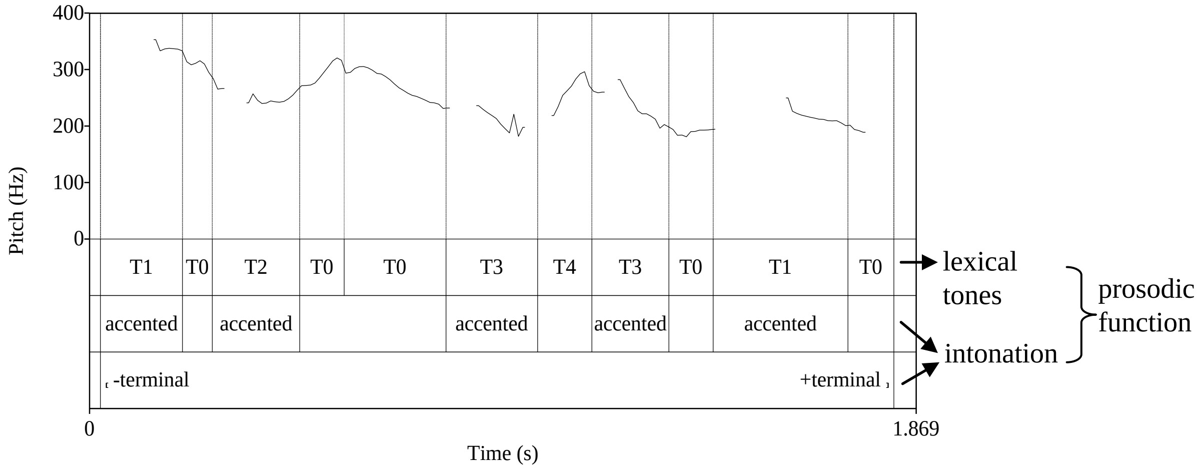


Figure 1.13: The representation of prosodic function of a neutral Mandarin utterance

As seen in Figure 1.13, the representation of the prosodic form can be distinguished into different levels. The raw F0 contour directly represents **the physical level** of the form of prosody, as which is the physical acoustic signal in the listeners’ perception of prosody. The raw F0 contour of speech prosody can be transcribed into symbolic coding by a series of INTSINT symbols, as seen in the following Figure 1.14:

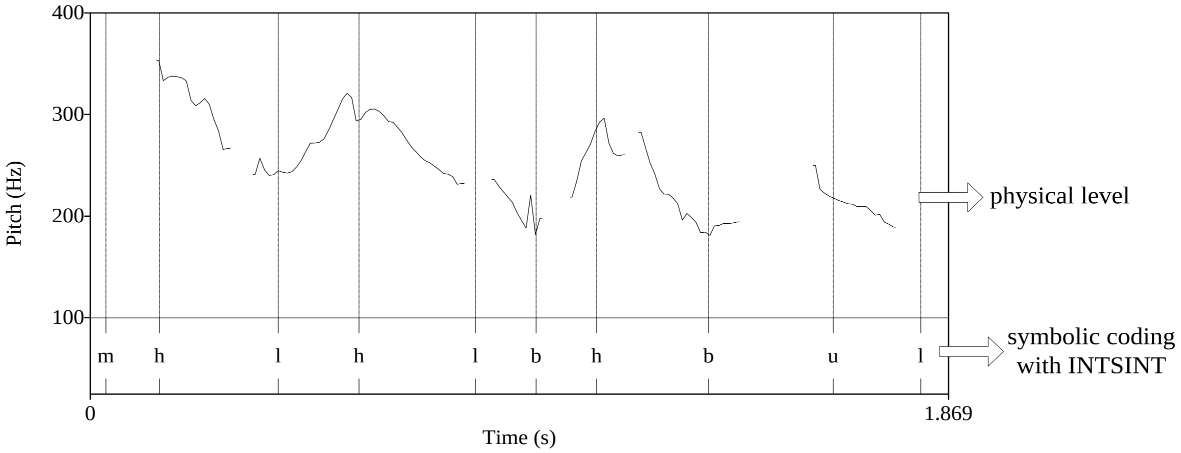


Figure 1.14: Symbolic annotation of the F0 movement with INTSINT alphabet

Such symbolic coding can provide a narrow transcription of the pitch movement contour, by reducing the observable complexity of a large quantity of raw F0 data to a more simplified model, which only retains the necessary prosodic information in speech. The surface pitch pattern annotated by the sequence of INTSINT symbols provides the **surface phonological representation** of prosodic form. With the implementation of Momel-INTSINT system, each discrete symbol at the surface phonological level can automatically derive its corresponding fundamental frequency value, as seen in the follow Figure 1.15:

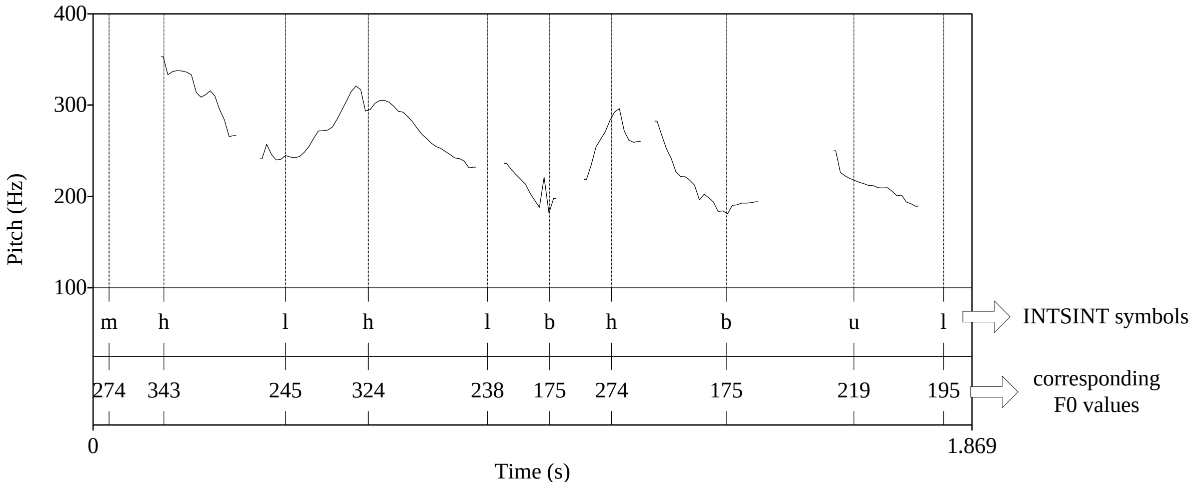


Figure 1.15: The sequence of discrete symbols and the corresponding F0 values in coding the prosodic form of an utterance

The F0 values are directly related to the acoustic signal and represent **the phonetic level** of the prosodic form. The symbolic coding of prosody with the INTSINT alphabet and the corresponding F0 values serve as the interface between the physical manifestation and the abstract cognitive representation of prosodic form.

The symbolic coding of the utterance pitch contour is directly related to the physical level of the acoustic data. At the same time, such level of representation is also interpretable to **the underlying phonological representation** of prosody, which serves as an intermediate phase in relating the linguistically prosodic functions with prosodic forms. The interaction between the surface phonological level and the underlying phonological level in representing the prosodic form shall be discussed in details in section §4.4.1.2.

The manual annotation of the surface prosodic form with INISINT symbols was used as the input of Prozed tool in generating the synthesized stimuli, as compared to the original utterance for evaluation. It can be seen in the following window that a satisfactory result of synthesized utterance (in green line) is derived, which closely follows the pitch movement of the original one (in red line).

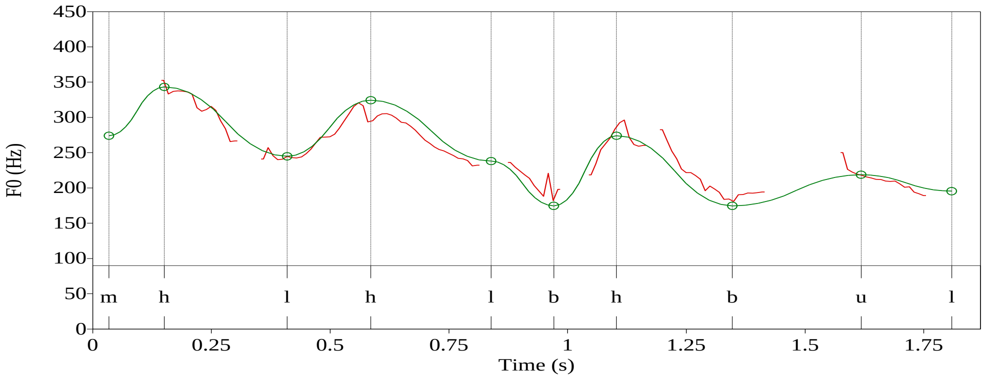


Figure 1.16: Synthesized utterance (in green colour) in comparison to the original utterance (in red colour)

The complete procedure of the functional analysis of prosody, and the annotations of prosodic form, as well as the mapping proposal between the two aspects of representation is presented in the following Figure 1.17:

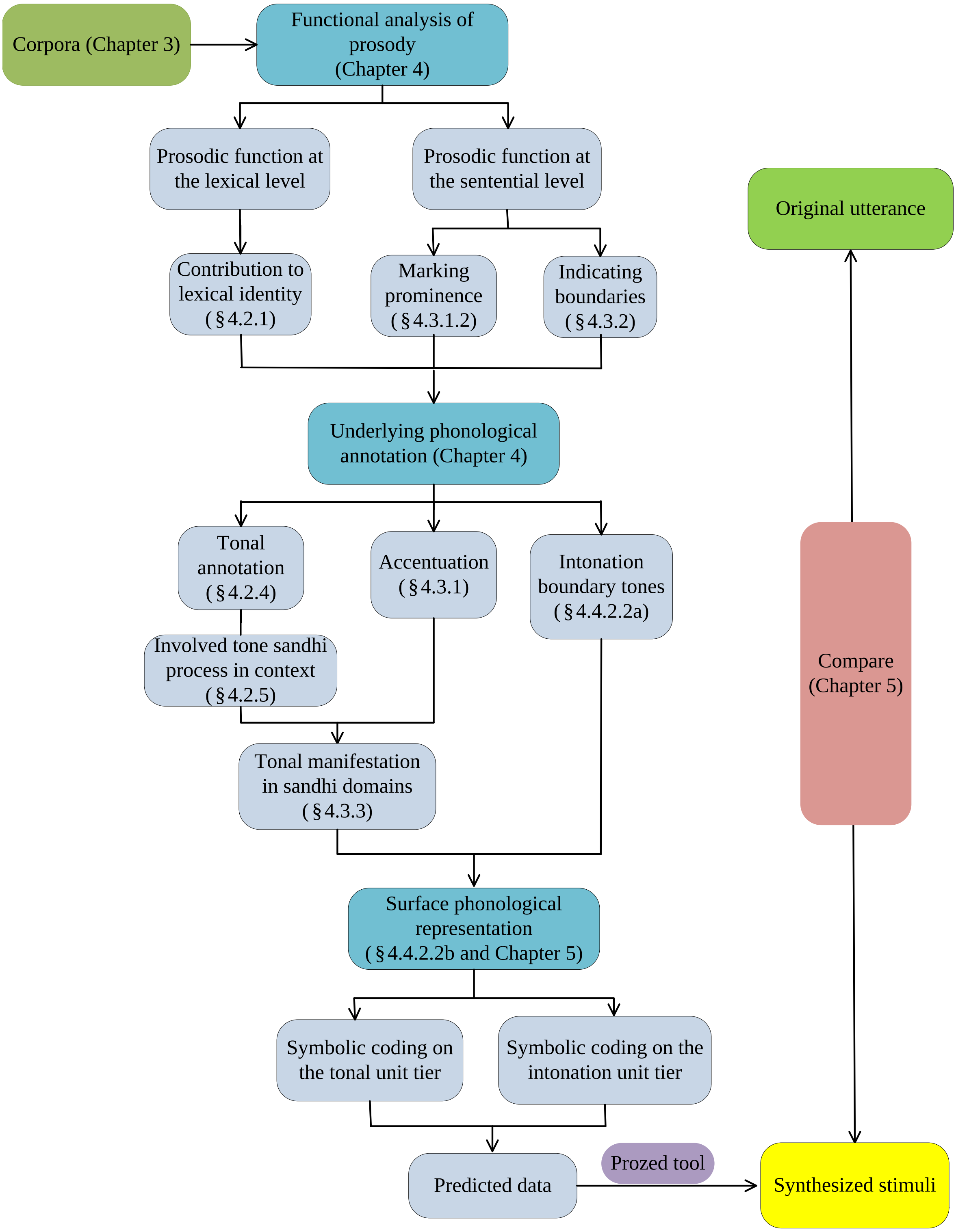

Corpora (Chapter 3)
Functional analysis of prosody (Chapter 4)
Prosodic function at the lexical level
Prosodic function at the sentential level
Original utterance
Contribution to lexical identity (§4.2.1)
Marking prominence (§4.3.1.2)
Indicating boundaries (§4.3.2)
Underlying phonological annotation (Chapter 4)
Tonal annotation (§4.2.4)
Accentuation (§4.3.1)
Intonation boundary tones (§4.4.2.2a)
Compare (Chapter 5)
Involved tone sandhi process in context (§4.2.5)
Tonal manifestation in sandhi domains (§4.3.3)
Surface phonological representation (§4.4.2.2b and Chapter 5)
Symbolic coding on the tonal unit tier
Symbolic coding on the intonation unit tier
Predicted data
Prozed tool
Synthesized stimuli

Figure 1.17: A scheme of the mapping proposal between the functional aspect and formal aspect of prosody

It is shown in this study that the symbolic coding of Mandarin prosodic form is derived from the functional analysis of prosody at both the lexical level and the sentential level. The 60 Mandarin utterances of both spontaneous and read speech styles are coded with the INTSINT system, with the synthesized stimuli compared with the original utterances. The comparison results will be presented in Chapter 5 and Appendix III, in which, one can find that the synthesized prosody can satisfactorily resemble the pitch contours of the original utterances.

CHAPTER 2

Literature Review of Prosodic Study

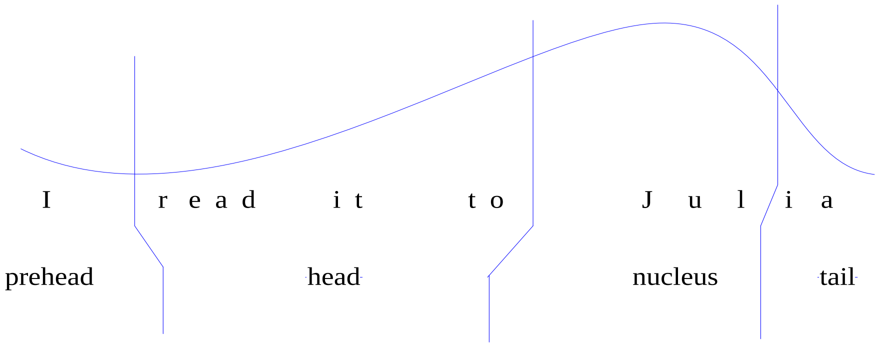
2.1 Introduction

This chapter reviews the previous literature on prosodic melody. The traditional studies in melodic prosody mainly centered on non-tonal European languages, in which the melodic contour directly represents the intonational pattern of the language speech. The prosodic theories and methods of three important schools are reviewed, namely, the contour analysis of the British School, the perceptual approach (or IPO approach) of the Dutch School, and the level approach as well as the autosegmental-metrical approach of the American School. The literature review of these studies in speech prosody aims to provide helpful background for the present study.

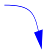
2.2 The British School

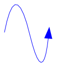
The approach of the British school in intonation study is pedagogically oriented, and mainly represented by Palmer (1922, 1933), Armstrong & Ward (1926), Halliday (1967, 1970), Crystal (1969), O’Connor & Arnold (1973) and etc. The British approach describes intonation patterns with the contours, and the intonation analyses are mainly based on the impressionistic judgment of the researchers.

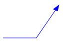
Within this framework, connected speech branches into tone groups, and each group is further divided into four parts, as prehead, head, nucleus and tail. “Prehead” refers to the pitch stretch on unstressed syllable(s) preceding the head at the beginning portion of a tone group; “head” is the pitch curve starting from the first stressed syllable up to the nuclear tone; “nucleus” denotes the pitch movement of the most prominent tone(s); “tail” refers to the pitch stretch following the nuclear syllable until the end of the tone group. Among these four components, pre-head, head and tail can all be optional, while the nucleus which contributes most to the pitch movement of intonation, is an obligatory part of a tone group. In an utterance with more than one syllable, the nuclear tone often situates on the last stressed syllable of a tone group. The pattern of (prehead)-(head)-nucleus-(tail) is the basic intonation model of the British school, and has been widely accepted as a standard structure in analyzing English intonation. Thus, the method is also called “nuclear approach” (Palmer 1922). According to the prehead-head-nucleus-tail pattern of the British school, a sentence contour can be divided into the following structure for further analysis (Ladd 1996: 210):



Different patterns of English intonation are specified by the contour of nuclear tones. Palmer (1922) distinguished intonation patterns with five nuclear tones, such as high-falling ↘ , low-falling ↘ , high-rising ↗ , low-rising ↗ , and falling-rising ∨ , and two head tones, such as superior head ¯, scandent head / and inferior head _. Different expressive meanings in intonation are conveyed by these specific nuclear tones and heads. Thus, the intonation of a sentence can be shown iconically with Palmer’s symbols as follows:

I ¯don’t  care!

Is ¯there  something wrong?

_ Good morning  ! Can I / help you?

Later in 1933, Palmer expanded his categorization of intonation patterns by employing six combined tones of heads and nuclei, and gave metaphorical names vividly reflecting the intonation movement. They are *cascade* (superior head + low-falling nucleus), *dive* (inferior head + high-falling nucleus), *ski-jump* (scandent head + low-falling nucleus), *snake* (superior head + rising-falling nucleus), *swan* (scandent head + low-rising nucleus) and *wave* (superior/inferior head + high-rising nucleus).

The various contour patterns in speech intonation were further generalized by O’Connor & Arnold (1973) into two preheads (low and high), four heads (low, high, falling, rising), seven nuclear tones (low fall, high fall, low rise, high rise, fall rise, rise fall, mid-level), and one continuing tail. The different combinations of prehead, head, nucleus and tail can represent various intonation patterns in speech.

Halliday (1967, 1970) also made great contribution to intonation studies of British English by highlighting the informational function of intonation. In his intonation system, three important notations, namely, tonality, tonicity and tone are employed. Tonality delimits speech into small tone groups, which correspond to the information units of speech, and such grouping does not require any coincidence with grammatical boundaries. Tonicity is the most salient part of a tone group, and carries the new and focal message of the information unit. Tone refers to a number of typical tonal contours in speech. There are seven primary tones distinguished by Halliday (1970: 9), that is, five simple tones: Tone1 falling, Tone2 high-rising, Tone3 low-rising, Tone4 falling-rising, Tone5 rising-falling, and two compound tones: Tone13 falling plus low-rising, Tone53 rising-falling plus low-rising. The above primary tones could represent the typical intonation patterns in English. Halliday also recognizes a large number of secondary tones and differentiates them into two types, one type refers to the pitch stretch of pre-tonic segment, and the other type refers to the tones subdivided from primary tones.

In Halliday's system, there are three prosodic units with hierarchical relations: syllable, foot and tone group. Tone group, marked by double slashes // at the boundary, is a basic unit for analyzing English intonation, and consists of a succession of feet, each marked by / at the boundary. Each foot is a basic rhythmic unit, and usually has its first constituent syllable carrying the salient beat. The tonicity of a tone group, marked in underlines, is often located at the initial syllable of a foot. In some cases, a foot may also start with a silent beat, marked by symbol /\. With Halliday's tone system, the intonation of a sentence can be represented (ibid: 27-28) as follows:

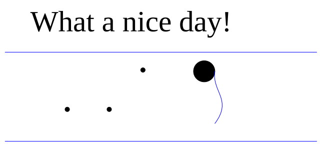
//1/\ I'll/ see what/ I can/ do//.

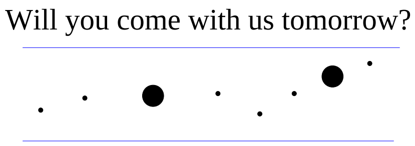
//2 Did you play/ tennis// 1/\ or /golf// ? ('which ?')

//2 Did you play/ tennis// 2/\ or /golf// ? ('yes or no?')

The number located at the beginning of each tone group represents the type of the primary tone, for instance, 1 indicates the intonation pattern of the sentence is a primary contour of Tone1.

Among the intonation transcription systems developed by the British school, the "interlinear-tonetic" (Cruttenden 1997) system is the most popular and iconic one. In this system, there is a top line and a bottom line, representing a speaker's pitch range, and between the lines there is a sequence of dots, each of which corresponds to a syllable. Big dots represent accented syllables, and small dots represent unaccented ones. Glides, in certain cases, denote lengthened sounds, or long vowels or diphthongs. With this transcription system, the pitch movement of a sentence intonation could be indicated vividly by the movement of a series of dots in the sentence. For example,





Because of its iconicity in describing intonation, the contour approach of the British school has applied widely in pedagogical textbooks, and contributed greatly to the motivation of the English language teaching.

2.3 The Dutch School

The IPO (Institute for Perception Research) approach of the Dutch school (’t Hart & Cohen 1973; ’t Hart & Collier 1975; ’t Hart *et al*. 1990) employs perceptual analyses in intonation study. According to the IPO proposal, an F0 contour is composed of a number of perceptually salient pitch changes, which defines the main tendency of an intonation pattern. These important perceptual changes in pitch movement are intentionally produced by speakers. Contrary to traditional views, it is believed by the Dutch School that intonation has no intrinsic meanings, but rather “influences the syntactic analysis, which in turn affects the semantic interpretation” of an utterance (’t Hart *et al*. 1990: 110). They proposed that speakers take “look-ahead” strategy by making use of the melodic and textual information comprehensively for the successive advancement of speech prosody.

The IPO approach intends to model intonation based on the native listeners’ perception. According to ’t Hart *et al*. (1990: 5), “psychoacoustic thresholds and communicative relevance constitute the lower and upper boundaries that delimit the province of our perceptual quest”. The principle of the perceptual method is to analyze intonation based on native speakers’ judgment according to their listening habits of what the mother tongue should sound like, which is “listener’s internal representation of the intonation system” (ibid: 66).

Intonation patterns are not distinguished according to any linguistic categories, such as statements and questions, as it is claimed that “such a strategy would be contrary to the IPO approach" (ibid: 67). Different categories of pitch patterns are defined only on their distinct phonetic properties as perceived by native listeners. Patterns with perceptual equivalence are classified into the same category.

The Dutch school shapes its phonetic model of intonation in the form of straight-line interpolation, which contains the "perceptually relevant pitch movements" with all the irrelevant details excluded. The original data on F0 contour is reduced into a small number of valuable points which represent the important pitch movement in perception. These valuable points are then interpolated in sequence with a series of straight lines, forming an acceptable approximation of the original F0 contour. Such stylization of F0 contour is a basic principle in intonation modelling of the Dutch school. The resulting stylized pattern formed with "a smallest possible number of straight-line segments" (ibid: 43) has perceptual equality with the original F0 contour. The following graph shows the original intonation curve of an utterance and its close-copy stylization in linear segments:

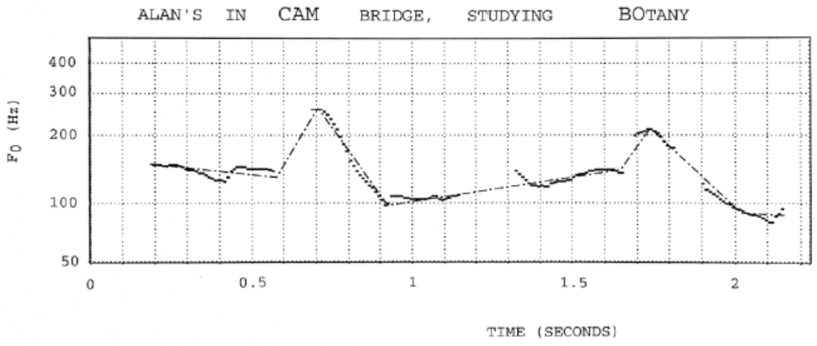


Figure 2.1: The stylized contour and the original contour

Among the linear stylizations, another important process, standardization is applied, intending to sort out a number of categories which could represent the basic and prototypical intonation patterns of a language. The standardization obeys two principles. One is that the standardization shall not delimit the categorical differences across intonation patterns; the other is that the resulted standardized pattern is a fair and acceptable representation of the original intonation contour (ibid: 48). Stylization and standardization are two important principles of the Dutch school in analyzing and modelling the intonation contour of a language. In the following graph, a standardized intonation pattern is indicated as a simplified contour marked in solid thick lines. The down-going direction of three parallel straight lines (topline, midline and baseline) in the graph mimics the declination phenomena widely observed in natural speech of English.

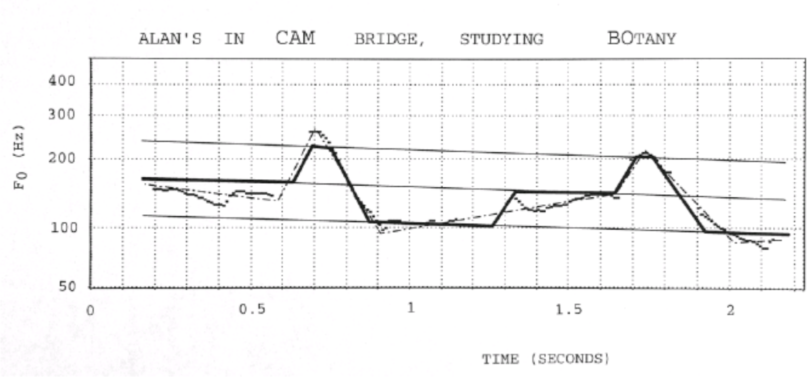


Figure 2.2: The resembling of intonation declination with three base lines

The data on the linear stylized model of the Dutch school are in logarithmic values. The perceptual equality is achieved among native speakers based on their linguistic tolerance of irrelevant information.

2.4 The American School

2.4.1 The level approach

With the level approach of the American school, a sequence of distinct pitch levels is employed to represent the pitch movement of intonation. Pike (1945) defined four pitch levels, labelled from number 1 to 4, corresponding to extra-high, high, medium, and low level within a speaker’s pitch range.

--------------------------------------1 extra high

--------------------------------------2 high

--------------------------------------3 medium

--------------------------------------4 low

The four pitch levels have no definite values, but are defined relatively to each other. The distance between each two levels is not uniform, but varies from time to time. The initial point and final point of each contour is marked with a numeral of a corresponding pitch level, which is linked together by a “-“. In an utterance, each new primary contour in intonation starts on a stressed syllable, which is marked with a °before the pitch numeral. The intonation contour of a sentence can be represented with the four pitch levels, exemplified by Pike (1945) as follows:

The boy in the house is eating peanuts rapidly.
3- ^{o}2-3 3- ^{o}2-3 3- ^{o}2—3 ^{o}2--3 ^{o}2- -4

Good morning Tom.
3- ^{o}2--4- -4-3

The doctor bought a car.
3- ^{o}2-4-3 4- ^{o}2-4

2.4.2 The Autosegmental-metrical approach

By reducing the number of pitch levels from four to two, the autosegmental-metrical (AM) theory, first proposed by Pierrehumbert (1980), maintains the original approach in analyzing the pitch movement in terms of levels, while at the same time, it eliminates the problem of Pike’s excessive scales, which often causes excessive complexity.

The AM theory benefits from the basic framework of autosegmental phonology (Leben 1973) and metrical phonology (Liberman 1975; Liberman & Prince 1977). The autosegmental phonology originated analyzing the contrast of lexical tones in African languages, claiming that tones are not intrinsic features in the segmental string, but rather features in an autonomous tier for independent analyses. Metrical phonology demonstrates the binary relations of weak and strong accents, which contribute to a language's rhythmic features. The AM theory, according to Ladd (1996), is more biased towards the autosemental than the metrical phonology, as in the AM approach speech is grouped into a number of prosodic units with hierarchical relations, and within each unit the intonation data is analyzed at a level autonomous from the segmental string.

According to the framework of the AM theory, the intonation contour is represented by a string of phonological events, namely, "pitch accent", "phrase accent" and "boundary tone". Pitch accent is a single H tone or L tone or a combination of the two. For each pitch accent, there is an obligatory central tone, which is usually associated with the accented syllable in English speech. Thus, in a single-tone pitch accent, the H or L tone is the central tone marked with an asterisk, as H* or L*, while in a double-tone accent, in addition to such a starred tone, there is an unstarred one, called "leading tone or trailing tone" according to its position relative to the starred tone. A pitch accent with two different tones, such as H*+L or L+H* forms a contour tone. In Pierrehumbert's first version of the AM theory (1980), seven pitch accents are specified: H*, L*, H*+L, H+L*, L*+H, L+H* and H*+H. In addition, there are two phrase accents, indicated as H- and L-, and also two boundary tones as H% and L%, which all associate with edges of prosodic domains, thus, they are jointly referred as "edge tone" by Ladd (1996).

The above pitch events are the "building blocks" of the intonation contour in English. According to Pierrehumbert (1980: 13), the tonal sequence could formulate a finite-state intonation grammar as the following transition network shows,

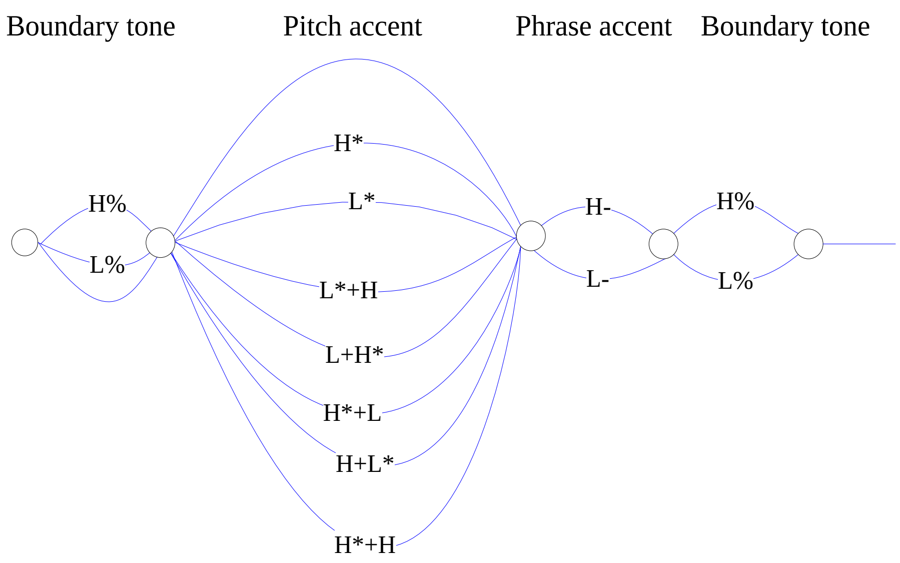
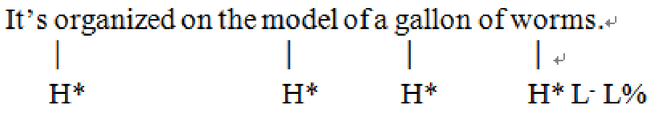


Figure 2.3: The intonation grammar of autosegmental-metrical approach

The intonation pattern of a sentence could be represented with these salient pitch events, that is: one or more pitch accents, a phrase accent and a boundary tone. There could be more than one pitch accent in a phrase when the nuclear accent is located on the final stressed syllable of the unit. When the nuclear accent is early in the phrase, the following stressed syllables do not carry pitch accents any more, and their F0 contour is derived from the phrase accent. The transcription of a sentence intonation contour with the AM model is exemplified by Pierrehumbert (1980: 19) as follows:



```
It's organized on the model of a gallon of worms.
  |                  |        |         |
  H*                 H*       H*        H* L- L%
```

With the AM approach, the various nuclear patterns used by the British school can be transcribed as a simple combination of pitch accent, phrase accent and boundary tone. The detailed comparison made by Pierrehumbert (1980 appendix) on twenty-two patterns were turned into a table by Ladd (1996: 82) as follows:

the AM approach	the British school
H* L L%	fall
H* L H%	fall-rise
H* H L%	stylished high rise
H* H H%	high rise
L* L L%	low fall
L* L H%	low rise (narrow pitch range)
L* H L%	stylished low rise
L* H H%	low rise
L+H* L L%	rise-fall
L+H* L H%	rise-fall-rise
L+H* H L%	stylished high rise (with low head)
L+H* H H%	high rise (with low head)
L*+H L L%	rise-fall (emphatic)
L*+H L H%	rise-fall-rise (emphatic)
L*+H H L%	stylished low rise
L*+H H H%	low rise
H+L* L L%	low fall (with high head)
H+L* L H%	low fall-rise (with high head)
H+L* H L%	stylished high rise (low rise?) with high head
H+L* H H%	low rise (high range)
H*+L H L%	stylished fall-rise (calling ‘contour’)
H*+L H H%	fall rise (high range)

Table 2.1: British intonation patterns decoded with the AM system

In contrast to the approach of the British School, the F0 contour is not divided into strict prenuclear and nuclear regions in the AM framework. Pierrehumbert (1980) believes that prenuclear and nuclear tones are not phonetically distinct from each other, and can both be represented by pitch events.

Within the AM framework, the basic unit for intonation analysis is identified as an “intonational phrase” (IP), and the unit is usually specified by features, such as containing at least one compulsory nuclear accent, a boundary tone at the right edge, and a pitch reset at the beginning of a following IP. In most cases, an IP corresponds to a syntactic unit in English.

Later on, an “intermediate phrase” was proposed by Beckman & Pierrehumbert (1986) as another prosodic unit at the lower hierarchical level, such that the phrase consists of one or more pitch accents and ends with a phrase accent L^{-} or H^{-}.

Since the first appearance of Pierrehumbert’s doctoral dissertation in 1980, the framework of AM theory has undergone several stages of revision. For instance, according to a revised version in Beckman & Pierrehumbert (1986, 1988), one pitch accent type, H*+H is eliminated from the original seven accents, as they claim that H*+H could be analyzed as “involving ordinary H* accents produced in an elevated but compressed pitch range” (Beckman & Pierrehumbert 1986: 306). In two later versions of the AM theory, represented by Silverman *et al*. (1992) and Beckman & Ayers (1997), two further accent types, H*+L and H+L* were cancelled, with the first one integrated into H* accent and the second replaced by H+!H* accent. Therefore, according to the later versions of AM framework, there are now altogether only five pitch accents, namely, H*, L*, L*+H (the ‘scooped’ accent), L+H*, and H+!H*.

The AM theory reveals the phonological features of English intonation by specifying a tonal sequence of pitch accent, phrase accent and boundary tone. This tonal inventory forms a consistent F0 contour through transition rules. Pitch accents, which usually locate on accented syllables together with edge tones, define the general frame of an intonation contour, while the F0 contour of unaccented syllables is derived from the interpolation of adjacent tones associated with accented syllables.

Researchers working on AM theory try to connect the intonational phonology defined by H and L pitch events with the expressive functions of intonation. According to Pierrehumbert & Hirschberg (1990), the various pitch accent types perform different discourse functions in English, as they propose that pitch accents “convey information about the status of the individual discourse elements”. Thus, for instance, a H* pitch accent is used to express introduction or continuation, and a L* accent could indicate completion and assertive confidence. As reported in Gili Fivela (2008), a boundary tone can also denote the relationship between adjacent intonational phrases, for instance, a low boundary tone denotes that the current intonational phrase is conclusive and has no relation with a subsequent one, while a high boundary tone indicates that the relevant information will be followed up in the upcoming phrase (Pierrehumbert & Hirschberg 1990: 307).

However, there are also abundant counter-examples in real data, for example, low terminals often associate with incomplete topics. Therefore, the relationship between pitch events and their corresponding discourse functions is actually in “very general and schematic terms”. The discourse meanings of tonal events are with great flexibility (Gili Fivela 2008).

2.4.3 The ToBI labelling system

Tone and Break Indices (ToBI) system is first introduced for intonation transcription in Silverman *et al*. (1992). This system is developed on the basis of Pierrehumbert’s AM framework (Pierrehumbert 1980; Beckman & Pierrehumbert 1986) and the categorized study of prosodic junctures by Price *et al*. (1991). Since its introduction, it has been widely applied in intonation transcription across languages.

The ToBI system requires a deep understanding of the prosodic grammar of each target language before making any intonation analyses. The ToBI for Mainstream American English (MAE_ToBI) is a most developed transcription system, and is also an ideal example for the ToBI system to apply in other languages. Six obligatory parts compose the MAE_ToBI system. They include speech signals of an utterance presented in waveform, an F0 contour and four annotation layers: a tonal layer for tagging the intonation contour, a word layer for annotating the orthographic form of each word, a break-indices layer with numerals marking the strength of word boundary, and a miscellaneous layer for making supplementary comments (Beckman *et al*. 2005).

Among the composing parts, the tonal tier and break indices tier are the most important two tiers, which also compose the name of the ToBI system. The following information reveals the details of the two tiers.

On the tonal tier, the pitch accents H*, L*, L+H*, L*+H, H+!H*, phrase accents H- and L-, boundary tones H%, L% and %H are used to transcribe the prominent pitch properties of an intonation contour. Other labels, such as L+!H*, !H- and !H* are used to transcribe the downstep phenomena in speech, with the exclamation mark denoting the start of a compressed pitch range. The pitch restart in the contour is marked by a label, %r. Uncertainty on tonal type or tonal occurrence is indicated by a question mark. Prosodic events such as a delayed peak is marked with a symbol <, and a maximum F0 point is labelled as HiF0 within an intermediate phrase.

On the break-indices tier, there are five index values from number 0 to 4, specifying the different boundary degrees. 0 indicates the very close inter-word juncture, which is often the phonetic break within a clitic group; 1 denotes the common boundary of words within a phrase; 2 is often used in ambiguous cases, when there is no definite indication whether the juncture should be specified by a index 3 or index 4 due to mismatch between actual perception and tonal marks in prosodic groupings; 3 indicates the boundary of an intermediate phrase ending with a phrase accent; and 4 marks the boundary of an intonational phrase delimited by a H% or L% boundary tone. The categories of the index values indicate "the metrical hierarchy of the prosodic groupings" (Jun 2005: 2). The numerical break indices are specified mainly based on annotator's perceived sense with the help of prosodic cues such as pause, syllable duration and intonation.

The following graph quoted from Beckman *et al*. (2005: 20) is an example of the MAE_ToBI system in annotating the intonation of an American English utterance:

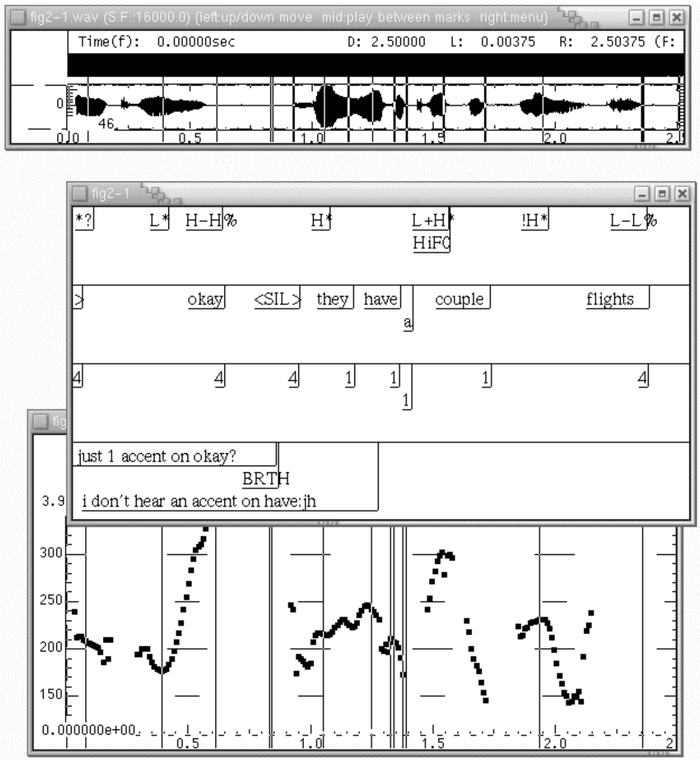


fig2-1.wav (S.F.:16000.0) {left:up/down move mid:play between marks right:menu}
Time(f): 0.00000sec D: 2.50000 L: 0.00375 R: 2.50375 (F:
0
46
0.0
0.5
1.0
1.5
2.0
2.5
fig2-1
*?
L*
H-H%
H*
L+H*
HiF0
!H*
L-L%
>
okay
<SIL>
they
have
a
couple
flights
4
4
4
1
1
1
1
4
3.9
300
250
200
150
0.000000e+00
0.5
1.0
1.5
2.0
2

just 1 accent on okay?

BRTH

i don't hear an accent on have:jh

Figure 2.4: A labeling example with MAE_ToBI system

It is important to note that the labels on tonal tier of the ToBI system are not symbolic representation of the actual pitch movement, as the labels are employed to "tag" not to "code" the intonation contour (Beckman *et al*. 2005: 37). The ToBI system does not provide a phonetic model, but intends to reveal the intonation phonology of the target language. In the system, tags are used for "retrieving phonologically relevant portions of the fundamental frequency and audio signals" in order to "keep track of possible phonological analysis of tune-text association at a stage when a research team has looked at enough data to make plausible guess…of contrasting intonation and prosodic patterns" (ibid: 39). However, as admitted by the research groups of ToBI system, a close phonetic model is still in need for intonation studies. Within ToBI conventions, a machine-readable system is proposed (Pitrelli *et al*. 1994, Beckman & Ayers 1997), so that annotated corpus can be shared by researchers for comparison between different languages.

2.4.4 Pan-Mandarin system

Pan-Mandarin system is developed from the original ToBI system for transcribing Chinese intonation. According to Peng *et al*. (2005), there are eight tiers in Pan-Mandarin system: a words tier, a romanization tier, a syllable tier, a stress tier, a sandhi tier, a tones tier, a break indices tier, and a code tier.

The word tier transcribes the acoustic data into corresponding Chinese characters for convenience of native researchers' observation;

The romanization tier is similar to the orthographic tier in the MAE_ToBI system. It transcribes each syllable in its orthographic form with Chinese Pinyin alphabet. The canonical tone of each syllable is also transcribed with the numerical notation system proposed by Chao (1930). 55 represents Tone1, 35 represents Tone2, 21 represents Tone3, and 51 represents Tone4;

The syllable tier transcribes the actual phonetic output of the acoustic syllable, e.g. when a disyllabic word, **tāmen** is uttered with contraction and sound swallowing, this word is transcribed according to its actual phonetic output as **tām**;

The stress tier transcribes the stress degree marked on each syllable. It deserves notice that the transcription is based on the actual phonetic output of the lexical tones. S3 indicates syllables with canonical lexical tones; S2 indicates syllables with substantial reduced tones; S1 indicates syllables which lose their distinctive tonal features in weakly-stressed position, and S0 indicates syllables which are inherently toneless.

The sandhi tier transcribes the actual phonetic output of lexical tones in an utterance. The allotones are transcribed with Chao's tonal numerals;

The tone tier transcribes the boundary tones of intonation with H% and L% tags, and marks the modification in global or local pitch range. The tag %reset indicates a reset at pitch level at the beginning of a new prosodic phrase, %q-raised denotes the overall raised pitch range in questions, especially in echo questions, %e-prom indicates the beginning of an expansion in local pitch range due to emphatic prominence, and %compressed marks the beginning of a compressed pitch range, which usually appears after a broadened pitch range under focus.

The break-indice tier transcribes the hierarchical levels of prosodic junctures. There are six values of break indices ranging from 0 to 5. 0 marks the boundary between contracted syllables, requiring on stress tier a S0 or S1 annotation at left or right; 1 marks the default syllable boundary within a multisyllabic word; 2 marks a minor phrase boundary, which is followed by at least a S2 on the stress layer; 3 marks a major phrase boundary; 4 marks a breath group boundary, where there is usually a pitch reset in between; and 5 marks a prosodic group boundary with a prolonged pause.

The code tier indicates the specific dialectal/regional language of speakers.

The following is an example from Peng *et al.* (2005: 262) on Pan-Mandarin transcription system of a sentence:

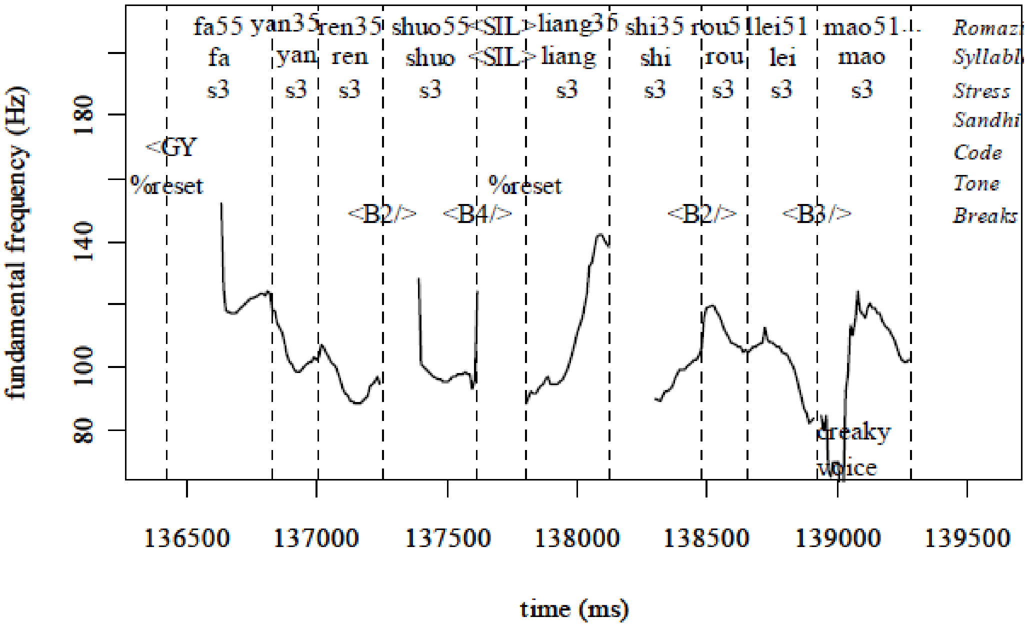


Figure 2.5: The transcription of a Mandarin utterance with Pan-Mandarin ToBI system

One problem of the ToBI labeling system is its subjective and categorical transcription method. Although the labelling in the system is dependent on the transcription work of experienced annotators, the subjective interpretation could be controversial. Furthermore, due to the “gradient nature” of speech, Xu & Wang (2005) contend that the transcription of break index cannot be dependent on a strict “categorical ranking of prosodic boundaries”.

2.5 Summary

The traditional studies in melodic prosody were mainly centered on non-tonal European languages. The prosodic discussions in those schools were focused on stylizing a number of contours representing intonation patterns. However, there exist several problematic issues with respect to the above schools of prosodic study.

In the intonation study of the British school, a number of typical nuclear tones were distinguished to represent the intonation patterns in speech. Due to those various principles decided by individual researchers in distinguishing nuclear contours, it is difficult to form a consistent intonation grammar. Moreover, the impressionistic nature of the British approach also prevents it from further development into an automatic model for intonation synthesis.

For the IPO approach by the Dutch school, the perceptual method in distinguishing intonation categories often tends to be subjective and variable across speakers, therefore, a consistent number of agreements have to be achieved before making any generalization of the melodic patterns, and supplementations always need to be made in face of divergence in spoken speech. Furthermore, according to Gili Fivela (2008), perceptual categories and phonological categories are not always in corresponding relations. Due to human’s limited perceptual ability and the gradient nature of acoustic signals, the thresholds between different intonation patterns, which are intentionally conveyed by speakers, may not be perceived categorically by listeners, and vice versa.

Concerning the ToBI system of the American school, according to Marotta (2008), the general limit of the AM approach and the ToBI transcription system is their lack of distinctive function in representing the scaling differences of prosodic varieties of the same language. Due to the limited number of categories in pitch accents and edge tones, the system is restricted in representative capacity. She proposed to improve the transcription system by increasing the number of tonal categories, i.e., introducing new pitch accents to represent the perceptual differences in pitch scaling, as seen in her studies on the prosody of Italian varieties (Marotta & Sardelli 2003, 2007; Marotta *el al*. 2004; Marotta 2008). She also commented that the use of the AM framework and the ToBI system should be attentatively conducted when applying it to different languages, as the system was originally proposed for the prosodic annotation of American English by Pierrehumbert (1980), but now it has been employed by reasearchers in the prosodic study of other languages, on the basis of "a supposed but still not proved equivalence" of the tonal categories in different linguistic systems.

The literature review of the theories and methods of important prosodic schools can provide a useful framework for the comparison of prosodic features in tonal and non-tonal languages. In the present study, the melodic form and function is explored in a typical tonal language, Mandarin, in which the functional aspect and formal aspect of prosody will be distinctly represented and discussed.

CHAPTER 3

Speech Corpora and Labeling Methodology

3.1 Introduction

The speech data employed in the study stem from Modern Mandarin. The data include utterances selected from casual-conversation corpus and read-speech corpus. The two styles of speech can reveal an overall picture of Mandarin prosodic features. This chapter presents details of the two speech corpora, the criteria of data selection and the segmental labeling method.

3.1.1 Daily-conversation corpus

The casual-speech corpus, named *Chinese Spontaneous Conversation Corpus* (Li *et al*. 2001, Li 2002), is a series of recordings of daily dialogues between native Beijing speakers. There are in total 12 dialogue units between 2 native speakers of the same gender, and each unit lasts around 1-hour. Each pair of participants in the conversation unit were not confined with any speech topic by the staff recorder, and the participants engaged in the speech interaction in a quite relaxed and daily mode.

Despite the fact that the speech in the corpus is varied in utterance length and speech rate, and is sometimes accompanied by laugh, sign, cough, background noise, and long silence, etc., the corpus presents us the behavior of native speakers in the natural context, especially their use of prosody for the linguistic, pragmatic and emotional reasons in mutual interaction.

3.1.2 Read-speech corpus

The data of clear speech style in this study stems from the *Chinese MULTEXT corpus* (Komatsu 2009). It is developed as the Chinese version of the multilingual prosodic database, which already contains the speech recordings of five languages, such as English, French, Italian, German and Spanish (Campione 1998, Campione & Veronis 1998). The data expansion to other languages is still in process, such as Japanese (Kitazawa 2004), the languages spoken in central and easten Europe (Erjavec 2004), and Korean (Kim *et al*. 2008). The same passages are translated into different languages, while the expressions of food, the names of cities and people, etc., are changed to adapt the passages into the local culture. Each passage has a real-life topic in content, such as talking on vacation experience, preparing for a long-distance travel, booking gifts for holidays, etc.

In the Chinese corpus, there are in total 40 passages and 10 Mandarin speakers, with most speakers reading 15 passages. The organization of passages read by each speaker is shown in the following table, where *f1, m1, f2…*on the vertical line represent the 10 speakers, while *o0-o4, o5-o9, p0-p4…*on the horizontal line represent the 8 sets of passages (the total 40 passages are grouped into 8 sets, with each set containing 5 passages), and the set which have been read by speakers are in grey colour, as seen in the following table from Komatsu (2009):

	Groups of passages							
Speakers	*o0-o4*	*o5-o9*	*p0-p4*	*p5-p9*	*q0-q4*	*q5-q9*	*r0-r4*	*r5-r9*
f1	▨	▨	▨	▨	▨	▨	▨	▨
m1		▨	▨	▨				
f2			▨	▨	▨			
m2				▨	▨	▨		
f3					▨	▨	▨	
m3						▨	▨	▨
f4	▨						▨	▨
m4	▨	▨						▨

f5				▨	▨	▨		
m5	▨	▨	▨					

Table 3.1: The organization of the *Chinese MULTEXT* corpus

175 passages have been read and recorded, with each passage read by around 4 speakers. The complete corpus includes 58 minutes' recording, with the mean length of each passage around 20 seconds.

3.2 Data selection and labeling

From the above two corpora, only fluent utterances with more than 7 syllables were selected as data for the present study, specifically, the ones without hesitation, repair or other fluency-interrupting factors, such as overlapping, laughter, signs, grunts, cough, among others. The requirement of utterance length with at least 7 syllables is under the same criteria for the selection of Italian data in the rhythmic study of natural languages, as seen in Bertinetto & Bertini (2010).

Besides the criteria on length and fluency of selected utterances, the neutral emotion is also an important condition. The selected data should not carry emotional prosody of anger, joy, fear, etc., as such kind of data is more relevant with the prosodic study of paralinguistic features, and severs for special pragmatic purpose in communication. In this study only the linguistic prosodic function in Mandarin speech is taken into consideration, therefore, the data selection are controlled to avoid strong affective emotions.

As the prosodic analysis is always required with the phonetic labeling of speech recordings, the data from the two speech corpora were annotated at the phonemic level. For the spontaneous speech corpus, 607 utterances produced by 7 speakers (4 females and 3 males) were manually labeled at the segmental level with SAMPA symbols. The correspondence between IPA representation and SAMPA symbols of Mandarin phonemes is presented in the following table:

Consonants:

Articulation manner		IPA	SAMPA
plosive	non-aspirated	p	p
		t	t
		k	k
	aspirated	t^{h}	t_h
		p^{h}	p_h
		k^{h}	k_h
affricate	non-aspirated	ts	ts
		tʂ	ts`
		tɕ	ts\
	aspirated	tʂ^{h}	ts_h`
		tɕ^{h}	ts_h\
		ts^{h}	ts_h
fricate		s	s
		f	f
		ʂ	s`
		ʐ	z`
		ɕ	s\
		x	x
		v (derived from *u*)	v
nasal		m	m
		ŋ	N
		n	n
liquid		l	l

Table 3.2: Mandarin consonants in IPA and SAMPA annotations

Vowels:

Tongue position	IPA	SAMPA
low	a	a
mid	ɚ	@`
mid-high	o	o
	ɤ	7
high	i	i
	ʅ	i`

	ɿ	i_d
	u	u
	y	y

Table 3.3: Mandarin vowels in IPA and SAMPA annotations

The following window shows the labelling of a Mandarin utterance in three tiers, phoneme tier, pin-yin tier and the utterance text tier.

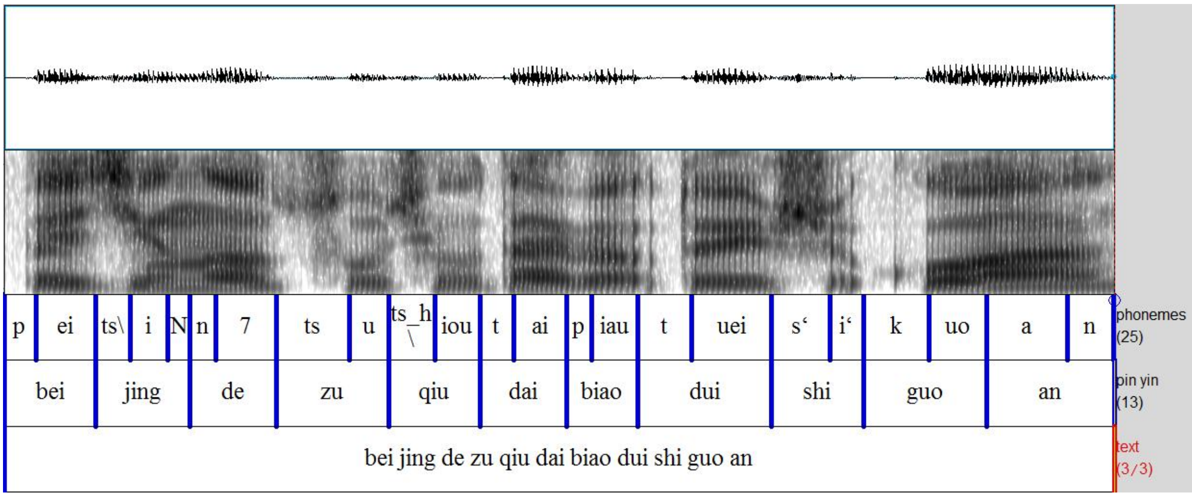


Figure 3.1: The annotation textgrid window of a Mandarin utterance

In connected speech, one can detect several unpredicted sound changes, especially deletions and shifts. Due to hypoarticulation, the phonetic output does not correspond to the speaker's phonological intention. Such phenomenon is related to speech rate and casualness of speaking style.

In the selected 607 utterances of the spontaneous corpus, there occur 61 cases of "irregular" vowel deletion, and 396 cases of "irregular" consonant deletion (323 onset deletions and 73 coda deletions). The classification of "irregular" deletion is defined in contrast to the classification of "regular" deletion of phoneme targets. The former deletion indicates the loss of segments due to speakers' hypoarticulation, whereas the latter involve the speaker's routine deletion of segments in casual speech pronunciations.

Regular deletions involve frequent words' allomorphs. Due to frequency of usage, these may become an intentional target. According to the data from spontaneous corpus, it is found that Beijing speakers often produce the bound morpheme 们 (a grammatical particle often used in plural personal pronouns) as *m* instead of the citational *mən*. For instance, *uo mən* (“we”) can be realized as *uo m*, *ta mən* (“they”) as *ta m*, *ni mən* (plural “you”) as *ni m*. Among the selected spontantous data, there are 63 cases in which the plural particle appeared, and only 11 cases where *mən* is not shortened. Therefore, *m* should be considered as an intentional deletion, due to the weak status of this particle in casual speech, except for instances of contrastive or emphatic focus on the plural particle.

A similar case involved the bound-morpheme 么 *mə*, which can be found in words such as 什么 *ʂən mə* (question word “what”), 怎么 *tsən mə* (question word “how”), or 那么 *n*a *mə* (conjunction “then”). The speakers often omit the vowel. There are 36 such cases, and 52 cases where the citational syllable is pronounced. In sum, *m* can be considered as an acceptable pronunciation for both *mən* and *mə*. Due to frequency of usage, certain words with deleted segments may become the intentional targets.

Occasionally, the phonological constitution of a segment is modified, namely, a phonological V is realized as a C. For instance, a high back vowel /u/, as part of the nucleus of an onsetless syllable, is sometimes produced as a labial-dental [v], especially in sequences such as /uən/, /uan/, /uaŋ/, /uai/ and /uei/. One and the same speaker can interchangeably articulate the two sounds. There are 32 such cases, while in 80 cases /u/ was not changed.

The following table details the data selected from the corpus. By “phonological segments” we mean the intended phonemes, by “phonetic segments” the phonemes actually produced, due to irregular deletions and shifts:

	Phonological segments	Phonetic segments
Vowels	10264	10203
Consonants	6976	6580
Total	17240	16783

Table 3.4: Number of phonological and phonetic segments in the selected data

The labeling of 607 utterances from the above spontaneous corpus was based on the author's manual work. The manual labeling of speech data was time-consuming and labor-intensive, therefore, an automatic tool for the phonetic transcription can largely reduce the aligment task.

For the annotation of the second Mandarin corpus, the read-speech *Chinese MULTEXT corpus*, an automatic transcription tool is explored. The tool is named as *SPPASS* (SPeech Phonetization Alignment and Syllabification), which is described in Bigi and Hirst (2012), and developed as an automatic tool for the alignment of speech sounds with utterance, word, syllabic and phonemic segmentations. The tool has been currently implemented in languages such as French, English, Italian, and Mandarin Chinese. The automatic tool can be applied to other languages by simply adding a pronunciation dictionary of the new language for the data training procedure.

The acoustic training is an important step to perform the automatic alignment in SPPASS, which requires a dictionary of words, with each word expanded with the constituent phones. In the Mandarin dictionary, there are in total 350 tokens stored, with the Pinyin form of each word/morpheme token having a corresponding annotation with SAMPA.

Pinyin form	Sampa
a	a
ai	ai
an	an
ar	a@`
ba	pa
bai	pai
ban	pan
bang	paN
bao	pau
bei	pei
ben	p7n
bi	pi
bian	pian

biao	piau
bie	pie
bing	piN
bu	pu
cai	ts_hai
…	…

Table 3.5: The Mandarin dictionary for SPPASS automatic alignment

The following window presents the annotation output of a Mandarin read-speech utterance from SPPASS implementation, with the “#” symbol indicating a pause in the speech flow.

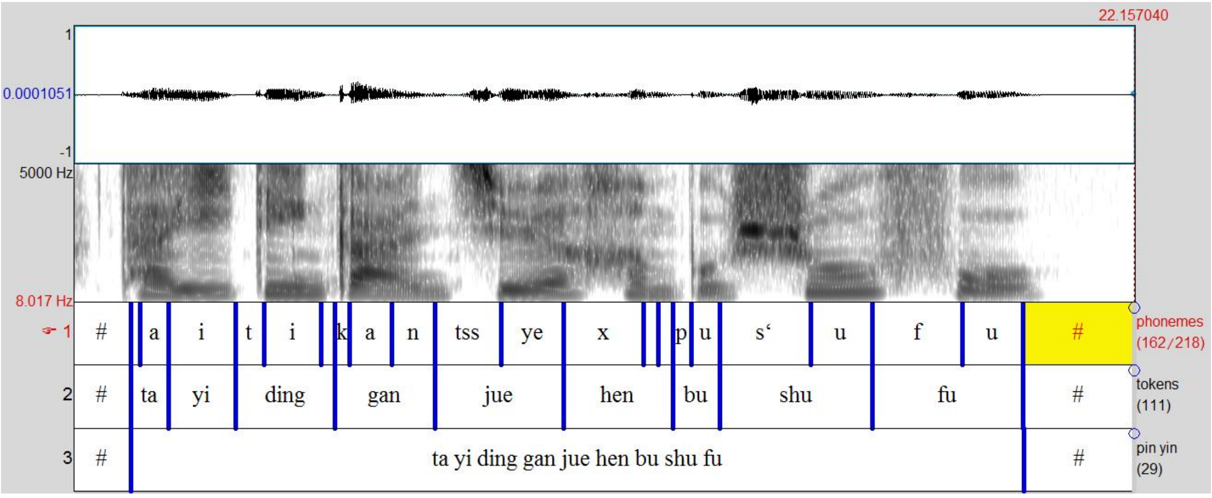

22.157040
1
0.0001051
-1
5000 Hz
8.017 Hz
1
# a i t i k a n tss ye x p u s‘ u f u #
phonemes (162/218)
2
# ta yi ding gan jue hen bu shu fu #
tokens (111)
3
# ta yi ding gan jue hen bu shu fu #
pin yin (29)

Figure 3.2: The annotation of a Mandarin utterance with SPPASS automatic alignment tool

The automatic labeling of the Mandarin corpus is subject to manual correction, with the result used for the further improvement of the data training of the alignment tool.

3.3 Summary

This chapter has provided some background information of the two speech corpora employed in the study. The utterance selection and the labeling methodology of the corpora at the phonemic level are also introduced.

In connected speech, it is normal to detect many sound changes, which include the segmental deviations, and tone sandhi at the supra-segmental level. According to Ladeforged & Johnson (2010), “a phoneme is not a single sound, but a name for a group of sounds”; the same is of tonemes, where each toneme, as an abstract phonological category, may include a number of phonetic tonal realizations. Chapter 4 discusses the tone sandhi phenomenon in details, together with rhythmic accent and intonation in the formal and functional representation of speech prosody.

CHAPTER 4

Three Functional Components of Mandarin Prosody

4.1 Introduction

In this chapter, three features involved in Mandarin prosody are discussed: lexical tone, accent and intonation. These three components closely interact in the prosodic form of Mandarin, and together play a salient functional role in speech communication.

4.2 Lexical tone

4.2.1 The functional role of lexical tone

Tone plays a salient role in tonal languages, as it yields lexical identity together with the segmental level. Words with identical segmental structures can be distinguished from one another by only commuting the attached tones. The mispronunciation of a tone in tonal language is analogous to the mispronunciation of lexical accent in certain non-tonal languages, like English or Italian.

4.2.2 The paradigmatic annotation of lexical tone

Tones are mainly distinguished in perception by differences in pitch, although it is believed that duration and intensity also contribute to the tonal identification. The annotation of lexical tones mainly depends on the description of pitch features.

For tonal transcriptions, Chao (1930) introduced a notation system of “tone-letters”. In this system, a tone is represented by its simplified pitch graph attached to the vertical reference line on the right. Several examples from Chao (ibid) are shown as follows:

	Tone symbol	Tone Name
Straight tone	˩˩	11
Circumflex tone	˩˥˧	153
Short tone	˨	2

Table 4.1: Tonal annotation

With this notation system Chao (ibid) also labeled the intonation of English by having each constituent syllable of an utterance attached with a tone symbol. He also examined the system in transcribing the lexical tones in Cantonese and Lhasa Tibetan. It can be noted that Chao used only one numeral in representing short tones, while he added two more numerals in representing complex tones. The difference in duration between tones was taken into account in his annotation of tones.

Since another work of Chao on tonal annotation published in 1968, this system became widely recognized as a classical tradition in transcribing Mandarin Chinese tones, and adopted in the International Phonetic Alphabet for tonal transcriptions. The following table shows the four canonical tones transcribed with tone symbols:

Tone	Description	Pitch	Tone symbol
Tone 1	High-level	55	˥˥
Tone 2	High-rising	35	˧˥
Tone 3	Low-dipping	214	˨˩˦
Tone 4	High-falling	51	˥˩

Table 4.2: Mandarin tones transcribed with tone-symbols

The relative pitch height and contour forms of the four canonical tones can also be seen in one graph representation as seen in Figure 1.2 (section §1.1.1.1).

The above description of tones is based on the pitch features in isolation forms: Tone 1 starts from a high pitch point and continues the same pitch level till the end; Tone 2 starts from the mid pitch level of the speaking range and rises to the upper limit toward the end; Tone 3 starts at a half-low pitch point and falls down to the bottom limit of the voice, then rises to the half-high pitch level; the two subsequent pitch movement of Tone 3 makes it longer in duration than the other tones. However as it shall be noted in the following text, Tone 3 is often actually produced with “half” of the tonal features, which can be either the “low-dipping” without final rise or the “half rise” without the initial falling. Tone 4 starts from a high pitch and falls abruptly to the bottom.

The neutral tone in standard Chinese is also represented with numerical values by Chao (1968), based on the assumption that the neutral tone (represented as “T0”) gets its pitch feature, determined from the preceding full tone. Therefore, T0 becomes a mid tone [3] when it is preceded either by a Tone 1 [55] or a Tone 2 [35]; T0 becomes a half high tone [4] when preceded by a Tone 3 [214]; T0 becomes a low tone [1] when preceded by a Tone 4 [51]. An illustrated example is given in Chen (2004: 22) with the neutral-tone clitic *de*, which functions as the nominalizer here, and derives its pitch feature from the preceding full-tone syllable as seen in the following Example 4.1:

zi	**de**	*the purple one*
214.	4	
hong	**de**	*the red one*
35.	3	
xin	**de**	*the new one*
55.	3	
da	**de**	*the big one*
51.	1	

Example 4.1: The tonal value of neutral-tone syllable, *de*

The tonal transcription with such tone-letter system can help provide a clear picture of the four Chinese lexical tones for the purpose of either non-native language acquaintance or further research purpose for general tonal annotation. However, due to the limited possibilities of five-point scale, when dealing with a large speech flow with higher requirement on annotation accurateness, the digits mode will not suffice to interpret the complex phonetic data. In the following section, a functional phonological system in representing lexical tones will be discussed. To begin with, the autosegmental property of tones is discussed.

4.2.3 The autosegmental property of tone

4.2.3.1 Basic principle

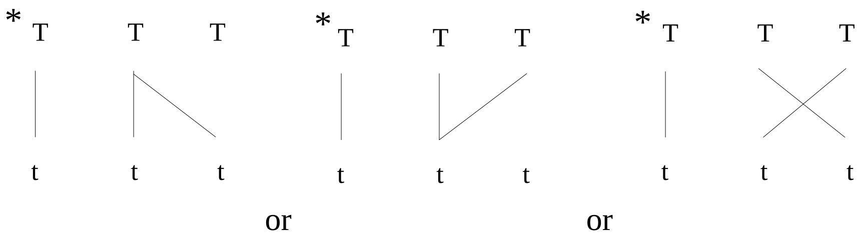
According to the autosegmental phonology (Goldsmith 1976), tone is not regarded as part of the segments; instead, tone is independent on a “tonal tier”, which is in parallel relationship with the segments on the “segmental tier”. The two tiers are linked by “association lines”, that is, the tonal features, labeled as H(igh), L(ow) and M(id) on one tier, are associated with the segmental matrices on the other tier under the principle of *Wellformedness Condition* as follows (ibid:27):

a) All vowels are associated with at least one tone;

b) All tones are associated with at least one vowel;

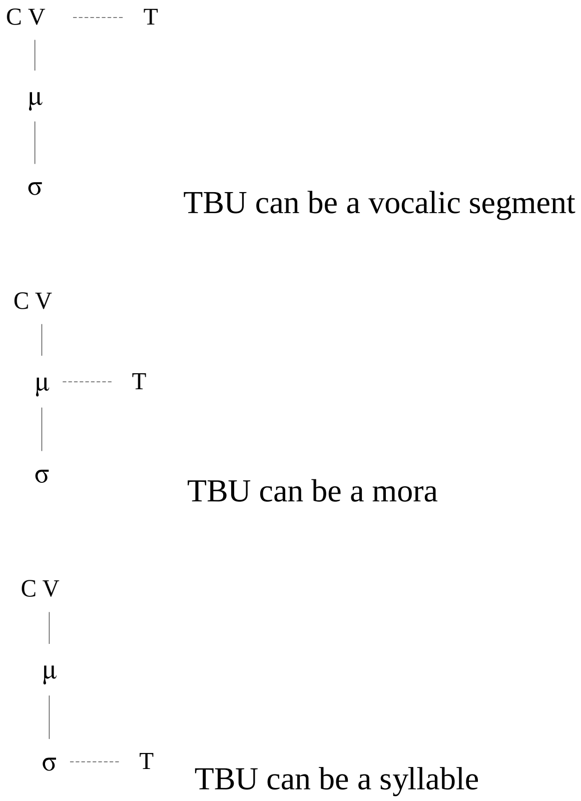
c) Association lines do not cross.

Accordingly, the following association between *T* (= “Tone”) and *t* (= “Tone Bearing Unit”) are examples which violate the conditions regulated in the above principle:



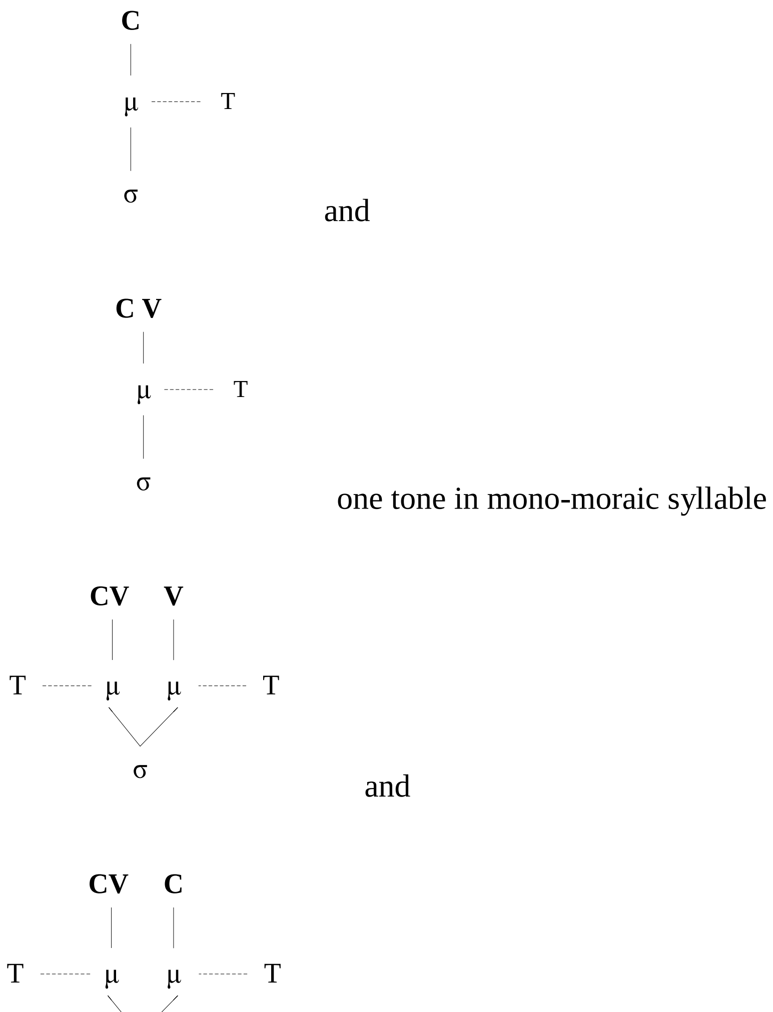
Example 4.2: Three cases of violating the principle of *Wellformedness Condition*

In fact, according to Yip (2002) it is not always clear whether the tone bearing unit (TBU) corresponds to a vocalic segment, or a prosodic entity, e.g., a syllable or a mora. In languages where each syllable bears only one tone, and the syllables are all mono-moraic with the CV phonotactic structure, the tone bearing unit of such languages can be either the vowel, or the mora, or the syllable, as illustrated in the following examples (ibid: 73), where *T* represents tone, *μ* represents mora and *σ* represents syllable.



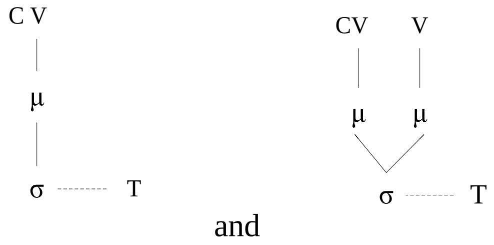
Example 4.3: TBU could be either a vowel, or a mora or a syllable in mono-moraic languages

For languages which have both light syllables (mono-moraic) and heavy syllables (bi-moraic), and the number of tones differ according to the syllable weights. Then the TBU of such languages should be the mora instead of the syllable, which can be clarified with the following examples (ibid: 73):


C
μ T
σ
and
C V
μ T
σ
one tone in mono-moraic syllable
CV V
T μ μ T
σ
and
CV C
T μ μ T
σ
two tones in bi-moraic syllable

Example 4.4: TBU is a mora in languages where tones are related with syllable weight

There are also languages where the number of tones remains the same regardless of the different syllable weights. The TBU of such language is a syllable.

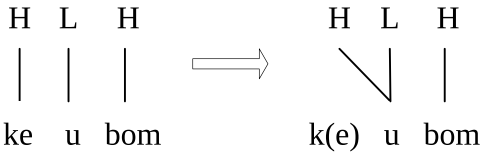


Example 4.5: TBU is a syllable in languages where tones are not conditioned by syllable weight

In standard Chinese, there are the following syllabic structures as (C)V, (C)VV, (C)VN, and (C)VVN, which can all bear the same number of tones regardless of the syllable weight. However, when confronted with the weak neutral-tone syllable, which often derives a **short** tone from the preceding stressed syllable and is more likely subject to the influence of phrasal intonation, in standard Chinese there exists the difference between such **short** tone on weak syllable and the **full long** tone on accented syllable. Thus, the TBU in standard Chinese is the syllable regardless of syllable weight among the accented syllables, while when taking into account unaccented syllables, the number of tones attached to a syllable changes depending on whether the syllable is accented or unaccented. Such cases belong to neither of the examples mentioned in Yip (2002). The particular case of Chinese tones may support the proposal that "quantity otherwise plays no part in Mandarin phonology" (Fox 2000:195).

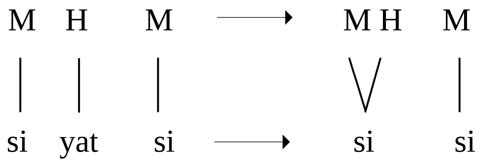
Considering how tones and TBU are associated with one another, the *Universal Association Convention* is formulated as mapping tones onto TBUs based on the principle of "from left to right" (Goldsmith 1976).

In the framework of autosegmental phonology, the autosegmental properties of tones provide support in analyzing the feature of tonal stability in tonal languages. When a vowel is lost, the associated tone survives and is attached to the neighboring vowel, as found in African languages of Lomongo (Lovins 1971), Efik (Fox 2000), and Pirahã (Yip 2002), as well as in two southern Chinese dialects, Cantonese and the Min dialect of Xiamen (Chen 2004). The following two examples respectively illustrate such tonal stability under segmental loss in languages of Efik and Cantonese:



(Fox 2000: 220)

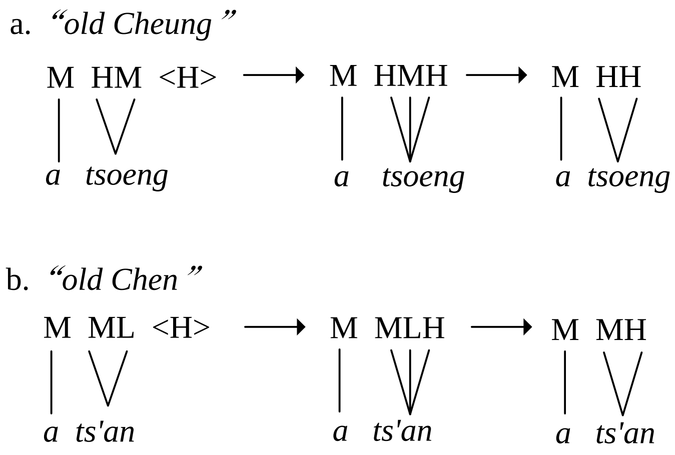
Example 4.6: The tonal stability in case of the loss of the vowel in langage Efik



(Chen 2004: 60)

Example 4.7: The tonal stability in Cantonese

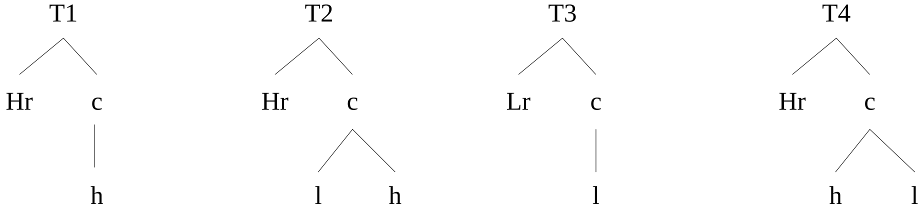
The phenomenon of floating tones in tonal languages is accounted for by the autosegmental phonology. As found in Mbam-Nkan (Fox 2000) and in the Cantonese dialect (Chen 2004), where a floating tone, not associated with a particular segment, is to “dock” onto a neighboring TBU. Such docking of a floating tone is subject to the tonotactic conditions of the language. As can be seen in the following examples of Cantonese (Chen 2004: 58), the inserted floating H tone displaces the existing M tone in case *a* of (4.8) and the L tone in case *b*, due to fact that Cantonese does not allow the existence of the complex tone, *HMH or *MLH. Therefore, HMH is simplified to HH, while MLH is simplified to MH.



Example 4.8: The floating tone in Shanghai dialect

Chen (2004) claims that in Mandarin Chinese such floating status of tone as independent from the aligned syllable can also be evidenced by the case of neutral-tone syllables, which derive the tonal value from the floating tone of the previous syllable.

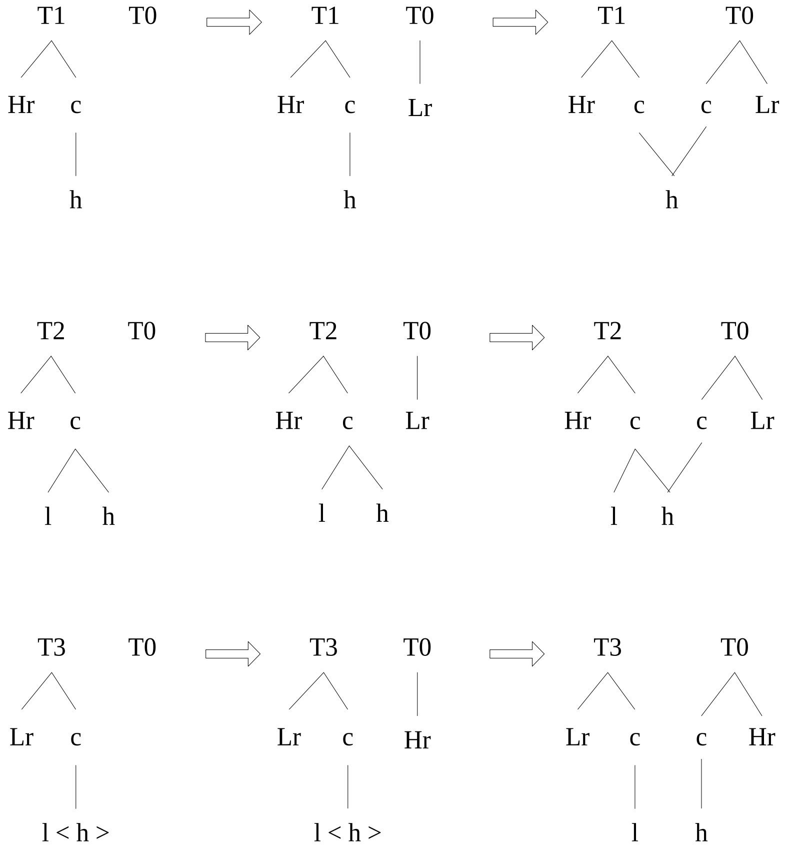
As Chen distinguished contour and register in the descriptions of Mandarin tones, his approach of geometrical representation of tones shall be first briefly introduced. According to his geometrical representation of tones, both register (pitch height) and contour (pitch shape) are employed in representing tonal property. Accordingly, T1, as a High level tone, is represented with a high register (Hr) and a high tone (h) for the contour (c) feature, thus [Hr, h]; T2, as a high-rising tone, is represented with a high register (Hr) and a low-high (l-h) contour tone, thus [Hr, l-h]; T3, as a low-dipping tone ("214") is represented with a low register (Lr) and a low (l) contour, thus [Lr, l]; T4, as a high-falling tone, is represented with a high register (Hr) and a high-low (h-l) contour tone, thus [Hr, h-l]. The geometrical representation of the four Mandarin tones is illustrated respectively in the following:

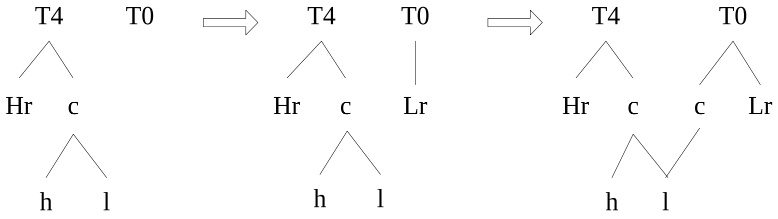


Example 4.9: Four Mandarin tones represented in geometrical structure

In a combination of a full tone syllable and a toneless syllable, Chen proposed that the surface contour feature (c) floats to the following toneless syllable; while the pitch register (Hr/Lr) of the toneless syllable is determined by a “polarity principle”, to be specific, the adjacent pitch registers are dissimilatory, thus, if a previous syllable has a Hr, then the subsequent syllable shall have a Lr, and vice versa. As T1, T2 and T4 are all represented with Hr in the geometrical representation, their following T0 shall be represented with a Lr; while the case of T3 is the contrast, where its register is a Lr, thus the following T0 shall be in a Hr. Chen claimed that such proposal of register dissimilation is subject to the underlying pitch-control mechanism, which requires the tonal targets to be farther apart in order for the “more exact fining tuning” (Chen 2004: 81).

The following figures illustrate that when a toneless syllable (T0) is preceded by different full tones, it derives the tonal feature from the floating tonal value of the left syllable. Such floating tone process is represented in the following:





Example 4.10: The floating tone process in Mandarin, where full tone floats its tonal value to the following toneless syllable.

Chen claimed that the predicted pitch value of a toneless syllable from such tonal assimilation process is quite close to the proposal of Chao (1968). However, people may question about the legitimate rule of "register dissimilation" in his approach.

Another important principle formulated in the autosegmental phonology is the *Obligatory Contour Principle*. It is proposed to prohibit identical tonal sequences in adjacent position on the tonal tier. The principle was introduced in Leben (1973), and stated in Goldsmith (1976:76) as follows:

> At the melodic level of the grammar, any two adjacent tonemes must be distinct. Thus HHL is not a possible melodic pattern; it automatically simplifies to HL.

Such principle meets the needs to explain the tonal phenomena observed in some languages, such as Margi (Kenstowicz 1994), Shona (Odden 1986), and the Tianjin dialect of China (Chen 2004), where the succession of identical tones in the underlying tonal string is avoided, with the tonal segments undergoing the alternation process such as tonal deletion, tonal dissimilation, or tonal absorption process.

However, such *Obligatory Contour Principle* has been debated as a controversial principle, due to the fact that it can be easily violated by many dispreferable cases from languages where in fact adjacent identical tones are commonly acceptable, such as in Kishambaa (Myers 1977) Therefore, the principle is not regarded as an universal rule, and “does not amount to a blanket prohibition” as concluded in Odden (1995).

In the following section §4.2.3.2, detailed introduction on a wide range of tone processes in Western African languages based on the independent tonal behavior will be presented. Such review of earlier studies on African tonal languages can provide helpful reference for the exploration of the tonal features in Mandarin Chinese.

4.2.3.2 Tone rules in Western African languages

In works by Hyman (1975, 2004), Hyman & Schuh (1972, 1974), and Schuh (1978), detailed discussions were given on tone processes in Western African languages. The authors explicitly turned away from the direct question on the underlying forms of tonal representation, and rather focused on how to map the underlying tonal representation onto the surface phonetic one, and to explore the possible universal tone rules. Tone processes in their opinion include both diachronic sound changes and synchronic phonological rules. Specifically, natural diachronic tonal universals can be generalized into five types, as: (1) downdrift, (2) low-raising, (3) spreading, (4) absorption and (5) simplification. The synchronic tone changes are attributed to five rules: (1) downstep, (2) shifting, (3) copying, (4) polarization, (5) dissimilation, (6) replacement and (7) displacement.

Among the five diachronic tone rules, ***downdrift*** and ***low-raising*** are considered as unavoidable phonetic universals, which have been widely studied in African languages, and been found in non-tonal languages as well. To be specific, ***downdrift*** occurs in a sequence of H-L-H, where due to the presence of a L tone, the second H tone becomes lower in pitch value than the first H tone, so that H-L-H (graphed as [¯ _ ¯]) may become H-L-M [¯ _ -].

In contrast, the ***low-raising*** tone process involves the raising of L tone at pitch level when it is followed by a High tone, thus, in a L-H sequence, the L tone is raised to a Mid tone due to the presence of the subsequent H tone, resulting a M-H sequence.

The third type of tone process is ***spreading***, which occurs when adjacent syllables have different tones and the earlier tone tends to extend its tonal feature rightwards, thus into a larger domain. Such spreading is never (or quite rarely) towards the left. The spreading tone process is actually a progressive assimilatory process, as can be observed from the Gwari language quoted in Hyman & Schuh (1974:88) as follows:

/ òkp á/ → [òkp**ǎ**] *"length"*

/s úkN ù/ → [s úk **û**] *"bone"*

Example 4.11: The tone-spreading process in Gwari

Absorption is a subtype of tone spreading process. It is the case in which a preceding syllable spreads part of its tonal feature onto the subsequent one, and such case should be under the condition that the subsequent syllable has a tone whose initial point has identical height with the end point of the tone in the preceding syllable. In the tone absorption process, only the early half of the contour feature is kept in the preceding syllable, while the second half feature is absorbed into the tone of the following syllable. Taking a two-tone sequence, R-H, as an example, where the Rising tone ends at the same pitch height with the starting point of the High tone, thus they meet the tone-absorption condition, and the R tone (which results from the combination of l-h) loses the second half of its contour feature, 'h', by having it absorbed onto the H tone of the second syllable; therefore, such R-H sequence turns out to be a L-H sequence. The same principle can be applied to the F-L sequence, and having it turn out as H-L sequence. It should be noted that in both tone spreading and absorption processes, no tonal targets is deleted, but rather it involves rightward expansion.

Another natural diachronic tone rule is ***simplification***, namely, contour simplification. All complex tones tend to be leveled, as can be seen from two samples from the language Gwari (ibid: 91):

[òj**ě**] + [bmy álō] → [òj**ē** bmy álō] *"cloth is good"* [ōz **â**] + [bmy ál ō] → [ōz**ā** bmy ál ō] *“person is good”*

Example 4.12: The tone-simplification process in Gwari

In Example 4.12, the tone [ě] and tone [â] in the above two cases is respectively leveled as tones [ē] and [ā]. However, Hyman & Schuh (1972, 1974) claimed that the implementation of the simplification rule may actually vary across languages.

The above diachronic rules are natural tone processes, due to the fact that they meet the phonetic nature of sound changes, but there are also synchronic rules in explaining the tone process, with rules which may seem quite unnatural but in fact quite natural in the historical phonological account conditioned by the language grammatical information.

Among the synchronic rules of tonal reconstruction, ***downstep*** indicates the loss of a low tone when it is between two high tones. Such example can be found in language Twi (Schachter and Fromkinn1968), as quoted in Hyman & Schuh (1974: 92):

[m é] + [ɔ̀b ó] → [m éˈb ó] *“my stone”*

Example 4.13: The downstep process of tone in Twi

In Example 4.13, the low tone [ɔ̀] when situated between two high tones, [é] and [ó], it gets deleted.

According to Hyman and Shuh (1974), although both tone processes ***downstep*** and ***downdrift*** involve cases where L tone is changed when situated between two H tones, it is not always possible to explain downstep as straight forwardly as the downdrift phenomenon.

Tone ***shifting*** involves the shifting movement of a tone from one syllable to the next, as illustrated with an example from the language Mbui Bamileke in Hyman and Shuh (1974:94):

/lɔ̀ɔ́ + /bə̀sə́η/ → [lɔ̀ɔ bə́*sə́η*] *“look for the birds”*

Example 4.14: The tone shifting process in Mbui Bamileke

In Example 4.14, the high tone [ɔ́] in the first word has shifted rightwards to the second word, and replaced the low tone [ə̀].

Tone ***copying*** refers to the case where a syllable without underlying tone obtains the tone from its adjacent and generally preceding syllable. Such toneless syllable is usually a grammatical morpheme, such as a pronoun in Western African languages. An example of tone copying can be seen as follows from Hyman and Shuh (1974: 96):

/lɔ̀ɔ́/ + /wa/ → [lɔ̀ɔ̀w á] *"look for me"*

Example 4.15: the tone copying process

In Example 4.15, the toneless pronoun /wa/ has copied the high tone from the preceding syllable [ɔ́], thus resulted in a [w á]. It seems interesting that the tone of [ɔ́] has changed to a [ɔ̀].

Polarization indicates that a toneless morpheme receives the opposite tone of the neighboring syllable, e.g., a toneless grammatical morpheme shall take a H tone when it is preceded before a L tone.

Tone ***Dissimilation*** involves a tonal change due to the presence of an identical tone in the adjacent syllable.

Tone ***Replacement*** is often used as a grammatical device in distinguishing grammatical differences. For example, in the Igbo language, an imperative is expressed by replacing the tone of the original syllable of the host verb with a low tone, and adding a suffix (ibid: 102):

/r ɨ́ *"to eat"* → /r ɨ̀ é/ *"To eat!"*

Example 4.16: The tone replacement of [ɨ] in impertative expression in Igbo

Finally, tone ***displacement*** refers to the case in which tonal contrasts are realized several syllables away from their original syllable position. It can be seen as follows from (ibid: 103):

/m ánj **è** mənəm ón ó/ → [m ánj **é** mənə ꜜm ón ó]

Example 4.17: The tone displacement process of tone [è]

In Example 4.17, the expected downstep process of the low tone [è], when between two high tones, [mán] and [mə́], is displaced two syllables after, instead of the expected downstep [mánj(**è**)ˈmə́nə́m ón ó].

As a review of the tone processes described in Hyman & Schuh (1974), the two authors summarized a non-exhaustive list of possible tonal reconstructions in sequences of bi-tones and tri-tones as follows:

Tonal sequences	**Derived tonal sequences**
H-H	Same
L-L	Same
H-L	H-F (***spreading***)
L-H	L-R (***spreading***), or further L-M (***simplification***)
H-H-H	Same
L-L-L	Same
H-H-L	H-H-F (***spreading***)

H-L-H	H-F-H (***spreading***), or further H-H-D (***simplification***);
	or further H-H-H (***simpification***);
	or further H-H-R (***spreading)****;*
	directly H-D-H (***downstep***);

H-L-L	H-F-L(***spreading***), or further H-H-L (***simplification***)

L-H-H	L-R-H (***spreading***) or futher L-M-H (***simpification***);
	or further L-L-H (***absorption***) , or even →L-L-R (***spreading***);

L-H-L	L-R-L(***spreading)***, or futher L-M-L (***simplification***); or futher L-L-F (***spreading and absorption***)
L-L-H	L-L-R(***spreading***), or futher L-L-M(***simplification***)

Table 4.3: A summary of possible tone processes with different tonal sequences

As seen in the above discussions, with evidence from tonal languages in Western Africa, Hyman and Schuh have generalized various types of tone processes. They distinguished two types of tone changes, one attributed to the universal phonetic nature of tones, the other manipulated by the phonological rules derived from historical innovation. Such study also inspired linguists to explore the tone processes in other tonal languages, such as Chen (2004), in which the rich and complex tone patterns across Chinese dialects, such as Tianjin and Cantonese are studied by examining the possible tone processes.

However, due to the fact that different tone processes may occur all at once, it arouses difficulty in deciding which operation shall take place first as seen in the above Table 4.3 . This leads to a one-to-many relation between the underlying tones and the surface tone forms, which makes it intricate in studying the complex tone phenomena through the derivation of tone processes.

4.2.4 The phonological representation of tone: level vs. contour systems

In earlier studies of tones, contour tones are represented by either the *level* system, as employed in the autosegmental phonology (Goldsmith 1976), in which contour tones are decomposed into levels, such as *High*, *Mid* and *Low*; or by the use of *contour* system, as seen in the discussion of tone processes in Western African languages (Hyman and Shuh 1974), where *Rising* and *Falling* are mostly used in the contour tonal representation.

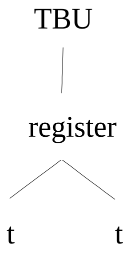
In the above section §4.2.3.1, besides the two mentioned representations of tones, which focus on the features of pitch movement, some linguists, such as Chen (2004), incorporated another entity in the tonal representation, the pitch register, as seen in Example 4.9 and Example 4.10. Such approach to tonal analysis can also be found in the earlier studies of Yip (1980, 1989), Bao (1990b) and Duanmu (1990), the property of each tonal category consists of two feature entities, that is, pitch register (height) and pitch contour.

Yip proposed that the *register* entity defines the pitch range into two halves, as [+Upper] and [-Upper], while the tone features, distinguished as [+high] and [-high], sub-divide each register and result in four tones as follows (Yip 2002: 43):

+Upper	+ high	55	extra-high
	- high	44	high
-Upper	+ high	33	mid
	- high	11	low

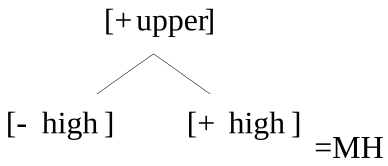
Example 4.18: Four tones distinguished by [±upper] registers

In Yip's tonal representation, pitch register dominates the alignment branching for pitch contours, with the relationship of the two entities, *register* and *t* ("tone feature") represented as in the following model:

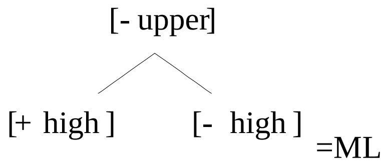


Example 4.19: subodinative relationship between *register* and *t*

Therefore, a tone with a high rising feature is represented by the model as:

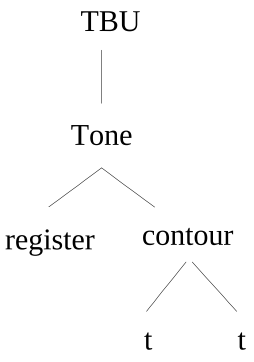


While a low-falling tone is represented as:



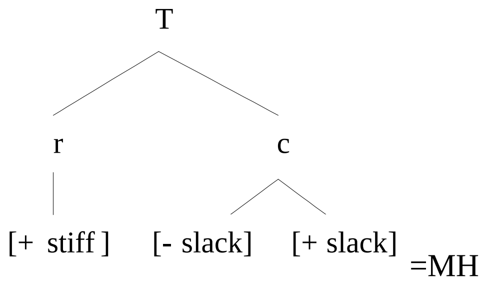
Example 4.20: The representation of a high-rising tone and a low-falling tone in Yip's study

In the model proposed by Bao (1990b), pitch register and pitch contour are posited in parallel as sister nodes aligned with the whole TBU, as can be seen in the following representation:



Example 4.21: Parallel relationship between *register* and *contour*

In Example 4.21, register and contour are two sister branches of a tone. According to the framework of such model, a high rising tone is represented as:



Example 4.22: The representation of a high-rising tone in Bao's study

Bao (1990b) used the term [±stiff], equivalent to Yip's [±upper], while [±slack] is equivalent to [±high].

In the above two mentioned register approaches, contour tones are in fact decomposed into level ones, with the pitch height defined by either the upper or juxtaposed register value. As mentioned in the section §4.2.3.1, the four standard Chinese tones are represented in Chen (2004) as, Tone 1 [Hr, h], Tone 2 [Hr, l-h], Tone 3 [Lr, l] and Tone 4 [Hr, h-l], where the contour tones, Tone 2, Tone 4 and Tone3 are represented as the sequences of level tones.

Linguists who are in favor of the level system in tonal representation, such as Anderson

(1978), claimed that contour tones should always be decomposed into levels in representation, as contours have no legitimate status in tonal phonology.

In contrast with the above claim of regarding contour tones as the juxtaposition of level tones, there are arguments by Pike (1948) and Newman (1986), who claimed that contour tones should be regarded as unitary feature of tonemes, uninterruptable by boundaries. According to Pike (1948), contour tones are the "basic tonemic units", and it is not necessary to align the initial and end points of contour tones with the level tones.

The debate between level and contour systems in tonal representation cannot be settled by claiming that one is appropriate while the other is not, as no strong evidence has been found within any phonological or phonetic theory. In general, level systems are less restricted than contour ones, with reduced number of tonal segments in the tonal inventory.

In this study of Standard Chinese, both the features of tones and intonation are discussed, with the pitch contour of utterances modeled and evaluated by means of the synthesis system. Therefore, a level system will be adopted in this study, due to the fact that it is sufficient for tonemic distinctions in both tonal and intonational representations.

In my study, two level tone targets, *High (H)* and *Low (L)* are employed to represent the underlying pitch movement of both tones and intonation.

The four Chinese tones may thus be represented as sequences of level tones by having High/Low aligned with the onset and offset pitch points, as seen in the following table:

Tone 1	[55]	HH
Tone 2	[35]	LH
Tone 3	[214]	LLH
Tone 4	[51]	HL

Table 4.4: The underling phonological representation of four Mandarin tones with the *Level* system

It is to be noted that in previous studies, such as Chen (2004), the level *mid* tone is also employed, accordingly in his study for the representation of Tone 2, the high-rising pitch feature was annotated as *MH* in order to highlight the high pitch register of the tone. However, I assume that what is salient for tonal contrast in actual speech is the overall rising and falling pitch movement of tones, rather than the high start or how far it can rise or fall. In the speech flow, a canonical high tone may not necessarily remain in the upper register of the pitch range, nor have higher phonetic values than a canonical low tone. Sandhi tones and canonical tones are not distinguished by register values, but rather by the tendencies of pitch movement. Therefore, the distinction of pitch register in tonal contrast is not employed for the tonal representation in the present study, but rather the pitch movement of tone is highlighted in annotation.

In the above Table 4.4 , Tone 1 as a level tone is represented by a sequence of two identical level targets, *HH* instead of only one *H*. Although it makes no difference in representing the static pitch level of Tone 1, it is believed that such bitonal-target sequence can be more appropriate than a mono-target representation in making the *timing slot* seem equivalent with those of other dynamic tones, such as Tone 2, *LH* and Tone 4, *HL*.

The canonical form of Tone 3 is represented as *LLH* with three level-tone targets, which is due to the general agreement that Tone 3 is longer in duration than the other three tones in isolated productions. The low-dipping feature of Tone 3 is highlighted with two adjacent *L* level targets to distinguish from the *LH* of Tone 2. In the latter sections of the study, it can be seen that Tone 3 is mostly realized as a “half T3” without the terminal rising glide, particularly, *LL* in most tonal contexts of connected speech, except the cases before another T3, or at the sentence boundary or under the phrasal focus.

The distinction of pitch register in tonal contrast is not employed for the tonal representation in this study. It is assumed that in actual speech flow, a canonical high tone may not necessarily remain in the upper register of the pitch range. Therefore, only the pitch movement of tone is highlighted in annotation.

However, in different tonal contexts, all of the citational forms of tones as annotated in Table 4.4 present tonal variations, as discussed in details in the subsequent section.

4.2.5 Tone sandhi in context

In connected speech, the tonal features of syllables can be quite different from their underlying forms as uttered in isolations. The deviational phenomena of tones are called "tone sandhi", where "sandhi" is a Sanskrit word which means "joining". Tone sandhi can be either "phonologically- or lexically-conditioned tonal alternations" (Peng *et al*. 2005).

Tone sandhi is contrasted with citation tones, with the latter referring to the tones mostly produced in isolated form or at the prepausal location, e.g., at the end of an utterance. The deviant forms of tones are found when located in the tonal contexts, especially when the tone bearing syllables are juxtaposed with each other. The deviational degrees of the allotones are sensitive to factors such as the prosodic effect, the functional load of the syllable in lexical-syntactic context, the speech style, the speech rate, etc. For instance, dynamic tones can be smoothed out as a flattened pattern in fast speech; a citational rising tone LH can be actually found with a falling pitch movement, when it is overcome by a preceding HL or a following HL tone, or by the declining intonation contour.

4.2.5.1 Tone variation rules

According to Boersma (1998), "while there is function in phonology, there is also organization in phonetics". Some common tone changes found in speech are included in the sandhi rules by Chao (1948, 1968: 27-28). The following are his major conclusions:

a) The low falling-rising form of Tone3 [214] loses its final rise and becomes a low-falling tone before any tone except another Tone3, namely, [214] → [21] / __ [X], X is a non-T3 tone;

b) Tone3 becomes a rising tone when it is before another Tone3, namely, [214] → [14] / __ [214];

c) Tone2 changes to Tone1 when it is at the middle position of a trisyllabic phrase, whose initial tone has a high-level ending, such as Tone1 or Tone2, namely, [35] → [55] / 5] __ [X'] X' is any full tone;

d) The pitch movement of Tone 4 does not fall to the bottom level when situated before another T4, but rather falls to a mid pitch level, namely, [51] → [53] / __ [51].

Another tone rule specifies the value of the neutral tone, which has been presented in the early section §4.2.2. In connected speech, the pitch feature of a neutral tone greatly depends on its preceding full tone, and appears as the post extension of pitch movement after a dynamic rising or falling tone. Chao (1968) made an estimation of the pitch value of a neutral tone according to the immediately preceding tone as follows:

T0 → half low tone / T1_; T0 → half high tone / T3_;

T0 → middle tone / T2_; T0 → low tone / T4_.

Such tone rule is also similar to the tone spreading process of Western African languages, discussed in section §4.2.3.2. In the following Table 4.5, the tone spreading process in Mandarin is shown as a list of Mandarin bi-tonal combinations of a full tone and a following neutral one, where the preceding syllable spreads part of its tonal feature onto the neighboring neutral-tone syllable, which results in the value of the tonal sequence.

Tone combinations	Underlying representation	Tone spreading process
T1+T0	HH+Neutral	HH+H
T2+T0	LH+ Neutral	LM+H
T3+T0	LLH+ Neutral	LL+H

T4+T0	HL+ Neutral	HM+L

Table 4.5: Tone spreading process in Mandarin

A neutral tone in Chinese is always aligned with an unstressed syllable which has shorter duration and lower intensity. As the timing slot of a neutral tone syllable is shorter than that of a non-neutral tone, in the table 4.5 it is shown that a normal tone syllable consists of two tonal targets, while a neutral-tone syllable receives only one tonal target spreaded from the preceding syllable.

This agrees with the Chinese classical literature in regarding the pitch feature of a neutral tone as depending on its preceding full tone, and appearing to be the post extension of pitch movement after a dynamic rising or falling tone.

In speech, besides the tone sandhi under phonological prediction, there are also morphologically-conditioned sandhi phenomena, such as morphemes like *yi* [55] (“first, one”), *bu* [51] (“no, negation meaning”), *qi* [55] (“seven”), *ba* [55] (“eliciting particle”), which have tonal alternation according to the different lexical meanings carried with them.

For example, when *yi* means “the 1st” the cardinal meaning, its tone always keeps the original Tone 1 form despite the influence from the surrounding tones, like in the words, yi[55]-ban[55] (“1st class”), yi[55]-ceng[35] (“1st floor”), di[51]-yi[55]-ming[35] (“the first one”), yi[55]-dui[51] (“1st team”), etc. By contrast, when *yi* means “one” the common ordinal number, its tone often changes to Tone 4 or Tone 2, like in words, yi[51]-ren[35] (“one person”) instead of *yi[55]-ren[35], and yi[35]-fen[51] (“one part”) instead of *yi[55]-fen[51].

Such kind changes are morphotonemic alternations, also called by many scholars ‘tone sandhi’, but such kind of morphological sandhi process is different from the phonological tone rules in this study.

However, when confronted with the actual tonal data in running speech, various allotonal phenomena in Chinese are beyond the interpretive capacity of general tone rules. For instance, the canonical contrasts of different tones can be smoothed out as a flattened pattern in fast speech; a citational rising tone, LH can actually be realized as a falling tone, HL when it is over-manipulated by a preceding HL or a following HL tone, or by neither of the above two but the declining intonation contour.

Due to the unpredictable property of tones in connected data, tone sandhi analyses, according to Chen (2004), should focus on exploring the hidden principles under the bewildering surface forms. In the following sections, I will discuss the relevant factors involved in tone sandhi process and their interactions, and try to explore a potential way in accounting for the surface deviations from the underlying input of tones.

4.2.5.2 Tonal co-articulation subject to physical limit

Xu (1999) claims that the deviant features of lexical tones in context are due to the speakers' physical constraint in articulation. As the vibration rate of the vocal cords is subject to certain physical limits, the completion of each articulation behavior requires a minimum duration. For instance, when preceded by a tone with conflicting pitch direction in an early syllable, the current tone can not present its canonical pitch feature if the aligned syllable is too short in duration, as it also takes time for laryngeal articulators to change from one state to another.

Multisyllabic words and phrases involve a more complicated tone process, due to different surroundings. As summarized in Xu (1993, 1994, 1997), there could be two types of tonal contexts, namely, "compatible context" and "conflicting context" distinguished according to the similarity in contour direction and pitch height of adjacent tones. The following is an idealized schematic representation of the compatible and conflicting context:

Compatible context

¯\ _	/ \ /	_ / ¯	\ / \	¯\ /	/ \ _	_ / \	\ / ¯
HFL	HFR	LRH	FRF	HFR	RFL	LRF	FRH

Conflicting context

¯/ _	/ / /	_\ ¯	\ \ \	¯/ /	/ / _	_\ \	\ \ ¯
HRL	RRR	LFH	FFF	HRR	RRL	LFF	FFH

H: high tone or T1; L: low tone or T3;
R: rising tone or T2; F: falling tone or T4.

Example 4.23: Compatible context vs. conflicting context

Xu concluded that the degree of deviation of a tone from its underlying form is quite context-sensitive, as the tonal context could exert "carry-over effect" on a following tone and "anticipatory effect" on a preceding one. From Xu's observation of F0 data, the deviational degree of a tone in a conflicting context is much greater than the identical one in a compatible context. For instance, under conflicting tonal environment, a rising tone could actually realize as a tone with a slight falling contour, or sometimes present as a flattened form, while a falling tone could actually show a slight rising configuration. In certain cases, the pitch shape of a dynamic tone can become completely reversed in direction under a conflicting tonal surrounding.

In fast speech, the two contexts are found to exert a greater effect on tonal deviation. According to Xu (1999), with increased speaking rate, more compromise is achieved between adjacent tonal features due to great gestural co-production. In extremely fast speech, adjacent tones often overlap to a flattened contour.

Based on the fundamental principle of human's articulatory mechanism and language's communicative functions, Xu (2009) proposed two kinds of timing in speech, namely, obligatory timing and informational timing. The first one indicates the obligatory time required for articulation movement, and the latter indicates the time encoded for communicative functions, for example, the emphatic focus and boundary markers in speech are made to stand out through the lengthened time duration.

Xu observed that the deviational features of lexical tones in Chinese speech are unlikely due to economy in effort distribution (Lindblom 1990), but rather to the speakers' physical constraint in articulation. Therefore, even with full muscular force, articulators cannot move faster. The completion of each articulation behavior requires a minimum duration, which is the obligatory timing in speech. For instance, when preceded by a tone with conflicting pitch direction in an early syllable, the current tone could not present its canonical pitch feature if the aligned syllable is too short in duration, as it takes time for laryngeal articulators to change from one state to another. Therefore, the minimum duration for pitch movement can well explain the phenomena that in fast speech, dynamic tones present only flat patterns on F0 contour.

According to Xu's perception experiments, tonal context plays an important role in people's perception of tones. Under a compatible tonal context, the tonal contour of the middle syllable is close to its underlying form, so it is always easy for subjects to recognize the tones correctly, while in a compatible context, the F0 performance of the tested tone is greatly distorted, so the subjects have to make judgment relying on the surrounding environment, but the severe deviation cannot always be compensated by the context. As it will be discussed in a later section, speakers in actual speech often have to use other cues, such as syntactic structure, prosodic organization and pragmatic context to compensate for tonal loss in mutual interaction.

Summarizing Xu's study on tonal contexts, it can be seen that the mismatching between the surface phonetic patterns and the underlying phonological tonal category is, to a certain degree, conditioned by the human's physical limits in articulation. Such phonetically motivated tonal phenomena have been widely discussed and shown to be an important factor in causing the surface output deviating from the canonical underlying input. Therefore, it needs no further evidence from this study to prove this universally-acknowledged mechanism.

For the study on the divergent tone sandhi phenomena, Shih (1986) distinguished two categories of tone processes, obligatory and optional tone process. The former refers to the process conditioned by default phonological tone rules. Optional tone sandhi refers to an unpredicted tonal change, triggered by casualness of speech style, fast speech rate, or related with grammatical and pragmatic contexts.

The distinction between obligatory tone sandhi and optional tone sandhi provides a plausible direction in analyzing complex tonal variations, as obviously most tone sandhi phenomena in Mandarin often take place subconsciously and even naturally in native speakers' connected speech. Such tone sandhi should be attributed to the category of optional tone process influenced by a variety of factors. In the following of the present study, it is demonstrated that in Mandarin Chinese besides the phonetically motivated tonal changes based on differences in speech rate and style; the optional tone processes are more influenced by syntactic, prosodic and pragmatic factors of the speech.

4.2.5.3 Tone sandhi domain

In connected speech, there exists a vast diversity of tonal undershoot. Despite such bewildering surface forms of tones, it is found that tone sandhi phenomena in many cases can be approachable and even predictable by locating the rational domains in which tone sandhi processes take place. In the study of Yip (2002: 120), it was discussed that tone coarticulation takes place mostly within a certain prosodic-syntactic unit and between "domain partners". Most studies on tone sandhi domain discussed the interaction of syntactic structure and prosodic phrasing with tone sandhi process.

a. Tone sandhi and the syntactic structure

This has been studied by Wu (2004), Zee (2004), and Cheng (1970, 1973) in the interaction between tone sandhi phenomena and the utterance syntactic structure. A conclusion has been agreed on that in the majority of cases, the formation of tone sandhi domain in speech is obligatory and sensitive to the surface syntactic bracketing, however not necessarily isomorphic to syntactic units.

Syntactic structure, as defined by Rossi (2004) is, "the linear organization and hierarchy of syntactic constituents, and the relationships by which syntactic functions are defined."

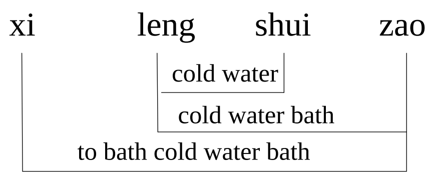
According to Wu (2004), such grammatical level plays an important role in segmenting utterances in speech into various syntactic domains, in which tone sandhi starts from the immediate constituent structure, and goes on from underlying forms to surface forms by steps in the order of grammatical constraints. Such tone sandhi process is implemented in successive steps one after another just like the Domino rule. He gave an example of a phrase, which consists of four Tone 3 syllables as follows:

Xi	leng	shui	zao
T3	T3	T3	T3
LLH	LLH	LLH	LLH
To bathe	cold	water	bath
Verb	adj.	noun	noun

'To take a cold bath.'

Example 4.24: A phrase, *xi leng shui zao*, with four Tone 3 syllables

The grammatic structure of the phrase is:



Example 4.25: The syntactic structure of the phrase

With the presence of four adjacent Tone 3 syllables in one phrase, its surface contour is derived from the successive tone sandhi processes as follows,

$$\text{T3 T3 T3 T3} \xrightarrow{\text{Step 1}} \text{T3 } \underline{\text{T2 T3}} \text{ T3} \xrightarrow{\text{Step 2}} \text{T3 } \underline{\text{T2 T2 T3}} \xrightarrow{\text{Step 3}} \text{T3 } \underline{\text{T2 TR T3}} \xrightarrow{\text{Surface contour}} \underline{\text{T3 T2 TR T3}}$$

Example 4.26: Successive T3 process conditioned by the syntactic structure

It can be seen in Example 4.26 that the phonological tone sandhi of Tone 3 starts first within the minimal domain of the compound, *leng3-shui3* (“cold water”), thus results in *leng2-shui3*; then the domain of tone sandhi proceeds to the larger domain of the trisyllabic compound, *leng2-shui3-zao3* (“cold-water bath”) where another round of T3 sandhi process occurs, which leads to *leng2-shui2-zao3* in step 2; Wu claimed that in step 3, due to the phonetic nature of cophonation, the second Tone 2 in *leng2-shui2-zao3* becomes a transitional (TR) tone, with its contour assimilated by the neighboring tonal contours; finally the surface contour of the phrase is generated after three successive steps by having the sandhi domains correspond to different levels of syntactic structures in each step.

Wu highlighted in his study the interaction between tone sandhi and grammatical structure by assuming that a phrase with the same tonal constituents but different syntactic structures may result in different surface contours.

The relationship between tone sandhi and syntactic organization was also discussed by Zee (2004), who claimed that in the majority of cases, the formation of tone sandhi domain in speech is obligatory, which is determined by the utterance syntactic structure, and may also be motivated by the semantic meaning. However, he clarified that tone sandhi domains are not necessarily isomorphic with syntactic domains. In his study on the sandhi rules of the Shanghai dialect, he found that tone sandhi often takes place across words of different syntactic categories, with function words in most cases prosodically affiliated with their preceding or following content words. Zee listed in total 29 cases of the relation between the prosodic host and the immediate neighboring prosodic clitics in the formation of tone sandhi domain as follows (2004: 523-524):

a) The subject noun, verb and object noun do not form a tone sandhi domain with each other;

b) A noun and the preceding adjective obligatorily form a tone sandhi domain;

c) A noun and the post-nominal genitive or postposition obligatorily form a tone sandhi domain;

d) …

The summaries of Zee (2004: 523-524) can be found in the appendix IV of the present dissertation. His study offers great insight on the interplay between tonal implementation and syntax, as tone sandhi domain takes partial reference from the surface syntactic bracketing.

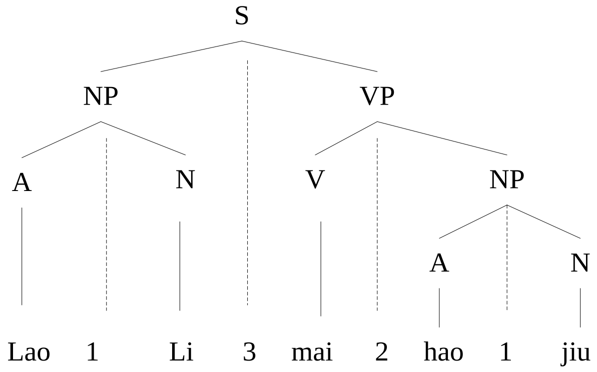
Cheng (1970, 1973) assumed that tone sandhi is sensitive to the strength of syntactic junctures by providing the following example:

Lao	Li	mai	hao	jiu
T3	T3	T3	T3	T3
LLH	LLH	LLH	LLH	LLH
Old	Li	buys	good	wine

' Old Li buys good wine.'

Example 4.27: A sentence, *lao li mai hao jiu*, with five Tone3 syllables

The syntactic structure of the sentence is:



Example 4.28: The syntactic tree of the sentence

In example 4.28, the numerals 1 to 3 indicate the hierarchy of syntactic junctures. Based on Cheng's idea, tone sandhi process obligatorily occurs among constituents which are syntactically close. Therefore, the first cycle of tone sandhi takes place between the syllables with the lowest strength of juncture (indicated with numeral "1"), namely in [Lao Li]$_{NP}$ and [hao jiu]$_{VP}$, thus, we derive the tonal pattern of the utterance as LH+LLH+LLH+LH+LLH;

then a second cycle of tone sandhi takes place between the remaining Tone 3 syllables, namely, between “li” and “mai” with the juncture strength of “3”, finally resulting as LH+LH+LLH+LH+LLH. Such surface contour pattern resulting from the prediction of tone process related with syntactic junctures corresponds to the default and natural tonal articulation in normal speech.

However, Cheng assumed that there could be other surface contours when the speech rate is changed. Tone sandhi is sensitive to tempo, as it was shown that in very slow speech the LLH tone dissimilatory process could be blocked at juncture 3 between subject and predicate, but still obligatory at juncture 1. When speech rate is very fast, tone sandhi process occurs simultaneously over the entire utterance, with all LLH changed to LH tone across junctures, thus as LH+LH+LH+LH+LH. Cheng’s hypothesis of relating the tone sandhi process with the syntactic juncture strength was argued against by Shih (1986), in which Shih claimed that many examples in speech prove that tone sandhi process can be independent of syntactic junctures, as tone sandhi may apply across a strong syntactic juncture, but not a weak one, depending on how the cyclic sandhi process implements. It was demonstrated in Shih’s study (1986) that tone dynamics in speech is more influenced by the prosodic organization, which is sensitive to syntactic information but not necessarily isomorphic to syntactic structures.

b. Tone sandhi and prosodic structure

In spoken language, it is quite normal that there are no clearly defined boundaries between words, phrases or even sentences according to the corresponding syntactic units as in written language. For word recognition, prosodic structure plays a vital role by segmenting continuous speech into metrical units and locating boundaries at hierarchical levels. In the psycholinguistic literature, it is suggested that the utterance’s prosodic structure plays an important organizing role in speech recognition (Rossi 2004). According to Chen (2004), each prosodic constituent, mora, syllable, foot, word, phrase, and utterance are all potential domains for phonological rules.

Tonal manifestation is closely related to the accentual status of the aligned syllable in speech. An unaccented syllable tends to be distributed with short duration and weak strength due to speaker’s careless articulation. When the weak syllable is syntactically and semantically close to the adjacent strong syllable, the tone of the prosodically weak syllable tends to be articulated together with the adjacent strong syllable. Such tone process trigged by the accentual pattern of syllables may also lead to the complete neutralization of tones on the prosodically weak syllables, as seen in Chen (2004: 91). He demonstrated the diverse tone processes attributed by different accentual patterns of two Mandarin phrases:

Phrase a:

(.	.	X)	
xiang	qi	lai	“want to get up”
Tone3	Tone3	Tone2	base tones
Tone2	Tone3	Tone2	tone sandhi

vs.

Phrase b:

(X	.	.)	
xiang	qi	lai	“remember”
Tone3	Tone3	Tone2	base tones
Tone3	º	º	tone sandhi

Example 4.29: Different prosodic structures resulting in different tonal patterns

In Example 4.29, *phrase a* and *phrase b* are both trisyllabic phrases which consist of identical component syllables and lexical tones, but the two phrases are distinguished from each other on their internal accentual patterns. In *phrase a*, the accentual pattern is right-prominent, with the third syllable being the most prominent syllable, while the former two syllables are weak syllables. The accentual pattern of *phrase a* is represented on the metrical grid as (. . X); while the accentual pattern of *phrase b* is left-prominent, represented as (X . .). The different accentual patterns of the two phrases lead to different tone processes. Due to the nature of human articulatory mechanism, tones are produced within a progressive and assimilatory process. In *phrase b*, the tone features of the two weak syllables, *qi* and *lai* were neutralized and assimilated to the tonal feature of the first syllable, *xiang*, resulting in the tone feature of the first syllable extending to the domain of the whole phrase. The left-prominent accentual pattern of *phrase b* triggers the "tonal reduction process" of the second and third weak syllables; Whereas in phrase *a*, the tonal dissimilation process of two adjacent Tone 3 syllables takes place in sequence between the first and second syllables in the phrase, resulting the T3+T3 → T2+ T3. While the prosodically salient third syllable has its tonal feature keeps unchanged.

Due to the tone process occurring in the accentual unit, the underlying tone feature of each syllable is not necessarily aligned with one syllable, instead, it can get spanned to a tonal unit, which may consist of more than one syllables, as seen in the above *phrase b*.

Details on the interaction between tone and accent in Mandarin speech will be discussed in the following section on rhythmic accent, especially in §4.3.3. The operation of Chinese tone sandhi is sensitive to the accentual status of the aligned syllable, with the domain of tone sandhi conditioned by the metrical units in speech.

4.2.5.4 Tone sandhi and speech context

In daily conversation among native speakers, it is well known that there exists an extensive diversity of tonal undershoot. In previous studies such as by Tseng (1981), it was observed that among the spontaneous data only a small portion of syllables in speech have lexical tones which can be predicted relying by the Chinese tonal phonology, while in most cases, there could be no general match between the phonological prediction and the final phonetic output of lexical tones. Despite that, a substantial portion of an utterance might be produced with incomplete acoustic information. She found that listeners often have no difficulty in understanding the utterance meaning. She concluded from her study on spontaneous data that native speakers can make the most extensive use of limited acoustic cues together with their native language background for mutual communication.

According to Patel (2008), speech prosody is functionally driven. This makes it basically distinguished from the musical notes, as the latter require accuracy of pitch while the former is more flexible due to contextual and other factors in actual communication.

In real speech, native speakers do not have to solely rely on the acoustic cues of F0 parameters for tonal recognition. The speech context, either grammatical or ungrammatical, is also a salient factor in causing tonal sandhi, as it provides a large part of anticipated information for the purpose of communication. The effect of speech context on tone sandhi phenomenon can be summarized as the speakers' balance between the ease of production and the ease of perception. To be specific, due to the speech production mechanism, speakers would use the "minimum production efforts" to reach the "minimization of perceptual confusion" (Boersma 1998). In some cases, the canonical form of tones could become redundant in speech exchanges.

It is proposed by Kochanski *et al*. (2003) that people usually plan their speech in advance and minimize their physical effort in production as long as the communicative meanings can be correctly conveyed and perceived. People's rough plan constructed in mind before speaking is proved by evidence that they tend to inhale a larger volume of air before a long utterance (Wilder 1981; Winkworth *et al*. 1995). The preplanning system in muscular forces can be evidenced by Yip's study (2002) on children's early production of utterances: "the added challenge of articulation means that motor control of the appropriate musculature must precede accurate production, and thus mastery of the system" (Yip 2002: 301).

When speech participants share the common cultural and language background, they tend to relax their speaking style. More importantly, the speakers also save their energy in articulating the exchanged information, which, in their consideration already exists in the listener's mind. Therefore, no big effort is needed for capturing the partner's attention on such shared information. For example, frequent words are more likely to present tonal variations than the rare ones. In some cases, the allotones of such common words may derive from routine deletions of tonal targets due to frequency of usage, as the simplified production of tonal targets may become the intentional effort in casual speech style, which is similar to the phenomenon of segmental deviation in casual speech as seen in Zhi *et al*. (2011).

Such sandhi phenomena in context have been defined by Shih (1986) as the **neutralization rule** which conditions the tone sandhi process in the Chinese language family. She claimed that in polysyllabic words some tones get lost, due to the reduction in their functional load of tonal distinction. It is claimed that in many cases the citational tone of the word initial syllable is maintained, while the tones in medial positions are partially neutralized.

The tonal neutralization process belongs to the optional tone sandhi phenomenon in Mandarin speech, especially in the case of unaccented syllables, in which the tonal feature deviates completely from the citational form, or realized as the interpolation between adjacent tonal features. This can be observed in the phonetic data of the following Figure 4.1 of a spontaneous utterance:

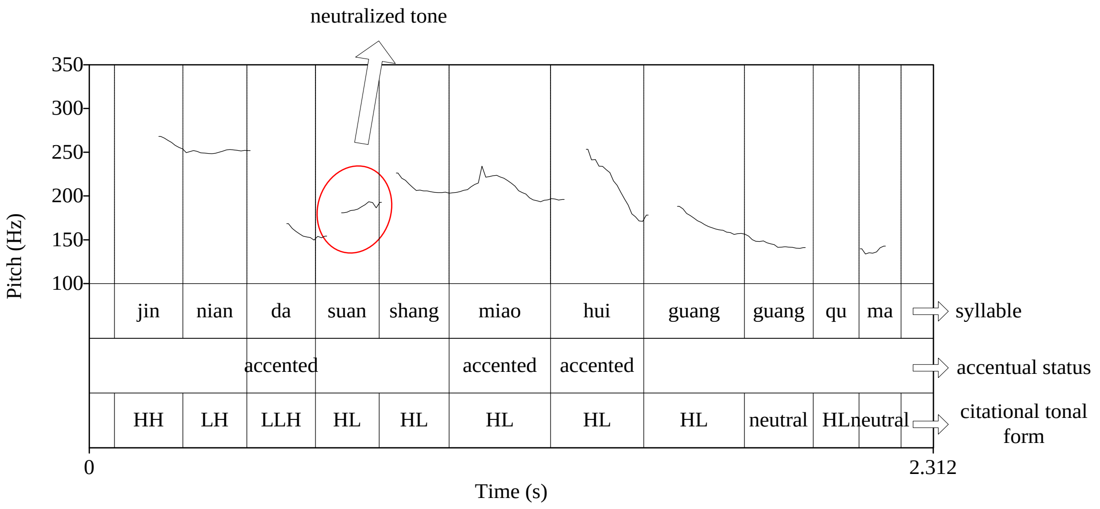


Figure 4.1: The phonetic picture of the sample Utterance (1)

In Figure 4.1, the fourth syllable, *suan*, which has a falling HL tone in the underlying phonological form, is actually realized as a rising pitch movement in the phonetic F0 data. Such pitch feature is a complete reverse form of its citational tone, and it is actually interpolated between the final pitch point of the preceding syllable, *da* and the following initial pitch point of the syllable *shang*. The sandhi form of the syllable *suan* is attributed to the neutralization tone process conditioned by its unaccented status, and to the unsalient functional load as the non-initial position in the lexical word, *da suan* (verb. “plan”).

The reduced functional load leading to tone sandhi phenomena has also been discussed in Xu (2011), who claimed that there exist both prosodic codes and syntactic codes for speech information transmission, and “if a function is already syntactically coded, there is no need to also encode it prosodically, and vice versa”, while of course redundant coding can still be found in speech, this is not universal.

Besides, face-to-face daily communications also involves a large number of non-verbal gestures, such as facial expressions (showing emotions of excitement, happiness, sadness, angry...), body gestures (nodding/shaking head, directing with hands…), etc. Such communicative gestures are shared among people with the same language background, and they contribute significantly in mutual comprehension along with the speech content.

When the above pragmatically governed coding is enough to achieve the purpose of speech communication, speakers are not likely to add further redundant codings for the same purpose. Thus, less articulatory efforts leading to undershooting of tonal articulation is quite understandable in casual speech.

Some tone sandhi phenomenon can be categorized as linguistically irrelevant tones, as according to Jassem’s words (1952: 23),

> There are differences between sounds which are linguistically irrelevant not because they are indetectable by the ear, but because they do not constitute a common property of the speech community. These are individual differences.

Summarizing, in the above discussion on tones, some important issues on Chinese tonology were discussed, including the underlying phonological annotation of tones, tone process in isolated words and fluent speech flow.

It is proposed that the surface tonal pattern in connected speech cannot be analyzed at the syllabic level, but rather at a higher level of tone sandhi domain, where the component syllables form one unitary domain for the lexical word prosody.

This study agrees with Shih's proposal that obligatory tone sandhi process should be distinguished from optional tone processes, as the former refers to the tonal alternations, which can be regulated by some default phonological rules. This study proposes that the dissimilatory sandhi process of Tone 3 as well as the tone spreading process in full tone and neutral tone sequence are obligatory tone process, which can be regulated by phonological tone rules as discussed in section §4.2.5.1.

The optional tone process, in contrast, is difficult to specify in written tone rules, as it can vary according to a number of influential factors in real speech. Tone variations in connected speech can be generally attributed to speaker's balance between ease of speech production and ease of speech perception. It can be triggered under several conditions, i.e., the accentual pattern of relevant syllables in speech prosody, the syntactic structure in influencing the tone sandhi process, and the influence from speech contextual factors.

In the following discussion on rhythmic accent, more details will be presented on the interactive relation between tones and the accentual status in prosody; it is proposed that tone dynamics in speech flow is closely related with the prosodic units, as defined by sentence accents.

4.3 Rhythmic accent

4.3.1 Accentual hierarchy

4.3.1.1 Level I

According to Fox (2000), accent can be distinguished into two levels of manifestation, Leve I and Level II. The lower level of accentuation mainly contributes to the speech rhythmic feature, while the higher level of accentuation contributes to sentential intonation. The domains of the accentual contrast at two levels extend to different units. At level I, accentuation is realized in a basic unit of foot, which consists of the accented syllable and the following unaccented ones. In English, accents at Level I is mainly manifested as the rhythmic beats in speech. The level I accentuation does not coincide with the “word-stress”, as such equation could be misleading due to the fact that words may either have several accents or no accent at all. The accentuation at this level contributes to the rhythmic structure of language, indicated by the assumed regularity of the occurrence of beats.

In a classical study, Pike (1945) distinguished two rhythmic types of languages, namely, syllable-timed and stress-timed rhythmic tendencies based on the isochronous occurrence of syllables and stress-intervals, respectively. The term “dichotomy” is used in his categorization of language rhythmic behavior based on the perceived equal duration of each syllable or each inter-stress interval. However, such predicted categorization was related to two extreme rhythmic behaviors, which could not be supported by the real fact. Instead, researchers found from their experimental data that the perceived regularity of natural languages’ rhythmic features could be closely related to the timing variations of relevant segmental events.

The acoustic measurement of vocalic and consonantal properties have been conducted in previous studies, such as in Ramus *et al*. (1999), where the percentage of vocalic intervals, (*%V*) and the standard deviations of consonantal intervals, (*ΔC*) were measured as two crucial variables in categorizing the rhythmic classes of different languages. In the study of Grabe & Low (2002), a mathematical formula, *Pairwise Variability Index* (raw *PVI*), was proposed for computing the durational variance of successive pairs of intervals, and a normalized PVI was used for the calculation of intervocalic variability by normalizing the speech rate effect on vowels, with the *rPVI* and *nPVI* formulas seen as follows, where m stands for ‘number of intervals’, and d for ‘duration’ (in ms).

$$rPVI = \frac{1}{m-1}\sum_{k=1}^{m-1}\left|d_k - d_{k+1}\right| \qquad \textbf{Formula 4.1}$$

$$nPVI = \frac{100}{m-1}\sum_{k=1}^{m-1}\left|\frac{2\,(d_k - d_{k+1})}{d_k + d_{k+1}}\right| \qquad \textbf{Formula 4.2}$$

In Grabe & Low's study (2002), the rhythmic features of 18 languages were comparatively analyzed according to the values computed by the above two formulas. Languages with low vocalic rPVI variations and low intervocalic nPVI values, such as Mandarin, French and Spanish, were grouped as the "syllable-timed" languages, whereas the languages with high variable correlates indicated by the two indexes were characterized as the "stress-timed" languages, such as English, German and Dutch. The following figure from their study provided the measurement results across 18 languages, where the filled circle, '●' corresponds to the prototypical syllable-timed languages; and the empty circle, '○' corresponds to the stress-timed languages; the filled square '■' corresponds to the mora-timed language, and the empty square, '□' corresponds to the mixed or unclassified languages in the traditional grouping method.

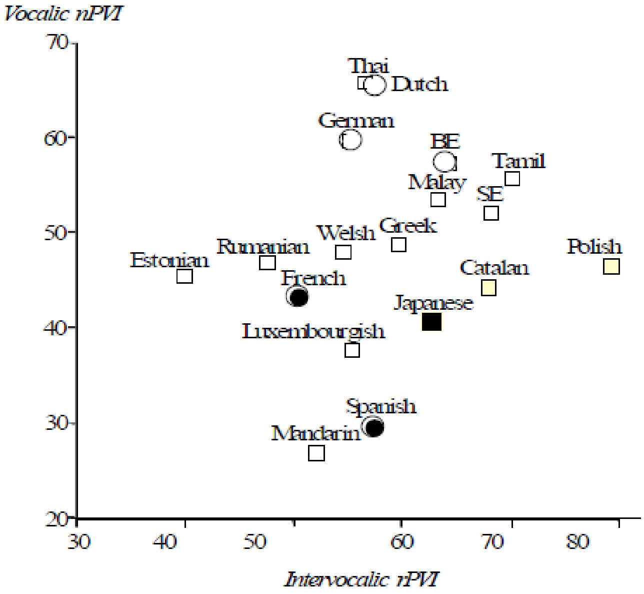


Figure 4.2: Rhymic tendencies of 18 languages based on the rPVI and nPVI indications

In the study of Bertinetto & Bertini (2008, 2010), a modification of PVI formula was proposed for the computation of language rhythmic behaviors, which is named as the "Control-Compensation Index" (CCI) as follows:

$$CCI = \frac{1}{m-1}\sum_{k=1}^{m-1}\left|\frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}}\right|$$ **Formula 4.3**

The revised model takes into account the number of segments (represented by ‘n’ in the formula) in each vocalic interval and consonantal interval. It is believed that the phonotactic complexity of a language significantly contributes to its speech durational behavior. The terms, “compensating” and “controlling” were proposed to use in indicating the gradient tendencies of language rhythm, instead of the strict categorization with “syllable-timed” and “stress-timed” terms. According to the durational fluctuations of vocalic and consonantal intervals respectively indicated by the CCI model, an idealized scheme is presented in the following Figure 4.3:

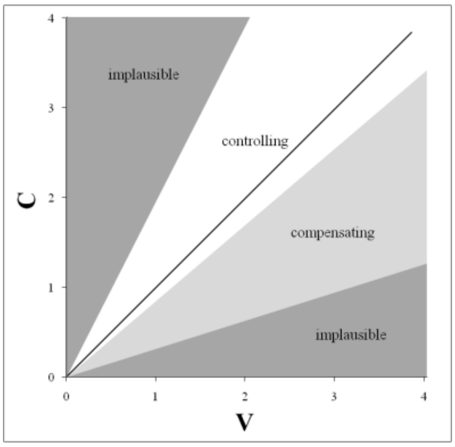


Figure 4.3: A schematic representation of major rhythmic types according to CCI model

A perfectly “controlling” language would present tendentially identical C and V local durational fluctuations, thus falling on the bisecting line, or at least it should exhibit stronger stability in the V intervals. By contrast, “compensating” languages fluctuate more in the V than in the C component, for they presuppose substantial V-reduction.

In Zhi *et al.* (2011), a rhythmic study was conducted by comparing the features between spontaneous Beijing Mandarin (SBC) and Pisa Italian, in both read speech (RPI) and spontaneous speech (SPI) styles. The following Table 4.6 presents the Chinese data and the

Italian data (read and spontaneous) employed in the CCI calculation.

language	number of phonological segments		number of speakers
SBC	vowels	10264	7
	consonants	6976	
SPI	vowels	2812	10
	consonants	3587	
RPI	vowels	1646	7
	consonants	1929	

Table 4.6: The vocalic and consonantal data of Spontaneous Beijing Chinese (SBC), Spontaneous Pisa Italian (SPI) and Read Pisa Italian (RPI)

Based on the CCI computation, the rhythmic tendencies of two languages are presented in the following Figure 4.4:

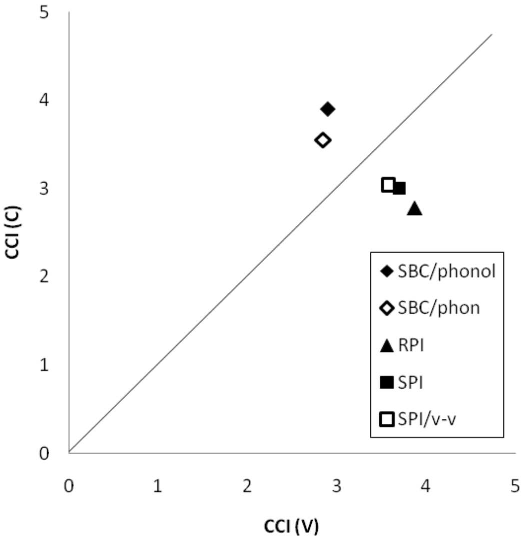


Figure 4.4: Rhythmic tendencies of Spontaneous Beijing Chinese (SBC) in two analyses (phonological and phonetic); Read Pisa Italian (RPI), Spontaneous Pisa Italian (SPI) with prevocalic glides assigned to C-intervals (SPI) or V-intervals

(SPI/bis)

In comparison with Italian, Beijing Chinese presents high stability in the vocalic components, and in the figure SBC falls to the left of the bisecting line, which both indicates a strong rhythmic controlling behavior. Interestingly, this arises despite the presence of so many two- and three-V sequences, which might in principle introduce a great deal of compression among the relevant V segments. Despite this, and despite the virtual absence of C clusters, there is more “local fluctuation” (as measured by CCI, in analogy with PVI) in the C than in the V intervals. This is evidently due to the longer duration of aspirated consonants as opposed to non-aspirated ones.

As can be noted in Figure 4.4, the rhythmic computations of SBC were conducted in two analyses (phonological and phonetic), due to the fact that in spontaneous speech the actual phonetic output does not always correspond to the speaker’s phonological intention. In the CCI model, phonetically inaudible, but phonologically intended segments are assigned zero duration but count as one component in the relevant interval, for one may assume that such segment is actually a part of the speaker’s articulatory plan. However, a double computation with both the phonological segments and the actual phonetic segments were measured and presented in the Figure 4.4. There are actually no big differences.

In Italian, the exist the ongliding diphthongs (or called rising diphthongs) such as *jV* and *wV*, and offgliding diphthongs (or falling diphthongs) such as *Vj* and *Vw*. In Marotta’s study (1988), the two types of diphthongs are found to differ not only structurally but also phonetically, with the evidence that the rising diphthongs as a whole have shorter duration than the falling diphthongs, as the glide in the GV context, is often produced consistently shorter than it is in the VG context. She proposed that for the falling diphthongs, due to the lack of sequences such as *VGC or even *VGC$_1$C$_2$ in Italian, the offglide in such diphthongs should be regarded as in the coda position of the syllable, and it forms together with the preceding vowel as a complex rhyme of the syllable. In the situation of the rising diphthongs, when the onglide is a /j/ phoneme, it is considered as a single part of the syllable onset, in the case of either being freely or together with the preceding consonant; while when the onglide is a /w/, there emerges different treating in two contexts: a) before a vowel [ɔ, ○], the onglide /w/ appeals to be the head of the syllable nucleus, b) after consonant [k, g], /w/ belongs to the onset part of the syllable. The glide status of Italian is also discussed in Loporcaro and Bertinetto (2005). In the present study, the glides in Italian are treated as part of the consonantal onset or coda as appropriate, and computed in the CCI algorithm as belonging to the C interval.

In Mandarin Chinese, there is no final agreement in previous studies (Bao 1990a, 2001, 2003; Van de Weijer & Zhang 2008; Wan 2002; Yip 2003) on the phonological status of the initial segments /i/, /y/ or /u/ in multi-vowel sequences such as /ia/, /iau/, /iou/, /ya/, /ye/, /ua/, /uo/, /uai/, /uei/. In the CCI computations, we regard: the relevant diphthongs and triphthongs with an initial /i/ or /y/ as vocalic sequences, as we find no convincing evidence to consider /i/ or /y/ as a glide. As for vowel sequences beginning with /u/, two alternative phonological entities are considered, [u] and [v]. In Mandarin speakers' productions, a high back vowel /u/, as part of the nucleus of an onsetless syllable, is sometimes produced as a labial-dental [v], especially in sequences such as /uən/, /uan/, /uaŋ/, /uai/ and /uei/. One and the same speaker can interchangeably articulate the two sounds. Therefore, in the CCI computations, the [u] and [v] are considered as belonging to the vocalic or a consonantal interval, respectively.

To strengthen our comparison with the Chinese data, we ran a double computation of the Italian data, by treating glides as part of the C intervals (as in previous studies) or as part of the V intervals as in the Chinese case. Such double counting in data dealing is suggested by Bertinetto & Bertini (2010) as:

> While applying the CCI algorithm one should thus carefully consider the phonological structure of the languages under study, possibly adopting a double counting in critical instances. Glides are a case in point: their treatment as either C or V segments varies from language to language. It is thus advisable to apply the algorithm in both ways, in order to ensure

cross-linguistic comparison.

In the above Figure 4.4, it can be noted that no substantial difference emerged between the SPI (marked as filled square) and SPI/v-v (marked as empty square) analysis. Such statistically irrelevant difference can be expected, as the number of Italian diphthongs in the present data was 145, with the sum amounts to only 5.2% of the V intervals. The small amount of diphthong presence yields a non-significant contrast. Whereas the Chinese data has a large number of two- and three-vowel sequences, which were 2360 and 626, amounting to 45.0% of the V intervals.

In the study of Zhi *et al*. (2011), the factor of speech rate on influencing the rhythmic tendencies were inspected. The productions of the two languages, SBC (spontaneous Beijing Chinese) and SPI (spontaneous Pisa Italian), were divided respectively into three tempo-groups (measured in segments per second). The three groups (slow, medium and fast) of each language data and the number of utterances involved in each group are listed in the following table:

Language	Group	Speech rate (*segm/sec)*	Number of utterances
SBC	1	slow ≦ 16.1 (average: 14.3)	205
	2	medium >16.1, < 18.8 (average: 17.4)	203
	3	Fast ≧ 18.8 (average: 20.5)	199
SPI	1	slow ≦ 14.7 (average: 13.4)	78
	2	medium > 14.7, < 16.8 (average: 15.7)	74
	3	Fast ≧ 16.8 (average: 18.7)	81

Table 4.7: The three tempo-groups: Group 1, 2, 3 of SBC and SPI

The CCI computations on the rhythmic behaviors of the three tempo-groups of the two respective languages are shown in the following Figure 4.5. One can note that at the highest speed, the Italian speakers tend to converge towards the bisecting line, showing an increased controlling behavior, as the faster one speaks, the less room there is for local durational fluctuations. While it is interesting to note that Chinese speakers seem to preserve a relatively larger control over the single segments' articulation even at the highest rate. The result shows that speed accelerations exerted a tendentially linear effect on both V and C intervals of the two languages, as predicted for their controlling rhythmic behaviors.

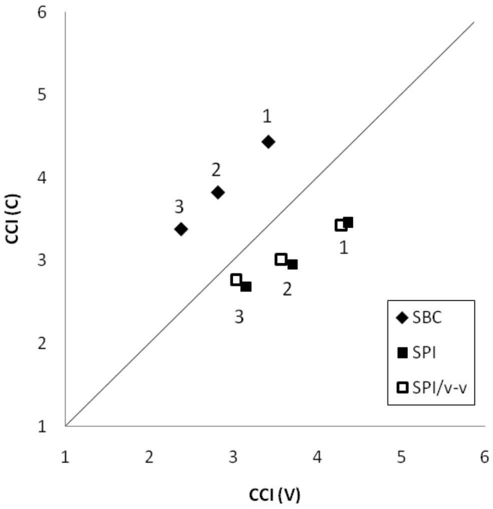


Figure 4.5: Slow (1), medium (2) and fast (3) speech-rate for SBC as compared with SPI and SPI/v-v

The CCI model in its full conception describes the speech rhythm of natural languages at two levels, at the lower phonotactic level, it considers the role of phonotactics in the rhythmic architecture; at the higher sentential level, it analyzes the coupling of two oscillators, the sentence accent and the syllable-peak oscillators. Adopting suggestions by O'Dell & Nieminen (1999), the oscillators' coupling is expressed by the following formula:

$$I(n) = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2} n$$ **Formula 4.4**

Where I stands for 'duration of the inter-accentual intervals', n for 'number of syllable-peaks', ω_1 and ω_2 for the angular frequency – or velocity – of the two oscillators. The rhythmic component at Level-II refers to the accentual prominence at the intonational level.

In the following section, the rhythmic organization at Level-II in Mandarin speech will be analyzed, with the accented syllables at the sentential level derived from the result of a perception experiment.

4.3.1.2 Level II

In Mandarin speech, accent does not fulfill a distinctive function at the lexical word level. The accented syllables perceived in Mandarin speech are mainly for expressing the intonational prominence at the sentential level. A perception experiment has been conducted, with the aim to identify the position of sentence accents in fluent Mandarin speech. The basic experimental procedure was to ask native subjects to listen to Mandarin utterances and point out the salient syllables of each utterance according to the subjects' own perception.

a. Perception experiment

- Stimuli

The experiment stimuli consist of 60 Mandarin sentences, with 30 selected from the spontaneous speech corpus, and the other 30 from the read speech corpus. The length of each utterance is between 7～13 syllables, as longer sentences may increase the difficulty for listeners to identify sentence accents.

15 native Mandarin speakers voluntarily participated in the perception experiment. The subjects were students from mainland China, who were taking courses or doing research in a university in France, and their average age was 27 years old. 13 subjects were born in mainland of China, and 2 in Taiwan, with all of the subjects having Mandarin as their first mother language. The details of the participants can be seen in Appendix I.

The 15 subjects were divided into three groups, Group I, Group II and Group III, and the

60 sentence stimuli were also divided into three lists, List A, List B and List C, with each list containing an equal number of spontaneous and read speech sentences. Each list of stimuli was perceived by two groups of subjects, with the experiment arrangement of the data and subjects in the following Table 4.8:

	Group I	Group II	Group III
List A	▨	▨	
List B	▨		▨
List C		▨	▨

Table 4.8: The experiment arrangement: Group I subjects perceive the sentences of List A and B, Group II perceives List A and C, and Group II perceives List B and C.

- Pre-experiment training

As most of the subjects who participated in the experiment are non linguistic majors, a short pre-experiment training seems necessary to guide each participant in understanding the basic conception of sentence accents, and familiarize with the experiment procedure in locating the most salient syllables.

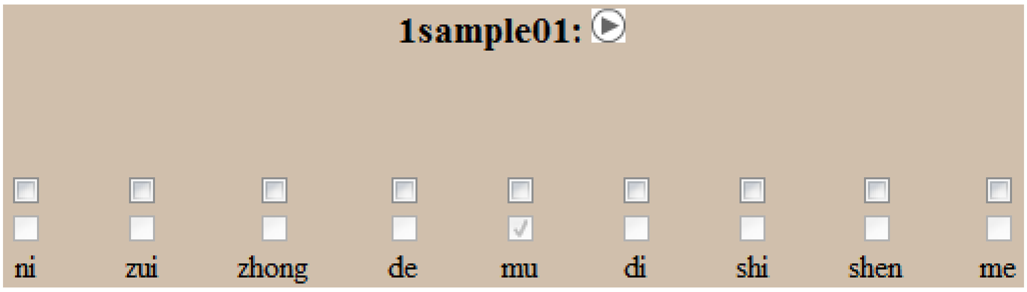
In the pre-experiment training session, each subject was given a sheet with the following instruction, which aimed to help subjects in understanding the procedure and purpose of the perception experiment.

Experiment Instructions

> This study aims at examining the sentence accents in fluent speech. You will have to point out the syllables bearing a sentence accent based on your own perception. You are going to listen to one utterance at a time. The sentences are 40. You can either do the whole task at one time or divide it into two sections. After the first 20 sentences you will be offered the chance

to rest. A short training for locating the accents is necessary. Two sample utterances will be used to this purpose.

UTTERANCE 1 [with strong emphatic stresses]

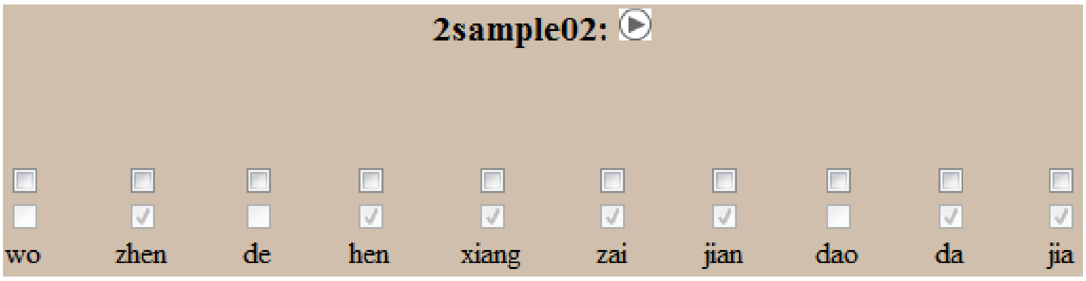


1sample01:

□	□	□	□	□	□	□	□	□
□	□	□	□	☑	□	□	□	□
ni	zui	zhong	de	mu	di	shi	shen	me

Click on the arrow to listen to the recorded version. In this sentence, the highlighted syllable is “mu”, which stands out most from the others. If you agree on this, then click on the corresponding box in the top tier for confirmation. You can now move on to the next sentence by clicking on the arrow at the bottom right corner of the screen.

UTTERANCE 2 [with neutral intonation]



2sample02:

□	□	□	□	□	□	□	□	□	□
□	☑	□	☑	☑	☑	☑	□	☑	☑
wo	zhen	de	hen	xiang	zai	jian	dao	da	jia

Most of the sentences in this experiment look like this, with no strong emotions or emphases. You will however be provided with some hints, pointing out the potentially accented syllables, as shown on the bottom tier. You are expected to identify the most salient syllables according to your

> perception, by selecting the most prominent syllables among the ones highlighted. You are however free to select any syllable, even among the ones that are not highlighted. Try and click on any box of the top tier to verify that you can activate your own selection. For any given sentence of the experiment, you will first have to click on the arrow to listen to it. On the screen you will see a sequence of boxes on two tiers. The bottom tier contains a few highlighted syllables, while the top tier is empty. By clicking on any empty box of the top tier you will make your own choice. You can listen to each sentence as many times as you like, and can modify your decisions until you are happy with them. Click on the bottom right arrow to move on to the next sentence.

- Experiment and result

Each group of subjects was required to perceive two lists of stimuli, with each list taking about 30 minutes for each subject to make the whole judgment. Therefore, to avoid the subjects being exhausted during the experiment, and to ensure the performance efficiency in the perception task, each subject was offered the chance to postpone the continuation of the experiment after the first list, and perceive the second list in another time. As a result, 2/3 of subjects chose to finalize the experiment in two sessions in different days within one week, while 1/3 of subjects completed the task at one time.

In the study of Bertinetto *et al*. (2012), the same perception experiment was also conducted among 15 native Italian speakers, who were asked to listen to the equal number of Italian utterances which were selected under the same criteria of the Mandarin ones. The two perception experiments were conducted for a comparative study on the role of sentence accents in the two languages.

In the following table from Bertinetto *et al*. (2012), the subjects’ perceptual judgment on the stimuli utterances were presented according to four criteria, 60%, 70%, 80% and 90%, with the percentage indicating the degree of agreement reached among the subjects. It can be seen that with the tightest criteria (90%), there is much less inter-subject convergence among Chinese subjects, as compared with Italian subjects.

	Spontaneous				Read			
	60%	70%	80%	90%	60%	70%	80%	90%
BC	29.2	21.6	13.0	5.3	37.6	25.5	13.8	3.8
PI	26.1	23.1	19.5	14.0	26.2	25.0	21.9	17.1

Table 4.9: Percentage of highlighted syllables in BC (Beijing Chinese) and PI (Pisa Italian)

To have a clear view of the results derived from the comparison between the two languages, I drew the following two figures on the basis of Table 4.9. The left figure indicates the comparison of native subjects' judgment on the spontaneous data between BC and PI, and the right figure indicates the comparison of the read-speech data between BC and PI.

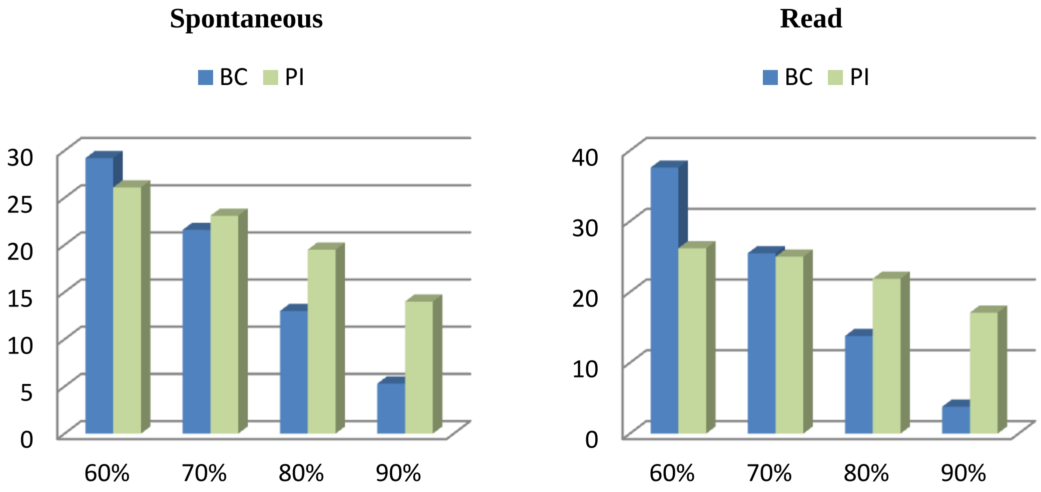


Figure 4.6: A comparison of salient syllables in BC and PI (in percentage) pointed out by native subjects under different criteria, 60%, 70%, 80%, and 90%

In Figure 4.6, it is seen that in both languages, alongside with the tightening of the criteria, from 60% to 90%, the percentages of salient accented syllables decrease. However, such declination is more dramatic in Mandarin speech (both spontaneous and read data) than

Italian speech.

Such phenomenon led Bertinetto *et al*. (2012) to conclude that sentence accent is part of prosodic competence of an Italian speaker, but a fairly elusive feature among Mandarin speakers.

b. The elusive identification of sentence accents

For Mandarin subjects, it is not easy to identify the sentence accents purely dependent on auditory perceptions. Two reasons may explain such accent deafness of Mandarin speakers.

Firstly, the distinctness of accents at the perceptive level is related to its functional load in the language. According to Jassem (1952: 20),

> Speakers usually fail to recognize differences between sounds which are functionally non-distinctive in their own language.

Accent is frequently employed in a stress language like Italian as it functions significantly at the lexical level for word identity, and a sentential peak is formed with the accent at the lexical level superimposed at the sentential level. Therefore, accent is part of the prosodic habit of Italian speakers, and it is no wonder that accent can be easily identified based on native subjects' auditory perceptions. However, the situation is different in a non-stress language like Mandarin, where accent is employed in quite seldom cases at the lexical level, but only at a higher sentential level, where together with other phonetic correlates accent contributes to the intonational prominence. Therefore, due to its less frequent employment in Mandarin, it is not surprising that accents are minimally salient in Mandarin subjects' auditory perception.

Secondly, according to Fox's study (2000), accent has ambiguity in both its phonetic definition and perceptual identification, as it is a "place-holder" to cover a number of phonetic features. In Mandarin speech, accent has close correlation with another suprasegmental feature, pitch; in particular, the lexical pitch contrast on each syllable in Mandarin is closely related with the accents in speech. Fox proposed that the accentual study shall not be constrained with the elusive phonetic feature of accentual phenomena, but rather on its functional role in defining the speech structure.

I agree with Beckman's (1986) highlighting on the **organizational** role of accents in speech. At the sentential level, accent contributes to chunking speech into hierarchical prosodic units. Such prosodic domains are conceived as the unit where phonological rules apply. As discussed in previous studies, such as Chen (2004), tone sandhi often takes place in certain prosodic domain, which is related to metrical prominence. Tones in metrically prominent position often remain unchanged, and present the feature of "tonal stability", while tones associated with weak status in metrical structure tend to assimilate to the tones on the prominent syllables, leading to tone sandhi, such as modification, neutralization, or even the complete loss of tone features. Tone sandhi is an accent-sensitive phenomenon with its domain related to the metrical structure of speech. In the following section of the present study, the interaction between tone sandhi and metrical unit will be discussed with the perceived sample utterances in the experiment. The locations of sentence accents pointed out by native subjects are analyzed in the organizational role of chunking various units in speech.

4.3.2 The prosodic structure of Mandarin speech

The theory of prosodic hierarchy has been widely applied in discussing the hierarchical phonological structures of different languages, such as in works of Selkirk (1978, 1980, 1981, 1984, 1986), Nespor & Vogel (1982, 1983, 1986), Hirst (1977, 1988, 2005), Beckman & Pierrehumbert (1986, 1988), etc. According to the basic principle of prosodic hierarchy, speech can be segmented into chunks at hierarchical levels, as seen in the following Figure 4.7:

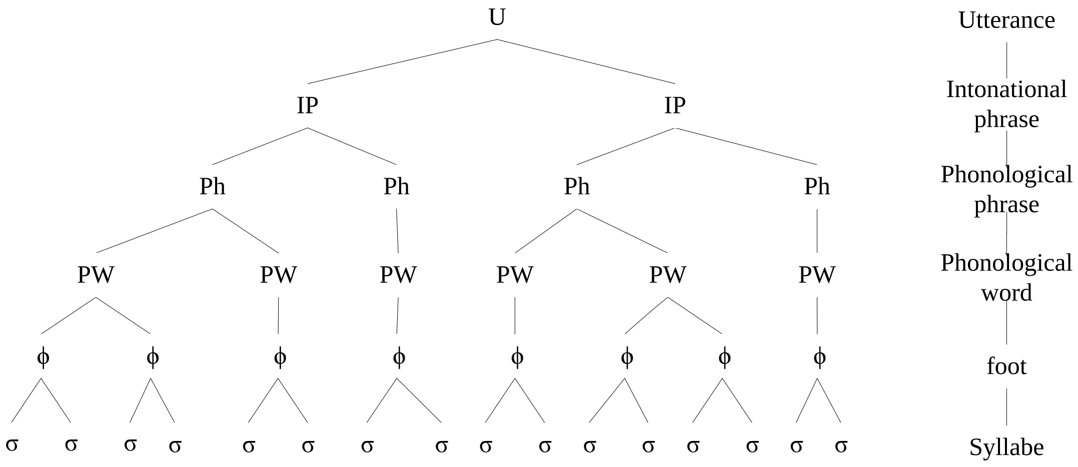


Figure 4.7: The hierarchical prosodic units of an utterance

The above phrasings can be highly different across languages. The terms used in defining prosodic domains are also alternatively used by different linguists, e.g., intonational phrase (IP) is also called *intonation unit* in Hirst (1977); phonological phrase (Ph) is distinguished into *major phrase* and *minor phrase* in Selkirk & Tateishi (1988), or *accentual phrase* and *intermediate phrase* in Beckman & Pierrehumbert (1986); phonological word (PW) is named *prosodic word* in Selkirk (1980). Other terms such as *clitic group* (Nespor & Vogel 1986), or *minimal rhythmic unit* (Chen 2004) are also employed in prosodic phrasing for different research purpose.

In this study, Mandarin speech is partitioned into five levels of units. From top to bottom they are **utterance (U)**, **intonation unit (IU)**, **accentual phrase (Σ)**, **tonal unit (TU)** and **syllable (σ)**.

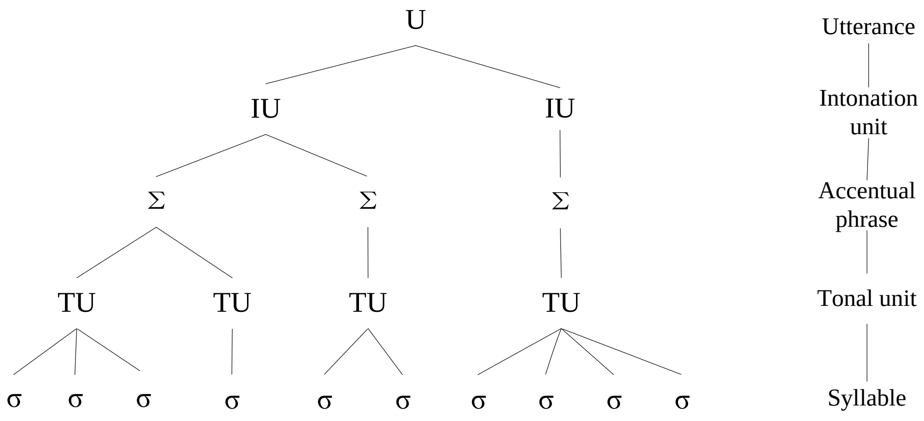


Figure 4.8: The prosodic structure of a Mandarin utterance

- Utterance (U) is the top hierarchical level of the prosody. The utterance boundary is marked by a long pause, and often accompanied with downstep at pitch level as well as the lengthening of utterance-final syllables. A Chinese utterance could correspond to a full syntactic sentence with the "subject + predicate" structure or more commonly, with a "topic + comment" structure as proposed by Li & Thompson (1976). It is also possible that in actual speech, a prosodic utterance only consists of one isolated word or phrase with no corresponding syntactic structure at all.

- Intonation unit (IU) is the subset category of a prosodic utterance. It is defined as a unit with a continuous and smooth pitch contour, and often set off by possible phonetic cues, such as a short pause, final lengthening and laryngealized voicing. However, the pitch reset between two IUs is often a more reliable phonetic cue in bounding an IU juncture. One utterance may contain one or more IUs, depending on speech rate. According to Ladefoged & Johnson (2011), in slow and formal speech, speakers may break up one utterance into several intonation units, while in fast and casual speech, one utterance may correspond to only one intonation unit. An IU may not necessarily correspond to a syntactic unit.

- Accentual phrase (Σ) is the domain defined by sentence accents. It corresponds to the *stress group* or *phonological phrase* called by other linguists. An accentual phrase is basically formed by the syllables bearing phrasal accents and the following unaccented ones. The domain of an accentual phrase can vary from an accented monosyllable to a phrase stretching across word boundaries.

- Tonal unit (TU) is the unit often employed in intonation studies, such as in Beach (1938), Jassem (1952) and Hirst (1988). TU was defined in Hirst (ibid: 157) as the unit in which "smallest linguistically pitch contours occur" in the intonation of English and French. In his later prosodic studies with Prozed for speech synthesis (Hirst 2005, 2008), he proposed that TU is not strictly defined to any corresponding phonological entities, but only corresponds to the domain where annotation can be conducted on the short-term variation in speech melody.

In this study, the formation of a tonal unit is a subcategory of an accentual phrase. It is distinguished from the accentual phrase, as it is conditioned by the principle that a tonal unit should keep the integrity of the lexical items involved in the domain, that is, only the adjacent constituents which are close in lexical relations and semantic meanings can be grouped into one tonal unit; on the contrast, if two adjacent constituents are irrelevant in lexical/semantic grounds, they are assigned into two TUs. In this way, TU insures the lexical and semantic cohesion among the syllables of the unit.

The formation principle of TU is similar to the *sense unit*, as defined by Selkirk (1984) and later employed in Hung (1987). In their studies, the sense unit is formed under the condition that,

> Two constituents Ci, Cj form a sense unit if (a) or (b) is true of the semantic interpretation of the sentence:
>
> (a) Ci modifies Cj (a head)
>
> (b) Ci is an argument of Cj (a head)
>
> from Selkirk (1984: 291)

They claimed that such meaning-based approach is essential in defining the prosodic organization in speech. Hung (1987) further proved that the sense unit is more appropriate in constructing a domain for implementing tone sandhi rules than the prosodic foot, which is formed according to the syntax-related principle of IC (immediate constituency) and DM (duple meter) proposed by Shih (1986: 110) in studying the tone sandhi domain.

In the present study, the idea of *sense unit condition* is adopted as an important reference in the formation of TU. As TU is conditioned by both meaning and prosodic structure, it corresponds to the cognitive unit where speakers make the phonetic plan in connected speech. Therefore, it is believed that a tonal unit is the minimal domain where tone sandhi process operates.

- Syllable (σ) is the lowest hierarchical category in speech prosody. It may either correspond to a free monosyllabic lexical word, or a constituent syllable of a polysyllabic word or compound.

In the following, Chinese prosodic structure is analyzed with the stimuli utterances which have been perceived by the native subjects, and the relative correlation will be explored between the utterance prosodic hierarchy and the tone process involved in the metrical unit.

4.3.3 Prosodic grouping and tone-sandhi domain

In the following section, the relationship between prosodic units and tone dynamics in speech flow are analyzed in details with three sample utterances from our data.

Utterance (1): "ta de hai zi men zao jiu deng zhe chi le"

ta	de	hai	zi	men	zao	jiu	deng	zhe	chi	le
She	possessive	child	diminutive	plural	early	already	wait	imperfective	eat	particle

Her children have been long waiting for eating (the cake).

The above Utterance (1) is from the read-speech corpus, and it was produced by a female speaker with neutral emotion. The accentual hierarchy of the utterance (1) is represented on a metrical grid with three levels: syllable level, word level and intonation unit level.

$X_{(90\%)}$		$X_{(80\%)}$			$X_{(80\%)}$		$X_{(80\%)}$		$X_{(90\%)}$		→ IU level
X		X			X	X	X		X		→ word level
X	X	X	X	X	X	X	X	X	X	X	→ syllable level
ta	de	hai	zi	men	zao	jiu	deng	zhe	chi	le	relative prominence

Figure 4.9: Metrical grid of utterance (1)

At the syllable level, all the audible beats within an utterance are indicated by ‘X’ in the grid representation, as it is believed that “every syllable participates in the rhythmic organization of the utterances” (Selkirk 1984).

At the word level, each Chinese content word, no matter with monosyllabic or multisyllabic structure, is supposed to have accent on each constituent syllable. In contrast to most content words, the grammatical words, which always have neutral tones, are normally unaccented syllables. Therefore, in Mandarin except for a few exceptions, all the content words can be marked with accented syllables at the word level in the grid representation. In utterance (1), all content words are marked as default strong syllables ‘X’ at the word level, while the grammatical words, the possessive marker *de*, the suffix-like nominal diminutive *zi*, the plural marker *men*, the imperfective marker *zhe*, and the utterance final particle, *le* (expressing current state), are represented in the default unaccented syllables.

At the intonation-unit level, the syllables which bear sentence accents are marked on the grid. Such accented syllables contribute to the intonational prominence of the utterance. Although the utterance syntactic structure may provide potential locations of sentential prominent syllables, speech intonation can unpredictably assign prominence on any syllables for various discourse purposes. In this study, the marked accented syllables at the IU level are those with the highest accordance according to the judgment by two groups of native subjects in the perception experiment in section §4.3.1.2. In this utterance, four syllables were perceived as bearing sentence accents with the subject’s agreement reached more than 70% at the IU level, namely, the syllables *ta* (90% of agreement), *hai* (80%), *zao* (80%), *deng* (80%), *chi* (90%).

According to the above discussion on sentence prosodic structure defined by sentential accents, the hierarchical level of the utterance can be represented in the following tree:

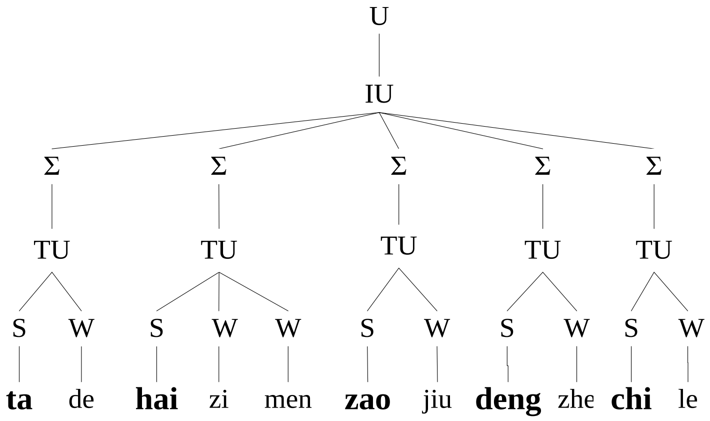


Figure 4.10: Prosodic hierarchy of utterance (1)

As seen in Figure 4.10, the whole utterance corresponds to one IU. The five accentual phrase units (Σ) in the utterance are formed by the accented syllables, *ta*, *hai*, *zao*, *deng*, and *chi* and their representative following unaccented ones. A tonal unit (TU), as a subcategory of an Σ, is defined according to the lexical and semantic cohesion among the syllables. If the syllables are close in lexical-semantic meanings, they are categorized into one TU; otherwise, they are defined into different TUs. The formation of a TU is conditioned by both the lexical-syntactic relations of the component syllables, and the metrical structure of the unit. In this utterance, all the Σs equal the TUs, as the component syllables of each Σ are close in lexical-syntactic meanings, and meet the formation principle of a TU. The binary branching nodes, represented as "S" and "W" in the metrical tree corresponds to the strong and weak syllables in the utterance, and indicates the relative prominence among the syllables.

I now move on to the relationship between accents and tones. Firstly, the citational tonal form of each syllable is represented in the following Figure 4.11, where the default neutral-tone syllables (the grammatical particles, *de*, *men*, *zhe*, *le* and the nominal diminutive, *zi*) are not marked with tones. Thus, the citational tonal pattern of the utterance syllables can be represented as follows:

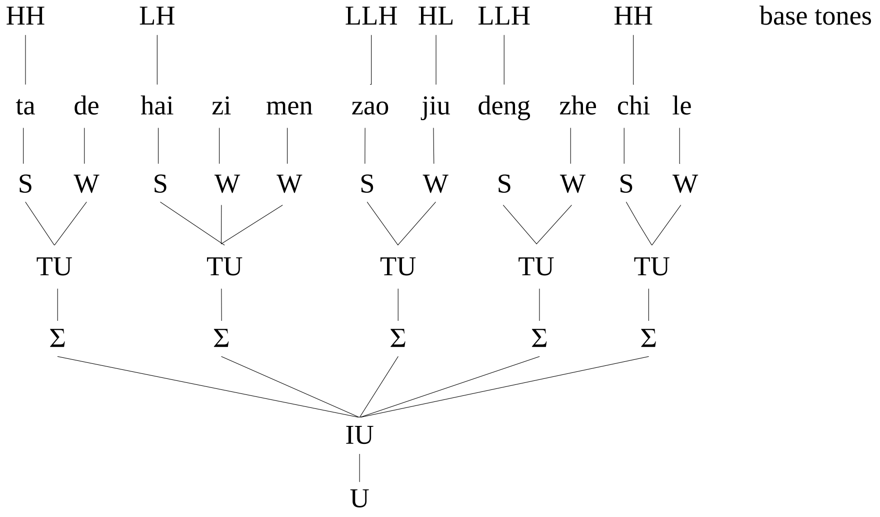


Figure 4.11: The citational tonal form of each component syllable in the utterance (1)

Figure 4.11 shows us the underlying phonological forms of component tones in an utterance.

The above tonal pattern reveals the citational forms of tones in isolated articulations. In tonal context, there shall occur various tone sandhi processes as discussed in section §4.2.5. In the case of a full tone followed by a neutral tone, the *tone spreading process* takes place; the earlier tone extends its tonal feature rightwards into a larger domain formed by the accented syllable and the unaccented one. Such tone spreading process is a progressive assimilatory process within a tonal unit. The implementation of tone spreading processes of the utterance can be represented as follows:

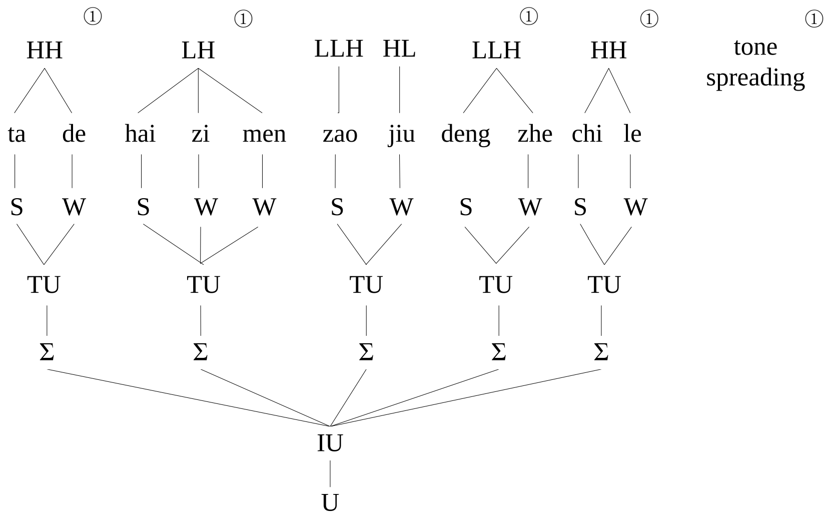


Figure 4.12: Tone spreading process within TUs

Tonal realization is closely related to its accentual status in prosody. An unaccented syllable tends to be related with short duration and weak strength due to speaker's careless articulation. Within a tonal unit, the tone on the prosodically weak syllable tends to be articulated together with the tonal feature of the adjacent strong syllable. Such tone process trigged by the accentual pattern of syllables may also lead to complete loss of tones on the prosodically weak syllables as discussed in section §4.2.5.3, see *tonal reduction process* (Chen 2004: 91).

In the sample utterance (1), the Tone 3 of the accented syllable *zao* and the Tone 4 of the unaccented syllable *jiu* can be produced together as one tonal pattern within the TU, as seen in the following Figure 4.13:

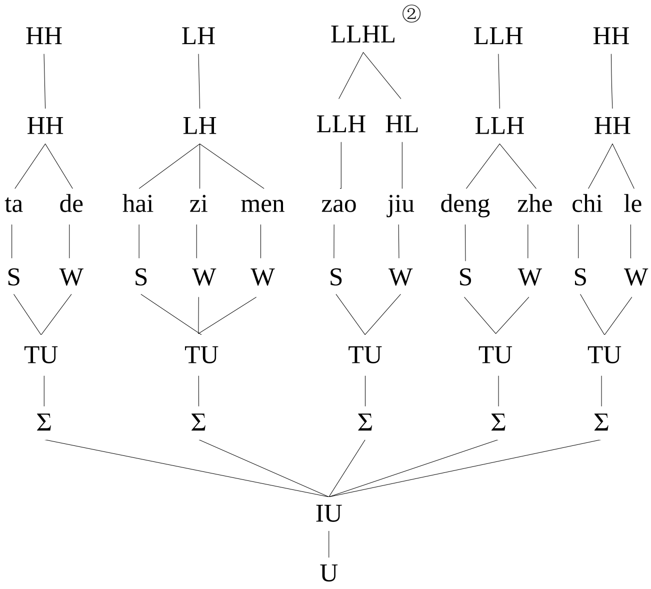


Figure 4.13: Tonal coarticulation of accented syllable and unaccented ones

With the above sample utterance (1), it is shown that the basic tones undergo the tone spreading and tone 'merging' process within the tonal units. Due to different tone processes, the citational tone feature of each syllable often gets spanned to a tonal unit, which can be larger than a syllable in connected speech. Therefore, it is normal that no correspondence can be found between the underlying tonal features and the actual phonetic tones within the unit of a syllable.

The overall pitch pattern of the utterance contributed by the constituent lexical tones can be represented by the tonal movement in the TUs. In the utterance (1), the overall pitch pattern is derived from the tonal processes within five tonal units, as seen in the following Figure 4.14:

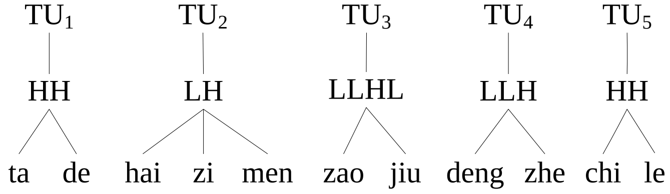


Figure 4.14: The tonal pattern of five tonal units in the utterance

Such contour pattern is the underlying phonological representation of the utterance prosodic pattern. It is directly derived from the representation of lexical prosodic functions, and can be employed as an intermediate phase for the mapping between prosodic function and surface level of prosodic form, which will be further discussed in the following section on the overall prosody of an utterance.

It is proposed in the present section that the relation between tonal manifestation and accent should be attributed to the utterance hierarchical units and the involved tone processes among units.

Here follows another sample utterance (2) to demonstrate the relationship between tone and accent.

Utterance (2): "fang jia le gei song dao lao-lao jia qu le".

fang	jia	le	gei	song	dao	lao-lao	jia	qu	le
set	holiday	perfective	give	send	resultative	grandma	home	away	perfective

Holiday is on, (my daughter) has been sent to her grandma's.

The above utterance is taken from the spontaneous speech corpus, and has a "topic-comment" structure, which is distinguished from the "subject-predicate" structure, with details seen in Li & Thompson (1976).

The metrical grid of the utterance is represented in the following Figure 4.15. At the word level, all the syllables of content words are marked with 'X' in the grid, while the perfective marker *le*, the resultative marker *dao*, and the second syllable of addressing term, lao-lao ("grandma") are all neutral-tone weak syllables. At the IU level, the syllables *jia*, *song*, *lao*, and *jia* were pointed out by native subjects as the most salient syllables, with the percentage of agreement on the accented syllables demonstrated in the grid.

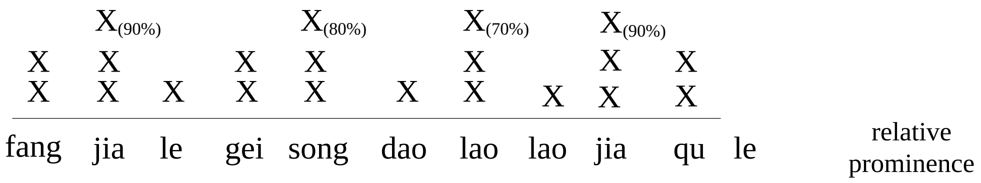


Figure 4.15: Metrical grid of Utterance (2)

Based on the above metrical grid, the hierarchical level of prosodic units as segmented by sentence accents, can be represented in the following prosodic tree:

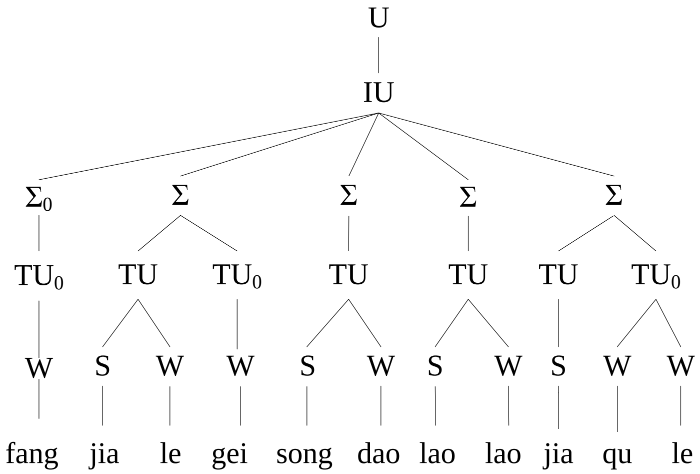


Figure 4.16: The prosodic structure of Utterance (2)

In Figure 4.16, it can be seen that the utterance (2) has only one IU, which corresponds to the whole utterance. The sentence accent syllables, *jia*, *song*, *lao*, *jia* define the accentual phrase units with their respective following unaccented syllables. Within each Σ, TU are formed based on the semantic-syntactic closeness of the component syllables. In Utterance (2), two Σs are respectively subcategorized into more than one TU.

The citational tonal forms of the component syllables in the utterance can be seen in the following Figure 4.17:

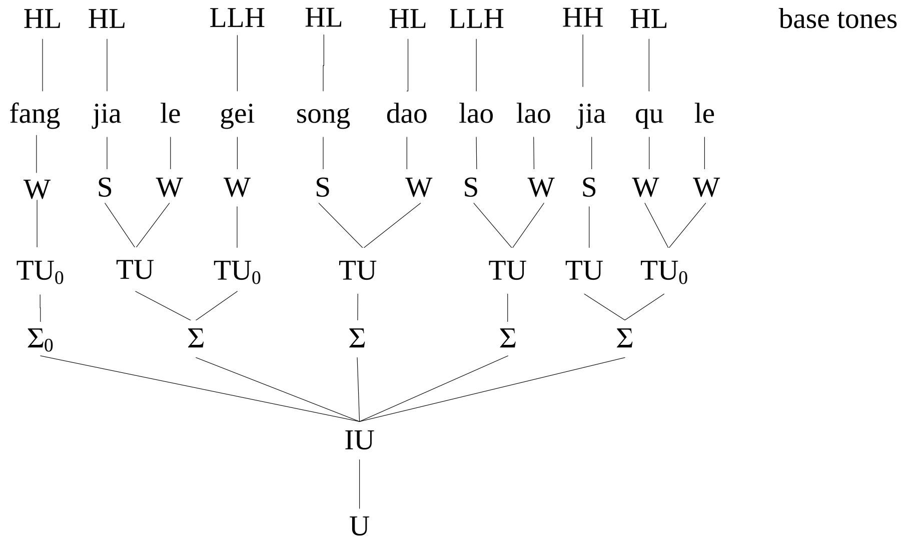

HL HL LLH HL HL LLH HH HL
base tones
fang jia le gei song dao lao lao jia qu le
W S W W S W S W S W W
TU0 TU TU0 TU TU TU TU0
Σ0 Σ Σ Σ Σ
IU
U

Figure 4.17: The citational tones of the component syllables in Utterance (2)

The above base tones undergoes the tone spreading process within the tonal unit formed by accented tonal syllables and the unaccented neutral tones as follows:

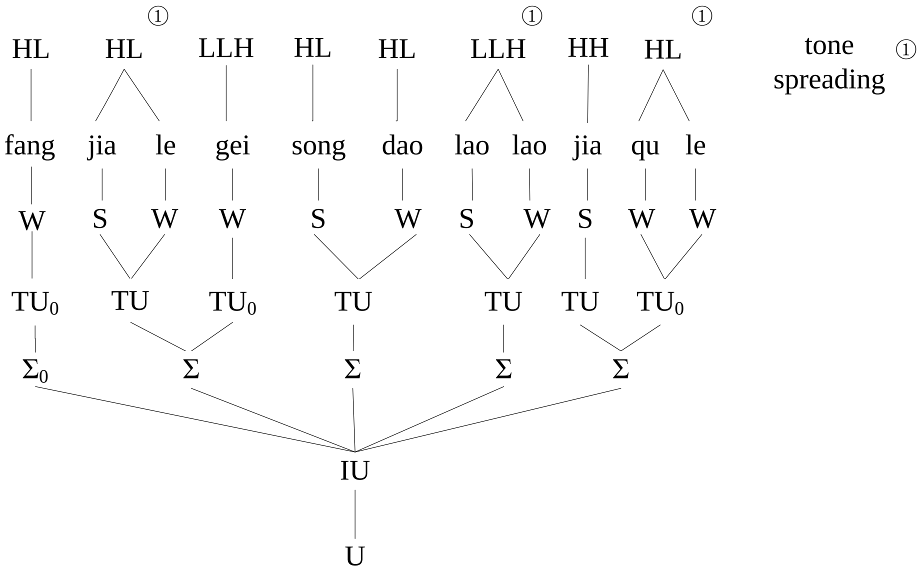


Figure 4.18: The tone-spreading process of Utterance (2)

In the earlier discussion with sample utterance (1), it was proposed that within the same tonal unit, the tonal features of the accented syllable and the tones of the following unaccented syllables can be produced together as one tonal pattern within the unit. In sample utterance (2), such proposal is evidenced from the acoustic signal in the following Figure 4.19:

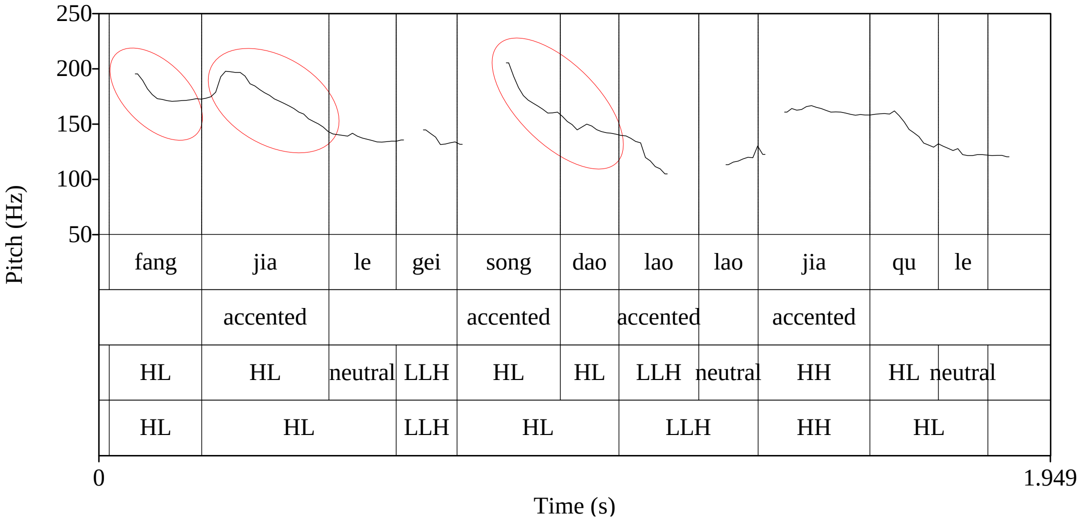


Figure 4.19: The acoustic signal and the tonal annotation of Utterance (2)

As can be seen from the acoustic signal of utterance (2), the two adjacent syllables *song* and *dao*, which both carry the identical base tones HL, merge their tonal features as a single falling movement, namely, HL+HL→HL which extends over the entire tonal unit (TU), composed by the accented syllable *song* and the unaccented syllable *dao*. Such tone process also verifies the sandhi rule of Tone 4 proposed by Chao (1968), reviewed in section 4.2.5.1, such that the pitch movement of Tone 4 does not fall to bottom level when situated before another Tone 4, but rather falls to a mid pitch level. In this utterance, it is seen that the second Tone 4 starts from the final pitch point of the earlier tone and continues the falling movement till the TU boundary. The two syllables are grouped within one tonal unit as they form a metrical foot at the prosodic level, and at the same time exhibit lexical integrity as a verb compound (verb + resultative verb). The two adjacent Tone 4 within one TU implement the tonal reduction process, reducing to one HL tone. The speaker does not raise the pitch for the starting point of the second Tone 4 due to its unaccented prosodic status, and less important lexical content.

Interestingly enough, within this same utterance, we can also find the contrastive situation. This occurs when the two Tone 4 are situated across TU boundary and have the accentual pattern of unaccented + accented, such as the syllable *fang* and the following syllable *jia* at the initial part of the utterance. It can be seen that such adjacent HL tones do not form a single falling movement; instead, the second accented Tone 4 syllable starts again from a high pitch point for the new HL tonal implementation. Due to the accented prosodic status, speakers raise the pitch level in order to mark out the salient position of the following syllable *jia*.

The two different realizations of Tone 4 in the above utterance are an optional tone process, dependent on the accentual status of the syllable. Such tone process at the phonetic level is for the ease of articulation, thus mostly as a tonal assimilation process, which takes place spontaneously in connected speech.

The tone process of two identical tones converging into one unique movement within one TU can also be found in the case of two adjacent rising tones, that is, LH+LH→LH, which is illustrated in the following utterance (3).

Utterance (3): “na xie pian zi xian zai hai neng kan”

na	xie	pian	zi	xian-zai	hai	neng	kan
demonstrative	classifier	video	diminutive	now	still	can	watch

Those videos can still be watched nowadays.

The Utterance (3) is from the spontaneous speech corpus, and the metrical grid of the utterance can be seen in the following:

X$_{(70\%)}$		X$_{(80\%)}$		X$_{(90\%)}$				X$_{(90\%)}$	
X	X	X		X	X	X	X	X	
X	X	X	X	X	X	X	X	X	relative prominence
nei	xie	pian	zi	xian	zai	hai	neng	kan	

Figure 4.20: Metircal grid of Utterance (3)

On the above metrical grid, at the word level only the nominal diminutive *zi* is a default weak syllable, while the other syllables of the content words are all strong syllables, marked with X. At the IU level, four syllables are indicated as sentential accented syllables according to the native speakers' perceptions.

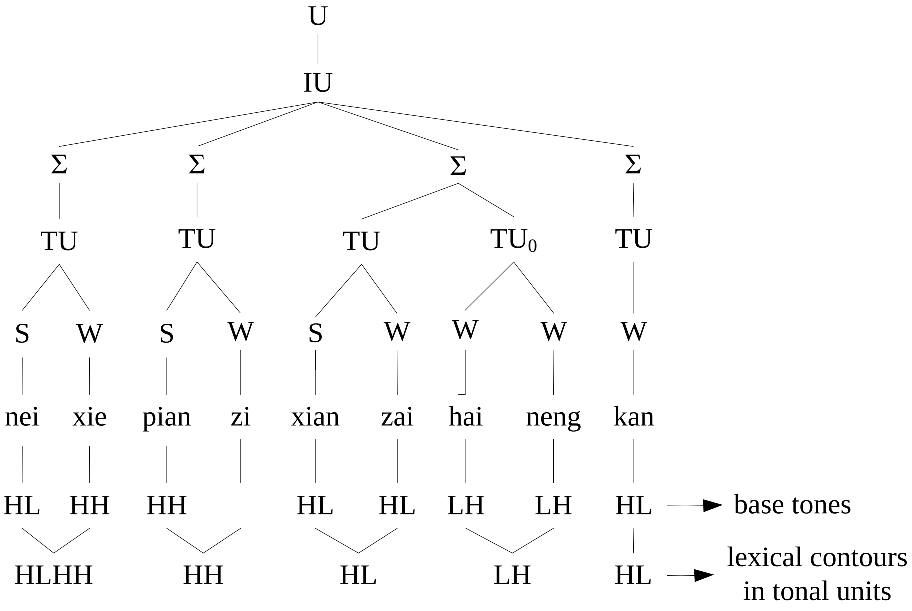


Figure 4.21: The prosodic structure of Utterance (3)

The utterance tonal pattern is represented by the tonal contours realized in the composing tonal units of the utterance, each of which groups syllables according to their prosodic status under the condition of lexical integrity. The following shows the underlying phonological representation of the utterance prosody and the acoustic signal:

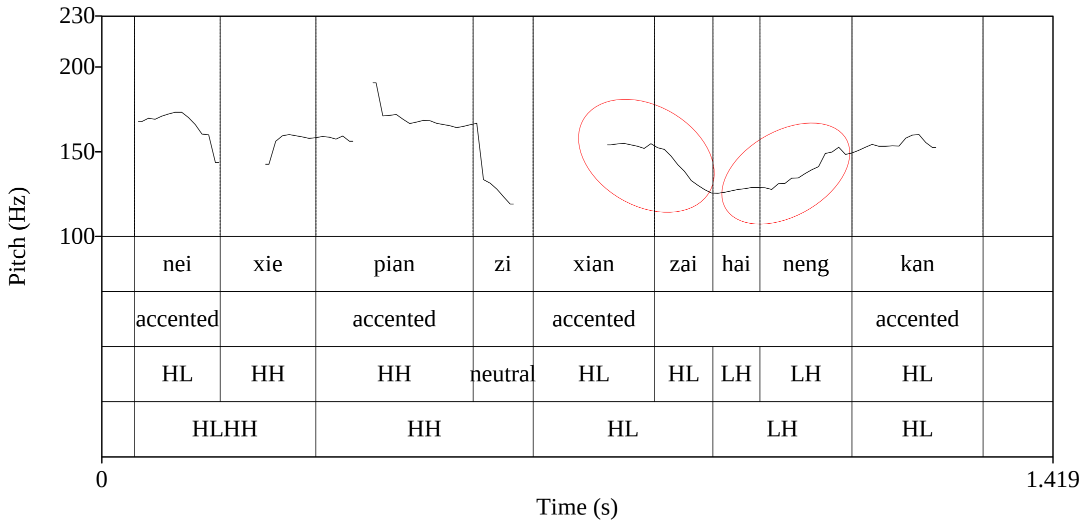


Figure 4.22: The acoustic signal and tonal annotation of Utterance (3)

In Figure 4.22, the two adjacent syllables, *hai* and *neng*, which carry the same rising LH tones, merge their tonal features as a single rising movement LH within the same tonal unit, thus, LH+LH→LH. The tonal process of two adjacent HL tones realized as one HL pattern is also observed within the tonal unit of *xian-zai*, where the two syllables realize the falling HL pitch movement through the entire TU domain.

It can be observed from the sample utterances (2) and (3) that such tone process occurs among the syllables of one tonal unit, with the accentual pattern of syllables as an important condition.

The above discussion reveals that in connected speech, tonal realization is influenced by the accentual status of the syllable. According to Xu (2005), the differences in F0 contour between strong and weak syllables mainly result from differences in pitch targets and target-approximation speed. The higher speed in articulation movement enables a pitch target to be reached more fully within the allocated time interval, and the longer duration provides more available time for approaching the target. The accented and non-accented syllables are also differentiated in terms of pitch target values, as the latter only has a mid-level static target, which is “half way between the maximum and minimum F0 value of full tones or stressed syllables. Therefore, it is expected that a syllable under accentuation is marked out with more complete tonal features than the unaccented ones.

Most sandhi phenomena found in Mandarin speech are optional tonal changes, which are susceptible to the influence of the speech accentual patterns, and the interaction between utterance syntactic structure and prosodic organization. Such optional tone process is conditioned by the processing of tonal units.

In the above section, the role of rhythmic accent in the overall speech prosody was discussed. The organizational function of rhythmic accents in segmenting speech in different chunks is the focus of the section. With the detailed analysis of three sample utterances, it is demonstrated the close correlation between the tonal manifestation and the hierarchical prosodic units, defined by sentence accents. In the following section, the function of intonation and the representation of prosodic form will be discussed for further exploration.

4.4 Intonation

4.4.1 The melodic events of Mandarin speech prosody

In Mandarin speech, there exist the binary uses of pitch by syllabic tones at the lexical level and intonation at the sentential level. These two features are "phonetically intertwined in the tempo and pitch contour of an utterance" (Beckman 1986: 28), which is quite unlike the F0 pattern of a non-tonal language, where all the pitch events on F0 contour contribute to the post-lexical function of intonation. In Chinese, although tone and intonation are independent in linguistic functions and phonological categories, the phonetic output of the two are inseparable from each other on F0 contour. In the present study, the discussion of intonation is conducted together with lexical tones, as they both compose the overall melodic contour of Mandarin speech.

4.4.1.1 The functional representation of prosody

In Mandarin speech, the prosodic events can be decomposed into two levels of functional representation. At the lexical level, tonal contour contributes to the distinction of word identity; at the sentential level, intonational prosody contributes to marking sentence prominence and defining prosodic boundaries. The functional representation of the prosodic events in a Mandarin utterance with neutral emotion can be demonstrated in the following Figure 4.23:

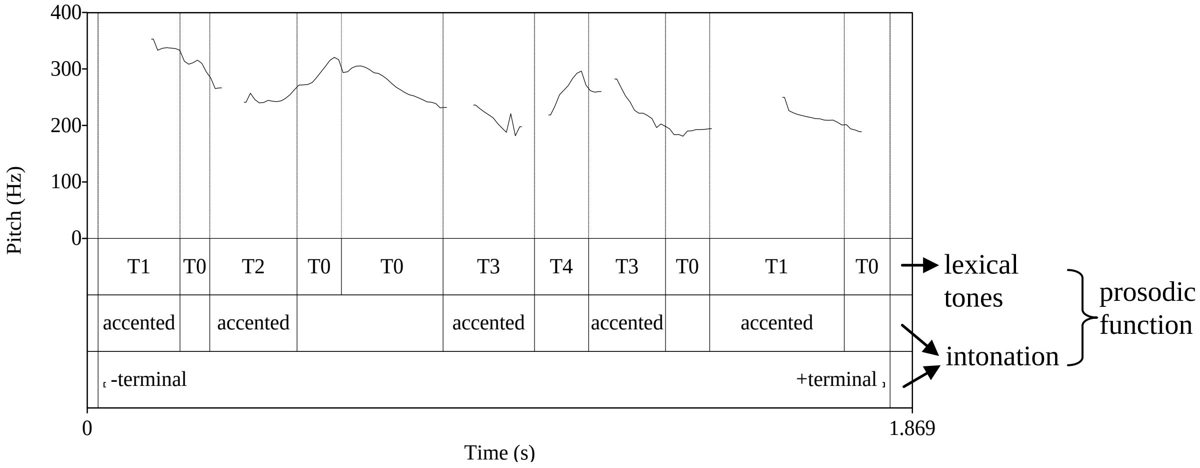


Figure 4.23: The functional representation of prosodic events

In section §4.2, the prosodic function at the lexical level for contributing to word identity has been discussed. Here I shall discuss the linguistic function of intonation at the sentential level, namely, the "weighting" function and the "grouping" function, with terms adopted from Gårding (1989).

a. The weighting function

The weighting function of intonation marks out the highlighted status of certain information from the utterance background. In section §4.3.1.2, a perception experiment was conducted with native subjects, in which listeners were asked to point out the salient syllables according to their own perception. In what follows, the data with more than 70% of agreement will be used. The reason for this choice is that this criterion provided the largest number of usable data.

To examine the features of accented syllables as compared to those of unaccented ones, two acoustic parameters are studied, syllable duration and the aligned tonal feature.

The duration of each syllable is normalized according to the Formula 4.5:

$$nT^{i}_{syllable} = \frac{T^{i}_{syllable}}{T_{mean}}, \quad i = 1, \cdots, m$$

Formula 4.5

Where $T^{i}_{syllable}$ represents the *i*th syllable duration in each utterance, *m* is the number of syllables in the utterance, and $T_{mean} = \frac{1}{m}\sum_{i=1}^{m} T^{i}_{syllable}$ represents the average duration of the utterance. In this study, 596 syllables (192 accented + 404 unaccented) in 60 utterances were normalized.

In the following two figures, the normalized duration of syllables based on the distinction between accented vs. unaccented categories is presented. The left figure shows that the average duration of accented syllables is longer than that of unaccented syllables, with the same result also found in the right figure when the syllables are distinguished with respect to tonal categories (Tone 1, Tone 2, Tone 3 and Tone 4). The atonal syllables (Tone 0) are all default unaccented syllables.

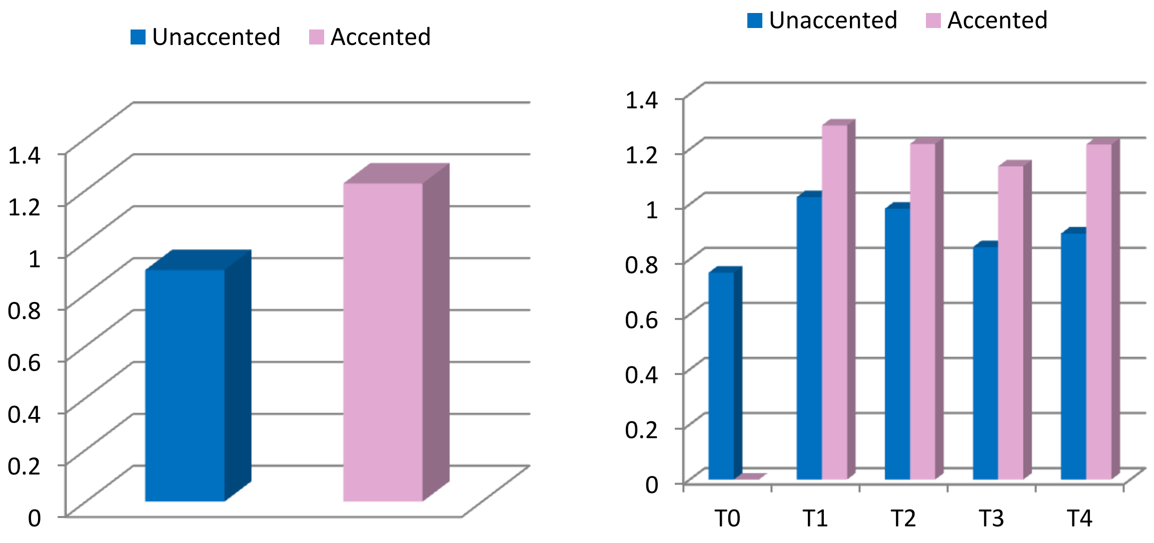


Figure 4.24: Duration of accented vs. unaccented syllables: the left figure revealed the comparison based on the overall average duration, the right figure

distinguished the respective tonal categories.

A computation is conducted on the correlation between accented syllables and the corresponding syllable duration. The following table reveals the syllabic information in each utterance for the correlation calculation, with the complete table found in **Appendix II**.

Utterance	Syllable	Tone	Accented	Sandhi	Normalized duration
(1)	ran	T2	accented	no	0.89853
	hou	T4	unaccented	no	0.56653
	ne	T0	unaccented	no	0.89932
	ta	T1	unaccented	no	0.68350
	zhao	T3	accented	no	1.47622
	wo	T3	accented	no	0.91015
	gan	T4	accented	yes	0.97240
	ma	T2	unaccented	no	1.35727
…	…	…	…	…	…

Table 4.10: The detailed information of each syllable in the target utterance

The result reveals that the Pearson correlation coefficient *r* between the accented syllables and the corresponding syllable duration is 0.487**, with the effect being statistically significant (** = $p \leq 0.01$). Therefore, it is confirmed that duration serves as an important intonational cue in marking out the accented status of prominent syllables.

In section §4.3.3, it was claimed that tone sandhi is an accent-related phenomenon. Tones of weak syllables often get neutralized and assimilated to the tonal features of adjacent strong syllables. The domain of tone sandhi process is conditioned by rhythmic grouping under the lexical-semantic cohesion principle.

In the present section, the 58 sandhi-tone syllables in the data are analyzed as a function of the involved accented and unaccented syllables, with the percentage of the two categories as in the following Figure 4.25.

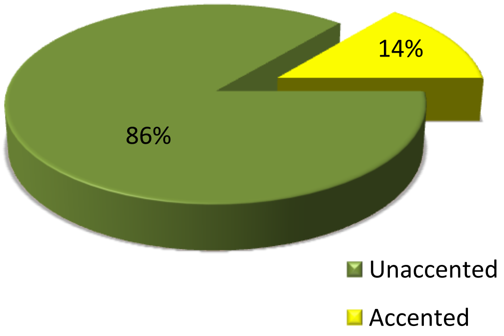


Figure 4.25: The percentage of accented sandhi-tone syllables and unaccented sandhi-tone syllables

From the above Figure 4.25, it can be seen that among the syllables with sandhi tone features, 86% are unaccented syllables, while 14% are accented syllables. Accordingly, the correlation between the accented status of syllables and the occurrence of tone-sandhi phenomenon is computed, with the Pearson correlation coefficient between the two as $r = -0.129$** (** = $p \leq 0.01$). The negative correlation coefficient reveals that the accented syllables are unlikely to undergo the tone-sandhi process, whereas the unaccented syllables are more likely to have sandhi tone features. Although the number of sandhi-tone syllables takes a small portion, only 9.7% of the whole number of syllables in the data, the correlation result provides evidence for the proposal that accentuation influences the aligned tonal manifestation of the syllable. The correlation between accented syllables and tone-sandhi phenomenon confirms the tonal stability of strong syllables, and the variable features of weak syllables.

b. The grouping function

Another important linguistic function of intonation is to indicate prosodic boundaries. The boundary markers of intonation serve for prosodic phrasing and indicate the finality or continuity between IUs in connected speech, as according to Cutler *et al*. (1997):

> Prosodic cues to the presence of a boundary have been the most reliable source of significant effects on parsing decisions.

The final syllable of an IU is often lengthened by speakers to mark finality. The comparison of duration between sentence final and non-final syllables is revealed in the following Figure 4.26. The duration of sentence-final syllables is always longer than that of non sentence-final syllables, as examined in three conditions: unaccented, accented, and average.

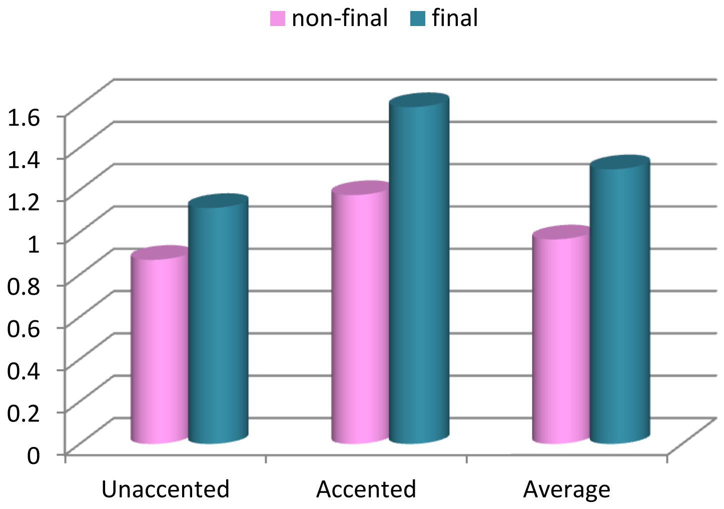


Figure 4.26: A comparison on duration of final syllables vs. non-final syllables in three conditions: unaccented, accented and average.

A correlation test is conducted on syllable position (sentence-final) and corresponding syllable duration. It emerges that the Pearson correlation efficient between the two is r = 0.311**, which is statistically significant (** = $p \leq 0.01$), revealing that the lengthened duration is closely related to final position in the sentence.

The pitch reset between two IUs is also a reliable cue in binding the prosodic juncture, as speakers often drop their pitch level toward the end of an IU, with a final low tone indicating finality. When the speakers start a new IU, they often initiate with a higher tone in marking the start of a new topic.

The declination in an IU has been observed as the gradual drifting-down tendency of the global pitch level. The declining effect in Mandarin declarative utterances was reported by G årding (1987), Shen (1990), Tseng (1981), Shih (2000), and Xu & Wang (1997).

Shen (1990: 26) stylized three intonation types of Mandarin utterances, as in the follow Figure 4.27:

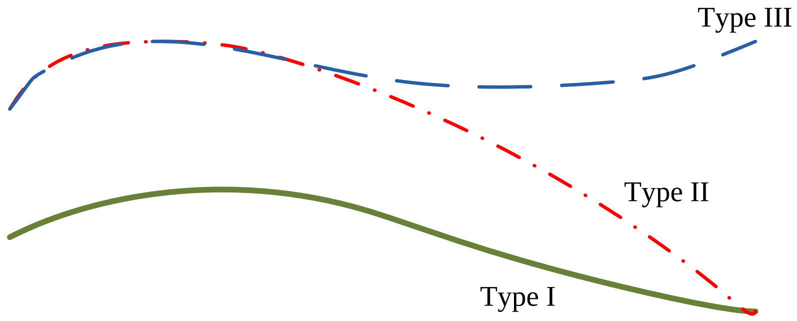


Figure 4.27: Three intonation patterns of Mandarin utterances in Shen (1990:26)

Type I represents the intonation pattern of declarative utterances, which starts from the mid key level, and ends in a low register. Type II generalizes the intonation pattern of unmarked questions and particle questions, which is featured with a mid-high key at the initial, and drops to a low register at the end, like in declarative utterances. Type III refers to the intonation pattern of A-not-A questions, which starts with a mid-high key level, and stays in the high pitch register till the end of the utterance. In Shen's study (1990), the declining intonation pattern was found in assertive sentences, unmarked questions and particle questions, but not in A-not-A interrogative sentences. Shen distinguished the intonation pattern of interrogative sentences from that of statements, as he claimed that in general the interrogative pattern has a higher initial pitch register than the declarative pattern. However, in Schack's (2000) study, the difference between questions and statements distinguished by pitch registers were not evidenced.

In an experimental study on F0 trajectories of Mandarin utterances, Shih (2000: 243-268) observed the declining effect in Mandarin statements, and proposed an exponentially decaying model in tracing the downtrend Mandarin pitch patterns, based on her discussion on the possible impact of sentence length, final lowering and prominence on the declination pattern. She found in the study that the declination rate is faster at the beginning of utterances and slows down as the utterances progress, accordingly, in her modelling, longer sentences were controlled with higher initial pitch parameters than shorter sentences. She also observed the impact of focus location on the declination pattern, that is, the declination slope is steeper in the post-focus portion than in the other locations. For the final lowering effect in Mandarin utterances, Shih reported that no significant evidence was found when the utterance ends with a high-level tone (Tone 1) syllable. However, she did not specify the final-lowering impact on utterances with other tonal types of syllables at the final position, and did not clarify the relationship between the declination tendency and the specific final-lowering effect, i.e., the additional lowering at the utterance end.

In this study, the declination phenomenon in Mandarin utterances is investigated by conducting a comparison on the normalized average F0 value between sentence final and non-final syllables. It is proposed that due to the down-drifting tendency of the global pitch level in each IU, the final syllable should have a lower normalized average F0 value than the syllables at the non-final positions.

To begin with the experiment, a normalization procedure is conducted on the F0 value of each syllable (596 syllables in total). For each syllable, only the average F0 value in the voiced portion is considered and normalized. The following formulas are used in normalizing the F0 value into the logarithm Z-score value. Such method of normalization is based on the suggestion of Zhu (2005: 52-57), in which six normalizing methods of raw F0 data were compared, proving that the logarithm Z-score normalization of F0 value yields the best normalized performance.

$$nF_{syllable}^{i} = \frac{\log_{10}(F_{syllable}^{i}) - \bar{F}_{mean}}{F_{std}}$$

Formula 4.6

$$\bar{F}_{mean} = \frac{1}{m}\sum_{i=1}^{m}\log_{10}(F^{i}_{syllable})$$

$$F_{std} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\log_{10}(F^{i}_{syllable}) - \frac{1}{m}\sum_{i=1}^{m}\log_{10}(F^{i}_{syllable})\right)^{2}}$$

In the above Formula 4.6, $nF^{i}_{syllable}$ represents the normalized F0 value of the *i*th syllable in an utterance, *m* the number of syllables in the utterance.

The following table presents the F0 information of each syllable in the given utterance, including the maximum F0 value, the minimum F0 value, and the mean F0 value in the voiced portion of the syllable. In this study, only the mean F0 value was selected and normalized, with the logarithm Z-score value presented in the final column of the table. The details for the 60 utterances can be found in **Appendix II**.

Utterance	syllable	Tone	Max F0	Min F0	Mean F0	Normalized Mean F0
(1)	ran	T2	204	161.5	175.6	-0.492
	hou	T4	218.5	204	214.8	1.050
	ne	T0	215.5	188.3	200.7	0.530
	ta	T1	227.9	221.2	222.2	1.309
	zhao	T3	221.6	179.5	202.5	0.599
	wo	T3	212.1	187.7	197.1	0.392
	gan	T4	175.6	162.4	171.3	-0.682
	ma	T2	174.3	147.8	154.7	-1.462
	ne	T0	166.4	152.4	159.2	-1.243
…	…	…	…	…	…	…

Table 4.11: The F0 information and normalized mean F0 of each syllable

With the average normalized F0 values of the 60 utterances computed according to Formula 4.6, a comparison is conducted between final and non-final syllables. In the following Figure 4.28, it can be seen that syllables at the final position have lower average pitch values than syllables at non-final positions. Such tendency can be observed in each tonal category, such as T0, T1, T2 and T4, while the distance is comparatively less obvious in T3 category.

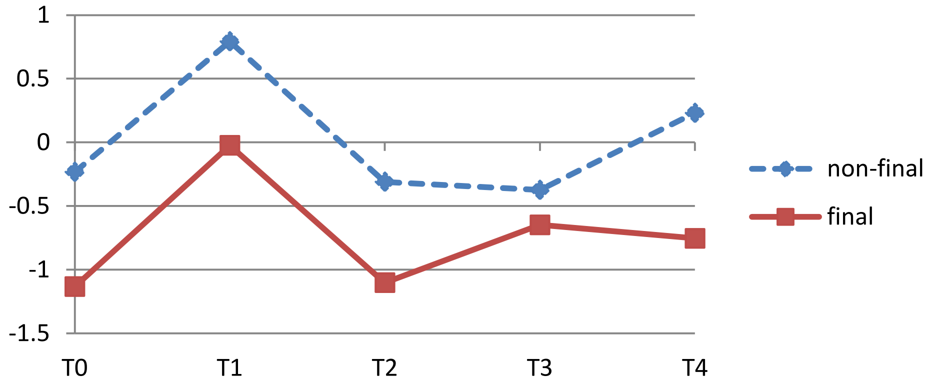


Figure 4.28: A comparion on normalized F0 value between final syllables and non-final syllables

A comparison between the normalized pitch values of final and non-final syllables in accented and unaccented conditions is also computed. The following Figure 4.29 shows that non-final syllables in the average have higher pitch values than final syllables, in both accented and unaccented positions.

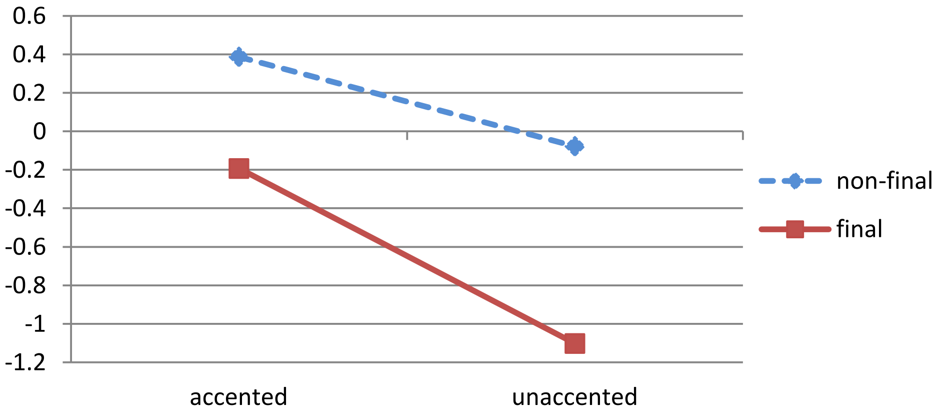


Figure 4.29: Normalized F0 value of non-final syllables and final syllables in accented and

unaccented conditions

The above Figure 4.28 and Figure 4.29 provide evidence that there does exist a declination effect in Mandarin speech. It should be noted that in the 60 utterances employed in the study, there are 58 statements and only 2 particle questions. Sentences with a question particle in Mandarin are believed to have the same intonation pattern as statement. Therefore, the experimental study lends support to the fact that Mandarin speech has a clear declination pattern at global pitch level.

To further prove the declination effect in Mandarin utterances, a correlation test is carried out between the sentence-final syllable and the normalized F0 value, which shows that the Pearson correlation coefficient *r* between the two is -0.238** (** = $p \leq 0.01$). This indicates that syllables at the final position of a sentence are more likely to have lower average pitch value than non-final syllables, which lends support to the global declining phenomenon in Mandarin speech.

Summarizing, in this section I have discussed the relevant features of prosodic function in Mandarin speech. For the functional representation in Mandarin prosody, two levels are distinguished: prosodic function at the lexical level and at the sentential level as follows:

At the lexical level: distinguishing word identity with Tone 1/Tone 2/ Tone 3/Tone 4 ;

At the sentential level: marking prominence with sentence accents ;
indicating prosodic boundaries with boundary tones

In the following section, I shall move on to the representation of the prosodic form of Mandarin speech. The aim is to pave the way for the discussion on mapping the abstract functional level of speech prosody with the complex picture of prosodic form.

4.4.1.2 The formal representation of prosody

Hirst (2000, 2005) proposed a multi-level organization for the form-function interface, and for the representation of prosodic form. A number of different levels are distinguished, that is: **the physical level**, **the phonetic level**, **the surface phonological level**, and **the underlying phonological level**, as demonstrated in Figure 4.30 .

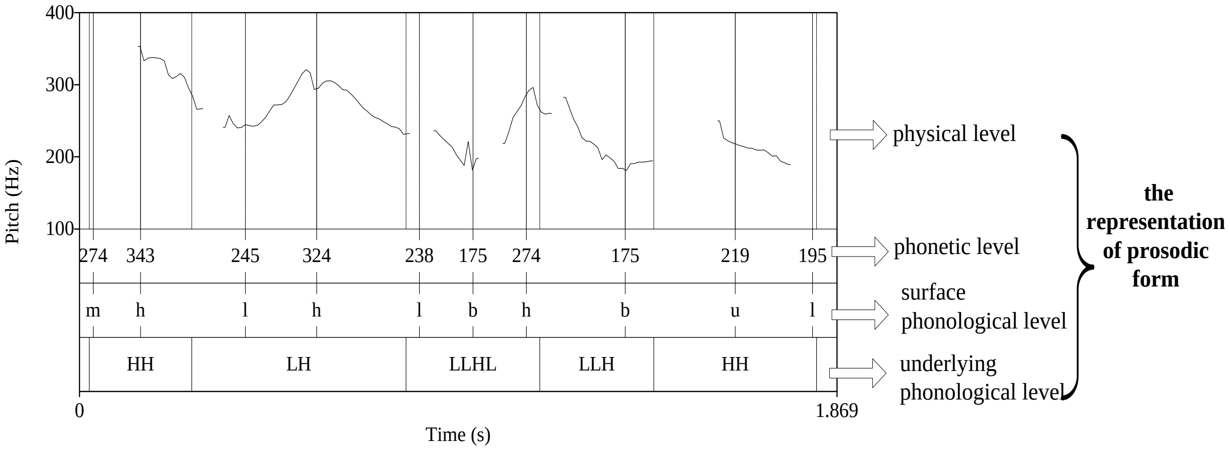


Figure 4.30: The representation of prosodic form at multi levels

- The physical level refers to the physical acoustic signal of the prosodic form;

- The phonetic level corresponds to the quantitative values of fundamental frequency, which are directly related to the acoustic signal;

- The surface phonological level refers to the sequence of discrete symbols, which are derived by having surface pitch pattern annotated with the INTSINT coding scheme; this level is still directly related to the observable features of the acoustic signal;

- The underlying phonological level serves as an intermediate phase between the representations of prosodic form and the linguistically significant functions.

The proposal of multi-level distinction serves for the mapping between the representation of form and function in prosody, which can be summarized in the following Figure 4.31. The theoretical status of each level in the figure is interpretable with respect to the adjacent levels:

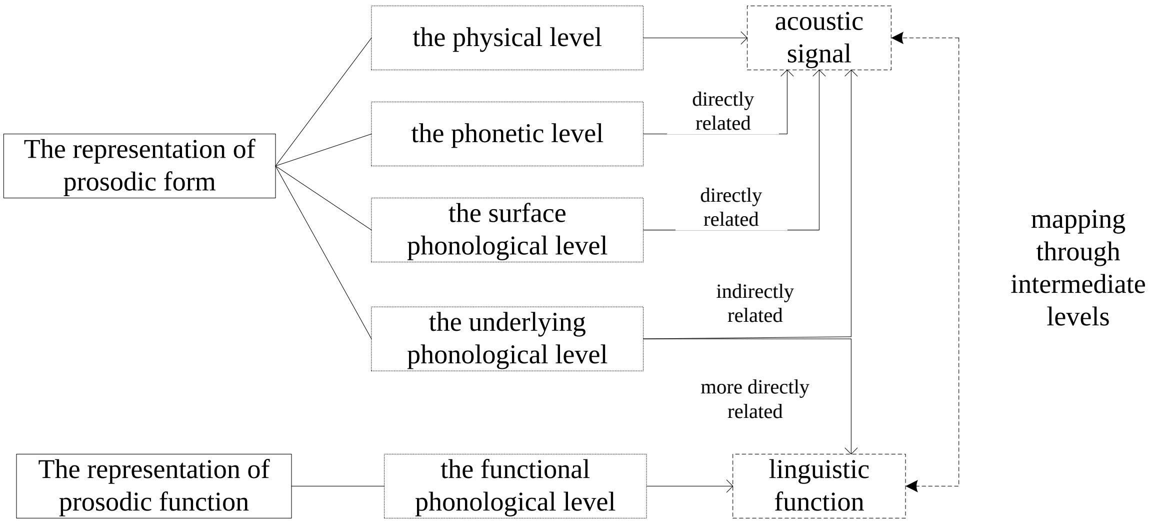

The representation of prosodic form
the physical level
the phonetic level
the surface phonological level
the underlying phonological level
acoustic signal
directly related
directly related
indirectly related
more directly related
mapping through intermediate levels
The representation of prosodic function
the functional phonological level
linguistic function

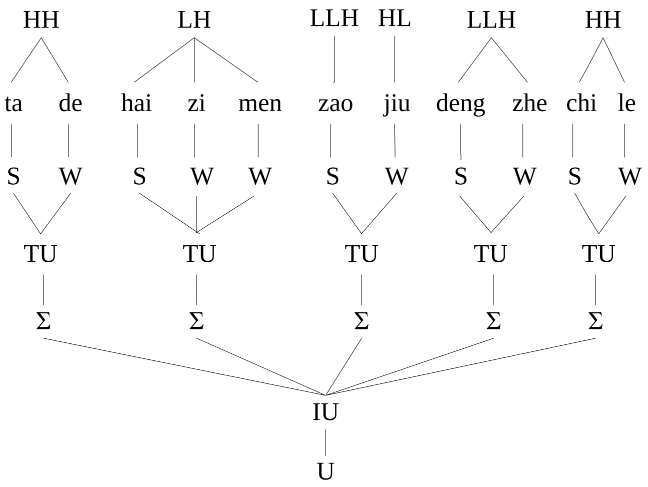
Figure 4.31: Multi-level mapping between the representation of prosodic form and the representation of prosodic function

a. At the underlying phonological level

As seen in the above Figure 4.31, the underlying phonological level of prosodic form is an intermediate phase which relates the linguistic function of prosody with the underlying formal representation.

In the previous discussion on lexical tones and tone sandhi process, the lexical prosodic form at the underlying phonological level was presented. In section §4.24, the *level* system was employed in annotating the citational forms of syllabic tones with two level targets, H and L targets. Accordingly, the four citational tones are represented as HH (Tone 1), LH (Tone 2), LLH (Tone 3) and HL (Tone 4). In section §4.3.3, the interaction between tone sandhi process and the sandhi domain conditioned by rhythmic grouping was discussed. It was concluded that the underlying form of lexical contours could best be represented in tonal units, as seen in the representation of the previously discussed utterance (1) as follows:

Utterance (1):



At the lexical level of prosodic form, the syllabic tones undergo the tone sandhi process within tonal units (TU) which could be larger than one syllable. The assembling of contours in TUs represents the underlying phonological level of the lexical prosodic form. Such level is directly related to the prosodic function of lexical tones.

For the underlying annotation of the intonation [±terminal] tones, the H/L level targets were also employed, with the initial tone of an intonation unit represented as [H (a high initial tone) or [L (a low initial tone), and the final boundary tone represented as H] (a high final tone) or L] (a low final tone).

As most of the sample utterances used in the study are neutral productions with no strong emotion or exaggerated emphasis, the melodic form of the utterance prosody is mostly contributed by the lexical tonal contours, with the intonational boundary tones for prosodic phrasing superimposed at the initial and final part of the utterance. As the sample utterance (1) is a statement, the terminal tone of the intonation unit (IU) is annotated with a L tone, while the initial tone of the IU is annotated with a H tone. Therefore, at the underlying phonological level, the pitch contours on tonal units together with the intonation boundary tones, compose the formal representation of a Mandarin IU.

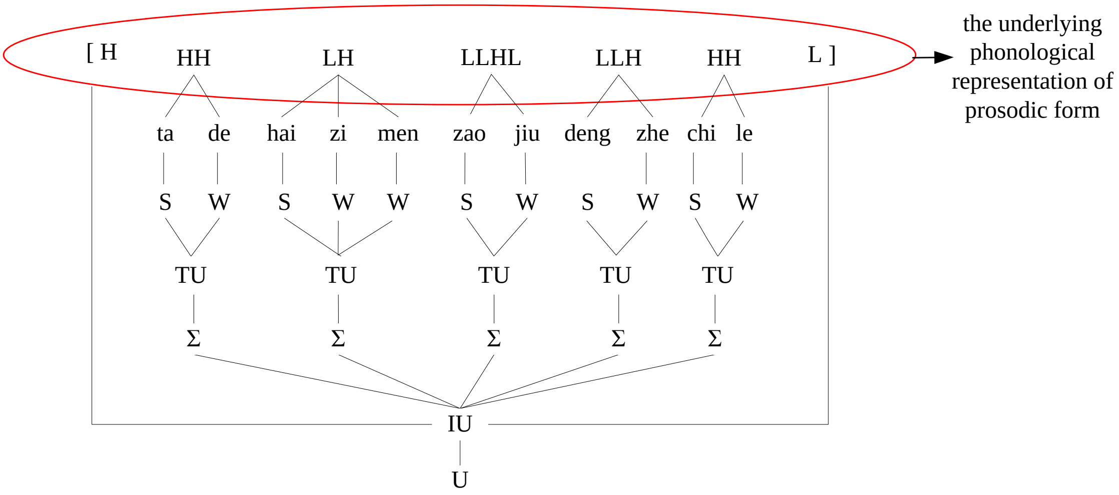


Figure 4.32: The underlying phonological representation of the prosodic form of Utterance (1)

The prosodic form at the underlying phonological level of an utterance is derived from the analysis of prosodic function, with the procedure demonstrated in the following Figure 4.33:

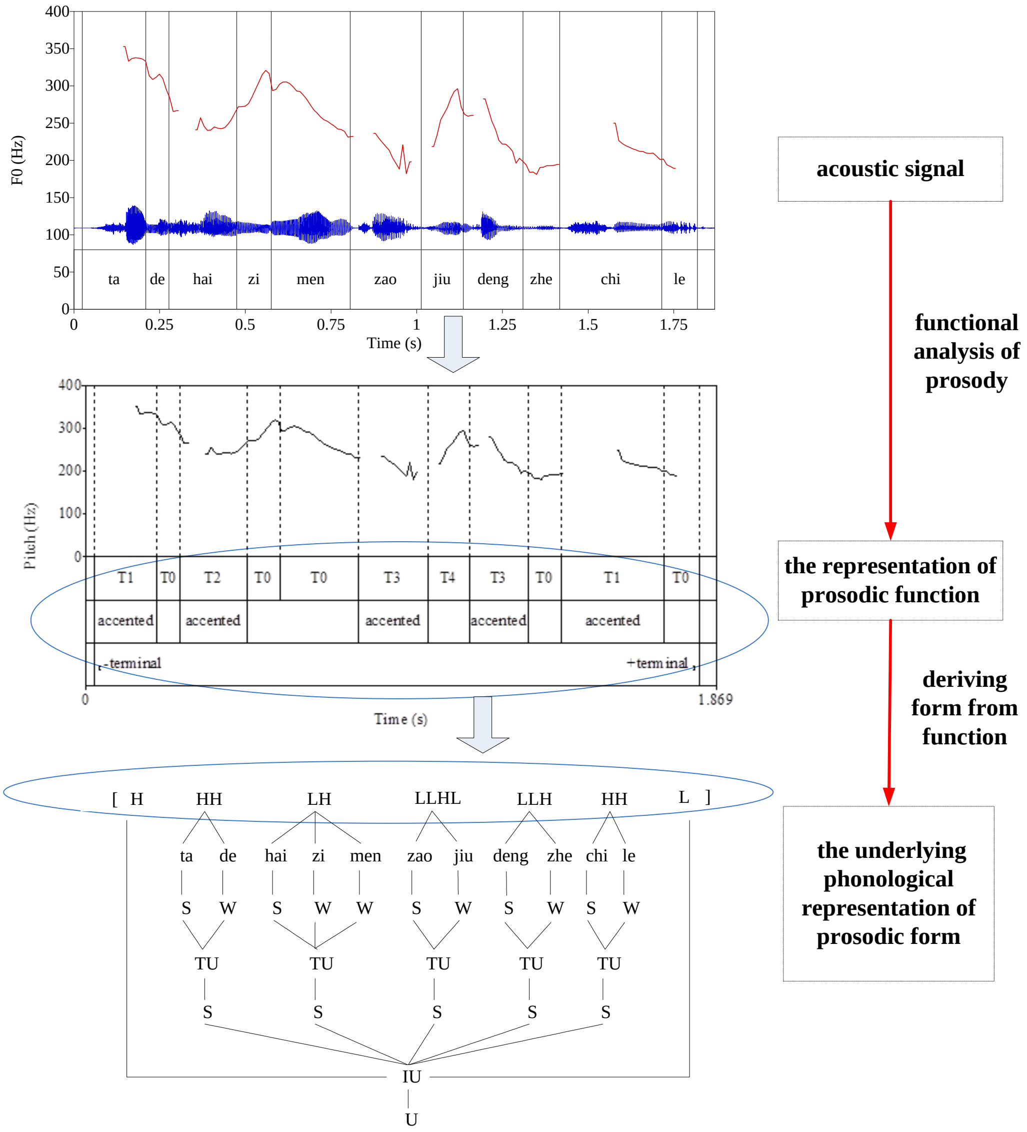


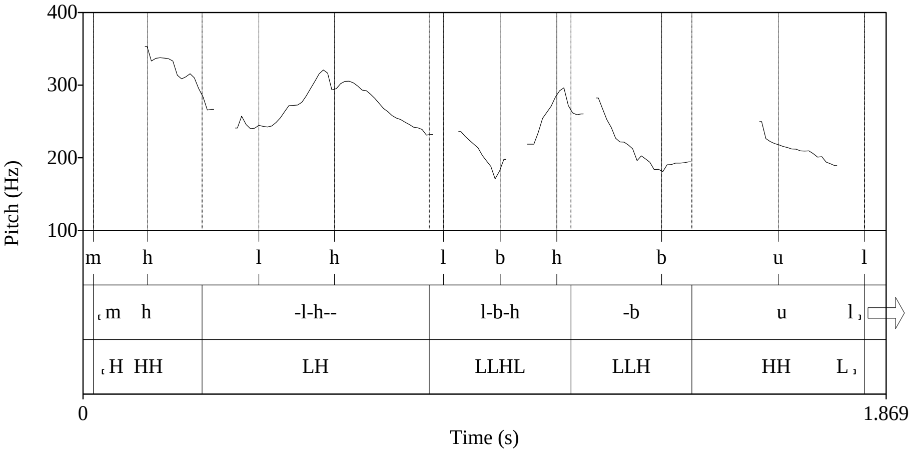
Figure 4.33: Deriving the underlying phonological representation of prosodic form based on the functional analysis of the raw acoustic signal

b. At the surface phonological level

Based on the annotation at the underlying phonological level, the surface melodic contour of a Mandarin utterance should be regarded as the assembling of the tonal contours in each tonal unit (TU) and the boundary tones of the overall intonation unit (IU). At the surface phonological level, the pitch property of the two units is annotated with the INTSINT system into symbolic representation. The resulted symbolic coding is related to the phonetic property of acoustic signal and also interpretable with respect to the underlying phonological representation.

The principle of annotation at the surface phonological level is to reduce the observable complexity of the acoustic signal to a simplified model, represented as a sequence of the INTSINT tonal symbols. Although the INTSINT system has not been formally applied in the annotation of pitch contour of a tonal language like Mandarin Chinese, it is believed that the surface pitch property of Mandarin can be studied in the same way as that of non-tonal languages. The annotation of Mandarin speech contour with the INTSINT system is explored below.

The surface pitch contours of 60 Mandarin utterances are coded into symbolic representation, thus the observable complexity of raw F0 data is represented as a sequence of tonal symbols. In the following Figure 4.34, the pitch contour of sample utterance (1) is manually annotated as:



the symbolic representation with INTSINT

Figure 4.34: The manual annotation of the surface contour pattern with INTSINT symbols

In Figure 4.34, the pitch contour of utterance (1) is annotated with INTSINT symbols, which results in a sequence of symbols, such as “[m h l h l b h b u l]”. The “[m” and “l]” symbols respectively code the initial and final boundary tones of the utterance. It is believed that speakers generally start from their mid-level pitch range in producing a neutral utterance, therefore, the [m symbol indicates the default initial tone of the utterance prosody. The final boundary tone of the utterance is defined according to the actual pitch tendency of the intonation unit. A low-tone target, l] is marked at the final boundary of the utterance, based on the obvious declining phenomena of the overall pitch pattern.

The peaks of pitch contour are generally labeled with *h* tones, instead of *t* tones, as it is aware that speaker seldom reaches the top level of his/her pitch range in producing a neutral emotional utterance. In this study, the INTSINT symbol *t* is only employed for annotating the peak of pitch contour in utterance with strong emphatic focus. The valleys in the utterance prosodic contour are annotated with *l* or *b* tones. The *u* or *d* is employed in marking the smaller changes of upward or downward pitch movement in the contour.

In the second labeling tier of Figure 4.34, it can be seen that the timing precision of target points aligned within each tonal unit is determined via the use of “dummy” targets represented by the symbol “-”. As explained in Hirst (2011a), when a tonal target sits in the middle of a tonal unit, it is coded as [X]; when the tonal target is at the third quarter of the tonal unit, it is coded as [-X], where the dummy target itself “-“ does not correspond to a pitch target, but can be equally spaced apart from the other target within the duration of the unit; when there are two tonal targets X and Y in the same tonal unit, with one at the first quarter and the other at the third quarter of the duration, the annotation is [XY]. The alignment between timing and the *i*th tonal target is determined with respect to the initial and final boundaries of the relevant tonal unit.

There is a mathematical formula defined in Hirst (2011a) for calculating the timing relation of the *i*th target in the sequence of a number of targets in a tonal unit, as follows:

$$\boldsymbol{t = start + \frac{(2i-1)[end-start]}{2n}} \qquad \textbf{Formula 4.7}$$

In Formula 4.7, *start* refers to the starting time of the tonal unit, and *end* refers to the ending time of the unit. However, as clarified by Hirst (2011a), the actual degree in timing precision of the targets can be either generally or precisely calculated. It is left to users to decide in their manual annotation. In the present study, the use of dummy targets is roughly estimated for locating the target tonal segments with respect to tonal unit junctures. The precision is left for evaluation by comparing the synthesized stimuli with the original utterances.

The surface phonological representation of speech melodic contour captures the salient properties of the acoustic data by means of the INTSINT symbolic coding. At the same time, such level of representation is also within the framework of the underlying phonological representation. The assembling of tonal contours in TUs as well as the boundary tones of the overall IU contain all the necessary information for expressing the functional contrasts of the original utterance.

The surface representation of melodic form is interpretable at the underlying phonological representation, as observed in the following matching between the two levels:

[H	HH	LH	LLHL	LLH	HH	L]	underlying phonological level
[m	h	lh	lbh	b	u	l]	surface phonological level

The surface phonological representation of prosodic form is derived within the framework of the underlying phonological level; at the same time, it directly follows the physical property of the acoustic data.

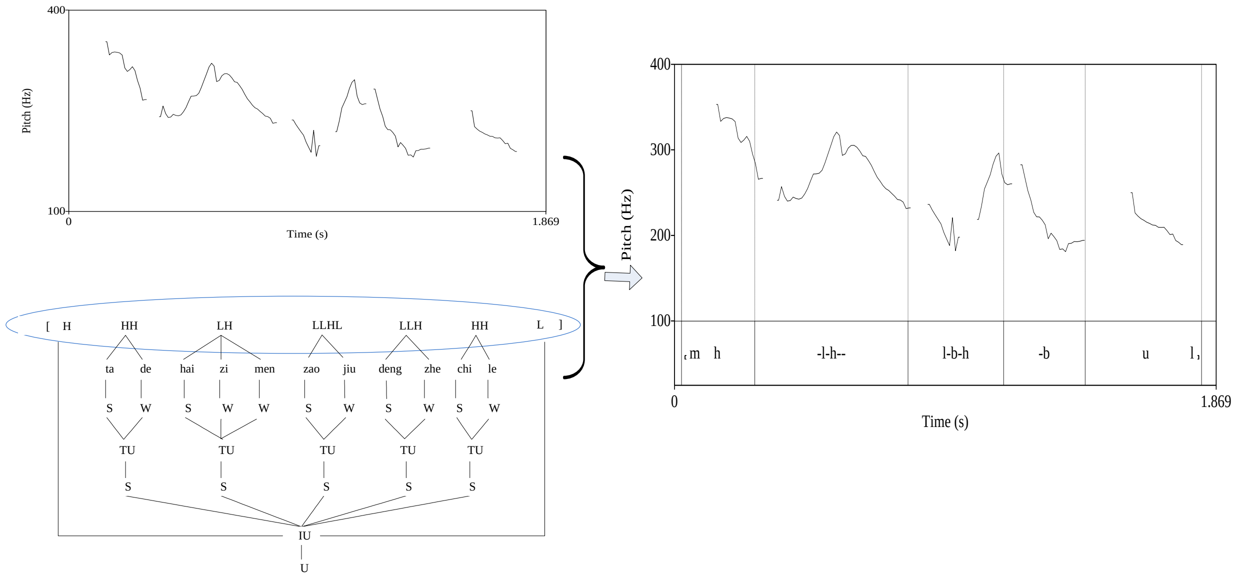


Figure 4.35: The surface representation of prosodic form is directly related to the acoustic signal and also interpretable to the underlying phonological representation

The annotation of a Mandarin utterance with INTSINT symbols is the surface coding of prosodic form. The procedure of such labeling is based on the principal of "extracting linguistic information from measurable physical data" (Hirst *et al*. 2000). The predicted data from such symbolic representation will be subject to evaluation with speech synthesis system presented in the following Chapter 5.

4.5 Summary

In this chapter, three important factors, lexical tone, rhythmic accent and intonation, which contribute to the function and form of Mandarin prosody, are discussed.

The focus of tonal study is on the association between the phonological function of tones and the phonetic nature of tone process in the speech flow. In order to understand the complex tone sandhi phenomenon, the discussion extends to various factors, such as the phonological tone rules, the unit for tonal manifestation conditioned by prosodic grouping and syntactic structure, the related functional load of tones in speech context, the distinction between optional tone sandhi and obligatory tone sandhi, etc. The importance of defining the domain in which the general rules of tone process apply is highlighted in the study.

The study of rhythmic accent was in alignment with discussions of tones and intonation, due to their close correlation on the formal and functional aspect of prosody. Accentual phenomena are closely related to the tonal manifestation. It is found that unaccented syllables are often related with modified or partially lost sandhi forms, while accented syllables are often associated with citational tone forms, and unlikely undergo a tone sandhi process. Accent is also in close correlation with intonation, as it is employed as an important means in marking intonational prominence and defining hierarchical structure in speech prosody. A metrical approach is conducted on the prosodic structure of each utterance, with the accented syllables at the sentential level alternating with the unaccented ones. The rhythmic grouping serves as a potential domain where tone sandhi process operates. In the present study, based on the structural analysis of speech prosody distinguished by sentential accents, the hierarchical structure of speech prosody is discussed. It is claimed that the prosodic organization in speech corresponds to the domain of functional relevance.

In this chapter the form and function of Mandarin speech prosody are proposed to be mapped with each other by derivation through a multi-level system. The theoretical status of each level is interpretable with respect to the adjacent levels. In Mandarin speech, the melodic form consists of lexically-specified pitch contours and the intonational use of pitch at the sentential level, which can be defined within two phonological structures, *intonation unit (IU)* and the subcategorical *tonal unit (TU*), with the former defined as in Daniel (2012), “the domain of long-term variation in both duration and pitch”, and the latter “the domain of short term pitch control”. For the representation of prosodic form, the INTSINT system is employed to code F0 contour into symbolic representation, thus deriving the predicted modelling of utterance prosody, which will be subject to speech synthesis in the following chapter.

CHAPTER 5

Analysis by Synthesis of Mandarin Prosody

5.1 Introduction

How to relate the physical acoustic information of the speech form in an appropriate way with its specified function is still a poorly understood problem in analyzing and modelling the prosody of natural languages.

To provide a potential solution to the mapping between the function and form of speech prosody, Hirst (2000, 2005) postulated a multi-level organization for the form-function interface, encouraging linguists to define mapping rules between the two aspects through an "analysis-by-synthesis" paradigm (Hirst 2011a) as shown in the following Figure 5.1:

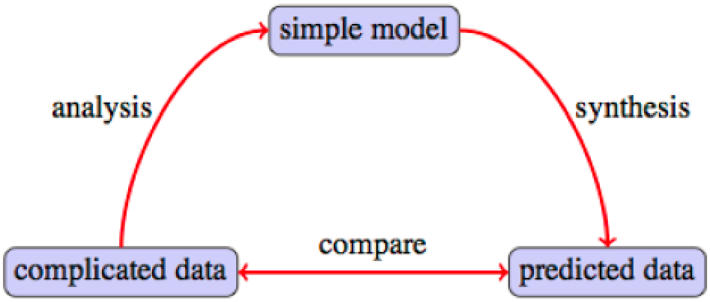


Figure 5.1: The "analysis-by-synthesis" paradigm

The paradigm in Figure 5.1 shows a reversible procedure in which the observable complexity of surface prosodic phenomenon (the raw F0 contour) is transposed into a stylized model, represented as a sequence of tonal segments (INTSINT symbols); as input, the symbolic representation (the “simple model”) is synthesized into acoustic data; such data, as “predicted” from the earlier step of analysis, is then compared to the original F0 contour; conversely, the comparison result is employed to evaluate the model. The cyclic process allows linguists to test and evaluate theoretical proposals directly with acoustic data with the help of speech synthesis technology.

The paradigm is based on using essential scientific criteria for the description of natural languages, as according to Hirst (2011a: 56):

> Scientists are confronted with the task of describing huge quantities of observable data. If they can reduce the complexity of the description by showing that the observable complexity is determined by some simpler, more abstract principle, then they have added to our knowledge of the data.

To complete the analysis-by-synthesis circle, a prosody editor for linguists, *Prozed* was designed in Hirst and Auran (2005), and further developed in Hirst (2011a, 2012). The implementation of Prozed is through a plugin in Praat. With this tool, linguists can manipulate the prosody of utterances by directly controlling the symbolic representation of prosodic form, and derive the immediate acoustic output for evaluating the coded alignment derived from the data analysis. In this chapter, the surface level of prosodic form coded with the INTSINT alphabet is tested as input for the speech synthesis module with the Prozed tool.

5.2 Evaluating the predicted data with Prozed

In Chapter 4, the functional analysis of Mandarin speech prosody at the lexical level and at the sentential level has provided a basic framework in which the underlying form of prosody is represented with annotations of tonal contours and intonation boundary tones. The two components manifest together at the surface level of prosodic form, which are symbolically represented with INTSINT annotation system. The symbolic coding of surface prosodic form is the predicted data which will be subject to the evaluation of Prozed synthesis system, with the generated stimuli compared to original utterances. The evaluation is aimed to improve the understanding of mapping between prosodic form and prosodic function.

The implementation of Prozed requires prosodic annotation via two interval tiers in Praat TextGrid: the annotation on Tonal Unit (TU) tier and on Intonation Unit (IU) tier. The general definition of TU and IU in Prozed implementation is the unit for encoding short-term pitch control, and the unit for encoding the long-term pitch variation, respectively (Hirst 2011a). The designer of the program made it clear that the two units do not have to strictly correspond to any phonological entities. The units can be experimented by different linguists with different assumptions for coding the pitch melody.

In the present study on Mandarin prosody, TU was discussed in section §4.3.3 as the unit conditioned by prosodic grouping. It serves as the domain where tone sandhi process operates. The pitch movement of surface melodic contour contributed by lexical tones in each TU and the intonation boundary tones of the overall IU are annotated by the INTSINT symbolic representation, as seen in the following window of annotation:

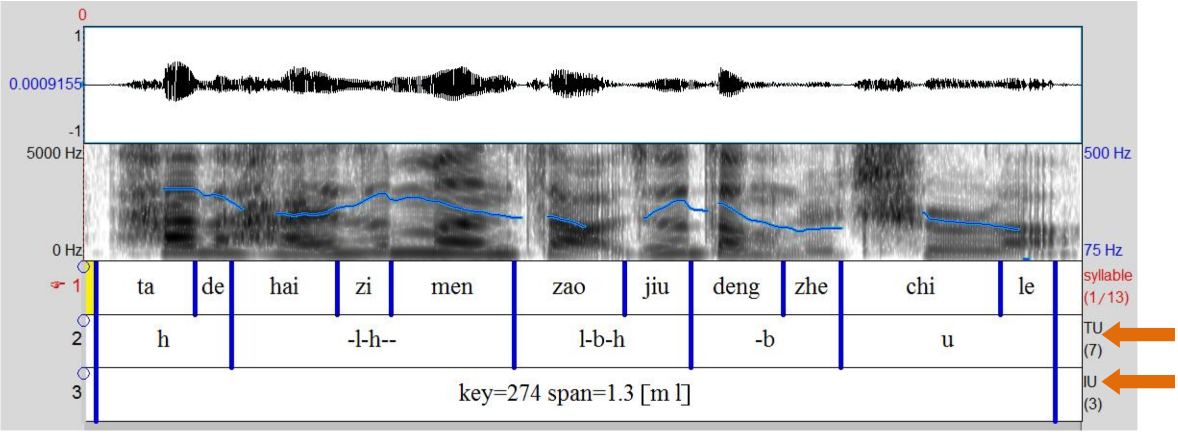


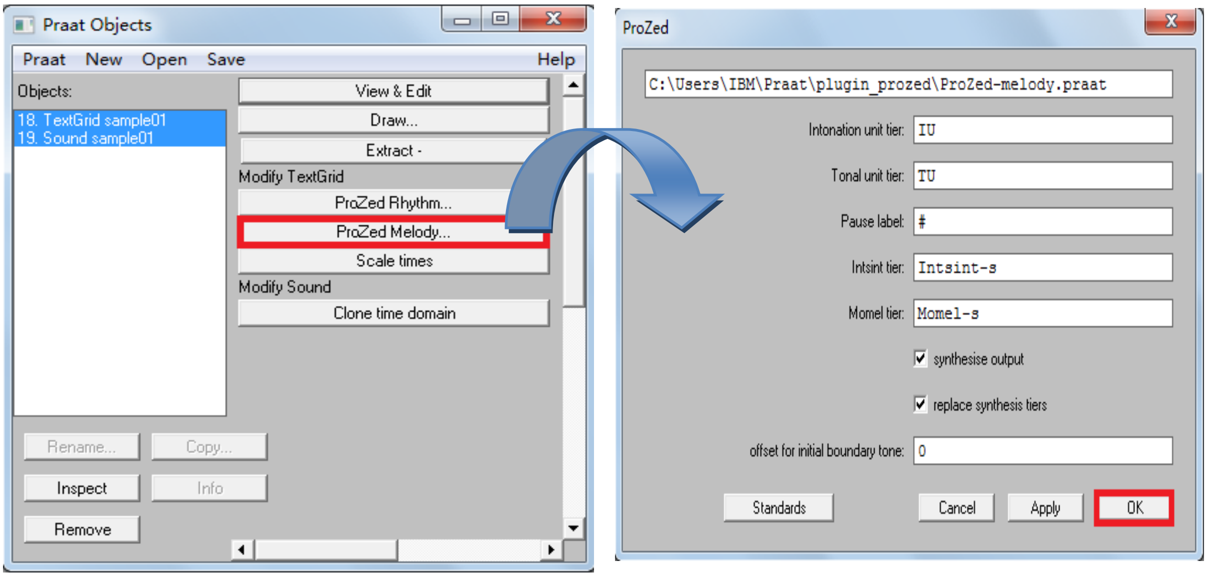
Figure 5.2: Annotation of a Mandarin utterance on TU tier and IU tier

The above Figure 5.2 shows a TextGrid window for annotating the prosody of a sample

Mandarin utterance. The symbolic representation of the utterance melodic contour is annotated on the TU interval tier (as “h-l-h…”) and on the IU interval tier (as “[m l]”). The symbolic coding of utterance prosody is the surface representation of prosodic form, which is derived through intermediate levels from functional analysis of prosody, with the details been discussed in section §4.4.2.2.

In the annotation window of the above Figure 5.2, besides the annotation of INTSINT tonal segments, two parameters *key* and *span*, are specified on the IU interval tier. The *key* parameter corresponds to the mid pitch level in the speaker’s current pitch range; the *span* parameter indicates the pitch range between the maximum and minimum pitch values which are symmetrical above and below the speaker’s *key* level. The acoustic values of the key parameter (in Herz value) and the pitch span (in octave value) of each IU can be automatically derived with the MOMEL algorithm.

The surface coding of the utterance melodic contour specified on TU and IU tiers in the above TextGrid are subject to Prozed implementation for speech synthesis. With the corresponding Sound file and TextGrid file selected together in Praat, the plugin of Prozed can be implemented, with the procedure indicated in the following two windows.



Objects:
18. TextGrid sample01
19. Sound sample01

C:\Users\IBM\Praat\plugin_prozed\ProZed-melody.praat

Intonation unit tier: IU

Tonal unit tier: TU

Pause label: #

Intsint tier: Intsint-s

Momel tier: Momel-s

☑ synthesise output

☑ replace synthesis tiers

offset for initial boundary tone: 0

Figure 5.3: The implementation window of Prozed

After the implementation step with Prozed, two new tiers, named “INTSINT-S” tier and “MOMEL-S” tier can be automatically generated and added to the original TextGrid. An output window of the Mandarin utterance TextGrid after Prozed application can be seen in the following Figure 5.4. On the INTSINT-S tier, the coded tonal symbols on TU tier and IU tier are displayed as a linear sequence of pitch targets, in which the boundary tones *[m* and *l]* are placed at the two extremes of the intonation unit, while the other targets are located relative to the boundary of each component unit of the TU tier. On the MOMEL-S tier, the corresponding fundamental frequency values aligned with the pitch targets on the INTSINT-S tier are presented.

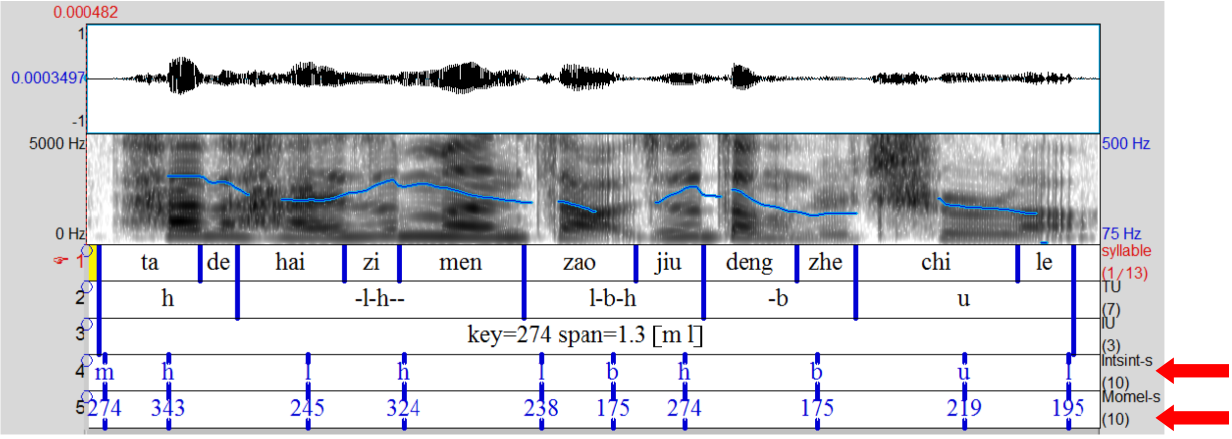


Figure 5.4: Output of the Prozed implementation of a sample utterance

The fundamental frequency values of the pitch targets on the MOMEL-S tier are defined according to the following formulas (Hirst 2012), in which the F0 values of absolute tones *t(op)*, *m(id)* and *b(ottom)* tones are determined by the *key* (274Hz) and *span* (1.3octave) parameters. The values of the relative tones, *h(igher)*, *l(ower)*, *s(ame)*, and of iterative tones, u(pstepped) and d(ownstepped), are defined based on the immediate *p(receding)* tone values:

Absolute tones:

$$\begin{aligned} t(op) &= key * \sqrt{2^{span}} \\ m(id) &= key \\ b(ottom) &= key / \sqrt{2^{span}} \end{aligned}$$

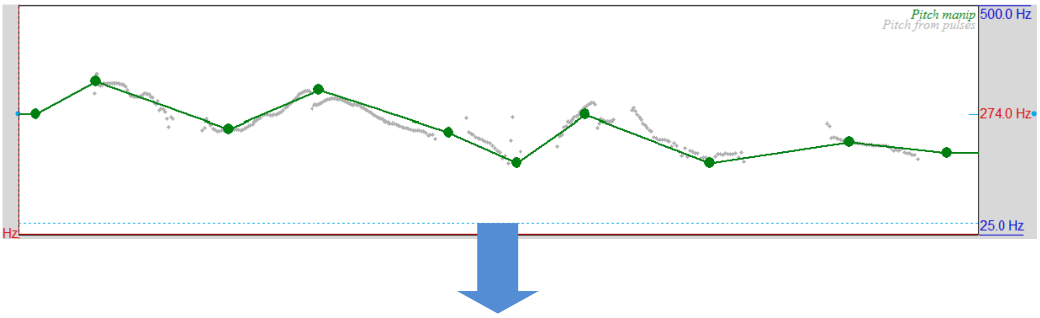
Relative tones:

$$\begin{aligned} h(igher) &= \sqrt{p*t} \\ s(ame) &= p \\ l(ower) &= \sqrt{p*b} \end{aligned}$$

Iterative tones:

$$\begin{aligned} u(pstepped) &= \sqrt{b*\sqrt{p*t}} \\ d(ownstepped) &= \sqrt{t*\sqrt{p*b}} \end{aligned}$$

The resulting pitch targets from Prozed application is in the form of a PitchTier, as seen in the top window of the following Figure 5.5. The green-colour points are the resulting pitch targets, which can be interpolated with a quadratic spline curve, leading to a smooth and continuous pitch contour in the second window of Figure 5.5. The synthesized output of the utterance is in the form of a PitchTier, which can provide the direct acoustic evidence for users to compare the synthesized version with the original utterance.



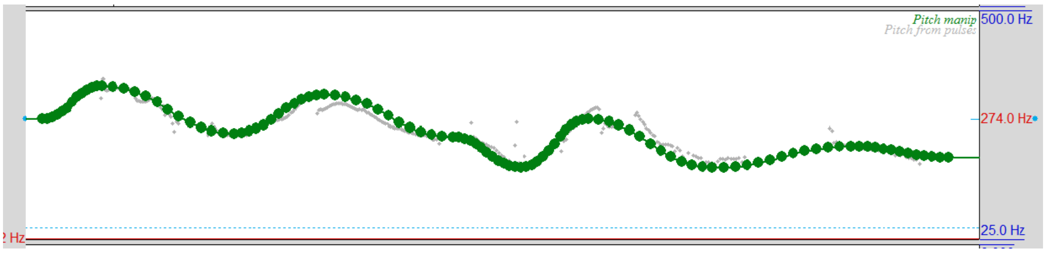


Figure 5.5: Synthesized output of the utterance prosody

In Figure 5.5, one can find that the synthesized pitch contour (the green-colour contour) closely follows the pitch movement of the original utterance (the discontinuous grey-colour curve). The synthesized output of the Mandarin utterance according to the author's perception also correctly conveys the prosodic function both at the lexical (tones) and at the sentential (sentence accents and boundaries) levels of the original utterance.

The evaluation of the predicted annotation of the 60 Mandarin utterance prosody is conducted by means of the comparison between the synthesized output and the original utterance, as presented in **Appendix III**. It is shown that a satisfactory synthesized result can be obtained from the surface representation of prosodic form with the INTSINT system. The symbolic representation of speech melody provides an access in reducing the observable complexity from a large quantity of data to a more simplified model, which only retains all the necessary information for expressing the functional contrasts in speech. The symbolic representation at the surface phonological level of prosodic form can also be mapped through multi levels to the functional level of prosody.

In the present study, the 60 Mandarin utterances (30 read-speech utterances and 30 spontaneous utterances) are subject to functional analysis. The annotation at the functional level provides the basic framework for the underlying phonological representation of prosodic form. The lexical contours aligned on each tonal unit together with the intonation boundary tones are represented at the surface level with the INTSINT alphabet. These are the predicted manual annotations, and are tested with the Prozed tool for deriving the synthesized output.

As presented in **Appendix III**, a satisfactory result can be obtained by comparing the 60 pairs of synthesized stimuli and the original utterances. One sample utterance is shown in the following Figure 5.6, with the annotation of the utterance prosody conducted in the Textgrid of Praat. There are in total six interval tiers used for the annotation as follows:

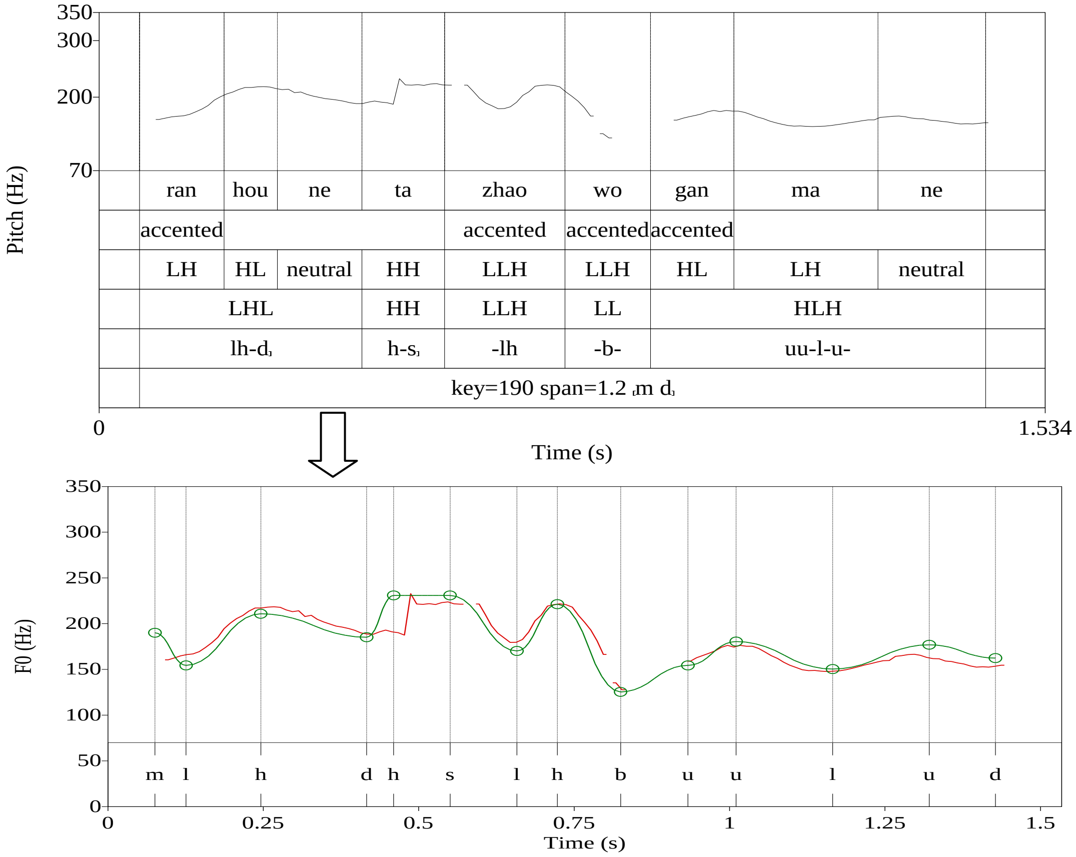


Figure 5.6: Annotation textgrid and the derived synthesis window

- The first tier codes the component syllables of the utterance;

- The second tier indicates the sentence accents, with the syllables perceived as the most salient ones according to the perception experiment marked out as [accented];

- The third tier codes the citational forms of the syllabic tones in the utterance;

- The fourth tier presents the tonal pattern of each rhythmic unit, derived from the various phonological and phonetic tone processes occurred in the units;

- The fifth tier shows the manual symbolic coding of the surface prosodic form on TU with the INTSINT alphabet;

- The sixth tier codes the *key* and *span* parameters and the boundary tones of IU of the utterance.

In the above Figure 5.6, the symbolic annotation on the fifth (TU tier) and sixth (IU tier) interval tiers in TextGrid underwent Prozed for synthesis implementation, leading to the following window with synthesized contour marked in green colour, in comparison to the original pitch contour in red colour. One can see that the synthesized contour, resulting from the interpolation of annotated symbols at the surface phonological level, can closely capture the features of pitch movement of the original utterance. The proposal of deriving the formal representation from the functional analysis of prosody makes the strategy of the whole dissertation explicit, as shown in Figure 5.7. The same graph has also been previously presented in section §1.3.2.

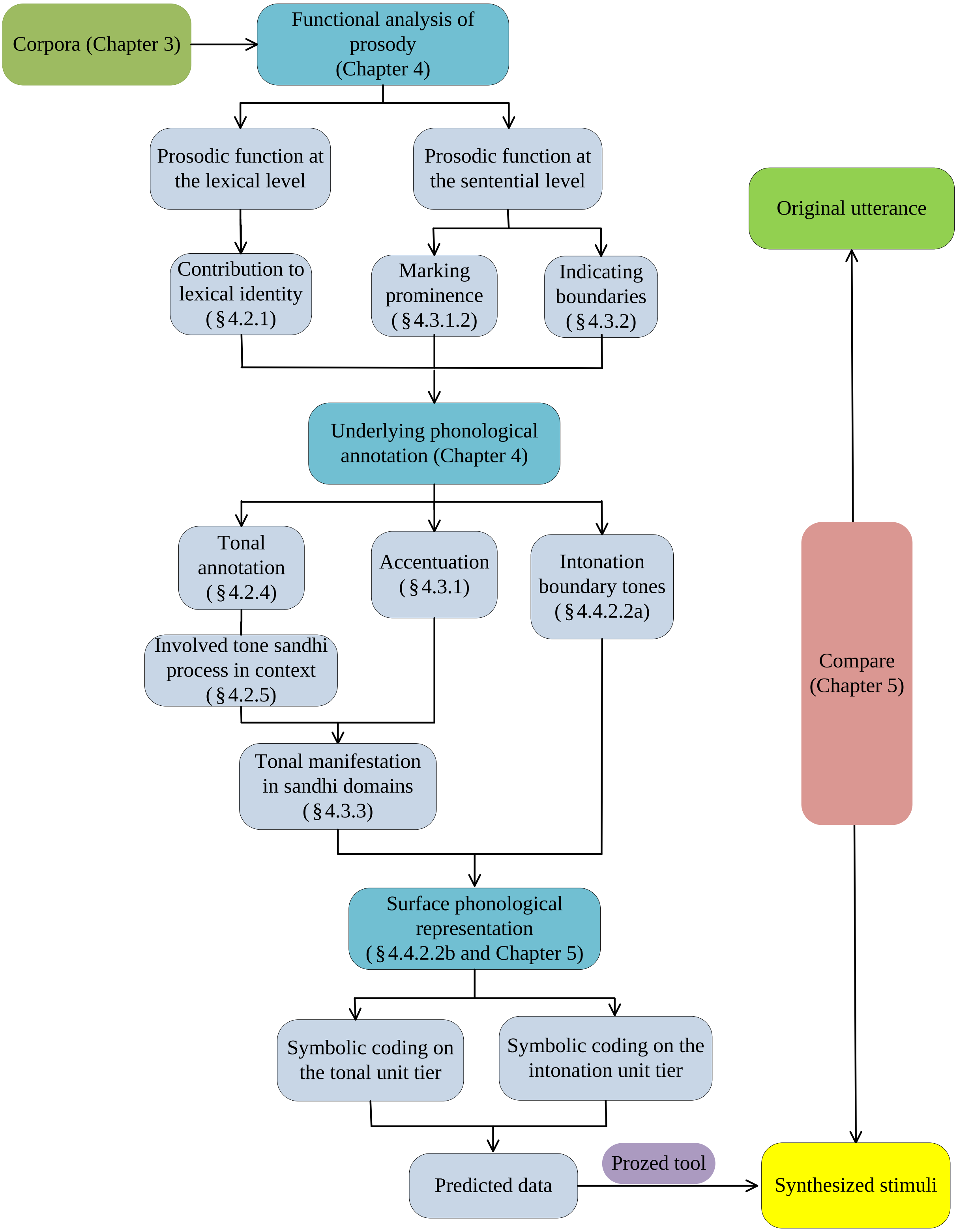

Corpora (Chapter 3)
Functional analysis of prosody (Chapter 4)
Prosodic function at the lexical level
Prosodic function at the sentential level
Original utterance
Contribution to lexical identity (§4.2.1)
Marking prominence (§4.3.1.2)
Indicating boundaries (§4.3.2)
Underlying phonological annotation (Chapter 4)
Tonal annotation (§4.2.4)
Accentuation (§4.3.1)
Intonation boundary tones (§4.4.2.2a)
Compare (Chapter 5)
Involved tone sandhi process in context (§4.2.5)
Tonal manifestation in sandhi domains (§4.3.3)
Surface phonological representation (§4.4.2.2b and Chapter 5)
Symbolic coding on the tonal unit tier
Symbolic coding on the intonation unit tier
Prozed tool
Predicted data
Synthesized stimuli

Figure 5.7: A mapping scheme between the functional aspect and formal aspect of prosody

5.3 Discussion

In the present study, the formal representation and the functional analysis of utterance prosody are interrelated to each other by means of derivation through a multi-level system. In the following Figure 5.8, the representation of prosodic function and the representation of prosodic form are annotated in the left top textgrid. The predicted data evaluated by means of speech synthesis, as the stimuli (the red-colour disconnected contour) in comparison to the original utterance (the green-colour smooth and continuous contour) is presented in the bottom window.

The theoretical status of the representation at each level can be interpretable to the other: the function of an utterance prosody is annotated at the lexical level (with the lexical tones represented as T1, T2, T3, T4, T0) and at the sentential level (with the sentence accents and boundary tones marked respectively as [accented] and [±terminal]). Such **representation of prosodic function** provides a basic frame for the coding of prosodic form. At the **underlying phonological level**, the tone process involved in each tonal unit is considered, with the tonal features of each unit annotated with the H, L sequence. At the **surface phonological level**, the utterance pitch contour yields to symbolic representation with the INTSINT system. Such level of representation serves as an intermediate phase, as being related to the surface phonetic property of the acoustic signal, and also interpretable with respect to the underlying phonological representation. The **phonetic level** consists of the quantitative F0 values, with each value corresponding to the target point annotated at the above surface phonological level. The **physical level** is the acoustic pitch contour, resulted from the interpolation of the F0 targets, which are derived from the Momel-INTSINT system at the phonetic level.

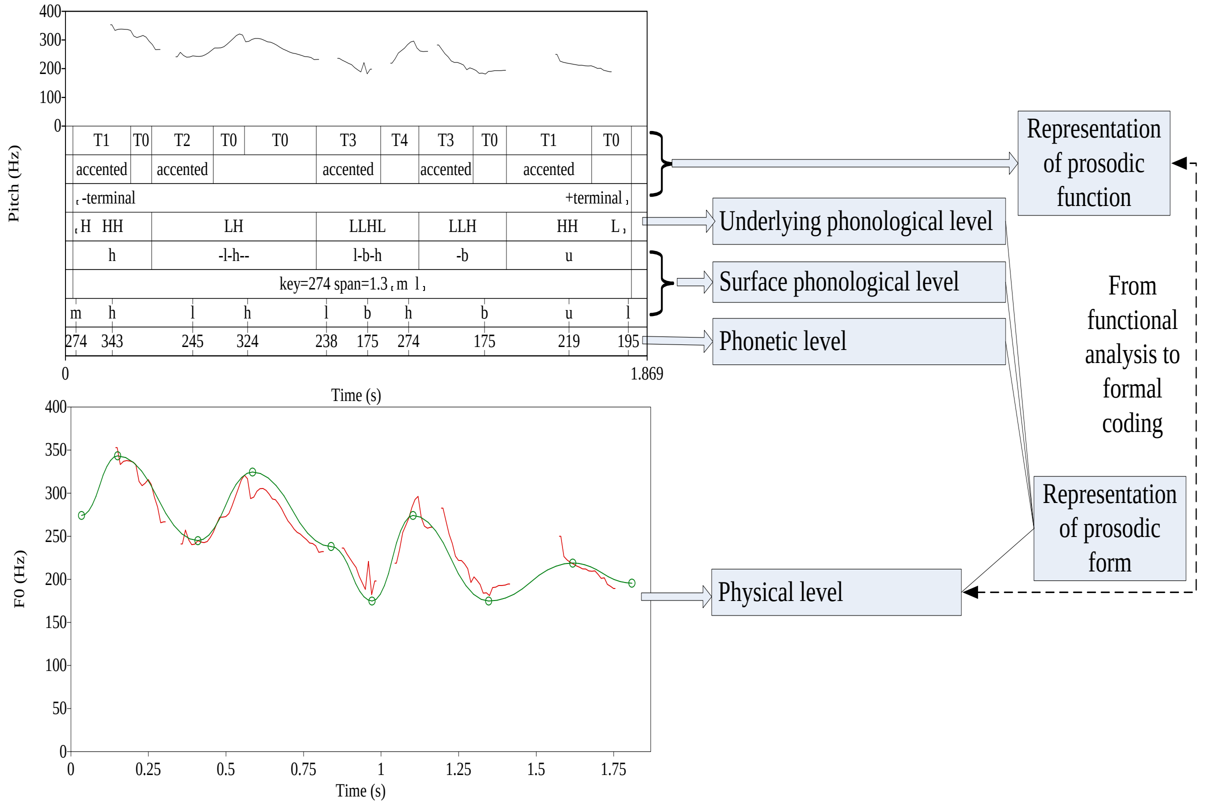


Figure 5.8: Annotation textgrid and the mapping between prosodic form and function

In this study, a high-quality synthesis result of Mandarin data can be derived based on the functional analysis and the surface coding of prosodic form. Such attempt in deriving prosodic form from functional representation provides a plausible method for associating the two aspects of prosody with each other. The synthesis stimuli of the 60 Mandarin utterances are perceived by the author as satisfactory, as the lexical tones and the sentence accents of each stimulus can well follow the melodic features of the original one. The comparison between the 60 pairs of synthesized stimuli and the original data are listed in **Appendix III**. It is shown that the synthesized contour closely captures the contour movement of each utterance.

In Prozed implementation, two interval tiers are required for annotation. On IU tier, the global parameters of *key* and *span* are annotated. The referential pitch span of a neutral utterance is 1 octave, but can rise to 2 octaves when the utterance has a strong emphatic focus. The average pitch span of the selected 60 utterances (all plain emotionless utterances) is 1.29, while a detailed categorization of pitch span is also computed and presented in Figure 5.9 based on the distinction of speech style (read and spontaneous), the gender of speakers (female and male), and the average value. It can be seen that the pitch span of read speech is narrower than that of spontaneous speech; female speakers in general employ a larger pitch span than male speakers.

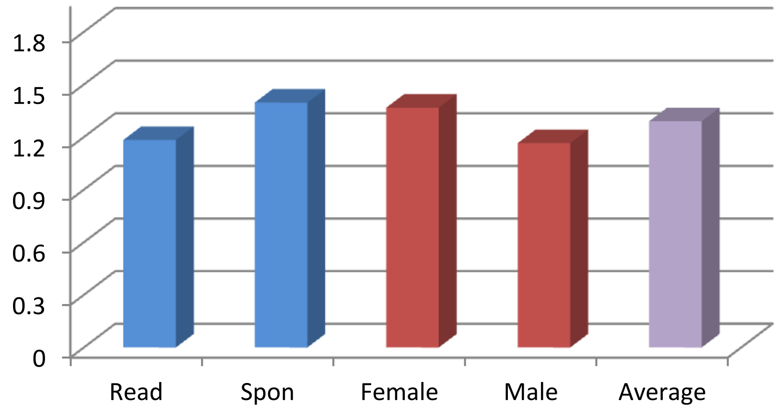


Figure 5.9: The pitch span of speech data based on the distinction of speech style (read vs. spontaneous), the gender of speakers, and the average span value

The parameters of key pitch values in the speech data are also computed in semitones and compared in the following Figure 5.10, in which semitone = 12log$_2$ (F0/F$_{\text{reference}}$), F$_{\text{reference}}$ = 1. It is seen that the key level in speech only distinguishes between female and male speakers.

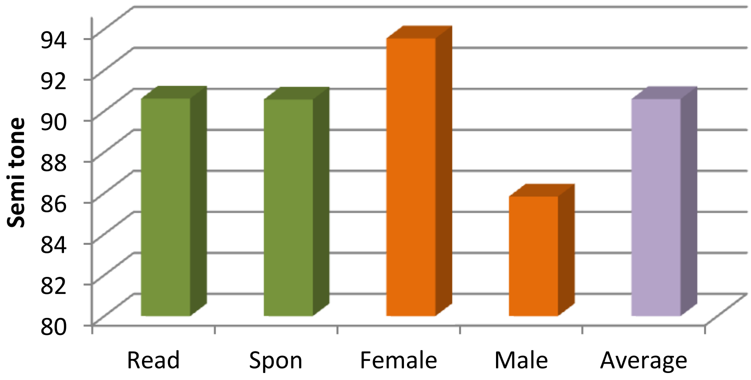


Figure 5.10: The key pitch level of Mandarin speech data, distinguished in read vs. spontaneous speech styles, the gender of speakers and the average key value

For the melodic annotation on TU tier, it seems necessary to “find the timing system in speech, to be specific, the basic speech units in order to capture the phenomena in different timing” (Hirst 2011b). In Mandarin utterances, due to the nature of the human articulatory mechanism, tones are produced in natural assimilatory process, in which the default tone target of each syllable often gets spanned to a tonal unit, larger than a syllable in speech prosody. The main focus of this tonal study is based on the relationship between utterance hierarchical structure and phonetic nature of the tone process. In testing Prozed with Mandarin utterances, it merged that a high-quanlity synthesis could be generated by coding the tonal contour movement within each tone-sandhi unit. The surface melodic form of Mandarin utterances contributed by both tonal contours and the intonational boundary tones are annotated with INTSINT tonal segments. A statistic calculation on the usage of each INTSINT segment in the annotation of Mandarin speech data is presented in the following Figure 5.11.

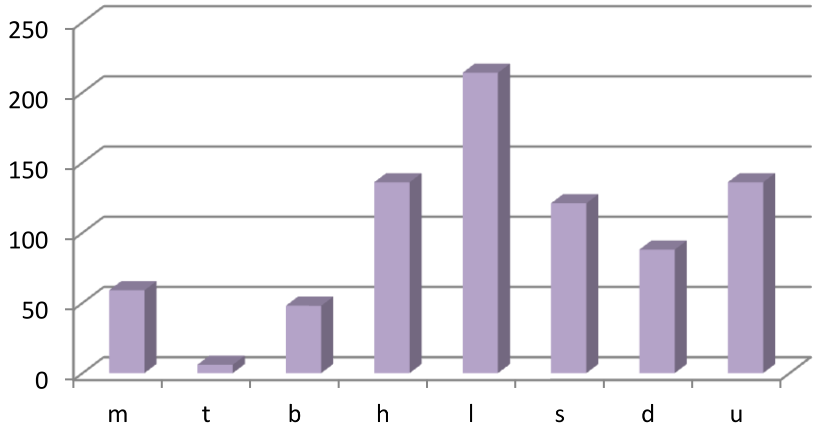


Figure 5.11: The use of INTSINT symbols in the annotation of 60 Mandarin utterances

It is shown that the absolute tones *m*, *t* and *b* are least employed in the annotation, while the relative tones *h*, *l*, *s*, and iterative tones, *d*, *u* are more frequently used. To further specify the use of INTSINT codes in annotating the intonation boundary tones of IUs in the data, another figure is presented for a review of the symbols employed for initial and final tones of IU as follows,

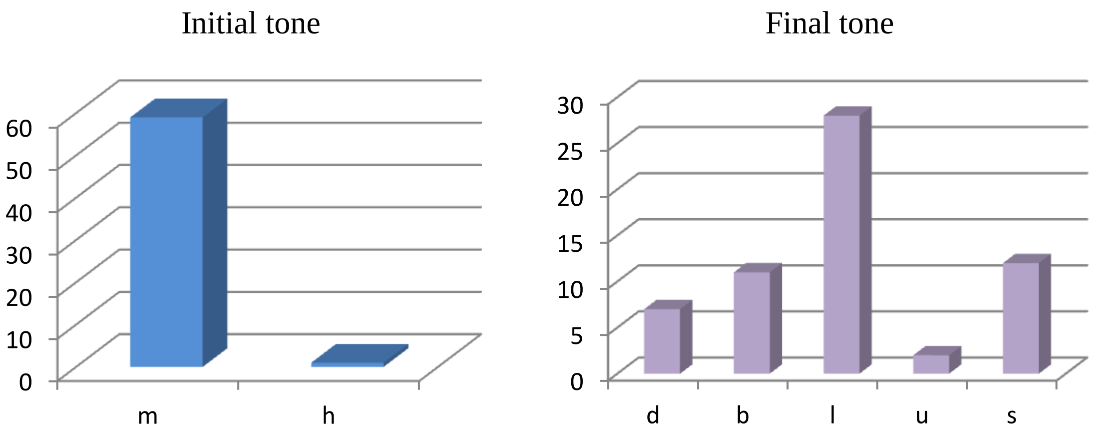


Figure 5.12: The use of INTSINT symbols in annotating the boundary tones on IU tier

For the annotation of initial tones, the most employed target is *m*, as it is held in the study that speakers generally start from their mid key level in articulating a neutral utterance.

The final tone of each utterance is annotated based on the actual pitch tendency of the IU. When the unit ends with a low tone, the INTSINT symbol, *l*, *d* or *b* is used for labelling the low final boundary tone. Such final-lowering phenomenon is observed in the IUs which have a low-tone syllable such as Tone 3 and Tone 4, and also the weak Tone 0 syllable at the end. When the IU has a Tone 1 syllable at the end, the final-lowering effect is not significant, with most final boundary tones either annotated with *u*, or *s* (following the preceding high-tone target), which indicates that the pitch level at the IU end still remains in the upper pitch level due to the previous high-tone syllable, instead of falling to a low tone. The same result was also observed in Shih's study (2000): the utterance ending with a high tone is not influenced by the final-lowering impact. In the present study, among the 13 utterances which end with a Tone 1 syllable, only 3 utterances were annotated with a lower final tone, while the other 10 utterances were annotated with a high final boundary tone. However, it deserves notice that the final-lowering of the utterance is not identical to the declination phenomenon indicating the **global** drifting-down of pitch values discussed in section §4.4.2.2b, whereas final-lowering only refers to the relative additional lowering at the final part of the utterance.

5.4 Summary

In this chapter, an analysis-by-synthesis study is conducted. The overall melody of each utterance is coded with INTSINT symbols on two tiers, tonal unit (TU) and intonation unit (IU) tier. In Mandarin speech, TU and IU can respectively correspond to the domains where the tone process occurs and where the intonation pattern displays itself. Through a cyclic multi-level derivation, the symbolic representation of prosody at the surface phonological level is not only closely related to the acoustic data, but also strongly related to the corresponding underlying phonological representation of lexical tones and intonation contour. The abstract prosodic model, i.e. the predicted coding of Mandarin prosody, is tested through the speech synthesis technology, the Prozed tool. By comparing the original acoustic signal with the synthesized stimuli, it is shown that a satisfactory synthesis of the original utterance could be obtained from the INTSINT symbolic coding, as the synthesized data not only captures the pitch features of the original utterance at the physical level, but is also interpretable with respect to the prosodic functions.

CHAPTER 6

Conclusion

In the present study, the relationship between the functional aspect and the formal aspect of Mandarin speech prosody is analyzed by defining a plausible way in which the two can be mapped onto each other.

Mandarin, as a typical tone language, employs prosody at both the lexical level and the sentential level, with the former contributing to the distinction of word identity; and the latter contributing to the intonational function of marking sentential prominence (the “weighting function”) and indicating prosodic boundaries (the “grouping function”).

The study centers on the features of three salient components of Mandarin speech prosody, namely, lexical tone, rhythmic accent and intonation.

Pitch contour serves as the primary acoustic parameter of tone and intonation. The binary uses of pitch information makes the whole picture of Mandarin prosodic form more complicated than that of non-tonal languages. The lexical and the intonational use of prosody are represented at the underlying phonological level by means of two level targets, H(igh) and L(ow) tone targets. Accordingly, the citational forms of four tones are represented as HH (Tone 1), LH (Tone 2), LLH (Tone 3) and HL (Tone 4). The intonational boundary tones are also represented with the H/L targets, with the initial tone of an intonation unit represented as [H (a high initial tone) or [L (a low initial tone), and the final boundary tone represented as H] (a high final tone) or L] (a low final tone). In speech context, the citational forms of tones are subject to tone-sandhi process which can be triggered by a combination of various factors, such as the phonological tone sandhi rules, human’s physical limit, the accent-sensitive tonal manifestation, and the influence of semantic-pragmatic context, etc. In the present study, the tone dynamics in Mandarin utterances are investigated by means of locating the domain of tone sandhi. In Mandarin, the canonical timing unit of a tone is the syllable, while its phonetic realization is related to a rhythmic unit, in which syllables with close syntactic-semantic relationship are subject to tonal unification. A “tonal unit” is thus formed based on the prosodic grouping under the lexical-semantic cohesion principle. At the surface phonological level, the pitch contours of tonal units together with the intonational boundary tones compose the surface melodic form of a Mandarin utterance. It is observed from the acoustic data of the the global F0 pattern and final syllable duration of stimuli, that Mandarin declarative utterances present a general declining tendency in pitch and lengthened duration towards the end of each IU.

The study of Mandarin rhythmic feature is conducted under the framework of Control/Compensation rhythmic model. At the phonotactic Level-I, Mandarin presents strong controlling behavior, with a higher stability of vocalic intervals as opposed to consonantal intervals. When speech rate is taken into account in the CCI computation, it is found that in Mandarin there emerges a tendentially linear effect on both the V and C intervals when speed accelerates. With the highest speech rate Chinese speakers still preserve the controlling behavior over the single segment’s articulation. For the study of rhythmic behaviors at the sentential Level-II in Mandarin, a perception experiment was conducted with native subjects identifying the most salient syllables in the utterance stimuli. The results revealed that native subjects have little convergence in the SA judgement, which indicated that SAs might be a fairly elusive feature in Chinese prosody. The controlling rhythmic feature of Mandarin is mostly contributed by the Level-I behavior, while the Level-II plays a weak role. The rhythmic behaviors of Mandarin at the two levels also were compared with another controlling language, Italian, in a series of studies as seen in Bertinetto *et al*. (2012), Bertini *et al*. (2011), and Zhi *et al*. (2010).

For the analysis of Mandarin prosodic form, the study attempts to find a plausible way in which prosodic form can be related with the specified prosodic function through intermediate levels of derivation. The analysis of Mandarin prosody at the underlying phonological level provides an intermediate phase which relates the linguistic function of prosody with the formal representation.

The surface prosodic form of each utterance is annotated into a series of symbols with the INTSINT alphabet. The symbolic coding of each utterance prosodic form is then subject to the Prozed tool for speech synthesis, with the implementation via a plugin to Praat software. The generated stimuli are compared to the original utterances for evaluation in Appendix III. It is shown that a satisfactory synthesis of the original utterances can be obtained from the manually annotated symbols at the surface phonological level. The synthesized data can not only closely capture the pitch feature of the original utterance at the physical level, but also contains all the necessary information for expressing the prosodic function.

The analysis-by-synthesis study of Mandarin speech prosody is conducted by means of a cyclic procedure. First, the complicated raw data (F0 contour) is simplified into a symbolic representation with the INTSINT annotation system, with the symbol sequence derived from the functional analysis; Then such stylization of the raw contour with the INTSINT symbols is synthesized into acoustic data; Furthermore, the synthesized stimuli are compared to the original data to evaluate the modelling proposal. Each step of the analysis-by-synthesis paradigm can be a reversible process, to evaluate the analyses in an objective way, associating the acoustic data with underlying phonological representation. The procedure implements a method in which the representation of prosodic form and representation of prosodic function can be mapped to each other through intermediated phases.

The symbolic coding of 60 Mandarin utterances in the present study was conducted by means of the author’s manual annotation. It is hoped that the findings and relative discussion in this study, by deriving the representation of prosodic form from the analysis of prosodic function, will provide useful insights for future understanding and development of an automatic annotation and modelling of Mandarin speech prosody within the MOMEL-INTSINT system and with a larger amount of speech data.

Bibliography

[1] Anderson, S. R. (1978). Tone features. In V. A. Fromkin (ed.), *Tone: A Linguistic Survey*. New York: Academic Press, 133-175.

[2] Armstrong, L. E. and Ward, I. C. (1926). *A Handbook of English Intonation*. Cambridge: Heifer.

[3] Auran, C. (2004). *Prosodie et anaphore dans le discours en anglais et en français: cohésion et attribution référentielle*. Doctoral dissertation, Université de Provence.

[4] Bao, Z. M. (1990a). Fanqie languages and reduplication. *Linguistic Inquiry*, 21: 317-350.

[5] Bao, Z. M. (1990b). *On the Nature of Tone*. Doctoral dissertation, MIT, Cambridge: Mass.

[6] Bao, Z. M. (2001). The Asymmetry of the medial glides in middle Chinese. *Proceedings of the 7th International and 19th National Conferences on Chinese Phonology*, 11: 7-27.

[7] Bao, Z. M. (2003). Tone, accent and stress in Chinese. *Journal of Linguistics*, 39(1): 147-166.

[8] Beckman, M. E. (1986). *Stress and Non-Stress Accent*. Dordrecht: Foris.

[9] Beckman, M. E. and Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology*, 3: 255-309.

[10] Beckman, M. E. and Ayers, G. M. (1997). *Guidelines for ToBI Labelling*. Department of Linguistics, Ohio State University. Available at http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf

[11] Beckman, M. E., Hirschberg, J. and Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S. A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. New York: Oxford University Press, 8-55.

[12] Bernard, S. A. (2006). The use of authentic materials in the teaching of reading. *The Reading Matrix*, 6(2): 60-69.

[13] Bertinetto, P. M. (1989). Reflections on the Dichotomy 'Stress' vs. 'Syllable-timing'. *Revue de Phonétique Appliquée*, 91-92-93: 99-130.

[14] Bertinetto, P. M. (1999). Boundary strength and linguistic ecology. Mostly exemplified on intervocalic /s/-voicing in Italian. *Folia Linguistic*, 33: 267-286.

[15] Bertinetto, P. M. and Bertini, C. (2008). On modelling the rhythm of natural languages. *Proceedings of 4th International Conference on Speech Prosody*, Campinas, 427-430.

[16] Bertinetto, P. M. and Bertini, C. (2010). Towards a unified predictive model of natural language rhythm. In M. Russo (ed.), *Prosodic Universals. Comparative Studies in Rhythmic Modelling and Rhythm Typology*. Naples: Aracne, 43-77.

[17] Bertinetto, P. M., Bertini, C., Floquet, O. and Giordano, R. 2012. The Control/Compensation model meets Brazilian Portuguese. *The 2012 GSCC Conference*, Belo Horizonte.

[18] Bertinetto, P. M., Bertini, C. and Zhi, N. (2012). Rhythm in Mandarin Chinese and Italian: the role of sentence accents. *Proceedings of 6th International Conference on Speech Prosody*, Shanghai, 520-523.

[19] Bertini, C., Bertinetto, P. M. and Zhi, N. (2011). Chinese and Italian speech rhythm, normalization and the CCI algorithm. *Proceedings of the Interspeech 2011 Conference*, Florence, 1853-1856.

[20] Bertrand, R. D. (1999). *l'hétérogénéité de la parole: analyse énonciative de phénomènes*

prosodiques et kinésiques dans l'interaction interindividuelle. Doctoral dissertation, Université de Provence.

[21] Bigi, B. and Hirst, D. (2012). Speech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. *Proceedings of 6th International Conference on Speech Prosody*, Shanghai, 19-22.

[22] Boersma, P. (1998). *Functional Phonology. Formalizing the Interactions between Articulatory and Perceptual Drives*. Doctoral dissertation, University of Amsterdam.

[23] Boersma, P. and Weenink, D. (1992-2012). *Praat: a system for doing phonetics by computer*. Available at http://www.praat.org.

[24] Bruce, G. and Gårding, E. (1978). A prosodic typology for Swedish dialects. In E. Gårding, G. Bruce and R. Bannert (eds.), *Nordic prosody, Travaux de l'Institut de linguistique de Lund XIII*. Lund: Gleerups, 219-228.

[25] Campione, E. (1998). *MULTEXT Prosodic Database* [CD-ROM]. Paris: European Language Resources Association. Available at http://www.elra.info/

[26] Campione, E. and Veronis, J. (1998). A multilingual prosodic database. *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, 3163-3166.

[27] Chao, Y. R. (1930). A system of "tone-letters". *Fang Yan*, 1980 (2): 81-83.

[28] Chao, Y. R. (1933). Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology*, 4: 121-134.

[29] Chao, Y-R. (1948). *Mandarin Primer*. Cambridge, MA: Harvard University Press.

[30] Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

[31] Chen, M. Y. (2004). *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press.

[32] Cheng, C.C. (1970). Domains of phonological rule application. In J. M. Sadock and A. L. Vanek (eds.), *Studies Presented to Robert B. Lees by His Students*. Edmonton: Linguistic Research, 39-59.

[33] Cheng, C.C. (1973). *A Synchronic Phonology of Mandarin Chinese*. The Hague: Mouton.

[34] Cho, H. (2009). *Propriétés acoustiques de la structure prosodique du coréen*. Doctoral dissertation, Université de Provence.

[35] Cho, H. and Rauzy, S. (2008). Phonetic pitch movements of accentual phrases in Korean read speech. *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil.

[36] Clarke, D. (1990). Communicative theory and its influence on materials production. *Language Teaching*, 25(1): 73-86.

[37] Cruttenden, A. (1997). *Intonation*. (Second Edition). Cambridge: Cambridge University Press.

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

[39] Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

[40] Dasher, R. and Bolinger, D. L. (1982). On preaccentual lengthening. *Journal of the International Phonetic Association*, 12: 58-71.

[41] Dellwo, V. (2010). *Influences of Speech Rate on yhe Acoustic Correlates of Speech Rhythm: an Experimental Phonetic Study based on Acoustic and Perceptual Evidence*. Doctoral dissertation, Universität Bonn.

[42] D'Imperio, M. (2006). Preface. *Italian Journal of Linguistics*, 18: 3-18.

[43] Duanmu, S. (1990). *A Formal Study of Syllable, Tone, Stress, and Domain in Chinese*

Languages. Doctoral dissertation, MIT.

[44] Duanmu, S. (2007). *The Phonology of Standard Chinese*. Second Edition. Oxford: Oxford University Press.

[45] Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual morphosyntactic specifications lexicons and corpora. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, 1535-1538. Available at http://nl.ijs.si/ME/

[46] Estruch, M. (2000). Évaluation de l'algorithme de stylisation mélodique MOMEL et du système de codage symbolique INTSINT avec un corpus de passages en Catalan. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, 19: 45-61.

[47] Fernandez-cruz. R. (2000). *L'analyse phonologique et acoustique du portugais parlé par des communautés noires de l'Amazonie*. Doctoral dissertation, Université de Provence.

[48] Fox, A. (2000). *Prosodic Features and Prosodic Structure. The Phonology of Suprasegmentals*. Oxford: Oxford University Press.

[49] Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In P. F. MacNeilage (ed.), *The Production of Speech*. New York: Springer-Verlag.

[50] Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura, (ed.), *Vocal Fold Physiology. Voice Production, Mechanisms and Function*. Raven Press, New York, Vol. 2: 347-355.

[51] Gårding, E. (1983). A generative model of intonation. In A. Cuder and D. R. Ladd (eds.), *Prosody: Models and Measurements*. Springer-Verlag, 11-25.

[52] Gårding, E. (1984). Chinese and Swedish in a generative model of intonation. In C. C. Elert, I. Johansson and E. Strangert (eds.), *Nordic Prosody III*. Almqvist and Wiksell, Stockholm, 79-91.

[53] Gårding, E. (1987). Speech act and tonal pattern in Standard Chinese: constancy and variation. *Phonetica*, 44: 13-29.

[54] Gårding, E. (1989). Intonation in Swedish. *Working Papers* 35: 63-88. Department of Linguistics, Lund Unversity.

[55] Gårding, E. (1993). On parameters and principles in intonation analysis. *Working papers* 40: 25-47.

[56] Gili Fivela, B. (2008). *Intonation in Production and Perception: The Case of Pisa Italian*. Alessandria: Edizioni dell' Orso.

[57] Giordano, R. (2005). Analisi Prosodica e transrizione intonativa in INTSINT. In F. A. Leoni and R. Giordano (eds.), *Italiano Parlato: Analisi di un Dialogo*. Naples: Liguori Editore, 231-256.

[58] Goldsmith, J. A. (1976). An Overview of Autosegmental Phonology. *Linguistic Analysis*, 2(1): 23-68.

[59] Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology*. Oxford: Blackwell.

[60] Grabe, E. and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, 7: 515-546.

[61] Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

[62] Halliday, M. A. K. (1970). A Course in Spoken English: Intonation. London: Oxford University Press.

[63] 't Hart, J. and Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1: 309-327.

[64] 't Hart, J. and Collier, R. (1975). Integrating different levels of intonation analysis. *Journal*

of Phonetics, 3: 235-255.

[65] ’t Hart, J., Collier, R. and Cohen, A. (1990). *A Perceptual Study of Intonation. An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[66] Herment, S., Loukina, A., Tortel, A., Hirst, D. and Bigi, B. (2012). AixOx, a multi-layered learners corpus: automatic annotation. *Proceedings of the 4th International Conference on Corpus Linguistics*, Ja ѐn, Spain.

[67] Hirst, D. (1977). *Intonative Features. A Syntactic Approach to English Intonation*. Mouton, The Hague.

[68] Hirst, D. (1983). Structures and categories in prosodic representations. In A., Cutler, and D. R. Ladd (eds.), *Prosody: Models and Measurements*. Springer-Verlag, 93-109.

[69] Hirst, D. (1988). Tonal units as phonological constituents: the evidence from French and English intonation. In H. Van der Hulst and N. Smith (eds.), *Autosegmental Studies in Pitch Accent*. Dordrecht: Foris, 151-165.

[70] Hirst, D. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l’Institut de Phontique d’Aix*, 15: 71-85.

[71] Hirst, D. and Di Cristo, A. (1998). *Intonation Systems. A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

[72] Hirst, D., Di Cristo, A. and Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonational systems. In M. Horne (ed.), *Prosody: Theory and Experiment*. Dordrecht: Kluwer Academic Press, 51-88.

[73] Hirst, D. (2001). Automatic analysis of prosody for multi-lingual speech corpora. In E. Keller, G. A. Baill, J. T. Monaghan and M. Huckvale (eds.) *Improvements in Speech Synthesis*. Chichester, England: John Wiley and Sons, 320-328.

[74] Hirst, D. (2004). Lexical and non-lexical tone and prosodic typology. *International Symposium on Tonal Aspects of Languages: with Emphasis on Tone Languages*, Beijing, 81-88.

[75] Hirst, D. and Auran, C. (2005). Analysis by synthesis of speech prosody: the prozed environment. *Proceedings of the 9th Interspeech Conference,* Lisbon, 3225-3228.

[76] Hirst, D. (2005). Form and function in the representation of speech prosody. In K. Hirose, D. Hirst, Y. Sagisaka (eds.), *Quantitative prosody modelling for natural speech description and generation* (=Speech Communication 46 (3-4)), 334-347.

[77] Hirst, D. (2007). A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. *Proceedings of the 16th International Congress on Phonetic Sciences*, Saarbrcken, 1233-1236.

[78] Hirst, D. (2011a). The analysis by synthesis of speech melody: from data to models. *Journal of speech Sciences*, 1(1): 55-83.

[79] Hirst, D. (2011b). *Tutorial in Summer School of Phonology and Phonetics in Tongji University,* Shanghai.

[80] Hirst, D. (2012). Prozed: a speech prosody analysis-by-synthesis tool for linguists. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hongkong, 15-18.

[81] Hogg, R. M. and Mccully, C. B. (1987). *Metrical Phonology: a Course book.* Cambridge: Cambridge University Press.

[82] Hyman, L. M. and Schuh, R. G. (1972). Universals of tone rules. *Working Papers in Language Universals*, 10: 1-50.

[83] Hyman, L. M. and Schuh, R. G. (1974). Universals of tone rules: evidence from West Africa. *Linguistic Inquiry*, 5: 81-115.

[84] Hyman, L. M. (1975). *Phonology: Theory and analysi*s. New York: Holt, Rinehart and

Winston.

[85] Hyman, L. M. (2004). Universals of Tone Rules: 30 Years Later. P*roceedings of the International Conference on Tone and Intonation*, Santorini.

[86] Hyman, L. M. (2009). How (not) to do phonological typology: the case of pitch-accent. *Language Sciences*, 31: 213-238.

[87] Jun, S. A. (2005). *Prosodic Typology: The Phonology of Intonation and Phrasing*. New York: Oxford University Press.

[88] Kenstowicz, M. (1994). *Phonology in Generative Grammar*. Oxford: Blackwell.

[89] Kim, S., Hirst, D., Cho, H., Lee, H. and Chung, M. (2008). Korean MULTEXT: A Korean Prosody Corpus. *Proceedings of Speech Prosody*, Campinas, Brazil.

[90] Kingdon, R. (1958). *The Groundwork of English Intonation*. London: Longman.

[91] Kitazawa, S. (2004). Japanese MULTEXT prosodic corpus database [CD-ROM]. Shizuoka University.

[92] Kochanski, G. and Shih, C. (2001). Automated modelling of Chinese intonation in continuous speech. *The 7th European Conference on Speech Communication and Technology: 2nd Interspeech Event*, Aalborg. Denmark, 911-914.

[93] Kochanski, G. and Shih, C. (2003). Prosody modelling with soft templates. *Speech Communication*, 39: 311-352.

[94] Kochanski, G., Shih, C. and Jing, H. Y. (2003). Hierarchical structure and word strength prediction of Mandarin Prosody. *International Journal of Speech Technology*, 6: 33-43.

[95] Komatsu, M. (2009). Chinese MULTEXT: recordings for a prosodic corpus. *Sophia Linguistica*, 57: 359-369.

[96] Kratochv Ĭ, P. (1968). *The Chinese language today*. London: Hutchinson University Library.

[97] Ladefoged, P. and Johnson, K. (2011). *A Course in Phonetics*. 6th Edition. Wadsworth, Cengage Learning.

[98] Ladd, D. R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

[99] Leben, W. R. (1973). *Suprasegmental Phonology*. Bloomington: Indiana University Linguistics Club.

[100] Levelt, W. J. M. (1989). *Speaking: from Intention to Articulation*. Cambridge, Mass.: MIT Press.

[101] Li, A. J., Yin, Z. G., Wang, M. L., Xu, B. and Zong, C. Q. (2001). A spontaneous conversation corpus CADCC. *Oriental COCOCSDA Workshop*, South Korea.

[102] Li, A. J. (2002). Chinese prosody and prosodic labeling of spontaneous speech. *Proceedings of the 1st International Conference on Speech Prosody*, Aix-en-Provence.

[103] Li, C. N. and Thompson, S. A. (1976). Subject and topic: a new typology of language. In: C. N. Li (ed.), *Subject and Topic*. London /New York: Academic Press, 457-489.

[104] Liberman, M. (1975). *The Intonational System of English*. Dissertation, MIT. Distributed by Indiana University Linguistics Club, Bloomington, 1978.

[105] Liberman, M. and Prince, A. S. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8:249-336.

[106] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*. Dordrecht: Kluwer Academic Press. 413-415.

[107] Louw, J. A. and Barnard, E. (2004). Automatic modelling with INTSINT. P*roceedings of the 15th Annual Symposium of the Pattern Recognition*. Association of South Africa, Grabouw, 107-111.

[108] Maghbouleh, A. (1998). ToBI accent type recognition, *Proceedings of ICSLP*-98, Sydney,

639-642.

[109] Marotta, G. (1988). The Italian diphthongs and the autosegmental framework. In P. M. Bertinetto and M. Loporcaro (eds.), *Certamen Phonologicum: Papers from the 1987 Cortona Phonology Meeting*. Turin: Rosenberg & Sellier, 399–430.

[110] Marotta, G., Calamai, S. and Sardelli, E. (2004). Non di sola lunghezza. La modulazione di f0 come indice sociofonetico. In A. De Dominicis, L. Morie and M. Stefani (eds.), *Costituzione, gestione e restauro di corpora vocali. Atti delle XIV Giornate del GFS*, Roma, Esagrafica, 210-215.

[111] Marotta, G. and Sardelli, E. (2003). Sulla prosodia della domanda con soggetto postverbale in due varietà di italiano toscano. In P. Cosi, E. Magno Caldognetto and A. Zamboni (eds.), *Studi di fonetica in ricordo di F. Ferrero*. Padova: Unipress, 205-212.

[112] Marotta, G. and Sardelli, E. (2007). Prosodic Parameters for the Detection of Regional Varieties in Italian. *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saarbrücken, Germany.

[113] Marotta, G. (2008). Phonology or not phonology? That is the question in intonation. *Proceedings of the International Symposium for the 30th Anniversary of the Phonetics Laboratoriy of UAB*, Bachelona.

[114] Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings ICASSP* 2000, Istanbul, 1281-1284,

[115] Mora Gallardo, E. (1996). *Caractérisation prosodique de la variation dialectale de l'espagnol parlé au Venezuela*. Doctoral dissertation, Université de Provence.

[116] Najim, Z. (1995). *Prosodie de l'arabe standard parlé au Maroc: analyse historique, sociolinguistique et expérimentale*. Doctoral dissertation, Université de Provence.

[117] Nespor, M. and Vogel, I. (1982). Prosodic domains of external sandhi rules. In H. Harry van der and S. Norval (eds.), *The Structure of Phonological Representations*. Dordrecht: Foris, 225-265.

[118] Nespor, M. and Vogel, I. (1983). Prosodic structure above the word. In A. Cutler and D. R. Ladd (eds.), *Prosody: Models and Measurements*. Berlin: Springer-Verlag, 123-140.

[119] Nespor, M. and Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

[120] Nesterenko, I. (2006). *Analyse formelle et implémentation phonétique de l'intonation du parler russe spontané en vue d'une application à la synthèse vocale*. Doctoral dissertation, Université de Provence.

[121] Nicolas, P. (1995). *Contribution de la prosodie à l'amélioration de la parole de synthèse: cas du texte lu en français*. Doctoral dissertation, Université de Provence.

[122] O'Connor, J. D. and Arnold, G. F. (1973). *Intonation of Colloquial English*. Second Edition. London: Longman.

[123] O'Dell, M. L. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey (eds.*), Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1075-1078.

[124] Odden, D. (1986). On the role of the obligatory contour principle in phonological theory. *Language*, 62: 353-383.

[125] Odden, D. (1995). Tone: African languages. In J. A. Goldsmith (ed.), *The handbook of phonological theory*. Cambridge, MA: Blackwell, 444-475.

[126] Ohala, J. J. (1992). The segment: primitive or derived? In G. Docherty and D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge: Cambridge University Press, 166-183.

[127] Palmer, H. E. (1922). *English Intonation with Systematic Exercises*. Cambridge: Heffer.

[128] Palmer, H. E. (1933). *A New Classification of English Tones*. Tokyo: Kaitakusha.

[129] Patel, A. D. (2008). *Music, Language, and the Brain*. New York: Oxford University Press.

[130] Peacock, M. (1997). The Effect of Authentic Materials on the Motivation of EFL Learners. *English Language Teaching Journal*, 51(2): 144-156.

[131] Peng, S. H., Chan, M. K. M., Tseng, C. Y., Huang, T., Lee, O. J. and Beckman, M. E. (2005). Towards a pan-Mandarin system for prosodic transcription. In S. A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. New York: Oxford University Press, 230-270.

[132] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Doctoral dissertation, Cambridge, Mass., MIT.

[133] Pierrehumbert, J. B. and Beckman, M. (1988). *Japanese Tone Structure*. Cambridge: MIT Press.

[134] Pierrehumbert, J. B. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan and M. E. Pollack (eds.), *Intentions in Communication*. Cambridge, MA: MIT Press, 271-311.

[135] Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.

[136] Pike, K. L. (1948). *Tone Languages. A Technique for Determining the Number and Type of Pitch Contrasts in a Lanugage, with Studies in Tonemic Substitution and Fusion.* Ann Arbor: University of Michigan Press.

[137] Pitrelli, J., Beckmann, M. E. and Hirschberg, J. (1994). ToBI (Tones and Break Indices). *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, 18-22.

[138] Portes, C. (2004). *Prosodie et économie du discours: spécificité phonétique, écologie discursive et portée pragmatique du patron d'implication*. Doctoral dissertation, Université de Provence.

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

[140] Prom-on, S., Xu, Y. and Thipakorn, B. (2009). Modelling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 125:405-424.

[141] Prom-on, S., Liu, F. and Xu, Y. (2011). Functional modelling of tone, focus, and sentence type in Mandarin Chinese. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 1638-1641.

[142] Prom-on, S. and Xu, Y. (2012). Pitch target representation of Thai tones. *Proceedings of the 3rd International Symposium on Tonal Aspects of Languages*, Nanjing.

[143] Schack, K. (2000). Comparison of intonation patterns in Mandarin and English for a particular speaker. *University of Rochester Working Papers in the Language Science*s, Vol. Spring (1). Available at: http://www.ling.rochester.edu/wpls/s2000n1/schack.pdf

[144] Schuh, R. G. (1978). Tone rules. In V. A. Fromkin (ed.), *Tone: A linguistic survey*. New York: Academic Press, 221-256.

[145] Selkirk, E. O. (1978). On prosodic structure and its relation to syntactic structure. In T. Fretheim (ed.), *Nordic Prosody II*. Trondheim: TAPIR, 111-140.

[146] Selkirk, E. O. (1980). Prosodic domains in phonology: Sanskrit revisited. In M., Aronoff and M. L. Kean (eds.), *Juncture*. Saratoga, CA: Anma Libri, 107-129.

[147] Selkirk, E. O. (1981). On the nature of phonological representation. In J. Anderson, J. Aver and T. Myers (eds.), *The Cognitive Representation of Speech*, North Holland, Amsterdam.

[148] Selkirk, E. O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press.

[149] Selkirk, E. O. (1986). On derived domains in sentence phonology. *Phonology*, 3: 371-405.

[150] Selkirk, E. O. and Tateishi, K. (1988). Constraints on Minor Phrase formation in Japanese. *CLS* 24: 316-336.

[151] Senior, R. (2005). Authentic Responses to Authentic Materials. *English Teaching Professional*, 38: 71.

[152] Shen, X. N. (1990). *The Prosody of Mandarin Chinese*. Berkeley: University of California Press.

[153] Shih, C. (1986). *The prosodic domain of tone sandhi in Chinese*. Doctoral dissertation, University of California, San Diego.

[154] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. *Proceedings of ICSLP*, Banff, Canada, 867-870.

[155] Tseng, C. Y. (1981). *An Acoustic Phonetic Study on Tones in Mandarin Chinese*. Doctoral dissertation, Brown University: Institute of History and Philology, Academia Sinica.

[156] Van de Weijer, J. and Zhang, J. (2008). An X-bar approach to the syllable structure of Mandarin. *Lingua*, 118: 1416-1428.

[157] Wallace, C. (1992). *Reading* .Oxford: Oxford University Press.

[158] Wan, L. P. (2002). *Alignments of Prenuclear Glides in Mandarin*. Taipei: Crane Publishing Co. LTD.

[159] Wennerstrom, A. (2001). *The Music of Everyday Speech Prosody and Discourse Analysis*. Oxford: Oxford University Press.

[160] Wightman, C. W. and Campbell, W. N. (1994). Improved Labeling of Prosodic Structure. *IEEE Transactions on Speech and Audio Processing*, 2(4): 469-481.

[161] Wilder, C. N. (1981). Chest wall preparation for phonation in female speakers. In D. M. Bless and J. H. Abbs (eds.), *Vocal Fold Physiology: Contemporary Research and Clinical Issues*. San Diego, CA: College-Hill Press, 109-123.

[162] Winkworth, A. L., Davis, P. J., Adams, R. D., and Ellis, E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research*, 38: 124-144.

[163] Wu, Z. J. (2004). From traditional Chinese phonology to modern speech processing – realization of tone and intonation in Standard Chinese. In G. Fant, H. Fujisaki, J. Cao and Y. Xu (eds.), *From Traditional Chinese Phonology to Modern Speech Processing*. Beijing: Foreign Language Teaching and Research Press.

[164] Xu, Y. (1993). *Contextual tonal variation in Mandarin Chinese*. Doctoral dissertation, the University of Connecticut.

[165] Xu Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, 95: 2240-2253.

[166] Xu Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25: 61-83.

[167] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, 27(1): 55-105.

[168] Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46(3-4): 220-251.

[169] Xu, Y. (2007). Speech as articulatorily encoded communicative functions. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrucken, Germany, 25-30.

[170] Xu, Y. (2009). Timing and coordination in tone and intonation - an articulatory-functional perspective. *Lingua*, 119(6): 906-927.

[171] Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, 38(3): 329-336.

[172] Xu, Y. (2011). Functions and mechanisms in linguistic research - lessons from speech prosody. *Proceedings of the 4th ISCA Tutorial and Research Workshop on Experimental*

Linguistics, Paris, France.

[173] Xu, Y. and Prom-on, S. (2012). Analysis and synthesis of speech prosody based on articulatory dynamics and communicative functions: from concept to practice. In Q. W. Ma, H. W. Ding and D. Hirst (eds.), *Abstract Book of Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China, 6.

[174] Xu, Y. and Wang, Q. E. (1997). What can toen studies tell us about intonation? *Proceedings of ESCA Workshop on Intonation*, Athens, Greece, 337-440.

[175] Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: evidence from Mandarin Chinese. *Speech Communication*, 33(4): 319-337.

[176] Xu, Y. and Wang, M. (2005). Tonal and duratonal variations as phonetic coding for syllable grouping. *Journal of the Acoustical Society of America*, 117: Pt. 2, 2573.

[177] Yip, M. (1980). The Tonal Phonology of Chinese. Doctoral dissertation, MIT.

[178] Yip, M. (1989). Contour tones. *Phonology*, 6: 149-174.

[179] Yip M. (2002). *Tone*. Cambridge: Cambridge University Press.

[180] Yip, M. (2003). Casting doubt on the onset-rime distinction. *Lingua*, 113: 779-816.

[181] Zhi, N., Hirst, D. and Bertinetto, P. M. (2010). Automatic analysis of the intonation of a tone language. Applying the momel algorithm to spontaneous standard Chinese. *Proceedings of the 11th Interspeech Conference*, Makuhari, Japan.

[182] Zhi, N., Bertinetto, P. M. and Bertini, C. (2011). The speech rhythm of Beijing Chinese, in the framework of CCI. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hongkong, 2316-2319.

[183] Zhu, X. N. (2005). *An Experimental Study in Shanghai Tones*. Shanghai Educational Publishing House.

Appendix

Appendix I Perception Experiment

a． Stimuli

Three lists of sentence stimuli, List A, List B and List C employed in the perception experiment were presented as follows, with each list consisting of 15 spontaneous and 15 read utterances selected from two different speech corpora.

List A

(1)

ran hou ne ta zhao wo gan ma ne - Romanization (Chinese Pinyin spelling system)

然后呢她找我干嘛呢 - Characters

"Then why did she look for me?" - Translation

(2)

wo shuo nei ge mi huang se ni ken ding neng chuan

我说那个米黄色你肯定能穿

"I say that that you can wear that (clothes) of beige color for sure."

(3)

fei shour chang duanr nar dou he shi

肥瘦儿长短哪儿都合适

"It fits (you) well in the size."

(4)

ta gen ta nei ge zhi zi yi kuair hui qu

他跟他那个侄子一块儿回去

“He and his niece go back together.”

(5)

gao zhong dao da xue tong yang ye shi yi ge zhuan bian

高中到大学同样也是一个转变

“It is as well a kind of transition from high school to university.”

(6)

zai lao lao jia zai nai nai jia dou shi xiang bo bo

在姥姥家在奶奶家都是香饽饽

“(The grandchild) is popular at both houses of two grandmothers”

(7)

na jin nian chun jie hai gao huo dong ma

那今年春节还搞活动吗

“Is there then any activity organized for this Spring Festival?”

(8)

an yi qian zhei biao zhun tian tian guo nian

按以前这标准天天过年

“According to the living criteria of the past years, we are in festivals every day.”

(9)

shi jian chang er qie di wen bi jiao di

时间长而且低温比较低

"The duration (of this winter) is long, and the temperature is comparatively lower."

(10)

ta zhu chi de shi jian jiu gou chang de le

他主持的时间就够长的了

"He had hosted the program for quite a long time."

(11)

yi shi tou lanr mei you kai deng

一时偷懒儿没有开灯

"I did not turn on the light due to a moment of laziness"

(12)

wo zai Sanya guo de fei chang yu kuai

我在三亚过得非常愉快

"I quite enjoy myself in SanYa*."*

(13)

zhei li tian qi hen re hen qing lang

这里天气很热很晴朗

"The weather here is quite nice and warm."

(14)

ming tian bi xu na dao wo de xing li

明天必须拿到我的行李

“I must get my luggage by tomorrow.”

(15)

jin van gong fa sheng liu qi an jian

今晚共发生六起案件

“There occurred six criminal cases tonight.”

(16)

wo nüer da suan ming tian kao dan gao

我女儿打算明天烤蛋糕

“My daughter plans to bake a cake tomorrow.”

(17)

wo fu qin rang ta dai shang gou

我父亲让她带上狗

“My father asked her to take the dog along with her.”

(18)

qi ta chi cun si zhong yan se dou you huo

其他尺寸四种颜色都有货

“(Clothes) of other sizes are in store with four colours.”

(19)

bei jing de zu qiu dai biao dui shi *Guo-An*

北京的足球代表队是国安

“The representative of Beijing football team is ***Guo-An*** *team.”*

(20)

feng da de ji hu yao ba wo chui xia shan qu

风大得几乎要把我吹下山去

“The wind was so heavy that it almost blew me off the mountain.”

List B

(21)

fan zheng *Huan Le Zong Dong Yuan* wo dao shi kan

反正《欢乐总动员》我倒是看

*“I watch the **Huan-Le-Zong-Dong-Yuan** program anyway.”*

(22)

zhu yao kan nei ge *mo fang xiu*

主要看那个《模仿秀》

*“I watch mostly the **Imitation Show**.”*

(23)

zhe hui huan le zhu chi ren le ha

这回换了主持人了哈

“The host has been changed this time, hasn’t it?”

(24)

yuan lai zhe yuanr li mei zhuang ji ge suo

原来这院儿里没装几个所

“There had not been many institutes situated in this ground in the past.”

(25)

suo yi zui hou nong yi liang tour bu zhen

所以最后弄一两头儿不真

“Therefore, in the end none of the two sides seem satisfied.”

(26)

jin nian da suan shang *Miao-Hui* guang guang qu ma

今年打算上庙会逛逛去吗

“Do you plan to have a visit of the ***Miao-Hui*** *this year?”*

(27)

you ti gao banr you ji chu banr liang ge ban

有提高班儿有基础班儿两个班

“There are two kinds of tutorial classes, at intermediate level and elementary level.”

(28)

xian zai shi zen me ge gai nian

现在是怎么个概念

“Now it is such a conception.”

(29)

ta men de cang shu liang xiang dang de da

他们的藏书量相当的大

“They have quite a large store of books.”

(30)

dan shi yi fen lei yi hou jiu shang le jia le

但是一分类以后就上了架了

“However, after the (books) have been categorized, they shall be arranged on shelves.”

(31)

ma fan nin bang wo lian xi yi xiar

麻烦您帮我联系一下儿

“Please help me to get contact with (them).”

(32)

wo nü er da suan ming tian kao dan gao

我女儿打算明天烤蛋糕

“My daughter plans to bake a cake tomorrow.”

(Note: In the read-speech corpus, same texts were read by different speakers. the present sentence shares identical text with sentence (16), but they were produced by two different speakers.)

(33)

ta zhi yi yao yong shou lai jiao ban

她执意要用手来搅拌

“She insisted to mix (it) by hands.”

(34)

ta de hai zi men zao jiu deng zhe chi le

她的孩子们早就等着吃了

“Her children have long been waiting for eating (the cake).”

(35)

wo fu qin rang ta dai shang gou

我父亲让她带上狗

“My father asked her to take the dog along with her.”

(Note: the sentence (35) and sentence (17) were indentical in text, but produced by different speakers in the read corpus.)

(36)

yuan yin zhi yi shi zuo wei bu shu fu

原因之一是座位不舒服

“One of the reasons is that the seat is uncomfortable.”

(37)

geng zao de shi you ren da han

更糟的是有人打鼾

“What is worse is that someone snores.”

(38)

er qie zhe li de hai jian zhi mei ji le

而且这里的海简直美极了

“Moreover the sea here is fabulous.”

(39)

er wo de xing li que qu le luo ma

而我的行李却去了罗马

“However, my luggage has arrived in Rome.”

(40)

wo hai xu yao yi xie ying ji yao

我还需要一些应急药

“I still need some medicine for emergency.”

List C

(41)

fei shour ma ye jue dui bu shou

肥瘦吗也绝对不瘦

“It is not small at all in consideration of the size.”

(42)

fang jia le gei song dao lao lao jia qu le

放假了给送到姥姥家去了

“The holiday is on, and (the grandchild) has been sent to her grandmother’s.”

(43)

dou fan ying ta nei ge yan de bu cuo

都反映他那个演得不错

“It is reflected that he had performed well in that (film).”

(44)

xian zai kai ge zhong ge yang de banr

现在开各种各样的班儿

“Now there are all kinds of classes.”

(45)

jiu deng yu shi gei ta tu ran jiu huan le

就等于是给她突然就换了

“It can be said that she has been changed for a sudden.”

(46)

zhao mu qian kan ying gai yi bai nian mei wen ti

照目前看一百年没问题

“It can be seen for now that there would be no problem in a hundred of years.”

(47)

mei tian ni shuo nei zhi piao dei duo shao

每天你说那个支票得多少

“According to you how many checks would be there every day?”

(48)

nei xie pian zi xian zai hai neng kan

那些片子现在还能看

“Those CD-ROMs can still be watched.”

(49)

que shi jiu you jing ji li yi

确实就有经济利益

“There does have economic benefit.”

(50)

jiu shuo zhei ge fang xing liang bi jiao shao

就说那个发行量比较少

“It is said that the publishing amount is quite small.”

(51)

ta de jiao sheng ke yi bao hu ta

它的叫声可以保护她

“The bark of (dog) may protect her.”

(52)

na lao yu fu zhang de fei chang gao da

那老渔夫长得非常高大

“That fisherman is quite big.”

(53)

ta jiu hao xiang bian le yi ge ren shi de

他就好像变了一个人似的

“He seems to change to another person.”

(54)

er qie zhei li de hai jian zhi mei ji le

而且这里的海简直美极了

“Moreover the sea here is fabulous.”

(Note: identical text with sentence (39), but different in productions with respect to articulators.)

(55)

er wo de xing li que qu le luo ma

而我的行李却去了罗马

“However, my luggage has arrived in Rome.”

（Note: identical text with sentence (40), but different productions with respect to articulators.）

(56)

hai suan shi yi chang jing cai de bi sai

还算是一场精彩的比赛

“It can be said that it was a wonderful game.”

(57)

wo yao ding shi he sheng dan cui bing

我要订十盒圣诞脆饼

“I would like to order ten boxes of Christmas crispy cakes.”

(58)

qing bang wo jie fu wu zhong xin

请帮我接服务中心

“Please help me to connect to the service center.”

(59)

wo shi shi san hao song qu de

我是十三号送去的

“I had delivered it on the 13th.”

(60)

wo men yao wei yi ge da xing hui yi ding can

我们要为一个大型会议订餐

“We need to order food for a big conference.”

b. Subjects participated in the perception experiment

	Name	**Gender**	**Age**	**Time**	**Major**	**Playing**

				Abroad		**Musical Instruments**
Group I	A	Male	23	6 months	Literature	No
	B	Female	26	10 months	Literature	No
	C	Female	32	3 years	Economics	No
	D	Female	27	13 months	Philosophy	No
	E	Male	26	1 month	Music	Yes
Group II	F	Female	28	4 years	History	Yes
	G	Male	26	14 months	History	No
	H	Male	29	4 months	History	No
	I	Male	25	5 years	History	No
	J	Female	30	3 years	Philosophy	Yes
Group III	K	Male	29	2 years	Philosophy	No
	L	Female	29	7 years	History	No
	M	Female	23	2 weeks	Literature	Yes
	N	Female	25	6 months	Literature	No
	O	Female	30	1 year	Philosophy	No

c. ***Sentence accents of each utterance according to the result from perception experiment, as represented by the following metrical grids***

List A

(1)

X$_{(70\%)}$				X$_{(90\%)}$	X$_{(70\%)}$	X$_{(70\%)}$			IU level
X	X		X	X	X	X	X		word level
X	X	X	X	X	X	X	X	X	syllable level
ran	hou	ne	ta	zhao	wo	gan	ma	ne	

(2)

				X$_{(80\%)}$							X$_{(80\%)}$
X	X	X		X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X	X
wo	shuo	nei	ge	mi	huang	se	ni	ken	ding	neng	chuan

(3)

X$_{(80\%)}$		X$_{(80\%)}$		X$_{(90\%)}$			
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
fei	shour	chang	duar	nar	dou	he	shi

(4)

X$_{(90\%)}$					X$_{(90\%)}$			X$_{(90\%)}$		
X	X	X	X		X		X	X	X	
X	X	X	X	X	X	X	X	X	X	X
ta	gen	ta	nei	ge	zhi	zi	yi	kuar	hui	qu

(5)

X$_{(100\%)}$			X$_{(80\%)}$		X$_{(90\%)}$							
X	X	X	X	X	X	X	X	X	X		X	X
X	X	X	X	X	X	X	X	X	X	X	X	X
gao	zhong	dao	da	xue	tong	yang	ye	shi	yi	ge	zhuan	bian

(6)

X$_{(80\%)}$				X$_{(80\%)}$				X$_{(100\%)}$				
X	X		X	X	X		X	X	X	X	X	
X	X	X	X	X	X	X	X	X	X	X	X	X
zai	lao	lao	jia	zai	nai	nai	jia	dou	shi	xiang	bo	bo

(7)

			X$_{(80\%)}$			X$_{(90\%)}$			
	X	X	X	X		X	X	X	
X	X	X	X	X	X	X	X	X	X
na	jin	nian	chun	jie	hai	gao	huo	dong	ma

(8)

X$_{(90\%)}$ X$_{(100\%)}$

X	X	X	X	X	X	X		X	X
X	X	X	X	X	X	X	X	X	X
an	yi	qian	zhei	biao	zhun	tian	tian	guo	nian

(9)

		$X_{(100\%)}$			$X_{(80\%)}$				$X_{(90\%)}$
X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
shi	jian	chang	er	qie	di	wen	bi	jiao	di

(10)

$X_{(100\%)}$							$X_{(90\%)}$	$X_{(80\%)}$		
X	X	X		X	X		X	X		
X	X	X	X	X	X	X	X	X	X	X
ta	zhu	chi	de	shi	jian	jiu	gou	chang	de	le

(11)

		$X_{(90\%)}$				$X_{(90\%)}$	$X_{(80\%)}$
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
yi	shi	tou	lanr	mei	you	kai	deng

(12)

		$X_{(100\%)}$				$X_{(100\%)}$	$X_{(70\%)}$		
X		X	X	X		X	X	X	X
X	X	X	X	X	X	X	X	X	X
wo	zai	san	ya	guo	de	fei	chang	yu	kuai

(13)

				$X_{(90\%)}$			$X_{(70\%)}$	
X		X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X

zhei	li	tian	qi	hen	re	hen	qing	lang

(14)

$X_{(70\%)}$		$X_{(100\%)}$	$X_{(80\%)}$						
X	X	X	X	X		X		X	
X	X	X	X	X	X	X	X	X	X
ming	tian	bi	xu	na	dao	wo	de	xing	li

(15)

$X_{(90\%)}$					$X_{(80\%)}$			
X	X	X	X	X	X		X	X
X	X	X	X	X	X	X	X	X
jin	wan	gong	fa	sheng	liu	qi	an	jian

(16)

$X_{(90\%)}$	$X_{(90\%)}$				$X_{(90\%)}$	$X_{(70\%)}$		$X_{(70\%)}$	
X	X		X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
wo	nü	er	da	suan	ming	tian	kao	dan	gao

(17)

$X_{(80\%)}$	$X_{(100\%)}$				$X_{(90\%)}$		
X	X		X	X	X		X
X	X	X	X	X	X	X	X
wo	fu	qin	rang	ta	dai	shang	gou

(18)

				$X_{(80\%)}$				$X_{(100\%)}$		
X	X	X	X	X		X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X
qi	ta	chi	cun	si	zhong	yan	se	dou	you	huo

(19)

$X_{(70\%)}$			$X_{(70\%)}$						$X_{(90\%)}$	$X_{(70\%)}$
X	X		X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X
bei	jing	de	zu	qiu	dai	biao	dui	shi	guo	an

(20)

$X_{(80\%)}$	$X_{(90\%)}$		$X_{(70\%)}$			$X_{(80\%)}$		$X_{(80\%)}$		$X_{(70\%)}$	
X	X		X	X		X	X	X		X	
X	X	X	X	X	X	X	X	X	X	X	X
feng	da	de	ji	hu	yao	ba	wo	chui	xia	shan	qu

List B

(21)

		$X_{(100\%)}$				$X_{(90\%)}$		$X_{(80\%)}$		$X_{(90\%)}$
X	X	X	X	X	X	X	X	X		X
X	X	X	X	X	X	X	X	X	X	X
fan	zheng	huan	le	zong	dong	yuan	wo	dao	shi	kan

(22)

		$X_{(80\%)}$			$X_{(70\%)}$		$X_{(100\%)}$
X	X	X	X		X	X	X
X	X	X	X	X	X	X	X
zhu	yao	kan	nei	ge	mo	fang	xiu

(23)

		$X_{(100\%)}$				$X_{(70\%)}$		
X		X		X	X	X		
X	X	X	X	X	X	X	X	X
zhe	hui	huan	le	zhu	chi	ren	le	ha

(24)

			$X_{(80\%)}$		$X_{(70\%)}$	$X_{(90\%)}$			
X	X	X	X		X	X	X		X
X	X	X	X	X	X	X	X	X	X
yuan	lai	zhe	yuanr	li	mei	zhuang	ji	ge	suo

(25)

		$X_{(90\%)}$							$X_{(80\%)}$
X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
suo	yi	zui	hou	nong	yi	liang	tour	bu	zhen

(26)

		$X_{(90\%)}$			$X_{(90\%)}$	$X_{(90\%)}$				
X	X	X	X	X	X	X	X			
X	X	X	X	X	X	X	X	X	X	X
jin	nian	da	suan	shang	miao	hui	guang	guang	qu	ma

(27)

	$X_{(90\%)}$				$X_{(80\%)}$			$X_{(70\%)}$		
X	X	X	X	X	X	X	X	X		X
X	X	X	X	X	X	X	X	X	X	X
you	ti	gao	banr	you	ji	chu	banr	liang	ge	ban

(28)

$X_{(90\%)}$	$X_{(70\%)}$		$X_{(80\%)}$			$X_{(90\%)}$	
X	X	X	X			X	X
X	X	X	X	X	X	X	X
xian	zai	shi	zen	me	ge	gai	nian

(29)

$X_{(70\%)}$					$X_{(90\%)}$	$X_{(70\%)}$	$X_{(70\%)}$
X	X	X	X	X	X	X	X

X	X	X	X	X	X	X	X	X	X
ta	men	de	cang	shu	liang	xiang	dang	de	da

(30)

$X_{(90\%)}$								$X_{(90\%)}$		$X_{(80\%)}$	
X	X	X	X	X	X	X	X	X		X	
X	X	X	X	X	X	X	X	X	X	X	X
dan	shi	yi	fen	lei	yi	hou	jiu	shang	le	jia	le

(31)

$X_{(70\%)}$		$X_{(80\%)}$	$X_{(80\%)}$		$X_{(70\%)}$			
X		X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X
ma	fan	nin	bang	wo	lian	xi	yi	xiar

(32)

$X_{(80\%)}$	$X_{(90\%)}$				$X_{(80\%)}$				$X_{(90\%)}$
X	X		X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
wo	nü	er	da	suan	ming	tian	kao	dan	gao

(33)

	$X_{(90\%)}$				$X_{(100\%)}$			
X	X	X		X	X		X	X
X	X	X	X	X	X	X	X	X
ta	zhi	yi	yao	yong	shou	lai	jiao	ban

(34)

$X_{(90\%)}$		$X_{(80\%)}$			$X_{(80\%)}$		$X_{(80\%)}$		$X_{(90\%)}$	
X		X			X		X		X	
X	X	X	X	X	X	X	X	X	X	X
ta	de	hai	zi	men	zao	jiu	deng	zhe	chi	le

(35)

	X$_{(90\%)}$				X$_{(80\%)}$		X$_{(90\%)}$
X	X		X	X	X		X
X	X	X	X	X	X	X	X
wo	fu	qin	rang	ta	dai	shang	gou

(36)

X$_{(80\%)}$					X$_{(90\%)}$			X$_{(80\%)}$	
X	X	X	X	X	X	X	X	X	
X	X	X	X	X	X	X	X	X	X
yuan	yin	zhi	yi	shi	zuo	wei	bu	shu	fu

(37)

X$_{(80\%)}$	X$_{(70\%)}$						X$_{(90\%)}$
X	X		X	X	X	X	X
X	X	X	X	X	X	X	X
geng	zao	de	shi	you	ren	da	han

(38)

		X$_{(80\%)}$				X$_{(90\%)}$			X($_{90\%)}$	
X	X	X			X	X	X	X	X	
X	X	X	X	X	X	X	X	X	X	X
er	qie	zhe	li	de	hai	jian	zhi	mei	ji	le

(39)

X$_{(80\%)}$			X$_{(90\%)}$					X$_{(90\%)}$	X$_{(80\%)}$
X	X		X		X	X		X	X
X	X	X	X	X	X	X	X	X	X
er	wo	de	xing	li	que	qu	le	luo	ma

(40)

X$_{(70\%)}$ X$_{(90\%)}$ X$_{(90\%)}$ X$_{(70\%)}$

X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X
wo	hai	xu	yao	yi	xie	ying	ji	yao

List C

(41)

$X_{(90\%)}$				$X_{(80\%)}$			$X_{(80\%)}$
X	X			X	X	X	X
X	X	X	X	X	X	X	X
fei	shour	ma	ye	jue	dui	bu	shou

(42)

	$X_{(90\%)}$			$X_{(80\%)}$		$X_{(70\%)}$		$X_{(90\%)}$		
X	X			X		X		X		
X	X	X	X	X	X	X	X	X	X	X
fang	jia	le	gei	song	dao	lao	lao	jia	qu	le

(43)

$X_{(90\%)}$			$X_{(70\%)}$						$X_{(80\%)}$
X	X	X	X	X		X		X	X
X	X	X	X	X	X	X	X	X	X
dou	fan	ying	ta	nei	ge	yan	de	bu	cuo

(44)

		$X_{(100\%)}$				$X_{(70\%)}$		$X_{(100\%)}$
X	X	X	X	X	X	X		X
X	X	X	X	X	X	X	X	X
xian	zai	kai	ge	zhong	ge	yang	de	ban

(45)

				$X_{(90\%)}$	$X_{(70\%)}$	$X_{(90\%)}$
X	X	X	X	X	X	X

X	X	X	X	X	X	X	X	X	X	X
jiu	deng	yu	shi	gei	ta	tu	ran	jiu	huan	le

(46)

	X$_{(80\%)}$					X$_{(80\%)}$			X$_{(70\%)}$		
	X	X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X	X
zhao	mu	qian	kan	ying	gai	yi	bai	nian	mei	wen	ti

(47)

	X$_{(70\%)}$				X$_{(80\%)}$	X$_{(80\%)}$		X$_{(90\%)}$	
X	X	X	X	X	X	X	X	X	
X	X	X	X	X	X	X	X	X	X
mei	tian	ni	shuo	nei	zhi	piao	dei	duo	shao

(48)

X$_{(70\%)}$		X$_{(80\%)}$		X$_{(90\%)}$				X$_{(90\%)}$
X	X	X		X	X	X	X	X
X	X	X	X	X	X	X	X	X
nei	xie	pian	zi	xian	zai	hai	neng	kan

(49)

X$_{(100\%)}$		X$_{(80\%)}$		X$_{(70\%)}$		X$_{(80\%)}$	
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
que	shi	jiu	you	jing	ji	li	yi

(50)

X$_{(80\%)}$		X$_{(100\%)}$							X$_{(70\%)}$
X	X	X		X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
jiu	shuo	zhei	ge	fang	xing	liang	bi	jiao	shao

(51)

		$X_{(100\%)}$				$X_{(80\%)}$		
X		X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X
ta	de	jiao	sheng	ke	yi	bao	hu	ta

(52)

	$X_{(70\%)}$					$X_{(90\%)}$	$X_{(80\%)}$	$X_{(70\%)}$	$X_{(70\%)}$
X	X	X	X	X		X	X	X	X
X	X	X	X	X	X	X	X	X	X
na	lao	yu	fu	zhang	de	fei	chang	gao	da

(53)

				$X_{(90\%)}$					$X_{(70\%)}$	
X		X	X	X		X		X	X	
X	X	X	X	X	X	X	X	X	X	X
ta	jiu	hao	xiang	bian	le	yi	ge	ren	shi	de

(54)

		$X_{(80\%)}$			$X_{(70\%)}$				$X_{(80\%)}$	
X	X	X			X	X	X	X	X	
X	X	X	X	X	X	X	X	X	X	X
er	qie	zhei	li	de	hai	jian	zhi	mei	ji	le

(55)

			$X_{(100\%)}$					$X_{(80\%)}$	$X_{(70\%)}$
X	X		X		X	X		X	X
X	X	X	X	X	X	X	X	X	X
er	wo	de	xing	li	que	qu	le	luo	ma

(56)

$X_{(90\%)}$ $X_{(90\%)}$ $X_{(70\%)}$

X	X	X	X		X	X		X	X
X	X	X	X	X	X	X	X	X	X
hai	suan	shi	yi	chang	jing	cai	de	bi	sai

(57)

		$X_{(70\%)}$	$X_{(80\%)}$		$X_{(80\%)}$		$X_{(80\%)}$	
X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X
wo	yao	ding	shi	he	sheng	dan	cui	bing

(58)

	$X_{(100\%)}$		$X_{(70\%)}$				$X_{(80\%)}$
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
qing	bang	wo	jie	fu	wu	zhong	xin

(59)

			$X_{(90\%)}$				
X	X	X	X	X	X		
X	X	X	X	X	X	X	X
wo	shi	shi	san	hao	song	qu	de

(60)

		$X_{(70\%)}$				$X_{(100\%)}$				$X_{(80\%)}$	$X_{(90\%)}$
X	X	X		X		X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X	X
wo	men	yao	wei	yi	ge	da	xing	hui	yi	ding	can

Appendix II Details of each syllable in the 60 Mandarin utternaces

Utterance	Syllable	Tone	Accented	Sandhi	Duration (s)	Normalized duration	Max F0	Min F0	Mean F0	Normalized F0
(1)	ran	T2	a	no	0.137	0.899	204	161.5	175.6	-0.492
	hou	T4	un	no	0.086	0.567	218.5	204	214.8	1.050
	ne	T0	un	no	0.137	0.899	215.5	188.3	200.7	0.530
	ta	T1	un	no	0.104	0.684	227.9	221.2	222.2	1.309
	zhao	T3	a	no	0.225	1.476	221.6	179.5	202.5	0.599
	wo	T3	a	no	0.139	0.910	212.1	187.7	197.1	0.392
	gan	T4	a	yes	0.148	0.972	175.6	162.4	171.3	-0.682
	ma	T2	un	no	0.207	1.357	174.3	147.8	154.7	-1.462
	ne	T0	un	no	0.188	1.236	166.4	152.4	159.2	-1.243
(2)	wo	T3	un	no	0.052	0.317	149.7	148.3	149	-0.153
	shuo	T1	un	yes	0.126	0.773	187.3	162.9	176.2	1.111
	nei	T4	un	no	0.129	0.792	198.4	170.7	188.8	1.632
	ge	T4	un	no	0.120	0.738	165.2	149.6	157.2	0.251
	mi	T3	a	no	0.238	1.458	151.2	132.2	143.6	-0.432
	huang	T2	un	no	0.214	1.308	179.1	128.6	146.5	-0.281
	se	T4	un	no	0.217	1.331	222.3	151.1	180.7	1.301
	ni	T3	un	yes	0.104	0.636	153.9	148.6	151.1	-0.048
	ken	T3	un	no	0.113	0.692	119.6	117.8	118.2	-1.900
	ding	T4	un	yes	0.144	0.882	140.9	126.6	136.4	-0.820
	neng	T2	un	no	0.200	1.224	146.6	129	135.7	-0.858
	chuan	T1	a	no	0.302	1.852	178.5	145.9	156.1	0.198
(3)	fei	T2	a	no	0.306	1.223	324.6	195.5	247.8	0.788
	shour	T4	un	no	0.269	1.077	357.3	180.2	247	0.763
	chang	T2	a	no	0.310	1.239	275.1	172.7	224.6	0.023
	duanr	T3	un	no	0.178	0.711	259.6	237.7	249.2	0.832
	nar	T3	a	no	0.287	1.147	253.6	196.4	232.6	0.295
	dou	T1	un	no	0.154	0.617	263.8	183.1	233.4	0.322
	he	T2	un	no	0.224	0.895	191.3	167.1	174.8	-1.929
	shi	T4	un	no	0.273	1.092	218	174.3	194.6	-1.093
(4)	ta	T1	a	no	0.202	1.013	280.8	270.5	273.2	1.111
	gen	T1	un	no	0.167	0.838	276.7	273.7	275.7	1.147
	ta	T1	un	no	0.138	0.692	293.9	272.7	281.7	1.233
	nei	T4	un	no	0.112	0.561	272.7	209.2	248.4	0.732
	ge	T4	un	no	0.116	0.580	190.8	177.2	184.2	-0.458
	zhi	T2	a	no	0.329	1.654	214.1	167.8	182	-0.506
	zi	T0	un	no	0.184	0.926	240.4	183.7	215	0.158
	yi	T2	un	no	0.102	0.512	206.6	176.7	186.5	-0.409
	kuair	T4	a	no	0.316	1.587	265.6	169.3	209.9	0.062
	hui	T2	un	no	0.300	1.507	158.8	130	143.7	-1.446
	qu	T4	un	no	0.225	1.130	157.8	117.2	137.4	-1.625
(5)	gao	T1	a	no	0.172	1.059	211.2	194	198.9	1.441

	zhong	T1	un	no	0.163	1.000	199.4	188.8	192.3	1.299
	dao	T4	un	no	0.158	0.972	189.8	156.5	166.6	0.694
	da	T4	a	no	0.185	1.136	144	129.3	135.1	-0.190
	xue	T2	un	no	0.135	0.827	144.1	130.7	136.4	-0.149
	tong	T2	a	no	0.267	1.640	175.4	127.2	141.4	0.003
	yang	T4	un	no	0.206	1.268	185.2	173.6	181.2	1.048
	ye	T3	un	yes	0.084	0.513	voiceless			
	shi	T4	un	no	0.120	0.738	110.1	100.2	106.2	-1.205
	yi	T2	un	no	0.104	0.642	115.8	92.3	103.1	-1.330
	ge	T4	un	no	0.097	0.593	128.9	115.3	122.3	-0.609
	zhuan	T3	un	yes	0.175	1.075	voiceless			
	bian	T4	un	no	0.250	1.540	115.7	106.3	111.4	-1.003
(6)	zai	T4	a	no	0.157	0.919	202.1	165.9	180.5	1.421
	lao	T3	un	no	0.167	0.978	107.3	94.8	101.7	-0.483
	lao	T0	un	no	0.141	0.827	79.8	73.9	76.8	-1.416
	jia	T1	un	no	0.202	1.183	157.3	150.6	152.8	0.868
	zai	T4	a	no	0.147	0.861	173.9	140.9	156	0.937
	nai	T3	un	no	0.178	1.045	140.9	72	103.2	-0.435
	nai	T0	un	no	0.111	0.648	83.2	69.9	75.4	-1.477
	jia	T1	un	no	0.174	1.022	150	140.8	143.2	0.653
	dou	T1	a	no	0.189	1.106	152	138.4	147	0.740
	shi	T4	un	no	0.067	0.391	138.4	124.2	129.7	0.324
	xiang	T1	un	no	0.286	1.677	139.1	120.3	126.1	0.231
	bo	T1	un	no	0.259	1.516	135.1	117.5	126.3	0.236
	bo	T0	un	no	0.141	0.827	72.7	72.7	72.7	-1.598
(7)	na	T4	un	yes	0.112	0.654	277.8	245.7	250.9	0.395
	jin	T1	un	no	0.187	1.095	268.2	252.3	262.6	0.576
	nian	T2	un	no	0.195	1.143	252.3	232.6	239.9	0.216
	chun	T1	a	no	0.183	1.074	264.8	232.6	247.5	0.340
	jie	T2	un	no	0.151	0.885	227.1	215.3	219.7	-0.134
	hai	T2	un	no	0.124	0.724	226.5	219.3	223.5	-0.066
	gao	T3	a	no	0.186	1.088	388.5	365.1	378.2	2.028
	huo	T2	un	yes	0.221	1.298	162	147.7	152.2	-1.595
	dong	T4	un	no	0.207	1.213	199.2	183	193.7	-0.635
	ma	T0	un	no	0.141	0.825	183	164.9	171.2	-1.126
(8)	an	T4	a	no	0.153	0.876	283.2	193.1	237.1	1.188
	yi	T3	un	no	0.095	0.542	193.1	162.2	173.9	-0.456
	qian	T2	un	no	0.190	1.089	229	174.8	189.8	0.008
	zhei	T4	un	no	0.114	0.651	232.7	188.9	209.2	0.524
	biao	T1	a	no	0.185	1.060	219.3	210.3	217.3	0.725
	zhun	T3	un	no	0.149	0.853	216.5	168.4	193.4	0.108
	tian	T1	un	no	0.175	1.003	209	205.6	207.2	0.473
	tian	T1	un	no	0.197	1.127	221.9	202.8	213.1	0.622
	guo	T4	un	no	0.193	1.108	186.3	132.8	158.7	-0.940
	nian	T2	un	yes	0.295	1.690	132.8	117.9	123.9	-2.253
(9)	shi	T2	un	no	0.175	0.833	223.1	215.4	220.9	0.765
	jian	T1	un	no	0.244	1.160	264.2	248.1	257.2	1.784
	chang	T2	a	no	0.268	1.275	198	166.4	178.9	-0.648

	er	T3	un	no	0.122	0.579	207.4	174.7	194.8	-0.078
	qie	T3	un	no	0.182	0.863	207.1	161.1	179.4	-0.630
	di	T1	a	no	0.244	1.158	232.8	210.2	220.8	0.762
	wen	T1	un	no	0.182	0.866	218	205.1	216.1	0.617
	bi	T3	un	yes	0.190	0.901	217.5	177.7	188	-0.316
	jiao	T3	un	no	0.186	0.885	190.5	133.4	151.5	-1.762
	di	T1	a	no	0.312	1.481	194.9	176	183.1	-0.493
(10)	ta	T1	a	no	0.223	1.280	232	220.7	224	2.044
	zhu	T3	un	no	0.173	0.990	184.7	141.9	156.9	-1.333
	chi	T2	un	no	0.156	0.892	173	160.1	164.1	-0.908
	de	T0	un	no	0.098	0.559	189.8	162.1	179.2	-0.073
	shi	T2	un	yes	0.155	0.890	197.9	181.9	189.2	0.443
	jian	T1	un	no	0.185	1.060	194.1	182.3	189.3	0.448
	jiu	T4	un	no	0.091	0.523	191.9	180.2	187.6	0.362
	gou	T4	a	no	0.226	1.297	197.3	137.5	165.8	-0.810
	chang	T2	a	yes	0.363	2.084	voiceless			
	de	T0	un	no	0.152	0.872	180.4	172.7	177.3	-0.174
	le	T0	un	no	0.096	0.553	voiceless			
(11)	yi	T4	un	no	0.071	0.426	303.6	268.7	293.9	0.635
	shi	T2	un	no	0.164	0.985	305.8	265.8	298.9	0.760
	tou	T1	a	no	0.166	1.000	327.9	318.2	324.3	1.365
	lanr	T3	un	no	0.174	1.047	320	199.2	244.1	-0.743
	mei	T2	un	no	0.146	0.876	250.4	198.2	221	-1.481
	you	T3	un	no	0.096	0.577	250.4	210.8	233.2	-1.082
	kai	T1	a	no	0.230	1.384	276.4	254.4	271.9	0.057
	deng	T1	a	no	0.284	1.705	305.2	280.3	288.2	0.489
(12)	wo	T3	un	no	0.080	0.479	243.2	219.6	237.6	-0.293
	zai	T4	un	yes	0.149	0.891	245.6	237.6	243.1	-0.149
	san	T1	a	no	0.218	1.303	324.5	310	316.8	1.510
	ya	T4	un	no	0.134	0.800	323.7	247.3	286.8	0.887
	guo	T4	un	no	0.160	0.958	263.8	234.7	248.9	-0.002
	de	T0	un	no	0.123	0.734	231	218.5	224.7	-0.642
	fei	T1	a	no	0.188	1.124	319.3	302.9	316	1.494
	chang	T2	a	yes	0.188	1.124	276.8	211.2	235.8	-0.340
	yu	T2	un	no	0.155	0.925	211.2	190.3	200.6	-1.353
	kuai	T4	un	no	0.278	1.663	240.1	188.5	208.5	-1.111
(13)	zhei	T4	un	no	0.132	0.654	330.2	280.3	320.6	1.444
	li	T3	un	no	0.134	0.665	285.6	228.6	254.1	-0.167
	tian	T1	un	no	0.228	1.132	332	314.2	326.6	1.572
	qi	T4	un	no	0.109	0.538	284.3	240.5	265.2	0.129
	hen	T3	a	no	0.223	1.106	280.7	197.1	246	-0.392
	re	T4	un	no	0.280	1.386	288.6	227.7	263.1	0.074
	hen	T3	un	no	0.214	1.060	228.7	187.9	208.8	-1.528
	qing	T2	a	no	0.271	1.341	249.3	214.8	227.6	-0.930
	lang	T3	un	no	0.226	1.119	281.1	204.1	252.8	-0.203
(14)	ming	T2	a	no	0.150	0.976	293.7	243.7	265.1	0.431
	tian	T1	un	no	0.204	1.324	321.2	300.5	317.3	1.436
	bi	T4	a	no	0.121	0.783	298.2	262.8	279.2	0.721

	xu	T1	a	no	0.156	1.015	329.2	319.6	326.1	1.589
	na	T2	un	yes	0.169	1.098	319.6	209.9	247.4	0.045
	dao	T4	un	no	0.118	0.767	221.7	208.9	215.3	-0.733
	wo	T3	un	no	0.100	0.647	208.9	199.3	205.2	-1.001
	de	T0	un	no	0.113	0.734	207.5	200.3	203	-1.061
	xing	T2	un	yes	0.246	1.594	235.7	208.4	211.8	-0.824
	li	T3	un	no	0.164	1.063	267.6	202	220.4	-0.602
(15)	jin	T1	a	no	0.170	1.065	153.4	145.6	151.5	1.330
	van	T3	un	no	0.186	1.170	119.8	105.7	110.5	-0.680
	gong	T4	un	no	0.176	1.105	128.6	118.6	122.6	-0.018
	fa	T1	un	no	0.138	0.865	142.5	135.7	138.6	0.763
	sheng	T1	un	no	0.160	1.004	149.9	144.3	146.4	1.112
	liu	T4	a	no	0.136	0.854	152.9	108.1	137	0.689
	qi	T3	un	yes	0.147	0.925	111.3	104.1	107.1	-0.879
	an	T4	un	yes	0.124	0.779	106.9	101.3	103.7	-1.085
	jian	T4	un	no	0.196	1.233	105.9	97.5	101.3	-1.234
(16)	wo	T3	a	no	0.134	0.965	148	121.1	132.8	0.918
	nü	T3	a	no	0.124	0.896	146.3	121.8	132.9	0.927
	er	T2	un	no	0.074	0.537	133.9	123	129.2	0.615
	da	T3	un	yes	0.143	1.035	131.2	121.7	124.7	0.224
	suan	T4	un	no	0.155	1.116	118.3	111	114.4	-0.728
	ming	T2	a	no	0.103	0.745	126.3	112.6	121	-0.109
	tian	T1	a	no	0.165	1.194	142.3	134.5	139.2	1.438
	kao	T3	un	no	0.125	0.903	107.4	103.5	105.5	-1.623
	dan	T4	a	yes	0.170	1.226	112.6	108.2	109.8	-1.181
	gao	T1	un	no	0.191	1.383	117.9	116.4	117	-0.480
(17)	wo	T3	a	no	0.134	0.926	141.5	121.1	126.7	-0.347
	fu	T4	a	no	0.171	1.184	162	145.4	155.2	1.572
	qin	T0	un	no	0.117	0.810	139.3	132.7	136.3	0.344
	rang	T4	un	no	0.072	0.497	132.7	124.2	127.4	-0.295
	ta	T1	un	no	0.118	0.816	140.2	137.3	138.9	0.523
	dai	T4	a	no	0.150	1.041	144.5	115.1	130.2	-0.089
	shang	T4	un	no	0.206	1.429	129.9	103.8	109.7	-1.709
	gou	T3	un	yes	0.187	1.299	voiceless			
(18)	qi	T2	un	no	0.174	1.170	142.9	130.8	134.3	1.571
	ta	T1	un	no	0.140	0.943	141.4	131.5	133.6	1.510
	chi	T3	un	no	0.150	1.005	102.8	101.9	102.4	-1.618
	cun	T4	un	no	0.185	1.243	120.7	114.8	117.5	0.000
	si	T4	a	no	0.167	1.119	126.1	113.2	120.7	0.316
	zhong	T3	un	no	0.128	0.860	113.1	103	108.1	-0.981
	yan	T2	un	no	0.131	0.880	118.9	109.7	113.5	-0.408
	se	T4	un	no	0.154	1.036	120.9	110	113.4	-0.418
	dou	T1	a	no	0.139	0.932	123.3	116.6	121.1	0.354
	you	T3	un	yes	0.094	0.634	116.6	113.3	114.3	-0.325
	huo	T4	un	yes	0.175	1.179	voiceless			
(19)	bei	T3	a	no	0.139	0.908	122.9	113.3	118.8	-0.387
	jing	T1	un	no	0.165	1.076	148.4	138.5	142.1	1.507
	de	T0	un	no	0.108	0.707	148.8	105.2	133.3	0.831

	zu	T2	a	no	0.172	1.121	125.6	121.7	123.5	0.023
	qiu	T2	un	no	0.138	0.897	133.2	124.8	127.8	0.385
	dai	T4	un	no	0.133	0.865	144.1	122.8	136.5	1.082
	biao	T3	un	no	0.129	0.844	111.5	98.9	103.3	-1.865
	dui	T4	un	no	0.182	1.185	121.6	108.4	113.7	-0.851
	shi	T4	un	yes	0.139	0.908	voiceless			
	guo	T2	a	no	0.187	1.218	117.8	113.9	116.5	-0.594
	an	T1	a	no	0.195	1.271	158.3	110.9	121.7	-0.132
(20)	feng	T1	a	no	0.248	1.419	361.1	342.6	356.7	1.546
	da	T4	a	no	0.156	0.892	357.5	274.5	320.5	0.859
	de	T0	un	no	0.110	0.633	251.5	236.9	247.5	-0.800
	ji	T1	a	no	0.139	0.798	299.8	289.9	296.3	0.355
	hu	T1	un	no	0.170	0.974	331.9	307.3	324.1	0.931
	yao	T4	un	no	0.102	0.587	307.3	243	272	-0.194
	ba	T3	a	no	0.166	0.950	296	257.8	275.2	-0.119
	wo	T3	un	no	0.137	0.786	295.7	207.3	241	-0.971
	chui	T1	a	no	0.195	1.114	303.9	281.2	297.3	0.377
	xia	T4	un	no	0.178	1.021	272.1	222	235.1	-1.130
	shan	T1	a	no	0.280	1.607	335.6	301	321.9	0.887
	qu	T4	un	no	0.213	1.220	251.7	195.8	213.7	-1.743
(21)	fan	T3	un	no	0.121	0.747	200.9	190.3	195	-0.068
	zheng	T4	un	yes	0.097	0.598	244.1	211.5	228.6	0.635
	huan	T1	a	no	0.214	1.326	303.8	285.4	294	1.748
	le	T4	un	no	0.113	0.701	298.4	237.8	271.3	1.393
	zong	T3	un	no	0.207	1.280	198.5	165.6	174.9	-0.549
	dong	T4	un	no	0.190	1.179	169.1	156.8	160.2	-0.937
	yuan	T2	a	no	0.171	1.058	260.8	155.7	172.3	-0.615
	wo	T3	un	no	0.138	0.853	271.5	196.4	242.4	0.895
	dao	T4	a	no	0.141	0.874	195.2	160.4	173	-0.597
	shi	T4	un	no	0.102	0.629	160.4	148.6	153.3	-1.132
	kan	T4	a	no	0.284	1.756	174.6	160	166.3	-0.772
(22)	zhu	T3	un	yes	0.117	0.718	187.9	187.3	187.7	-0.423
	yao	T4	un	yes	0.093	0.571	208.7	186.9	195.3	-0.203
	kan	T4	a	no	0.220	1.347	250.5	232.2	238.7	0.909
	nei	T4	un	no	0.090	0.552	232.2	192.5	212.2	0.257
	ge	T4	un	no	0.059	0.361	192.5	179.1	184.5	-0.518
	mo	T2	a	no	0.163	1.003	243.8	179.2	208.4	0.157
	fang	T3	un	no	0.173	1.060	152.4	145.1	147.5	-1.759
	xiu	T4	a	no	0.389	2.389	280.1	249.6	269.5	1.581
(23)	zhe	T4	un	no	0.143	0.782	265.9	253	258.6	0.926
	hui	T2	un	no	0.141	0.775	267.3	246.9	255.9	0.890
	huan	T4	a	no	0.306	1.681	323.3	181.8	282.6	1.236
	le	T0	un	no	0.100	0.548	181.8	164.2	171.1	-0.513
	zhu	T3	un	no	0.173	0.948	147.5	145.2	146.3	-1.059
	chi	T2	un	no	0.243	1.335	238.2	215.8	228	0.488
	ren	T2	a	no	0.172	0.943	180.2	149.1	156.1	-0.833
	le	T0	un	no	0.150	0.821	253.2	180.2	222.2	0.398
	ha	T0	un	no	0.213	1.167	177.1	78.6	127.7	-1.533

(24)	yuan	T2	un	no	0.139	0.717	235.3	193.6	208.2	0.345
	lai	T2	un	no	0.153	0.788	287.5	235.3	267.7	1.318
	zhe	T4	un	no	0.137	0.706	302.2	224.1	250.7	1.064
	yuanr	T4	a	no	0.242	1.247	260.4	222.3	241.5	0.919
	li	T3	un	no	0.157	0.811	142.9	134	139.2	-1.214
	mei	T2	a	no	0.221	1.141	228.8	140.2	163.9	-0.582
	zhuang	T1	a	yes	0.273	1.411	254	182.3	204.7	0.279
	ji	T3	un	no	0.212	1.094	147.1	136.9	142	-1.137
	ge	T4	un	yes	0.113	0.584	149.9	145.9	147.4	-0.992
	suo	T3	un	yes	0.291	1.502	voiceless			
(25)	suo	T3	un	no	0.176	1.061	259.2	241.9	251.7	1.243
	yi	T3	un	no	0.096	0.579	256.9	227.5	240.4	0.870
	zui	T4	a	no	0.189	1.134	254.9	234	241.9	0.920
	hou	T4	un	no	0.132	0.795	256.4	234	241.2	0.897
	nong	T4	un	no	0.196	1.179	233.4	204.9	222.5	0.242
	yi	T2	un	no	0.104	0.626	221.1	204.3	210.3	-0.216
	liang	T3	un	no	0.196	1.182	203.6	158.5	174	-1.755
	tour	T2	un	no	0.158	0.950	204.8	184.7	196.6	-0.763
	bu	T4	un	no	0.153	0.923	205.7	180.4	189.3	-1.070
	zhen	T1	a	no	0.261	1.572	214.4	198	206.4	-0.368
(26)	jin	T1	un	no	0.188	0.957	279	253.2	262.5	1.534
	nian	T2	un	no	0.175	0.893	253.2	248.3	250.8	1.337
	da	T3	a	no	0.188	0.960	159.5	152.1	155	-0.734
	suan	T4	un	yes	0.175	0.893	193.3	180.3	187.2	0.078
	shang	T4	un	no	0.192	0.978	221.6	203.5	208.4	0.540
	miao	T4	a	no	0.278	1.418	223.6	193.9	209	0.552
	hui	T4	a	no	0.255	1.303	252.3	170.4	212	0.614
	guang	T4	un	no	0.275	1.407	190.5	156.4	168.4	-0.377
	guang	T0	un	no	0.189	0.963	147.6	140.2	142.5	-1.096
	qu	T4	un	yes	0.126	0.643	139.3	139	139.3	-1.194
	ma	T0	un	no	0.115	0.586	142.4	134.2	137.4	-1.253
(27)	you	T3	un	no	0.100	0.586	170.4	164.8	167.5	-1.031
	ti	T2	a	no	0.266	1.558	228.2	216.4	220.4	0.909
	gao	T1	un	no	0.185	1.082	248.1	228.2	238	1.452
	banr	T1	un	no	0.206	1.203	232.6	207.4	226.9	1.115
	you	T3	un	no	0.097	0.566	207.4	145.1	165	-1.137
	ji	T1	a	no	0.155	0.907	225.1	211.6	213.5	0.685
	chu	T3	un	no	0.143	0.839	177.3	153.5	162.3	-1.254
	banr	T1	un	no	0.221	1.292	205.1	174.6	193	-0.029
	liang	T3	a	no	0.127	0.745	197.8	151.9	176.8	-0.649
	ge	T4	un	yes	0.111	0.649	voiceless			
	ban	T1	un	no	0.269	1.575	203.2	179.7	192.1	-0.062
(28)	xian	T4	a	no	0.285	1.696	287.4	241.3	257.9	0.945
	zai	T4	a	no	0.170	1.010	262.9	212.9	240.8	0.736
	shi	T4	un	yes	0.132	0.787	227.6	216.6	225.4	0.534
	zen	T4	a	no	0.155	0.926	284.1	241	262.4	0.998
	me	T0	un	no	0.090	0.536	241	169.1	201.9	0.197
	ge	T4	un	no	0.123	0.735	153.8	142.5	146.4	-0.784

	gai	T4	a	no	0.187	1.116	161.6	119.3	136	-1.009
	nian	T4	un	no	0.200	1.193	119.3	100.9	111.5	-1.616
(29)	ta	T1	a	no	0.197	1.307	195.9	188.2	192.1	1.749
	m	T0	un	no	0.085	0.564	195	175.7	188.7	1.552
	de	T0	un	no	0.085	0.564	166.5	147.7	159.5	-0.300
	cang	T2	un	no	0.183	1.219	156	147.1	149.9	-0.983
	shu	T1	un	no	0.100	0.665	167.9	159.9	165.6	0.114
	liang	T4	un	no	0.158	1.048	166.6	126.5	150.8	-0.917
	xiang	T1	a	no	0.242	1.612	176.4	171.2	173.9	0.653
	dang	T1	a	no	0.176	1.169	167.4	152.4	156.8	-0.488
	de	T0	un	no	0.121	0.806	159	152.8	154.7	-0.636
	da	T4	a	no	0.158	1.048	162.4	143.7	153.2	-0.743
(30)	dan	T4	a	no	0.166	1.088	245.9	212.2	231.6	0.979
	shi	T4	un	no	0.098	0.645	245.6	215.3	230.3	0.959
	yi	T4	un	no	0.115	0.758	245.9	179.5	217.2	0.748
	fen	T1	un	no	0.224	1.470	250.6	229.4	239.2	1.096
	lei	T4	un	no	0.143	0.939	260.4	135	197.4	0.403
	yi	T3	un	yes	0.119	0.783	111.9	105.7	109.7	-1.717
	hou	T4	un	no	0.175	1.151	193.4	148.7	172.6	-0.081
	jiu	T4	un	no	0.128	0.841	156.5	151	153.8	-0.498
	shang	T4	a	no	0.243	1.598	188.2	134.6	166.2	-0.218
	le	T0	un	no	0.105	0.690	164.6	134.8	148.4	-0.626
	jia	T4	a	no	0.208	1.369	220.1	212	214.6	0.704
	le	T0	un	no	0.102	0.669	110.6	105.9	108.7	-1.750
(31)	ma	T2	a	no	0.130	0.822	270.2	220.8	234.2	0.286
	fan	T2	un	no	0.150	0.949	271.8	243.1	261.5	0.954
	nin	T2	a	no	0.179	1.134	243.1	226.7	233.9	0.278
	bang	T1	a	no	0.146	0.923	267	242.9	258	0.873
	wo	T3	un	no	0.085	0.536	242.9	205.1	219.6	-0.104
	lian	T2	a	no	0.263	1.665	251.8	194.4	209.3	-0.396
	xi	T4	un	no	0.158	0.999	277.3	243.4	263.2	0.993
	yi	T2	un	no	0.083	0.525	243.4	95.6	196.2	-0.787
	xiaer	T4	un	no	0.229	1.448	214.3	70.4	158.1	-2.096
(32)	wo	T3	a	no	0.099	0.584	351	287	322.1	1.539
	nü	T3	a	no	0.152	0.897	349	243.1	307.3	1.150
	er	T2	un	no	0.131	0.771	280.8	240.1	255.5	-0.378
	da	T3	un	no	0.160	0.943	291.6	223.8	244.8	-0.732
	suan	T4	un	yes	0.151	0.893	243	228.4	232.7	-1.152
	ming	T3	a	no	0.184	1.085	278.9	228.6	249.4	-0.578
	tian	T1	un	no	0.170	1.004	305.9	289.2	301.5	0.992
	kao	T3	un	no	0.171	1.010	voiceless			
	dan	T4	un	no	0.212	1.251	246.7	230.3	237.5	-0.983
	gao	T1	a	no	0.265	1.562	280.4	254.7	272.1	0.143
(33)	ta	T1	un	no	0.161	0.879	312.2	297.5	305.4	0.967
	zhi	T2	a	no	0.159	0.870	307.5	260.7	275.8	0.314
	yi	T4	un	no	0.135	0.738	337.2	307.5	329.2	1.448
	yao	T4	un	no	0.171	0.936	321.2	293.6	307.4	1.009
	yong	T4	un	no	0.152	0.828	293.6	234.2	263.3	0.017

	shou	T3	a	no	0.288	1.572	228	194	211.5	-1.386
	lai	T2	un	no	0.161	0.879	259.7	207.9	236.8	-0.662
	jiao	T3	un	no	0.175	0.955	250.5	219.8	234.3	-0.730
	ban	T4	un	no	0.246	1.344	285.5	204.1	225.4	-0.978
(34)	ta	T1	a	no	0.186	1.138	337.4	333.6	336.6	1.562
	de	T0	un	no	0.067	0.413	314.3	288	304.9	1.043
	hai	T2	a	no	0.198	1.213	265.3	240.5	246.7	-0.069
	zi	T0	un	no	0.101	0.616	321.6	284.6	306.8	1.075
	men	T0	un	no	0.230	1.412	305.1	232.8	272.8	0.459
	zao	T3	a	no	0.207	1.270	235.9	200.2	220.4	-0.661
	jiu	T4	un	yes	0.131	0.803	298.2	233.9	272.9	0.461
	deng	T3	a	no	0.166	1.017	282.7	197.3	232.9	-0.372
	zhe	T0	un	no	0.107	0.658	193.5	184.1	191.3	-1.405
	chi	T1	a	no	0.298	1.825	249.5	200	215.9	-0.769
	le	T0	un	no	0.104	0.636	200	190.8	194.3	-1.323
(35)	wo	T3	un	no	0.083	0.504	252.2	242.2	249.1	-0.676
	fu	T4	a	no	0.180	1.098	341.6	307.9	333.9	1.415
	qin	T0	un	no	0.127	0.775	311.6	292.3	304.1	0.748
	rang	T4	un	no	0.141	0.860	292.3	257.4	276.5	0.069
	ta	T1	un	no	0.123	0.752	302.2	278	292.3	0.465
	dai	T4	a	no	0.180	1.098	317.4	271.3	297.3	0.586
	shang	T4	un	no	0.200	1.221	269.5	205.2	231.3	-1.205
	gou	T3	a	no	0.277	1.693	326.6	71.9	225	-1.402
(36)	yuan	T2	a	no	0.129	0.946	148.8	127.5	134.7	-0.444
	yin	T1	un	no	0.131	0.966	152.9	146.1	150.8	1.149
	zhi	T1	un	no	0.144	1.059	151.1	146.9	148.5	0.932
	yi	T1	un	no	0.090	0.658	152.8	147.5	149.5	1.027
	shi	T4	un	no	0.134	0.987	144.1	133.4	138.2	-0.082
	zuo	T4	a	no	0.174	1.276	145.1	131.7	136.8	-0.225
	wei	T4	un	no	0.094	0.688	134.5	123.8	131.4	-0.793
	bu	T4	un	no	0.142	1.047	122.4	120.4	121.2	-1.933
	shu	T1	a	no	0.176	1.296	150.3	131.5	142.7	0.370
	fu	T2	un	yes	0.146	1.077	voiceless			
(37)	geng	T4	a	no	0.170	1.186	170.1	136.6	148.2	1.124
	zao	T1	a	no	0.154	1.069	154.8	147.9	151.6	1.280
	de	T0	un	no	0.093	0.644	146.6	139.6	142.6	0.859
	shi	T4	un	no	0.117	0.818	115.8	114.4	115.2	-0.610
	you	T3	un	no	0.132	0.919	114.9	98.4	108.1	-1.048
	ren	T2	un	no	0.138	0.964	130.2	109.6	118.9	-0.392
	da	T3	un	no	0.118	0.818	107.9	102.2	104	-1.314
	han	T1	a	no	0.227	1.583	130.9	125	127.7	0.099
(38)	er	T3	un	no	0.084	0.638	153.8	144.8	150.2	1.634
	qie	T3	un	no	0.139	1.057	135	119.1	127.9	0.159
	zhe	T4	a	no	0.172	1.308	157.4	128.7	148.7	1.542
	li	T3	un	no	0.058	0.439	128.7	119.9	123.4	-0.169
	de	T0	un	no	0.099	0.754	140.6	137.9	139	0.923
	hai	T3	un	no	0.174	1.319	129.4	107.2	118.6	-0.534
	jian	T3	a	no	0.181	1.371	120.7	109.2	112.6	-1.010

	zhi	T2	un	no	0.117	0.887	129.6	126.3	128.4	0.195
	mei	T3	un	no	0.145	1.102	132.6	101.5	115.9	-0.745
	ji	T2	a	no	0.185	1.403	111	107.2	109.4	-1.275
	le	T0	un	no	0.095	0.722	119.8	108.8	116.2	-0.721
(39)	er	T3	a	no	0.078	0.622	148	137.4	143.1	1.777
	wo	T3	un	no	0.093	0.742	142.2	107.9	126.4	0.207
	de	T0	un	no	0.086	0.684	114.6	112.2	113.2	-1.188
	xing	T2	a	no	0.214	1.699	135.4	118.1	124.1	-0.025
	li	T3	un	no	0.072	0.572	137.8	121.7	133.1	0.861
	que	T4	a	no	0.154	1.223	136.4	126.4	132.7	0.822
	qu	T4	un	no	0.145	1.155	128.9	126.7	128.3	0.396
	le	T0	un	no	0.079	0.624	126.7	107.4	114.8	-1.011
	luo	T2	a	no	0.150	1.194	121	106.9	113.6	-1.144
	ma	T3	a	no	0.187	1.486	121.2	113.3	117.7	-0.695
(40)	wo	T3	un	no	0.064	0.517	122.6	114.8	119.8	-1.199
	hai	T2	a	no	0.127	1.029	144	129.5	136.5	0.251
	xu	T1	a	no	0.187	1.518	161.4	154	157.3	1.828
	yao	T4	un	no	0.073	0.597	154	134.4	143.8	0.830
	yi	T4	un	no	0.070	0.573	134.4	117	123.3	-0.879
	xie	T1	un	no	0.161	1.305	147.6	139.6	142.5	0.730
	ying	T4	a	no	0.161	1.305	145.8	114.1	132.1	-0.113
	ji	T2	un	no	0.169	1.372	127.2	120.3	122.5	-0.951
	yao	T4	a	no	0.097	0.785	129.4	124.9	127.6	-0.498
(41)	fei	T2	a	no	0.291	1.412	246.5	166.3	188.1	0.701
	shour	T4	un	no	0.280	1.360	276.9	179.6	219.2	1.876
	ma	T0	un	no	0.139	0.677	179.6	158	166.6	-0.231
	ye	T3	un	no	0.130	0.631	158.1	140.5	147.2	-1.182
	jue	T2	a	no	0.165	0.803	161.9	149.4	153.7	-0.850
	dui	T4	un	no	0.150	0.728	192.3	165.9	181.5	0.427
	bu	T2	un	no	0.186	0.905	173.5	140.7	154.8	-0.795
	shou	T4	a	no	0.305	1.484	217.1	129.3	172.9	0.054
(42)	fang	T4	un	no	0.190	1.161	192.4	170.5	175.3	1.221
	jia	T4	a	no	0.260	1.589	197.5	144.4	173.1	1.148
	le	T0	un	no	0.138	0.845	144.4	133.5	137.4	-0.190
	gei	T3	un	no	0.124	0.759	142.8	131.9	135.7	-0.262
	song	T4	a	no	0.212	1.295	195.9	158.3	170.8	1.070
	dao	T4	un	no	0.120	0.732	149.5	140.3	144.7	0.110
	lao	T3	a	no	0.164	0.999	140.3	71.4	103.1	-1.854
	lao	T0	un	no	0.122	0.745	119.7	112.1	116.4	-1.151
	jia	T1	a	no	0.229	1.399	166.1	158.1	161.8	0.757
	qu	T4	un	no	0.140	0.857	158	129.8	140	-0.082
	le	T0	un	no	0.102	0.621	130	121.3	124.4	-0.766
(43)	dou	T1	a	no	0.168	0.934	215.2	204	207.4	1.442
	fan	T3	un	no	0.195	1.082	168.9	96.1	123.5	-0.719
	ying	T4	un	yes	0.129	0.715	156.2	132	147.3	0.015
	ta	T1	a	no	0.185	1.023	208.6	201	204.2	1.377
	nei	T4	un	no	0.146	0.811	204.8	162	187.9	1.030
	ge	T4	un	no	0.098	0.542	162	146.3	151.8	0.141

	yan	T3	un	no	0.195	1.083	146.3	74.6	106.9	-1.321
	de	T0	un	no	0.115	0.639	123	116.5	118.9	-0.877
	bu	T2	un	no	0.201	1.116	146.2	136.7	141.8	-0.143
	cuo	T4	a	no	0.371	2.056	137.2	95.2	117	-0.945
(44)	xian	T4	un	no	0.177	0.921	235.7	230.3	233	1.262
	zai	T4	un	no	0.169	0.883	228.5	207.3	217.8	0.786
	kai	T1	a	no	0.212	1.104	241.9	217	222.5	0.937
	ge	T4	un	no	0.196	1.020	231.2	189.9	212	0.596
	zhong	T3	un	no	0.165	0.862	176.3	154.2	160.4	-1.372
	ge	T4	un	no	0.141	0.736	194.9	191.9	193.1	-0.063
	yang	T4	a	no	0.135	0.704	191.9	170.6	180.5	-0.539
	de	T0	un	no	0.133	0.692	162.5	153.2	156.6	-1.541
	banr	T1	a	no	0.399	2.079	196	190.1	193	-0.067
(45)	jiu	T4	un	no	0.128	0.811	181.6	169.1	177.5	0.035
	deng	T3	un	no	0.131	0.824	170.2	63.3	137.3	-1.113
	yu	T2	un	no	0.123	0.778	180.4	54.8	140.5	-1.010
	shi	T4	un	no	0.163	1.028	198.6	171.4	187.8	0.288
	gei	T3	un	no	0.128	0.806	178.4	158.6	169.4	-0.173
	ta	T1	un	no	0.125	0.791	174.4	168.9	171.1	-0.129
	tu	T1	a	no	0.211	1.330	251.5	240.4	248.6	1.542
	ran	T2	a	no	0.190	1.202	252.8	163	200.5	0.580
	jiu	T4	un	no	0.120	0.761	294.7	213.9	258.2	1.711
	huan	T4	a	no	0.253	1.596	197.4	134.9	161.7	-0.381
	le	T0	un	no	0.170	1.074	134.9	128	130.2	-1.350
(46)	zao	T4	un	no	0.225	1.239	244.4	228	238.4	1.675
	mu	T4	a	no	0.228	1.258	282.6	226.6	249.9	1.847
	qian	T2	un	no	0.199	1.098	158.6	131.5	141.4	-0.236
	kan	T4	un	no	0.223	1.229	163.1	127.1	145.1	-0.142
	ying	T1	un	no	0.105	0.578	159.1	126.3	139.5	-0.286
	gai	T1	un	no	0.167	0.921	166.2	164.2	165.3	0.335
	yi	T4	a	no	0.125	0.689	177.2	159.4	167.9	0.392
	bai	T3	un	no	0.143	0.789	118	110.2	113.1	-1.053
	nian	T2	un	yes	0.195	1.075	118	110.1	112.5	-1.072
	mei	T2	a	no	0.148	0.818	126.8	117.3	119.2	-0.861
	wen	T4	un	no	0.152	0.836	130.6	124.2	128	-0.600
	ti	T2	un	yes	0.267	1.471	voiceless			
(47)	mei	T3	un	no	0.168	1.241	123.3	111.6	115.6	-1.512
	tian	T1	a	no	0.192	1.416	186.7	179.6	183.3	1.192
	ni	T3	un	no	0.064	0.476	179.9	129.9	145.3	-0.171
	shuo	T1	un	no	0.097	0.717	161	127.6	144.9	-0.187
	nei	T4	un	no	0.135	1.000	161.1	147.7	156.2	0.254
	zhi	T1	a	no	0.160	1.179	177.2	170.7	175.2	0.927
	piao	T4	a	no	0.169	1.247	172.6	164.7	169.4	0.729
	dei	T3	un	no	0.096	0.709	110	107.7	108.6	-1.878
	duo	T1	a	no	0.162	1.199	166.7	133.2	154.4	0.186
	shao	T3	un	yes	0.111	0.818	161.8	161.8	161.8	0.460
(48)	nei	T4	a	no	0.128	0.910	173.3	166.9	170.8	1.125
	xie	T1	un	no	0.142	1.012	160	154.6	158.8	0.406

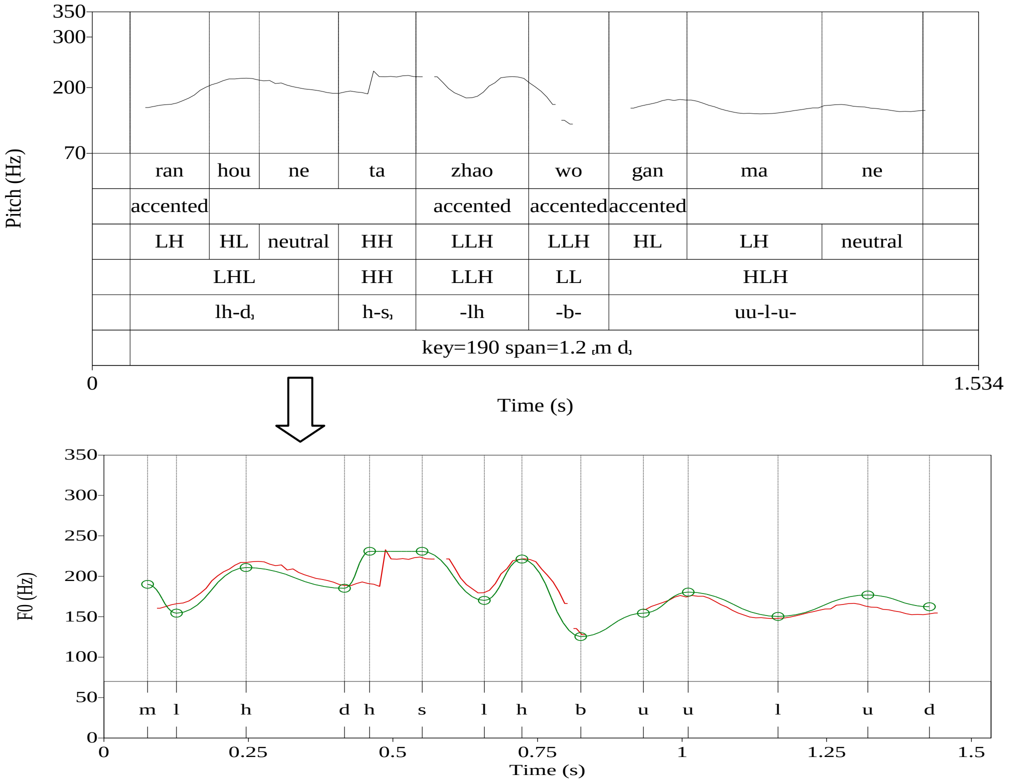
	pian	T1	a	no	0.235	1.669	178.8	164.5	168.1	0.968
	zi	T0	un	no	0.090	0.639	166.3	166	166.2	0.855
	xian	T4	a	no	0.181	1.287	154.7	152.9	153.9	0.096
	zai	T4	un	no	0.087	0.618	153.2	126	138.7	-0.930
	hai	T2	un	no	0.070	0.499	128.9	125.6	127.3	-1.777
	neng	T2	un	no	0.137	0.976	150.7	128.2	138.9	-0.916
	kan	T4	a	no	0.196	1.390	160.1	152.8	155.1	0.173
(49)	que	T4	a	no	0.231	1.587	218.4	176.8	198.4	1.659
	shi	T2	un	no	0.163	1.119	172.1	161.2	167.8	0.670
	jiu	T4	a	no	0.138	0.947	170.4	160.3	165.3	0.581
	you	T3	un	yes	0.067	0.456	148.5	135.4	141.4	-0.342
	jing	T1	a	no	0.157	1.078	128.4	126.1	127.4	-0.958
	ji	T4	un	yes	0.106	0.728	139.9	128.1	136.5	-0.550
	li	T4	a	no	0.193	1.322	140.2	110.6	125.2	-1.061
	yi	T4	un	yes	0.111	0.763	voiceless			
(50)	jiu	T4	a	no	0.165	1.145	158	142.7	150.4	0.342
	shuo	T1	un	no	0.115	0.796	162.1	142	154.7	0.456
	zhei	T4	a	no	0.164	1.133	182.6	150.1	169.4	0.823
	ge	T4	un	no	0.058	0.402	231.4	227.5	229.3	2.047
	fa	T1	un	no	0.194	1.342	126.2	118.6	121.5	-0.521
	xing	T2	un	yes	0.092	0.638	119.8	118.4	119	-0.605
	liang	T4	un	no	0.168	1.164	121.2	105.9	114.6	-0.757
	bi	T3	un	no	0.169	1.169	111.2	108.7	109.7	-0.934
	jiao	T3	un	yes	0.125	0.868	112.9	111.2	112	-0.850
	shao	T3	a	yes	0.194	1.343	voiceless			
(51)	ta	T1	un	no	0.106	0.639	313.7	262	303.6	0.937
	de	T0	un	no	0.112	0.675	325.1	280.6	316.9	1.092
	jiao	T4	a	no	0.251	1.511	365.5	245.8	313.7	1.056
	sheng	T1	un	no	0.230	1.385	253.7	230.2	234.9	0.012
	ke	T3	un	yes	0.076	0.459	231.2	226.6	229.6	-0.071
	yi	T3	un	no	0.119	0.718	226.6	217.7	221.4	-0.202
	bao	T3	a	no	0.187	1.126	222.8	63.5	130.7	-2.103
	hu	T4	un	no	0.187	1.124	234.6	218.3	224.2	-0.156
	ta	T1	un	no	0.226	1.363	220.2	192.2	200.2	-0.565
(52)	na	T4	un	no	0.087	0.504	328.9	313.1	325	1.185
	lao	T3	a	no	0.192	1.109	313.1	232	256.6	-0.268
	yu	T2	un	no	0.128	0.744	305.8	235.5	274	0.135
	fu	T1	un	no	0.210	1.216	346.7	320.8	340.1	1.464
	zhang	T3	un	no	0.167	0.967	289.3	228.4	245.4	-0.543
	de	T0	un	no	0.109	0.628	234.8	231.2	232.7	-0.869
	fei	T1	a	no	0.215	1.247	339.7	335	337.6	1.418
	chang	T2	a	yes	0.209	1.211	264.5	224.2	237.5	-0.744
	gao	T1	a	no	0.176	1.021	252.4	237.7	243.1	-0.600
	da	T4	a	no	0.234	1.353	256.1	191	221.3	-1.178
(53)	ta	T1	un	no	0.130	0.874	350.1	342.8	346.6	1.427
	jiu	T4	un	no	0.103	0.695	324.9	277.1	311.5	0.958
	hao	T3	un	no	0.131	0.881	244	227.6	236	-0.261
	xiang	T4	un	yes	0.190	1.282	290	247.3	272.9	0.377

	bian	T4	a	no	0.224	1.507	378.9	308.9	360.8	1.603
	le	T0	un	no	0.108	0.726	308.9	228.2	256.2	0.100
	yi	T2	un	yes	0.119	0.799	228.2	208.4	221.6	-0.538
	ge	T4	un	no	0.149	1.005	220	202.7	212.1	-0.730
	ren	T2	un	no	0.165	1.110	218.4	195.6	203.3	-0.916
	shi	T4	a	no	0.156	1.052	233.1	210.2	224.4	-0.482
	de	T0	un	no	0.159	1.069	210.4	112.3	176.5	-1.537
(54)	er	T3	un	no	0.078	0.473	323.3	284.2	312.3	1.344
	qie	T3	un	no	0.141	0.858	318.8	256.6	292.2	0.920
	zhei	T4	a	no	0.131	0.794	317.7	298.8	313.4	1.367
	li	T3	un	no	0.110	0.666	298.8	231.3	258.3	0.134
	de	T0	un	no	0.110	0.668	271.9	258.5	266.7	0.338
	hai	T3	a	no	0.290	1.758	236.8	205.8	217.9	-0.951
	jian	T3	un	no	0.166	1.006	236.8	204.4	211.7	-1.135
	zhi	T2	un	no	0.144	0.875	262.5	212.1	238.3	-0.380
	mei	T3	un	no	0.219	1.328	266.9	190.6	226.6	-0.701
	ji	T2	a	no	0.246	1.493	217	187.5	200.3	-1.488
	le	T0	un	no	0.178	1.080	458.8	212.9	275.8	0.552
(55)	er	T3	un	no	0.112	0.692	312.2	249.2	282.6	1.194
	wo	T3	un	no	0.107	0.658	310.9	225.9	278.5	1.072
	de	T0	un	no	0.124	0.765	218.2	205	212.6	-1.180
	xing	T2	a	no	0.256	1.584	255.1	219.5	230.4	-0.510
	li	T3	un	no	0.150	0.927	315.5	255.1	294.9	1.549
	que	T4	un	no	0.155	0.960	266.2	235.8	250.7	0.195
	qu	T4	un	no	0.148	0.914	256.6	235.4	249.5	0.155
	le	T0	un	no	0.123	0.760	235.4	209.2	220.9	-0.861
	luo	T2	a	no	0.198	1.221	239.9	203.7	212.8	-1.173
	ma	T3	a	no	0.246	1.519	248.7	192.1	232.3	-0.441
(56)	hai	T2	un	no	0.132	0.887	120.9	116.1	119.1	0.007
	suan	T4	a	no	0.170	1.149	166.3	138.4	147.3	0.796
	shi	T4	un	no	0.110	0.742	132.2	122.4	126.3	0.225
	yi	T4	un	yes	0.100	0.674	125	121.7	123.4	0.139
	chang	T3	un	no	0.215	1.450	voiceless			
	jing	T1	a	no	0.195	1.316	150.8	139.6	146.2	0.768
	cai	T3	un	no	0.141	0.953	145.4	132.1	138.3	0.562
	de	T0	un	no	0.071	0.476	voiceless			
	bi	T3	un	yes	0.129	0.868	66	61.9	63.7	-2.315
	sai	T4	a	no	0.220	1.485	114.2	112.3	113.2	-0.181
(57)	wo	T3	un	no	0.112	0.632	178	118	126.3	-0.462
	yao	T4	un	yes	0.076	0.427	132.8	118.9	125.8	-0.498
	ding	T4	a	no	0.207	1.167	167.8	126	144.8	0.807
	shi	T2	a	no	0.204	1.149	145.5	130.8	137	0.293
	he	T2	un	no	0.280	1.576	126.3	116.4	119.4	-0.983
	sheng	T4	a	no	0.200	1.123	177.3	146.4	156.4	1.522
	dan	T4	un	no	0.148	0.836	141.8	129.2	134	0.088
	cui	T4	a	no	0.181	1.019	174.9	120.2	146	0.884
	bing	T3	un	no	0.190	1.071	126.7	104.3	111.1	-1.652
(58)	qing	T3	un	no	0.209	1.197	129.1	120.7	126.4	-0.832

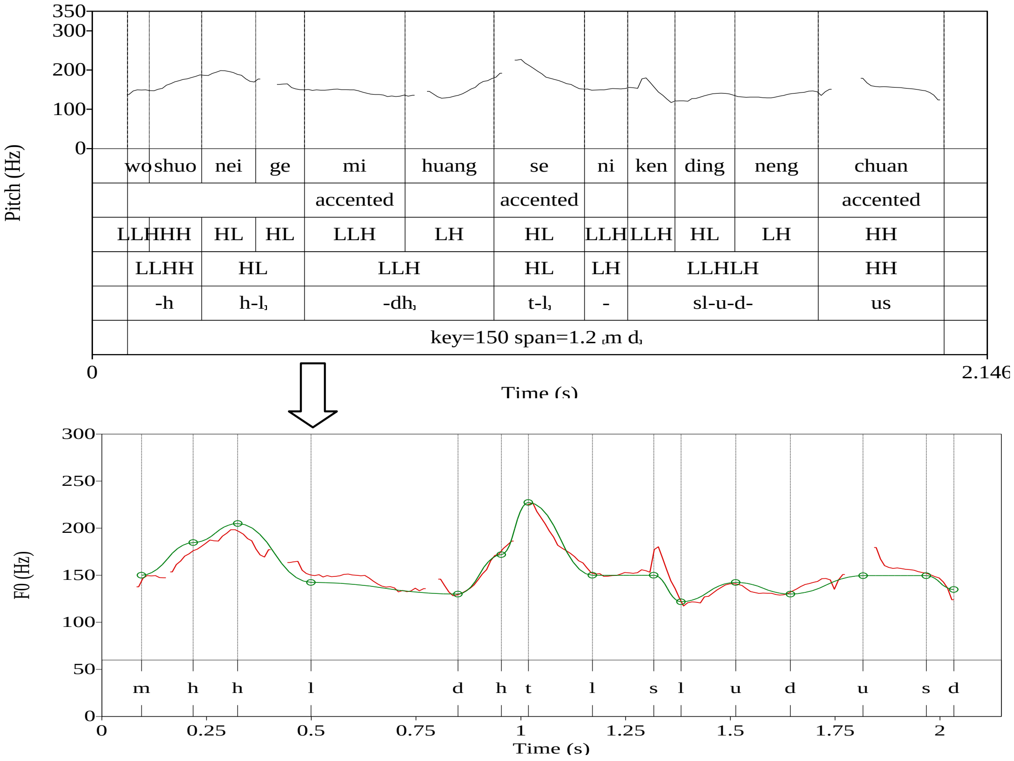
	bang	T1	a	no	0.168	0.967	155.2	146.7	152	1.961
	wo	T3	un	no	0.125	0.718	149.4	107.3	124.2	-1.098
	jie	T1	a	no	0.202	1.161	139.3	132.5	135.2	0.187
	fu	T2	un	yes	0.149	0.855	129.7	126.1	127.7	-0.677
	wu	T4	un	yes	0.054	0.311	131	123.9	129.8	-0.430
	zhong	T1	un	no	0.178	1.023	137.2	131.6	134.8	0.142
	xin	T1	a	no	0.308	1.768	153.2	135.2	140.3	0.748
(59)	wo	T3	un	no	0.053	0.381	120.3	116.4	119	-0.048
	shi	T4	un	no	0.145	1.049	140.6	130.9	135.3	0.643
	shi	T2	un	no	0.127	0.922	141.9	140.1	141	0.864
	san	T1	a	no	0.174	1.260	152.5	137.2	142.7	0.929
	hao	T4	un	no	0.168	1.218	162	122.2	139.3	0.799
	song	T4	un	no	0.173	1.252	125.7	106.5	114.1	-0.274
	qu	T4	un	no	0.132	0.956	93.7	92.2	93.2	-1.363
	de	T0	un	no	0.133	0.964	93.8	87.7	90	-1.550
(60)	wo	T3	un	no	0.098	0.592	118.8	112.4	116	-0.774
	m	T0	un	no	0.116	0.699	126.5	117.7	121.7	-0.203
	yao	T4	a	no	0.145	0.875	148.2	126.5	144.6	1.852
	wei	T4	un	no	0.169	1.017	146.8	119.4	136.9	1.200
	yi	T2	un	no	0.148	0.891	120.7	112.7	117.7	-0.601
	ge	T4	un	no	0.170	1.024	137.9	120	126.4	0.249
	da	T4	a	no	0.176	1.060	148.5	121.1	136	1.121
	xing	T2	un	no	0.208	1.253	124.2	108.5	117.5	-0.621
	hui	T4	un	no	0.158	0.952	136	111.7	125.1	0.126
	yi	T4	un	no	0.078	0.470	111.7	102.8	108.8	-1.538
	ding	T4	a	no	0.202	1.217	129.5	106.9	113.8	-1.003
	can	T1	a	yes	0.324	1.952	139.3	119.6	125.8	0.192

Appendix III Annotations of speech data and the derived synthesis result

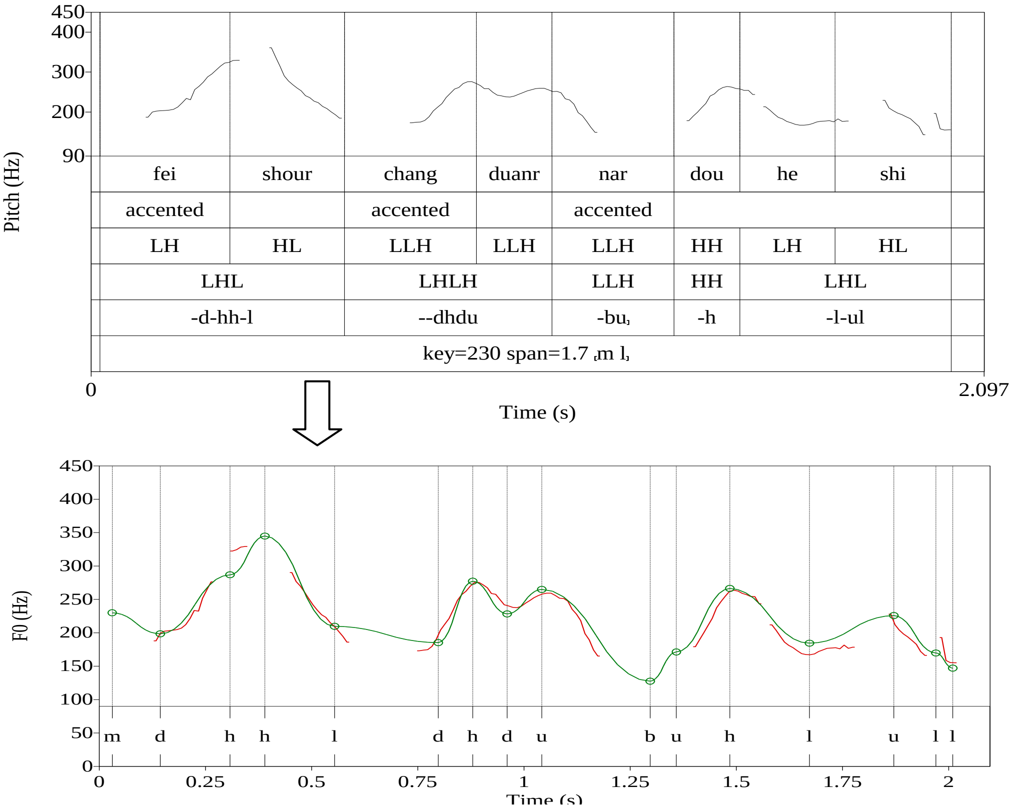
(1)



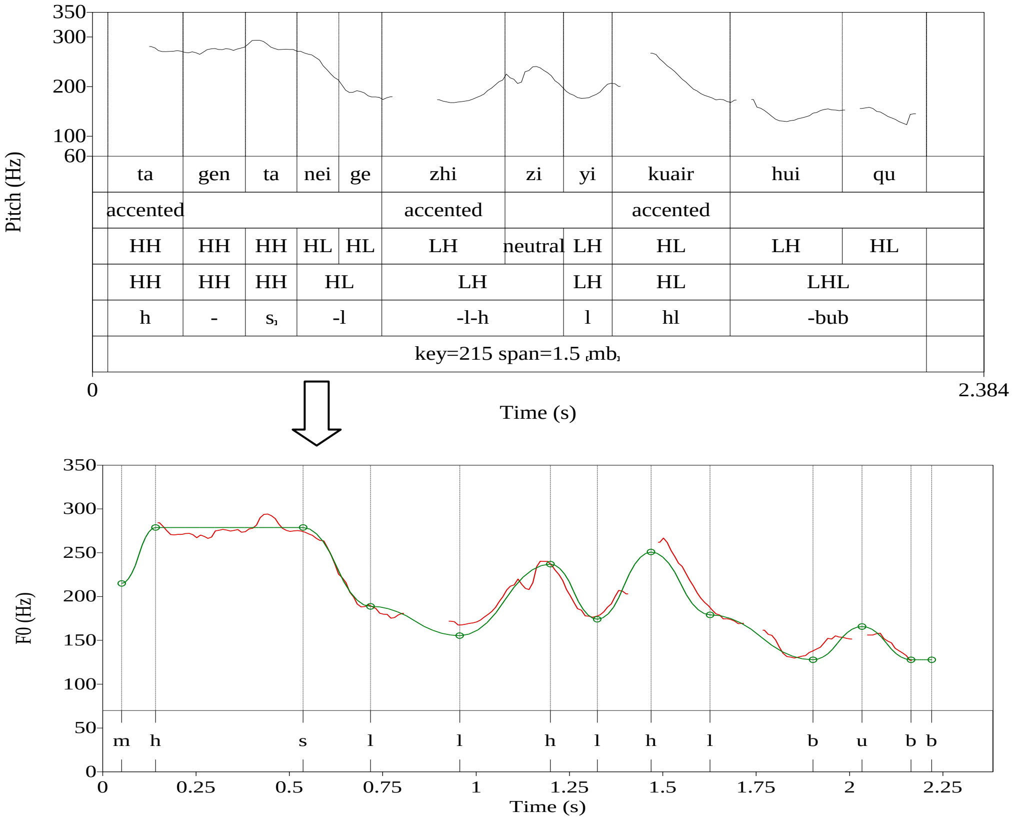
(2)



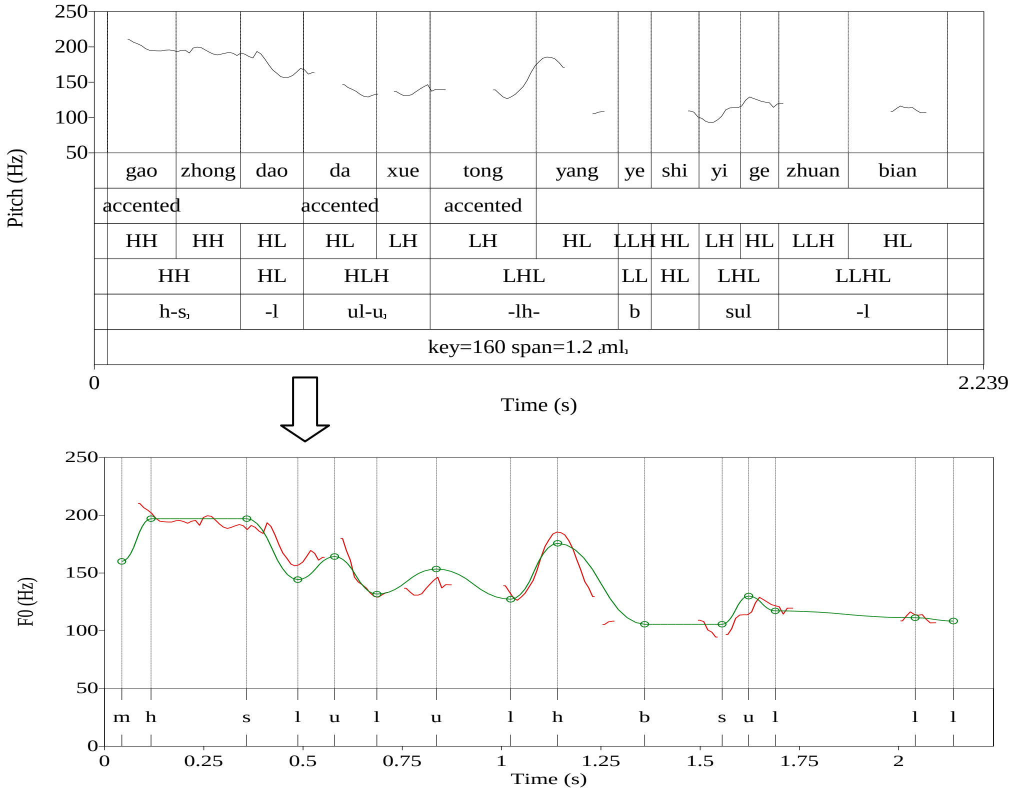
(3)


450
400
300
200
90
Pitch (Hz)
fei shour chang duanr nar dou he shi
accented accented accented
LH HL LLH LLH LLH HH LH HL
LHL LHLH LLH HH LHL
-d-hh-l --dhdu -bu˒ -h -l-ul
key=230 span=1.7 ˛m l˒
0
2.097
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m d h h l d h d u b u h l u l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2
Time (s)

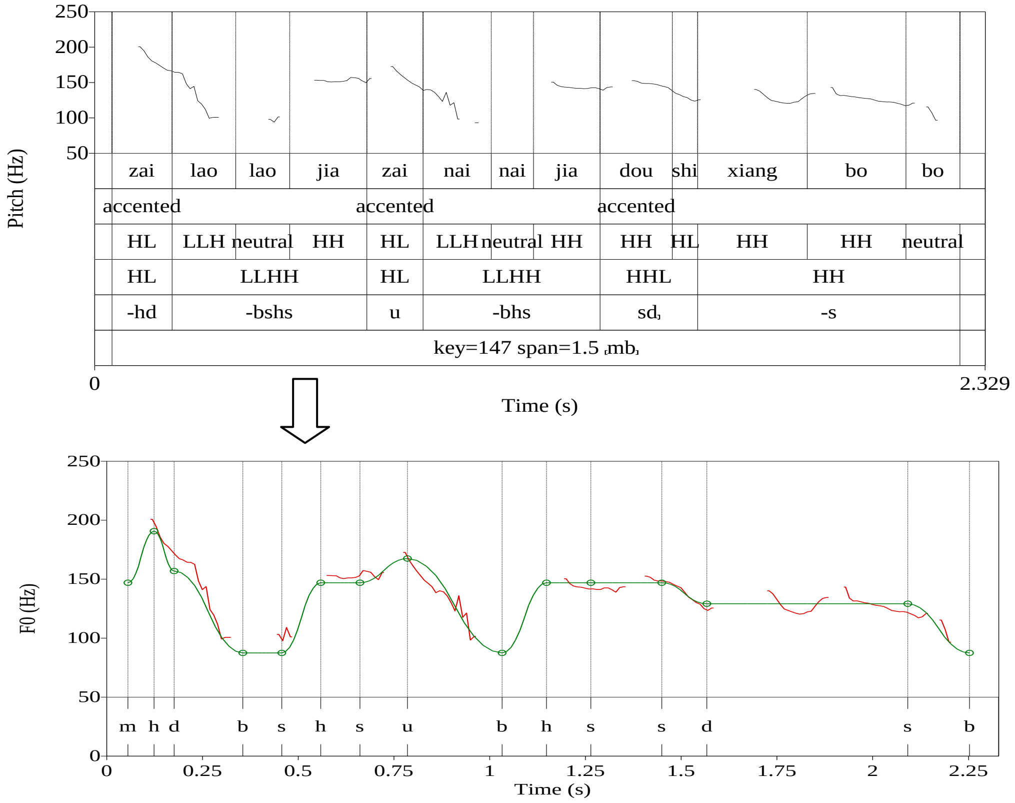
(4)


350
300
200
100
60
Pitch (Hz)
ta gen ta nei ge zhi zi yi kuair hui qu
accented accented accented
HH HH HH HL HL LH neutral LH HL LH HL
HH HH HH HL LH LH HL LHL
h - s˒ -l -l-h l hl -bub
key=215 span=1.5 ˛mb˒
0
2.384
Time (s)
350
300
250
200
150
100
50
0
F0 (Hz)
m h s l l h l h l b u b b
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25
Time (s)

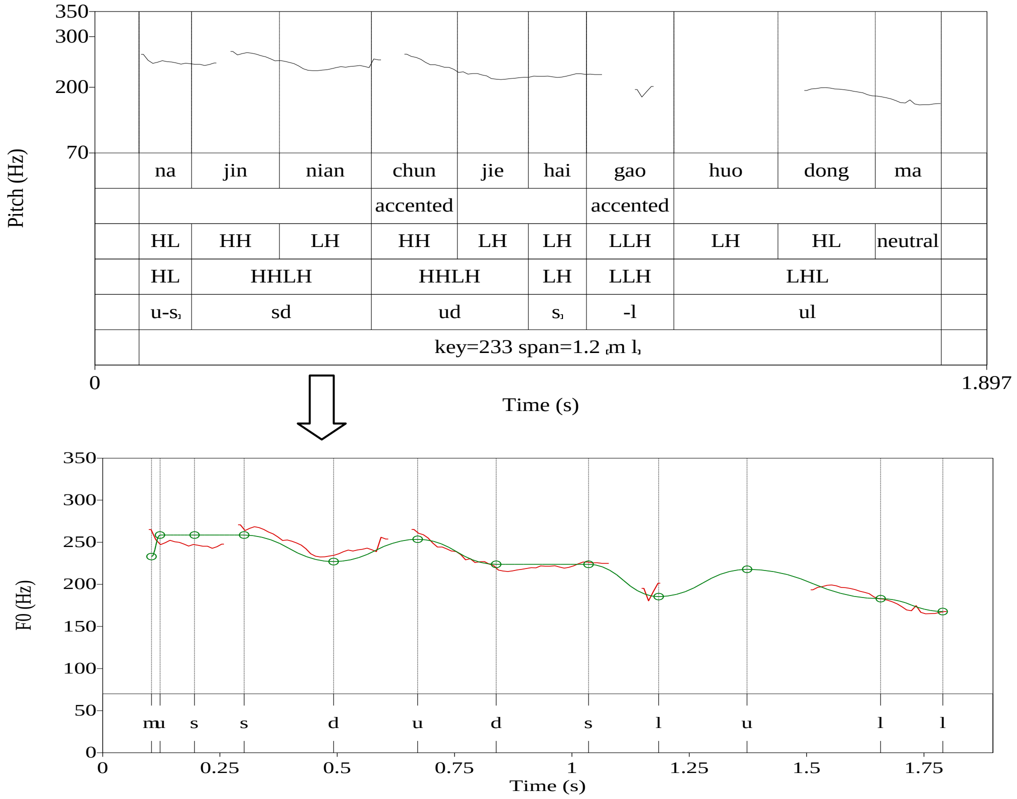
(5)


250
200
150
100
50
Pitch (Hz)
gao zhong dao da xue tong yang ye shi yi ge zhuan bian
accented accented accented
HH HH HL HL LH LH HL LLH HL LH HL LLH HL
HH HL HLH LHL LL HL LHL LLHL
h-s, -l ul-u, -lh- b sul -l
key=160 span=1.2 ,ml,
0
2.239
Time (s)
F0 (Hz)
m h s l u l u l h b s u l l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2
Time (s)

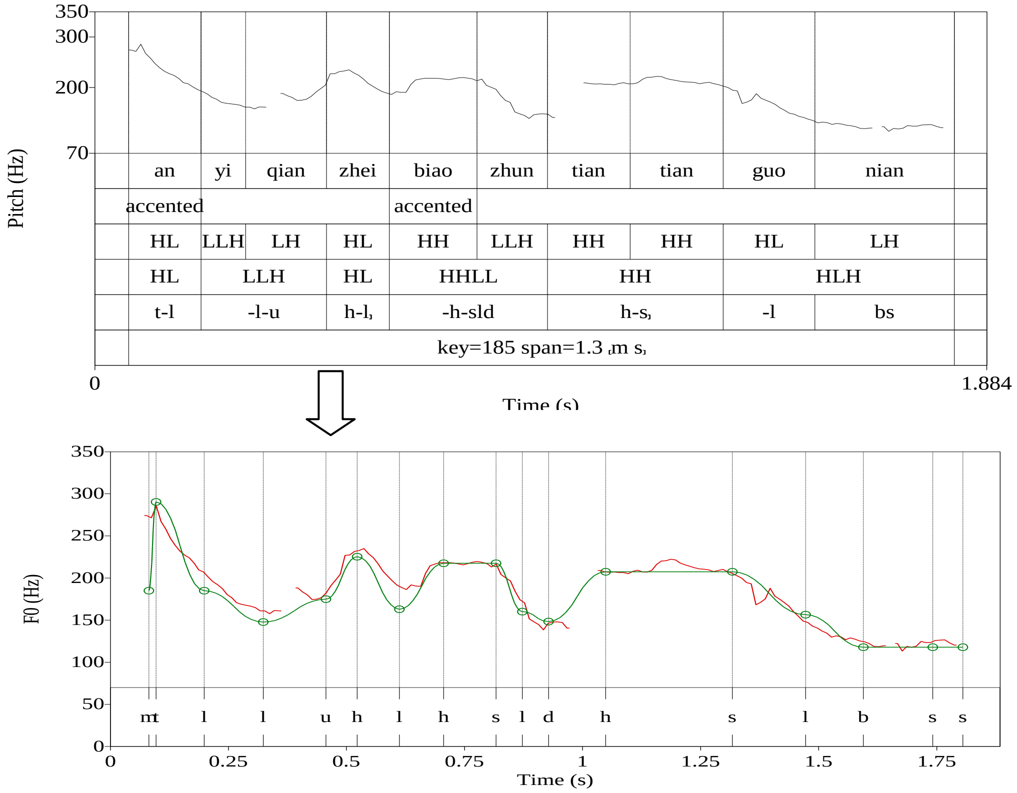
(6)


250
200
150
100
50
Pitch (Hz)
zai lao lao jia zai nai nai jia dou shi xiang bo bo
accented accented accented
HL LLH neutral HH HL LLH neutral HH HH HL HH HH neutral
HL LLHH HL LLHH HHL HH
-hd -bshs u -bhs sd, -s
key=147 span=1.5 ,mb,
0
2.329
Time (s)
F0 (Hz)
m h d b s h s u b h s s d s b
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25
Time (s)

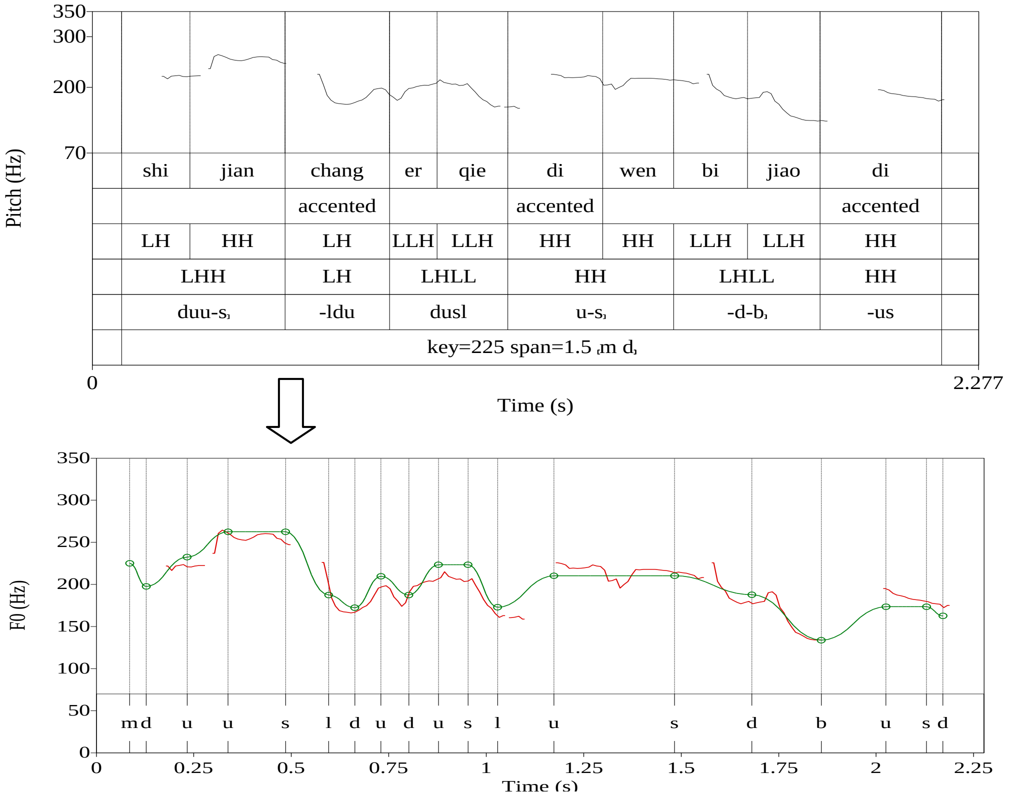
(7)


Pitch (Hz)
350
300
200
70
na
jin
nian
chun
jie
hai
gao
huo
dong
ma
accented
accented
HL
HH
LH
HH
LH
LH
LLH
LH
HL
neutral
HL
HHLH
HHLH
LH
LLH
LHL
u-s,
sd
ud
s,
-l
ul
key=233 span=1.2 ,m l,
0
1.897
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
mu
s
s
d
u
d
s
l
u
l
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

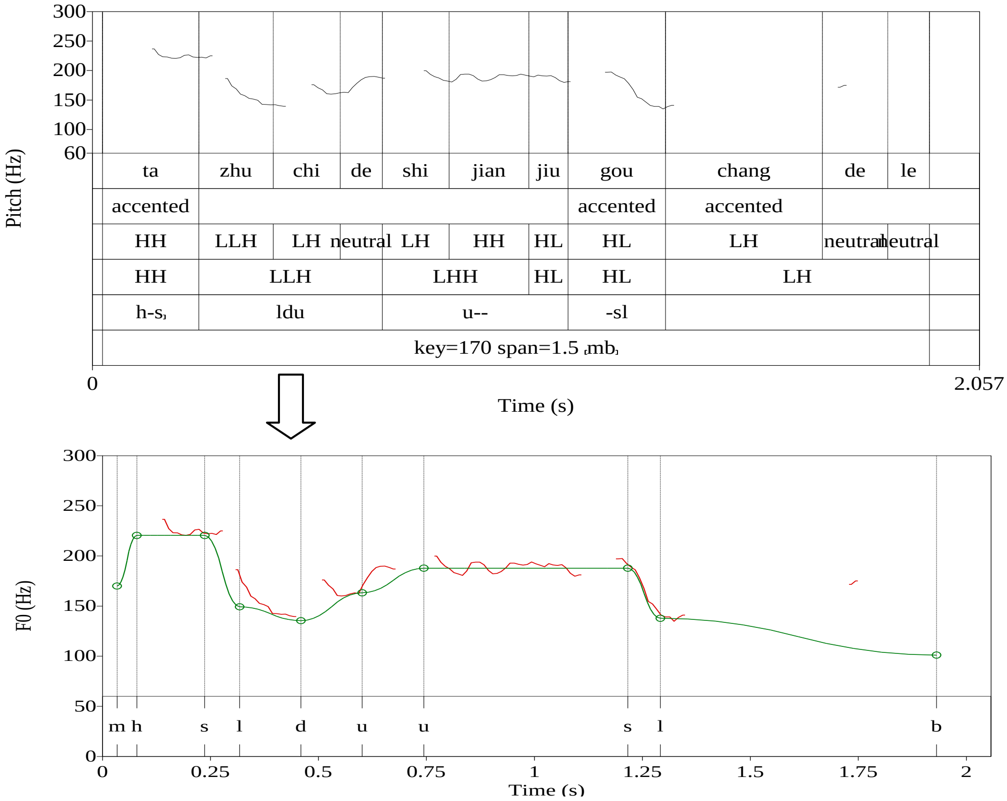
(8)


Pitch (Hz)
350
300
200
70
an
yi
qian
zhei
biao
zhun
tian
tian
guo
nian
accented
accented
HL
LLH
LH
HL
HH
LLH
HH
HH
HL
LH
HL
LLH
HL
HHLL
HH
HLH
t-l
-l-u
h-l,
-h-sld
h-s,
-l
bs
key=185 span=1.3 ,m s,
0
1.884
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
mt
l
l
u
h
l
h
s
l
d
h
s
l
b
s
s
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

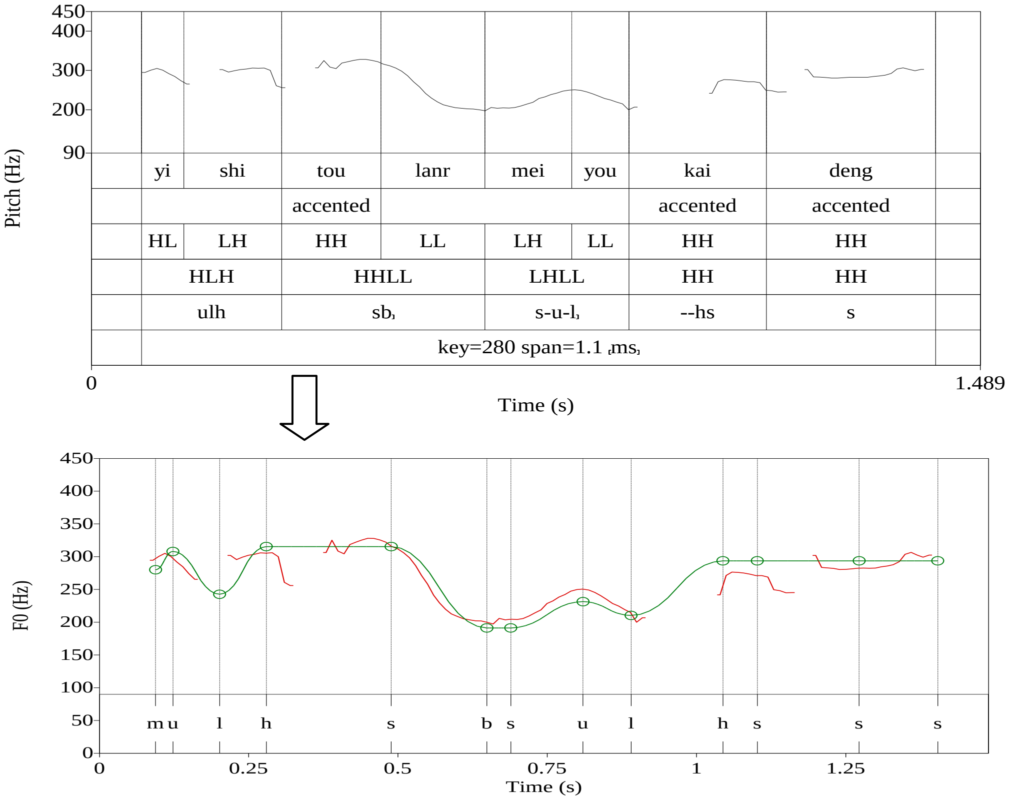
(9)


Pitch (Hz)
350
300
200
70
shi jian chang er qie di wen bi jiao di
accented accented accented
LH HH LH LLH LLH HH HH LLH LLH HH
LHH LH LHLL HH LHLL HH
duu-sˌ -ldu dusl u-sˌ -d-bˌ -us
key=225 span=1.5 ˌm dˌ
0
2.277
Time (s)
F0 (Hz)
350 300 250 200 150 100 50 0
m d u u s l d u d u s l u s d b u s d
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25
Time (s)

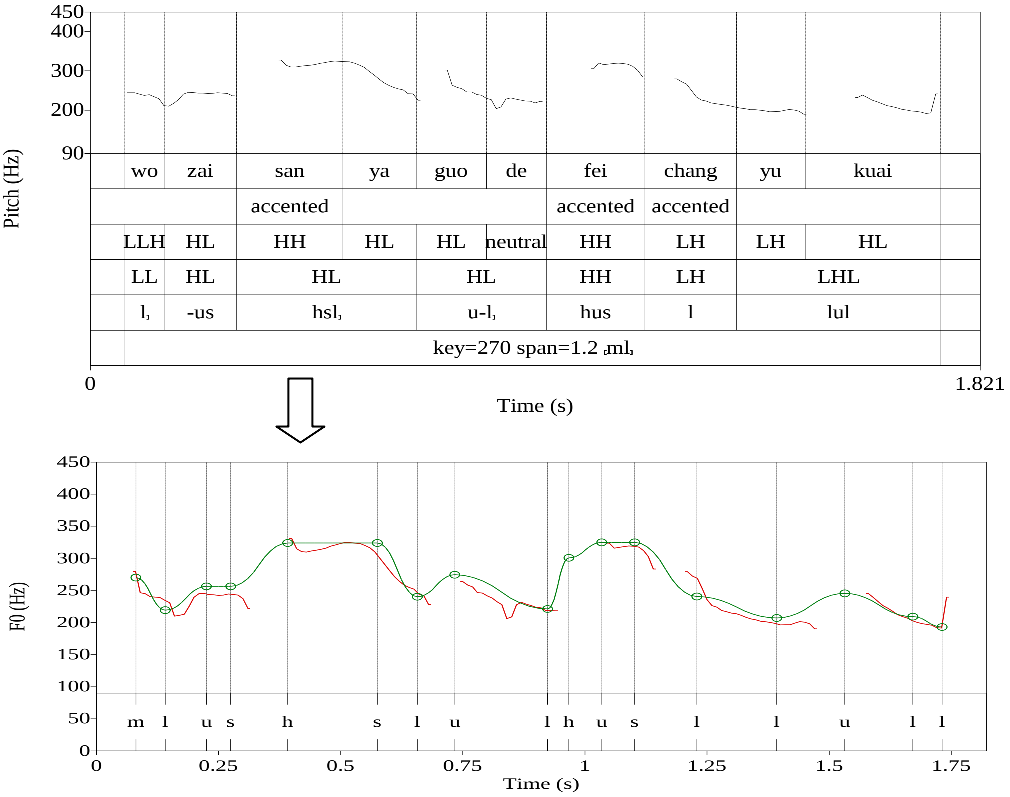
(10)


Pitch (Hz)
300
250
200
150
100
60
ta zhu chi de shi jian jiu gou chang de le
accented accented accented
HH LLH LH neutral LH HH HL HL LH neutral neutral
HH LLH LHH HL HL LH
h-sˌ ldu u-- -sl
key=170 span=1.5 ˌmbˌ
0
2.057
Time (s)
F0 (Hz)
300 250 200 150 100 50 0
m h s l d u u s l b
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2
Time (s)

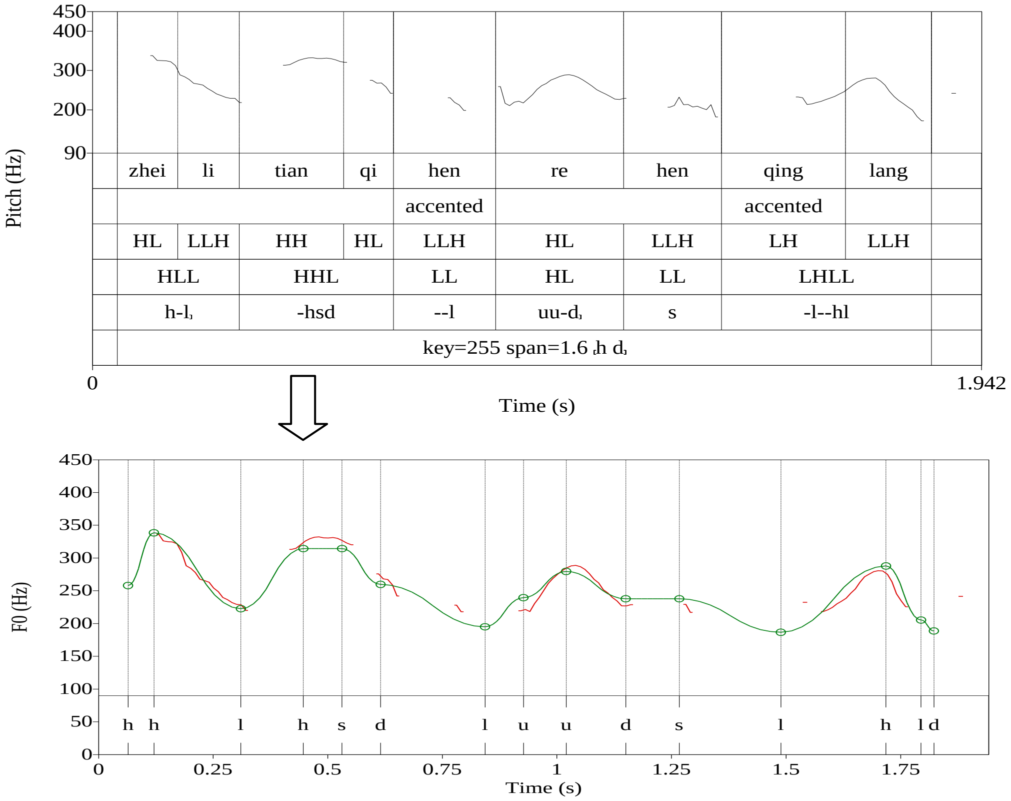
(11)


Pitch (Hz)
450
400
300
200
90
yi
shi
tou
lanr
mei
you
kai
deng
accented
accented
accented
HL
LH
HH
LL
LH
LL
HH
HH
HLH
HHLL
LHLL
HH
HH
ulh
sb˺
s-u-l˺
--hs
s
key=280 span=1.1 ˻ms˺
0
1.489
Time (s)
F0 (Hz)
450
400
350
300
250
200
150
100
50
0
m u
l
h
s
b
s
u
l
h
s
s
s
0
0.25
0.5
0.75
1
1.25
Time (s)

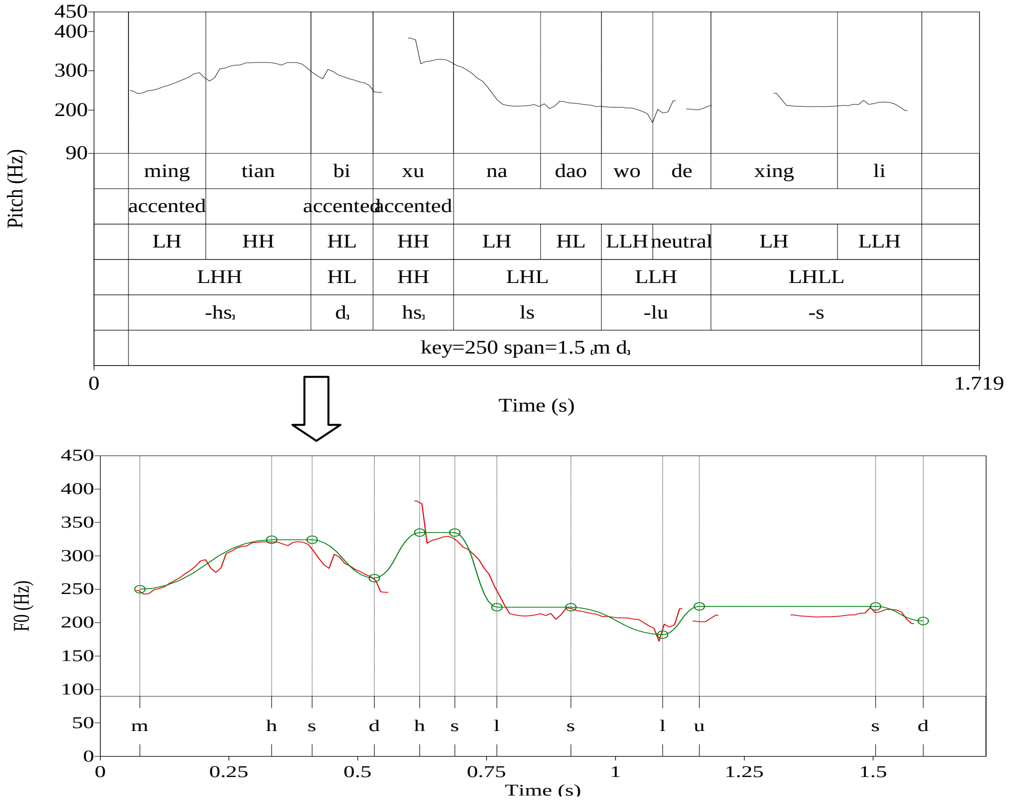
(12)


Pitch (Hz)
450
400
300
200
90
wo
zai
san
ya
guo
de
fei
chang
yu
kuai
accented
accented
accented
LLH
HL
HH
HL
HL
neutral
HH
LH
LH
HL
LL
HL
HL
HL
HH
LH
LHL
l˺
-us
hsl˺
u-l˺
hus
l
lul
key=270 span=1.2 ˻ml˺
0
1.821
Time (s)
F0 (Hz)
450
400
350
300
250
200
150
100
50
0
m
l
u
s
h
s
l
u
l
h
u
s
l
l
u
l
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

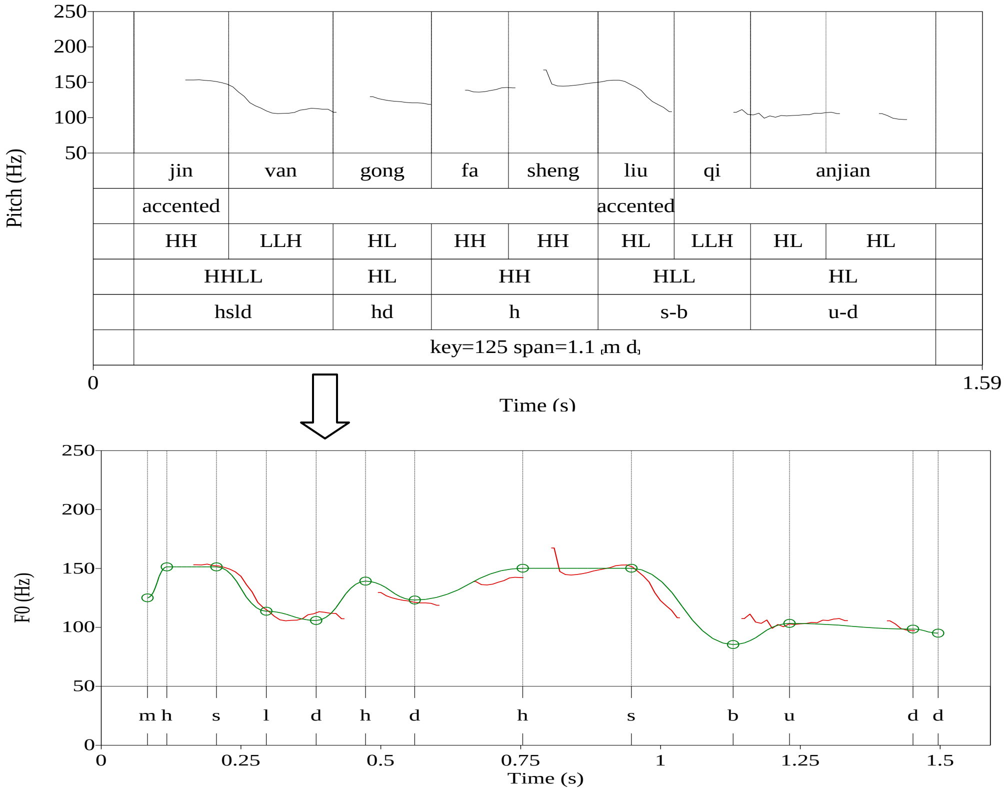
(13)


450
400
300
200
90
Pitch (Hz)
zhei
li
tian
qi
hen
re
hen
qing
lang
accented
accented
HL
LLH
HH
HL
LLH
HL
LLH
LH
LLH
HLL
HHL
LL
HL
LL
LHLL
h-lˌ
-hsd
--l
uu-dˌ
s
-l--hl
key=255 span=1.6 ˌh dˌ
0
1.942
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
h
h
l
h
s
d
l
u
u
d
s
l
h
l
d
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

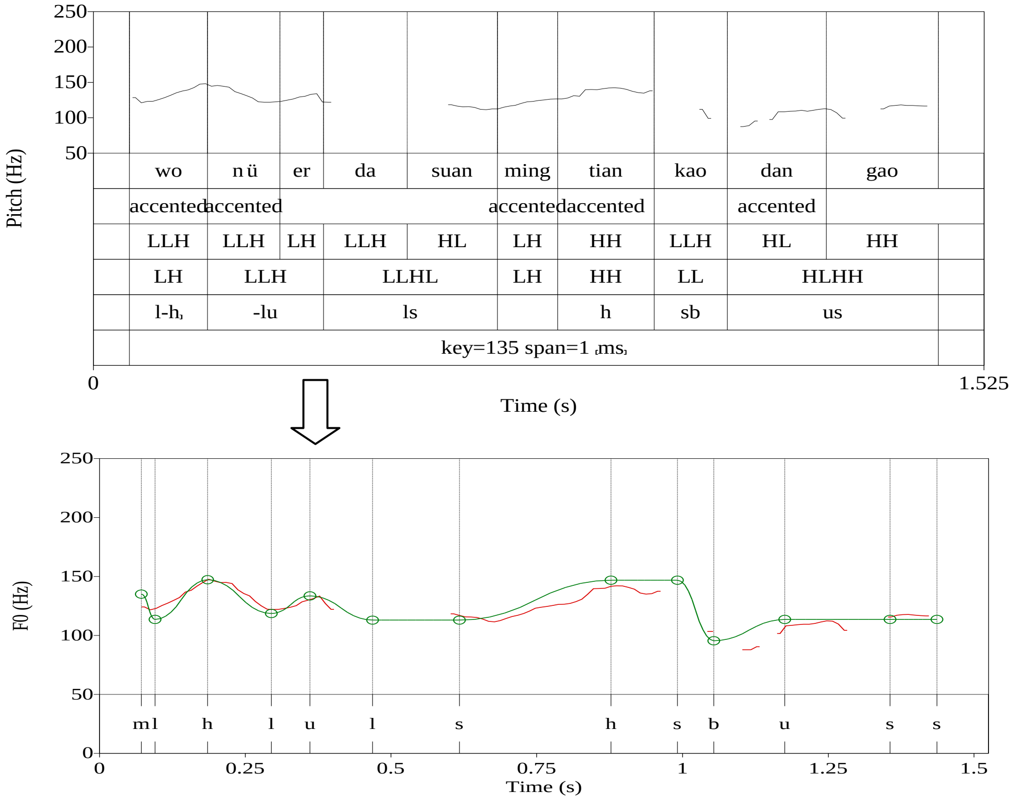
(14)


450
400
300
200
90
Pitch (Hz)
ming
tian
bi
xu
na
dao
wo
de
xing
li
accented
accented
accented
LH
HH
HL
HH
LH
HL
LLH
neutral
LH
LLH
LHH
HL
HH
LHL
LLH
LHLL
-hsˌ
dˌ
hsˌ
ls
-lu
-s
key=250 span=1.5 ˌm dˌ
0
1.719
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m
h
s
d
h
s
l
s
l
u
s
d
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

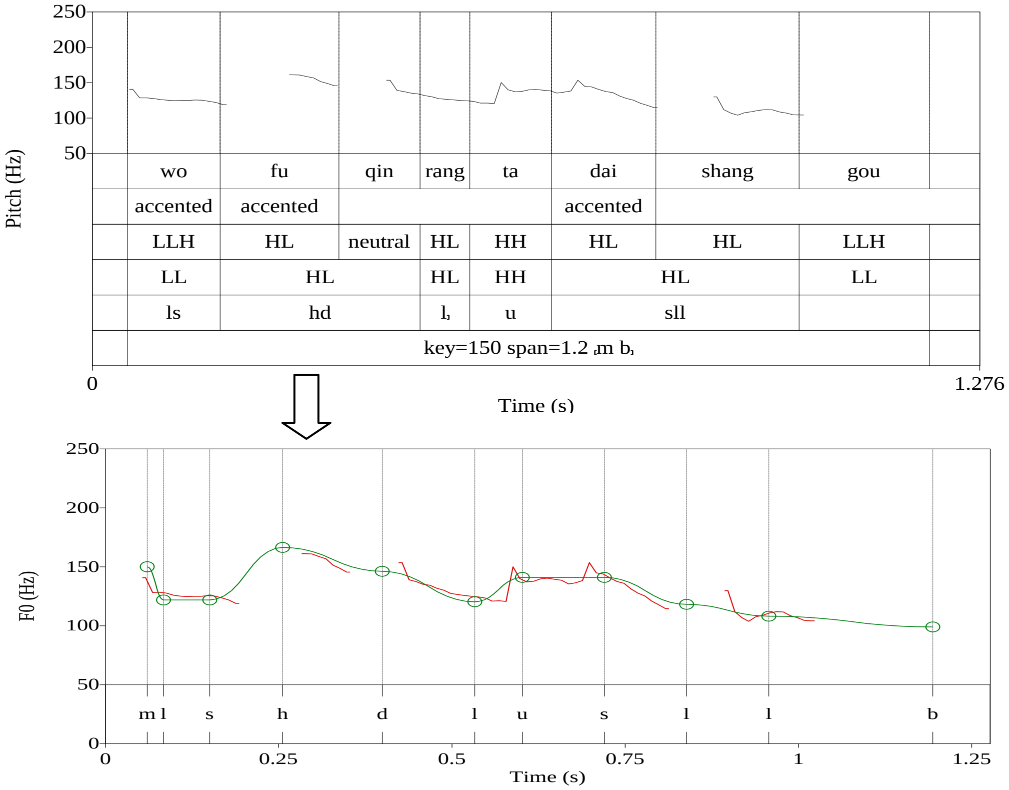
(15)


250
200
150
100
50
Pitch (Hz)
jin
van
gong
fa
sheng
liu
qi
anjian
accented
accented
HH
LLH
HL
HH
HH
HL
LLH
HL
HL
HHLL
HL
HH
HLL
HL
hsld
hd
h
s-b
u-d
key=125 span=1.1 ₍m d₎
0
1.59
Time (s)
250
200
150
100
50
0
F0 (Hz)
m h
s
l
d
h
d
h
s
b
u
d
d
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

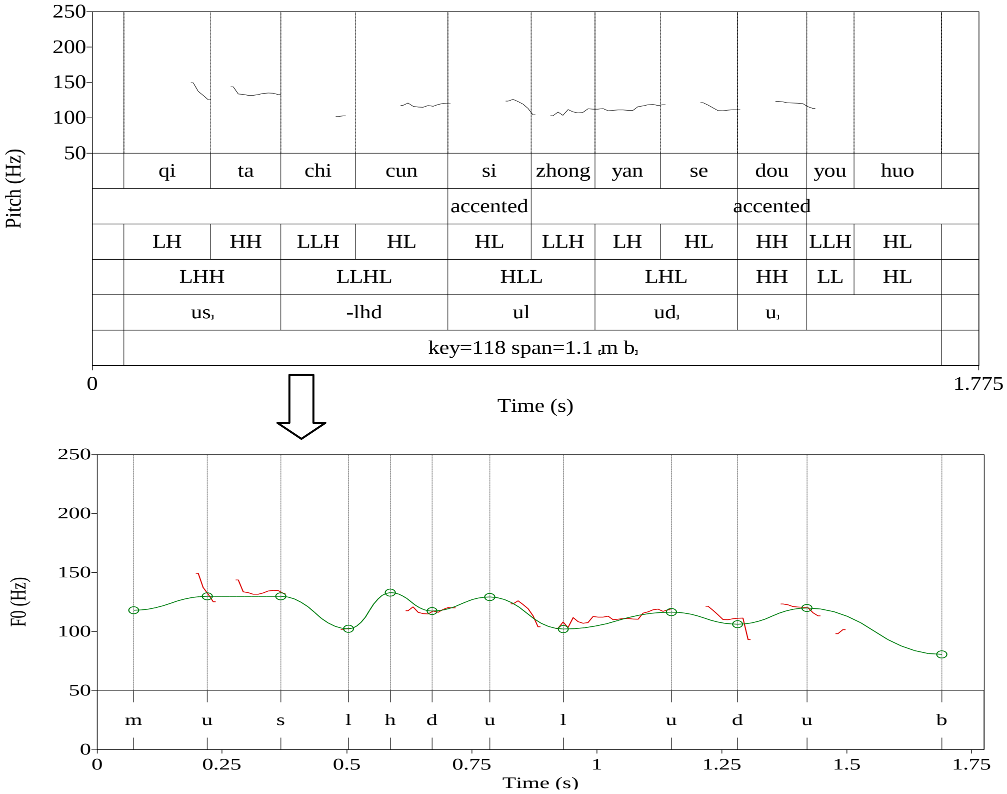
(16)


250
200
150
100
50
Pitch (Hz)
wo
nü
er
da
suan
ming
tian
kao
dan
gao
accented
accented
accented
accented
accented
LLH
LLH
LH
LLH
HL
LH
HH
LLH
HL
HH
LH
LLH
LLHL
LH
HH
LL
HLHH
l-h₎
-lu
ls
h
sb
us
key=135 span=1 ₍ms₎
0
1.525
Time (s)
250
200
150
100
50
0
F0 (Hz)
m l
h
l
u
l
s
h
s
b
u
s
s
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

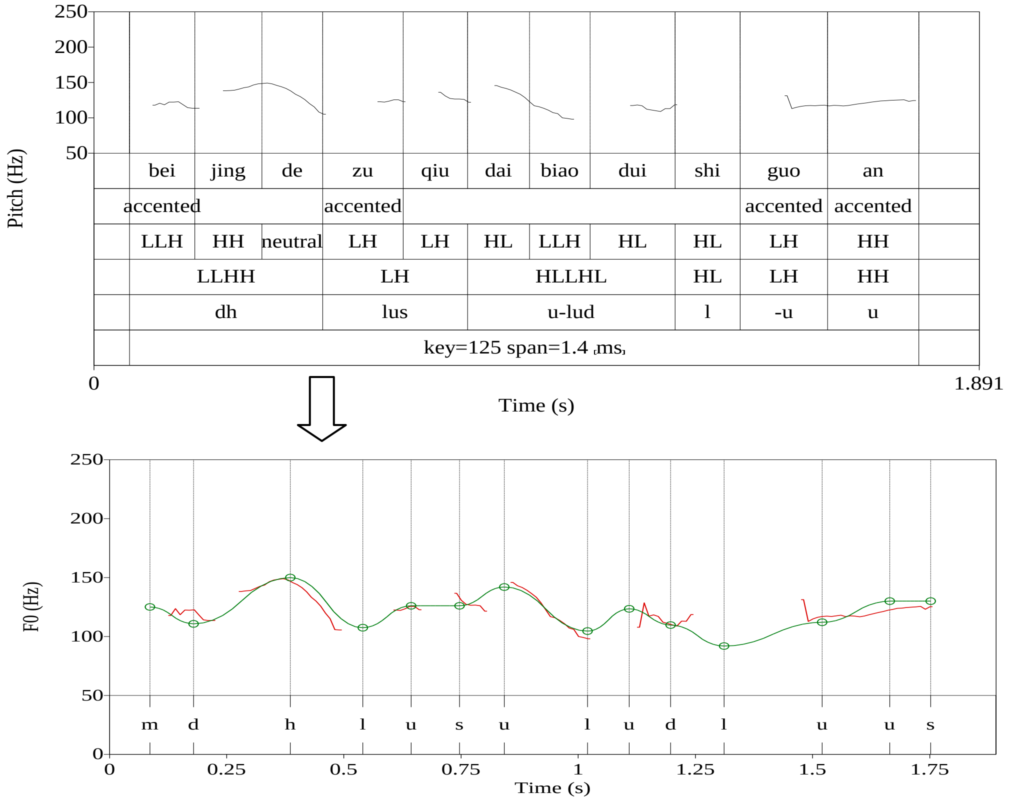
(17)


Pitch (Hz)
250
200
150
100
50
wo fu qin rang ta dai shang gou
accented accented accented
LLH HL neutral HL HH HL HL LLH
LL HL HL HH HL LL
ls hd l⌟ u sll
key=150 span=1.2 ⌞m b⌟
0
1.276
Time (s)
F0 (Hz)
250
200
150
100
50
0
m l s h d l u s l l b
0
0.25
0.5
0.75
1
1.25
Time (s)

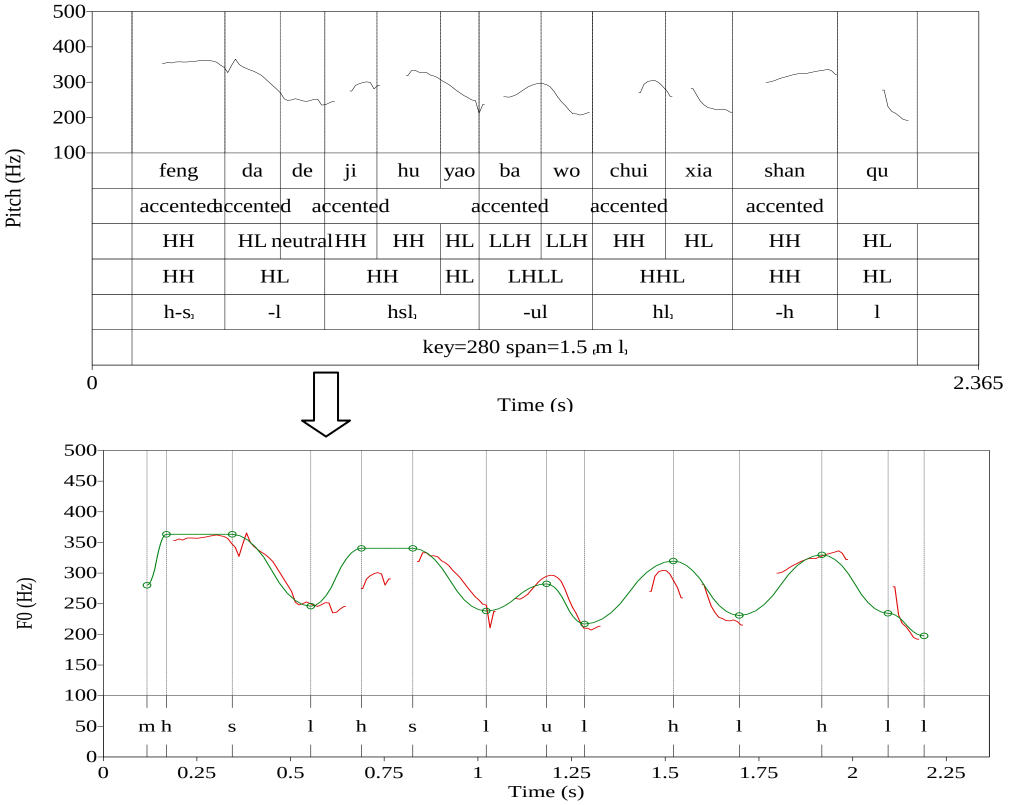
(18)


Pitch (Hz)
250
200
150
100
50
qi ta chi cun si zhong yan se dou you huo
accented accented
LH HH LLH HL HL LLH LH HL HH LLH HL
LHH LLHL HLL LHL HH LL HL
us⌟ -lhd ul ud⌟ u⌟
key=118 span=1.1 ⌞m b⌟
0
1.775
Time (s)
F0 (Hz)
250
200
150
100
50
0
m u s l h d u l u d u b
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

(19)

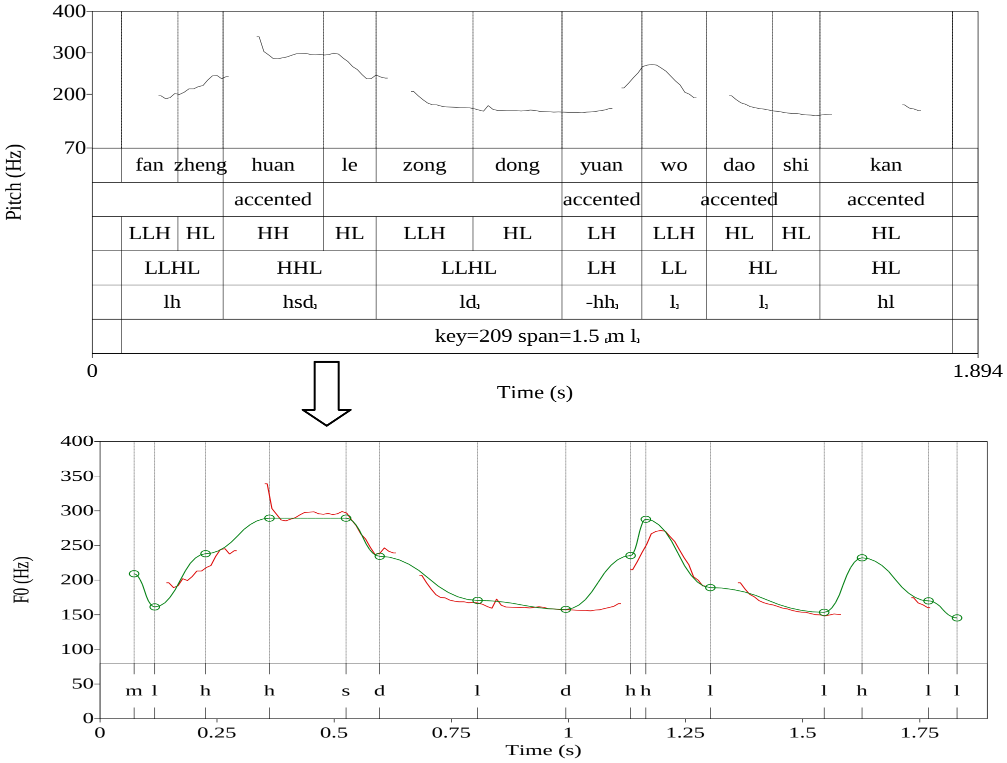

Pitch (Hz)
250
200
150
100
50
bei jing de zu qiu dai biao dui shi guo an
accented accented accented accented
LLH HH neutral LH LH HL LLH HL HL LH HH
LLHH LH HLLHL HL LH HH
dh lus u-lud l -u u
key=125 span=1.4 ˌmsˌ
0
1.891
Time (s)
F0 (Hz)
250
200
150
100
50
0
m d h l u s u l u d l u u s
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

(20)

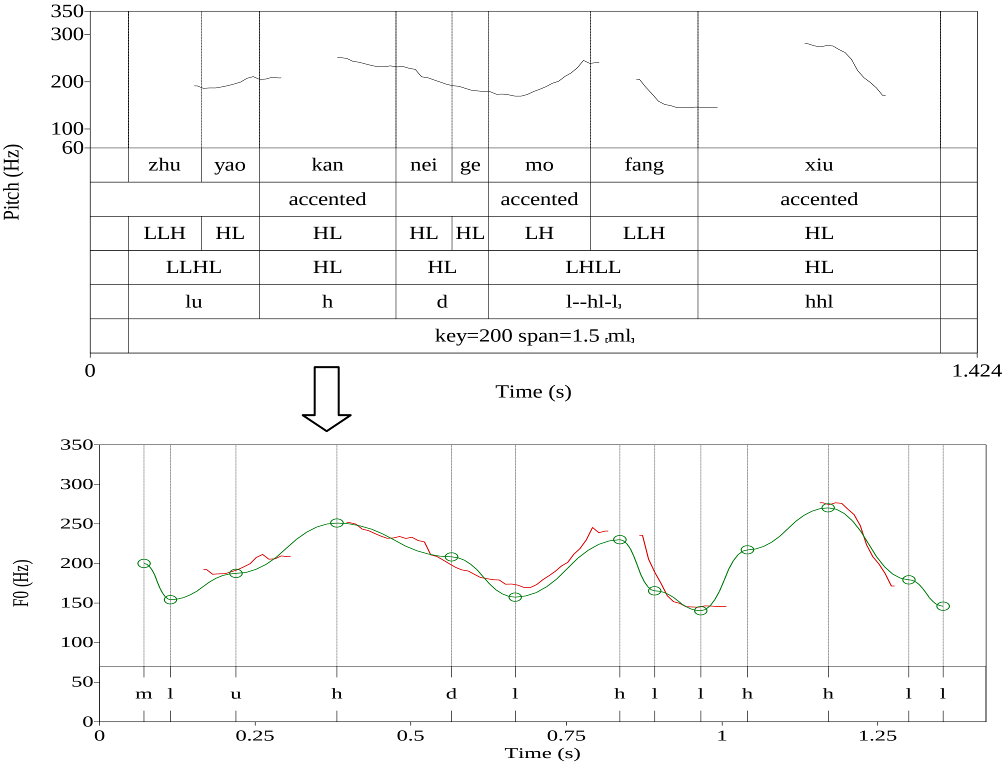

Pitch (Hz)
500
400
300
200
100
feng da de ji hu yao ba wo chui xia shan qu
accented accented accented accented accented accented
HH HL neutral HH HH HL LLH LLH HH HL HH HL
HH HL HH HL LHLL HHL HH HL
h-sˌ -l hslˌ -ul hlˌ -h l
key=280 span=1.5 ˌm lˌ
0
2.365
Time (s)
F0 (Hz)
500
450
400
350
300
250
200
150
100
50
0
m h s l h s l u l h l h l l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
2.25
Time (s)

List B

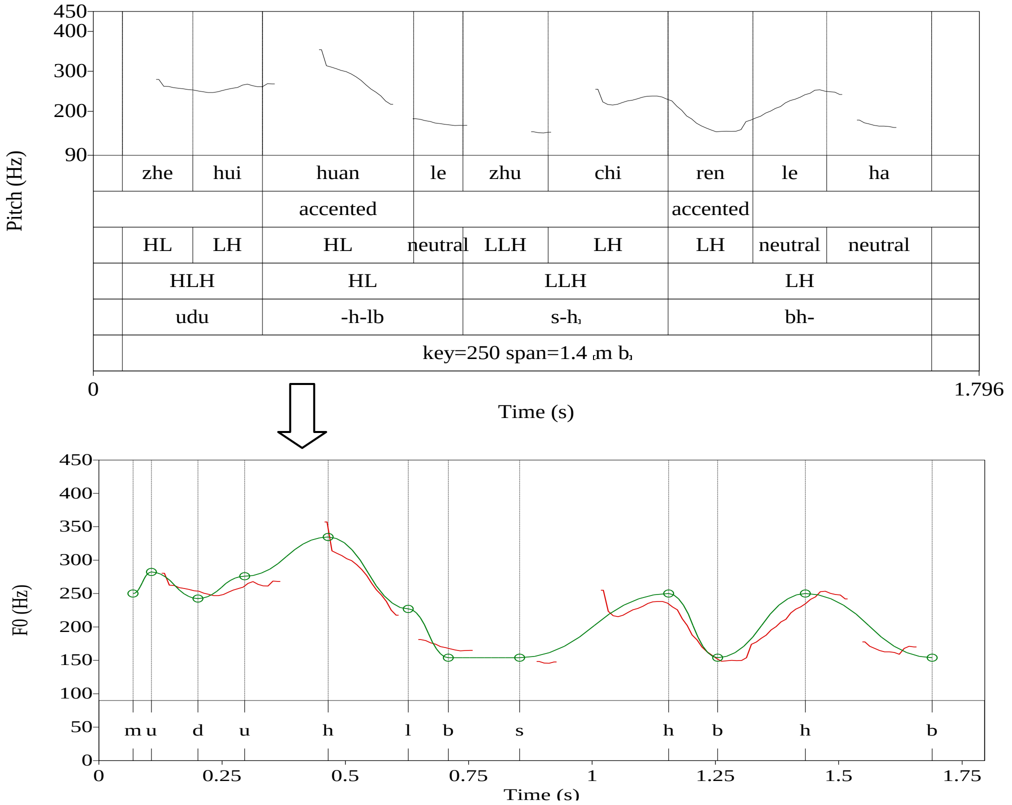
(21)


Pitch (Hz)
400
300
200
70
fan zheng huan le zong dong yuan wo dao shi kan
accented accented accented accented
LLH HL HH HL LLH HL LH LLH HL HL HL
LLHL HHL LLHL LH LL HL HL
lh hsd, ld, -hh, l, l, hl
key=209 span=1.5 ,m l,
0
1.894
Time (s)
F0 (Hz)
400
350
300
250
200
150
100
50
0
m l h h s d l d h h l l h l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

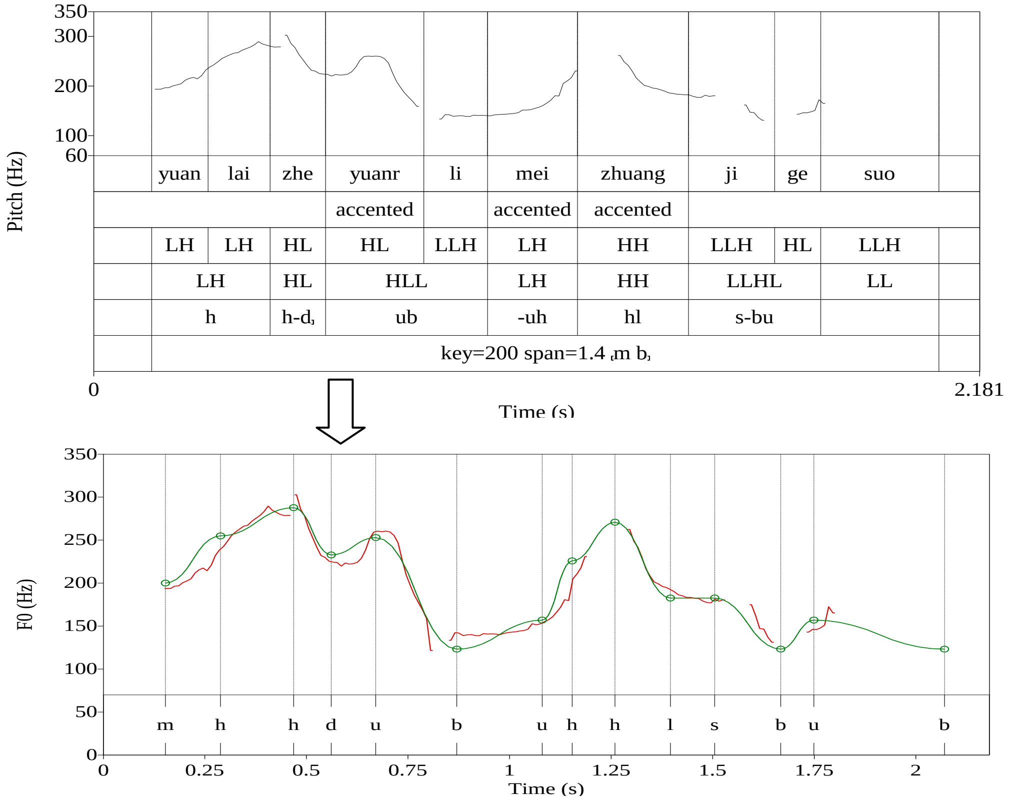
(22)


Pitch (Hz)
350
300
200
100
60
zhu yao kan nei ge mo fang xiu
accented accented accented
LLH HL HL HL HL LH LLH HL
LLHL HL HL LHLL HL
lu h d l--hl-l, hhl
key=200 span=1.5 ,ml,
0
1.424
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
m l u h d l h l l h h l l
0 0.25 0.5 0.75 1 1.25
Time (s)

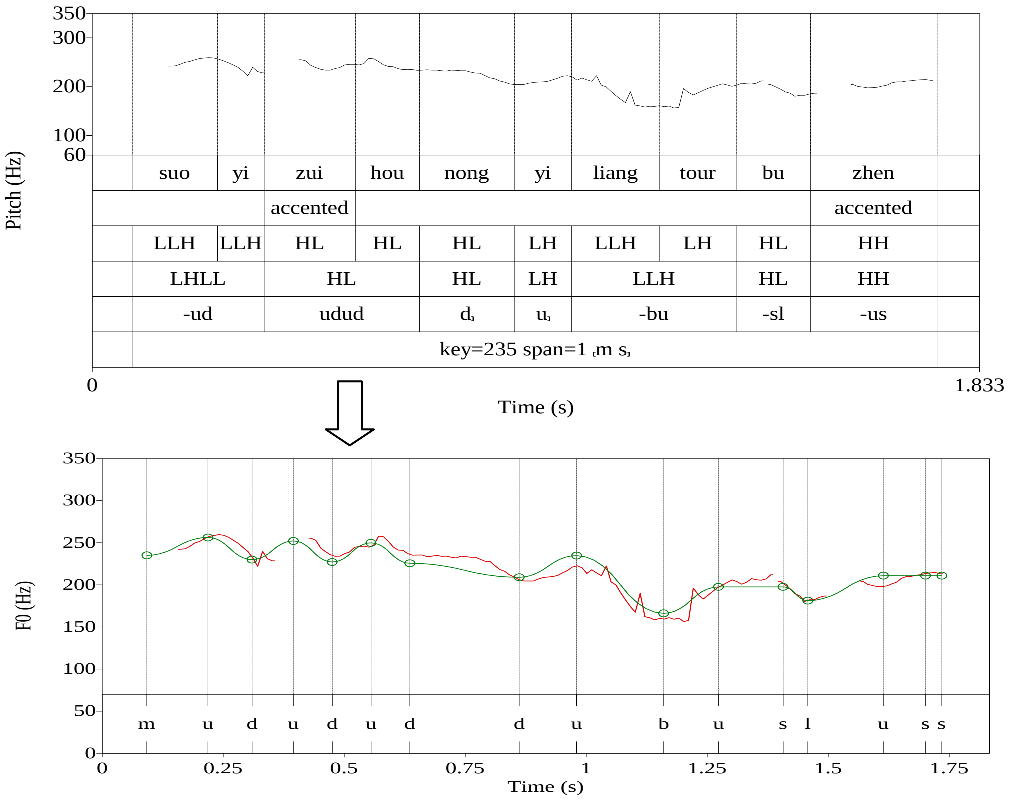
(23)


450
400
300
200
90
Pitch (Hz)
zhe
hui
huan
le
zhu
chi
ren
le
ha
accented
accented
HL
LH
HL
neutral
LLH
LH
LH
neutral
neutral
HLH
HL
LLH
LH
udu
-h-lb
s-h⌟
bh-
key=250 span=1.4 ⌞m b⌟
0
1.796
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m
u
d
u
h
l
b
s
h
b
h
b
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

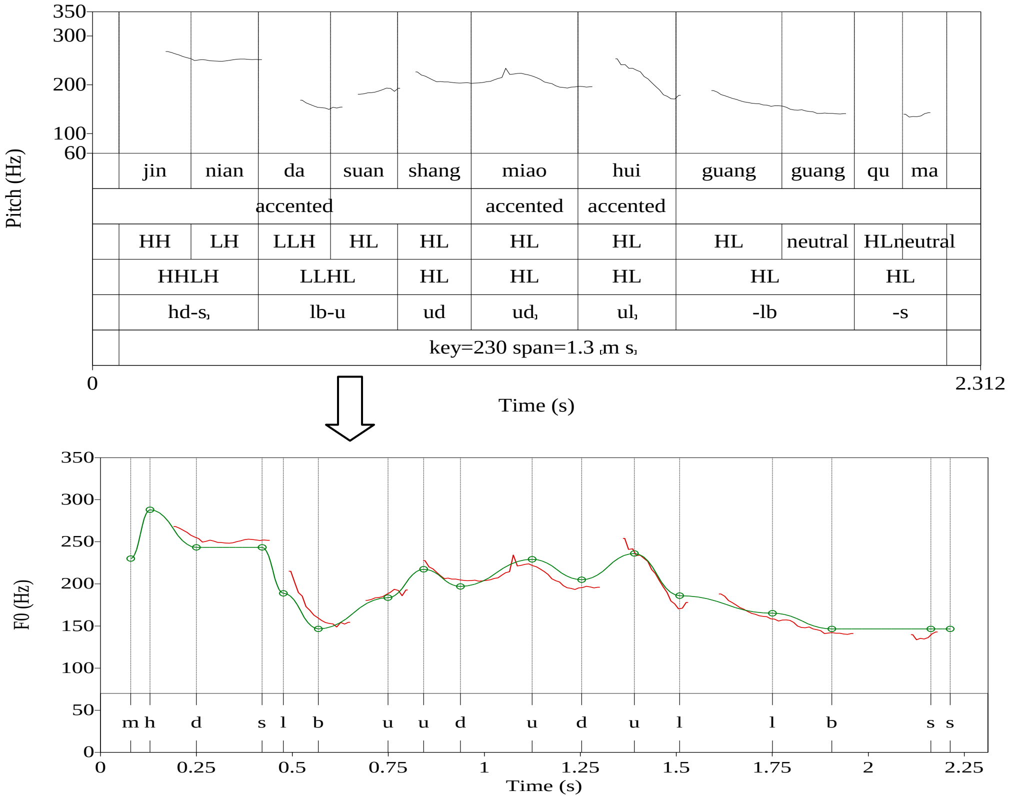
(24)


350
300
200
100
60
Pitch (Hz)
yuan
lai
zhe
yuanr
li
mei
zhuang
ji
ge
suo
accented
accented
accented
LH
LH
HL
HL
LLH
LH
HH
LLH
HL
LLH
LH
HL
HLL
LH
HH
LLHL
LL
h
h-d⌟
ub
-uh
hl
s-bu
key=200 span=1.4 ⌞m b⌟
0
2.181
Time (s)
350
300
250
200
150
100
50
0
F0 (Hz)
m
h
h
d
u
b
u
h
h
l
s
b
u
b
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
Time (s)

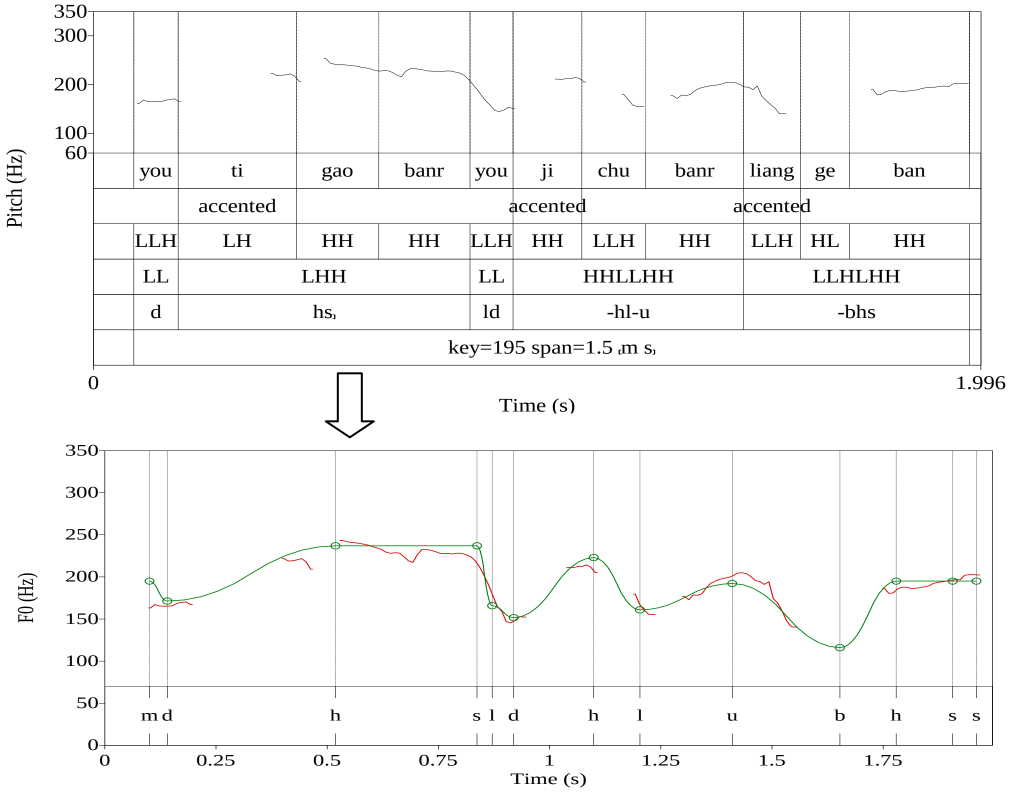
(25)


Pitch (Hz)
350
300
200
100
60
suo yi zui hou nong yi liang tour bu zhen
accented
accented
LLH LLH HL HL HL LH LLH LH HL HH
LHLL HL HL LH LLH HL HH
-ud udud d, u, -bu -sl -us
key=235 span=1 ˌm s,
0
1.833
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
m u d u d u d d u b u s l u s s
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

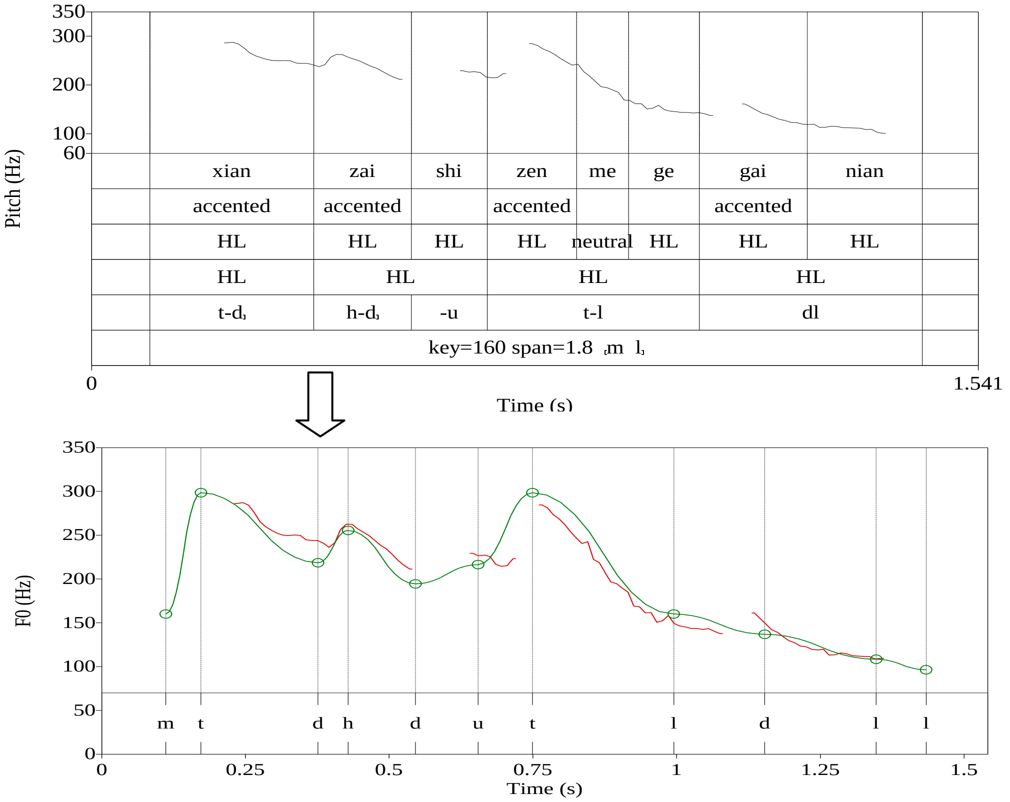
(26)


Pitch (Hz)
350
300
200
100
60
jin nian da suan shang miao hui guang guang qu ma
accented
accented
accented
HH LH LLH HL HL HL HL HL neutral HLneutral
HHLH LLHL HL HL HL HL HL
hd-s, lb-u ud ud, ul, -lb -s
key=230 span=1.3 ˌm s,
0
2.312
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
m h d s l b u u d u d u l l b s s
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25
Time (s)

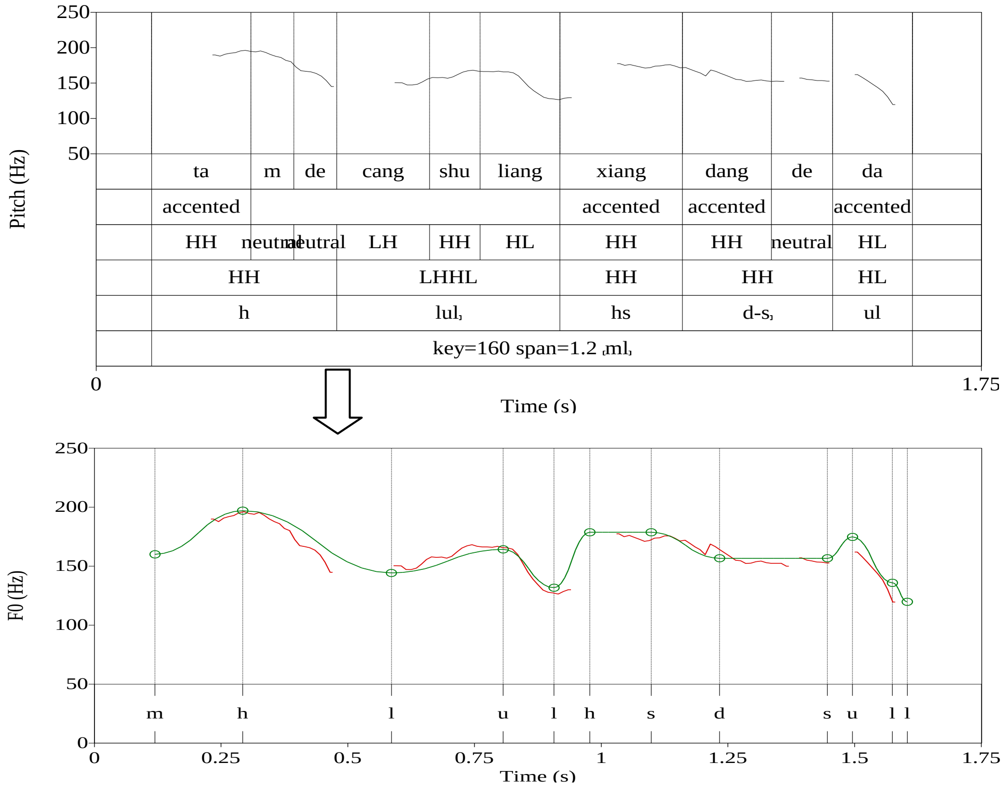
(27)


Pitch (Hz)
350
300
200
100
60
you ti gao banr you ji chu banr liang ge ban
accented accented accented
LLH LH HH HH LLH HH LLH HH LLH HL HH
LL LHH LL HHLLHH LLHLHH
d hs˒ ld -hl-u -bhs
key=195 span=1.5 ˌm s˒
0
1.996
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
m d h s l d h l u b h s s
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

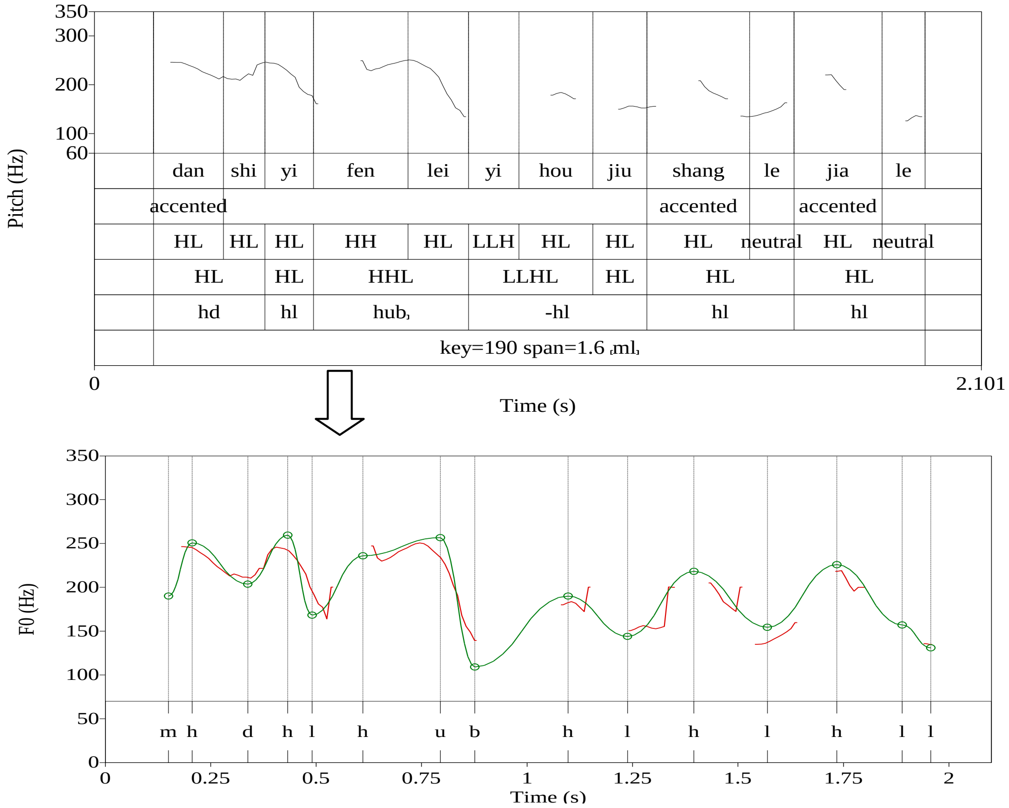
(28)


Pitch (Hz)
350
300
200
100
60
xian zai shi zen me ge gai nian
accented accented accented accented
HL HL HL HL neutral HL HL HL
HL HL HL HL
t-d˒ h-d˒ -u t-l dl
key=160 span=1.8 ˌm l˒
0
1.541
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
m t d h d u t l d l l
0 0.25 0.5 0.75 1 1.25 1.5
Time (s)

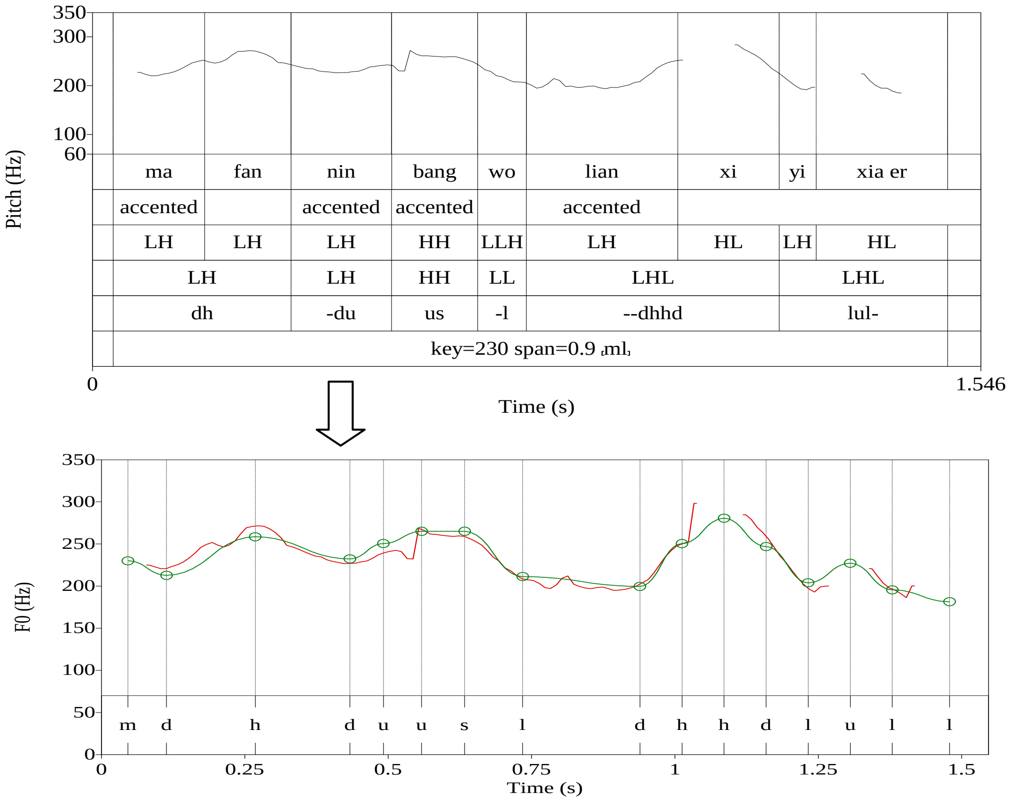
(29)


Pitch (Hz)
250
200
150
100
50
ta m de cang shu liang xiang dang de da
accented accented accented accented
HH neutral neutral LH HH HL HH HH neutral HL
HH LHHL HH HH HL
h lul, hs d-s, ul
key=160 span=1.2 ,ml,
0
1.75
Time (s)
F0 (Hz)
250
200
150
100
50
0
m h l u l h s d s u l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

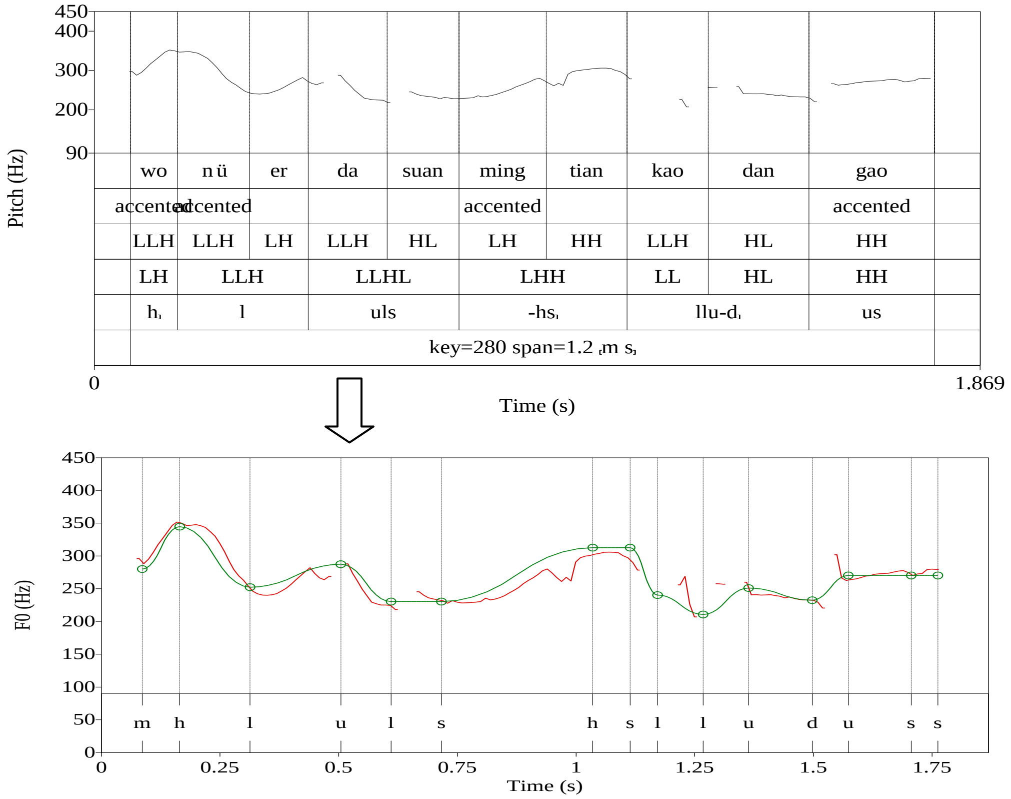
(30)


Pitch (Hz)
350
300
200
100
60
dan shi yi fen lei yi hou jiu shang le jia le
accented accented accented
HL HL HL HH HL LLH HL HL HL neutral HL neutral
HL HL HHL LLHL HL HL HL
hd hl hub, -hl hl hl
key=190 span=1.6 ,ml,
0
2.101
Time (s)
F0 (Hz)
350
300
250
200
150
100
50
0
m h d h l h u b h l h l h l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2
Time (s)

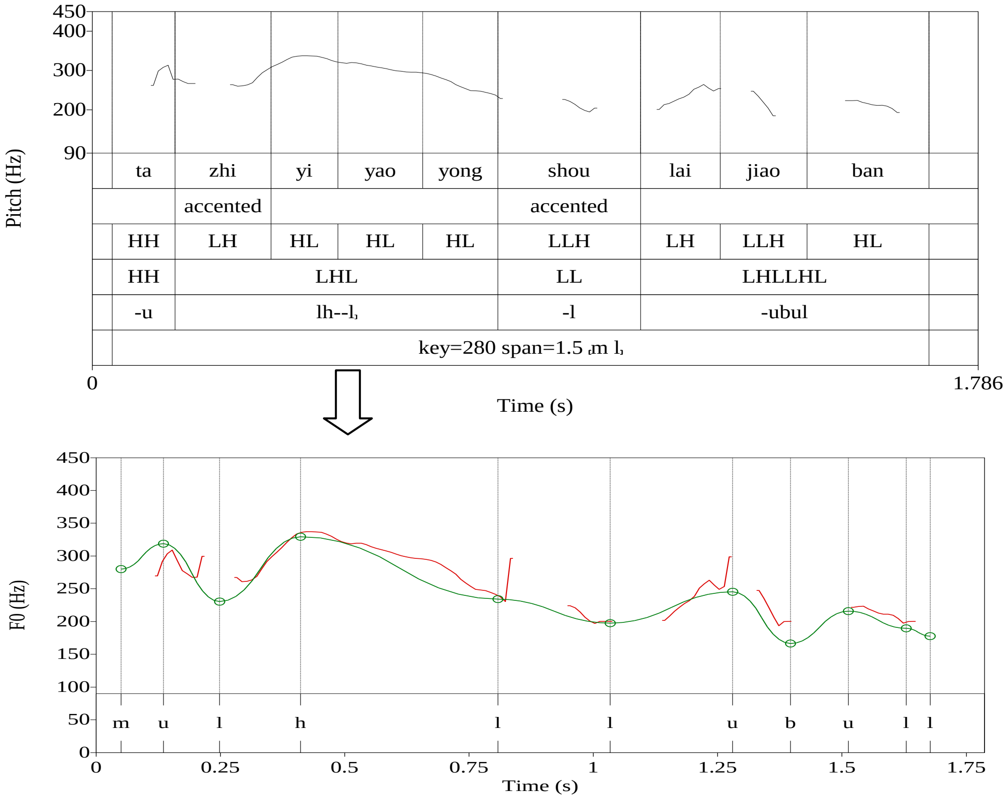
(31)


Pitch (Hz)
350 300 200 100 60
ma fan nin bang wo lian xi yi xia er
accented accented accented accented
LH LH LH HH LLH LH HL LH HL
LH LH HH LL LHL LHL
dh -du us -l --dhhd lul-
key=230 span=0.9 ,ml,
0 1.546
Time (s)
F0 (Hz)
350 300 250 200 150 100 50 0
m d h d u u s l d h h d l u l l
0 0.25 0.5 0.75 1 1.25 1.5
Time (s)

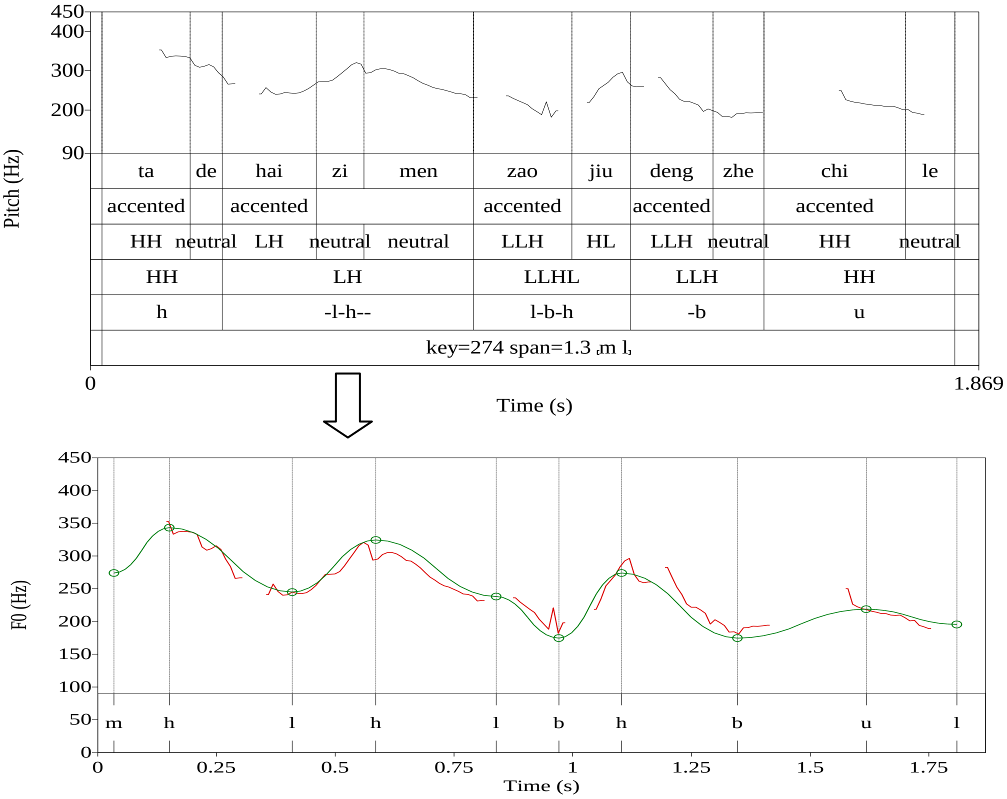
(32)


Pitch (Hz)
450 400 300 200 90
wo nü er da suan ming tian kao dan gao
accented accented accented accented
LLH LLH LH LLH HL LH HH LLH HL HH
LH LLH LLHL LHH LL HL HH
h, l uls -hs, llu-d, us
key=280 span=1.2 ,m s,
0 1.869
Time (s)
F0 (Hz)
450 400 350 300 250 200 150 100 50 0
m h l u l s h s l l u d u s s
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

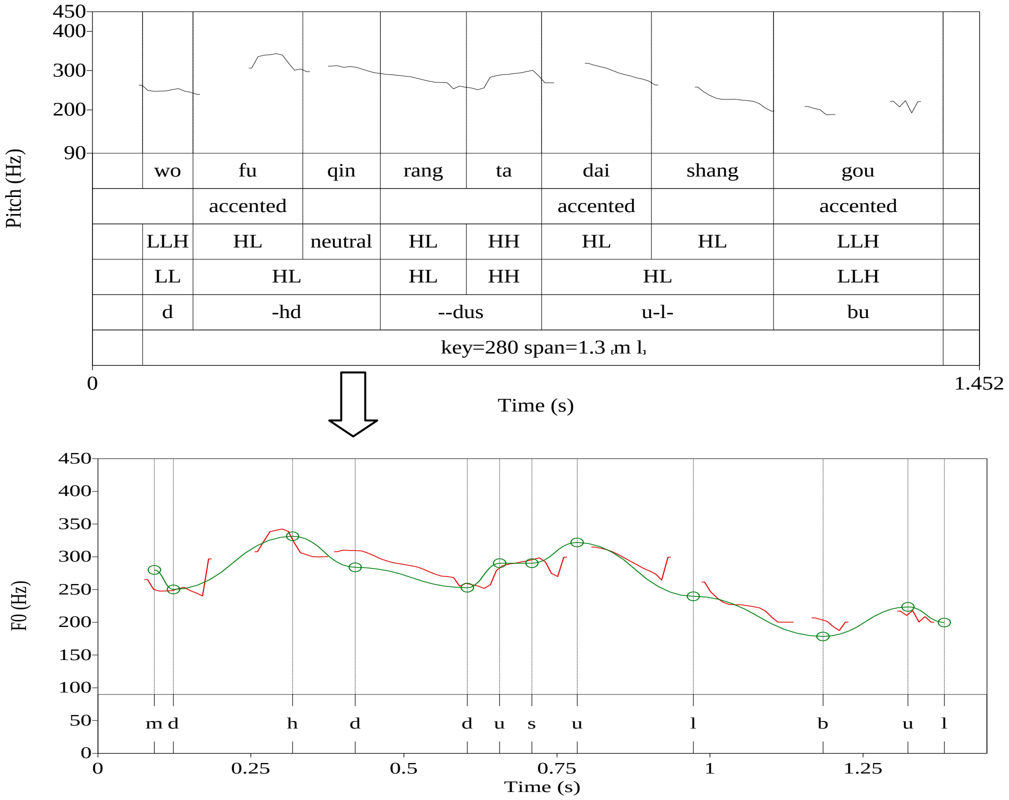
(33)


450
400
300
200
90
Pitch (Hz)
ta
zhi
yi
yao
yong
shou
lai
jiao
ban
accented
accented
HH
LH
HL
HL
HL
LLH
LH
LLH
HL
HH
LHL
LL
LHLLHL
-u
lh--l˼
-l
-ubul
key=280 span=1.5 ˻m l˼
0
1.786
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m
u
l
h
l
l
u
b
u
l
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

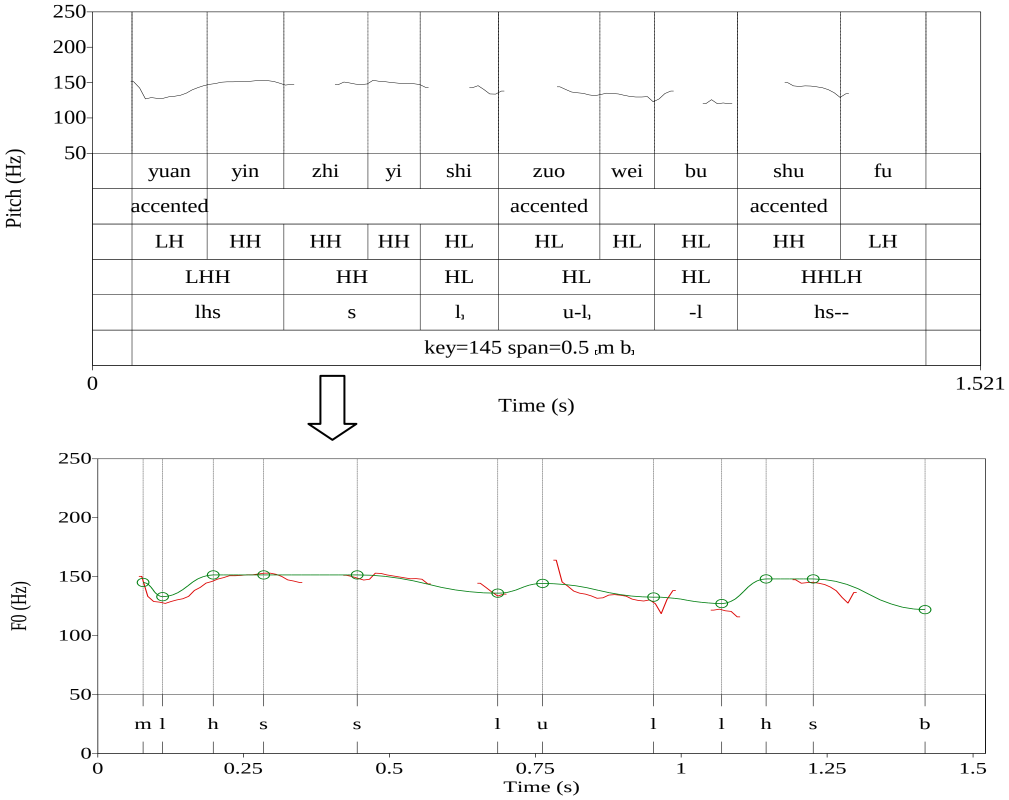
(34)


450
400
300
200
90
Pitch (Hz)
ta
de
hai
zi
men
zao
jiu
deng
zhe
chi
le
accented
accented
accented
accented
accented
HH
neutral
LH
neutral
neutral
LLH
HL
LLH
neutral
HH
neutral
HH
LH
LLHL
LLH
HH
h
-l-h--
l-b-h
-b
u
key=274 span=1.3 ˻m l˼
0
1.869
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m
h
l
h
l
b
h
b
u
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

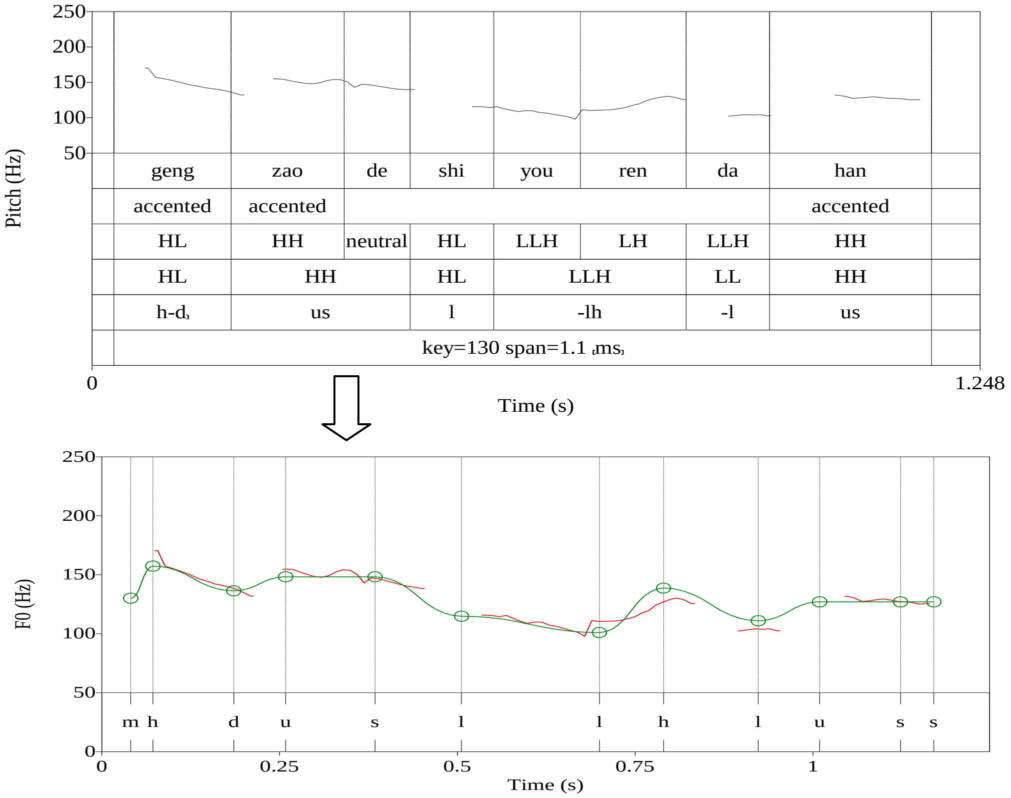
(35)


450
400
300
200
90
Pitch (Hz)
wo
fu
qin
rang
ta
dai
shang
gou
accented
accented
accented
LLH
HL
neutral
HL
HH
HL
HL
LLH
LL
HL
HL
HH
HL
LLH
d
-hd
--dus
u-l-
bu
key=280 span=1.3 ˌm lˌ
0
1.452
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m d
h
d
d
u
s
u
l
b
u
l
0
0.25
0.5
0.75
1
1.25
Time (s)

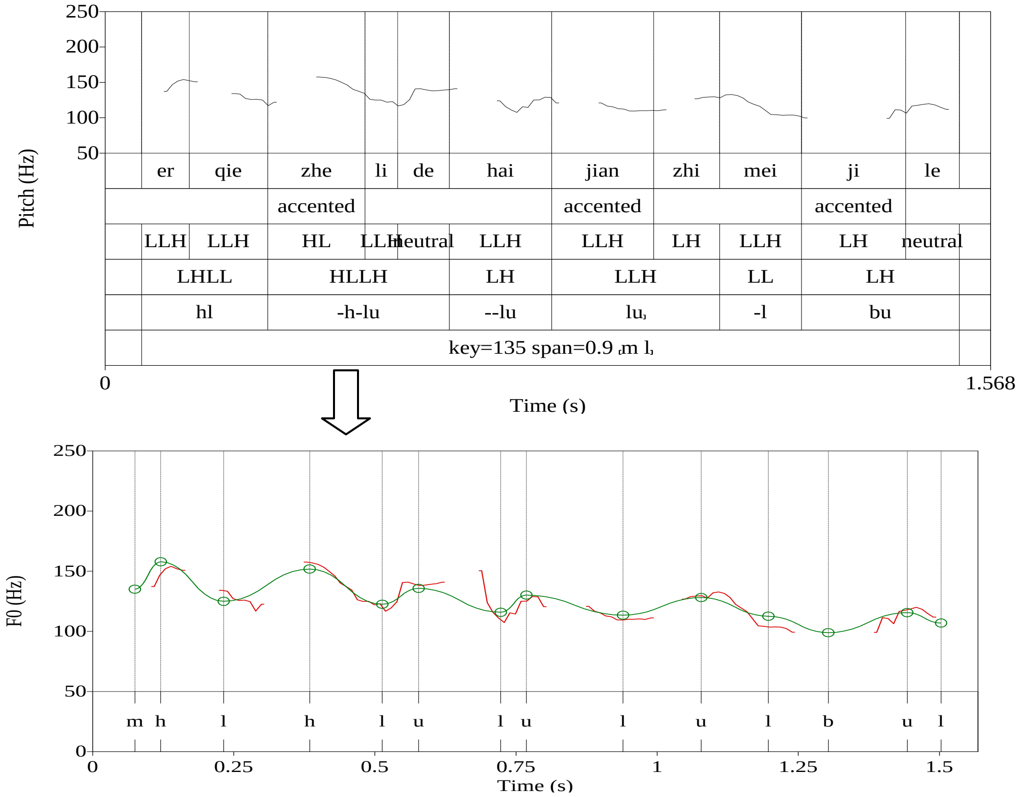
(36)


250
200
150
100
50
Pitch (Hz)
yuan
yin
zhi
yi
shi
zuo
wei
bu
shu
fu
accented
accented
accented
LH
HH
HH
HH
HL
HL
HL
HL
HH
LH
LHH
HH
HL
HL
HL
HHLH
lhs
s
lˌ
u-lˌ
-l
hs--
key=145 span=0.5 ˌm bˌ
0
1.521
Time (s)
250
200
150
100
50
0
F0 (Hz)
m l
h
s
s
l
u
l
l
h
s
b
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

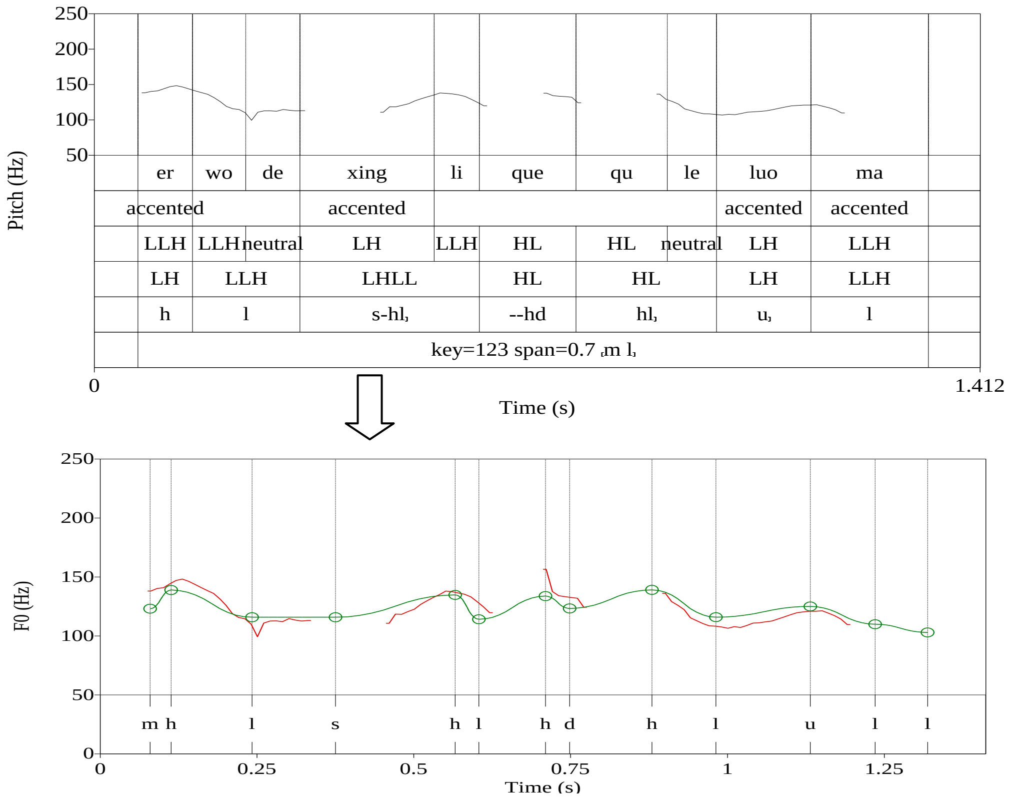
(37)


250
200
150
100
50
Pitch (Hz)
geng
zao
de
shi
you
ren
da
han
accented
accented
accented
HL
HH
neutral
HL
LLH
LH
LLH
HH
HL
HH
HL
LLH
LL
HH
h-dˌ
us
l
-lh
-l
us
key=130 span=1.1 ˌmsˌ
0
1.248
Time (s)
250
200
150
100
50
0
F0 (Hz)
m
h
d
u
s
l
l
h
l
u
s
s
0
0.25
0.5
0.75
1
Time (s)

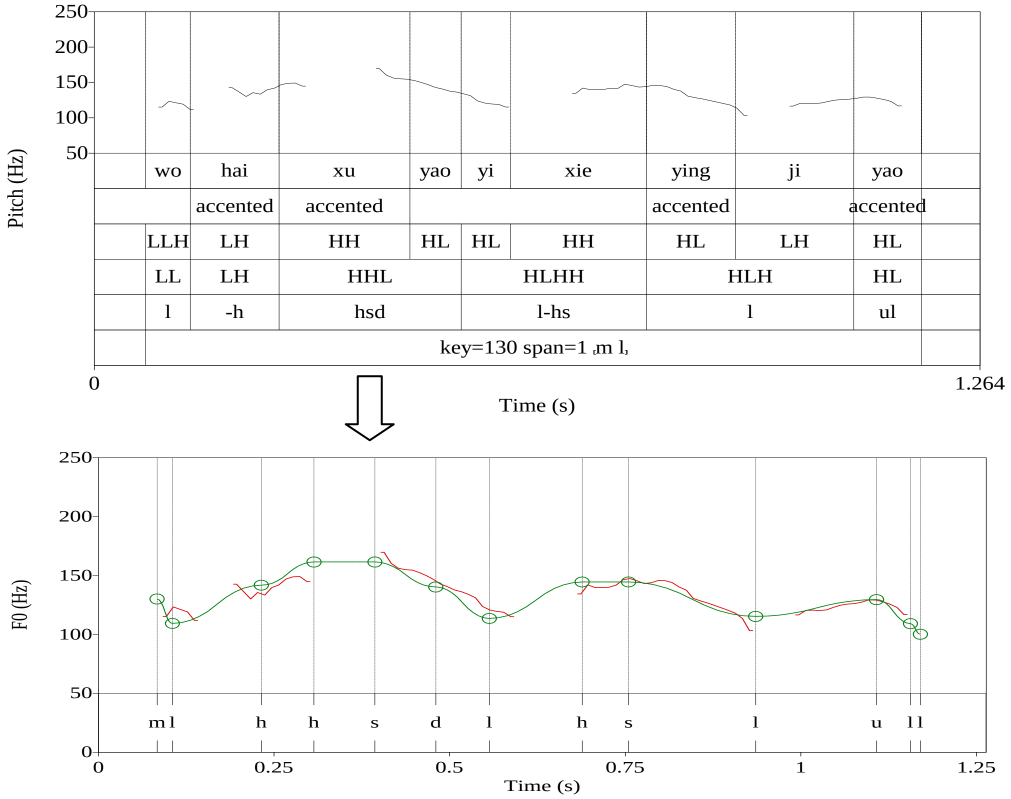
(38)


250
200
150
100
50
Pitch (Hz)
er
qie
zhe
li
de
hai
jian
zhi
mei
ji
le
accented
accented
accented
LLH
LLH
HL
LLH
neutral
LLH
LLH
LH
LLH
LH
neutral
LHLL
HLLH
LH
LLH
LL
LH
hl
-h-lu
--lu
luˌ
-l
bu
key=135 span=0.9 ˌm lˌ
0
1.568
Time (s)
250
200
150
100
50
0
F0 (Hz)
m
h
l
h
l
u
l
u
l
u
l
b
u
l
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

(39)

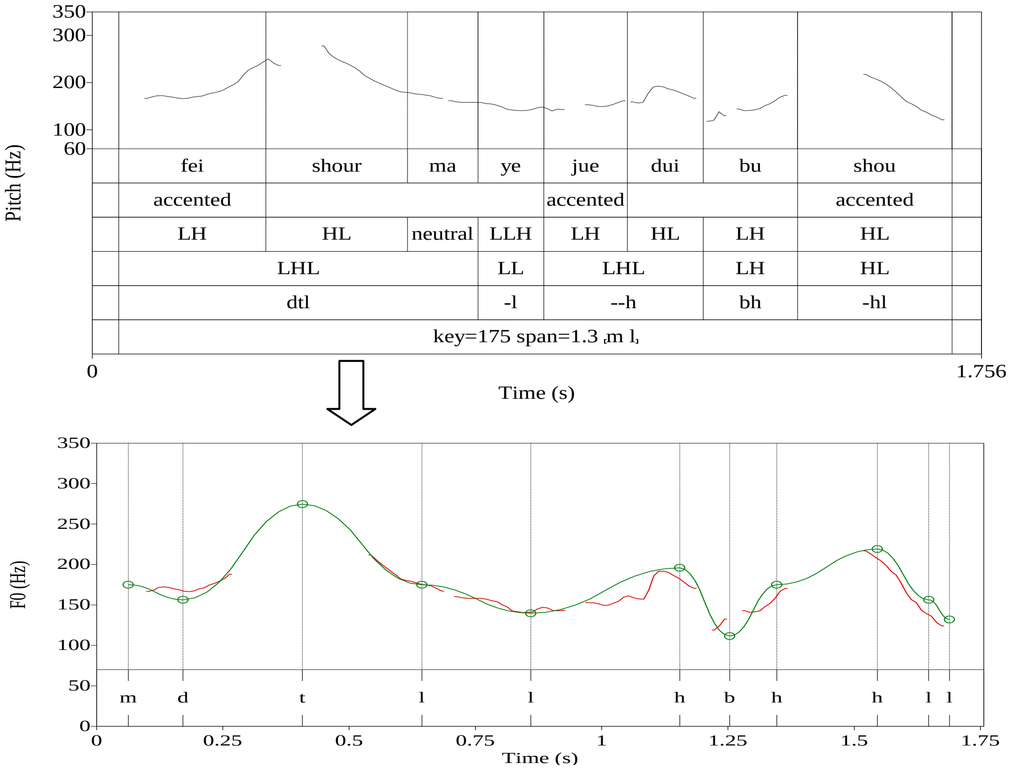

250
200
150
100
50
Pitch (Hz)
er
wo
de
xing
li
que
qu
le
luo
ma
accented
accented
accented
accented
LLH
LLH
neutral
LH
LLH
HL
HL
neutral
LH
LLH
LH
LLH
LHLL
HL
HL
LH
LLH
h
l
s-hl˒
--hd
hl˒
u˒
l
key=123 span=0.7 ˓m l˒
0
1.412
Time (s)
250
200
150
100
50
0
F0 (Hz)
m h
l
s
h
l
h
d
h
l
u
l
l
0
0.25
0.5
0.75
1
1.25
Time (s)

(40)

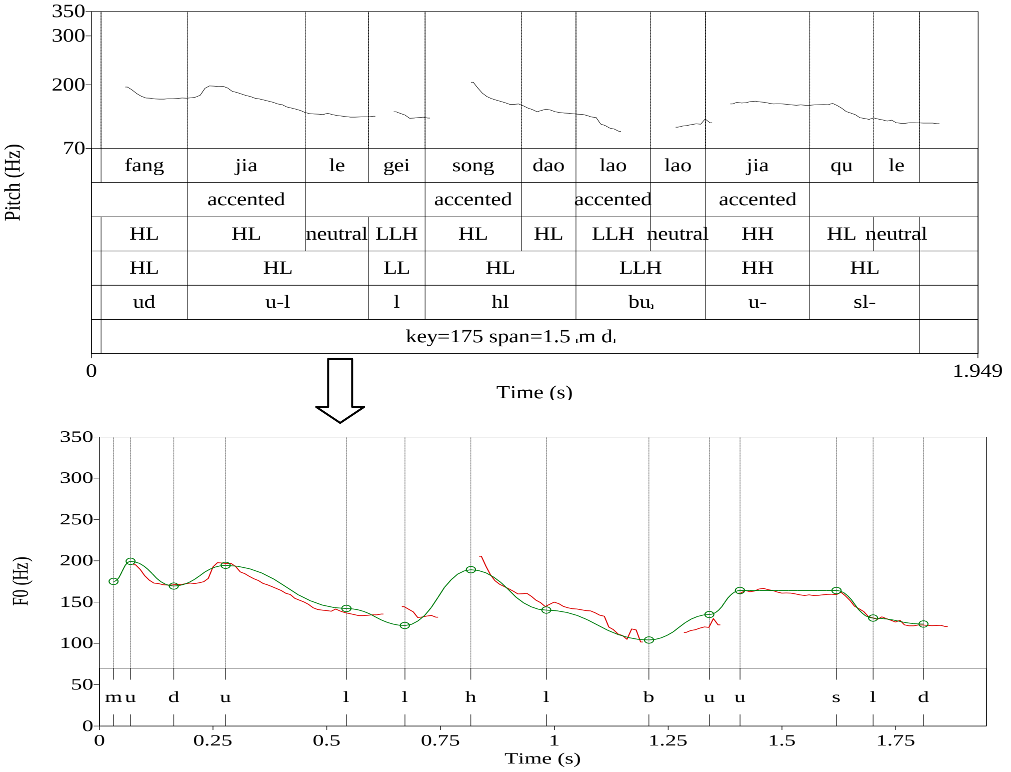

250
200
150
100
50
Pitch (Hz)
wo
hai
xu
yao
yi
xie
ying
ji
yao
accented
accented
accented
accented
LLH
LH
HH
HL
HL
HH
HL
LH
HL
LL
LH
HHL
HLHH
HLH
HL
l
-h
hsd
l-hs
l
ul
key=130 span=1 ˓m l˒
0
1.264
Time (s)
250
200
150
100
50
0
F0 (Hz)
m l
h
h
s
d
l
h
s
l
u
l l
0
0.25
0.5
0.75
1
1.25
Time (s)

List C

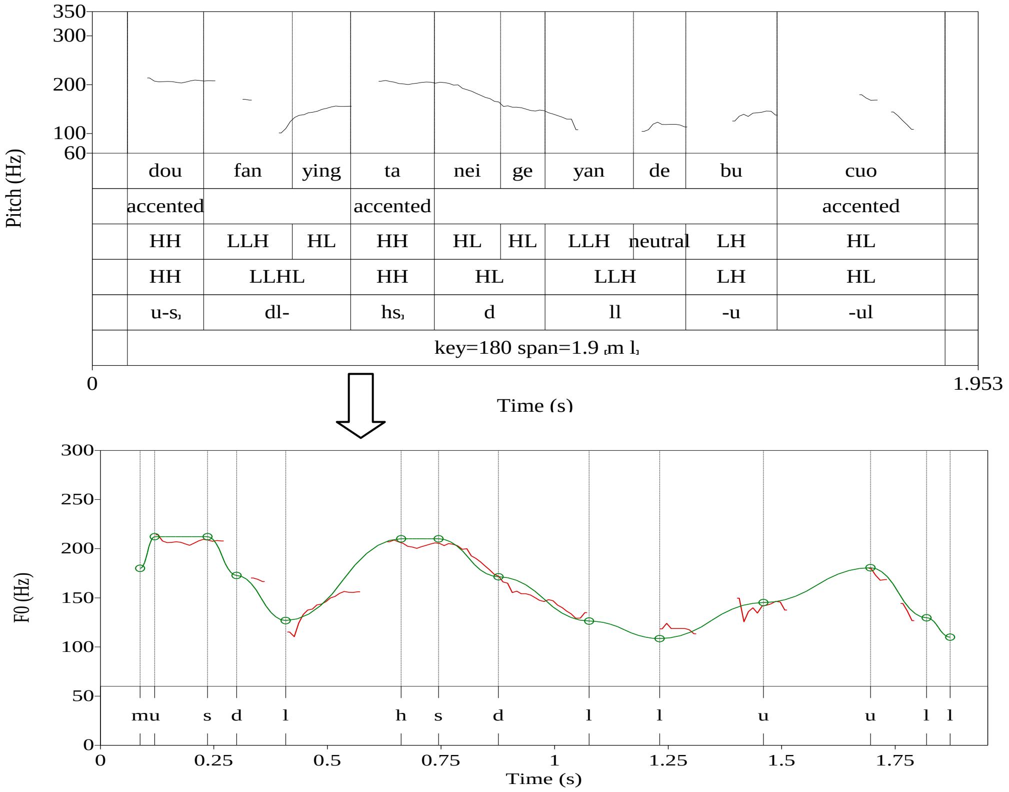
(41)


350
300
200
100
60
Pitch (Hz)
fei
shour
ma
ye
jue
dui
bu
shou
accented
accented
accented
LH
HL
neutral
LLH
LH
HL
LH
HL
LHL
LL
LHL
LH
HL
dtl
-l
--h
bh
-hl
key=175 span=1.3 ˌm lˌ
0
1.756
Time (s)
350
300
250
200
150
100
50
0
F0 (Hz)
m
d
t
l
l
h
b
h
h
l
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

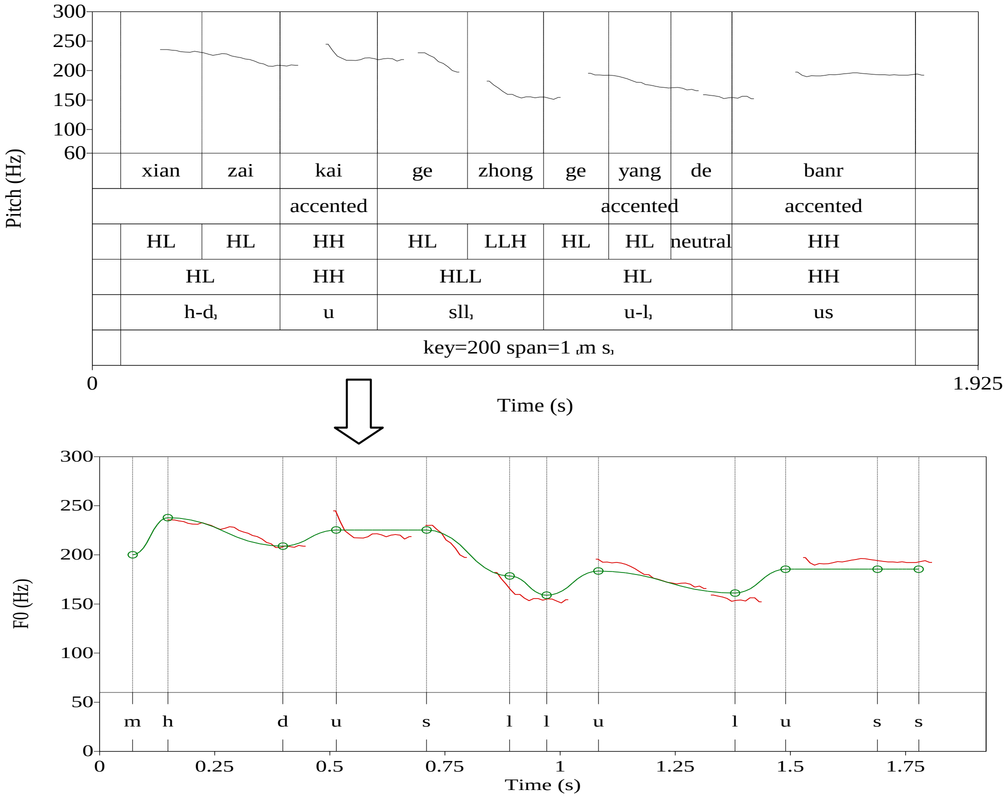
(42)


350
300
200
70
Pitch (Hz)
fang
jia
le
gei
song
dao
lao
lao
jia
qu
le
accented
accented
accented
accented
HL
HL
neutral
LLH
HL
HL
LLH
neutral
HH
HL
neutral
HL
HL
LL
HL
LLH
HH
HL
ud
u-l
l
hl
buˌ
u-
sl-
key=175 span=1.5 ˌm dˌ
0
1.949
Time (s)
350
300
250
200
150
100
50
0
F0 (Hz)
m
u
d
u
l
l
h
l
b
u
u
s
l
d
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

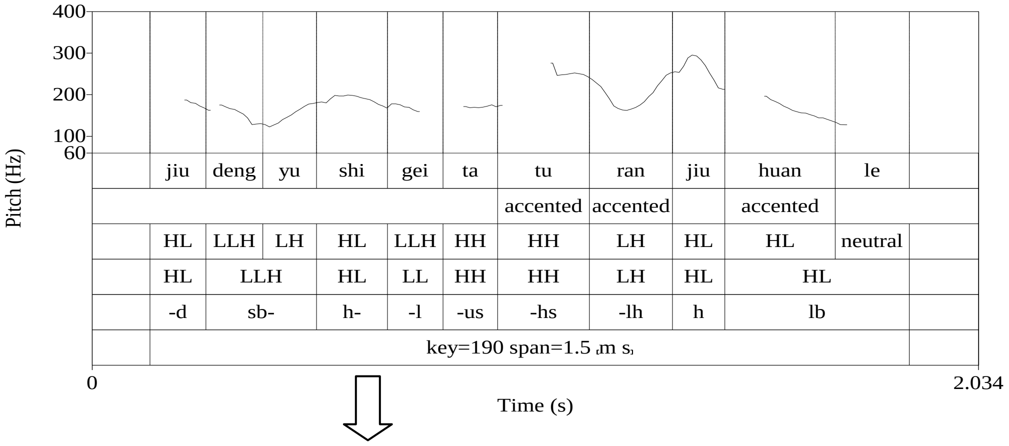
(43)

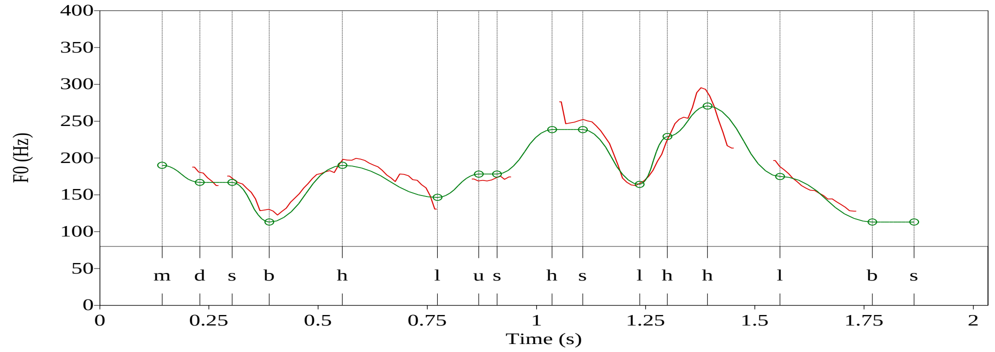

350
300
200
100
60
Pitch (Hz)
dou fan ying ta nei ge yan de bu cuo
accented accented accented
HH LLH HL HH HL HL LLH neutral LH HL
HH LLHL HH HL LLH LH HL
u-s˼ dl- hs˼ d ll -u -ul
key=180 span=1.9 ˻m l˼
0
1.953
Time (s)
300
250
200
150
100
50
0
F0 (Hz)
mu s d l h s d l l u u l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

(44)

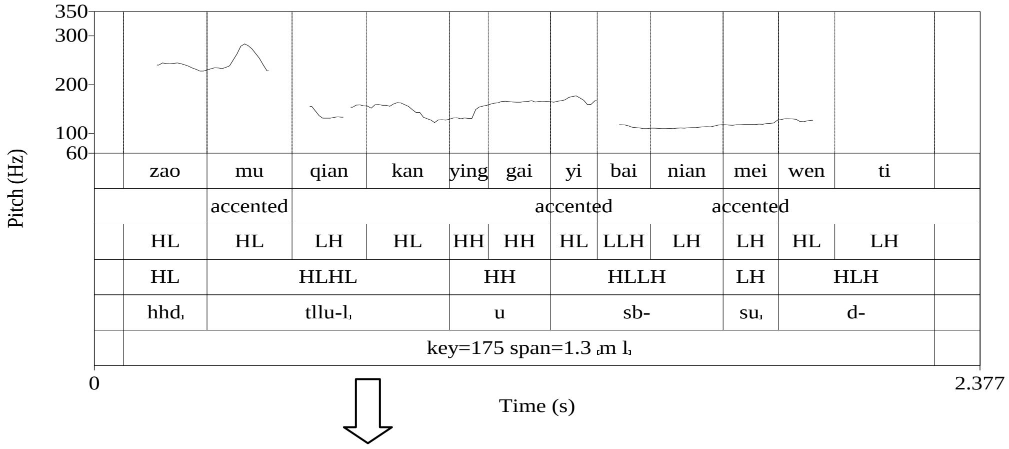

300
250
200
150
100
60
Pitch (Hz)
xian zai kai ge zhong ge yang de banr
accented accented accented
HL HL HH HL LLH HL HL neutral HH
HL HH HLL HL HH
h-d˼ u sll˼ u-l˼ us
key=200 span=1 ˻m s˼
0
1.925
Time (s)
300
250
200
150
100
50
0
F0 (Hz)
m h d u s l l u l u s s
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

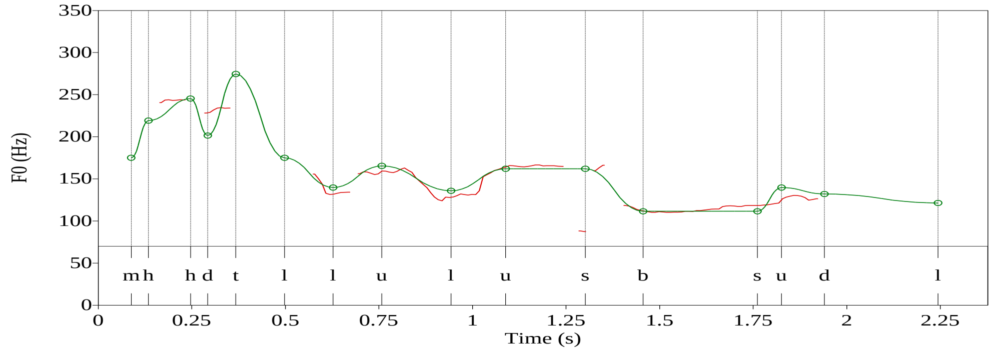
(45)


400
300
200
100
60
Pitch (Hz)
jiu
deng
yu
shi
gei
ta
tu
ran
jiu
huan
le
accented
accented
accented
HL
LLH
LH
HL
LLH
HH
HH
LH
HL
HL
neutral
HL
LLH
HL
LL
HH
HH
LH
HL
HL
-d
sb-
h-
-l
-us
-hs
-lh
h
lb
key=190 span=1.5 ⌊m s⌋
0
2.034
Time (s)

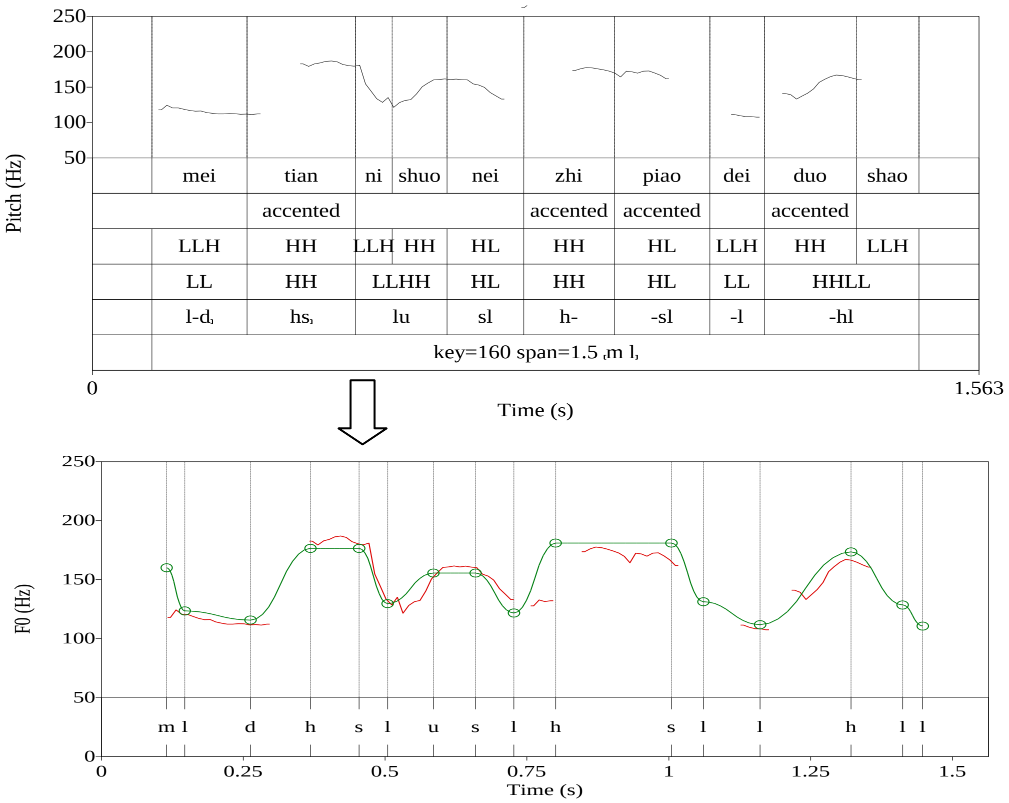

400
350
300
250
200
150
100
50
0
F0 (Hz)
m
d
s
b
h
l
u
s
h
s
l
h
h
l
b
s
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
Time (s)

(46)

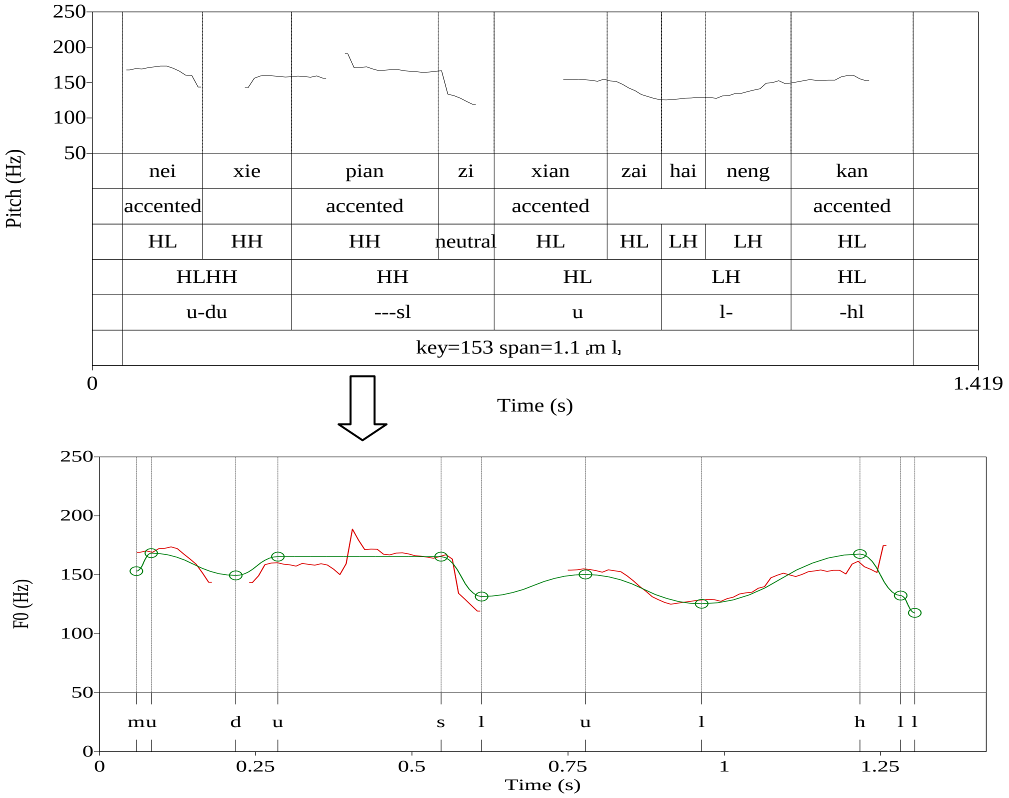

350
300
200
100
60
Pitch (Hz)
zao
mu
qian
kan
ying
gai
yi
bai
nian
mei
wen
ti
accented
accented
accented
HL
HL
LH
HL
HH
HH
HL
LLH
LH
LH
HL
LH
HL
HLHL
HH
HLLH
LH
HLH
hhd⌋
tllu-l⌋
u
sb-
su⌋
d-
key=175 span=1.3 ⌊m l⌋
0
2.377
Time (s)


350
300
250
200
150
100
50
0
F0 (Hz)
m
h
h
d
t
l
l
u
l
u
s
b
s
u
d
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
2.25
Time (s)

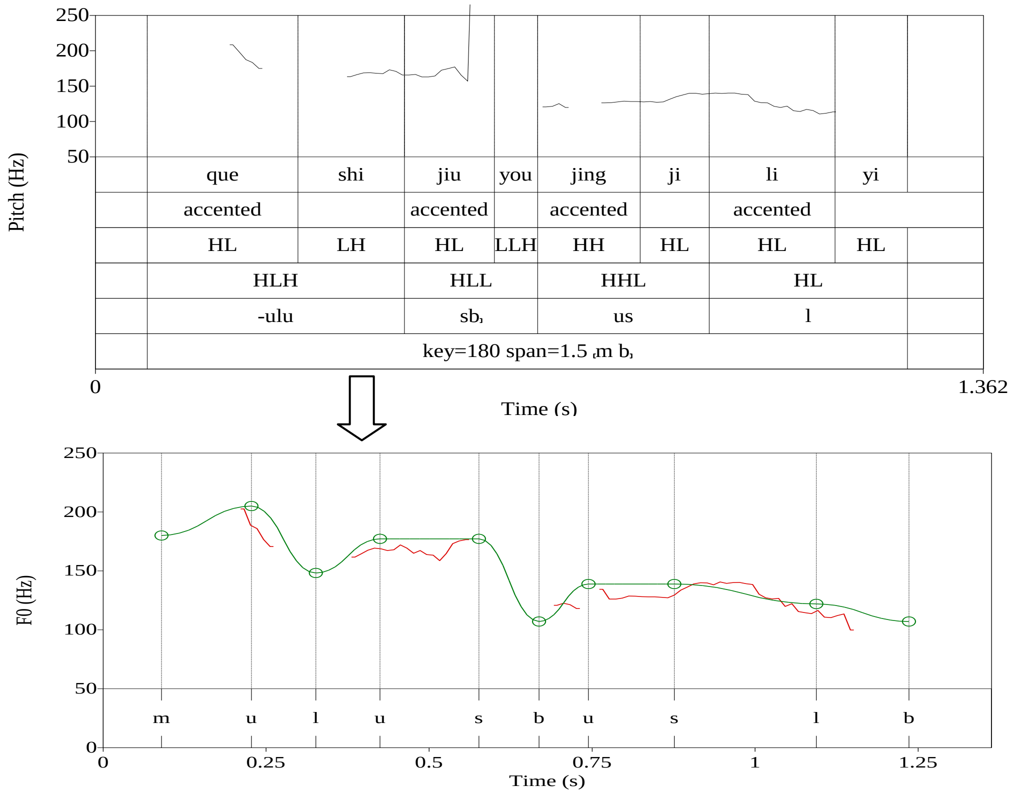
(47)


250
200
150
100
50
Pitch (Hz)
mei
tian
ni
shuo
nei
zhi
piao
dei
duo
shao
accented
accented
accented
accented
LLH
HH
LLH
HH
HL
HH
HL
LLH
HH
LLH
LL
HH
LLHH
HL
HH
HL
LL
HHLL
l-dˌ
hsˌ
lu
sl
h-
-sl
-l
-hl
key=160 span=1.5 ˌm lˌ
0
1.563
Time (s)
250
200
150
100
50
0
F0 (Hz)
m l d h s l u s l h s l l h l l
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

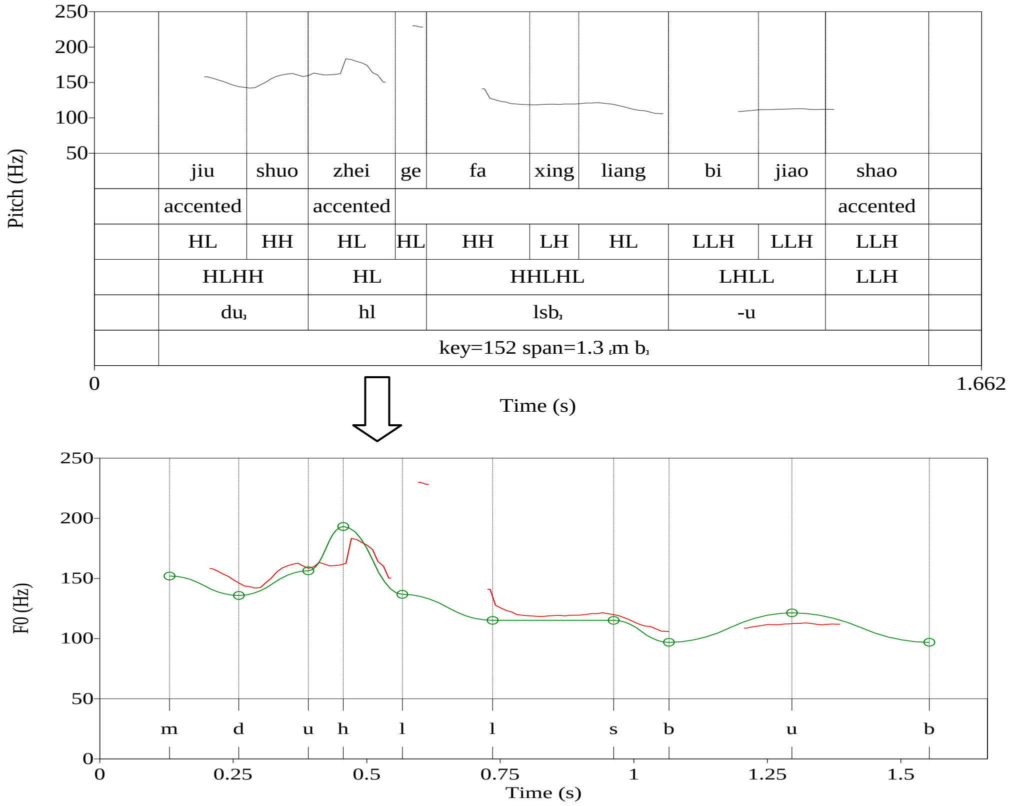
(48)


250
200
150
100
50
Pitch (Hz)
nei
xie
pian
zi
xian
zai
hai
neng
kan
accented
accented
accented
accented
HL
HH
HH
neutral
HL
HL
LH
LH
HL
HLHH
HH
HL
LH
HL
u-du
---sl
u
l-
-hl
key=153 span=1.1 ˌm lˌ
0
1.419
Time (s)
250
200
150
100
50
0
F0 (Hz)
mu d u s l u l h l l
0
0.25
0.5
0.75
1
1.25
Time (s)

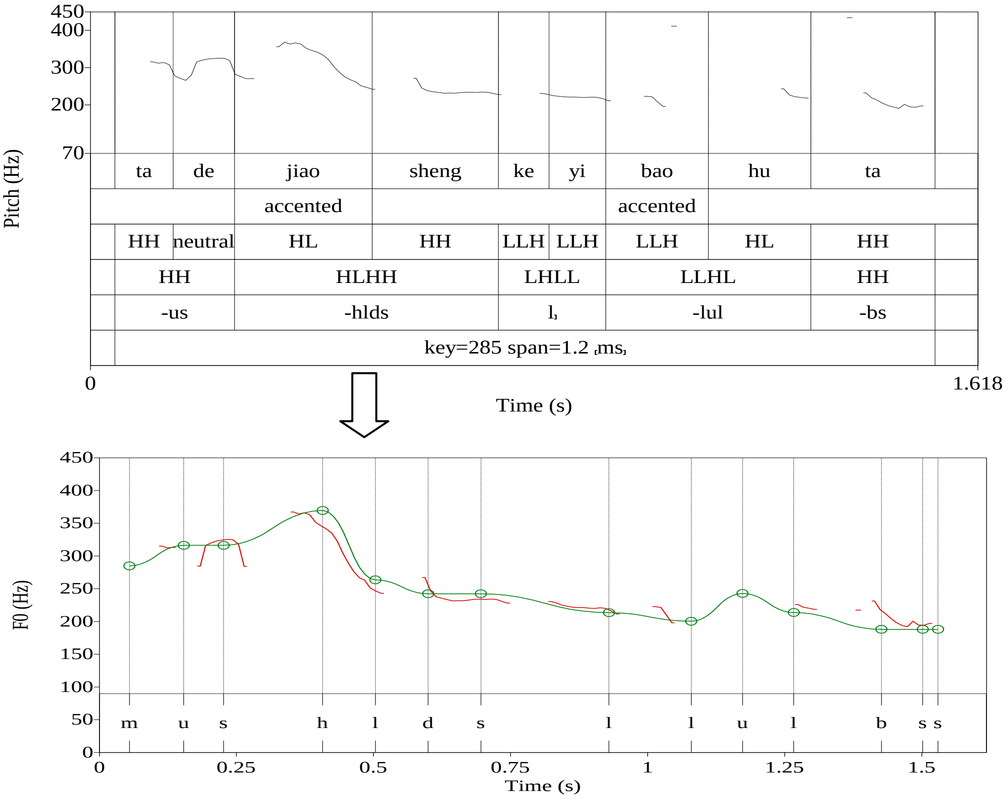
(49)


Pitch (Hz)
250
200
150
100
50
que
shi
jiu
you
jing
ji
li
yi
accented
accented
accented
accented
HL
LH
HL
LLH
HH
HL
HL
HL
HLH
HLL
HHL
HL
-ulu
sb˺
us
l
key=180 span=1.5 ˻m b˺
0
1.362
Time (s)
F0 (Hz)
250
200
150
100
50
0
m
u
l
u
s
b
u
s
l
b
0
0.25
0.5
0.75
1
1.25
Time (s)

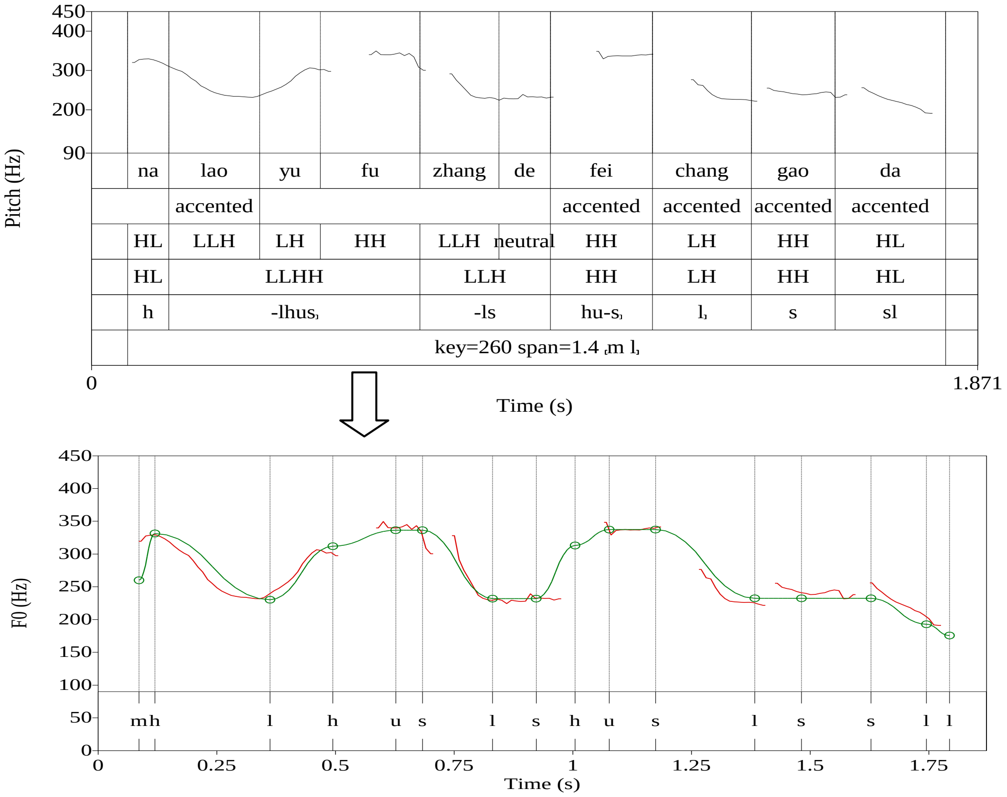
(50)


Pitch (Hz)
250
200
150
100
50
jiu
shuo
zhei
ge
fa
xing
liang
bi
jiao
shao
accented
accented
accented
HL
HH
HL
HL
HH
LH
HL
LLH
LLH
LLH
HLHH
HL
HHLHL
LHLL
LLH
du˺
hl
lsb˺
-u
key=152 span=1.3 ˻m b˺
0
1.662
Time (s)
F0 (Hz)
250
200
150
100
50
0
m
d
u
h
l
l
s
b
u
b
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

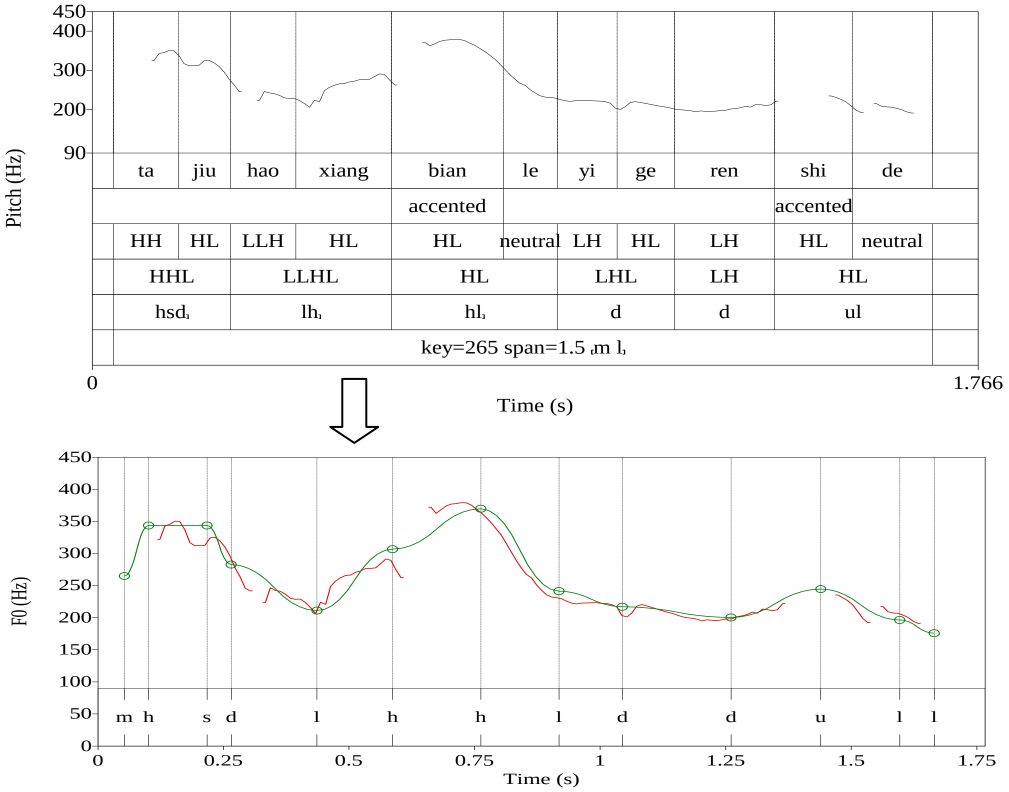
(51)


450
400
300
200
70
Pitch (Hz)
ta | de | jiao | sheng | ke | yi | bao | hu | ta
accented | accented
HH | neutral | HL | HH | LLH | LLH | LLH | HL | HH
HH | HLHH | LHLL | LLHL | HH
-us | -hlds | l˼ | -lul | -bs
key=285 span=1.2 ˻ms˼
0
1.618
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m u s h l d s l l u l b s s
0
0.25
0.5
0.75
1
1.25
1.5
Time (s)

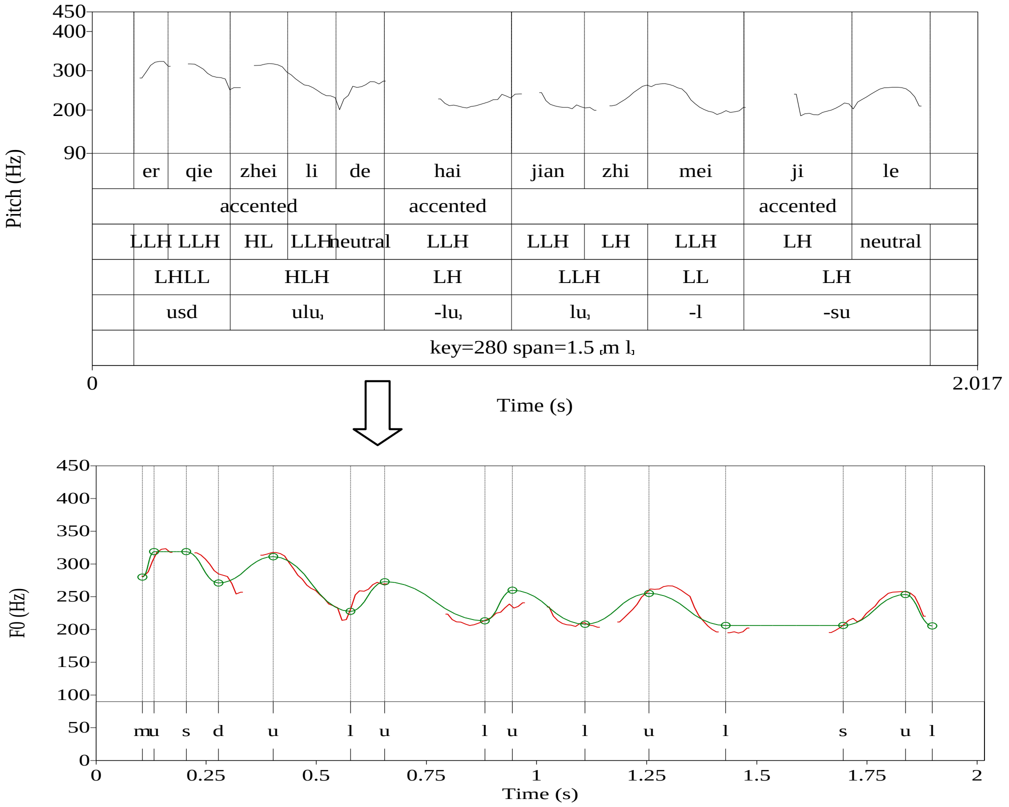
(52)


450
400
300
200
90
Pitch (Hz)
na | lao | yu | fu | zhang | de | fei | chang | gao | da
accented | accented | accented | accented | accented
HL | LLH | LH | HH | LLH | neutral | HH | LH | HH | HL
HL | LLHH | LLH | HH | LH | HH | HL
h | -lhus˼ | -ls | hu-s˼ | l˼ | s | sl
key=260 span=1.4 ˻m l˼
0
1.871
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
mh l h u s l s h u s l s s l l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

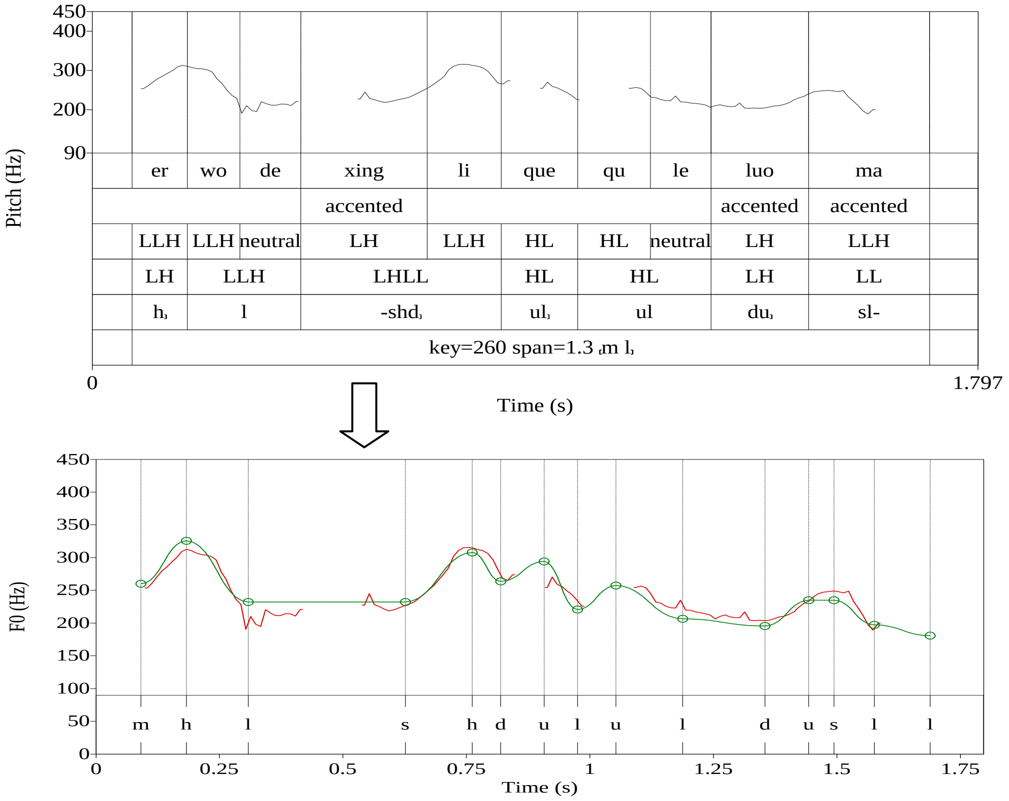
(53)


450
400
300
200
90
Pitch (Hz)
ta
jiu
hao
xiang
bian
le
yi
ge
ren
shi
de
accented
accented
HH
HL
LLH
HL
HL
neutral
LH
HL
LH
HL
neutral
HHL
LLHL
HL
LHL
LH
HL
hsd,
lh,
hl,
d
d
ul
key=265 span=1.5 ,m l,
0
1.766
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m
h
s
d
l
h
h
l
d
d
u
l
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Time (s)

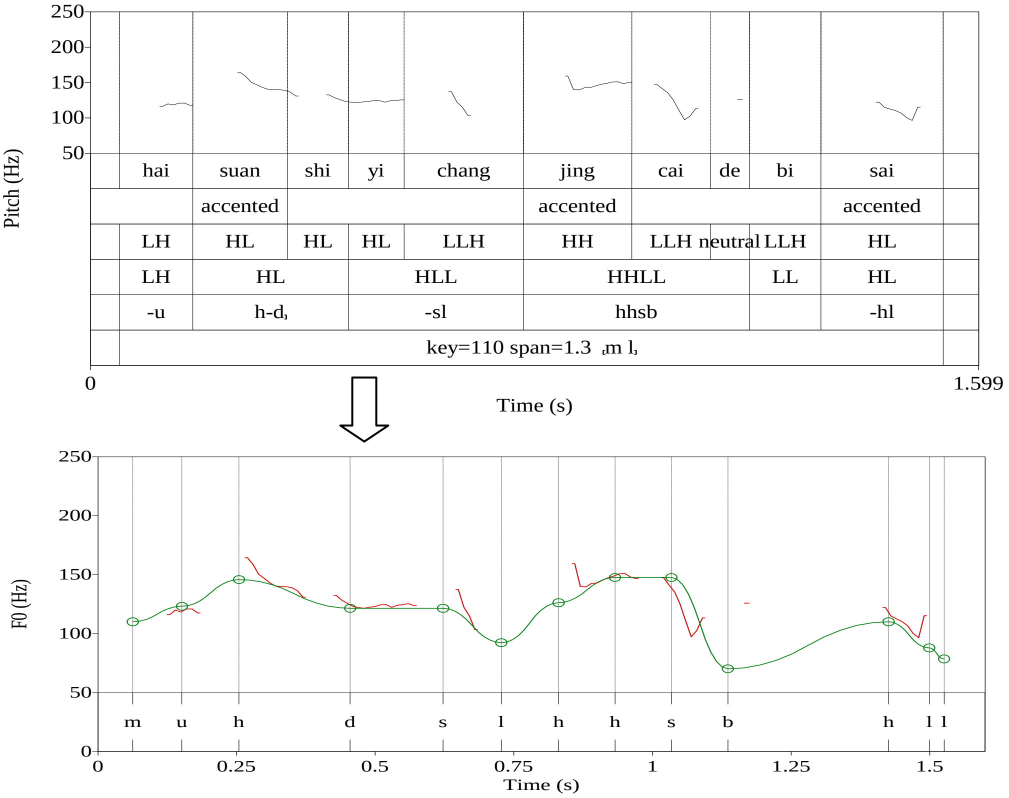
(54)


450
400
300
200
90
Pitch (Hz)
er
qie
zhei
li
de
hai
jian
zhi
mei
ji
le
accented
accented
accented
LLH
LLH
HL
LLH
neutral
LLH
LLH
LH
LLH
LH
neutral
LHLL
HLH
LH
LLH
LL
LH
usd
ulu,
-lu,
lu,
-l
-su
key=280 span=1.5 ,m l,
0
2.017
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
mu
s
d
u
l
u
l
u
l
u
l
s
u
l
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
Time (s)

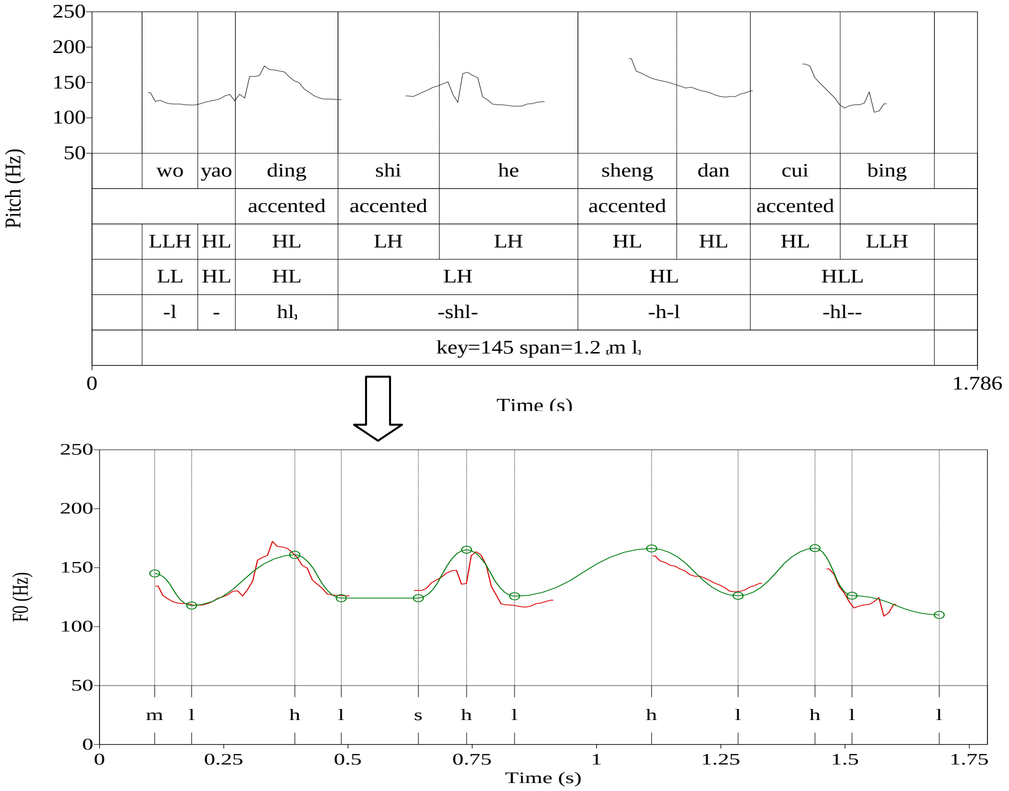
(55)


450
400
300
200
90
Pitch (Hz)
er wo de xing li que qu le luo ma
accented accented accented
LLH LLH neutral LH LLH HL HL neutral LH LLH
LH LLH LHLL HL HL LH LL
h˒ l -shd˒ ul˒ ul du˒ sl-
key=260 span=1.3 ˏm l˒
0
1.797
Time (s)
450
400
350
300
250
200
150
100
50
0
F0 (Hz)
m h l s h d u l u l d u s l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

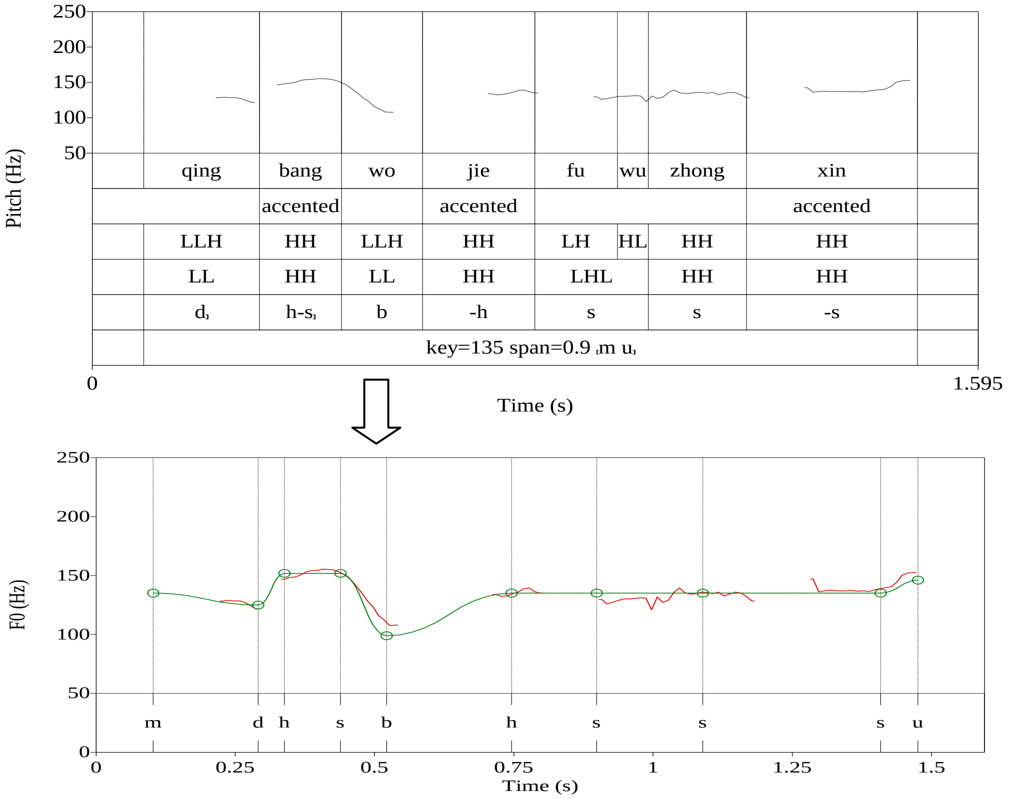
(56)


250
200
150
100
50
Pitch (Hz)
hai suan shi yi chang jing cai de bi sai
accented accented accented
LH HL HL HL LLH HH LLH neutral LLH HL
LH HL HLL HHLL LL HL
-u h-d˒ -sl hhsb -hl
key=110 span=1.3 ˏm l˒
0
1.599
Time (s)
250
200
150
100
50
0
F0 (Hz)
m u h d s l h h s b h l l
0 0.25 0.5 0.75 1 1.25 1.5
Time (s)

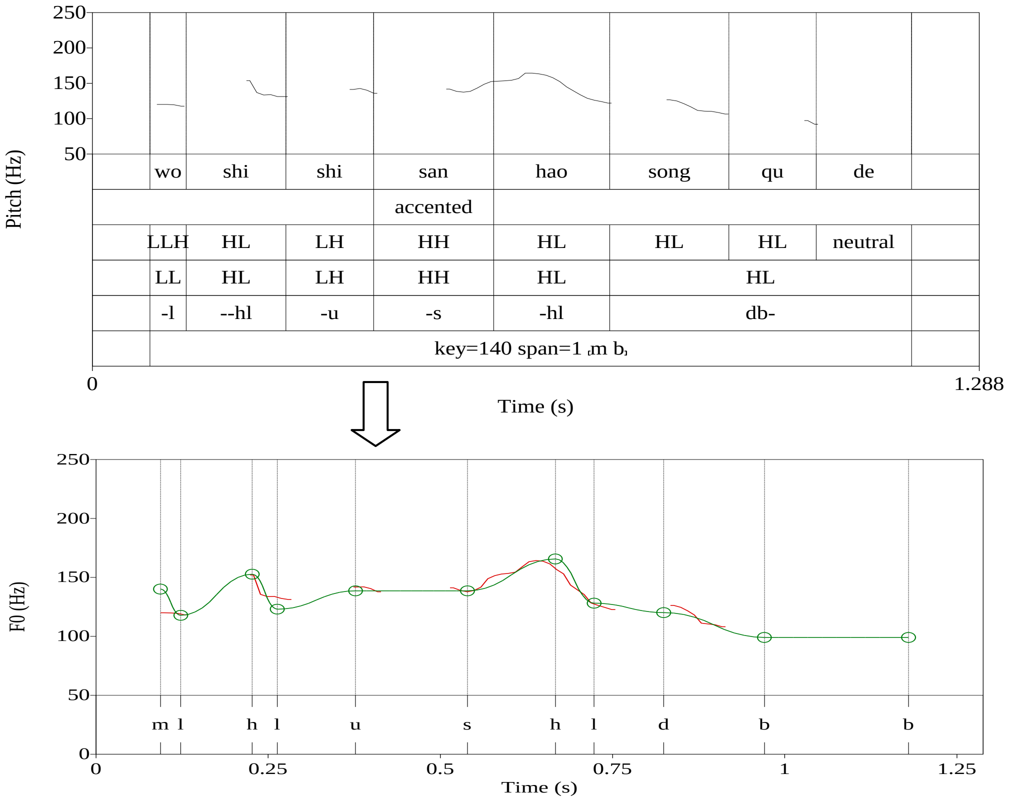
(57)


Pitch (Hz)
250
200
150
100
50
wo yao ding shi he sheng dan cui bing
accented accented accented accented
LLH HL HL LH LH HL HL HL LLH
LL HL HL LH HL HLL
-l - hl, -shl- -h-l -hl--
key=145 span=1.2 ,m l,
0
1.786
Time (s)
F0 (Hz)
250
200
150
100
50
0
m l h l s h l h l h l l
0 0.25 0.5 0.75 1 1.25 1.5 1.75
Time (s)

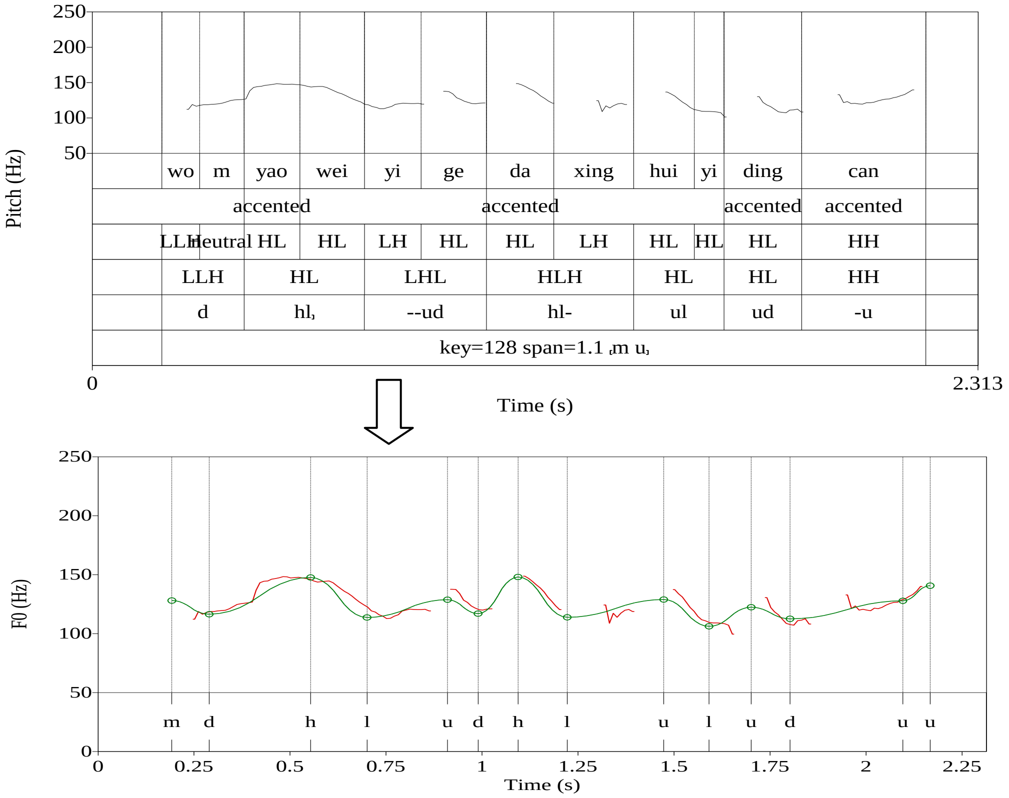
(58)


Pitch (Hz)
250
200
150
100
50
qing bang wo jie fu wu zhong xin
accented accented accented
LLH HH LLH HH LH HL HH HH
LL HH LL HH LHL HH HH
d, h-s, b -h s s -s
key=135 span=0.9 ,m u,
0
1.595
Time (s)
F0 (Hz)
250
200
150
100
50
0
m d h s b h s s s u
0 0.25 0.5 0.75 1 1.25 1.5
Time (s)

(59)


250
200
150
100
50
Pitch (Hz)
wo
shi
shi
san
hao
song
qu
de
accented
LLH
HL
LH
HH
HL
HL
HL
neutral
LL
HL
LH
HH
HL
HL
-l
--hl
-u
-s
-hl
db-
key=140 span=1 ⌞m b⌟
0
1.288
Time (s)
250
200
150
100
50
0
F0 (Hz)
m
l
h
l
u
s
h
l
d
b
b
0
0.25
0.5
0.75
1
1.25
Time (s)

(60)


250
200
150
100
50
Pitch (Hz)
wo
m
yao
wei
yi
ge
da
xing
hui
yi
ding
can
accented
accented
accented
accented
LLH
neutral
HL
HL
LH
HL
HL
LH
HL
HL
HL
HH
LLH
HL
LHL
HLH
HL
HL
HH
d
hl⌟
--ud
hl-
ul
ud
-u
key=128 span=1.1 ⌞m u⌟
0
2.313
Time (s)
250
200
150
100
50
0
F0 (Hz)
m
d
h
l
u
d
h
l
u
l
u
d
u
u
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
2.25
Time (s)

Appendix IV Zee's summary on the Shanghai tone sandhi domain

Zee (2004) claims the formation of tone sandhi domain is through the prosodic cliticization process in Shanghai dialect (523-524),

1) *The subject noun, verb and object noun do not form a tone sandhi domain with each other;*
2) *A noun and the preceding adjective obligatorily form a tone sandhi domain;*
3) *A noun and the post-nominal genitive or postposition obligatorily form a tone sandhi domain;*
4) *A noun and the preceding preposition or comparative do not form a tone sandhi domain, except where the noun is pronominalized;*
5) *A noun and the preceding quantifier and classifier do not form a tone sandhi domain;*
6) *A noun or pronoun and the preceding or following conjunction do not form a tone sandhi domain;*
7) *An object pronoun and the preceding verb, preposition, or comparative particle obligatorily form a tone sandhi domain. The prosodic cliticization of the verb and object pronoun is uninterrupted, where a complement is infixed in between them;*
8) *The object pronoun in the matrix sentence undergoes obligatory prosodic cliticization to the preceding verb, where it also functions as a subject pronoun in the embedded clause;*
9) *An object pronoun following a verb does not undergo prosodic cliticization, where it functions as subject pronoun in both the matrix sentence and embedded clause.*
10) *An object pronoun does not undergo prosodic cliticization, where it is contrastively stressed.*
11) *A verb and the following complement, preposition, perfective aspect particle, durative aspect particle, confirmation particle, or complementizer obligatorily form a tone sandhi domain;*

12) *A verb and the following complement, durative aspect particle, and perfective aspect particle obligatorily form a tone sandhi domain simultaneously;*

13) *A verb and the pre-verbal aspect particle do not form a tone sandhi domain;*

14) *A verb and the preceding auxiliary do not form a tone sandhi domain;*

15) *An adverb, excepting the negation [$_{\text{NEG}}$vəʔ 12] ‘not’ [勿], and the following verb optionally form a tone sandhi domain;*

16) *An adverb and the following adjective do not form a tone sandhi domain;*

17) *A quantifier and the following classifier obligatorily form a tone sandhi domain;*

18) *A nominalizer and the preceding noun, verb, or adjective obligatorily form a tone sandhi domain;*

19) *The negation [$_{\text{NEG}}$vəʔ 12] ‘not’ [勿]and the preceding verb and the following verb complement obligatorily form a tone sandhi domain;*

20) *The negation [$_{\text{NEG}}$vəʔ 12] ‘not’ [勿]and the following verb, adjective, adverb, auxiliary, or copula obligatorily form a tone sandhi domain;*

21) *The component syllables of an A-not-A question construction obligatorily form a tone sandhi domain;*

22) *The negation [$_{\text{NEG}}$məʔ 12] ‘not’ [没] does not form a tone sandhi domain with a preceding of following word of any syntactic category;*

23) *The copula does not form a tone sandhi with the preceding or following noun or pronoun;*

24) *The wh-question word [$_{\text{QUES}}$sa 34] ‘what’ [啥] and the following noun obligatorily form a tone sandhi domain;*

25) *The wh-question word [$_{\text{QUES}}$sa 34] ‘what’ [啥] and the preceding verb do not form a tone sandhi domain;*

26) *The phrase-initial or sentence-initial question word [$_{\text{QUES}}$ ʔaʔ 4] ‘whether or not’ [阿] and the following verb or auxiliary verb optionally form a tone sandhi domain;*

27) *The question particle [$_{\text{QUES}}$vaʔ 12] ‘yes-no’ [伐] and the preceding verb, adverb, or adjective obligatorily form a tone sandhi domain;*

28) *The phrase-final or sentence-final inchoative particle obligatorily forms a tone*

sandhi domain with any word that precedes it;

29) *A number word or personal name forms one or more tone sandhi domains, and the size of each tone sandhi domain does not exceed two syllables. The odd number syllable and the following even number syllable of a bisyllabic or polysyllabic number word or personal name obligatorily form a tone sandhi domain. The last odd number syllable of a number word remains prosodically unaffiliated, but the last odd number syllable of a person name is prosodically cliticized to the preceding tone sandhi domain.*