The Acoustic Cues at Prosodic Boundaries in Mandarin

Jiani Chen

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Committee:

Gina-Anne Levow

Richard Wright

Program Authorized to Offer Degree:

Linguistics

University of Washington

Abstract

The Acoustic Cues at Prosodic Boundaries in Mandarin

Jiani Chen

Chair of the Supervisory Committee:
Gina-Anne Levow
Department of Linguistics

Prosodic boundary labeling is an important task not only for its direct application in speech synthesis but also for constructing speech corpora for speech synthesis. However, due to the lack of quantitive study on Mandarin prosody, it is hard to provide a unified standard for prosodic boundary labeling.

In this work, I study the acoustic cues at prosodic boundaries in Mandarin with a large corpus and with quantitative methods. First of all, the study of acoustic cues at prosodic boundaries is done through experimental phonetics methods. Then, the acoustic cues are used as features to study their relation to different boundary types through automatic prosodic boundary labeling and feature ablation experiments. The one-way ANOVA results indicate that the baseline is reset only after intonational phrase boundaries, and it slightly declines after prosodic phrase boundaries. The results of ablation experiments employing a MaxEnt and an SVM

classifier, along with the ANOVA test results, demonstrate that silence duration is an essential acoustic cue at the prosodic boundaries. The results of ablation experiments also provide some information on acoustic-related acoustic cues: 1) Long-distance f0 variation (reset/declination) might be useful for measuring the degree of f0 variation after a prosodic boundary, and might be a useful acoustic cue for distinguishing different boundary types. 2) The pitch difference of the prosodic word after the boundary and a prosodic unit before the boundary might be more helpful to distinguish different boundary types. 3) The maximum pitch differences are not as useful as minimum pitch differences for distinguishing different boundary types. The f0 variation (reset/declination) at prosodic word boundaries and prosodic phrase boundaries might be mainly reflected in the variation of the minimum pitch; while at intonational phrase boundaries, the f0 variation might be mainly reflected in the variation of the mean pitch.

Furthermore, the results of the one-way ANOVA test and automatic prosodic boundary labeling both indicate that it is most difficult to distinguish prosodic word boundaries and prosodic phrase boundaries. Employing the proposed features in an SVM achieves substantially better results at distinguishing these two types of boundaries.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my advisor, Professor Gina-Anne Levow, for her patience, advice, and support throughout not only this thesis but my entire CMLS course load. I would also like to thank my reader Richard Wright for rekindling my enjoyment of phonetics in the experimental phonetics course. I would like to thank the entire UW Linguistics faculty and staff, too, for helping me get to this point in my studies. Finally, I would like to express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study.

Chapter 1. INTRODUCTION

Prosodic labeling is an important task not only for its direct application in speech synthesis but also for constructing speech corpora for speech synthesis.

A conventional way of prosodic labeling for constructing speech corpora is manual labeling. However, it is very time-consuming, and as the demand for speech corpora for speech synthesis is ever-growing, it is hard to rely only on manual labeling.

Another problem with manual prosodic labeling is that it is often inconsistent among annotators. The inconsistency is mainly due to the lack of a unified standard for prosodic labeling. The reason behind it is that prosody itself, which is very complicated in human speech, has not yet been fully studied. Although there has been a lot of research on Mandarin prosody, the definition of three typical Mandarin prosodic boundary labels - prosodic word (PW), prosodic phrase (PPH), and intonational phrase (IPH) is still controversial in Chinese academic linguistics. In addition, many of these researches only focus on qualitative analysis but ignore quantitative analysis. Thus, it is hard to give a general and definite description of PW, PPH, and IPH in Mandarin. Therefore, current manual prosodic boundary labeling relies heavily on the annotators' native language intuition rather than the acoustic cues, causing the labeling results to vary from person to person.

Nowadays, automatic prosodic boundary labeling, which can be more efficient than manual labeling, has been widely applied in speech synthesis as well as speech corpora construction. However, manually annotated corpora are still necessary for training a machine learning model. The inconsistency in the manual labeling brings difficulties to automatic prosodic boundary labeling when these manually annotated corpora are used.

Furthermore, although there have been many approaches to improve the accuracy of automatic labeling, it is still challenging to reach the accuracy level of manual labeling. Therefore, when it is applied to speech synthesis or speech corpora construction, manual post-editing is still a necessary step.

As to research on Mandarin automatic prosodic boundary labeling, many efforts have been made in feature engineering, but most of the features are not selected in consideration of the characteristics of prosodic boundaries in Mandarin. Even though some features might perform well in automatic prosodic boundary labeling, these features are hard to interpret, and thus cannot improve our understanding of Mandarin prosody.

In this work, I would like to study the acoustic cues at prosodic boundaries in Mandarin with quantitative methods. First of all, the study of acoustic cues at prosodic boundaries is done through experimental phonetics methods. Then, acoustic cues are used as features to study their relation to different boundary types through automatic prediction of boundary classes.

The remainder of this thesis is laid out as follows: chapter 2 details previous work on Mandarin prosody and automatic prosodic boundary labeling in Mandarin; chapter 3 provides details of the data and experimental framework; chapter 4 describes the details of the experiments and chapter 5 shows the experimental results and analysis. Finally, in chapter 6, I present my overall conclusions and describe future work.

Chapter 2. BACKGROUND

2.1 PROSODIC HIERARCHY IN MANDARIN

During communication, speech is divided into smaller units due to the physiology of speech and the meaning of a sentence. Between the units, breaks of different strength can be inserted. Broadly speaking, prosody includes tonal events (pitch accents, phrase accents, and boundary tones) and prosodic breaks. In a narrow sense, the study of prosody focuses on the prosodic structure, which is a prosodic hierarchy that describes a series of increasingly smaller prosodic units.

C-ToBI system was first proposed in 1996 for transcribing and annotating the prosodic structure of Mandarin read speech [Li, 1997]. It has later been developed for labeling spontaneous speech [Li, 2002]. Theoretically, the C-ToBI system includes the following eight tiers: a Pinyin tier, an initial and final tier, a tone and intonation tier, a break index tier, a stress index tier, a sentence function tier, an accent tier, a turn-taking tier, and a miscellaneous tier. However, due to the lack of a workable intonational theory for Mandarin, especially for Mandarin spontaneous speech, intonation transcription can be very ambiguous. Therefore, the Pinyin tier and the tone and intonation tier often fuse into a Pinyin and tone tier in practice. Intonation labeling is often omitted when annotating the corpora.

A typical Mandarin prosodic hierarchy includes prosodic word (PW), prosodic phrase (PPH), intonational phrase (IPH), and utterance (U). In C-ToBI system, boundaries after PW, PPH, IPH, and U correspond to break indices 1 to 4, respectively.

Tseng et al. [Tseng and Chou, 1999] [Tseng et al., 2005] also proposed adding the prosodic phrase group (PG) as a higher-level governing unit over the utterance (U) in the prosodic

hierarchy. They pointed out that the characteristics and manifestation of phrase- or sentence-f0 contours will vary according to their relative positions within the PG.

However, in this work, I will employ the typical Mandarin prosodic hierarchy, which only includes PW, PPH, IPH, and U. Graph 2.1 shows an example of a typical Mandarin prosodic hierarchy tree and the corresponding break indices at prosodic boundaries in the C-ToBI system.

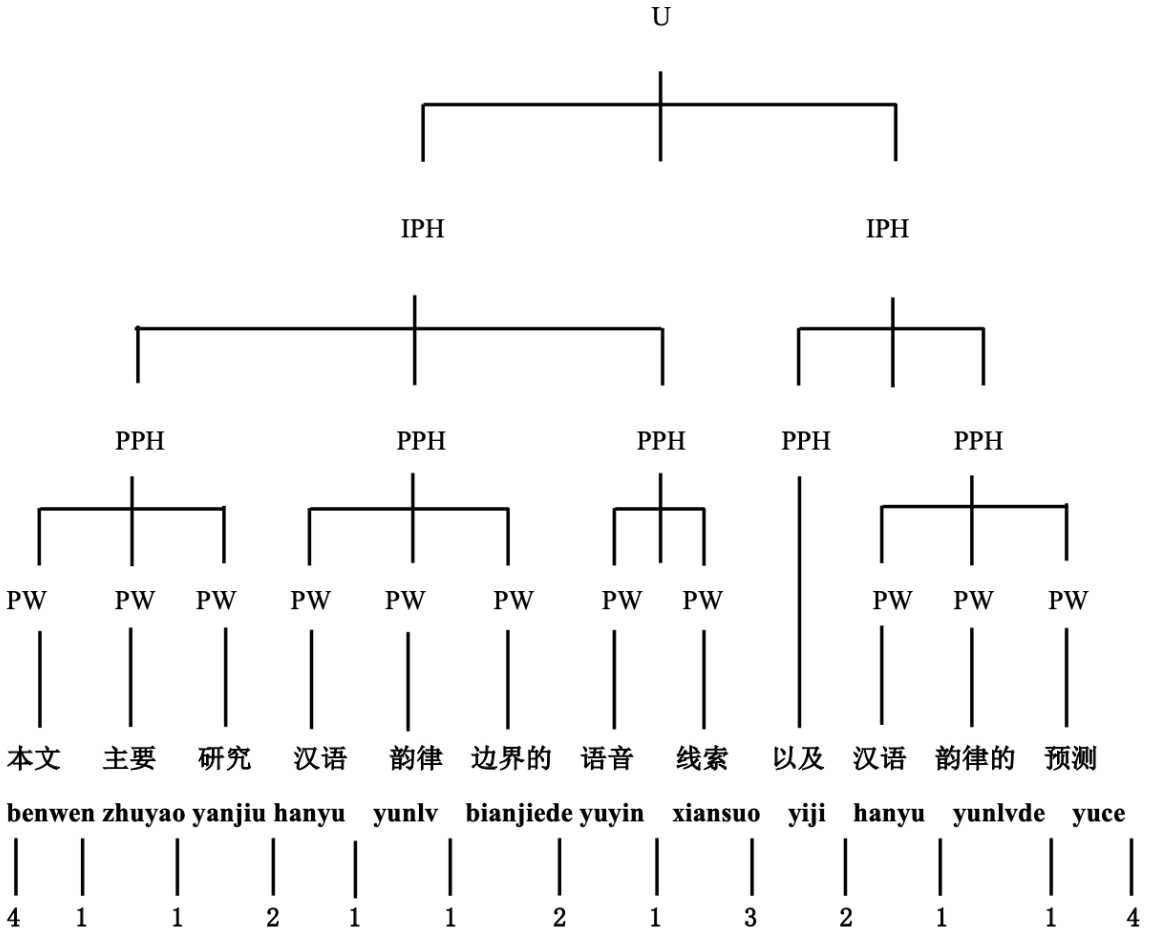

Figure 2.1: Four-level Mandarin prosodic hierarchy tree and the corresponding break indices at the prosodic boundaries.

2.1.1 *Prosodic Word*

In the Mandarin prosodic hierarchy, a prosodic word is a constituent higher than a syllable and a foot but lower than a prosodic phrase. A prosodic word is an f0 variation group [Lin, 2002]. The boundaries between characters/syllables and within a prosodic word are clitic boundaries. Feng [Feng, 1997] pointed out that a prosodic word should consist of at least one foot, and the foot in Mandarin is usually a two-syllable foot, which is because most Mandarin lexical words are disyllabic words. However, different from languages like English, a prosodic word in Mandarin is not always equal to a lexical word. Some one-character function words can form a prosodic word together with the nominal word before or after it. Also, some 4-character words (usually idioms) might be considered as two prosodic words. Sometimes, a prosodic word can also be part of a lexical word. Thus, the segmentation of prosodic words in Mandarin can be challenging. There has been much work focusing on the characteristics of Mandarin prosodic words and prosodic word segmentation. For example, Peng et al. [Peng et al., 2014] reported discrepancies found between lexical words parsed from text and prosodic words annotated from speech data, and proposed a statistical model to predict prosodic words from lexical words.

Prosodic word segmentation is a first step towards building prosody models from text. However, the segmentation of prosodic words in Mandarin is more related to the syntax. Therefore, the acoustic cues are less critical than lexical information. As the focus of this work is exploring acoustic cues that can distinguish prosodic word boundaries, prosodic phrase boundaries, and intonational phrase boundaries, the segmentation of prosodic words goes beyond the scope of this work, and will not be the focus of this work.

2.1.2 *Prosodic Phrase*

In the Mandarin prosodic hierarchy, a prosodic phrase is a constituent higher than a prosodic word but lower than an intonational phrase. The segmentation of a prosodic phrase is not only related to the syntax but also related to semantics and pragmatics. A prosodic phrase in Mandarin is not always equal to a syntactic phrase. There might be breaks between prosodic phrases to strengthen the rhythm of the sentence.

2.1.3 *Intonational Phrase*

In the Mandarin prosodic hierarchy, an intonational phrase is a constituent higher than a prosodic phrase but lower than an utterance. According to Ladd [Ladd, 1996], an intonational phrase is the largest phonological unit into which an utterance can be divided and has its own specifiable intonational structure. It matches with syntactic and discourse structures, so it is usually accompanied by a punctuation mark.

2.2 THE ACOUSTIC CUES AT PROSODIC BOUNDARIES IN MANDARIN

Since prosody is not only related to the syntax or the structure of a sentence but also related to the physiology of speech and the sentiments of the speaker, the acoustic cues at prosodic boundaries are crucial for segmenting the prosodic structure.

Precious research on Mandarin prosody had found that the acoustic cues at prosodic boundaries are more related to pitch and duration. Pitch-related acoustic cues include f0 reset or f0 declination. Duration-related acoustic cues include the lengthening of the last syllable before the boundary and the silence period at the boundary.

2.2.1 *Prosodic Word Boundary*

Syllable lengthening is found as a duration-related acoustic cue at prosodic word boundaries. Wang et al. [Wang et al., 2004] reported that there is vowel lengthening in the last syllable before prosodic word boundaries, and there is usually no silence period at prosodic word boundaries. As to pitch-related acoustic cue at prosodic word boundaries, Lin [Lin, 2002] concluded that f0 between prosodic words always reset.

2.2.2 *Prosodic Phrase Boundary*

Syllable lengthening is also found at prosodic phrase boundaries. Lin [Lin, 2000] reported that there might be vowel lengthening in the last syllable before a prosodic phrase boundary. Shih and Ao [Shih and Ao, 1994] also found that the vowel duration in the last syllable in a Mandarin prosodic phrase is lengthened. They also pointed out that vowel lengthening mostly occurs at the end of a prosodic phrase, rather than at the end of an utterance.

The silence period is another duration related acoustic cue at prosodic phrase boundaries. Lin [Lin, 2000] pointed out that at prosodic phrase boundaries there might be a silence period.

A vital pitch related acoustic cue at prosodic phrase boundaries is the reset of f0 baseline (a line which links the f0 bottom points of the prosodic words). He et al. [He et al., 2001] concluded that baseline declination is a critical characteristic of prosodic phrases. That is, the baseline declines within a prosodic phrase, and the baseline is reset at a prosodic phrase boundary.

2.2.3 *Intonational Phrase Boundary*

The duration-related acoustic cue at intonational phrase boundaries is the silence period. Wang et al. [Wang et al., 2004] found that the duration of the silence period at intonational phrase boundaries is longer than that at prosodic phrase boundaries.

Baseline reset is also found at intonational phrase boundaries. Wang et al. [Wang et al., 2004] pointed out that the baseline reset is more pronounced at intonational phrase boundaries than at prosodic phrase boundaries.

2.2.4 *Summary*

The above research has already explored some acoustic cues at prosodic boundaries in Mandarin. However, these conclusions are only based on a small number of corpora, and they haven't been evaluated statistically. Therefore, in this work, one of the aims is to verify the above findings on a larger corpus with statistical methods.

Besides, as to the pitch-related acoustic cues, some research did not quantify the f0 reset or declination. For example, Lin's [Lin, 2002] research did not detail how he measures the f0 reset between the prosodic words, so it is hard to tell whether the f0 reset between prosodic words is different from the baseline reset between prosodic phrases and intonational phrases. Thus, in my work, I would like to examine further the f0 variation (reset/declination) at prosodic boundaries in Mandarin.

2.3 AUTOMATIC PROSODIC BOUNDARY LABELLING

Automatic prosodic boundary labeling has wide application in speech synthesis as well as in constructing speech corpora for speech synthesis. Through automatic prosodic boundary labeling, we can also study the relation of the features used in the model to different boundary types, which can improve our understanding of prosody in return.

In general, automatic prosodic boundary labeling can be regarded as a sequence labeling task. Current automatic prosodic boundary labeling approaches can be classified into two categories.

One is the approaches that need feature engineering. For example, Qian et al. [Qian et al., 2010] and Sun et al. [Sun et al., 2009] selected various features and applied them in the Conditional Random Field (CRF) model. Ni et al. [Ni et al., 2012] proposed a range of features and tested them in various machine learning models, such as the CRF model, the Decision Tree model, and the support vector machine (SVM) model.

The other is the approaches that use models that are less dependent on feature engineering, such as neural networks. For example, Ding et al. [Ding et al., 2015] proposed to use stacking feed-forward and bidirectional long short-term memory (BLSTM) recurrent network to predict prosodic boundary labels directly from Chinese characters without any feature engineering. Xie et al. [Xie et al., 2018] suggested a prosodic boundary prediction method based on the "encoding-decoding" framework while using an effective position attention mechanism to further improve performance.

The deep learning approaches often show higher accuracy than traditional machine learning approaches, as they use word embeddings and take into account the distributional behavior of words. However, they might not reveal the nature of different prosodic boundary types because these approaches discover multiple levels of feature representations from the basic feature vectors, and thus the features in the neural networks model are less interpretable.

As the focus of this work is studying the acoustic cues at prosodic boundaries in Mandarin, automatic prosodic boundary labeling in this work is only used to study the relation of the features used in the model to different boundary types. In this sense, traditional machine learning models have more advantages. Therefore, in this work, I employ traditional machine learning models such as Maximum Entropy (MaxEnt) and support vector machine (SVM) in the automatic prosodic boundary labeling task.

2.3.1 *Features Used in Automatic Prosodic Boundary Labelling*

In previous researches, features used in automatic prosodic boundary labeling can be generally classified into two types: text features and acoustic features.

Text features refer to features that encoded information extracted from the text, such as the word's POS (part-of-speech) tag, the position of the word, and word count. These features are more closely related to the syntax or the structure of a sentence. For example, in Sun et al.'s [Sun et al., 2009] work, they used text features based on Zhao's [Zhao et al., 2003] research, including part-of-speech (POS), length in syllables and the words themselves surrounding the boundary.

Acoustic features refer to features which encoded acoustic information of the sentence, like pitch and duration. Since prosody is not only related to the structure of the sentence but also related to the factors such as the physiology of speech and the sentiments of the speaker, acoustic features can reflect other aspects of prosody. For example, in Qian et al.'s [Qian et al., 2010] work, a study on English automatic prosodic boundary labeling, acoustic features are applied in addition to text features: silence after word, duration of last syllable, and duration of last stressed syllable.

Since the focus of this work is studying the acoustic cues at prosodic boundaries in Mandarin, I will be essentially examining the relation of the acoustic features to different boundary types in Mandarin.

2.3.2 *Acoustic Features Used in Mandarin Prosody Researches*

The acoustic features used in previous research on Mandarin prosody can be generally classified into three types: duration-related features, pitch-related features, and energy-related features.

Fu et al. [Fu et al., 2018] applied the silence duration for each word as the acoustic features together with other text features in Chinese prosodic boundary labeling. The silence duration for

each word proved to have a better correlation with the prosodic boundaries than the acoustic features used in traditional methods which are extracted frame-by-frame.

Ni et al. [Ni et al., 2012] proposed various duration-related features, pitch-related features, energy-related features and lexical and syntactic related features in the prosodic break detection task, and analyzed the functions of different features in the Mandarin prosodic break detection task. Table 2.1, Table 2.2, and Table 2.3 list the duration-related features, pitch-related features, and energy-related features used in their work.

Feature Name	Feature Description
SilD	The silence duration after the syllable
SylDur	The duration of the syllable
SylDurRatio	The ratio between the following syllable duration and the current syllable duration
PDur	The duration of pitch discontinuing between the syllable and the following syllable

Table 2.1: Duration-related features

Type	Feature Name	Feature Description
Pitch Statistical	Pth_Max	The maximum of the syllable pitch
	Pth_Min	The minimum of the syllable pitch
	Pth_Range	The difference between Pth_Max and Pth_Min
	Pth_Mean	The mean of the syllable pitch
Pitch Contour	Con_Pth_a0, Con_Pth_a1, Con_Pth_a2, Con_Pth_a3, Con_Pth_a4, Con_Pth_a5	The coefficient of 5-order Legendre polynomial expansion
Pitch Comparison	PDlt	The difference between the last non-zero pitch value of the syllable and the first non-zero pitch value of the following syllable
	BPDlt	The difference between the minimum pitch of the syllable and the minimum pitch of the following syllable
	TPDlt	The difference between the maximum pitch of the syllable and the maximum pitch of the following syllable
	PMDlt	The difference between the mean pitch of the syllable and the mean pitch of the following syllable
	PRatio	The ratio between the last non-zero pitch value of the syllable and the first non-zero pitch value of the following syllable

Table 2.2: Pitch-related features

Type	Feature Name	Feature Description
Energy Statistical	Eng_Max	The maximum of the syllable energy
	Eng_Min	The minimum of the syllable energy
	Eng_Range	The difference between Eng_Max and Eng_Min
	Eng_Mean	The mean of the syllable energy
	EngRatio	The ratio between the mean of the syllable energy and the mean of the following syllable
Energy Contour	Con_Eng_a0, Con_Eng_a1, Con_Eng_a2, Con_Eng_a3, Con_Eng_a4, Con_Eng_a5	The coefficient of 5-order Legendre polynomial expansion

Table 2.3: Energy-related features

Although in this research, they are only concerned with whether the syllable is followed by a prosodic break or not (i.e., they did not make a distinction between different types of prosodic breaks), the features they proposed are still inspiring. However, one problem is that they did not take into consideration the characteristics of Mandarin prosody when selecting the features. Therefore, although they have compared the performance and analyzed the functions of the features, they were not able to give a sound explanation that has a theoretical basis, and the findings could not improve our understanding of Mandarin prosody in return.

Chapter 3. METHODOLOGY

3.1 DATA

The data used for the experiments in this work is ASCCD (Annotated Speech Corpus of Chinese Discourse) corpus. The texts of ASCCD contain 18 pieces of discourses, and each contains 2~5 sections and 300-500 syllables. The texts are read by 5 female speakers and 5 male speakers, who are F001, F002, F003, F004, F005, M001, M002, M003, M004, and M005 respectively.

Each speaker has 78 sound files, and each sound file is around 30 seconds. The recordings were made in a professional sound booth with CONDENSER CR 1-4 microphone (distance between the mouth of the speaker and the mic is about 20 centimeters), and CREATIVE SOUND BLASTER LIVE! soundcard. The speech signal was recorded in two channels: the speech waveform and the glottal impedance waveform. The speech waveform channel was recorded by SONY DTC-55ES recorder, and the glottal impedance waveform channel was recorded by KAY Laryngograph Model 6094. The sampling rate is 16kHz, and the bit rate is 16-bit rate.

Each sound file has a corresponding TextGrid file, which is annotated with segmental and prosodic annotations including canonical Pinyin and tone tier, initial / final tier of real pronunciation and stress tier. Stress is labeled by indices 1, 2, and 3, which correspond to prosodic word stress, prosodic phrase stress, and intonational phrase stress, respectively. The prosodic boundaries are labeled by break indices 1, 2, 3, and 4, which correspond to prosodic word break (PW), prosodic phrase break (PPH), intonational phrase break (IPH), and sentence/utterance break (U) respectively. Table 3.1 shows the distribution of the four types of break indices in the whole corpus.

	PW	PPH	IPH	U
Count	14758	7675	6957	3710

Table 3.1: The distribution of the four types of break indices in the whole corpus

3.2 EXPERIMENTAL FRAMEWORK

In this work, the experiments are divided into two parts.

The first part is experimental phonetics experiments. This part aims to study the acoustic cues at different prosodic boundaries in Mandarin with quantitative methods. I also would like to verify and concretize the previous findings of the acoustic cues at the prosodic boundaries in Mandarin mentioned in section 2.2 with a larger corpus.

Previous research on Mandarin prosody has found that the acoustic cues at prosodic boundaries in are more related to pitch and duration, as mentioned in Section 2.2. Therefore, the acoustic measures I select are duration and pitch. Duration-related acoustic cues are syllable lengthening before the prosodic boundary and the silence period at the prosodic boundary. Pitch-related acoustic cue is f0 rest or declination after prosodic boundaries. To further study the specific manifestation of f0 variation (reset/declination) after different types of prosodic boundaries, I not only measured the pitch difference of the syllables before and after the boundary but also measured the pitch difference of the prosodic words before and after the boundary and their combinations. Furthermore, I also measured the pitch difference of the syllable/prosodic word after the boundary and the syllable/prosodic word after last intonational phrase boundary (BI=3), so as to find out the degree of long-distance f0 reset or declination (i.e., the degree of f0 variation compared to last intonational phrase boundary.) Section 4.2 will detail the procedure of measurements and calculations.

The second part is to study further the relation of the acoustic cues to different boundary types with machine learning methods. In the first part of the experiments, I have explored several acoustic cues at prosodic boundaries in Mandarin. To find out which of these acoustic cues are more characteristic of each boundary type, and which of these acoustic cues can better distinguish different boundary types, I use these acoustic cues as features and apply them in the automatic prosodic boundary labeling task, using the MaxEnt (Maximum Entropy) model and the SVM (Support Vector Machine) model. I also ran ablation experiments to study the function of each feature (acoustic cue) and its relation to different boundary types. Section 4.3 will detail the procedure of automatic prosodic boundary labeling and ablation experiments.

3.3 DATA PREPROCESSING

First, I ran a Praat script to cut the sound files and the corresponding TextGrid files into smaller files at the sentence/utterance boundaries (BI=4), because in this work, I only focus on the acoustic cues at prosodic word boundaries (BI=1), prosodic phrase boundaries (BI=2) and intonational phrase boundaries (BI=3). Then, I removed the sound files in which the pitch track is incomplete. (That is, the pitch value cannot be obtained automatically from Praat.) Table 3.2 and Table 3.3 shows the number of files of the female speakers and male speakers before and after processing.

Speaker index	Unprocessed	After cutting	After removing
f001	78	421	418
f002	78	374	316
f003	78	303	244
f004	78	369	366
f005	78	271	266
Total	390	1738	1610

Table 3.2: The number of files before and after processing of the female speakers

Speaker index	Unprocessed	After cutting	After removing
m001	78	496	410
m002	80	370	267
m003	78	442	414
m004	78	345	157
m005	78	319	251
Total	392	1972	1499

Table 3.3: The number of files before and after processing of the male speakers

Chapter 4. EXPERIMENTS

4.1 FEATURE SETUP

In the experimental phonetics experiment, I select 24 pitch-related acoustic cues and 2 duration-related acoustic cues. And then, in the automatic prosodic boundary labeling task, these acoustic cues are used as the features for the classification task. Table 4.1 shows the details of the features.

Feature #1 to feature #24 are pitch-related features. For each prosodic boundary, I measured the pitch of two prosodic units after the boundary: 1) the syllable after the boundary and 2) the prosodic word after the boundary, and four units before the boundary: 1) the syllable before the boundary, 2) the prosodic word before the boundary, 3) the first syllable after last intonational phrase boundary (BI=3) and 4) the first prosodic word after last intonational phrase boundary (BI=3). I took the difference of the pitch of a unit after the boundary and the pitch of a unit before the boundary. Thus, there are 8 combinations in all. And for each combination, I took the difference of the maximum pitch, the mean pitch , and the minimum pitch of the two units. Thus, there are 24 types of pitch difference in total for each prosodic boundary.

Feature #25 and feature #26 are duration-related features. Feature #25 is the percentage of the duration of the final of the last syllable before the boundary and the duration of the syllable. Feature #26 is the duration of the silence period.

Feature Index	Feature Description
1	f0_next_syllable_maximum – f0_last_syllable_maximum
2	f0_next_syllable_mean – f0_last_syllable_mean
3	f0_next_syllable_minimum – f0_last_syllable_minimum
4	f0_next_syllable_maximum – f0_last_prosodic_word_maximum
5	f0_next_syllable_mean – f0_last_prosodic_word_mean
6	f0_next_syllable _minimum – f0_last_prosodic_word_minimum
7	f0_next_syllable_maximum – f0_boundary_next_syllable_maximum
8	f0_next_syllable_mean – f0_boundary_next_syllable_mean
9	f0_next_syllable_minimum – f0_boundary_next_syllable_minimum
10	f0_next_syllable_maximum – f0_boundary_next_prosodic_word_maximum
11	f0_next_syllable_mean – f0_boundary_next_prosodic_word_mean
12	f0_next_syllable_minimum – f0_boundary_next_prosodic_word_minimum
13	f0_next_prosodic_word_maximum – f0_last_syllable_maximum
14	f0_next_prosodic_word _mean – f0_last_syllable_mean
15	f0_next_prosodic_word_minimum – f0_last_syllable_minimum
16	f0_next_prosodic_word_maximum – f0_last_prosodic_word_maximum
17	f0_next_prosodic_word_mean – f0_last_prosodic_word_mean
18	f0_next_prosodic_word_minimum – f0_last_prosodic_word_minimum
19	f0_next_prosodic_word_maximum – f0_boundary_next_syllable_maximum
20	f0_next_prosodic_word_mean – f0_boundary_next_syllable_mean
21	f0_next_prosodic_word_minimum – f0_boundary_next_syllable_minimum
22	f0_next_prosodic_word_maximum – f0_boundary_next_prosodic_word_maximum
23	f0_next_prosodic_word_mean – f0_boundary_next_prosodic_word_mean
24	f0_next_prosodic_word_minimum – f0_boundary_next_prosodic_word_minimum
25	the percentage of the duration of the final of the last syllable before the boundary and the duration of the syllable
26	the duration of silence period

Table 4.1: Features used in the automatic prosodic boundary labelling task

4.2 Experimental Phonetics Experiments

4.2.1 *Acoustic Measurement*

When measuring the pitch, the pitch range for female speakers in the pitch settings is set to be 55-400 Hertz, and for male speakers, the pitch range is set to be 50-300 Hertz. I first randomly selected 5 sound files from the 78 sound files of each speaker to do the manual measurement in order to make more in-depth observations of the sound files and the corresponding annotations. In manual measurement, for each prosodic boundary, I only measured the maximum, the mean, and the minimum pitch of the two prosodic units after the boundary and the four units before the boundary. After the manual measurement, a Praat script is run on the whole preprocessed corpus to extract the maximum, the mean, and the minimum pitch of each syllable and each prosodic word in each sound file.

When measuring the pitch of the syllable in the manual measurement, I exclude the initial consonants if they are voiceless. The start point I selected for measuring the pitch of the syllable is where the waveform starts to be periodic, and the endpoint is 20-50ms before the start point of the next consonant (depending on when the amplitude of the waveform becomes very low). If the syllable to be measured doesn't have a consonant initial, then there is formant transition between the last vowel in the last syllable and the first vowel in the current syllable, so the endpoint of the last syllable and the start point of the current syllable is the mid-point of the formant transition. The value of the maximum, the mean and minimum pitch of the syllable is obtained directly from selecting Get maximum pitch, Get pitch, and Get minimum pitch from the Pitch menu in the SoundEditor window in Praat. While in automatic measurement, the start point and the endpoint of the syllable just comply with the annotation in the second tier (Pinyin tier) of the TextGrid files. The initial consonants are also excluded from the syllable if they are voiceless.

However, I did not exclude the last 20-50ms of the syllables as I did in the manual measurement, because the time when the amplitude of the waveform becomes very low varies a lot in different syllables and different sound files. Thus, it is hard to use the same standard for all the syllables and the sound files in automatic measurement.

The maximum, the mean, and the minimum pitch of the prosodic word depend on its component syllables. The maximum/minimum pitch of the prosodic word is the highest/lowest pitch of the component syllables. The mean pitch of the prosodic word is calculated using the following formula, in which syllable_duration is the duration of the part selected for measuring the pitch of the syllable:

$$\frac{\sum syllable\_mean\_pitch \times syllable\_duration}{\sum syllable\_duration} \tag{4.1}$$

In automatic measurement, the start point and the endpoint of the prosodic word comply with the annotation in the third tier (prosodic boundary tier) of the TextGrid files.

In order to measure syllable lengthening, I measured both the duration of the syllable and the duration of the final of the syllable, and then calculate the percentage of the duration of the final in the syllable. Because according to Lin [Lin, 2000] [Lin, 2002] and Wang et al. [Wang et al., 2004], the duration of the finals in the last syllable before the prosodic word and prosodic phrase boundary is lengthened. The following is the formula:

$$syllable\_final\_percentage = \frac{final\_duration}{syllable\_duration} \tag{4.2}$$

In manual measurement, I did not exclude the last 20-50ms of the syllable when measuring the duration of the syllable as I did when measuring the pitch of the syllable, because the lengthening of the syllable-final is often characteristic of the lengthening of those low amplitude waveform at the end of the syllable. In automatic measurement, the duration of the syllable is measured according to the second tier (Pinyin tier) in the TextGrid files, and the duration of the

final in each syllable is measured according to the first tier (initial/final tier) in the TextGrid files. It should be noted that if there is a noticeable lengthening in the low amplitude waveform at the end of a syllable, the lengthening period is marked with label "tl" (tile) in the first tier (initial/final tier) and the second tier (Pinyin tier) in the TextGrid files. So the duration of this period is also added to the duration of the final as well as the total duration of the syllable.

When measuring the duration of the silence period in manual measurement, if there is a period after a prosodic boundary when there is no waveform (ignoring the background noise), then it is counted as the silence period after the prosodic boundary. However, if there is a stop or affricate after the silence period, the last 5ms of the silence period (the 5ms before the release burst of the following stop or affricate) will be excluded from the duration of the silence period and will be included in the duration of the following stop or affricate. In automatic measurement, the duration of the silence period at a prosodic boundary is measured according to the second tier (Pinyin tier) in the TextGrid files. According to the TextGrid files in this corpus, the annotators usually leave 2ms - 6ms for the silence period before the release burst of a stop or affricate, and the rest are counted as the silence period at a prosodic boundary. The silence period after a prosodic boundary is labeled as "sil" or "silv" in the TextGrid files. The duration of the silence period after a prosodic boundary is the duration of the interval with label "sil" or "silv". If there is no interval with label "sil" or "silv" before or after the boundary, then there is no silence period at the boundary.

4.2.2 *Calculations*

In order to reduce the effect of inter-speaker and intra-speaker variation, I used z-score[1] to normalize the original value of pitch (in Hertz) and duration (in seconds). For comparison, 24 types of pitch difference and 2 types of duration-related acoustic correlates are calculated both using the original value and the normalized value.

Finally, the values were classified into three groups by the prosodic boundary label[2]. For each type of acoustic cue, I calculated the mean of each type of prosodic boundary and used SPSS[3] to run the one-way ANOVA[4] test to see whether there is statistical evidence that they are significantly different. Section 5.1 will discuss the results of the one-way ANOVA test.

4.3 AUTOMATIC PROSODIC BOUNDARY LABELLING

4.3.1 *Features*

Automatic prosodic boundary labeling is run on the preprocessed data. 24 pitch-related features and 2 duration-related features are used in this task.[5] The feature values are calculated using the normalized pitch and duration values.

It should be noted that since these features all have real-number values, the real-number valued original features should be transformed into a new form by quantization when the features are used as the input to the MaxEnt model because MaxEnt operates over discrete features rather than continuous features. For example, suppose the original value for feature #1 is

---

[1] z-score normalization: $z = \frac{x-\mu}{\sigma}$, where x is a value to normalize, μ and σ are mean and standard deviations which are estimated from the pitch or duration of all syllables or prosodic words for a speaker.

[2] BI=1: prosodic word, BI=2: prosodic phrase, BI=3: intonational phrase

[3] IBM SPSS Statistics for Macintosh, Version 26.0. Armonk, NY: IBM Corp.

[4] I used Brown-Forsythe model for one-way ANOVA test and Tamhane's T2 as the model for post hoc multiple comparisons. It is run both on the original values and the normalized values.

[5] See section 4.1 for details.

-0.9418610803381964, first, the feature value needs to be rounded to 2 decimal places, then the feature input to the MaxEnt model needs to be transformed into one-hot representation: #1_-0.94:1. As to the SVM model, the real-valued original features can be used as input directly.

It should also be noted that, 12 of the features: #7, #8, #9, #10, #11, #12, #19, #20, #21, #22, #23, #24 are long-distance features because these types of pitch difference are calculated using the pitch of the syllable/prosodic word after the boundary and the pitch of the first syllable/prosodic word after last intonational phrase boundary. As the values of these long-distance features are dependent on where the last intonational phrase boundary is, therefore they need to be treated differently in the testing stage, which will be explained in the next section.

4.3.2 *Training and Testing*

For each speaker, the processed files are divided into training data and testing data according to the ratio 7:3. Table 4.2, Table 4.3, and Table 4.4 shows the details of the number of training data and the testing data.

Speaker index	Training	Testing	Total
f001	292	126	418
f002	221	95	316
f003	170	74	244
f004	256	110	366
f005	186	80	266
Total	1125	485	1610

Table 4.2: The number of training data and the testing data of the female speakers

Speaker index	Training	Testing	Total
m001	287	123	410
m002	186	81	267
m003	289	125	414
m004	109	48	157
m005	175	76	251
Total	1046	453	1499

Table 4.3: The number of training data and the testing data of the male speakers

Speaker Gender	Training	Testing	Total
Female	1125	485	1610
Male	1046	453	1499
Total	2171	938	3109

Table 4.4: Total number of training data and the testing data

The training stage of the automatic prosodic boundary labeling task is similar to other classification tasks. In the training stage, the input training instances are the prosodic words with 26 features[6], and the output label is the prosodic boundary label after that prosodic word. In this task, there are three output labels in the training and testing data: 1, 2 and 3, corresponding to the prosodic word boundary (PW), the prosodic phrase boundary (PPH), and the intonational phrase boundary (IPH). Table 4.5 shows the distribution of the three types of prosodic boundary labels in the training and testing data.

LABEL / DATA	PW	PPH	IPH
Training data	8364	4358	3889
Testing data	3522	1834	1613
Total	11886	6192	5502

Table 4.5: The distribution of the three types of prosodic boundary labels in the training data and the testing data

The testing stage is a bit different from the training stage. The testing stage of the automatic prosodic boundary labeling task is similar to other sequence labeling tasks. Each test instance is a sentence which consists of several prosodic words. Each prosodic word with 26 features is used as the input to the classifier, and the classifier will predict the probability of each output label. Then I used the Viterbi algorithm to select the best output sequence. Since features: #7, #8, #9,

[6] See Section 3.4.2 for the details of the features.

#10, #11, #12, #19, #20, #21, #22, #23, #24 are long-distance features which are dependent on the output labels of previous instances, so these features are not present in the test files and should be added on the fly.

MaxEnt model and SVM model are both used for training and testing. The MaxEnt trainer is from Mallet[7]. The SVM trainer is from Scikit-learn[8]. When using the SVM model, I selected RBF (Radial basis function) as the kernel, and gamma is set to be 0.5. As to the parameter "cost", I tried several different values, and the one which has the highest test accuracy is used in the ablation experiments. The test results of "most common class assignment" are also provided as the baseline results for comparison.

Section 5.2 will discuss the results of the automatic prosodic boundary labeling task.

4.3.3 *Ablation Experiments I*

The first set of ablation experiments aims to study the function of each feature and its relation to different boundary types. This set of ablation experiments was run on both the MaxEnt model and the SVM model. Specifically, in each ablation experiment, one type of feature is removed from both the training data and the test data. Then, a new model is trained and tested on the new data with that type of feature removed. For example, to see the effect of feature #1, all features like #1_* are removed when using the MaxEnt model, and when using the SVM model, just simply remove feature #1 from the input vectors. Section 5.3 will discuss the results of the ablation experiments I.

[7] Mallet: http://mallet.cs.umass.edu
[8] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

4.3.4 *Ablation Experiments II*

In order to study further the role of different types of pitch-related features, I ran another three sets of ablation experiments on the SVM model only using pitch-related features (Feature #25 and #26 are excluded).

The first set of ablation experiment aims to compare the features which incorporate the pitch of the syllable after the boundary with the features which incorporate the pitch of the prosodic word after the boundary. In the first configuration, only Feature #1 to #12 are included in the training data and the test data. These features are the pitch difference of the syllable after the boundary and a prosodic unit before the boundary. In the second configuration, only Feature #13 to #24 are included. These features are the pitch difference of the prosodic word after the boundary and a prosodic unit before the boundary.

The second set of ablation experiments aims to examine further the role of long-distance features in distinguishing different boundary types. In the first configuration, only long-distance features (Feature #7, #8, #9, #10, #11, #12, #19, #20, #21, #22, #23, and #24) are included in the training data and the test data. In the second configuration, all the long-distance features are removed, and I only include twelve non-long-distance features (Feature #1, #2, #3, #4, #5, #6, #13, #14, #15, #16, #17, and #18).

The third set of ablation experiments aims to compare the minimum pitch based features with the mean pitch based features and the maximum pitch based features. In the first configuration, only minimum pitch based features (Feature #3, #6, #9, #12, #15, #18, #21, and #24) are included in the training data and the test data. In the second configuration, I only keep mean pitch based features (Feature #2, #5, #8, #11, #14, #17, #20, and #23). In the third

configuration, I only use maximum pitch based features (Feature #1, #4, #7, #10, #13, #16, #19, and #22).

The results of these three sets of ablation experiments will be discussed in section 5.4.

Chapter 5. RESULTS AND DISCUSSION

5.1 ONE-WAY ANOVA TEST

5.1.1 *Pitch-related Acoustic Cues*

From the result of the one-way ANOVA test, for both original pitch values and normalized pitch values, the means of three types of prosodic boundaries are all significantly different for all the 24 types of pitch difference, which is shown in Table 5.1 and Table 5.2.

		BI = 1	BI = 2	BI = 3	p-value	Mean difference (3-1)	Mean difference (3-2)	Mean difference (2-1)
next_syllable - last_syllable	maximum	7.08	14.59	36.5	3.06E-210	29.42	21.91	7.52
	mean	5.98	23.68	47.32	0	41.34	23.64	17.7
	minimum	3.39	28.71	53.72	0	50.32	25	25.32
next_syllable - last_prosodic_word	maximum	-26.7	-26.39	-5.72	3.53E-117	20.98	20.66	0.31*
	mean	-4.94	8.57	33	0	37.94	24.43	13.51
	minimum	15.86	40.03	65.98	0	50.12	25.95	24.17
next_syllable - boundary_next_ syllable	maximum	-18.7	-18.07	-6.64	3.60E-37	12.06	11.43	0.63*
	mean	-22.02	-19.35	-6.82	9.97E-63	15.2	12.53	2.67
	minimum	-23.86	-20.24	-6.57	6.61E-75	17.29	13.67	3.62
next_syllable - boundary_next_ prosodic_word	maximum	-33.99	-34.92	-21.99	1.07E-44	12	12.93	-0.93*
	mean	-13.01	-5.25	5.22	1.82E-118	18.23	10.48	7.76
	minimum	7.22	21.52	29.73	1.03E-144	22.51	8.21	14.3
next_prosodic_word - last_syllable	maximum	24.43	32.99	52.74	4.41E-203	28.31	19.75	8.56
	mean	-8.15	10.57	36.79	0	44.94	26.22	18.72
	minimum	-38.79	-12.61	19.78	0	58.57	32.39	26.17
next_prosodic_word - last_prosodic_word	maximum	-9.35	-7.99	10.52	2.58E-109	19.87	18.51	1.36*
	mean	-19.07	-4.54	22.47	0	41.54	27.01	14.53
	minimum	-26.32	-1.29	32.05	0	58.37	33.34	25.02
next_prosodic_word - boundary_next_ syllable	maximum	-1.35	0.33	9.6	5.54E-31	10.95	9.27	1.68
	mean	-36.15	-32.46	-17.35	1.48E-126	18.8	15.11	3.69
	minimum	-66.04	-61.56	-40.51	1.89E-159	25.53	21.05	4.48
next_prosodic_word - boundary_next_ prosodic_word	maximum	-16.64	-16.52	-5.75	1.68E-37	10.89	10.77	0.12*
	mean	-27.14	-18.36	-5.3	4.36E-264	21.84	13.06	8.78
	minimum	-34.96	-19.8	-4.2	2.58E-285	30.76	15.6	15.16

Table 5.1: The mean of each type of original pitch difference of each prosodic boundary type and corresponding one-way ANOVA p-value and post hoc multi-comparisons results (*. The mean difference is not significant at the 0.01 level)

		BI = 1	BI = 2	BI = 3	p-value	Mean difference (3-1)	Mean difference (3-2)	Mean difference (2-1)
next_syllable - last_syllable	maximum	0.15	0.29	0.68	9.65E-175	0.54	0.39	0.14
	mean	0.12	0.46	0.91	0	0.79	0.45	0.33
	minimum	0.07	0.54	1.03	0	0.96	0.49	0.47
next_syllable - last_prosodic_word	maximum	-0.52	-0.5	-0.16	6.16E-81	0.35	0.34	0.01*
	mean	-0.09	0.17	0.63	0	0.72	0.46	0.26
	minimum	0.31	0.76	1.27	0	0.96	0.52	0.45
next_syllable - boundary_next_ syllable	maximum	-0.35	-0.33	-0.14	7.63E-28	0.21	0.18	0.02*
	mean	-0.42	-0.36	-0.15	3.98E-54	0.27	0.21	0.07
	minimum	-0.46	-0.37	-0.14	6.03E-70	0.32	0.23	0.09
next_syllable - boundary_next_ prosodic_word	maximum	-0.65	-0.65	-0.45	6.00E-30	0.2	0.2	0*
	mean	-0.25	-0.09	0.09	8.25E-107	0.33	0.18	0.15
	minimum	0.14	0.4	0.56	7.56E-147	0.43	0.16	0.26
next_prosodic_word - last_syllable	maximum	0.49	0.65	1.01	2.48E-161	0.52	0.36	0.16
	mean	-0.15	0.2	0.7	0	0.86	0.5	0.36
	minimum	-0.74	-0.25	0.37	0	1.12	0.62	0.5
next_prosodic_word - last_prosodic_word	maximum	-0.17	-0.14	0.16	1.08E-70	0.33	0.3	0.03*
	mean	-0.37	-0.09	0.42	0	0.79	0.51	0.28
	minimum	-0.51	-0.03	0.62	0	1.12	0.65	0.48
next_prosodic_word - boundary_next_ syllable	maximum	0	0.04	0.18	8.17E-22	0.19	0.15	0.04
	mean	-0.7	-0.61	-0.35	1.18E-115	0.35	0.26	0.09
	minimum	-1.27	-1.16	-0.8	1.98E-167	0.47	0.36	0.11
next_prosodic_word - boundary_next_ prosodic_word	maximum	-0.3	-0.29	-0.12	4.41E-23	0.18	0.17	0.01*
	mean	-0.52	-0.35	-0.12	4.30E-243	0.41	0.23	0.18
	minimum	-0.68	-0.38	-0.09	2.22E-304	0.58	0.29	0.29

Table 5.2: The mean of each type of normalized pitch difference of each prosodic boundary type and corresponding one-way ANOVA p-value and post hoc multi-comparisons results (*. The mean difference is not significant at the 0.01 level)

From the post hoc multi-comparisons results in Table 5.1 and Table 5.2, mean difference (3-1)[9] and mean difference (3-2)[10] are statistically significant for all the 24 types of pitch difference. Among all these types of pitch difference, "next_prosodic_word – last_ prosodic_word (minimum) " and "next_ prosodic_word – last_syllable (minimum) " (corresponding to the feature #15 and #18 in Table 4.1)[11] have the largest mean difference (3-1), mean difference (3-2), and mean difference (2-1).

However, mean difference (2-1)[12] is not statistically significant for the following types of pitch difference: "next_syllable – last_prosodic_word (maximum)", "next_syllable – boundary_next_syllable (maximum)", "next_syllable – boundary_next_prosodic_word (maximum)", "next_ prosodic_word – last_ prosodic_word (maximum)" and "next_ prosodic_word – boundary_next_prosodic_word (maximum)" (corresponding to the feature #4, feature #7, feature #10, feature #16 and feature #22 in Table 4.1). Besides, the mean difference between prosodic word boundaries (BI=1) and prosodic phrase boundaries (BI=2) for all the 24 types of pitch difference are all small[13]. These might indicate that it is harder to distinguish prosodic word boundaries and prosodic phrase boundaries. Among all the 24 types of pitch difference, the following three types of pitch difference have comparatively larger mean difference between prosodic word boundaries and prosodic phrase boundaries, that is: "next_syllable – last_syllable (minimum)", "next_ prosodic_word – last_syllable (minimum)" and "next_prosodic_word – last_ prosodic_word (minimum)". It indicates that these three types

[9] The difference between the means of intonational phrase boundaries (BI=3) and prosodic word boundaries (BI=1)
[10] The difference between the means of intonational phrase boundaries (BI=3) and prosodic phrase boundaries (BI=2)
[11] The types of pitch difference are sorted in descending order according to the mean difference value.
[12] The difference between the means of prosodic phrase boundaries (BI=2) and prosodic word boundaries (BI=1)
[13] The largest mean difference between prosodic word boundaries (BI=1) and prosodic phrase boundaries (BI=2) in normalized value is 0.5, and the largest mean difference in original value is 26.17.

of pitch difference might be better acoustic cues to distinguish prosodic word boundaries and prosodic phrase boundaries.

The one-way ANOVA test result also indicates that different ways of measuring the pitch difference can lead to different conclusions of whether f0 after a prosodic boundary is reset or declines. For example, f0 is reset after all the three types of boundaries when measuring the pitch difference of the syllables before and after the boundary. However, f0 declines after prosodic word boundaries and prosodic phrase boundaries but is reset after intonational phrase boundaries when measuring the pitch difference of the prosodic words before and after the boundary. Besides, the pitch difference of the maximum pitch or mean pitch or minimum pitch can also lead to different conclusions. For example, when measuring the pitch difference of the syllable after a prosodic boundary and the prosodic word before the boundary, the maximum pitch declines after all the three types of prosodic boundaries, but the minimum pitch is reset after all the three types of boundaries.

However, we can find that, overall, for all the 24 types of pitch difference, the means of intonational phrase boundary (BI=3) are the largest, while the means of prosodic phrase (BI=2) boundary are the second largest, and the means of prosodic word boundary (BI=1) are the smallest. It indicates that if there is an f0 reset (the pitch of the unit after the boundary is higher than the pitch of the unit before the boundary) after the boundary, the unit after an intonational phrase boundary has the largest degree of reset. If f0 declines after the boundary, the declination after a prosodic word boundary is more pronounced than after a prosodic phrase boundary than after an intonational phrase boundary.

Another point that should be noted is about baseline declination and reset. Baseline reset means the lowest pitch of the prosodic word after the boundary is higher than the lowest pitch of

the prosodic word before the boundary. According to Wang et al. [Wang et al., 2004], baseline reset is found both after the prosodic phrase boundaries and intonational phrase boundaries, and the degree of reset is more pronounced after intonational phrase boundaries than after prosodic phrase boundaries. However, according to the result in Table 5.1 and Table 5.2, the baseline is reset only after intonational phrase boundaries. And after prosodic phrase boundaries, baseline slightly declines (or it can be interpreted as the degree of reset is small). Nevertheless, on the other hand, since "next_ prosodic_word – last_syllable (minimum)" has the largest mean difference (3-1), mean difference (3-2), and mean difference (2-1), it might indicate that the manifestation of baseline variation (reset/declination) at the three types of boundaries are different, and therefore baseline variation can be a critical acoustic cue to distinguish different prosodic boundaries in Mandarin.

5.1.2 *Duration-related Acoustic Cues*

From the result of the one-way ANOVA test, for both original pitch values and normalized pitch values, the means of three types of prosodic boundaries are all significantly different for the two types of duration-related acoustic cues, which is shown in Table 5.3 and Table 5.4.

	BI = 1	BI = 2	BI = 3	p-value	Mean difference (3-1)	Mean difference (3-2)	Mean difference (2-1)
syllable_final_ percentage	0.76	0.78	0.75	8.22E-14	-0.01	-0.03	0.02
silence_period_ duration (in seconds)	0.02	0.09	0.48	0	0.46	0.39	0.07

Table 5.3: The mean of each duration-related acoustic cue of each prosodic boundary type (original value) and corresponding one-way ANOVA p-value and post hoc multi-comparisons results

	BI = 1	BI = 2	BI = 3	p-value	Mean difference (3-1)	Mean difference (3-2)	Mean difference (2-1)
syllable_final_ percentage	0.17	0.25	0.13	4.15E-13	-0.04	-0.13	0.08
silence_period_ duration	-0.66	-0.31	1.6	0	2.26	1.91	0.35

Table 5.4: The mean of each duration-related acoustic cue of each prosodic boundary type (normalized value) and corresponding one-way ANOVA p-value and post hoc multi-comparisons results

For "syllable_final_percentage", prosodic phrase boundary (BI=2) has the largest mean value, while intonational phrase boundary (BI=3) has the smallest mean value, indicating that syllable-final lengthening is most pronounced in the last syllable before the prosodic phrase boundary than in the syllable before the prosodic word boundary and intonational phrase boundary.

For "silence_period_duration", intonational phrase boundary (BI=3) has larger mean value than prosodic phrase boundary (BI=2) than prosodic word boundary (BI=1). It indicates that, after intonational phrase boundaries, the silence period is the longest.

5.2 AUTOMATIC PROSODIC BOUNDARY LABELLING

5.2.1 *Results*

A total of 26 features are used in the automatic prosodic boundary labeling task[14], and the MaxEnt model and the SVM model are both used for training and testing.

When using the SVM model, I selected RBF (Radial basis function) as the kernel because the RBF kernel often works well in practice, and it is relatively easy to tune. The parameter gamma is set to be 0.5. As to the parameter "C" (cost), I have tried several different values, and the test accuracy is the highest when C=10. Table 5.5 shows different parameter settings of the SVM model and the corresponding test accuracy.

C (cost)	gamma	Overall Test Acc
0.01	0.5	0.72536
0.1	0.5	0.76309
1	0.5	0.78777
10	0.5	0.78864
100	0.5	0.78203

Table 5.5: Different parameter settings of SVM model and the corresponding test accuracy

Table 5.6 shows the test accuracy of the automatic prosodic boundary labelling task of the MaxEnt model and the best test accuracy of the SVM model (when C=10, gamma=0.5).[15] The test results of "most common class assignment "are also provided as the baseline results for comparison. Concretely, "most common class assignment" means labeling each test instance with its most frequently occurring label in the training data. To handle the test instances that do

[14] See Table 4.3 for the details of the features used in the task.
[15] "x-x Acc" is the accuracy of correctly distinguishing the two types of boundaries, e.g. "1-3 Acc" is the accuracy of correctly distinguishing prosodic word boundary (BI=1) and intonational phrase boundary (BI=3).

not appear in the training data, I set all the training instances that only appear once as "unknown". If the specific test instance does not appear in the training data, it will be labeled with the most common label of "unknown".

Model Name	Baseline	MaxEnt	SVM
Overall Test Acc	0.62936	0.6892	0.78864
1-3 Acc	0.85108	0.97715	0.99851
1-2 Acc	0.71443	0.69795	0.77552
2-3 Acc	0.74033	0.78805	0.88383
BI=1 Acc	0.88103	0.79472	0.92107
BI=2 Acc	0.30371	0.40403	0.43511
BI=3 Acc	0.45009	0.78301	0.90143

Table 5.6: The test accuracy of the automatic prosodic boundary labeling task of the Baseline model, the MaxEnt model and the SVM model (C=10, gamma=0.5)

5.2.2 *Discussion*

Overall, the MaxEnt model and the SVM model have higher test accuracy than the baseline model. And the SVM model has higher test accuracy than the MaxEnt model, which indicates that the SVM model performs better in this task. The better performance of the SVM model might be because the original features proposed in this work are real-valued features. The SVM model operates over real-valued continuous features directly, but the MaxEnt model interprets the numeric values as a string and operates over binary features, which might reduce the accuracy of the classification task.

For the baseline model, the MaxEnt model and the SVM model, the test accuracy of prosodic word boundary (BI=1) is the highest, while the accuracy of prosodic phrase boundary (BI=2) is the lowest. It indicates that it is most challenging to predict prosodic phrase boundaries

correctly. However, the accuracy of prosodic phrase boundary (BI=2) of the MaxEnt model and the SVM model are both higher than the baseline model, which indicates that the proposed features might be helpful to improve the accuracy of labeling prosodic phrase boundaries.

Besides, the three models all have the highest accuracy when distinguishing prosodic word boundary (BI=1) and intonational phrase boundary (BI=3). However, they all have the lowest accuracy when distinguishing prosodic word boundary (BI=1) and prosodic phrase boundary (BI=2). It indicates that it is easier to distinguish prosodic word boundaries from intonational phrase boundaries, but it is the hardest to distinguish prosodic word boundaries from prosodic phrase boundaries.

From the confusion matrix of the baseline model, the MaxEnt model and the SVM model (Table 5.6, Table 5.7 and Table 5.8) we can see that, the low accuracy when distinguishing prosodic word boundary (BI=1) and prosodic phrase boundary (BI=2) is mainly because many prosodic phrase boundaries (BI=2) are wrong labeled as prosodic word boundaries (BI=1).

However, the "1-2 Acc " of the SVM model is higher than the baseline model and than MaxEnt model, which indicates that employing the proposed features in an SVM achieve substantially better results at distinguishing prosodic word boundaries and prosodic phrase boundaries.

SYS / TRUE	BI=1	BI=2	BI=3
BI=1	3103	349	70
BI=2	1114	557	163
BI=3	600	287	726

Table 5.7: The confusion matrix of the baseline model

TRUE \ SYS	BI=1	BI=2	BI=3
BI=1	2799	688	35
BI=2	844	741	249
BI=3	60	290	1263

Table 5.8: The confusion matrix of the MaxEnt model

TRUE \ SYS	BI=1	BI=2	BI=3
BI=1	3244	277	1
BI=2	893	798	143
BI=3	6	153	1454

Table 5.9: The confusion matrix of the SVM model

5.3 Ablation Experiments I

5.3.1 *MaxEnt Model Results*

Table 5.10 shows the results of the ablation experiments of the MaxEnt model on all the 26 features. "Feature Index" is the index of the feature which removed from the training and testing data. The accuracy which drops compared to the accuracy of using all the 26 features is marked in red.

Feature Index	Overall Test Acc	1-3 Acc	1-2 Acc	2-3 Acc	BI=1 Acc	BI=2 Acc	BI=1 Acc
1	0.69321	0.97918	0.69756	0.79756	0.79671	0.40294	0.79727
2	0.69135	0.97583	0.70126	0.78898	0.79898	0.40349	0.78363
3	0.69235	0.97888	0.69858	0.79418	0.79699	0.40676	0.78859
4	0.68633	0.97804	0.69161	0.79091	0.78989	0.39804	0.78797
5	0.69178	0.97887	0.69643	0.79715	0.79642	0.40567	0.78859
6	0.69049	0.97839	0.69817	0.78952	0.79756	0.40185	0.78487
7	0.6945	0.97722	0.70256	0.79554	0.79642	0.41658	0.78797
8	0.69006	0.97791	0.69805	0.78912	0.79699	0.40185	0.78425
9	0.69364	0.97611	0.70289	0.79312	0.80324	0.40785	0.77929
10	0.69006	0.97651	0.69717	0.79279	0.79727	0.40076	0.78487
11	0.69307	0.979	0.69655	0.79826	0.79727	0.3964	0.80285
12	0.69206	0.97894	0.69507	0.79968	0.79699	0.39967	0.79541
13	0.69092	0.97934	0.69903	0.78782	0.79784	0.40294	0.78487
14	0.68905	0.97859	0.69628	0.78881	0.795	0.40076	0.78549
15	0.6925	0.97913	0.69913	0.79285	0.79756	0.40567	0.78921
16	0.68833	0.97832	0.69676	0.78648	0.79727	0.40076	0.77743
17	0.6945	0.97752	0.70293	0.794	0.80295	0.41003	0.78115
18	0.69178	0.97721	0.69894	0.79439	0.79841	0.40731	0.78239
19	0.6912	0.97837	0.69829	0.79206	0.79557	0.40676	0.78673
20	0.69077	0.97818	0.69685	0.79305	0.79671	0.40022	0.78983
21	0.69336	0.97906	0.70024	0.79506	0.79585	0.41658	0.78425
22	0.68876	0.97738	0.69683	0.78867	0.79387	0.40294	0.78425
23	0.69493	0.97849	0.70191	0.79513	0.79983	0.40785	0.79231
24	0.69235	0.97726	0.69829	0.79653	0.79642	0.40513	0.79169
25	0.69336	0.98041	0.70097	0.78836	0.80295	0.39695	0.79107
26	0.4906	0.74243	0.6218	0.57353	0.64963	0.27372	0.38996
No feature removed	0.6892	0.97715	0.69795	0.78805	0.79472	0.40403	0.78301

Table 5.10: The test accuracy of the ablation experiments of the MaxEnt model on all the 26 features

From the results, the overall test accuracy drops after removing feature #4, #14, #16, #22, and #26.

Among these four features, feature #26 has the most significant degree of decrease, which is much higher than the other three features. It means the duration-related feature "the duration of silence period" is much more important than other features, and indicates that the duration of silence period is the most crucial acoustic cue for distinguishing three types of prosodic

boundaries. Besides, "2-3 Acc" has the most considerable degree of decrease, indicating that feature #26 is more useful for distinguishing prosodic phrase boundaries and intonational phrase boundaries.

However, since feature #26 is dominant among all the 26 features, the accuracy of removing other features might be influenced by this dominant feature, and thus the results might not be credible. Therefore, I also ran the ablation experiments on 25 features (feature #26 excluded). Table 5.11 shows the results of the ablation experiments run on 25 features. The accuracy which drops compared to the accuracy of using all the 25 features is marked in red.

Feature Index	Overall Test Acc	1-3 Acc	1-2 Acc	2-3 Acc	BI=1 Acc	BI=2 Acc	BI=3 Acc
1	0.49132	0.73997	0.62394	0.57643	0.64991	0.27808	0.38748
2	0.49046	0.74248	0.62328	0.56954	0.65304	0.2759	0.37942
3	0.49218	0.74771	0.61718	0.58048	0.65133	0.26554	0.40236
4	0.49376	0.74594	0.61953	0.58593	0.65077	0.27208	0.40298
5	0.49218	0.74401	0.62112	0.57964	0.65048	0.27754	0.39058
6	0.49605	0.74383	0.62492	0.58681	0.65815	0.27317	0.39554
7	0.49003	0.74061	0.62316	0.57056	0.65048	0.27099	0.38872
8	0.4939	0.74319	0.62609	0.57634	0.65361	0.27045	0.39926
9	0.4896	0.74129	0.62182	0.57063	0.64991	0.27154	0.38748
10	0.49361	0.74518	0.62514	0.57432	0.65417	0.27317	0.39368
11	0.4896	0.73994	0.62068	0.57622	0.64679	0.27644	0.38872
12	0.48615	0.74352	0.61086	0.57633	0.64253	0.26827	0.39244
13	0.4962	0.74143	0.63022	0.58091	0.65872	0.28135	0.38562
14	0.48816	0.73864	0.61687	0.58223	0.64225	0.28626	0.38128
15	0.48902	0.73726	0.62145	0.57513	0.65247	0.27317	0.37756
16	0.49361	0.74704	0.62189	0.57598	0.65815	0.26063	0.39926
17	0.48257	0.73253	0.61767	0.56289	0.6448	0.26718	0.37322
18	0.48831	0.73813	0.61961	0.5746	0.6502	0.27099	0.3819
19	0.48988	0.74256	0.61933	0.5748	0.65077	0.2699	0.38872
20	0.49778	0.7507	0.62635	0.58028	0.6607	0.2759	0.3943
21	0.48744	0.73831	0.61607	0.57846	0.64424	0.26772	0.39492
22	0.49218	0.74407	0.6204	0.58138	0.64935	0.27863	0.39182
23	0.49376	0.74636	0.62244	0.58146	0.64566	0.28462	0.39988
24	0.49118	0.7451	0.62091	0.57274	0.65219	0.26936	0.39182
25	0.4873	0.73749	0.61914	0.57039	0.65247	0.26009	0.385
No feature removed	0.49031	0.74268	0.62143	0.57273	0.64935	0.27372	0.38934

Table 5.11: The test accuracy of the ablation experiments of the MaxEnt model on 25 features (feature #26 excluded)

The results of the ablation experiments on 25 features are quite different from the results of the ablation experiments on all the 26 features, which proves that feature #26 is dominant among all the 26 features.

From the results in Table5.11, the overall test accuracy decreases after removing feature #17, #12, #25, #21, #14, #18, #15, #9, #11, #19, and #7[16]. However, the magnitude of decrease in accuracies is rather small, and the accuracy of each ablation experiment is close, which indicates that these features may be highly correlated.

Among the pitch-related features, six (feature #7, #9, #11, #12, #19, and #21) are long-distance features that encode the difference of the pitch of prosodic unit after the prosodic boundary and the pitch of a prosodic unit after last intonational phrase boundary. It indicates that long-distance f0 variation (reset/declination) might be useful for measuring the degree of f0 variation after a prosodic boundary, and it might be a useful acoustic cue for distinguishing the three types of boundaries.

Besides, the accuracy of prosodic phrase boundary (BI=2) is affected most by removing the features[17], compared with prosodic word boundary (BI=1) and intonational phrase boundary (BI=3). It indicates that the proposed features might be helpful to improve the accuracy of labeling prosodic phrase boundaries when using the MaxEnt model.

Furthermore, according to the results of the ablation experiments, "1-2 Acc" and "1-3 Acc" is affected most by removing the features.[18] It indicates that the proposed features might be useful in distinguishing prosodic word boundaries from prosodic phrase boundaries and intonational phrase boundaries when using the MaxEnt model.

[16] The features are sorted in descending order according to the degree of decrease. See Table 4.3 for the details of the features.
[17] According to the number of ablation experiments in which the accuracy of the boundary type decreases.
[18] According to the number of ablation experiments in which "x-x Acc" decreases.

5.3.2 *SVM Model Results*

Table 5.12 shows the results of the ablation experiments of the SVM model on all the 26 features. "Feature Index" is the index of the feature which removed from the training and testing data. The accuracy which drops compared to the accuracy of using all the 26 features is marked in red.

Feature Index	Overall Test Acc	1-3 Acc	1-2 Acc	2-3 Acc	BI=1 Acc	BI=2 Acc	BI=3 Acc
1	0.78806	0.99894	0.77158	0.88976	0.91539	0.43566	0.91073
2	0.78706	0.99894	0.77062	0.88849	0.91709	0.42966	0.90949
3	0.78777	0.99894	0.771	0.8895	0.9188	0.42748	0.91135
4	0.78792	0.99894	0.77109	0.89016	0.91709	0.43184	0.91073
5	0.78749	0.99894	0.77153	0.88793	0.91709	0.43184	0.90887
6	0.78749	0.99894	0.771	0.88893	0.91738	0.43021	0.91011
7	0.78835	0.99894	0.77211	0.88902	0.91851	0.43075	0.91073
8	0.78763	0.99894	0.7713	0.88867	0.91709	0.43075	0.91073
9	0.78835	0.99894	0.77181	0.88976	0.91823	0.4313	0.91073
10	0.78749	0.99894	0.77043	0.8902	0.91596	0.4313	0.91197
11	0.7872	0.99894	0.77091	0.8881	0.91766	0.42857	0.91011
12	0.78964	0.99894	0.77339	0.89055	0.91794	0.43675	0.91073
13	0.78921	0.99894	0.77281	0.89037	0.91823	0.43457	0.91073
14	0.78806	0.99894	0.77205	0.88867	0.91794	0.43293	0.90825
15	0.78749	0.99894	0.77096	0.88858	0.91965	0.4253	0.91073
16	0.78864	0.99894	0.77234	0.88933	0.91936	0.43021	0.91073
17	0.7872	0.99894	0.77028	0.88972	0.91624	0.43075	0.91073
18	0.78878	0.99915	0.77196	0.89024	0.91823	0.4313	0.91259
19	0.78734	0.99894	0.77115	0.8881	0.91794	0.42912	0.90949
20	0.78806	0.99894	0.77143	0.88985	0.91709	0.43239	0.91073
21	0.78777	0.99894	0.77128	0.88959	0.91596	0.43457	0.90949
22	0.78864	0.99894	0.77205	0.89029	0.91766	0.43348	0.91073
23	0.78634	0.99893	0.77034	0.88714	0.91539	0.4313	0.90825
24	0.78864	0.99894	0.77196	0.89024	0.91794	0.43184	0.91197
25	0.78749	0.99894	0.77153	0.88771	0.91851	0.42912	0.90887
26	0.59966	0.81945	0.70431	0.67744	0.85321	0.13304	0.57657
No feature removed	0.78864	0.99851	0.77552	0.88383	0.92107	0.43511	0.90143

Table 5.12: The test accuracy of the ablation experiments of the SVM model on all the 26 features

From the results, the test accuracy decreases after removing feature #1, #2, #3, #4, #5, #6, #7, #8, #9, #10, #11, #14, #15, #17, #19, #20, #21, #23, #25, and #26. Similar to the ablation experiments results of the MaxEnt model, the accuracy after removing feature #26 (the duration of silence period) decreases most significantly. It also indicates that the duration of the silence period is the most important acoustic cue for distinguishing three types of prosodic boundaries. Also, similar to the ablation experiments result of the MaxEnt model, "2-3 Acc" has the greatest degree of decrease, indicating that feature #26 is more useful for distinguishing prosodic phrase boundaries and intonational phrase boundaries.

In order the reduce the dominant effect of feature #26, I also ran the ablation experiments on 25 features (feature #26 excluded). Table 5.11 shows the results of the ablation experiments run on 25 features. The accuracy which drops compared to the accuracy of using all the 25 features is marked in red.

Feature Index	Overall Test Acc	1-3 Acc	1-2 Acc	2-3 Acc	BI=1 Acc	BI=2 Acc	BI=3 Acc
1	0.59994	0.81937	0.70516	0.67705	0.85378	0.13522	0.57409
2	0.59923	0.81974	0.7043	0.67496	0.85434	0.13468	0.57037
3	0.59966	0.81956	0.70398	0.67768	0.85463	0.13141	0.57533
4	0.59808	0.81828	0.70361	0.67418	0.85207	0.12923	0.57657
5	0.59923	0.81828	0.70401	0.67903	0.85292	0.13359	0.57471
6	0.59937	0.81987	0.70381	0.67596	0.85548	0.13086	0.57285
7	0.59994	0.82023	0.70439	0.67705	0.85378	0.13522	0.57409
8	0.59894	0.81907	0.70456	0.67399	0.85406	0.13086	0.57409
9	0.59779	0.81807	0.70371	0.67378	0.85094	0.13359	0.57285
10	0.59836	0.81907	0.70421	0.67188	0.85434	0.12868	0.57347
11	0.59951	0.81928	0.70538	0.67434	0.85349	0.1325	0.57595
12	0.59793	0.81778	0.70402	0.6738	0.85236	0.13086	0.57347
13	0.60023	0.81822	0.70647	0.67779	0.85321	0.13577	0.57595
14	0.5998	0.82036	0.70394	0.67725	0.85378	0.13522	0.57347
15	0.59894	0.81852	0.70362	0.67772	0.85434	0.13141	0.57285
16	0.60109	0.81905	0.70684	0.6782	0.85548	0.13195	0.57905
17	0.60109	0.82165	0.70451	0.6788	0.85576	0.13631	0.57347
18	0.59808	0.81801	0.70325	0.67525	0.85576	0.12814	0.56975
19	0.59923	0.81926	0.70464	0.67438	0.85463	0.12923	0.57595
20	0.59793	0.81833	0.70352	0.67379	0.85179	0.13032	0.57533
21	0.59836	0.81835	0.70341	0.67575	0.85321	0.12923	0.57533
22	0.59966	0.81784	0.70574	0.67688	0.85406	0.12923	0.57905
23	0.59822	0.81873	0.70312	0.67569	0.85065	0.13304	0.57595
24	0.59722	0.8175	0.70219	0.67575	0.85094	0.12977	0.57471
25	0.59923	0.81998	0.70375	0.67586	0.85179	0.13359	0.57719
No feature removed	0.59966	0.81962	0.70416	0.67744	0.85321	0.13304	0.57657

Table 5.13: The test accuracy of the ablation experiments of the SVM model on 25 features (feature #26 excluded)

From the results in Table 5.12, the overall test accuracy decreases after removing feature #24, #9, #12, #20, #4, #18, #23, #10, #21, #8, #15, #2, #5, #19, #25, #6, and #11[19]. However, the magnitude of decrease in accuracies is rather small, and the accuracy of each ablation experiment is close, which indicates that these features may be highly correlated.

[19] The features are sorted in descending order according to the degree of decrease. See Table 4.3 for the details of the features.

Among the pitch-related features, feature #24, #9, #12, and #20 have comparatively larger degree of decrease. These four features are all long-distance features which encode the difference of the pitch of a prosodic unit after the boundary and the pitch of the prosodic unit after last intonational phrase boundary. It supports that long-distance f0 variation (reset/declination) might also be useful for measuring the degree of f0 variation after a prosodic boundary, and it might be a useful acoustic cue for distinguishing the three types of boundaries.

Besides, the accuracy of prosodic phrase boundary (BI=3) is affected most by removing the features, compared with prosodic word boundary (BI=1) and intonational phrase boundary (BI=2), which indicates that the proposed features might be helpful to improve the accuracy of labeling intonational phrase boundaries when using the SVM model.

Furthermore, according to the results of the ablation experiments, "1-3 Acc" and "2-3 Acc" is affected most by removing the features[20]. It indicates that the proposed features might be useful in distinguishing intonational phrase boundaries from prosodic word boundaries and prosodic phrase boundaries when using the SVM model.

5.3.3 *Discussion*

In general, the ablation experiments results of the MaxEnt model do not precisely match the ablation experiments results of the SVM model.

When looking at the ablation experiments results of the two models separately, for the MaxEnt model, the proposed features are more helpful to improve the accuracy of labeling prosodic phrase boundaries and are more helpful in distinguishing prosodic word boundaries from prosodic phrase boundaries and intonational phrase boundaries; while for the SVM model, the proposed features are more useful to improve the accuracy of labeling intonational phrase

[20] According to the number of ablation experiments in which "x-x Acc" decreases.

boundaries and are more useful in distinguishing intonational phrase boundaries from prosodic word boundaries and prosodic phrase boundaries. However, overall, the SVM model has better performance in distinguishing different types of prosodic boundaries.

Nevertheless, it is acceptable because firstly, the features are encoded in different ways in these two models - MaxEnt is using categorical features, but SVM is using numeric features. Besides, even the features are encoded in the same way, it is common that one feature is good for one model but is not good for the other, since different features suit different models. So, these all might be the reasons for these two models performing differently on the prosodic boundary labeling task.

However, the ablation experiments results of the MaxEnt model and the ablation experiments results of the SVM model did have some overlapping parts.

Firstly, for both the MaxEnt model and the SVM model, the accuracy after removing the duration-related feature – feature #26 decreases most significantly, which indicates that the duration of the silence period is the most critical acoustic cue for distinguishing three types of prosodic boundaries. Also, for the two models, "2-3 Acc" has the most considerable degree of decrease, indicating that feature #26 is more useful for distinguishing prosodic phrase boundaries and intonational phrase boundaries. Besides, the accuracy after removing another duration-related feature – feature #25 (the percentage of the duration of the final of the last syllable before the boundary and the duration of the syllable) also decreases for both MaxEnt model and SVM model, indicating that duration-related features are essential for distinguishing different boundary types.

Then, according to the overall test accuracy of the ablation experiments on 25 features, among 24 pitch-related features, the accuracy after removing feature

#11 (f0_next_syllable_mean – f0_boundary_next_prosodic_word_mean),

#12 (f0_next_syllable_minimum – f0_boundary_next_prosodic_word_minimum),

#15 (f0_next_prosodic_word_minimum – f0_last_syllable_minimum),

#18 (f0_next_prosodic_word_minimum – f0_last_prosodic_word_minimum),

#19 (f0_next_prosodic_word_maximum – f0_boundary_next_syllable_maximum),

and #21 (f0_next_prosodic_word_minimum – f0_boundary_next_syllable_minimum) drops for both MaxEnt model and SVM model, which indicates that these features suit both of the two models, and are useful for distinguishing the three types of prosodic boundaries. It should be noted that, four of the features are the difference of the minimum pitch of the prosodic units, which indicates that the variation of the minimum pitch might be a more useful acoustic cue for distinguishing different boundary types.

Besides, both feature #11, #12, #19, and #21 are long-distance features, which are the difference of the pitch of a prosodic unit after the boundary and the pitch of a prosodic unit after last intonational phrase boundary. It indicates that long-distance f0 variation(reset/declination) might also be useful for measuring the degree of f0 variation after a prosodic boundary, and it might be a useful acoustic cue for distinguishing the three types of boundaries.

Furthermore, feature #15 and #18 are the difference of the minimum pitch of the prosodic word after the boundary and the minimum pitch of the syllable/prosodic word before the boundary. It should also be noted that according to the post hoc multi-comparisons results in Section 5.1.1, "next_prosodic_word – last_ prosodic_word (minimum)" and "next_ prosodic_word – last_syllable (minimum)" (corresponding to the feature #15 and #18) also have the largest mean difference (3-1), mean difference (3-2), and mean difference (2-1). It indicates that different types of boundaries can be better distinguished by the pitch difference of the

prosodic word after the boundary and a prosodic unit before the boundary. It also implies that the f0 variation (reset/declination) of two adjacent prosodic units is mainly reflected in the minimum pitch difference of the prosodic word after the boundary and the prosodic unit before the boundary. It should also be noticed that feature #18 is the feature that describes the baseline variation (reset/declination), thus it might support the previous linguistic finding that baseline variation is an important acoustic cue at prosodic boundaries in Mandarin.

However, overall, the magnitude of decrease in accuracies in the ablation experiments on 25 features is rather small, and the accuracy of each ablation experiment is close. It indicates that these features may be highly correlated. Therefore, I ran another three sets of ablation experiments to examine further the role of different types of pitch-related features.

5.4 Ablation Experiments II

5.4.1 *Syllable After the boundary VS. Prosodic Word After the boundary*

Table 5.14 shows the test accuracy of the prosodic boundary labeling task of the SVM model in two different feature configurations. The first configuration includes feature #1 to #12. These features are the pitch difference of the syllable after the boundary and a prosodic unit before the boundary. The second configuration uses feature #13 to #24, which are the pitch difference of the prosodic word after the boundary and a prosodic unit before the boundary. For comparison, the test accuracy when using all the 24 pitch-related features is also included in the table, as shown in the column "Original".

	Original	Configuration 1	Configuration 2
Overall Test Acc	0.59923	0.55072	0.57426
1-3 Acc	0.81998	0.75746	0.78507
1-2 Acc	0.70375	0.68341	0.69513
2-3 Acc	0.67586	0.61993	0.6443
BI=1 Acc	0.85179	0.89012	0.88274
BI=2 Acc	0.13359	0.07143	0.04907
BI=3 Acc	0.57719	0.35462	0.49783

Table 5.14: The test accuracy of the prosodic boundary labeling task of the SVM model when using feature #1 to #12 and feature #13 to #24

The test accuracy when using feature #1 to #12 and feature #13 to #24 are both lower than the test accuracy when using all the 24 pitch-related features, which indicates that these two sets of features are both useful in distinguishing different boundary types.

Besides, the test accuracy when using feature #13 to #24 is higher than the test accuracy when using feature #1 to #12, which indicates that different types of boundaries can be better distinguished by the pitch difference of the prosodic word after the boundary and a prosodic unit before the boundary.

Moreover, the intonational phrase labeling accuracy when using feature #13 to #24 is much higher than the accuracy when using feature #1 to #12. However, the prosodic word labeling accuracy and the prosodic phrase labeling accuracy when using feature #13 to #24 are a bit lower than the accuracies when using feature #1 to #12. It indicates that feature #13 to #24 are more useful in predicting intonational phrase boundaries. It also indicates that the f0 variation at intonational phrase boundaries is mainly reflected in the pitch difference of the prosodic word after the boundary and a prosodic unit before the boundary.

5.4.2 *Long-distance Features VS. Non-long-distance Features*

Table 5.15 shows the test accuracy of the prosodic boundary labeling task of the SVM model in two different feature configurations. The first configuration only includes long-distance features (feature #7, #8, #9, #10, #11, #12, #19, #20, #21, #22, #23, and #24). The second configuration only keeps non-long-distance features (Feature #1, #2, #3, #4, #5, #6, #13, #14, #15, #16, #17, and #18). For comparison, the test accuracy when using all the 24 pitch-related features is also included in the table, as shown in the column "Original".

	Original	**Configuration 1**	**Configuration 2**
Overall Test Acc	0.59923	0.5193	0.57081
1-3 Acc	0.81998	0.70616	0.77967
1-2 Acc	0.70375	0.66335	0.69209
2-3 Acc	0.67586	0.65278	0.64451
BI=1 Acc	0.85179	0.96082	0.89267
BI=2 Acc	0.13359	0.01036	0.02726
BI=3 Acc	0.57719	0.13391	0.48605

Table 5.15: The test accuracy of the prosodic boundary labeling task of the SVM model when only using long-distance features and non-long-distance features

The test accuracy when only using long-distance features and non-long-distance features are both lower than the test accuracy when using all the 24 pitch-related features, which indicates that both long-distance features and non-long-distance features are useful for distinguishing different boundary types.

However, the test accuracy when only using long-distance features is lower than the test accuracy when only using non-long-distance features. The intonational phrase labeling accuracy when only using long-distance features is much lower than the accuracy when only using non-

long-distance features. It is mainly because the long-distance features proposed in this work are the pitch difference of the prosodic unit after the boundary and the prosodic unit before last intonational phrase boundary. Therefore, the overall test accuracy relies much on the intonational phrase labeling accuracy. If one previous intonational phrase boundary is wrongly predicted, then it will affect the prediction of latter prosodic boundaries. Thus, it indicates that long-distance features should be used together with other features to achieve better results in prosodic boundary labeling.

5.4.3 *Minimum Pitch VS. Mean Pitch VS. Maximum Pitch*

Table 5.16 shows the test accuracy of the prosodic boundary labeling task of the SVM model in three different feature configurations. The first configuration only includes minimum pitch based features (Feature #3, #6, #9, #12, #15, #18, #21, and #24). The second configuration only keeps mean pitch based features (Feature #2, #5, #8, #11, #14, #17, #20, and #23). The third configuration only uses maximum pitch based features (Feature #1, #4, #7, #10, #13, #16, #19, and #22). For comparison, the test accuracy when using all the 24 pitch-related features is also included in the table, as shown in the column "Original".

	Original	Configuration 1	Configuration 2	Configuration 3
Overall Test Acc	0.59923	0.56909	0.5622	0.51571
1-3 Acc	0.81998	0.77847	0.76549	0.70094
1-2 Acc	0.70375	0.69172	0.6807	0.65744
2-3 Acc	0.67586	0.639	0.67458	0.76142
BI=1 Acc	0.85179	0.90091	0.88643	0.97785
BI=2 Acc	0.13359	0.04144	0.00763	0.00055
BI=3 Acc	0.57719	0.44451	0.48481	0.09237

Table 5.16: The test accuracy of the prosodic boundary labeling task of the SVM model when only using minimum pitch based features, mean pitch based features, and maximum pitch based features

The test accuracy when only using minimum pitch based features, mean pitch based features and maximum pitch based features are all lower than the test accuracy when using all the 24 pitch-related features, which indicates that these three sets of features are all useful for distinguishing different boundary types. However, the test accuracy when only using maximum pitch based features is the lowest. It indicates that maximum pitch differences are not as helpful as minimum pitch differences and mean pitch differences.

The test accuracy when only using minimum pitch based features is a bit higher than the test accuracy when only using mean pitch based features. It indicates that minimum pitch based features might be more useful for distinguishing different boundary types.

Besides, the prosodic word labeling accuracy and the prosodic phrase labeling accuracy when using the minimum pitch based features are higher than the accuracies using the mean pitch based features. However, the intonational phrase labeling accuracy when using the mean pitch based features is higher. It indicates that the f0 variation at prosodic word boundaries and prosodic phrase boundaries might be mainly reflected in the variation of the minimum pitch;

while at intonational phrase boundaries, the f0 variation might be mainly reflected in the variation of the mean pitch.

Chapter 6. CONCLUSIONS AND FUTURE WORK

In this work, I studied the acoustic cues at different types of prosodic boundaries in Mandarin. One of my aims is to verify and concretize the findings in previous linguistic studies on the acoustic cues at the prosodic boundaries in Mandarin with a larger corpus and quantitative methods. I also would like to study the relation of the acoustic cues to different boundary types and explore the acoustic cues that can be useful to distinguish different boundary types through automatic prosodic boundary labeling task and feature ablation experiments.

6.1 CONCLUSIONS

The results of the one-way ANOVA test verified that the silence duration is an essential acoustic cue at the prosodic boundaries. However, the one-way ANOVA results seem to counter the previous linguistic study about baseline declination, as the baseline is reset only after intonational phrase boundaries, but it slightly declines after prosodic phrase boundaries. The one-way ANOVA test result also indicates that different ways of measuring the pitch difference can lead to different conclusions of whether f0 after a prosodic boundary is reset or declines. However, we can reach an overall conclusion that the pitch difference of the prosodic units before and after an intonational phrase boundary is the largest, while the prosodic phrase boundary is the second largest, and the prosodic word boundary is the smallest. It means if there is an f0 reset after the boundary, the unit after an intonational phrase boundary has the largest degree of reset. If f0 declines after the boundary, the declination after a prosodic word boundary is more pronounced than after a prosodic phrase boundary than after an intonational phrase boundary.

The results of automatic prosodic boundary labeling task and feature ablation experiments also verifies the importance of silence duration as a duration-related acoustic cue and indicates that duration-related features are essential for distinguishing different boundary types.

Besides, the results of feature ablation experiments also provide some other inspiring information on pitch-related acoustic cues.

Firstly, long-distance f0 variation (reset/declination) might be useful for measuring the degree of f0 variation after a prosodic boundary and might be a useful acoustic cue for distinguishing different boundary types. However, long-distance features should be used together with other features to achieve better results in prosodic boundary labeling.

Secondly, the pitch difference of the prosodic word after the boundary and a prosodic unit before the boundary might be more helpful to distinguish different boundary types.

Thirdly, the maximum pitch differences are not as useful as minimum pitch differences for distinguishing different boundary types. The f0 variation (reset/declination) at prosodic word boundaries and prosodic phrase boundaries might be mainly reflected in the variation of the minimum pitch; while at intonational phrase boundaries, the f0 variation might be mainly reflected in the variation of the mean pitch.

Last but not least, the f0 variation (reset/declination) of two adjacent prosodic units might be mainly reflected in the minimum pitch difference of the prosodic word after the boundary and the prosodic unit before the boundary. The results might also support the previous linguistic finding that baseline variation (reset/declination) is an important acoustic cue at prosodic boundaries in Mandarin.

Furthermore, the results of the one-way ANOVA test and automatic prosodic boundary labeling task both indicate that there is most confusion between prosodic word boundaries and

prosodic phrase boundaries, and it is most difficult to distinguish these two types of boundaries. Employing our proposed features in a SVM achieve substantially better results at distinguishing these two types of boundaries.

6.2 Future Work

In this work, I have found some acoustic cues which can be useful to distinguish different prosodic boundary types in Mandarin. However, since I only use the features in the MaxEnt model and the SVM model, it is hard to predict the performance of these features in other machine learning models since each machine learning model has its own characteristics. Thus, it is worth applying these features in other machine learning models and analyzing their performance in future work. Besides, in order to improve the accuracy of the prosodic boundary labeling task, there might be better ways to encode the proposed features in different models, and it can be a worth studying question as well.

Furthermore, in this work, I only focus on the acoustic cues at prosodic boundaries in Mandarin read speech, rather than spontaneous speech. However, the prosody of spontaneous speech could be very different from read speech. Therefore, in future work, I will also try to apply this experimental framework to study the acoustic cues at prosodic boundaries in Mandarin spontaneous speech, and compare the prosody of Mandarin read speech and the prosody of Mandarin spontaneous speech.

BIBLIOGRAPHY

[Ding et al., 2015] Ding, C., Xie, L., Yan, J., Zhang, W., and Liu, Y. 2015. Automatic prosody prediction for Chinese speech synthesis using BLST-RNN and embedding features. In *proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 98-102.

[Feng, 1997] Feng, S. 1997. *Interactions between Morphology Syntax and Prosody in Chinese*. Beijing: Peking University Press.

[Fu et al., 2018] Fu, R., Tao, J., Li, Y., and Wen, Z. 2018. Automatic prosodic boundary labeling based on fusing the silence duration with the lexical features. *Journal of Tsinghua University (Science & Technology)*, 58(1): 61-66.

[He et al., 2001] He, L., Chu M., Lv S., Qian, Y., and Feng, Y. 2001. A research on prosodic boundary labelling of Mandarin synthesis corpora. In *proceedings of 5th National Conference on Modern Phonetics*.

[Ladd, 1996] Ladd, D. Robert. 1996. *Intonational Phonology.* Cambridge: Cambridge University Press.

[Li, 1997] Li, Z. 1997. A Preliminary Study of the Prosodic Transcription System for Standard Chinese. In *Proceedings of Third National Conference of Intelligence Interface and Its Application.*

[Li, 2002] Li, A. 2002. Chinese prosody and prosodic labeling of spontaneous speech. In *proceedings of Speech Prosody*, pages: 39-46.

[Lin, 2000] Lin, M. 2000. Breaks and prosodic phrases in the utterances of Standard Chinese. *Contemporary Linguistics*. 2(4): 210-217.

[Lin, 2002] Lin, M. 2002. Prosodic structure and lines of F0 top and bottom of utterances in Chinese. *Contemporary Linguistics*. 4(4): 254-265.

[Ni et al., 2012] Ni, C., Zhang, A., Liu, W., and Xu, B. 2012. Automatic prosodic break detection and feature analysis. *Journal of Computer Science and Technology*, pages: 1184-1196.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. 2011. Scikit-learn: Machine Learning in Python. *JMLR*. 12(85): 2825−2830.

[Peng et al., 2014] Peng, H., Chen, C., Tseng, C., and Chen, K. 2014. Predicting prosodic words from lexical words - a first step towards predicting prosody from text. In *Proceedings of 2004 International Symposium on Chinese Spoken Language Processing,* pages: 173-176

[Qian et al., 2010] Qian, Y., Wu, Z., Ma, X., and Soong, Frank. 2010. Automatic prosody prediction and detection with Conditional Random Field (CRF) models. In *Proceedings of 2010 7th International Symposium on Chinese Spoken Language Processing*, pages: 135-138.

[Shih and Ao, 1994] Shih, C. and Ao, B. 1994. Duration study for the AT&T Mandarin Text-to-speech System. In *proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages: 29-32

[Sun et al., 2009] Sun, J., Yang, J., Zhang, J., and Yan, Y. 2009. Chinese prosody structure prediction based on Conditional Random Fields. In *proceedings of 2009 Fifth International Conference on Natural Computation*. 3: 602-606.

[Tseng and Chou, 1999] Tseng, C. and Chou, F. 1999. A prosodic labeling system for Mandarin speech database. In *Proceedings of the XIV International Congress of Phonetic Science*, pages: 2379-2382.

[Tseng et al., 2005] Tseng, C., Pin, S., Lee, Y., Wang, H., and Chen, Y. 2005. Fluent speech prosody: Framework and modeling. *Speech Communication.* 46: 284-309.

[Wang et al., 2004] Wang, B., Lv, S., and Yang, Y. 2004. The acoustic analysis of Mandarin prosodic boundaries. *ACTA ACUSTICA, Chinese version*. 29(1): 29-36.

[Xie et al., 2018] Xie, K. and Pan, W. 2018. Mandarin prosody prediction based on attention mechanism and multi-model ensemble. In book *Intelligent Computing Theories and Application*, pages: 491-502.

[Zhao et al., 2003] Zhao, S., Tao, J., and Jiang, D. 2003. Chinese prosody phrasing with extended feature. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*.