

**IMPROVING MISPRONUNCIATION DETECTION AND ENRICHING
DIAGNOSTIC FEEDBACK FOR NON-NATIVE LEARNERS OF MANDARIN**

A Dissertation
Presented to
The Academic Faculty

By

Wei Li

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

December 2019

Copyright © Wei Li 2019

**IMPROVING MISPRONUNCIATION DETECTION AND ENRICHING
DIAGNOSTIC FEEDBACK FOR NON-NATIVE LEARNERS OF MANDARIN**

Approved by:

Professor Chin-Hui Lee, Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor David V Anderson
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Elliot Moore II
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Sabato Marco Siniscalchi
School of Electrical and Computer
Engineering
University of Enna Kore

Professor Jin Liu
School of Modern Languages
Georgia Institute of Technology

Date Approved: October 09, 2019

To my parents and girlfriend, for their boundless love and support.

ACKNOWLEDGEMENTS

I would never have been able to reach the destination of my PHD journey without the support of my advisors, committee members, colleagues, and family members.

First, I would like to express my most sincere gratitude to my advisor, Prof. Chin-Hui Lee, who always gave me high level research insights, and continuous guidance on my weekly reports. More importantly, Prof. Lee always gave me great freedom to pursue the research topic that I am most interested in. I still remember the moment I met with Prof. Lee six years ago, when he was invited to give a talk in China. Inspired by his work, I wrote a research plan on computer-assisted pronunciation training (CAPT), and successfully got his approval of my PHD application. During my five years studying at Georgia Tech, Prof. Lee created an excellent environment in which to let me focus on CAPT problems, even though no specific funding on this research topic was available. I also appreciate his support of my startup in China, and when the company did not work well, he encouraged me to come back to Georgia Tech to finish my PHD thesis.

Second, I would like to show my great appreciation for my advisor Prof. Sabato Marco Siniscalchi, who gave me invaluable comments on my proposed methods and research papers. In short, he is my trusted mentor and best friend. Moreover, I am also grateful to Dr. Nancy Chen, who shared the non-native Mandarin corpus extensively used in this thesis. We kept weekly meetings for three years, and each time she listened to my research progress and gave me precious suggestions.

I would also like to thank my committee members for serving on my dissertation committee and giving me many suggestions: Prof. David V Anderson, Prof. Elliot Moore II, Prof. Jin Liu, and Prof. Siniscalchi.

During the last five years at Georgia Tech, I owe many thanks to my colleagues. Thanks to Dr. You-Chi Cheng and Dr. I-Fan Chen for helping me settle down at Atlanta. Thanks to Dr. Zhen Huang, Dr. Kehuang Li, and Sicheng Wang for your assistance of my research.

I would also like to extend my thanks to other friends at Georgia Tech: Dr. Di Wu, Dr. Zhong Meng, Dr. Zhen Wang, Dr. Yuting Hu, Hu Hu, Jun Qi, Yongliang He, Sam Li. Finally, thanks to Pat Dixon, Raquel Plaskett, Jennifer Lunsford, Dr. Daniela Staiculescu, and Tasha Torrence for their great administrative support.

Last but not least, I would like to thank my parents for always standing by my side. Although they are not well-educated, my parents convinced me that knowledge can change my fate. Finally, I want to express my deepest gratitude to my girlfriend, Huixiang Xu, for her constant love and support, which get me fully prepared for the rest of my life.

TABLE OF CONTENTS

Acknowledgments	v
List of Tables	xii
List of Figures	xiii
Chapter 1: Introduction	1
1.1 Overview	1
1.2 Main Contributions	4
1.3 Thesis Outline	5
Chapter 2: Background and Literature Survey	7
2.1 Mispronunciation Types	7
2.1.1 Phonetic Errors	7
2.1.2 Prosodic Errors	8
2.2 Mispronunciation Detection and Corrective Feedback Generation	9
2.2.1 Detecting and Diagnosing Phonetic Errors (phone substitution) . . .	9
2.2.2 Detecting and Diagnosing Prosodic Errors (tone substitution) . . .	12
2.2.3 Corrective Feedback Generation	13
2.3 Performance Measurement of Mispronunciation Detection	13

2.4	Speech Corpora	15
2.5	Summary	16
Chapter 3: Visualizing Non-native Pronunciation Through Phone and Its Attrib- tribute Detection		17
3.1	Introduction	17
3.2	Mandarin Phones and Speech Attributes	19
3.3	Speech Attribute and Phone Classifiers Training	21
3.4	Speech Attribute and Phone Feature Extraction	22
3.5	Experiments	23
3.5.1	Experimental Setup	23
3.5.2	Experimental Results and Discussions	24
3.6	Visible Non-native Pronunciation Analysis	27
3.7	Summary	30
Chapter 4: Improving Mispronunciation Detection of Mandarin Phones with Multisource Information and BLSTM-based Mispronunciation De- tectors		31
4.1	Introduction	31
4.2	Overview of The Phone Mispronunciation Detection Framework	33
4.2.1	Speech Attribute and Phone Feature Extraction	34
4.2.2	DNN-based Mispronunciation Verifier Construction	34
4.2.3	BLSTM-based Mispronunciation Verifier Construction	35
4.3	Experiments	36
4.3.1	Experimental Setup	37

4.3.2	Experimental Results and Discussions	39
4.4	Summary	44
Chapter 5: Improving Mispronunciation Detection of Mandarin Tones with Soft Target Tone Labels and BLSTM-based Models	45	
5.1	Introduction	45
5.2	Modeling Challenges of Non-native Lexical Tones	49
5.2.1	Variability in Pitch Contour of L2 Tone Productions	49
5.2.2	Prolonged and High-Variance Duration in L2 Tones	50
5.3	Overview of The Tone Mispronunciation Detection Framework	51
5.3.1	Acoustic Tonal Model Training with Hard Target	51
5.3.2	Mandarin Tones Soft Target Generation	52
5.3.3	Acoustic Tonal Model Training with Soft Target	53
5.3.4	Non-native Tone Feature Extraction	53
5.3.5	Tone Mispronunciation Verifier Construction	54
5.4	Experiments	54
5.4.1	Experimental Setup	55
5.4.2	Experimental Results and Discussions	57
5.5	Summary	67
Chapter 6: Diagnosing Non-native Phonetic Mispronunciations and Providing Articulatory-level Feedback with Knowledge-guided and Data-driven Based Decision Trees	68	
6.1	Introduction	68
6.2	Overview of The Decision Tree Based Diagnosis Framework	71

6.2.1	Speech Attribute Feature Extraction	71
6.2.2	Tree-based Mispronunciation Verifier Construction	72
6.3	Experiments	73
6.3.1	Experimental Setup	73
6.3.2	Experimental Results and Discussions	74
6.4	Summary	77
Chapter 7: Conclusions and Future Work		78
7.1	Summary and Contributions	78
7.2	Future Work and Directions	80
Appendix A: Phone-dependent Decision Trees		82
References		98
Vita		99

LIST OF TABLES

3.1	Speech attributes and their associated Pinyin initials (consonants)	20
3.2	Classification accuracy for each phone category in the test set, where the overall performance is 92.1%	24
3.3	Classification accuracy for each manner attribute in the test set, where the overall performance is 97.0%	25
3.4	Classification accuracy for each place attribute in the test set, where the overall performance is 94.9%	25
3.5	Classification accuracy for each aspiration attribute in the test set, where the overall performance is 96.0%	26
3.6	Classification accuracy for each voicing attribute in the test set, where the overall performance is 99.4%	26
4.1	The number of each phone's correct and mispronounced samples in our experiments, where the overall mispronunciation rates in the train and test sets are 3.3% and 19.2%, respectively	38
4.2	The mispronunciation detection performance for individual phones, where the precision is set the same as recall	43
5.1	Characteristic of pitch contours of Mandarin tones	46
5.2	Confusion Matrix of Mandarin tone accuracy (%) on iCALL corpus analyzed in [55]	49
5.3	The mean and standard deviation (in milliseconds) of the duration of the four lexical tones where the underlying phone's identity is ignored	51

5.4	The mean and standard deviation (in milliseconds) of the duration of the four lexical tones where the underlying phone's identity is /A/	51
5.5	The number of each tone's correct and mispronounced samples in our experiments, where the overall mispronunciation rates in the train and test sets are 26.3% and 53.6%, respectively	55
5.6	Tone posteriors produced by DNN and BLSTM for Figure 5.6.	59
5.7	Tone posteriors produced by DNN and BLSTM for Figure 5.7.	60
5.8	Tone posteriors of hard targets and proposed soft targets for Figure 5.8. . . .	61
5.9	Tone posteriors of hard targets and proposed soft targets for Figure 5.10. . .	63
5.10	Tones EERs (%) of different verifiers with features extracted from different tone models, where * denotes soft target training. If none, hard target is utilized.	66
6.1	The individual phone's diagnostic error rate (DER) of two different decision trees on test set, where the overall DERs of KGBDT and DDBDT are 4.1% and 4.6%, respectively	74

LIST OF FIGURES

2.1	Hierarchical structure of the metrics for evaluating the systems' mispronunciation detection and diagnostic performance (adapted from [8, 87]).	14
3.1	Overview of the phone and speech attribute classifiers training.	21
3.2	Segmental posterior histograms of Mandarin consonant /J/ computed on native (upper) and non-native data (lower).	27
3.3	Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /J/, /P/, and the human assigned labels are /Q/, /P/.	28
3.4	Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /B/, /P/, /T/, and the human assigned labels are /B/, /B/, /T/.	30
4.1	Overview of the phone mispronunciation detection framework.	34
4.2	Comparison of the overall mispronunciation detection performance of the baseline and our proposed systems.	40
4.3	Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /S/, /K/, and the human assigned labels are /C/, /K/.	41
4.4	Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /N/, /G/, /B/, and the human assigned labels are /N/, /G/, /P/.	42
5.1	Pitch contours of standard Mandarin lexical tones. Tone 5 is not depicted as it does not have a defined pitch contour.	46

5.2	One example of pitch contours of non-native lexical tones, where human assigned labels are Tone1 and Tone1	50
5.3	Overview of the tone mispronunciation detection framework.	52
5.4	An example of tone-based ERN.	53
5.5	Tone-dependent detection curves and EERs for DNN-based tone mispronunciation verifiers trained with segmental tone features vectors extracted from different acoustic tonal models	58
5.6	Pitch contours of non-native pronunciation, where human assigned labels are Tone3 and Tone4	59
5.7	Pitch contours of non-native pronunciation, where human assigned labels are Tone3 and Tone1	60
5.8	Pitch contours of non-native pronunciation, where human assigned labels are Tone4 and Tone1	61
5.9	Frame-level tone posteriors generated with different tone classifiers.	62
5.10	Pitch contours of native pronunciation, where human assigned labels are Tone3 and Tone3	63
5.11	Tone-dependent detection curves and EERs for BLSTM-based tone mispronunciation verifiers trained with vectors of a sequence of frame-level tone features extracted from different acoustic tone models	64
5.12	Frame-level tone posteriors generated with DNN (hard target) tone classifier.	65
6.1	Overview of the decision tree based phone mispronunciation diagnosis framework.	72
6.2	The KGBDT of phone /CH/ (left) and phone /D/ (right)	75
6.3	The DDBDT of phone /CH/ (left) and phone /D/ (right)	76
A.1	The KGBDT (left) and DDBDT (right) of phone /B/	82
A.2	The KGBDT (left) and DDBDT (right) of phone /C/	82
A.3	The KGBDT (left) and DDBDT (right) of phone /CH/	83

A.4	The KGBDT (left) and DDBDT (right) of phone /D/	83
A.5	The KGBDT (left) and DDBDT (right) of phone /J/	84
A.6	The KGBDT (left) and DDBDT (right) of phone /K/	84
A.7	The KGBDT (left) and DDBDT (right) of phone /P/	84
A.8	The KGBDT (left) and DDBDT (right) of phone /Q/	85
A.9	The KGBDT (left) and DDBDT (right) of phone /R/	85
A.10	The KGBDT (left) and DDBDT (right) of phone /S/	85
A.11	The KGBDT (left) and DDBDT (right) of phone /SH/	86
A.12	The KGBDT (left) and DDBDT (right) of phone /T/	86
A.13	The KGBDT (left) and DDBDT (right) of phone /X/	87
A.14	The KGBDT (left) and DDBDT (right) of phone /Z/	87
A.15	The KGBDT (left) and DDBDT (right) of phone /ZH/	87

SUMMARY

Computer assisted pronunciation training (CAPT) system has been designed to help students improve their speaking skills by providing automatic pronunciation scores and diagnostic feedback. Its mispronunciation detection performance highly depends on the quality of the ASR acoustic model trained with a non-native corpus, and the binary detectors verifying whether the current pronunciation is correctly articulated. Meanwhile, its diagnostic ability is dependent on the choice of the modeled units (e.g., phone, articulation manner, and place), and whether the made decision of selected verifiers/classifiers is interpretable. In this thesis, we show our effort to improve the mispronunciation detection of Mandarin and enrich diagnostic feedback for second language learners. The problem is tackled from the perspective of acoustic modeling, verification and feedback generation of Mandarin phones and tones.

For the acoustic modeling part, speech attributes and soft targets are respectively proposed to help resolve phone and tone's hard-assignment labels, which are not optimal for describing irregular non-native pronunciations. Subsequently, multisource information or better trained acoustic model can provide more accurate features for mispronunciation detectors. Experimental results show that enhanced features can bring consistent improvement for Mandarin phone/tone mispronunciation detection.

For the verification part, segmental pronunciation representation, usually calculated by frame-level averaging in a DNN, is now learned by the memory components in a BLSTM, which directly uses sequential context information to embed a sequence of pronunciation scores into a pronunciation vector to improve the performance of mispronunciation detectors. This improvement is observed both in the phone and tone's mispronunciation detection task.

For the feedback generation part, with the help of phone-, articulatory-, and tone-level posterior scores and interpretable decision trees, we can visualize nonnative mispronuncia-

tions and provide comprehensive feedback, including articulation manner, place, and pitch contour-related diagnostic information, to help L2 learners. Experimental results confirm that our proposed decision trees can provide accurate diagnostic feedback.

CHAPTER 1

INTRODUCTION

1.1 Overview

With accelerating globalization, more and more people with various native language (L1) backgrounds are willing or required to learn second languages (L2). Among them, English is the most popular foreign language, only in China the number of English learners has reached to 270 million in 2012 [1]. Mandarin is also a growing L2 language, and the Chinese Ministry of Education reported that there are roughly 100 million people around the world learning Mandarin [2, 3]. The shortage of qualified L2 language teachers has thus become a serious problem, and computer assisted language learning (CALL) systems can play a key role in alleviating the lack of qualified teachers and offering flexibility in terms of time and space constraints. As an essential component of CALL, the computer assisted pronunciation training (CAPT) subsystem is designed to help students improve their speaking skills by providing automatic pronunciation scores and diagnostic feedback.

Powered by automatic speech recognition (ASR) engines, CAPT not only measures pronunciation quality at a segmental level (e.g., individual phonetic units [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]), but mispronunciations at the supra-segmental level (e.g., stress [15, 16, 17, 18], tone [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30], and intonation [31, 32]) could be also detected. The mispronunciation detection performance highly depends on the quality of the ASR acoustic model trained with a non-native corpus, and the binary detectors verifying whether the current pronunciation is correctly articulated. In addition to accurately detect non-native learners' mispronunciations, a good CAPT system also needs to provide comprehensive corrective feedback for students to improve their pronunciation quality. This ability depends on the choice of the modeled units (e.g., phone, articula-

tion manner, and place), and whether the made decision of selected verifiers/classifiers is interpretable.

The quality of the non-native acoustic model is often heavily dependent on the quality of phone-level labeling of the non-native corpora used for training the phone models for pronunciation scoring. Labeling non-native speech data is intrinsically much more challenging than labeling native speech data. In [33, 34, 35], it was observed that L2 learners' mispronunciations contain many "distortion errors", i.e., the erroneous pronunciation is often between two canonical phones, rather than a straight-forward phonemic substitution. Therefore, standard forced-assigned human labeling of phone categories inevitably generates noisy phone labels for the acoustic model training. As phonetic annotation is a subjective task, even for linguistic experts, annotator subjectivity adds another layer of uncertainty to the ground-truth labels.

Faced with the challenge of inconsistency in non-native phone-based labeling, and imperfect phone modeling, automatic speech attribute transcription (ASAT) [36] is here utilized to extract speech attribute features (e.g., articulation manner, place) to enhance the quality of subsequent mispronunciation detectors. This idea is inspired by the sharing mechanism of speech attribute, i.e., each speech attribute feature is sharable among a group of phones, which allows it to pool more training data than an individual phonetic category. Therefore, speech attribute classifiers are not too sensitive to individual phone-level labeling errors and could be more robustly trained. Compared with a traditional phone based mispronunciation detection system, mispronunciation detectors trained with phone-attribute based features can bring consistent improvement.

Similar to phonemes, the acoustic distribution of non-native Mandarin tone productions often falls between two canonical tone categories or does not resemble any canonical tonal categories [37, 38, 39]. Consequently, forced-assigned standard tone categories in training tone models may also reveal to be suboptimal. Different from phones, where a set of broad phone units (manner and place) [36] can be used to describe how a phone is pro-

nounced, tone productions are mainly characterized by pitch contours, where pitch height and slope are used to differentiate tones' realizations [40, 41, 42]. However, different tones have different pitch heights and slopes so that broad tone units might not exist, and speech attribute-like features used in phone mispronunciation system are not suitable for alleviating the side effects brought by noisy tone labels. Thus, soft targets [43, 44, 45, 46] are here proposed for our tone model design. Unlike [43, 44, 45, 46], which apply soft targets for model compression, we use soft targets to help resolve the hard-assignments of non-native tone labels. Compared with hard targets (one-hot targets) [47], soft targets are the posterior probability of a given event, which are more suitable for describing non-native tone production. With soft targets and Kullback-Leibler (KL) divergence [48] training, we have observed a substantial relative equal error rate (EER) reduction from an already-strong speaker independent deep neural network (DNN) based CAPT system.

Another challenge to face when design CAPT systems is that mispronunciation detectors (verifiers) are trained with averaged frame-level pronunciation scores within each phone or tone's segment [4, 5, 49]. This averaged pronunciation vector ignores the dynamic changes of frame-level scores and might be suboptimal to model humans' perception, where a sequence of frame-level information is fed into a human brain to decide whether the current phone/tone is correctly pronounced. To tackle that issue, we leverage upon the memory components in the long short-term memory (LSTM) network [50] to store context information. Next, we use learned sequential information to embed a sequence of pronunciation scores into a pronunciation vector for the final mispronunciation detection. Experimental results show the effectiveness of the proposed method for phone and tone mispronunciation detection.

Lastly, how to generate instructive feedback still remains challenging. Although traditional pronunciation score or phone-based feedback can help learners to improve their pronunciations, its major assumption is that learners are aware of which articulatory movements (e.g., manner and place of articulation) have to be corrected in order to restore the

pronunciation of the canonical phone. Unfortunately, that is a challenging task for L2 beginners. To tackle this challenge, we take advantage of a large non-native corpus and deep learning techniques to train accurate articulatory classifiers, which allow us to directly measure pronunciation quality and give corrective feedback based on articulation manner and place. Specifically, knowledge-guided and data-driven decision tree based classifiers trained with articulatory-level features are proposed to analyze how inaccurate articulatory movements lead to detected mispronunciations. For example, if the unaspirated retroflex affricate phone /ZH/ is detected to be mispronounced as its dental counterpart /Z/, we can traverse its decision path from current leaf node to the root node to know lower expected retroflex posterior or higher unexpected dental posterior results in current mispronunciation. Consequently, its corrective feedback could be formulated as, “Try to move your tongue tip backwards so that the edges of your tongue are touching your hard palate”.

1.2 Main Contributions

The goal of this thesis is to improve the mispronunciation detection of Mandarin and enrich diagnostic feedback for second language learners. Our contributions are summarized as follows:

For the visualization of non-native pronunciation part, DNN-based speech attribute detectors are proposed to generate articulatory-level posteriors to analyze the characteristics of non-native pronunciations in addition to the conventional phone-related posterior scores. Experimental results show that our large non-native corpus can facilitate the full usage of DNN for much better estimation and detection, especially for each articulatory category, shared by a group of phones. More importantly, non-native pronunciation is often between two canonical phones, rather than belonging to predefined phone category. Therefore, articulatory features (a set of broad phone units) are more suitable for describing how non-native pronunciations are realized by L2 learners.

For the acoustic modeling part, speech attributes and soft targets are respectively pro-

posed to help resolve phone and tone's hard-assignment labels, which are not optimal for describing irregular non-native pronunciations. Subsequently, multisource information or better trained acoustic model can provide more accurate features for mispronunciation detectors. Experimental results show that enhanced features can bring consistent improvement for Mandarin phone/tone mispronunciation detection.

For the verification part, segmental pronunciation representation, usually calculated by frame-level averaging in a DNN, is now learned by the memory components in a BLSTM, which directly uses sequential context information to embed a sequence of pronunciation scores into a pronunciation vector to improve the performance of mispronunciation detectors. This improvement is observed both in the phone and tone's mispronunciation detection task.

For the feedback generation part, with the help of phone-, articulatory-, and tone-level posterior scores and interpretable decision trees, we can visualize non-native mispronunciations and provide comprehensive feedback, including articulation manner, place, and pitch contour-related diagnostic information, to help L2 learners. Experimental results confirm that our proposed decision trees can provide accurate diagnostic feedback.

1.3 Thesis Outline

The rest of this thesis is organized as follows:

In chapter 2, an overview of previous work pertinent to the proposed research is provided. We also give brief introduction of the corpora used in this study.

In chapter 3, multisource information, e.g., phone- and articulatory-level posteriors, are utilized to analyze non-native pronunciations. This data analysis motivates the design of our phones' mispronunciation detectors introduced in chapter 4.

In chapter 4, we show our efforts in improving Mandarin phone mispronunciation detection. Meanwhile, some examples are given to demonstrate the efficiency of proposed speech attribute feature and BLSTM-based verifier.

In chapter 5, we start with data analysis on non-native tone productions, then present our efforts on improving Mandarin tone mispronunciation detection. Some examples are given to show why soft target and longer sequential information is critical in verifying tone's pronunciation correctness.

In chapter 6, we investigate the phone-dependent decision tree's diagnostic ability. Both qualitative and quantitative analyses on the designed decision trees are presented.

In chapter 7, we conclude this thesis and discuss possible future work.

CHAPTER 2

BACKGROUND AND LITERATURE SURVEY

In this chapter, an overview of the mispronunciation types made by L2 learners is given in Section 2.1. Techniques related to mispronunciation detection and corrective feedback generation are introduced in Section 2.2. Finally, commonly adopted performance measurements and the description of native and non-native corpora used in this dissertation are summarized in Section 2.3 and 2.4, respectively.

2.1 Mispronunciation Types

2.1.1 Phonetic Errors

Mispronunciations that occur at a segmental level are most commonly categorized as insertion, deletion, and substitution errors[51, 52]. Insertion and deletion errors refer to a phoneme pronunciation inserted or omitted before or after another phoneme, respectively. These errors might result from L2 learners’ unfamiliarity with foreign language production. For example, consonant clusters are not allowed in Mandarin, so Chinese learners of English are likely to omit some consonants in an English syllable with consonant cluster [53], e.g., word “attempt” /ax t eh m p t/ might be mispronounced as /ax t eh m/ or /ax t eh m p/. Substitution errors occur when the canonical phone pronunciation is replaced with another one. It is well-known that the L2 learning process is heavily affected by a well-established habitual perception of phones and articulatory motions in the learners’ primary language (L1) [54], which often causes the abovementioned replaced mispronunciations. For example, aspiration does not exist in Romance language, such as French, Italian and Spanish, so L2 learners taking a Romance language as their mother language are more likely to mispronounce Mandarin aspirated phones /P/, /T/, and /K/ as their unaspirated counterparts /B/,

/D/, and /G/ [55]. Moreover, some mispronunciations just deviate a little from the canonical sound, rather than the absolute phone category substitution. We call these mispronunciations as distortion errors [33, 34, 35], and they can be corrected by slightly adjusting the current articulation manner and place.

2.1.2 Prosodic Errors

Mispronunciations that occur at the supra-segmental level can be categorized as stress, rhythm, and intonation errors [51, 52]. The information loss caused by supra-segmental mispronunciation would fail to convey correct word meaning, intent, and emotion of the speaker. Lexical stress is a specific emphasis assigned to a syllable where higher pitch, greater loudness, and longer duration can be observed. In some English words, it serves to disambiguate lexical terms by proper placement of primary stress, e.g., “\’insert” vs. “in\’sert”. For tonal languages, e.g., Mandarin and Vietnamese, lexical tones are used to differentiate syllable meanings, e.g., ma1 (mother), ma2 (hemp), ma3 (horse), and ma4 (scold) have the same toneless syllable, namely ma, yet different tone markers. Regarding rhythm, it refers to some temporal pattern of how a sequence of syllables is spoken. For stress-timed languages, e.g., English, stress occurs at the regular intervals, and the duration of stressed syllables are longer than unstressed counterparts. In contrast, for syllable-timed languages, e.g., French, each syllable is produced at a steady rate, which is unaffected by stress differences. Therefore, we can imagine that French L2 learners of English might mis-equalize each syllable’s duration. Finally, intonation describes how the voice rises and falls in speech; it is characterized by the variation in pitch and used to indicate the attitudes and emotions of the speakers and to signal the difference between statements and questions. Obviously, mis-realized pitch will cause miscommunication.

2.2 Mispronunciation Detection and Corrective Feedback Generation

In general, a CAPT system armed with appropriate evaluation and feedback components is able to detect abovementioned pronunciation errors and provide corrective feedback. In this thesis, we focus on detecting and diagnosing Mandarin phones and tones' substitution errors, which have been carefully labeled in our used non-native corpus iCALL [55]. Indeed, some researchers analyzed different non-native corpora and found that the number of substitution error is much larger than the number of insertion and deletion error [56, 57]. Below, we make a literature survey on phone and tone's error detection and feedback generation.

2.2.1 Detecting and Diagnosing Phonetic Errors (phone substitution)

ASR acoustic models have been a key component in CAPT systems because they can be used to calculate acoustic scores for assessing the pronunciation quality. The acoustic model used in ASR-based CAPT systems evolved from the original template-based model [58, 59] to the context-dependent Gaussian mixture hidden Markov model [60, 61, 62, 49]. Now the state-of-the-art mispronunciation detection performance is achieved by introducing DNN-HMM acoustic model [4, 5, 6, 11, 12, 13, 14]. A template-based detection framework was first introduced in [58], where an isolate-word template-based speech recognizer was built for calculating children's pronunciation scores. [59] improved its performance by extracting word-level features from the aligned path for extra verification. Subsequently, deep belief networks (DBN) posteriograms replacing Mel-frequency cepstral coefficients (MFCCs) were explored to achieve a better alignment path [63]. Although template-based solutions have achieved an acceptable accuracy, a superior mispronunciation detection result was delivered with a statistical model-based framework, where various approaches based on hidden Markov model (HMM) likelihood probability, or posterior probability have been proposed. For example, the log-likelihood ratio (LLR) was adopted in [60] as

a confidence score to measure the difference between native-like and non-native acoustic phone models. Subsequently, Witt and Young [61] took the ratio between the likelihood score from forced alignments and the likelihood score from open phone loop decoding into account and proposed the Goodness of Pronunciation (GOP) score. Along with its variations [62, 49], the GOP score has been widely used to detect mispronunciations. More recently, metric-related and discriminative training criteria have been proposed in [7, 64] to enhance the GMM-HMM or DNN-HMM based acoustic model. Consequently, more accurate GOP scores can be calculated and used in the CAPT systems.

While GOP scores can measure L2 learners' pronunciation quality and return a decision as to whether a current phone is mispronounced, an extra verification step could further enhance its performance. Specifically, mispronunciation detection can be formulated as a binary classification/verification task, where segmental phone features (e.g., phone posteriors and GOP scores) along with their corresponding segmental level pronunciation labels (correct/incorrect) are paired and used to train binary verifiers (e.g., ANN [65], SVM [49] and decision tree [66]). Trained classifiers assign posterior probabilities to each non-native production to see how likely the given phone is correctly pronounced. To alleviate the training data shortage problem, a new ANN-based verifier with shared hidden layers is proposed in [5], where the last output layer consists of M logistic regression classifiers, each for classifying one phone. In addition to using likelihood or posterior scores as input features, articulatory information, e.g., the position of the tongue [67] and acoustic landmarks [68], were also employed to train classifiers for binary mispronunciation detection.

Although the abovementioned model-based CAPT systems have attained satisfactory mispronunciation detection results, those systems only focus on binary detection and lack the capability of providing diagnostic feedback. Therefore, researchers in [69, 70] modified the training labels, and classifier approaches were used for discriminating confusion pairs including canonical phone and its most frequent mispronounced counterparts. For example, Sébastien et al. [71] investigated the effectiveness of audiovisual features for dis-

tinguishing confusion pairs from L2-learners of Swedish. A comprehensive comparison for detecting a confusion pair, including using different features (e.g., acoustic phonetic, MFCC, GOP), can be found in [70]. Meanwhile, another CAPT framework called extended recognition network (ERN) was proposed in [72, 73, 74]. The phone recognition network is first expanded by adding common phonetic error patterns. Then this ERN is used to force align learners’ utterances. By contrasting the canonical form with the forced aligned phone sequence, the ERN method can provide some diagnostic information related to phone substitution, i.e., phone /A/ has been substituted with phone /B/. The performance of these approaches is highly dependent on the quality of collected L1-dependent phone error patterns, which are obtained either from consulting with linguistic experts or comparing specific L1-L2 pair’s phonological difference [72]. Phonological rules were also automatically derived by aligning the manual transcriptions of L2 speech with the canonical pronunciations [73, 74]. More recently, L1-dependent mispronunciation tendencies, in terms of articulation manner and place, were summarized and added into ERN to detect “distortion errors”, which deviate a little from the canonical sound, rather than the absolute phone category substitution [34].

While the target specific L1-L2 based CAPT systems can use summarized error patterns to provide corrective feedback, the collection of common phonetic errors is time-consuming, and L1-dependent. Therefore, unsupervised mispronunciation pattern discovery is a research topic of rising interest. Wang et al. [8] used posteriogram as feature to perform unsupervised clustering to discover common error patterns directly from data. Mao et al. [75] analyzed the clusters and found that some could represent phone category substitution error, some clusters were interpreted to “distortion error”. Qian et al. [9] proposed a two-pass recognition framework. In the first pass, an ERN including canonical phone and its anti-phone is used to force align learners’ utterances to search regions where there are possible errors. In the second pass, free phone recognition is carried out in the detected region to reveal the phonetic identities of the detected mispronunciation errors.

2.2.2 Detecting and Diagnosing Prosodic Errors (tone substitution)

Generally speaking, different target languages have different prosodic error categories. For example, lexical stress misplacement can change the part of speech (POS) of some English words, e.g., “\’object” (noun) vs. “ob\’ject” (verb). In this thesis, our target language is Mandarin (tonal language), therefore automatic mispronunciation detection of lexical tone is our major goal. Mandarin tones are mainly characterized by their pitch contours, heights and durations. Thus, early work [21] adopted a template-based detection framework, where discrete pitch value and duration of the tested tone is compared with the canonical tone template to decide whether the current tone is mispronounced. In [22], the Euclidean distance was used to calculate the discrepancy between L2 learners’ and ideal contours in the energy and pitch domains. A superior mispronunciation detection result was achieved with statistical model-based frameworks [23, 24, 25, 26, 27, 28, 29, 30], in which prosody-related features (e.g., pitch, energy [24, 25, 26, 27, 28, 29, 30], or fundamental frequency variation [23]) are first extracted to train tone classifiers, e.g., decision tree [27], HMM [24, 25], GMM [23], and ANN [26]. Subsequently, given a test tone segment, its posterior or goodness of pronunciation (GOP) score is used to measure the quality of tone-level pronunciation. Inspired by recent findings on the role played by cepstral features in tone recognition within a DNN framework [76, 77], along with the understanding that tone pitch realization and perception are influenced by the underlying phone units [78, 79], several speech researchers have proposed the use of DNNs to map combined cepstral and tonal features to tone-related posteriors [28, 29, 30]. Those posteriors are then fed into verifiers to finally assess the pronunciation correctness of the current tone. Recently, a combination of pitch contours and tone-related posteriors was reported to achieve better result [26]. If the underlying phone sequence is assumed to be known, tone-based ERN [19] was proposed to force align non-native speech. Compared with phone-based ERN, the tone-based counterpart does not need to collect frequent error patterns, it just needs to be fully expanded with five Mandarin tone categories. Although a tone-based ERN approach can provide tone substitu-

tion feedback to non-native learners, [30] reported that most tone mispronunciations cannot be categorized to one of standard Mandarin tones. Therefore, [27] proposed a decision tree trained with pitch related features to interpret the types of tone errors that were made.

2.2.3 Corrective Feedback Generation

After mispronounced phone and tone are detected, a CAPT system with diagnostic function should be capable of providing feedback to L2 learners and help them improve their pronunciation. Neri et al. [80] analyzed the percentage of pronunciation errors before and after using CAPT system and found that an ASR-based pronunciation score is an effective feedback in correcting some errors made by non-native Dutch learners. However, when facing lower confidence scores, L2 learners are more likely to feel helpless, because they do not know what is wrong with their pronunciation and how to improve it with only numeric scores. Therefore, multimodal feedback, such as native-like pronunciation and articulatory animation, are proposed to more efficiently help non-native learner correct their mispronunciations. Specifically, synthesized or recorded native-like audio clips [81, 82] are provided to L2 learners for imitating. Meanwhile, articulatory animation including motions of the lips, tongue, and mouth as well as the opening of nasal passage is used to help learners understand the expected articulatory movement [82]. More recently, Engwall [83] utilized acoustic-to-articulatory inversion to estimate L2 learners' real articulation, and then compared it with its canonical counterpart. Different from articulatory-level explanation for phonetic errors, pitch contour and height are frequently used for describing non-native tone production and generating diagnostic feedback [84, 85, 86], e.g., please keep your pitch contour exhibit as a high-level straight line without slope for Mandarin Tone1.

2.3 Performance Measurement of Mispronunciation Detection

Many metrics have been proposed to measure the quality of mispronunciation detection systems. Most of them follow the hierarchical evaluation structure developed in [8] (see

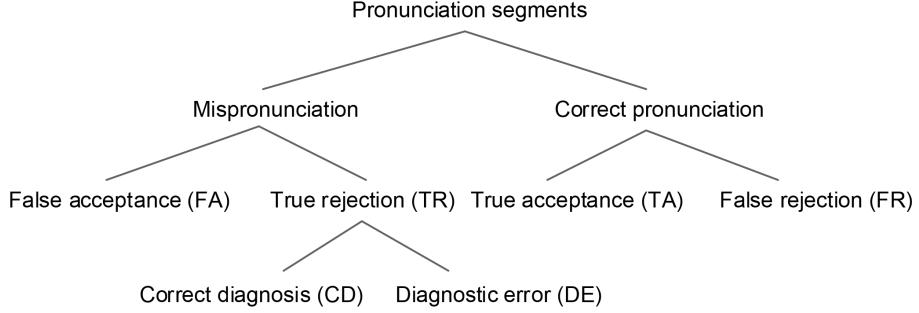


Figure 2.1: Hierarchical structure of the metrics for evaluating the systems' mispronunciation detection and diagnostic performance (adapted from [8, 87]).

Fig. 2.1). We use five commonly adopted metrics:

1. False Acceptance Rate (FAR): the percentage of the number of all the mispronounced segments that are accepted by the system as correct.

$$FAR = \frac{FA}{TR + FA} \quad (2.1)$$

2. False Rejection Rate (FRR): the percentage of the total number of correct pronunciations that are identified by the system as mispronounced.

$$FRR = \frac{FR}{TA + FR} \quad (2.2)$$

3. Precision: the ratio between the total number of detected mispronunciations and the number of true mispronunciations detected.

$$Precision = \frac{TR}{TR + FR} \quad (2.3)$$

4. Recall: the ratio between the total number of mispronunciations labeled by a human expert and the number of true mispronunciations detected.

$$Recall = \frac{TR}{FA + TR} \quad (2.4)$$

5. Diagnostic Error Rate (DER): the ratio between the total number of correctly detected mispronunciations and their number of incorrect diagnostic feedback.

$$DER = \frac{DE}{TR} \quad (2.5)$$

The trade-off curve between FAR and FRR is investigated to find the optimal operation point, where FAR is equal to FRR (e.g., equal error rate). If we treat mispronunciation detection as an information retrieval task, the precision-recall plot is used to evaluate system performance. [88] showed that this metric is more informative than FAR-FRR plot when evaluating binary classifiers on imbalanced datasets, where the number of positive samples is much smaller than its counterpart.

2.4 Speech Corpora

In this thesis, three Mandarin speech corpora are used to train acoustic models and mispronunciation versifiers. In order to keep the same characteristics of reading-style non-native speech, two native read speech corpora, (i) 863 LVCSR corpus [89], and (ii) THCHS-30 corpus [90], have been selected. The first native speech corpus is provided by the Chinese National Hi-Tech Project 863 for Mandarin large vocabulary continuous speech recognition system development. A total of 94,000 utterances spoken by 160 speakers for about 100 hours have been recorded. The second native corpus was freely released by Tsinghua University at 2015, it contains about 34 hours continuous Mandarin speech spoken by 60 speakers.

The non-native Mandarin speech corpus used in this study is iCALL [55] provided by the Institute for Infocomm Research (I2R) at Singapore. It contains 90,841 utterances spoken by 305 non-native learners with a total duration of 142 hours. These learners have different mother languages including Germanic, Romance, and Slavic. Each learner was asked to read 300 Pinyin prompts, ranging from short phrases to sentences. All audio

recordings are manually transcribed (surface pronunciation) by trained annotators, while the original Pinyin prompts were used as canonical pronunciations. By comparing the above surface and canonical transcriptions, we can get mispronunciation types at the segmental and supra-segmental levels. Its detailed analysis on the error patterns including phonetic and tonal errors as well as fluency scores is presented in [55].

2.5 Summary

In this chapter, we first give an overview of mispronunciations occurred at the segmental and supra-segmental level. The goal of this dissertation is to detect and diagnose phone and tone's mispronunciation, therefore a number of previous related work is discussed in the second section. Finally, we provide a description of system performance measurements and Mandarin corpora adopted in this thesis.

CHAPTER 3

VISUALIZING NON-NATIVE PRONUNCIATION THROUGH PHONE AND ITS ATTRIBUTE DETECTION

In this chapter, our goal is to investigate the feasibility of using multisource information, e.g., phone- and articulatory-level posteriors, to visualize and analyze non-native phonetic pronunciations. Section 3.1 first gives an introduction of the techniques used for generating phone and articulatory-level confidence scores. These scores can be used to indicate the activity levels for the speech events of interest. Second, Mandarin phones and articulatory features are introduced in Section 3.2. Next, our implemented deep learning based phone and speech attribute classifiers, and their confidence score extraction are described in Section 3.3 and 3.4, respectively. Finally, experimental setup and results are reported in Section 3.5, and some non-native pronunciations with visible speech cues are displayed in Section 3.6.

3.1 Introduction

The automatic speech attribute transcription (ASAT) project [36] was an attempt to mimic some human speech recognition capabilities with asynchronous speech event detection followed by bottom-up knowledge integration and verification. This framework has been successfully applied to a variety of existing applications in speech processing and information extraction [91, 92, 93, 94, 95, 96, 97]. As its front-end speech processing module, a bank of speech attribute detectors was designed to transcribe an input speech signal into a time series that describes the level of presence of all the speech attributes of interest, in the input speech utterance over time.

In previous studies [91, 92, 93, 94, 95, 96, 97], speech attributes were defined as phonetic features, e.g., articulation manner and place, which have been proven helpful in im-

proving robustness toward noise [98], speaking styles, and speaker population [99]. Many existing pattern classification techniques have been proposed to build speech attribute detectors, e.g., CRFs [91], HMMs [92], DNNs [93], and TDNNs [94]. The posterior or likelihood generated by attribute detectors can be used as an input feature for speech visualization [36], lattice rescoring [95], language recognition [96], and bottom-up detection-based ASR [97].

In CAPT research field, speech attribute information can be used to visualize non-native pronunciations and provide articulatory-level corrective feedback. For example, when the word “matter” is spoken by Chinese learners of English, the original low-front vowel /æ/ is likely to be mispronounced as its low-back counterpart [100]. We can use speech attribute detectors to visualize/estimate learners’ tongue position, and its articulatory-level feedback could be formulated as, “keep your current height of your tongue, but move it forward to create the acoustic characteristics of /æ/”. Indeed, it has been reported that L2 learners prefer to receive direct instruction on how to correct mispronunciation at a articulatory-level [101]. Attracted by the abovementioned preference, researchers have exploited articulatory information for L2 learning [83, 101, 102], where an acoustic-to-articulatory inversion method was adopted to directly provide feedback at an articulatory level, e.g., tongue position. In [67, 103, 104], rule-based acoustic-articulatory mapping tables were employed to overcome the difficulty of collecting physical articulatory measurements to map each phone to its corresponding articulators. However, past work in mispronunciation detection performance at the articulatory level have been suboptimal due to the use of shallow models, or the lack of large training non-native corpora [67].

In this thesis, a large non-native Mandarin corpus iCALL [55] is used to facilitate the full usage of deep neural networks (DNNs) for better modeling and estimation [36, 93]. Specifically, we adopt rule-based acoustic-articulatory mapping method and DNNs are trained to map frame-level acoustic features to phone- and speech attribute-related posteriors. These frame-based DNN posteriors are used as scores to measure pronunciation

quality, e.g., correctness of manner and place of current articulation. In addition to being used as non-native feedback generation, speech cues extracted from trained DNNs can also provide labeling candidates to annotators for achieving more consistent labeling results and reducing labeling time [105].

3.2 Mandarin Phones and Speech Attributes

We focus on learners of Mandarin Chinese with European first languages. Each Chinese character corresponds to one spoken syllable, consisting of an initial, usually a consonant, and a final, usually a vowel or sometimes a vowel followed by a nasal consonant. There are a total of 21 syllable initials and 38 syllable finals. In this study, we are concerned with the mispronunciation of 21 syllable initials, because initial errors are more prone to cause miscommunication in Mandarin when compared to the errors of syllable finals [106]. Moreover, after analyzing a large-scale, non-native Mandarin corpus, Chen [55] found that 90% of the top 10 mispronounced phones are syllable initials.

Each initial’s articulatory characteristic can be described using its corresponding speech attribute features [107, 108]. For example, when people pronounce the initial /B/, the airflow from the lungs is blocked by the place of articulation “labial”, causing a pressure difference to build up. Once the closure is opened, the released airflow produces a sudden impulse causing an audible sound or burst [107, 108]. This whole process is called the articulation manner “stop”. As articulation place and manner, the speech attribute features “labial” and “stop” are used to describe how /B/ is produced. In addition to manner and place of articulation, we also consider voicing and aspiration. When a phone is pronounced, voicing is used to describe if the vocal cord vibrates; whereas aspiration is used to describe whether there is a brief puff of air after an obstruction is released. Table 3.1 lists the mapping rules between speech attribute and Mandarin initials (consonants) denoted in Pinyin [109, 110].

Table 3.1: Speech attributes and their associated Pinyin initials (consonants)

category	Attribute	Phone set
Place	Labial	B,P,M,F
	Alveolar	D,L,N,T,C,S,Z,
	Retroflex	ZH,CH,SH,R
	Palatal	J,Q,X
	Velar	G,H,K,NG
	N/A	VOWELS
Manner	Stop	B,P,D,T,G,K
	Fricative	F,S,SH,X,H
	Affricative	Z,ZH,C,CH,J,Q
	Nasal	M,N,NG
	Liquid	L, R
	N/A	VOWELS
Aspiration	Aspirated	P,T,K,C,CH,Q
	Unaspirated	B,D,G,Z,ZH,J
	Others	F,H,L,M,N,R,S,SH,X,NG,
	N/A	VOWELS
Voicing	Voiced	M,N,L,R,NG,
		VOWELS
	Unvoiced	B,P,M,F,D,T,N,L,G,K,H,J,Q,X
		ZH,CH,SH,R,Z,C,S
Silence	Silence	SIL

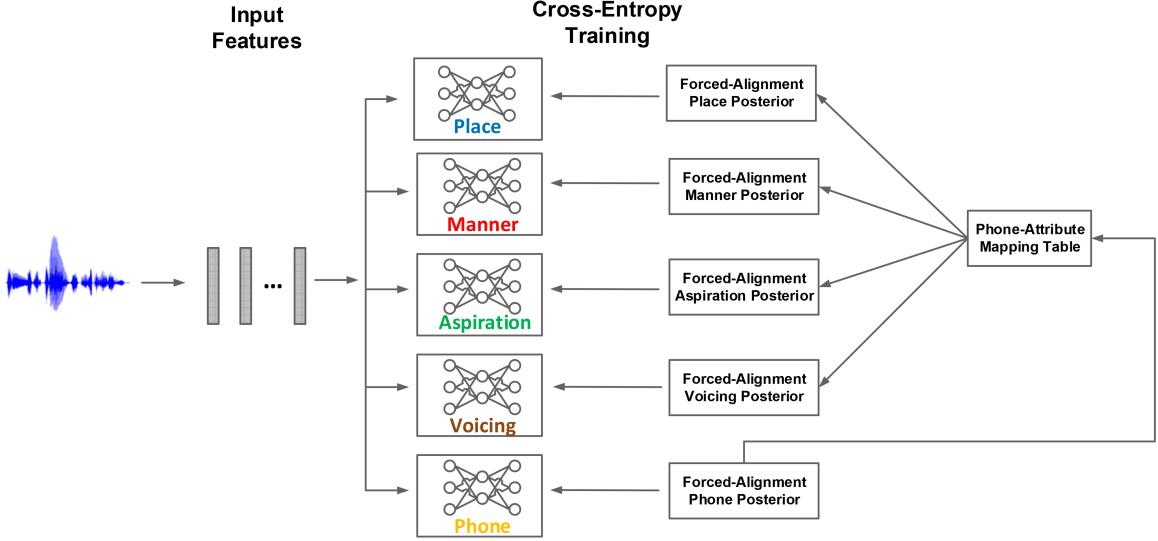


Figure 3.1: Overview of the phone and speech attribute classifiers training.

3.3 Speech Attribute and Phone Classifiers Training

After compiling the phone-attribute mapping table and obtaining the phone-level forced-alignment, a bank of speech attribute classifiers is trained to map acoustic features (e.g., FBANK) to corresponding articulatory-level forced-alignment as shown in Fig. 3.1. Specifically, a DNN-based classifier is separately trained for each articulatory-motivated attribute category described in Table 3.1. In addition, a DNN-based phone classifier is also trained with the senone labels derived from the GMM-HMM acoustic model. In DNNs, Eq. 3.1 is used to iteratively map the lower-level features into higher-level hidden features, which are fed into a softmax layer to predict the probability distribution over the predefined classes

$$a^l = \sigma(W^l a^{l-1} + b^l), \quad (3.1)$$

$$\hat{y} = \text{softmax}(W a^* + b), \quad (3.2)$$

where W^l and b^l are the weight matrix and bias vector of the l th hidden layer, a^l is its output after the activation function σ (e.g., sigmoid or relu), W and b are the weight matrix

and the bias vector of the nonlinear softmax layer, respectively, a^* is the last hidden layer's output, and a^0 is equal to the input feature vector.

The above parameters are learned by minimizing the cross-entropy function of the predicted and true distributions, as shown in Eq. 3.3

$$L(y, \hat{y}) = - \sum_{i=1}^N \sum_{j=1}^C y_i^j \log(\hat{y}_i^j), \quad (3.3)$$

where y_i^j and \hat{y}_i^j are the ground-truth and predicted probability distributions, respectively, N is the number of the training samples, and C denotes the class number, e.g., C is equal to 7 for the manner classification (stop, fricative, affricative, nasal, liquid, silence, vowels).

After the speech attribute and phone classifiers are well trained, we can use them to extract multisource information for visualizing and analyzing non-native phonetic pronunciations.

3.4 Speech Attribute and Phone Feature Extraction

A window of 11 speech frames centered on the current frame is fed into each DNN classifier, which in turn generates a set of confidence scores in terms of posterior probabilities that the current frame pertains to each speech attribute within the target category. We name these posteriors as frame-level speech attribute features. Similarly, a DNN phone classifier analyzes an expanded frame of the input speech signal and produces the posterior probability that pertains to each tied HMM state, often referred to as a senone. Subsequently, [5, 29] proposed to use Eq. 3.4 to calculate the posteriors of the current frame belonging to each phone category

$$P(P|o_t) = \sum_{s \in p} P(s|o_t), \quad (3.4)$$

where unit P is the target phone category, o_t is the input feature at frame t , and s is the senone label used to compute phone state posterior; $\{s \in p\}$ is the set of all senones corresponding to unit P . Finally, given the phone boundary t_s, t_e extracted from forced-

alignment, frame-level phone or speech attribute posteriors within each phone segment are averaged to produce log pronunciation posterior (LPP) for the target unit P as in Eq. 3.5.

$$LPP(P) = \log P(P|o; t_s; t_e) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P(P|o_t). \quad (3.5)$$

3.5 Experiments

Non-native Mandarin phone and speech attribute classification experiments are carried out on the largest non-native Mandarin corpus iCALL [55]. Specifically, two native speech corpora introduced in chapter 2 and a subset of iCALL with 100 speakers' data are mixed together to train our speech attribute and phone classifiers. The remaining part of iCALL with 195 speakers' data is used to evaluate our system performance.

3.5.1 Experimental Setup

The open source Kaldi toolkit [111] is used to train our phone classifier: a context-dependent Gaussian mixture hidden Markov (CD-GMM-HMM) acoustic model is initially trained using the maximum likelihood (ML) estimation. Then CD-DNN-HMM model is built using the state level phone alignments provided by the trained CD-GMM-HMM and human labeled transcriptions. The DNN acoustic model has six hidden layers, each having 2048 Sigmoid units. Its input is an augmented 11-frame vector, including 5 preceding, the current and 5 succeeding frames. Each frame is 40-dimension mean-normalized log-filter bank feature with up to second-order derivatives.

Regarding to speech attribute classifiers shown in Fig. 3.1, the same data used for the abovementioned phone classifier training is utilized to train each articulatory-motivated category described in Table 3.1. These classifiers have the same DNN architecture and input feature as its phone counterpart, except the dimension of output. For example, the output dimension of the classifier for articulation manner is equal to 7 (stop, fricative, affricative, nasal, liquid,silence, vowels). The phone-attribute mapping rules shown in Table

Table 3.2: Classification accuracy for each phone category in the test set, where the overall performance is 92.1%

Phone	Accuracy(%)	Phone	Accuracy(%)	Phone	Accuracy(%)
B	95.0	J	88.1	R	93.5
C	76.4	K	92.5	S	88.9
CH	86.6	L	96.6	SH	95.0
D	94.4	M	96.2	T	85.6
F	97.5	N	90.6	X	89.8
G	95.3	P	88.4	Z	88.1
H	98.2	Q	84.1	ZH	89.1

3.1 are used to convert phone-level alignment into speech attribute labels within different categories for training different classifiers, e.g., place, manner, aspiration, and voicing.

In this experiment, classification accuracy is used to evaluate the performance of phone- and articulatory-level classifiers:

$$accuracy(y, \hat{y}) = \frac{1}{N_{segments}} \sum_{i=1}^{N_{segments}} \mathbb{I}(y_i = \hat{y}_i) \quad (3.6)$$

where variable y_i is the i -th segment label assigned by human annotators; and the i -th predicted label is $\hat{y}_i = argmax_{p \in Q} LPP(p)$, where $LPP(p)$ is log pronunciation posterior of unit p and calculated by Eq. 3.5, Q is the predefined phone set or each articulation category shown in Table 3.1. $N_{segments}$ is the total number of the testing segments. Originally, the iCALL corpus only provided phone-level labels, we utilized phone-attribute mapping rules shown in Table 3.1 to convert each phone-level label to its articulatory counterpart for calculating the accuracy of articulatory-level classifiers.

3.5.2 Experimental Results and Discussions

Table 3.2 summarizes the classification accuracy (ACC) for each Mandarin phone (consonant). In this study, we are concerned with non-native Mandarin consonant classification and visualization because Chen [55] analyzed iCALL and found that 90% of the top 10 mispronounced phones are consonants.

Table 3.3: Classification accuracy for each manner attribute in the test set, where the overall performance is 97.0%

Manner	Accuracy(%)
Stop	96.5
Fricative	97.6
Affricative	97.4
Nasal	98.4
Liquid	94.5

Table 3.4: Classification accuracy for each place attribute in the test set, where the overall performance is 94.9%

Place	Accuracy(%)
Labial	94.9
Alveolar	95.7
Retroflex	93.4
Palatal	92.4
Velar	98.2

From this table we observe that the combination of DNN and large training corpora can bring good phone classification accuracy, namely its overall ACC is 92.1%. However, we can also observe that there exist significant differences among the performance of different phones. For example, the performance of phone /H/ is the highest 98.2% ACC, and the lowest counterpart is 76.4% for phone /C/. In this study, pronunciation quality might result in this observed difference. Namely, more frequently mispronounced phone category is more likely to have ambiguous pronunciations, which are difficult to be distinguished. Therefore, its comparatively low ACC is reasonable. We select five phones with the lowest ACC from Table 3.2, and compare them with top 5 mispronounced phones in iCALL [55]. The comparison result shows that these two groups of phones are the same. Moreover, phone /C/ was reported to have the largest mispronunciation rate in iCALL [55].

Tables 3.3-3.6 summarize the classification accuracy (ACC) for each speech attribute shown in Table 3.1. We can observe that the overall ACC of the different articulatory-level classifiers is promising, e.g., manner(97.0%), place (94.9%), aspiration (96.0%), and voicing (99.4%). Similar to previous work [36, 96], the manner of articulation is often easier

Table 3.5: Classification accuracy for each aspiration attribute in the test set, where the overall performance is 96.0%

Aspiration	Accuracy(%)
Aspirated	91.1
Unaspirated	95.0
Others	98.8

Table 3.6: Classification accuracy for each voicing attribute in the test set, where the overall performance is 99.4%

Voicing	Accuracy(%)
Voiced	98.8
Unvoiced	99.5

to be detected than the place of articulation, and voicing discrimination is the easiest task. After comparing the overall ACC in Tables 3.2-3.6, we find that articulatory-level classifiers achieve higher ACC than their phone counterpart. This observation might be caused by following reasons. First, the difficulty of classification task is partially dependent on the number of class. Obviously, discriminating Mandarin consonants is more challenging than manner classification, where only 7 speech attributes are needed to be classified. In addition, the observed superior ACC might be attributed to the sharing mechanism of the speech attribute, i.e., each speech attribute feature is shared by a group of phones, which allows it to pool more training data than an individual phonetic category, so that the speech attribute classifiers are not as sensitive to phone label inconsistency as discussed in our Introduction chapter, and could be more robustly trained.

To visualize the degree of phone label inconsistency, we compare the histogram of phone-dependent segmental posterior of native and non-native corpus in Fig. 3.2, where the phone label given by annotators is Mandarin consonant /J/, one of the top mispronounced Mandarin phones produced by L2 learners [55, 112]. Native speakers have standard pronunciations so that trained phone classifier can better map acoustic feature into expected labels in the training set, e.g., higher posteriors close to 1 are observed on the upper panel of Fig. 3.2. However, for non-native speakers, we see the posterior scores are more evenly

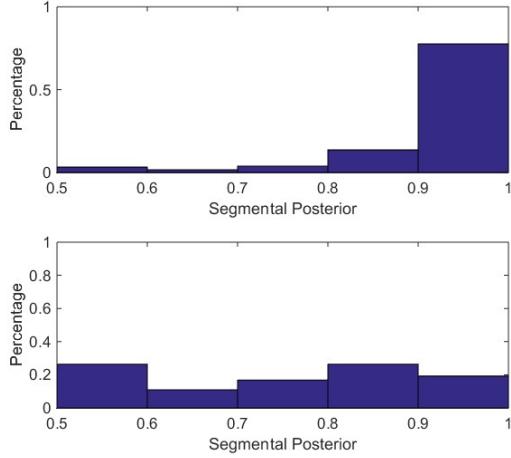


Figure 3.2: Segmental posterior histograms of Mandarin consonant /J/ computed on native (upper) and non-native data (lower).

spread out, i.e., the subfigure on the lower panel of Fig. 3.2. This shows that the aforementioned phone classifier achieving good mapping performance on the native task does not learn well between the non-native productions and force-assigned labels. Obviously, force-assigned labels are not optimal for labeling non-native productions with high variability and uncertainty. The above-mentioned segmental posterior is achieved by averaging frame-level phone posteriors, no logarithm is used.

3.6 Visible Non-native Pronunciation Analysis

The satisfied classification accuracy shown in Tables 3.2-3.6 opens another door to visualize and analyze non-native pronunciations in addition to the conventional spectrogram reading. Specifically, the speech cues extracted from trained phone- and articulatory-level classifiers can be used to describe/analyze how non-native pronunciations are articulated. After understanding the characteristic of non-native speech, researchers can design more powerful mispronunciation detectors and feedback generators for diagnosing non-native mispronunciations. Moreover, extracted speech cues can also offer speech insights to annotators who are usually not expert trained in spectrogram reading. Subsequently, they can label more non-native data with better consistency within a fixed labeling time. Fig.

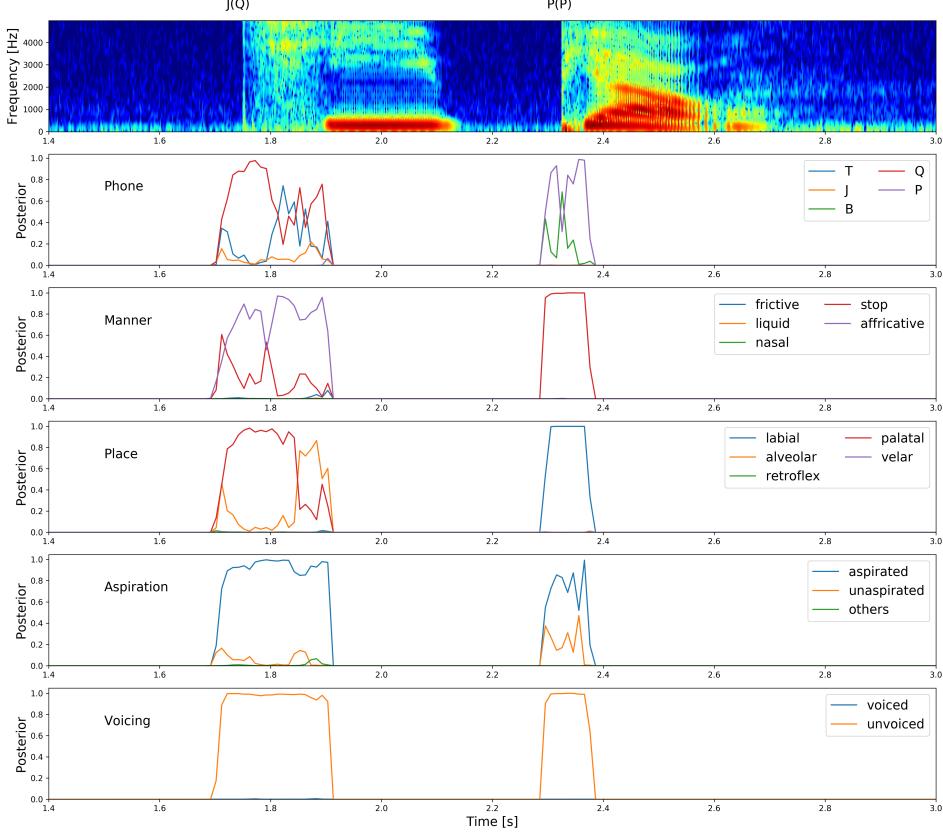


Figure 3.3: Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /J/, /P/, and the human assigned labels are /Q/, /P/.

3.3 is an example of using multiple types of information to visualize/analyze non-native pronunciations. In the upper panel, the spectrogram of two syllables and their corresponding consonants are shown, where the consonants outside the parentheses denote canonical pronunciations, and their counterparts represent surface ones. A subset of phone detection curves in the second panel show that phone /J/ is mispronounced as /Q/, and the second phone /P/ is correctly pronounced. This detection result is consistent with human labeling in the upper panel. Meanwhile, the articulatory-level detection curves shown in the panels 3-6 also make a correct detection. Namely, the unvoiced palatal affricate unaspirated phone /J/ is detected to be mispronounced as its aspirated counterpart /Q/. Finally, the speech attributes (e.g., stop, labial, aspirated, and unvoiced) related to phone /P/ all have dominant frame-level posteriors within the range of the second consonant.

Although phone detection curves alone can diagnose abovementioned mispronunciations, and provide phone-level substitution feedback, e.g., “pay attention, you mispronounced phone /J/ as /Q/”, they might not discover the reason underlying current mispronunciations. However when it’s analyzed at the articulatory-level, we can find that unvoiced phone /J/ and /Q/ share the same articulation manner, and place, namely affricate-palatal, and differ only in that /Q/ is aspirated and /J/ is not. Obviously, diagnostic related to “aspiration error” is closer to the nature of the detected substitution error. Once L2 learners grasp how to reduce brief puff of air after an obstruction in mouth is released, they are more likely to avoid following mispronunciations: /B/ → /P/, /D/ → /T/, /G/ → /K/, /Z/ → /C/, and /ZH/ → /CH/. For each pair, the phone before and after the arrow share the same articulation manner, and place, differing only in aspiration. In addition to providing a more systematic diagnostic, speech attribute classifiers achieve higher classification accuracy, thus their frame-level speech cues are more reliable than the counterpart generated from conventional phone classifier.

Finally, compared with phone features, speech attribute features are more powerful to describe non-native pronunciations. Take phone /B/ for example, its acoustic realization in English is voiced. Therefore, affected by their well-established articulatory motions, learners from English-speaking countries are more likely to mispronounce unvoiced Mandarin /B/ as its voiced counterpart, which is not in Mandarin phonetic inventory. One utterance spoken by American learner is shown in Fig. 3.4. Its spectrogram and corresponding canonical and surface consonants are shown in the upper panel. The phone-related detection curves shown in the second panel indicate that the first phone /B/ is good pronounced, and its corresponding speech attribute features “stop”, “labial”, and “unaspirated” all signify it is correctly pronounced, except voiced/unvoiced detection curves shown in the last panel. Obviously, phone and speech attribute features can tell a more complete story than phone features alone.

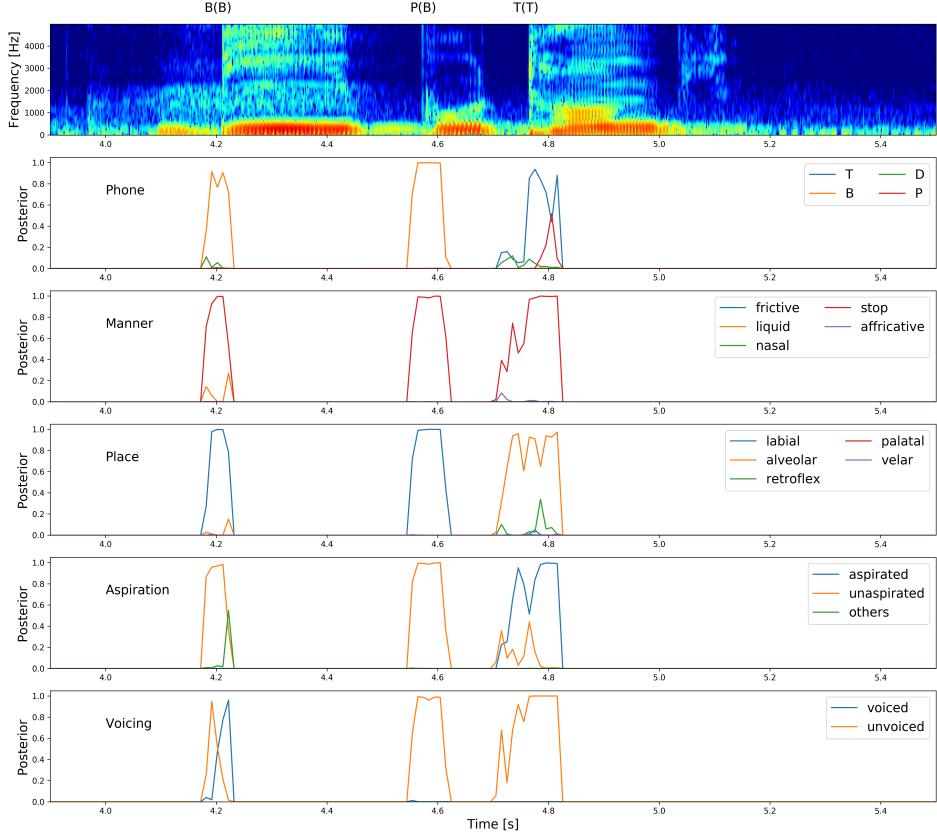


Figure 3.4: Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /B/, /P/, /T/, and the human assigned labels are /B/, /B/, /T/.

3.7 Summary

In this chapter, we first give an introduction of the automatic speech attribute transcription project and its applications in CAPT. Then we introduce the Mandarin phones and speech attributes. Their corresponding classifier training and feature extraction are described in the third and fourth section, respectively. Benefit from using large non-native corpus and deep learning techniques, the high classification accuracy shown in Tables 3.2-3.6 is achieved, it makes using phone- and articulatory-level features to visualize and analyze non-native pronunciations become possible and reliable. Finally, some visualized non-native pronunciations are displayed, the speech cues extracted from well-trained classifiers can be used to diagnose non-native mispronunciations and provide corrective feedback to L2 learners.

CHAPTER 4

IMPROVING MISPRONUNCIATION DETECTION OF MANDARIN PHONES WITH MULTISOURCE INFORMATION AND BLSTM-BASED MISPRONUNCIATION DETECTORS

In this chapter, our goal is to use multisource information and memory-based verifiers to improve the detection performance of phonetic mispronunciations produced by second language learners. First, speech attribute scores extracted from well-trained articulatory-level classifiers are concatenated with conventional phone scores as enhanced features to improve a baseline system based only on phone information. Next, pronunciation representation, usually calculated by frame-level averaging in a DNN, is now learned by BLSTM, which directly uses sequential context information to embed a sequence of pronunciation scores into a pronunciation vector to improve the performance of subsequent mispronunciation detectors. Finally, the combination of these two techniques achieves a significant improvement over the traditional DNN-based verifier trained with only phone-related features.

4.1 Introduction

A wealth of research has utilized confidence scores [60, 61, 62, 49, 5, 113] derived from automatic speech recognition (ASR) systems to provide pronunciation scores to the L2 learners. For example, the log-likelihood ratio (LLR) was adopted in [60] as a confidence score to measure the difference between native and non-native acoustic phone models. A variation of the posterior probability ratio, a “Goodness of Pronunciation (GOP)” [61] score, was also proposed to evaluate the L2 learners’ pronunciation quality. Along with its variations [62, 49], GOP score has been widely used to detect mispronunciations as well. After formulating mispronunciation detection as a binary classification task, GOP scores

between a canonical phone and other competing phones are combined into a feature vector, which is then fed into phone-dependent classifiers (e.g., [5, 49, 65, 66]). The posterior probabilities obtained from those classifiers are often used as pronunciation scores. However, when facing lower confidence scores, L2 learners are more likely to feel helpless, because they do not know what is wrong with their pronunciation and how to improve it when only given numeric scores. In [113], it was shown that L2 learners could improve their production of the targeted phones by receiving a corrective feedback about the mispronunciation error at the phone level. More recent research work has thus focused on how to use automatic mechanisms to generate finer detection results and corrective information, such as in the phone-based extended recognition network (ERN) [72, 73, 74] approach. To reduce the amount of resources needed for collecting frequent L1- dependent error patterns, a recent work [6] proposed an acoustic phonological model to automatically learn the acoustic-phonetic rules from canonical productions of words and annotated mispronunciation. Consequently, a multi-distributed DNN [6] leveraging learned phonological rules was then used to accomplish phone recognition and provide phone level corrective feedback.

Although the abovementioned phone-based CAPT systems have achieved satisfactory mispronunciation detection results, the performance is often heavily dependent on the quality of phone-level labeling of the non-native corpora used for training the phone models for pronunciation scoring. Labeling non-native speech data is intrinsically much more challenging than labeling native speech data. In [33, 34, 35] it was observed that L2 learners' mispronunciations contain many “distortion errors”, i.e., the erroneous pronunciation is often between two canonical phones, rather than a straight-forward phonemic substitution. Therefore, standard forced-assigned human labeling of phone categories inevitably generates noisy phone labels during acoustic model training. In addition, as phonetic annotation is a subjective task, even for linguistic experts, annotator subjectivity adds another layer of complexity to the ground-truth labels.

Faced with the challenges of inconsistency in non-native phone-based labeling and im-

perfect acoustic modeling, we propose to leverage upon automatic speech attribute transcription (ASAT) [36] and extract speech attribute information (e.g., articulation manner, place) to enhance the quality of the input features fed into subsequent mispronunciation detectors. This idea is inspired by the sharing mechanism of speech attribute, namely each speech attribute is shared by a group of phones, results in each speech attribute leveraging more training data from a group of phones, so that the side effect of labeling noise at the individual phone level is mitigated. In the previous chapter, with the help of large non-native corpus and deep learning techniques, we have achieved a good speech attribute classification accuracy. In fact, Speech attribute features have been used as complementary features to reduce word error rate in ASR, e.g., [98, 99], where articulatory motivated features are shown to help improve robustness toward noise, speaking styles and speaker population. In this chapter, we combine speech attribute and traditional phone features to improve the performance of phone-level mispronunciation detection.

In addition, we can observe that the frame-level detection curves shown in Fig. 3.3 exhibit dynamic changes within each phone segment. Traditional frame-level averaging for calculating input vector fed into DNN-based verifier will inevitably lose observed sequential information, which might be helpful for non-native mispronunciation detection. Therefore, BLSTM-based classifier is proposed here to make full use of sequential context information to embed a sequence of frame-level pronunciation scores into a vector for verifying whether the current phone is mispronounced or not. Compared with the pronunciation representation derived from traditional frame-level averaging [4, 5], the BLSTM embedded pronunciation vector is expected to contain much more context information and achieve better mispronunciation detection performance.

4.2 Overview of The Phone Mispronunciation Detection Framework

Fig. 4.1 shows the proposed mispronunciation detection framework, which consists of two blocks: (i) the speech attribute and phone feature extraction module and (ii) the phone-

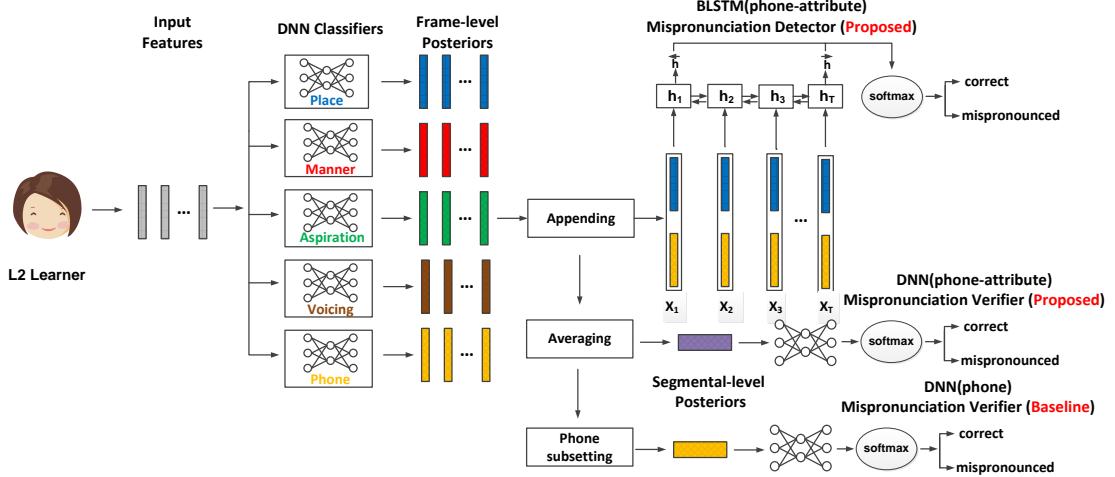


Figure 4.1: Overview of the phone mispronunciation detection framework.

dependent mispronunciation detector training modules, including DNN (phone), DNN (phone-attribute), and BLSTM (phone-attribute) mispronunciation verifiers.

4.2.1 Speech Attribute and Phone Feature Extraction

Speech Attribute and phone classifiers training have been introduced in Section 3.3. After the training of speech attribute and phone classifiers is completed, we can use those classifiers to extract multisource information for subsequent mispronunciation detection. Please consult Section 3.4 for detailed description of extracting frame-level and segmental-level features.

4.2.2 DNN-based Mispronunciation Verifier Construction

Similarly to [5, 49], the input features for supervised training consist of two parts, the first part is the log pronunciation posterior of each phone and speech attribute. The second part is the log posterior ratio between the canonical unit and each element in its corresponding full set Q . Therefore, the segmental feature vector of a canonical phone, P_i , is defined as

follows:

$$\begin{aligned} & [LPP(P_1), LPP(P_2) \dots LPP(P_M), \\ & LPP(P_1) - LPP(P_i), LPP(P_2) - LPP(P_i) \dots LPP(P_M) - LPP(P_i)], \end{aligned} \quad (4.1)$$

where M is the size of the predefined phone set or the number of speech attributes in each articulatory-motivated attribute category described in Table 3.1, e.g., $M=7$ for articulation manner category (vowel and silence are included). Subsequently, segmental features and binary labels (correct or mispronounced) are used to train DNN-based mispronunciation detectors.

4.2.3 BLSTM-based Mispronunciation Verifier Construction

Different from the DNN-based mispronunciation verifier, a sequence of frame-level posteriors is no longer required to be averaged when a recurrent architecture is used. The posterior sequence can in fact be directly fed into the BLSTM-based mispronunciation verifier, which will directly generate the verification score. Like DNN segmental feature, frame-level feature is also extended to contain posteriors and posterior ratios. Yet, a many-to-one BLSTM-based recurrent architecture is adopted to avoid the frame averaging step. Similar to sentiment analysis and text classification tasks [114, 115] in natural language processing, we force the BLSTM to map an input sequence into a fixed-size feature vector (the output of the hidden layer at the last time step), which is then passed to a softmax layer for binary classification. The LSTM transition equations are defined as follows:

$$i_t = \text{sigmoid}(W_i x_t + U_i h_{t-1} + b_i), \quad (4.2)$$

$$f_t = \text{sigmoid}(W_f x_t + U_f h_{t-1} + b_f), \quad (4.3)$$

$$o_t = \text{sigmoid}(W_o x_t + U_o h_{t-1} + b_o), \quad (4.4)$$

$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g), \quad (4.5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \quad (4.6)$$

$$h_t = o_t \odot \tanh(c_t), \quad (4.7)$$

where the operation \odot denotes the element-wise vector product. At each time step t , x_t is the current input feature vector, i_t , f_t , o_t and g_t are the gate functions defined in [50], c_t is the memory cell and h_t is the hidden layer representation. W_* and b_* denote the weight matrix and the bias vector of the corresponding gate functions in LSTM.

Given a calculated frame-level pronunciation posterior sequence $x = \{x_1, x_2 \dots x_T\}$, we adopt Eqs. 4.2 - 4.7 iteratively from $t = 1 \dots T$ to calculate the hidden layer at the last time step \vec{h} . Similarly, the backward hidden sequence's final output \overleftarrow{h} is calculated from $t = T \dots 1$ and combined with \vec{h} as a supervector, which is finally fed into Eq. 4.8 to predict the probability distribution over pre-defined classes:

$$\hat{y} = \text{softmax}(W[\vec{h}, \overleftarrow{h}] + b), \quad (4.8)$$

where W and b are the weight matrix and bias vector of the nonlinear softmax layer. The BLSTM's parameters are trained to minimize the cross-entropy function shown in Eq. 3.3. Finally, a threshold is needed for DNN-based or BLSTM-based verifiers to find optimal operation point, where precision is equal to recall.

4.3 Experiments

Non-native Mandarin phone mispronunciation detection is carried out on the largest non-native Mandarin corpus iCALL [55], where 305 European learners of Mandarin Chinese had been asked to read 300 Pinyin prompts, providing canonical transcriptions. The surface transcriptions of iCALL are labeled by three native experts. Upon comparing the above-mentioned surface and canonical transcriptions, we get each phone's pronunciation label,

e.g., correct or incorrect.

295 speakers' data in iCALL corpus is selected and split into three portions. First, a subset of iCALL with 100 speakers' data is mixed with two native speech corpora introduced in chapter 2 to train our speech attribute and phone classifiers. Next, another subset with 165 speakers' data is used to train DNN or BLSTM-based binary classifier. Lastly, the remaining part of iCALL with 30 speakers' data is used to evaluate our system performance. There is no speaker overlap in those three portions. Moreover, our L2 test set is made up of 5 different L1s, including English, French, Spanish, Italian and Russian. Such L1 diversity makes mispronunciation detection more challenging, because the error types made by different L2 learners are influenced by their L1s.

In this study, we are concerned with detecting mispronunciation of 21 consonants in Mandarin, because consonant errors are more prone to cause miscommunication in Mandarin when compared to vowel(s) or vowel(s) followed by nasals [106]. Moreover, after analyzing a large-scale non-native Mandarin corpus, Chen [55] found that 90% of top 10 mispronounced phones are consonants. Table 4.1 lists some statistics of 21 Mandarin consonants in our train and test sets. We can observe that the mispronunciation rate in our test set is much higher than the train set, with 3.3 samples mispronounced per 100 non-native productions. The higher mispronunciation rate in our test set is caused by data selection and refined phone label. Namely, we select speakers with higher mispronunciation rate as our research targets so that the test set can cover most mispronunciations in real life. Obviously, the result reported on this test set should be more convincing. Meanwhile, some original ignored mispronunciation samples are recalled by linguistic experts in Beijing Language and Culture University.

4.3.1 Experimental Setup

With regard to the phone and speech attribute classifiers shown in Fig. 4.1, their DNNs configuration and training procedure are the same as in the phone and speech attribute

Table 4.1: The number of each phone’s correct and mispronounced samples in our experiments, where the overall mispronunciation rates in the train and test sets are 3.3% and 19.2%, respectively

Phone	Train Set		Test Set	
	# Correct	# Mispronounced	# Correct	# Mispronounced
B	10,272	132	193	20
C	2,149	378	28	33
CH	5,652	431	105	48
D	16,486	281	250	24
F	5,492	18	163	3
G	9,921	107	185	6
H	9,912	21	210	1
J	12,779	540	281	45
K	4,335	93	93	25
L	12,033	25	220	0
M	8,509	7	168	1
N	4,849	9	74	2
P	2,821	168	42	15
Q	6,402	632	121	61
R	4,461	44	84	13
S	2,653	196	65	16
SH	13,873	487	245	57
T	7,662	514	96	63
X	11,044	464	210	47
Z	6,750	441	78	51
ZH	10,044	708	225	74

classification experiment described in Section 3.5.1. The open source Kaldi toolkit is used to train our phone and speech attribute classifiers.

The DNN-based mispronunciation verifiers are trained using the segmental feature vectors shown in Eq. 4.1 along with the corresponding phone level pronunciation labels (correct/incorrect). Different configurations, e.g., number of hidden layers (1, 2), or number of hidden nodes (128, 256, 512), have been evaluated on the development set (one-tenth of total data for verifiers’ training). The final selected DNN has two hidden layers each with 256 hidden nodes. In iCALL, the phonetic error rate is around 5%, namely the number of correct samples is much higher than that of the incorrect samples. This data imbalance will make trained mispronunciation detector have a high precision rate but a low recall rate.

Therefore, a cost-sensitive objective function is used to balance the ratio between correct and mispronounced samples.

Similarly to DNN’s segmental phone features, the frame-level features are first extended to contain posteriors themselves and posterior ratios between the canonical phone and remaining phone categories. BLSTM-based mispronunciation detectors are then trained on a sequence of frame-level vectors according to the phone labels (correct/incorrect). Each phone’s time boundaries are obtained from forced-alignment. BLSTM- and DNN-based verifiers are built with the KERAS toolkit [116]. The BLSTMs that have one hidden layer with 64 or 128 memory cells are investigated. Due to the limited amount of training samples, e.g., the number of training samples for each phone category is around 2k~16k, The architecture with 64 memory cells is finally selected. The Adam optimizing algorithm [117] is chosen to minimize the cross entropy described in Eq. 3.3. Before training the BLSTM models, cost-sensitive objective function is applied and data pre-processing step has been executed to deal with variable length of input sequences. Namely, zero-padding was performed to pad the shorter tone segment to make every training sample have the same length. 42-time step, equivalent to 0.42 seconds in duration, is selected as the maximum length.

In this experiment, the precision and recall introduced in Section 2.3 are used to evaluate the performance of mispronunciation detectors of Mandarin phone. The reader can treat it as an information retrieval task.

4.3.2 Experimental Results and Discussions

The Precision-Recall curve shown in Fig. 4.2 is used to illustrate the overall mispronunciation detection performance of the baseline and our proposed mispronunciation detectors for Mandarin consonants in the test set. These precision-recall curves are drawn with a single phone-independent threshold. From this Figure, we can observe that the BLSTM-phone based system increases equal precision-recall rate from 84.18% to 85.99%, compared with

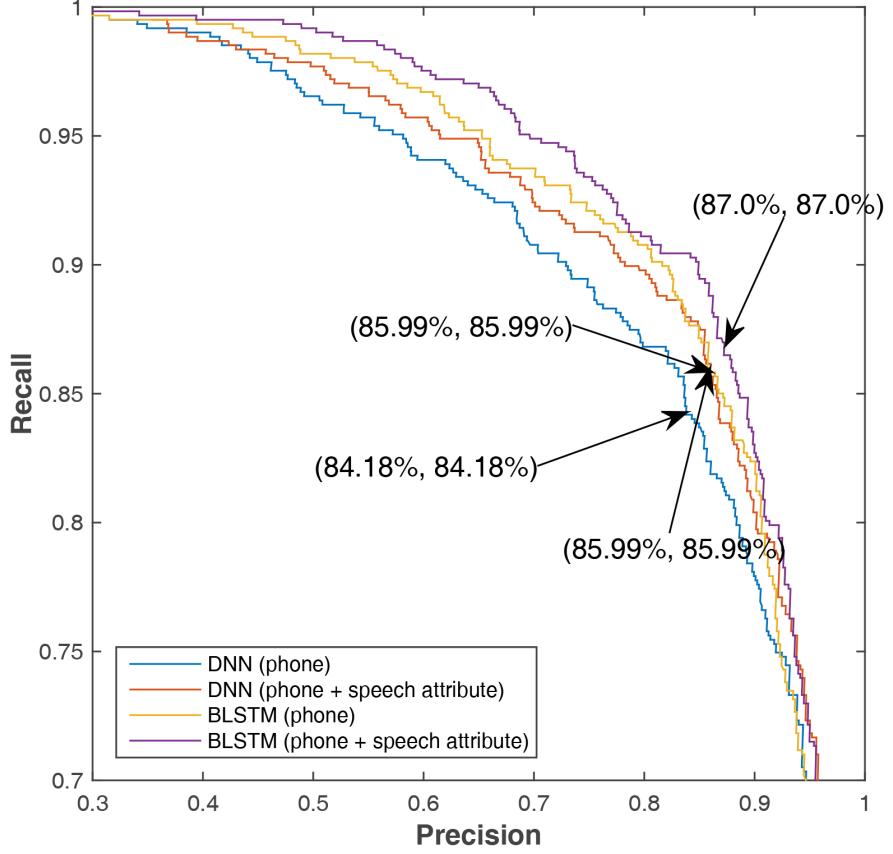


Figure 4.2: Comparison of the overall mispronunciation detection performance of the baseline and our proposed systems.

the DNN-phone based baseline, where direct frame-level averaging is used to calculate the pronunciation score vector. This observation confirms that the context information characterized and modeled by the BLSTM is helpful for verifying phone mispronunciations. Fig. 4.3 is an example of modeling context information to increase the recall rate. In the upper panel, the spectrogram of two syllables and their corresponding expected and surface consonants are shown outside and inside the parenthesis, respectively. A subset of phone detection curves in the second panel is used to visualize and analyze these two consonants. Let us focus on the mispronounced consonant /S/, the frame-level posterior scores for phone /C/ and /Z/ are dominant in its initial frames, then scores for phone /S/ begin to increase, and finally the posteriors of these three phones can be all observed at the end of current segment. Facing this dynamic change of frame-level posteriors, our proposed BLSTM-phone

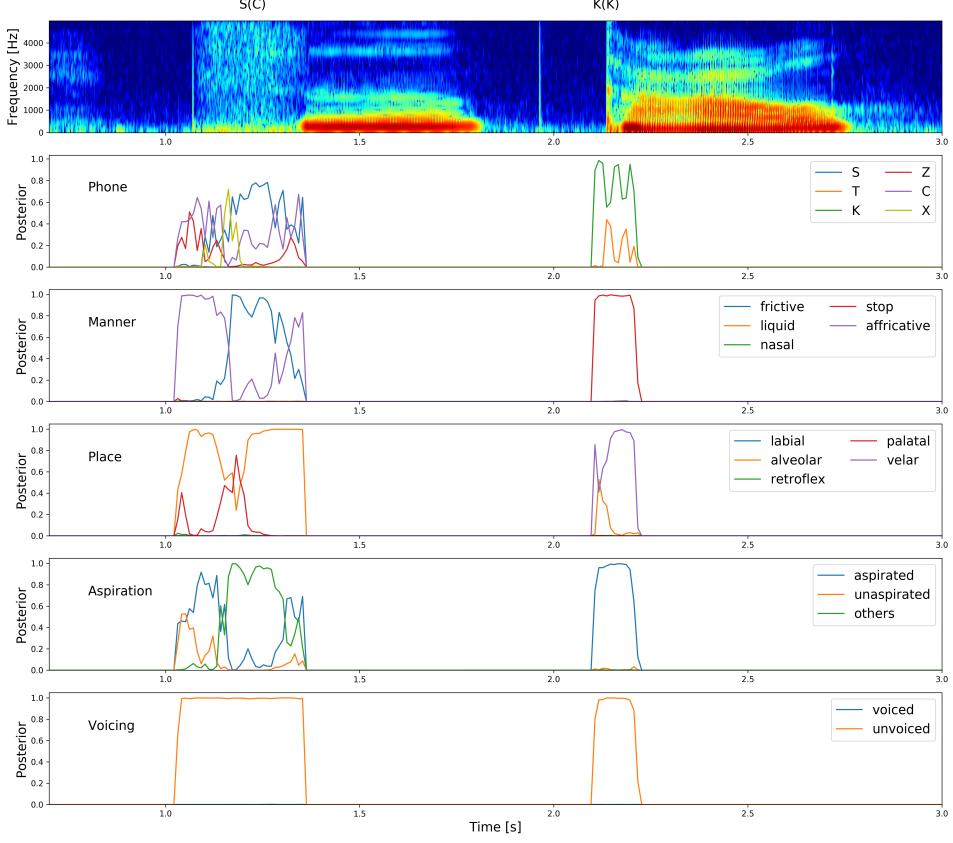


Figure 4.3: Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /S/, /K/, and the human assigned labels are /C/, /K/.

based verifier only assigns 0.15 confidence to judge it is correctly pronounced; however, the DNN-phone based system recognizes it as correct with 0.64 probability. Therefore, if the threshold is set to 0.5, this mispronounced sample will be ignored by the DNN-phone based system. Obviously, the BLSTM-phone based verifier is able to model abovementioned mispronunciation pattern, namely phone /S/ should be classified as mispronounced when some mis-articulations occur at its beginning, even though they are corrected in its latter part.

In addition, the precision-recall curves shown in Fig. 4.2 also show that mispronunciation systems trained with appended speech attribute features outperform systems trained with only conventional phonetic features, independently of using DNN or BLSTM based mispronunciation detectors. The BLSTM-phone-attribute detectors achieve the highest

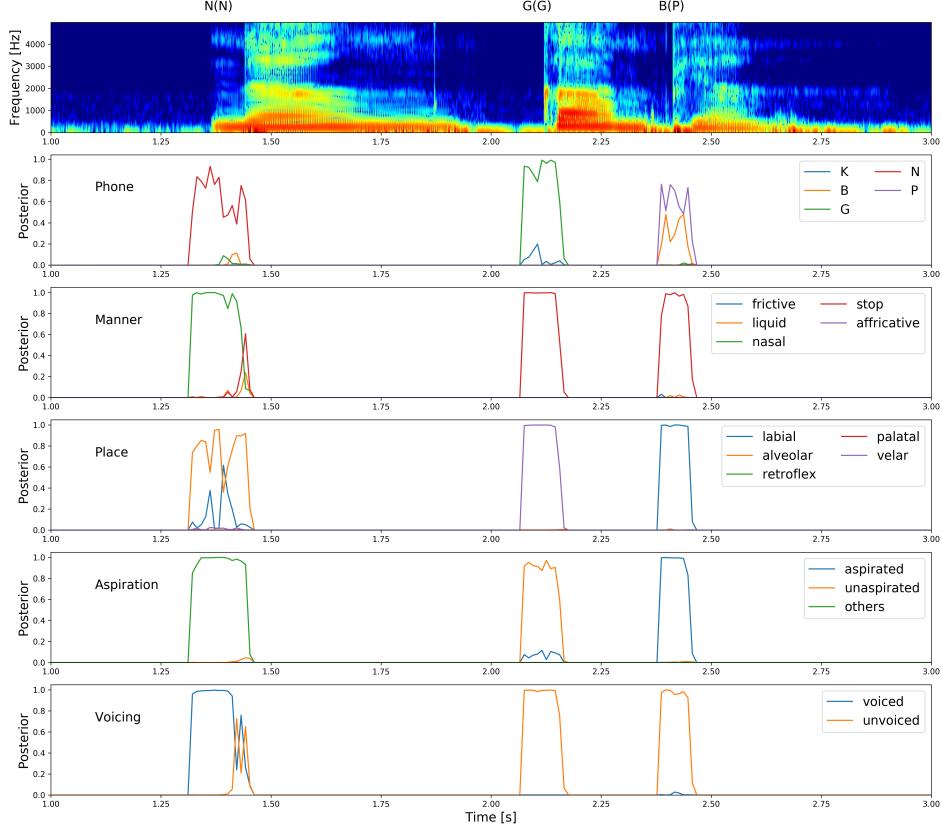


Figure 4.4: Frame-level posteriors of each phonetic and articulatory category for one non-native utterance, where the canonical consonants are /N/, /G/, /B/, and the human assigned labels are /N/, /G/, /P/.

equal precision-recall rate, 87.0%. Experimental results demonstrate that combining articulatory-motivated attributes and phone features is more robust to detect non-native speakers' mispronunciations. Fig. 4.4 is an example of using multiple types of information to detect mispronunciations. In the upper panel, the spectrogram of three syllables and their corresponding canonical and surface consonants are shown. A subset of phone detection curves in the second panel show that /B/ has evident frame-level confidence scores that make the DNN-phone based system assign some posterior to judge unaspirated stop labial phone /B/ as properly pronounced. However, this wrong preference is reduced, when the poor “unaspirated” score of /B/ is observed in the fifth panel. Obviously, speech attribute features can provide complementary information to traditional phone system, where the quality of phone features is not optimal, due to the phone label inconsistency discussed in our

Table 4.2: The mispronunciation detection performance for individual phones, where the precision is set the same as recall

Phone	DNN (phone)	DNN (phone+attribute)	BLSTM (phone)	BLSTM (phone+attribute)	# Training samples
B	85.0%	90.0%	85.0%	90.0%	10,404
C	87.8%	87.8%	90.9%	90.9%	2,527
CH	87.5%	89.5%	89.5%	89.5%	6,056
D	79.1%	79.1%	79.1%	83.3%	16,767
J	75.0%	77.2%	77.2%	79.5%	13,319
K	82.7%	82.7%	82.7%	79.3%	4,428
P	86.6%	93.3%	86.6%	86.6%	2,989
Q	81.9%	86.8%	86.8%	88.5%	7,034
R	69.2%	76.9%	76.9%	92.3%	4,505
S	87.5%	87.5%	93.7%	87.5%	2,849
SH	89.4%	94.6%	94.6%	94.6%	14,360
T	88.8%	92.0%	92.0%	92.0%	8,176
X	91.4%	91.4%	91.4%	91.4%	11,508
Z	90.1%	90.1%	92.1%	92.1%	7,191
ZH	87.8%	87.8%	85.1%	87.8%	10,752

introduction section. One example of visualizing the degree of phone label inconsistency is shown in Fig. 3.2 in Section 3.5.2.

Finally, the mispronunciation detection performance of different systems for 15 most frequently mispronounced consonants is summarized in Table 4.2. A phone-dependent threshold is used to find the optimal operation point for each phone, where precision is set the same as recall. Meanwhile, the training size (e.g., the total number of correct and mispronounced samples) in the test set for each phone category is also shown in the Table 4.2. We can observe that BLSTM-phone based system achieves higher equal precision-recall rate than DNN-phone system for most phones. A similar improvement tendency can also be found when DNN-phone and DNN-phone-attribute based systems are compared. However, the comparison between the fourth and fifth column shows that multi-source information is not always helpful when complex verifier (e.g., BLSTM) is used. For example, after appending speech attribute features, the equal precision-recall rate of BLSTM-based verifier for phone /S/ is reduced from 93.7% to 87.5%. We argue that this

performance degradation mainly results from the shortage of training samples. Namely, due to appended input, BLSTM-phone-attribute is more complex than BLSTM-phone system, thus needs more data to make it more robustly trained. However, Table 4.2 shows that phone /S/ has only 2,849 training samples. In order to resolve the abovementioned problem, we build one directional LSTM-phone-attribute system for phone /S/, and find that the equal precision-recall rate is increased to 93.7%. After adopting the same strategy for phone /K/, we observe that its performance is increased to 82.7%. This improvement tells us that the size of training data plays a critical role in applying complex verifier to detect non-native mispronunciations. Meanwhile, each phone category has a limited number of mispronounced samples in our test set as shown in Table 4.1; therefore we prefer to use the overall phone-independent performance shown in Fig. 4.2 to compare the capabilities of different mispronunciation detectors.

4.4 Summary

In this chapter, the phone label inconsistency discussed in the introduction section and the satisfied speech attribute classification accuracy achieved in the previous chapter motivate us to use speech attribute feature to enhance the traditional phone feature vector fed into mispronunciation detectors. The experimental results show that the combined features can bring higher equal precision-recall rate. Furthermore, after modeling dynamic changes of frame-level pronunciation scores, our proposed BLSTM-based verifier can deliver a meaningful improvement in both precision and recall. It supports our initial motivation that dynamic changes that exist in frame-level detection curves are useful for non-native mispronunciation detection.

CHAPTER 5

IMPROVING MISPRONUNCIATION DETECTION OF MANDARIN TONES WITH SOFT TARGET TONE LABELS AND BLSTM-BASED MODELS

In this chapter, we investigate the effectiveness of soft target tone labels and sequential context information for mispronunciation detection of Mandarin lexical tones. In conventional approaches, prosodic information (e.g., F0 and tone posteriors extracted from trained tone models) is used to calculate goodness of pronunciation (GOP) scores or train binary classifiers to verify pronunciation correctness. We propose three techniques to improve detection of mispronunciation of Mandarin tones for non-native learners. First, we extend our acoustic tonal model from a deep neural network (DNN) to a bidirectional long short-term memory (BLSTM) based recurrent neural network (RNN) in order to more accurately model contextual information from tone-level co-articulation and non-native tone production. Second, we characterize ambiguous pronunciations where L2 learners' tone realizations are between two canonical tone categories by replacing hard targets with posterior probabilities based targets, which are usually referred to as soft targets. Third, segmental tone features fed into verifiers are here extracted by a BLSTM-based RNN to exploit sequential context information and improve mispronunciation detection. Experimental results demonstrate that our proposed techniques result in a significant performance improvement over the traditional DNN-based verifier trained with tone feature extracted from DNN-based acoustic tonal models.

5.1 Introduction

Lexical tone error is a special type of mispronunciation in tonal languages, such as Mandarin Chinese, where each word is composed of one to several characters. Each character is pronounced as a basic syllable with five different tones including four lexical tones and

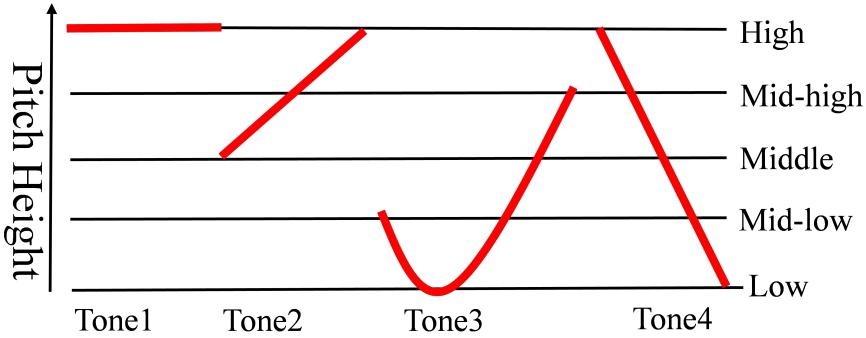


Figure 5.1: Pitch contours of standard Mandarin lexical tones. Tone 5 is not depicted as it does not have a defined pitch contour.

Table 5.1: Characteristic of pitch contours of Mandarin tones

Tone	Pitch Contour	English Equivalent
1	High-level	Singing
2	High-rising	Question-final Intonation; e.g., What?!
3	Dipping	No equivalent
4	Falling	Curt commands; e.g., Stop!
5	Undefined	Unstressed syllable

one neutral tone (5). Mandarin tones are characterized by their pitch contours shown in Fig. 5.1, where Tone1 keeps high-level pitch height, which is equivalent to singing in English; Tone2 starts at lower pitch height and keeps rising like in English questions; Tone3 has a dipping pitch contour, there is no equivalent in English, and Tone4 falls from the highest pitch height to the lowest pitch height like English curt commands. These tones characteristics are summarized in Table 5.1. The same base syllable when combined with different tones will result in different lexical meanings; e.g., ma1 (mother), ma2 (hemp), ma3 (horse) and ma4 (scold) have the same toneless base syllable, namely ma, yet different lexical tones. Some recent studies [106, 118] have proposed to use entropy-based functional load to measure the importance of tone for Mandarin communication and have shown that the information loss caused by tone level mispronunciation is as high as its

counterpart mispronunciation at the phone level. Therefore, tone classification and mispronunciation detection subsystems [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30] always play a key role in Mandarin CAPT systems.

Over the past two decades, many approaches have been proposed to detect tone-level mispronunciations. A template-based detection framework was first introduced in [21], where discrete pitch values and the duration of the tested tone are compared with the canonical tone template to decide whether the tone in question is mispronounced. In [22], Euclidean distance was adopted to calculate the discrepancy between L2 learners' and ideal contours in terms of energy and pitch. Although template-based solutions have obtained an acceptable accuracy, superior mispronunciation detection results were attained by leveraging upon statistical model-based frameworks [23, 24, 25, 26, 27, 28, 29, 30], in which prosody-related features (e.g., pitch, energy [24, 25, 26, 27, 28, 29, 30], or fundamental frequency variation [23]) are first extracted to train tone classifiers, e.g., decision tree [27], HMM [24, 25], GMM [23], and ANN [26]. Subsequently, given a test tone segment, its posterior or goodness of pronunciation (GOP) score is used to measure the quality of tone-level pronunciation. Inspired by recent findings, e.g., 1) cepstral features play an important role in tone recognition within a DNN framework [76, 77]; 2) tone pitch realization and perception are influenced by the underlying phone units [78, 79], DNNs were hence proposed to map the combined cepstral and tonal features to frame-level tone posteriors [28, 29, 30], which in turn are fed into tone verifiers to finally assess the pronunciation correctness of the current tone.

Although the abovementioned model-based CAPT systems have delivered satisfactory mispronunciation detection results, there is still potential for further improvement. Compared with phone-level co-articulation, tone realization is influenced by a broader temporal context (e.g., 2 or 3 neighboring syllables) [119]. In [19, 20], we have demonstrated that a better tone classification accuracy could be achieved by expanding the temporal information provided to the DNN-based acoustic models; nevertheless, the temporal informa-

tion captured by DNN is limited to the fixed context window spanning several consecutive speech frames. Furthermore, L2 learners tend to slow down their pronunciation speed and increase each syllable’s duration due to their unfamiliarity with tone production, especially for Tone2 and Tone3 [120, 121, 122]. In this chapter, we put forth the use of bidirectional long-short-term memory (BLSTM) based recurrent neural networks [50] in place of DNNs, because they are more suitable in handling long-term dependencies of acoustic and prosodic features and could thereby better model tone-level co-articulation and non-native tone production. More accurate tone posteriors extracted from BSSTM-based acoustic tone models could thus be fed into a verification module.

In addition, the performance of model-based CAPT system also heavily depends upon the quality of tone-level labeling of the non-native corpora employed in training the tonal models used in pronunciation scoring. However, tone pronunciations of non-native learners often fall between two canonical tone categories. The use of hard targets (one-hot target [47]) is thus suboptimal, and we propose the use of soft targets for tone modeling. Soft targets were used in [43, 44, 45, 46] to transfer knowledge acquired from large-size complex models to small-size DNNs. We instead use soft targets to help resolve the hard-assignments of non-native tone labels. Soft targets are more suitable for describing non-native tone realization than hard targets, since, for example, the non-native pitch contour may lie between two canonical tones or does not resemble any canonical tone at all.

Finally, similar to phonetic counterpart, frame-level tonal detection curves also exhibit dynamic changes within each tone segment. In order to make full use of this sequential context information at the verification stage, conventional frame-level averaging in feed-forward DNN-based verifiers [4, 5] is replaced by a BLSTM, employing learned context information to embed a sequence of frame-level pronunciation scores into a pronunciation vector. The proposed BLSTM-based embedded pronunciation vector is expected to contain more context information for subsequent mispronunciation detection. Our results show that it is especially useful for verifying pronunciation correctness of Tone2 and Tone3.

Table 5.2: Confusion Matrix of Mandarin tone accuracy (%) on iCALL corpus analyzed in [55]

Tone Categories		Non-native Production				
		Tone1	Tone2	Tone3	Tone4	Tone5
Canonical	Tone1	72.6	8.3	7.1	11.7	0.2
	Tone2	13.3	67.7	8.0	10.7	0.3
	Tone3	12.0	19.3	58.8	9.7	0.2
	Tone4	13.5	9.9	7.9	68.4	0.3
	Tone5	6.4	6.3	3.6	8.1	75.5

5.2 Modeling Challenges of Non-native Lexical Tones

In this thesis, the first language (L1) of all second language (L2) learners in the adopted iCALL corpus [55] is non-tonal. In [55], a difficulty of tone learning was quantified in a confusion matrix shown here in Table 5.2 for the reader's reference. Other studies in acoustics [122], phonetics [121], and language education [120] have also reported the difficulty for non-tonal L1 learners of Mandarin to perceive and/or produce lexical tones. In the next two subsections, we elaborate on specific challenges posed by characteristics of non-native lexical tone productions.

5.2.1 Variability in Pitch Contour of L2 Tone Productions

In iCALL, the acoustic distribution of non-native lexical tone productions often does not conform to the canonical pitch contours shown in Fig. 5.1. We select one non-native utterance, its pitch contour and human labeling are shown in Fig. 5.2, and we can verify that the pitch contour of the second syllable not only contains characteristic of Tone4, namely a falling slope is observed, but also exhibits similar pattern as Tone1, namely the beginning part of current tone's pitch contour is a straight line without salient slope and maintains at a high pitch level like canonical Tone1 behaves in Fig. 5.1. This pattern resembles a Tone1 followed by a Tone4 within the same syllable. (Note that each syllable is expected to take on only one lexical tone for monosyllabic tonal languages like Mandarin). Due to the fact that ground-truth tone label being ill-defined, it can result in inevitable labeling noise by

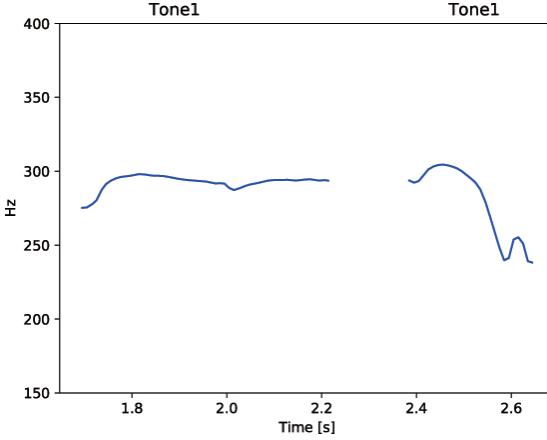


Figure 5.2: One example of pitch contours of non-native lexical tones, where human assigned labels are Tone1 and Tone1

an annotator. In this work, a computational model exploiting soft counting of the different class labels is proposed to better characterize the underlying heterogeneous acoustic distributions of non-native lexical tone productions using probability scores, addressing the technical challenges stemming from the high variability in non-native tone productions.

5.2.2 Prolonged and High-Variance Duration in L2 Tones

Studies have shown that non-native tone productions are often longer than those of native speakers [37, 38]. The L2 learners' difficulty of the tone production seems to make L2 speakers reduce their pronunciation speed causing a corresponding increase in the duration of the non-native tone realization. In this work, we approximate the time durations of each lexical tone in a native speech corpus (the Mandarin 863 speech recognition corpus [89]) and non-native speech corpus (iCALL [55]) through time boundaries generated by forced alignments using acoustic models trained with native and non-native corpora. We list these approximate time durations in Table 5.3 (for all phonemes) and Table 5.4 (only considering the phoneme /A/). In both cases, we observe that non-native tones are longer than those of native tones. The standard deviations (SD) of the durations are also higher for L2 tones. This implies that simply elongating a fixed-context window spans in a DNN

Table 5.3: The mean and standard deviation (in milliseconds) of the duration of the four lexical tones where the underlying phone’s identity is ignored

	Tone1		Tone2		Tone3		Tone4	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Non-native	302.3	122.6	319.2	123.2	302.7	131.6	263.9	98.0
Native	156.0	63.2	158.2	66.4	153.4	64.2	156.7	71.4

Table 5.4: The mean and standard deviation (in milliseconds) of the duration of the four lexical tones where the underlying phone’s identity is /A/

	Tone1		Tone2		Tone3		Tone4	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Non-native	244.0	109.7	289.2	125.6	262.1	126.2	234.7	98.4
Native	132.6	55.9	151.4	69.7	142.2	62.1	147.7	69.5

model might not be enough to characterize duration variability. Along with the fact that the tone realization is influenced by a broader temporal context (e.g., 2 or 3 neighboring tones) [119]. BLSTM-based recurrent models are thereby proposed to better characterize prolonged non-native tone productions.

5.3 Overview of The Tone Mispronunciation Detection Framework

Fig. 5.3 shows both the proposed and the baseline conventional tone mispronunciation detection frameworks, which consist of three blocks: (i) The frame-level tone-related feature extraction module, in which DNN-based or BLSTM-based acoustic tonal model trained with hard targets or soft targets is investigated; (ii) the baseline DNN-based tone-dependent mispronunciation verifier training module, which is based on averaged segmental tone scores; and (iii) the proposed BLSTM-based tone-dependent mispronunciation verification training module, where a sequence of frame-level features and the corresponding binary labels are used to train mispronunciation detectors.

5.3.1 Acoustic Tonal Model Training with Hard Target

Similar to the speech attribute and phone classifier training described in Section 3.3, the baseline acoustic tonal model is trained by minimizing the cross-entropy between the pre-

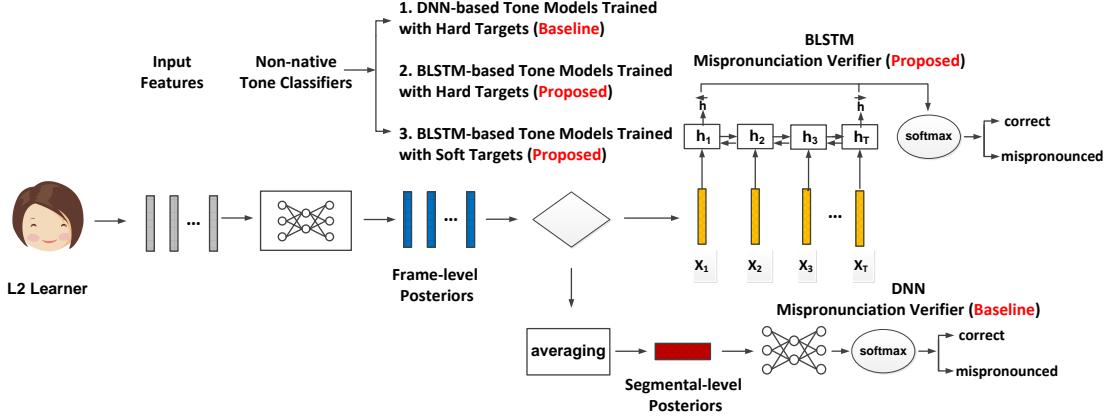


Figure 5.3: Overview of the tone mispronunciation detection framework.

dicted and true distributions as shown in Eq. 3.3. Their differences are: 1) the input features are now extended to contain pitch related features; 2) the ground-truth targets are replaced with tone-related senones. Due to the fact that each speech frame is labeled by one of tone-related senones, we call it as hard target (one-hot target).

5.3.2 Mandarin Tones Soft Target Generation

Upon hard targets based acoustic tonal models training completion, we use a tone-based extended recognition network (ERN) and the trained CD-BLSTM-HMM to decode each training utterance into a lattice, where each frame is annotated with senone labels (which share the same toneless phone but own different tone marks) and their probabilities. We refer to the decoded lattice as soft targets, which are subsequently used to retrain the CD-BLSTM-HMM model. Similar to the phone-based ERN [72, 73, 74] adopted for phone-level mispronunciation detection, a tone-based ERN is constructed by expanding each toneless syllable into five different tonal syllables. Fig. 5.4 shows an example of a tone-based ERN, where the toneless syllable sequence is “zhong guo”. The tone-based ERN used as a grammar (i.e., a language model) is designed to constrain the search space to keep the trained acoustic model focused on tone discrimination. Namely, each node in the decoded

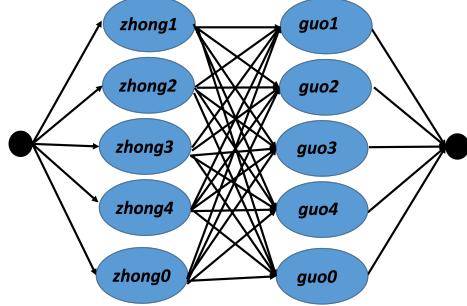


Figure 5.4: An example of tone-based ERN.

speech lattice is only allowed to be associated with the provided toneless syllable, yet with different tone marks.

5.3.3 Acoustic Tonal Model Training with Soft Target

After having generated soft targets, the proposed tone model is trained by minimizing the Kullback-Leibler (KL) divergence between the model's output and the soft targets as follows:

$$L^{(KL)}(\theta) = \sum_i \sum_j \hat{y}_i^j \log \frac{\hat{y}_i^j}{y_i^j}, \quad (5.1)$$

where y_i^j and \hat{y}_i^j are the j -th dimension value of predicted target and soft targets at a time i . Gradient descent is used to update parameter θ associated with deep models (e.g., DNN and BLSTM) to minimize the KL divergence loss.

5.3.4 Non-native Tone Feature Extraction

After the abovemetioned acoustic tonal models/classifiers are well trained, we can use them to extract tone posteriors for subsequent mispronunciation detection. Please consult Section 3.4 for detailed description of extracting frame-level and segmental-level features. The only difference is that now variable P in Eqs. 3.4 and 3.5 represents one element in the Mandarin tone set, which consists of the five predefined Mandarin tone categories and one additional category for representing non-tone units (e.g., consonants, silence).

5.3.5 Tone Mispronunciation Verifier Construction

Similar to phone mispronunciation detectors, DNN (baseline) and BLSTM (proposed) based tone mispronunciation verifiers are trained with extracted tone-related features and binary labels (correct or mispronounced). Please consult Sections 4.2.2 and 4.2.3 for the detailed training procedure.

5.4 Experiments

Non-native Mandarin tone mispronunciation detection is also carried out on the largest non-native Mandarin corpus iCALL [55], where the surface tone label representing the real tone production of non-native learner is given by three native experts. After comparing the surface and expected transcriptions, we can get each tone’s pronunciation label, e.g., correct or incorrect.

295 speakers’ data in iCALL corpus is selected and split into three portions. First, a subset of iCALL with 215 speakers data is mixed with two native speech corpora introduced in chapter 2 to train our acoustic tonal models/classifiers. Next, another subset with 50 speakers’ data is used to train DNN or BLSTM-based binary classifier. Lastly, the remaining part of iCALL with 30 speakers’ data is used to evaluate our system performance. Different from non-native data partition in phone mispronunciation experiment, we move 115 speakers’ data from the data used for training binary verifier to our acoustic model design. Because we find that the acoustic tonal model trained with more non-native data can generate more reliable soft targets.

In this study, we are concerned with detecting the mispronunciation of 4 lexical tones in Mandarin. Table 5.5 lists some statistics of these tones in our train and test sets. Similar to the phone mispronunciation rates shown in Table 4.1, our tone mispronunciation rate in the test set is also much higher than its counterpart in the train set. This higher mispronunciation rate in our test set is caused by the previously described data selection and refined tone

Table 5.5: The number of each tone’s correct and mispronounced samples in our experiments, where the overall mispronunciation rates in the train and test sets are 26.3% and 53.6%, respectively

Tone	Train Set		Test Set	
	# Correct	# Mispronounced	# Correct	# Mispronounced
Tone1	4,448	1,144	446	328
Tone2	3,883	1,531	301	381
Tone3	2,558	1,685	171	332
Tone4	6,687	1,940	550	661

labels. Each selected tone sample in our test set is double checked by Beijing Language and Culture University graduate students majoring in Chinese phonology. They utilized their auditory perception and analysis of observed pitch contours to refine the original tone labels in iCALL.

5.4.1 Experimental Setup

The open source Kaldi toolkit [111] is used to train our acoustic tonal models: a context-dependent Gaussian mixture hidden Markov (CD-GMM-HMM) acoustic model is initially trained using the maximum likelihood (ML) criterion. Then CD-DNN-HMM and CD-BLSTM-HMM models are built using hard targets, e.g., the state level tonal phone alignments provided by the trained CD-GMM-HMM and human labeled transcriptions. The DNN acoustic tonal model has six hidden layers each having 2048 Sigmoid units. Its input is a window of 21 consecutive speech frames centered at the current frame. Each frame contains 43-dimensional log mel filter bank (LMFB) coefficients, F0, the probability of voicing (POV) [123], and the corresponding derived velocity, and acceleration features. The BLSTM-based architecture has two hidden layers, and 320 memory cells for each layer.

Once the hard targets based acoustic tonal models training is finished, we use a tone-based extended recognition network (ERN) and the trained CD-BLSTM-HMM to generate soft targets, which are used to re-train CD-BLSTM-HMM acoustic tonal model. To have a

fair comparison, the CD-BLSTM-HMM trained with soft targets has the same architecture and input feature as its hard targets based counterpart.

Regarding to DNN-based mispronunciation verifiers, they are trained with segmental feature vectors including tone related posterior and posterior ratio, and their corresponding pronunciation labels (correct/incorrect). Different numbers of hidden layers (1, 2) are evaluated on the development set (one-tenth of the total data for verifiers' training). The structure of 2 hidden layers, each with 128 hidden nodes, achieves the best performance for tone-dependent verifiers trained with tone features extracted from the BLSTM models. However, for verifiers trained with tone features extracted from DNN model, the Tone3's optimal architecture is 1 hidden layer, each containing 128 neurons. The preferred shallow DNN verifiers can exhibit better generalization when trained with imperfect tone features, e.g., Tone3 posteriors and ratios extracted from DNN models. Similar to previous work [4, 5], shallow DNN-based verifiers always yield good performance, due to the limited training samples.

With regard to BLSTM-based mispronunciation detectors, they are trained on a sequence of frame-level vectors according to the tone labels (correct/incorrect). Each tone's time boundaries are obtained from forced-alignment. BLSTM and DNN verifiers are built with the KERAS toolkit [116]. The BLSTMs have one hidden layer each with 64 or 128 memory cells are investigated. Due to the limited training sample, e.g., the number of training samples for each tone category is around 4k~8k, 64 memory cells are preferred, except for Tone2 and Tone3, where 128 memory cells are selected, when the tone features are extracted from BLSTM-based acoustic tonal models. The Adam optimizing algorithm [117] is chosen to minimize the cross entropy described in Eq. 3.3. Before training the BLSTM models, data pre-processing step has been executed to deal with variable length of input sequences. Zero-padding was performed to pad shorter tone segments to make every training sample have the same length. 100-time step, equivalent to 1 second in duration, is selected as the maximum length.

In this experiment, we adopt false acceptance rate (FAR) and false rejection rate (FRR) as our measurements. The reason is that our test set has around 53.6% tonal errors (positive samples), which largely reduces the imbalance between positive and negative samples that existed in our phone mispronunciation detection experiment. Therefore, measurements that both take positive and negative samples into account are selected. Please consult Section 2.3 for the detailed measurement calculation.

5.4.2 Experimental Results and Discussions

After collecting the segmental tone features in the test set, well-trained DNN-based classifiers then assign a posterior probability to each non-native tone production to see how likely the given tone is correctly pronounced. Finally, the generated posteriors along with the tone labels (correct/mispronounced) are used to draw FAR-FRR detection curve for each lexical tone category as shown in Fig. 5.5. Meanwhile, the equal error rate (EER), where FAR is equal to FRR, is highlighted in this figure.

From Fig. 5.5, we can see that the BLSTM-based acoustic tonal model (orange curve) trained with hard targets achieves a lower EER than its DNN counterpart (blue curve); for instance, 22% and 13.7% EER reduction is respectively observed for Tone2 and Tone3. For Tone1 and Tone4, although EER reduction is observed, substantial improvements are no longer observed across all operation points. We argue that it may due to the imperfect tone labeling of training samples. Specifically, a more sophisticated data-driven model, such as BLSTM-based acoustic model, could be more sensitive to noisy training labels than a simpler connectionist model like a DNN [124]. Moreover, non-native tone labeling is a challenging task, as discussed in the introduction section; therefore, tone labeling errors are likely to happen, which in turn bring more challenges into the training procedure of BLSTM-based acoustic tonal models.

Facing the same issue of unavoidable tone labeling errors, Tone2 and Tone3 yet achieve significant EER reduction. Long-range dependencies among acoustic and prosodic features

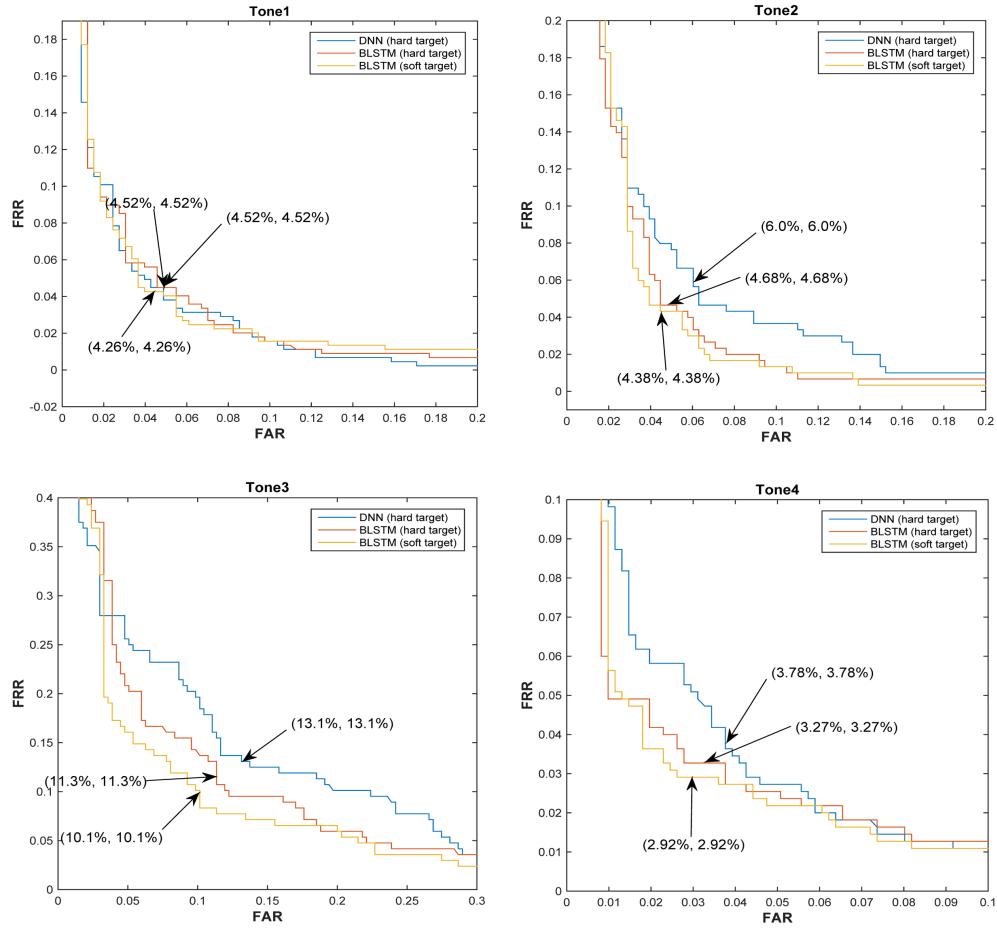


Figure 5.5: Tone-dependent detection curves and EERs for DNN-based tone mispronunciation verifiers trained with segmental tone features vectors extracted from different acoustic tonal models

captured by the BLSTM-based system seem to provide a substantial beneficial effect on Tone2 and Tone3 modeling. Fig. 5.6 is an example of a testing utterance, where human assigned labels are Tone3 and Tone4. In Table 5.6, tone-related posteriors produced by DNN and BLSTM are summarized, where each posterior characterizes how likely the non-native pronounced tone belongs to each tone category, e.g., the posteriors in lower right corner of Table 5.6 indicate that both DNN and BLSTM recognized that the second non-native tone pronunciation belongs to the Tone4 class with a posterior probability value (confidence score) very close to 1.0. For the first non-native tone production, although the human annotator and the BLSTM system agreed that it is pronounced as Tone3, the DNN-

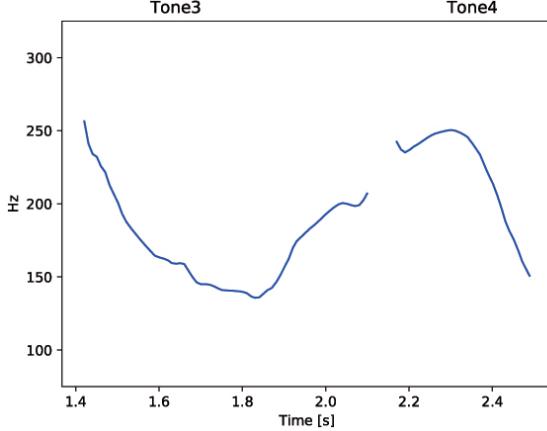


Figure 5.6: Pitch contours of non-native pronunciation, where human assigned labels are Tone3 and Tone4

Table 5.6: Tone posteriors produced by DNN and BLSTM for Figure 5.6.

	Tone3		Tone4	
	DNN	BLSTM	DNN	BLSTM
Tone1	0.01	0.01	0.04	0.01
Tone2	0.55	0.41	0.0	0.0
Tone3	0.24	0.56	0.0	0.0
Tone4	0.2	0.02	0.96	0.99

based system gives a higher preference to Tone2, assigning to it a confidence score equal to 0.55 (as shown in the second column of Table 5.6). We think that Tone2 is preferred by the DNN-based system because its pitch contour is highly similar to Tone2 at the end of the syllables, e.g., the rising slope in Fig. 5.1. Meanwhile, the DNN-based system also gives a 0.2 confidence score to Tone4 category. The reason is that the beginning part of this tone's pitch contour exhibits a falling slope as canonical Tone4 behaves in Fig. 5.1. The fixed-context input window of the DNN-based system may therefore fail at capturing all of the discriminative information existing at the current syllable/tone. That is, the duration of the pitch contour in Tone3 is around 0.7 second, which cannot be appropriately modeled by a DNN with only fixed context input.

In addition, the BLSTM architecture is also helpful in modeling tone-level long-range dependencies with coarticulation influenced by several (2 or 3) neighboring syllables. In

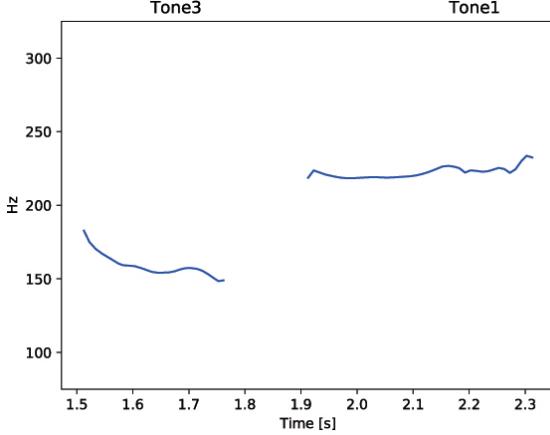


Figure 5.7: Pitch contours of non-native pronunciation, where human assigned labels are Tone3 and Tone1

Table 5.7: Tone posteriors produced by DNN and BLSTM for Figure 5.7.

	Tone3		Tone1	
	DNN	BLSTM	DNN	BLSTM
Tone1	0.52	0.0	0.95	1.0
Tone2	0.09	0.02	0.03	0.0
Tone3	0.26	0.97	0.02	0.0
Tone4	0.13	0.01	0.0	0.0

Fig. 5.7, a test utterance is shown, where the human assigned labels are Tone3 and Tone1. We analyze the syllable marked by Tone3. Although the BLSTM-based system generates a high 0.97 confidence score for Tone3, the DNN-based system signals a mispronunciation event by assigning a confidence score as low as 0.26 to it (see the second column in Table 5.7). The DNN-based system wrongly prefers Tone1 (a score of 0.52) over Tone3 since the non-native pitch contour of the tone in question is similar to canonical Tone1 (see Fig. 5.1), where pitch trajectories are both flat, with no significant rising or falling trajectories. However, the pitch contour of the tone in question can also be a tone-level co-articulation product of Tone2 or Tone3, when the preceding tone is Tone4 and the following tone is Tone1 [119]. The experimental evidence gathered in this section suggest that the context information captured by the input layer using a fixed context window does not allow the DNN to correctly model the entire pitch contour of the target tone and properly decide

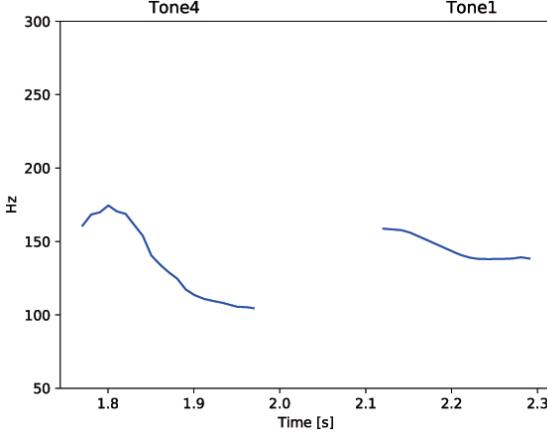


Figure 5.8: Pitch contours of non-native pronunciation, where human assigned labels are Tone4 and Tone1

Table 5.8: Tone posteriors of hard targets and proposed soft targets for Figure 5.8.

	Tone4		Tone1	
	hard target	soft target	hard target	soft target
Tone1	0.0	0.0	1.0	0.44
Tone2	0.0	0.0	0.0	0.0
Tone3	0.0	0.0	0.0	0.0
Tone4	1.0	1.0	0.0	0.56

whether the pitch contour is the canonical Tone1’s realization, or the product of tone-level coarticulation. Therefore, a BLSTM-based system, which by design can exploit long-range dependencies of prosodic features, is more suitable for modeling tone-level coarticulation and prolonged non-native tone production.

Although acoustic tonal models leveraging BLSTM properties have proven to significantly enhance tone level mispronunciation detection performance, there is still room for further improvement. In fact, soft targets can be considered as a probability-based transcription used to retrain BLSTM-based tone models (yellow curves in Fig. 5.5) and the resulting EERs are further reduced for each lexical tone category. Such an outcome indicates that soft targets could be more suitable for labeling irregular non-native tone pronunciations, namely L2 learners’ tone realizations often fall between two canonical tone categories and do not belong to a single category. Traditional hard target is not an optimal

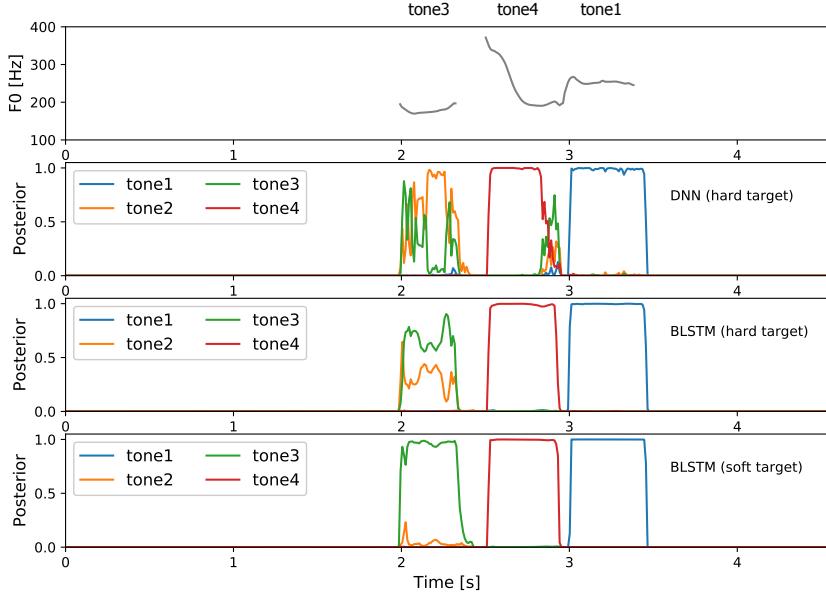


Figure 5.9: Frame-level tone posteriors generated with different tone classifiers.

choice for labeling the above-mentioned non-native tone realizations, as shown in the following example. The proposed soft target labeling strategy is displayed in Fig. 5.8, where one training utterance is shown, and in Table 5.8, where traditional hard targets and the proposed soft targets are reported. From Fig. 5.8, we can verify that the pitch contour of the second syllable not only contains characteristic of Tone4, namely a falling slope is observed, but also exhibits a similar pattern as Tone1, namely the end part of current tone's pitch contour is a straight line without slope and maintains at a high pitch level as canonical Tone1 behaves in Fig. 5.1. As a consequence, the proposed soft target labeling strategy allows the acoustic tonal models to assign a confidence score of 0.44, and 0.56 for Tone1, and Tone4, respectively. In contrast, the traditional hard target solution assigns a 1.0 posterior to Tone1, other tone categories are considered highly not probable, with a confidence score equal to zero. Finally, the frame-level detection curves generated by the DNN trained with hard targets, and the BLSTMs trained with hard and soft targets are compared in Fig. 5.9 to demonstrate the efficiency of the proposed soft target training. In Fig. 5.9, we can see that the frame-level scores generated from the BLSTM (soft target) are more consistent with the human labeling in the top panel of Fig. 5.9.

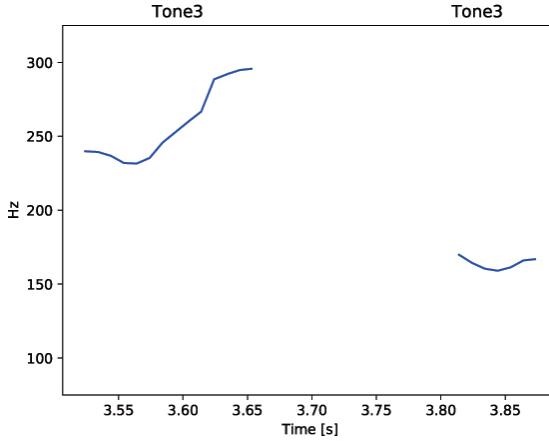


Figure 5.10: Pitch contours of native pronunciation, where human assigned labels are Tone3 and Tone3

Table 5.9: Tone posteriors of hard targets and proposed soft targets for Figure 5.10.

	Tone3		Tone3	
	hard target	soft target	hard target	soft target
Tone1	0.0	0.0	0.0	0.0
Tone2	0.0	0.38	0.0	0.0
Tone3	1.0	0.62	1.0	1.0
Tone4	0.0	0.0	0.0	0.0

In addition to giving a more accurate description of non-native tone productions, our proposed soft targets can also characterize tone sandhi [125], which refers to the phenomenon that, in continuous speech, some lexical tones may change their tone category in certain tone contexts. In Mandarin Chinese, the most common tone sandhi rule is “Tone3 Tone3 → Tone2 Tone3”, where the first Tone3 in a set of two third-tone syllables is converted to Tone2. In the non-native corpus design, we try our best to avoid the abovementioned tone sandhi. However, the native corpora used to train acoustic tonal models do not take the tone sandhi into consideration. Therefore, their original tone labels might not be optimal for tone modeling. In this study, we find that our soft target tone label can automatically reflect the mentioned change. In Fig. 5.10, two consecutive Tone3 spoken by a native speaker are shown, where the pitch contour of the first Tone3 keeps rising and ends at high-level pitch height like the canonical Tone2 shown in Fig. 5.1. Facing this tone

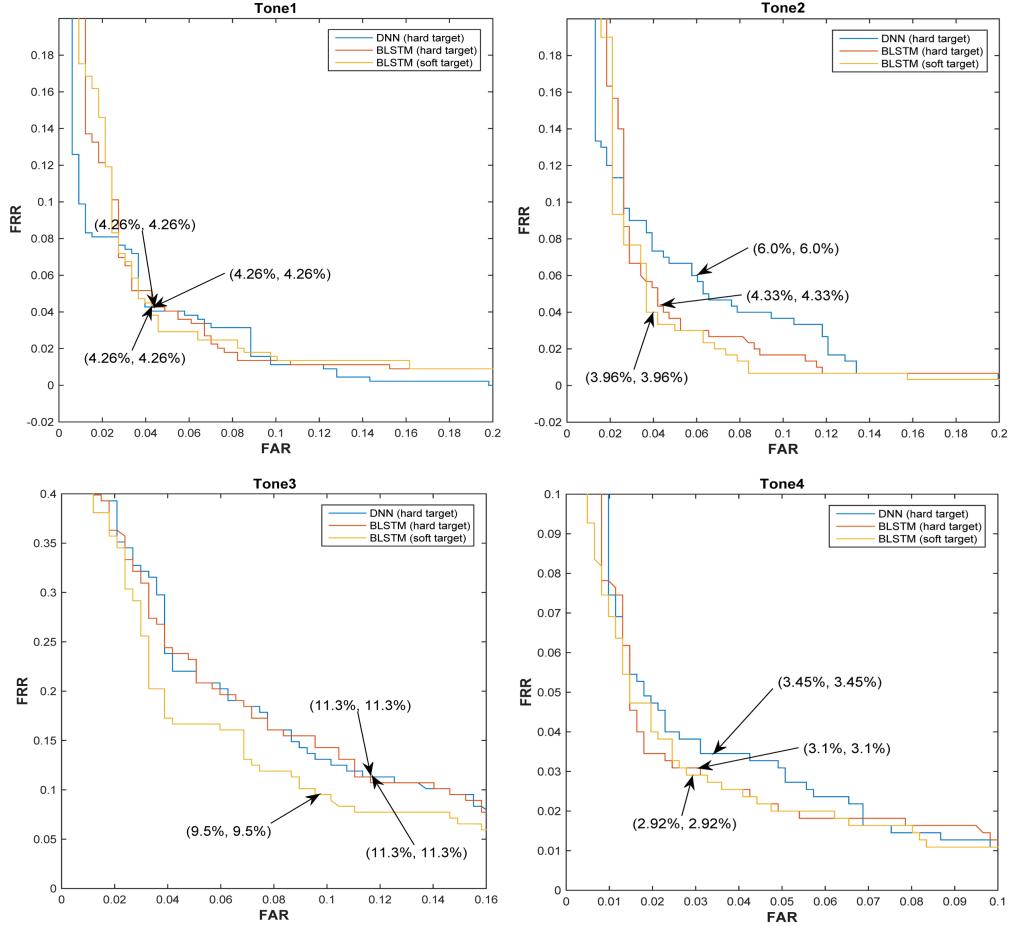


Figure 5.11: Tone-dependent detection curves and EERs for BLSTM-based tone mispronunciation verifiers trained with vectors of a sequence of frame-level tone features extracted from different acoustic tone models

sandhi, our generated soft targets shown in Table 5.9 assign 0.38 posterior to Tone2, and 0.62 to Tone3. Obviously, compared with conventional hard target labeling, proposed soft target can partially solve the tone label inconsistency caused by tone sandhi.

Now, let us investigate the improvement brought by better tone feature embedding learned from a sequence of frame-level tone related features. Unlike DNN-based mispronunciation verifiers trained with simple averaged frame-level features, our proposed BLSTM-based classifier makes full use of sequential context information to embed a sequence of pronunciation scores into a vector for verifying whether the current tone is mispronounced or not. Its corresponding detection curves and EERs are shown in Fig. 5.11,

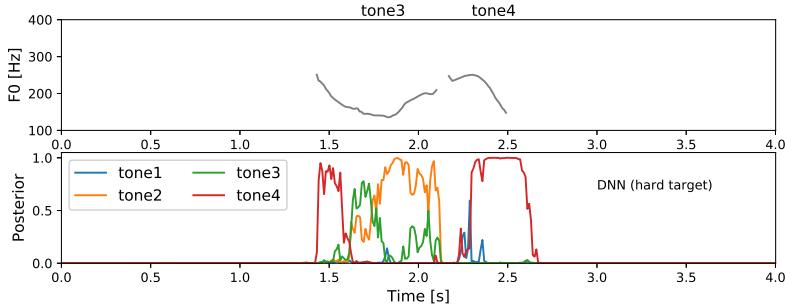


Figure 5.12: Frame-level tone posteriors generated with DNN (hard target) tone classifier.

in which EER reductions are observed after introducing BLSTM acoustic tonal models and soft target training, similar to DNN-based mispronunciation detectors. Contrasting Figs. 5.11 and 5.5, we can find that the BLSTM-based verifiers attain lower EERs than their DNN-based counterparts, especially when tone features are extracted from DNN-based tone models (blue curves). For example, the EER for Tone3 in Fig. 5.5 is equal to 13.1%, which is reduced to 11.3% in the Fig. 5.11. This improvement demonstrates that the sequential information of frame-level posterior scores learned by BLSTM block can bring extra gain over traditional DNN-based verifiers, where simple frame-level averaging results in the loss of useful sequential information.

One example is given in Fig. 5.12, where frame-level posteriors extracted from DNN-based acoustic model are presented. Regarding the detection curves of first Tone3, we can observe that the posterior probability scores for Tone4 are dominant in the initial frames, then posterior probability scores for Tone3 begin to increase (middle frames), and the final frames are more likely to be classified as Tone2. Facing this dynamic change of frame-level posteriors, our proposed BLSTM-based verifier assigns a 0.59 confidence score to judge that it is correctly pronounced; however, the DNN-based system recognizes it as a correct instance with only 0.44 probability. Therefore, if the threshold is set to 0.5, one false rejection will be generated. Obviously, the sequential information of frame-level posterior scores stored in the BLSTM block is beneficial to Tone3's verification. Moreover, we find that the improvement brought by sequential information becomes smaller when

Table 5.10: Tones EERs (%) of different verifiers with features extracted from different tone models, where * denotes soft target training. If none, hard target is utilized.

Tone Features Extracted from	Tones EERs (%)		Tone Verifiers	
	DNN	BLSTM	DNN	BLSTM
DNN	6.85	6.25		
BLSTM	5.94	5.74		
BLSTM*	5.41	5.16		

features fed into BLSTM-based verifier become long-range context-aware. Namely, benefit from long-range dependencies modeling ability, BLSTM-based acoustic model can produce more accurate tone features, so that the improvement brought by context-aware verifier is reduced. A similar tendency can also be observed on Tone2 category. Lastly, BLSTM-based verifiers achieve comparable Tone1 and Tone4 mispronunciation detection performance with DNN-based systems. This result indicates that tone features themselves have already contained good discriminant information, and long-term dependencies learned by BLSTM-based verifiers are no longer needed for verifying Tone1 and Tone4's correctness. It should be noted that training BLSTM-based mispronunciation detection systems is more challenging than training their DNN-based counterparts, especially with regard to the imperfect human labeling (correct/mispronounced) and inaccurate input tone features.

Finally, Table 5.10 summarizes Mandarin tones' averaged EERs of different verifiers with different input tone features extracted from the DNN-based tone model trained with hard targets and BLSTM-based tone models trained with hard and soft targets. The EER reductions in each column demonstrate that our proposed soft target trained BLSTM-based tone model can produce better tone feature for verification, thus each verifier utilizing this feature achieves its own best performance. Moreover, the EER reductions in each row demonstrate that the long-range dependencies learned by the BLSTM-based verifier are beneficial to tone verification. Lastly, combining the BLSTM-based tone model and verifier achieves the lowest EER, as shown in the bottom right corner of Table 5.10.

5.5 Summary

In this chapter, through a series of systematic experiments, we have shown that performance of Mandarin tone mispronunciation detection can be improved by properly choosing the DNN architecture and adjusting the training scheme to better match the inherent features/peculiarities of the spoken languages at hand, namely: (i) Effect of long-term acoustic and prosodic dependencies in non-native tone realizations, and (ii) effect of imprecise tonal pronunciation of L2 learners. In contrast to the fixed input context window of DNN-based tone modeling, the BLSTM-based neural architecture captured tone-level coarticulation and prolonged tone productions with long-range dependencies of acoustic and prosodic features and led to a significant EER reduction for Tone2 and Tone3. Subsequently, adding soft target training further reduces EERs (Fig. 5.5).

Finally, memory-based verifiers are investigated to model dynamic changes of frame-level posteriors. This sequential information makes the proposed BLSTM-based verifiers achieve the lowest EERs. Figs. 5.11 and 5.5 show that captured sequential information can make up for the inaccuracy of extracted tone features, especially for Tone3. That is, when tone features are extracted from DNN-based models, Tone3' EER of 13.1% in Fig. 5.5 is reduced to 11.3% (blue curve in the lower left panel of Fig. 5.11), comparable to the DNN-based verifier trained with features extracted from the BLSTM-based acoustic model (orange curve in the lower left panel of Fig. 5.5) where an EER of 11.3% is achieved for Tone3.

CHAPTER 6

DIAGNOSING NON-NATIVE PHONETIC MISPRONUNCIATIONS AND PROVIDING ARTICULATORY-LEVEL FEEDBACK WITH KNOWLEDGE-GUIDED AND DATA-DRIVEN BASED DECISION TREES

In this chapter, we propose a novel decision tree based framework to diagnose phonetic mispronunciations produced by L2 learners caused by using inaccurate speech attributes, such as manner and place of articulation. Compared with conventional score-based CAPT (computer assisted pronunciation training) systems, our proposed framework has three advantages: (1) each mispronunciation in a tree can be interpreted and communicated to the L2 learners by traversing the corresponding path from a leaf node to the root node; (2) corrective feedback based on speech attribute features, which are directly used to describe how consonants are produced using related articulators, can be provided to the L2 learners; and (3) by building the phone-dependent decision tree, the relative importance of the speech attribute features of a target phone can be automatically learned and used to distinguish it-self from other phones. This information can provide L2 learners speech attribute feedback that is ranked in order of importance. In addition to the abovementioned advantages, experimental results confirm that the proposed approach can detect most pronunciation errors and provide accurate diagnostic feedback.

6.1 Introduction

As has been discussed in the previous chapters, automatic speech recognition (ASR) systems can be used in the CAPT module to define ad-hoc confidence scores provided to the end learners. Meanwhile, these confidence scores can be also fed into binary verifiers to generate posterior probabilities to be used as a measure of how likely a current phone/tone is correctly pronounced. Indeed, it has been demonstrated that a significant improvement

in the pronunciation training can be achieved by providing simple pronunciation scores to the L2 learners, e.g., [80]. However, when facing lower confidence scores, L2 learners might have difficulty in self-correction, because they do not know what is wrong with their pronunciations and how to improve them with only numeric scores.

In [113], it was shown that L2 learners can improve their production of the targeted phones by receiving the feedback about the mispronunciation error at phone level. Nowadays, more and more research work has thus focused on how to use automatic mechanisms to generate finer detection results and corrective information. For example, an extended recognition network (ERN) [72, 73, 74] containing canonical phones and frequent erroneous patterns has been proposed to provide diagnostic feedback related to phone substitutions, i.e., phone /A/ is mispronounced as phone /B/. Nonetheless, a major assumption made by providing a feedback at a segment (a.k.a. phone) level is that learners are aware of which articulatory movements (e.g., manner and place of articulation [107, 108]) have to be corrected in order to restore the canonical phone pronunciation. Unfortunately, that is a challenging task for L2 beginners. For example, some phones in L2 are absent in L1: beginning Japanese learners of English might pronounce “lice” instead of “rice”. Facing the segmental level feedback “phone /R/ is mispronounced as phone /L/”, learners might fail to adjust their articulatory movements to correct this error, because the phone /R/ does not exist in Japanese. Moreover, even if the supposed target phone exists in the learners’ L1, the acoustic realizations of it might differ from those of the target language. For example, Mandarin unvoiced phone /B/ shares the same articulation manner and place as its English counterpart, but English phone /B/ is voiced. Finally, in [33, 34, 35], researchers pointed out that non-native pronunciations have many “distortion errors”, i.e., the erroneous sound is always between two canonical phones, rather than an absolute phoneme substitution. Obviously, phone-level feedback is not sufficient enough to give direct instructions and deal with such distortion errors.

Therefore, exploiting information at an articulatory-level is expected to enhance the

quality of the feedback and alleviate some of the problems mentioned above. For example, the Japanese learners could be instructed to correct their mispronunciation if they are given following feedback “make a sound similar to /L/ but roll your tongue more backwards to create the acoustic characteristics of /R/.” Indeed, it has been reported that L2 learners prefer to receive direct instruction on how to correct mispronunciation at a articulatory level [101]. Moreover, researchers have exploited articulatory information for L2 learning [83, 101, 102], where an acoustic-to-articulatory inversion method is adopted to directly provide feedback at an articulatory level, e.g., tongue position. In [67, 103, 104], rule-based acoustic-articulatory mapping tables were employed to overcome the difficulty of collecting physical articulatory measurements to map each phone to its corresponding articulators. However, past work in mispronunciation detection performance at the articulatory level have been suboptimal due to the use of shallow models, or the lack of large training non-native corpora [67]. In chapter 3, with the help of large non-native corpus and deep learning techniques, we have achieved a good speech attribute classification accuracy, and it allows us to directly measure pronunciation quality and give corrective feedback based on articulation manner and place. However, the decision boundary of each speech attribute feature in there is optimized for its own classification purpose, which is not directly related to phone-level mispronunciation detection. In this chapter, we aim to optimize the decision boundaries of each attribute feature for different target phones. Moreover, in order to inspect how inaccurate speech attribute features can lead to mispronunciations, a white-box classifier with interpretable output is investigated, instead of relying only on a fully black-box approach, e.g., the DNN or BLSTM described in the previous chapters.

In this study, two types of decision trees [126] trained with speech attribute features are used to discriminate articulatory characteristics of correct and incorrect phone-level pronunciations. The first type is a knowledge-guided decision tree in which the input uses only speech attribute features that belong to the target phone. The other is a data-driven decision tree, which is built by automatically selected “optimal” speech attribute features.

Both decision trees are “readable” models that allows each phone-level mispronunciation to become interpretable by traversing the corresponding path from a leaf node to the root node. We can thus find and analyze how speech attribute features result in mispronunciations. Subsequently, pertinent articulatory-level feedback could be formulated to help the L2 learners improving their pronunciations. Finally, through constructing the decision trees, we can automatically learn which speech attributes are more important for distinguishing one phone from others. This information helps rank the corrective feedback by importance. Detailed analysis and examples of the abovementioned decision trees will be given in our experimental section.

6.2 Overview of The Decision Tree Based Diagnosis Framework

Fig. 6.1 shows the training and testing procedure of our proposed decision tree based mispronunciation diagnosis. The training part consists of two blocks: (i) the speech attribute feature extraction module, and (ii) the phone-dependent mispronunciation detector training modules, including knowledge-guided and data-driven based decision trees. At the testing stage, non-native pronunciation is first fed into our multisource BLSTM-based mispronunciation detector described in chapter 4. If some mispronunciation is detected, well-trained decision trees will be used to analyze this non-native pronunciation and provide diagnostic feedback. The format of generated feedback depends on the feedback design module, where audio clips or articulatory animation video can be prepared in advance.

6.2.1 Speech Attribute Feature Extraction

The training of speech attribute classifiers have been introduced in Section 3.3. After the speech attribute classifiers are well trained, we can use them to extract articulatory segmental-level features, which are then fed into knowledge-guided or data-driven based decision trees. Please consult Section 3.4 for detailed description of extracting segmental-level features.

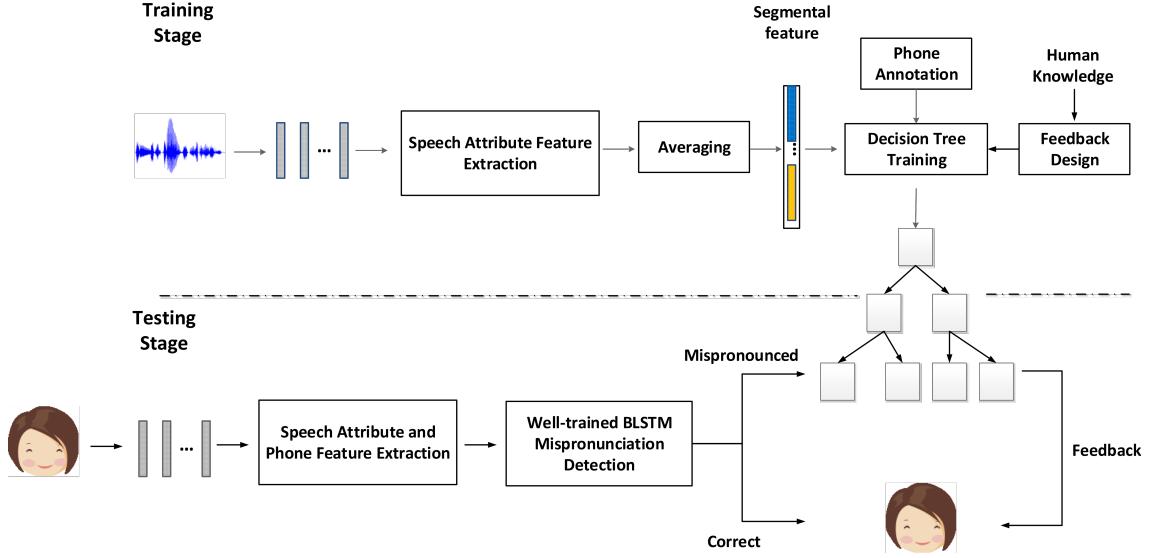


Figure 6.1: Overview of the decision tree based phone mispronunciation diagnosis framework.

6.2.2 Tree-based Mispronunciation Verifier Construction

We build a phone-dependent decision tree with the C4.5 algorithm [127] according to the annotated phone label (correct/mispronounced) and segmental-level pronunciation scores of the speech attributes. Each decision tree was constructed by using C4.5 algorithm selecting the speech attribute with the highest normalized information gain to split the current node set of samples into subsets. The resulting leaf nodes classify each phone segment into either the correct or incorrect (mispronounced) categories. With respect to the mispronounced category, we can traverse the corresponding path from the current leaf node to the root node to know how this mispronunciation has occurred. Since each non-leaf node of the decision tree is associated with one speech attribute and its corresponding splitting value, we can easily give quantitative and qualitative corrective feedback.

6.3 Experiments

Non-native phonetic mispronunciation diagnosis is carried out on the largest non-native Mandarin corpus iCALL [55]. Its data preparation is the same as the phone-level mispronunciation detection experiment described in chapter 4. In this study, we are concerned with diagnosing the top 15 most frequently mispronounced consonants, which cover around 97.8% consonant errors in our test set.

6.3.1 Experimental Setup

In this study, before constructing a phone-dependent decision tree implemented by WEKA [128], a data imbalance problem had to be addressed. For one target phone, the number of correct samples is much higher than that of the incorrect samples, leading to a biased decision tree with a high precision rate, but a low recall rate. Therefore, cost-sensitive objective function is used to balance the training samples. Although there are nearly 20 different speech attributes in our feature vector, the decision tree can automatically select the more important attributes for distinguishing the target phone from the others. This is called data-driven based decision tree (DDBDT). In addition to DDBDT, this study also investigates knowledge-guided based decision tree (KGBDT), where the feature vector only contains speech attributes related to the target phone. For example, if our target phone is /B/, only four attributes (labial, stop, unaspirated and unvoiced) are used to construct the feature vector, which is subsequently used for training decision tree.

In this experiment, the diagnostic error rate (DER) introduced in Section 2.3 is used to evaluate the performance of aforementioned decision trees. Specifically, at the optimal operation point (e.g., precision is equal to recall) shown in Fig. 4.2, the mispronunciation samples correctly detected by the BLSTM-phone-attribute verifier are collected. Then the DER is used to calculate the percentage of incorrect feedback generated by our knowledge-guided or data-driven based decision trees.

Table 6.1: The individual phone's diagnostic error rate (DER) of two different decision trees on test set, where the overall DERs of KGBDT and DDBDT are 4.1% and 4.6%, respectively

Phone	DER(%)	
	KGBDT	DDBDT
B	0.0	5.5
CH	2.3	4.7
C	3.5	3.5
D	5.0	5.0
J	5.2	5.2
K	4.7	4.7
P	7.1	7.1
Q	8.1	8.1
R	0.0	16.6
SH	1.8	1.8
S	0.0	0.0
T	1.8	3.6
X	4.5	4.5
ZH	1.5	1.5
Z	10.6	6.3

6.3.2 Experimental Results and Discussions

The KGBDT and DDBDT are separately constructed for each phone category. The DERs of different phones are summarized in Table 6.1. From this table, we can see that both the DDBDT and KGBDT can provide accurate diagnostic feedback, e.g., their overall DERs are both below 5%. Some reasons might lead to this observed phenomenon. First, the input speech attribute features themselves already have very good classification accuracy, as shown in chapter 3. Moreover, the generated feedback is at articulatory-level, where the number of units within each articulatory category is far less than the size of phone set. Therefore, it's comparatively easy to get the correct prediction/feedback. In order to offer an insight into how a specific decision tree is used to diagnose mispronunciations and give corrective feedback, we first give two phone-dependent KGBDT examples, as shown in Fig. 6.2.

A 10-fold cross validation is conducted to prune the decision tree to avoid overfitting

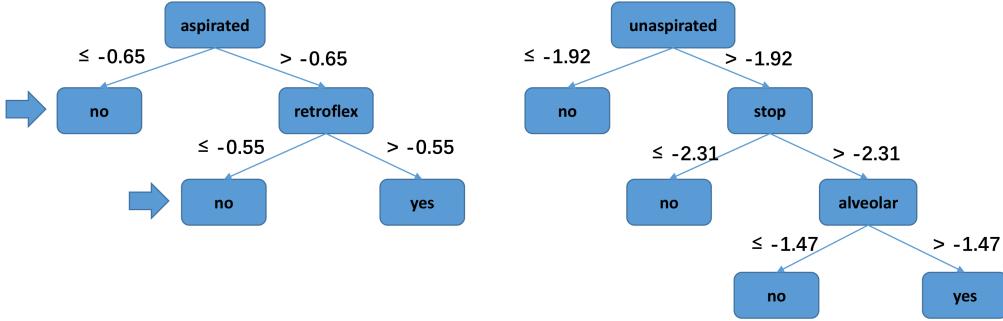


Figure 6.2: The KGBDT of phone /CH/ (left) and phone /D/ (right)

and unreasonable splitting so that only the most important speech attributes are selected. Take phone /CH/ in Fig. 6.2 for example, only “aspirated” and “retroflex” are used to build the knowledge-guided decision tree. This observation is consistent with Table 3.1, where the intersection between “aspirated” and “retroflex” sets only contains the phone /CH/. Moreover, we can see that the higher a node is, the more important the speech attribute associated with this node is for distinguishing /CH/ from other phones. Therefore, the speech attribute priority for phone /CH/ is: aspirated (aspiration) > retroflex (place). Facing phone /CH/ mispronunciation caused by a combination of inaccurate aspiration and place, aspiration related feedback should be given priority.

Each non-leaf node in the decision tree is associated with one speech attribute and its corresponding splitting value. Thus, how each input feature leads to a mispronunciation can be quantitatively and qualitatively analyzed. Take phone /D/ in Fig. 6.2 for example, the decision tree first checks the pronunciation score of unaspirated. If the score is smaller than -1.92 (this splitting value or decision boundary is automatically learned by using C4.5 algorithm and optimized for each target phone), we can conclude that too much aspiration contributes to one poor pronunciation. Obviously, the speech attribute associated with each tree node gives a qualitative description about which speech attribute results in a mispronunciation. Moreover, the pronunciation score of each speech attribute itself gives a quantitative description of the mispronunciation tendencies, i.e., the lower the attribute score is, the more likely this attribute is not well pronounced.

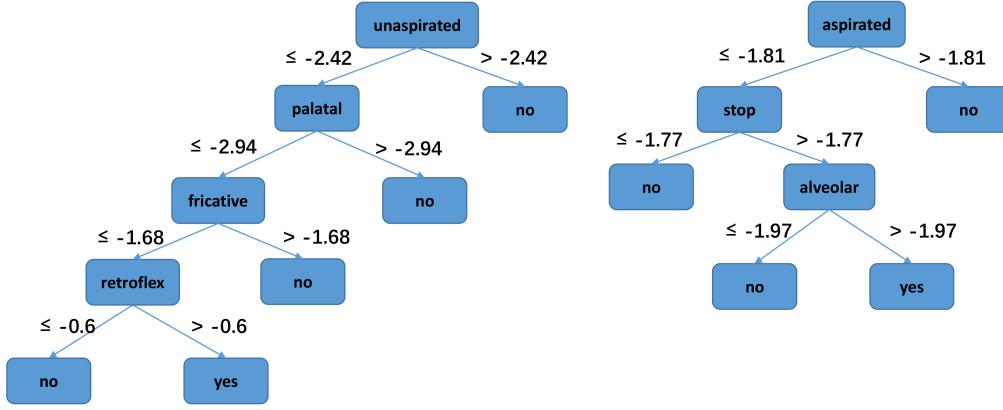


Figure 6.3: The DDBDT of phone /CH/ (left) and phone /D/ (right)

By traversing a corresponding path from a leaf node to the root node, each mispronunciation can be interpreted, i.e., the reason why this mispronunciation occurs and how to correct it can be communicated to the L2 learner. KGBDT classifies the phone segment as an incorrect category when target phone-related attributes' pronunciation scores are small, which shows that expected articulatory manner or place is not observed in the current phone segment. In Fig. 6.2, there are two possible error patterns, marked by arrows. The upper arrow points to phone /CH/ being mispronounced due to a lack of aspiration. Consequently, a feedback on teaching how to practice aspiration can be given. The second arrow shows that phone /CH/ is mispronounced due to the articulation place: the retroflex attribute has a low score although its aspiration is correct. Therefore, articulation place-based feedback can be given, e.g., try to move your tongue tip backwards so that the edges of your tongue are touching your hard palate.

In order to compare and contrast the different mispronunciation detection mechanisms of the DDBDT and KGBDT, the aforementioned decision trees' DDBDT counterparts are shown in Fig. 6.3. Through comparing Figs. 6.2 and 6.3, we find that speech attributes, such as palatal and fricative, are used to construct the DDBDT for phone /CH/. These speech attributes do not belong to phone /CH/, as shown in Table 3.1. The DDBDT thus classifies the phone segment into the incorrect category when these attribute scores are high,

which indicates that unexpected articulatory manner or place is observed in the current phone segment. This detection mechanisms is also observed in the DDBDT for phone /D/, where the attribute “aspirated” does not belong to phone /D/. The full report of each phone’s KGBDT and DDBDT are shown in appendix A.

6.4 Summary

In this chapter, speech attribute based decision trees are proposed to detect phonetic mispronunciations and provide articulatory level feedback based on the manner and place of articulation. Compared with conventional score-based and phone-based systems, our approach can tell the L2 learners why their mispronunciations occur and how to correct them using appropriate methods. While giving more intuitive and instructive feedback, our system also achieves a good diagnostic performance. Indeed, mispronunciation detection at an articulatory level can more accurately specify systematic L2 pronunciation errors [67]. For example, [55] reported that non-native Mandarin learners with Romance mother language are more likely to mispronounce aspirated phone /Q/ as unaspirated /J/. This seems like a simple substitution error, but the fact that aspiration does not exist in Romance languages might be the underlying reason for this de-aspiration error. Therefore, the feedback on teaching how to practice aspiration is more instructive than just telling non-native learners you have a phone-level substitution error.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The objective of this thesis is to improve mispronunciation detection of Mandarin and enrich diagnostic feedback for second language learners. In this chapter, we summarize our contributions and discuss some directions for future work.

7.1 Summary and Contributions

In this thesis, we show our efforts to tackle the three challenges laid out in the Introduction chapter. The first challenge is the inconsistency in non-native phone-based labeling, which makes trained phone model imperfect and its generated pronunciation scores still have room for improvement. Therefore, we investigate using speech attribute features to enhance the quality of original phone features. This idea is inspired by the sharing mechanism of the speech attribute, i.e., each speech attribute feature is sharable among a group of phones, which allows it to pool more training data than an individual phonetic category. Therefore, speech attributes are not too sensitive to individual phone-level labeling errors and could be more robustly trained. Our experiment results shown in chapter 3 demonstrate the efficiency of our preliminary thought, namely that higher classification accuracy is achieved for speech attribute classifiers, when compared to traditional phone classifier. It allows us to use speech attribute features to visualize and analyze non-native pronunciations in addition to conventional phone features. More importantly, mispronunciation detector trained with phone-attribute based features can bring consistent improvement over our baseline trained with only phone features.

Similar to phonetic pronunciations, non-native tone productions often fall between two canonical tone categories or do not resemble any canonical tonal categories. Consequently, the use of forced-assigned standard tone categories will result in the inconsistency in tone-

based labeling. Therefore, we propose to use soft targets, a kind of probabilistic transcriptions, to label aforementioned non-native tone productions. After introducing soft targets and Kullback-Leibler (KL) divergence training, our proposed system can achieve substantial relative equal error rate (EER) reduction from its counterpart trained with hard targets.

The second challenge is related to how to more accurately verify whether current phone/tone is mispronounced. In this study, we observe that the frame-level detection curves exhibit dynamic changes within each phone or tone segment. Traditional frame-level averaging used in DNN-based verifiers might lose this observed sequential information. Therefore, BLSTM is proposed to employ learned context information to embed a sequence of frame-level pronunciation scores into a pronunciation score vector. Compared with the pronunciation representation derived from simple frame-level averaging, the proposed BLSTM-based embedded pronunciation score vector is expected to characterize more context information for subsequent mispronunciation detection, which is especially useful for verifying the pronunciation correctness of Tone2 and Tone3 in our tone experiments. Meanwhile, we also observe that the BLSTM-based phone mispronunciation detectors can achieve a higher equal precision-recall rate than its DNN-based counterparts. These improvements give us some insight on modeling non-native pronunciations. Namely, due to being unfamiliar with the target language, the L2 learners' speech production process might not be stable enough, which results in its detection curves exhibit dynamic changes. Therefore, memory-based verifiers are more suitable for mispronunciation detection.

Lastly, we try to use articulatory information to tackle the challenge of generating more intuitive and instructive feedback for non-native learners. Specifically, decision trees trained with speech attribute features are used to analyze how speech attribute features result in phonetic mispronunciations. Subsequently, pertinent articulatory-level feedback could be formulated to help the L2 learners improve their pronunciations. Unlike traditional pronunciation scores or phone-level feedback, our generated feedback can directly tell the user which articulatory movements (e.g., manner and place of articulation) have

to be corrected in order to restore the pronunciation of the canonical phone. Moreover, the experimental results show that our proposed decision trees achieve a good diagnostic performance. Different from articulatory-level explanations for phonetic errors, pitch contour and height are frequently used for generating feedback for mispronounced tones. These speech cues have already been readable enough, therefore their optimization is not our focus in this thesis.

7.2 Future Work and Directions

Several future research directions arise from the work presented in this thesis. We here give a brief discussion of them.

Mispronunciation visualization and detection of Mandarin vowels using speech attribute features: In this thesis, we have demonstrated the feasibility of using articulatory-level posteriors to analyze non-native consonant productions. Obviously, we can further extend this framework to predict Mandarin vowels' speech attributes, e.g., the position of tongue, the shape of lip.

Language universal mispronunciation detection: Some work [36, 129] in the ASAT project have demonstrated the efficiency of cross-language speech attribute detection. Therefore a direct extension of the current work is to apply speech attribute detectors trained on Mandarin corpus to detect some mispronunciations made by other users, e.g., non-native learners of English.

Automatic error pattern discovery: With the help of a large non-native corpus and deep learning techniques, posteriors generated from well-trained phone, articulatory, and tone classifiers are proposed to analyze non-native productions. Naturally, unsupervised clustering on top of these posteriors is expected to find new error patterns, in addition to traditional substitution errors. After modeling the discovered patterns, we can enhance the diagnostic ability of the current CAPT system.

Appendices

APPENDIX A

PHONE-DEPENDENT DECISION TREES

In this study, we generate phone-dependent decision trees for each of the top 15 most frequently mispronounced consonants in Mandarin. Their corresponding knowledge-guided based decision tree (KGBDT) and data-driven based decision tree (DDBDT) are respectively shown in the left and right part of the following figures.

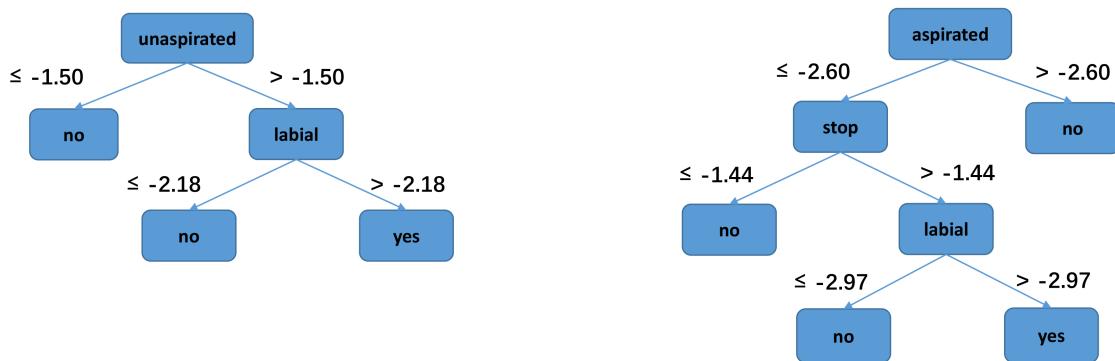


Figure A.1: The KGBDT (left) and DDBDT (right) of phone /B/

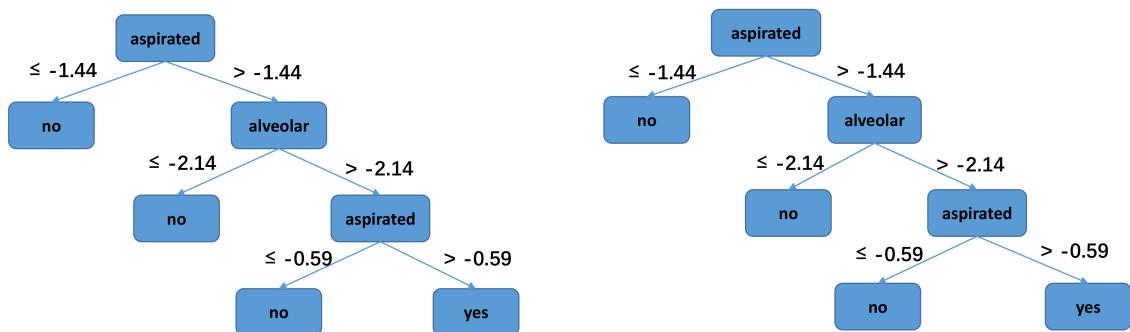


Figure A.2: The KGBDT (left) and DDBDT (right) of phone /C/

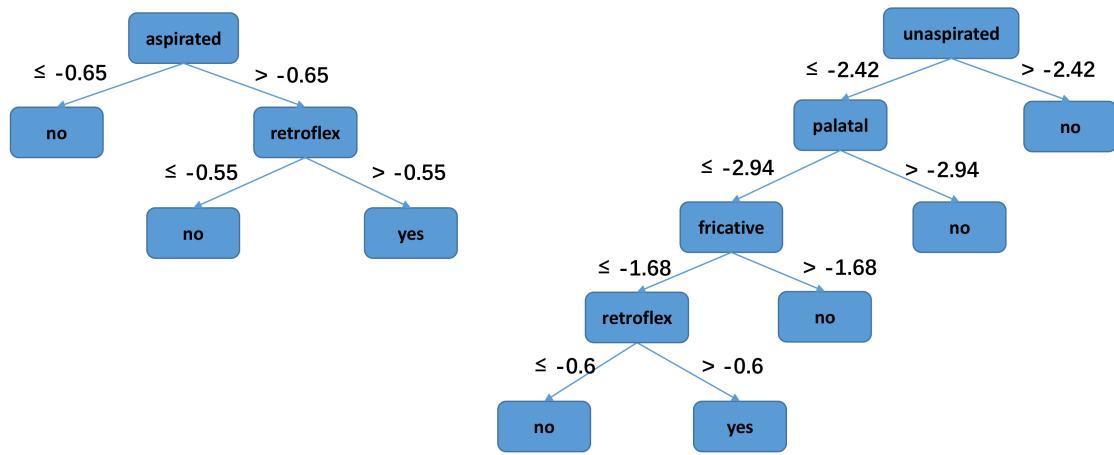


Figure A.3: The KGBDT (left) and DDBDT (right) of phone /CH/

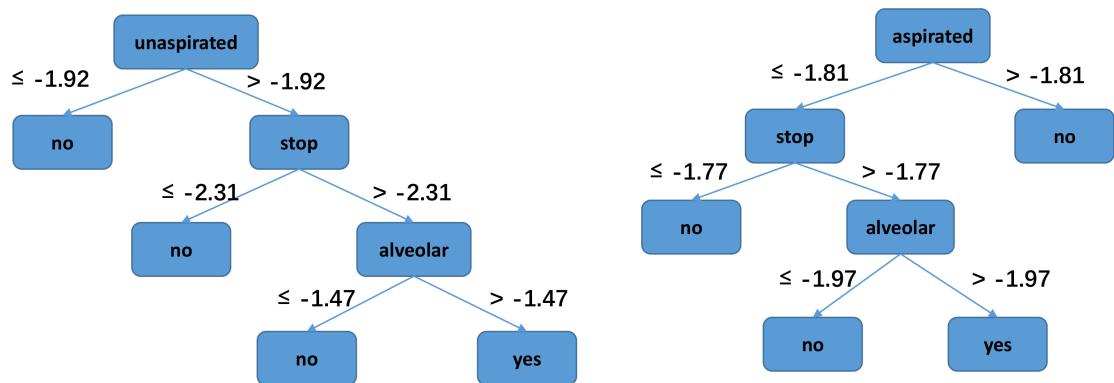


Figure A.4: The KGBDT (left) and DDBDT (right) of phone /D/

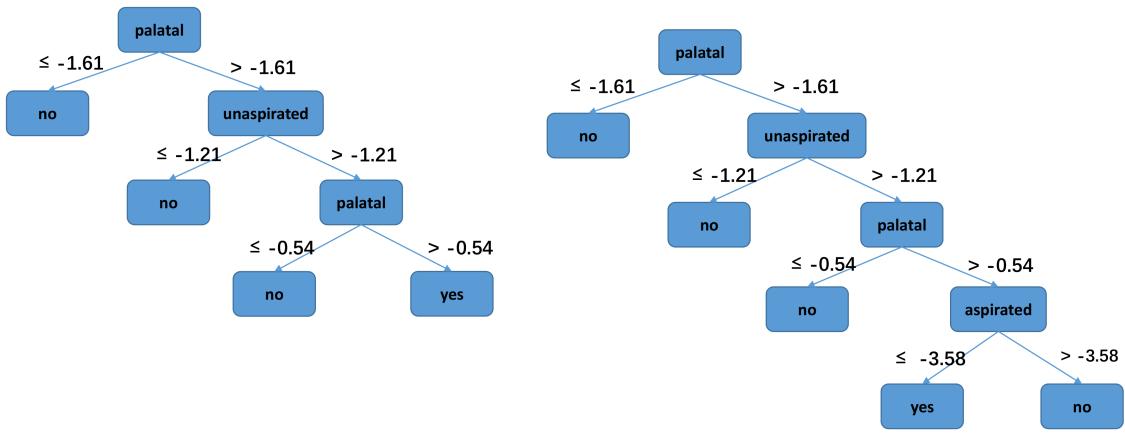


Figure A.5: The KGBDT (left) and DDBDT (right) of phone /J/

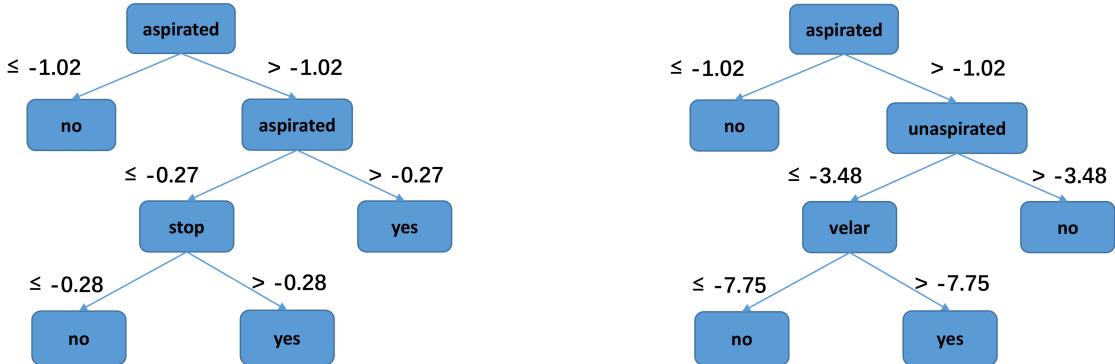


Figure A.6: The KGBDT (left) and DDBDT (right) of phone /K/



Figure A.7: The KGBDT (left) and DDBDT (right) of phone /P/

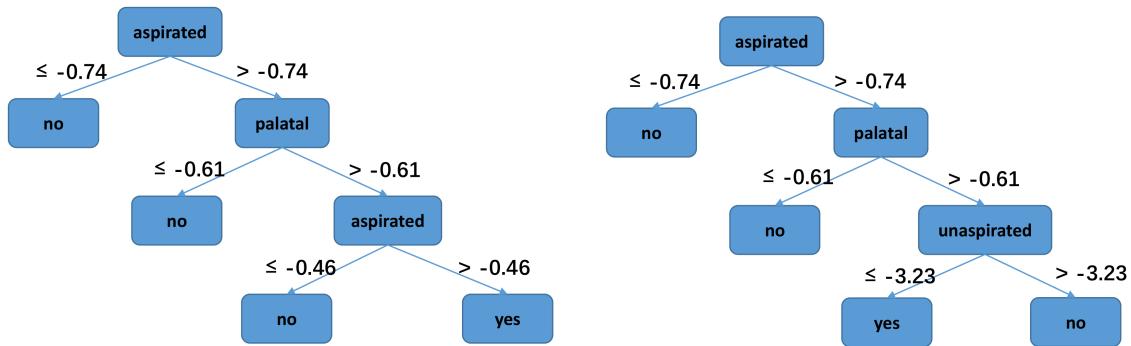


Figure A.8: The KGBDT (left) and DDBDT (right) of phone /Q/



Figure A.9: The KGBDT (left) and DDBDT (right) of phone /R/

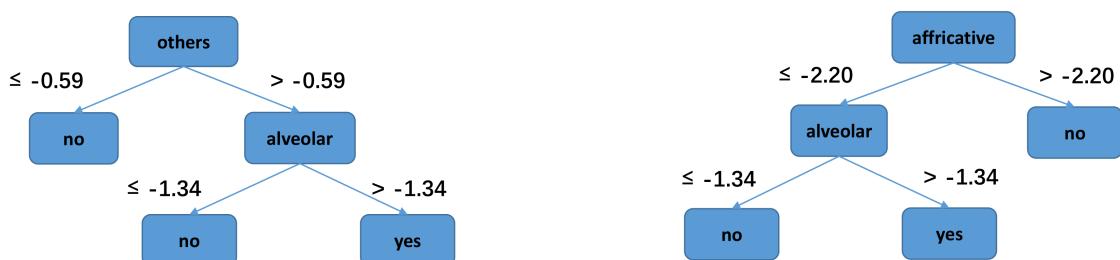


Figure A.10: The KGBDT (left) and DDBDT (right) of phone /S/

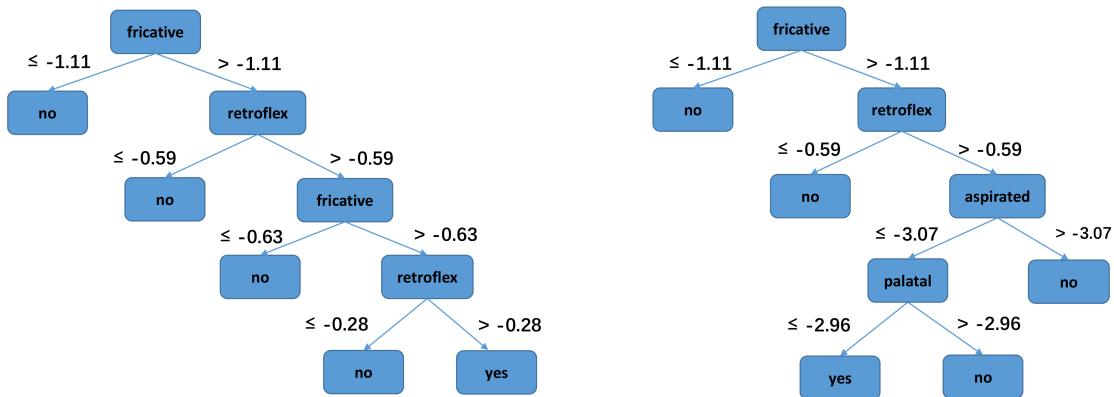


Figure A.11: The KGBDT (left) and DDBDT (right) of phone /SH/

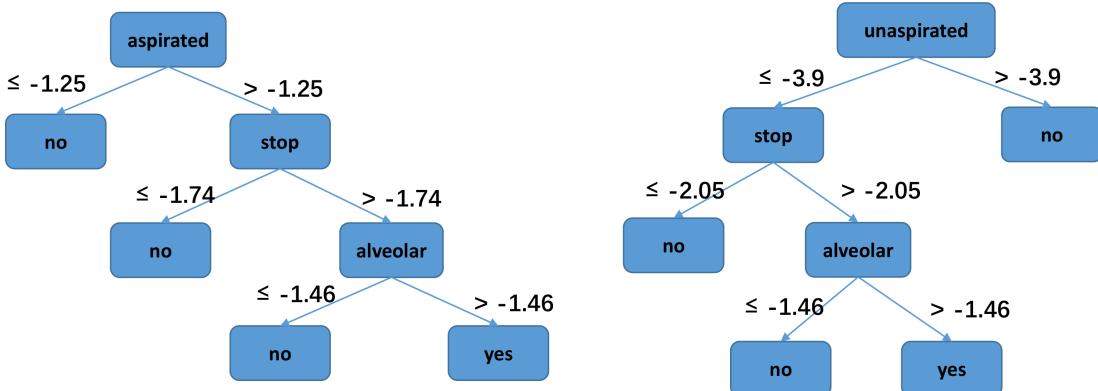


Figure A.12: The KGBDT (left) and DDBDT (right) of phone /T/

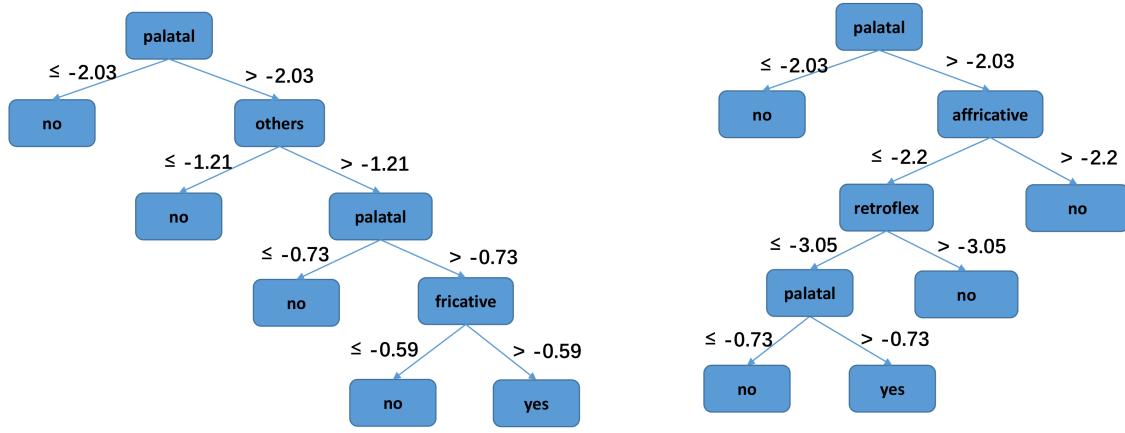


Figure A.13: The KGBDT (left) and DDBDT (right) of phone /X/

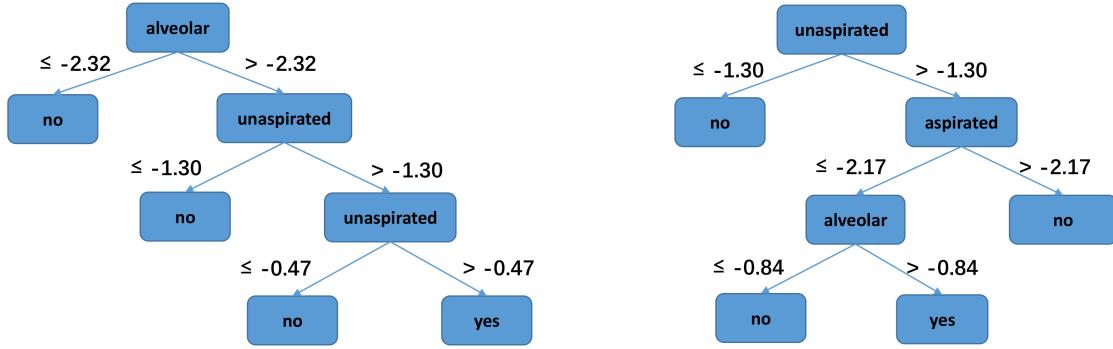


Figure A.14: The KGBDT (left) and DDBDT (right) of phone /Z/

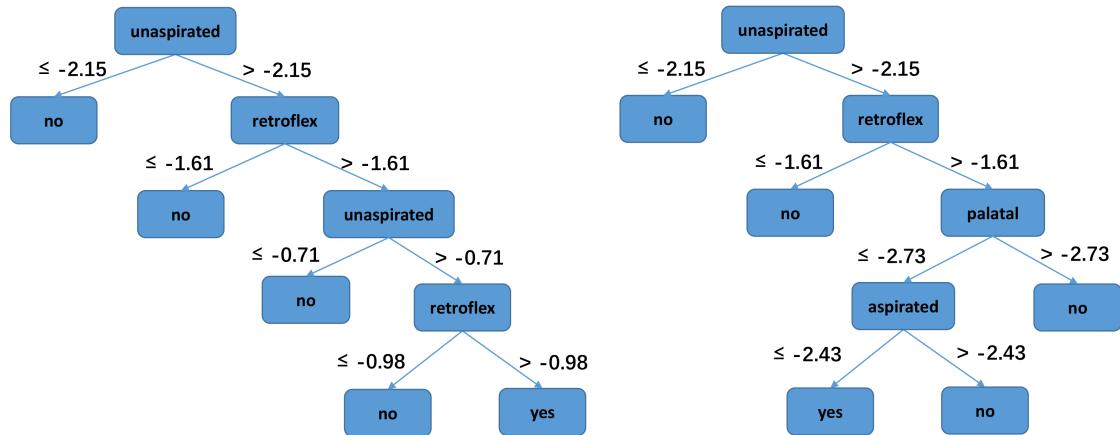


Figure A.15: The KGBDT (left) and DDBDT (right) of phone /ZH/

REFERENCES

- [1] *China sees number of teachers grow*, http://news.xinhuanet.com/english/china/2013-09/09/c_132705547.htm. Accessed : 2014-11-20, 2014.
- [2] *Chinese-as-a-second-language-growing-in-popularity*, <http://www.cctv-america.com/2015/03/03/chinese-as-a-second-language-growing-in-popularity>, 2015, 2015.
- [3] M. D. Swaine, “Chinese views and commentary on the ‘one belt, one road’initiative,” *China Leadership Monitor*, vol. 47, no. 2, p. 3, 2015.
- [4] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling,” in *Proc. ICASSP*, 2016, pp. 6135–6139.
- [5] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers,” *Speech Communication*, vol. 67, pp. 154–166, 2015.
- [6] K. Li, X. Qian, and H. Meng, “Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.
- [7] H. Huang, H. Xu, X. Wang, and W. Silamu, “Maximum f1-score discriminative training criterion for automatic mispronunciation detection,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 4, pp. 787–797, 2015.
- [8] Y. Wang and L. Lee, “Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 3, pp. 564–579, 2015.
- [9] X. Qian, H. Meng, and F. K. Soong, “A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 6, pp. 1020–1028, 2016.
- [10] L. Chen and J. R. Jang, “Automatic pronunciation scoring with score combination by learning to rank and class-normalized dp-based quantization,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 11, pp. 1737–1749, 2015.

- [11] R. Tong, N. F. Chen, B. Ma, and H. Li, “Context aware mispronunciation detection for mandarin pronunciation training,” in *Proc. INTERSPEECH*, 2016, pp. 3112–3116.
- [12] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, “Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning,” *IEICE Transactions*, vol. 100-D, no. 9, pp. 2174–2182, 2017.
- [13] W. Li, N. F. Chen, S. M. Siniscalchi, and C. Lee, “Improving mispronunciation detection for non-native learners with multisource information and lstm-based deep models,” in *Proc. INTERSPEECH*, 2017, pp. 2759–2763.
- [14] A. Lee, N. F. Chen, and J. Glass, “Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery,” in *Proc. ICASSP*, 2016, pp. 6145–6149.
- [15] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, “Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system,” in *Proc. INTERSPEECH*, 2002.
- [16] K. Li, X. Qian, S. Kang, and H. Meng, “Lexical stress detection for L2 english speech using deep belief networks,” in *Proc. INTERSPEECH*, 2013, pp. 1811–1815.
- [17] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, “Automatic lexical stress and pitch accent detection for L2 english speech using multi-distribution deep neural networks,” *Speech Communication*, vol. 96, pp. 28–36, 2018.
- [18] J. Tepperman and S. Narayanan, “Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners,” in *Proc. ICASSP*, 2005, pp. 937–940.
- [19] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, “Using tone-based extended recognition network to detect non-native mandarin tone mispronunciations,” in *Proc. APSIPA*, 2016, pp. 1–4.
- [20] J. Lin, W. Li, Y. Gao, Y. Xie, N. F. Chen, S. M. Siniscalchi, J. Zhang, and C. Lee, “Improving mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks,” *Signal Processing Systems*, vol. 90, no. 7, pp. 1077–1087, 2018.
- [21] L. Zhang, C. Huang, M. Chu, F. K. Soong, X. Zhang, and Y. Chen, “Automatic detection of tone mispronunciation in mandarin,” in *Proc. ISCSLP*, 2006, pp. 590–601.

- [22] J. Cheng, “Automatic tone assessment of non-native mandarin speakers,” in *Proc. INTERSPEECH*, 2012, pp. 1299–1302.
- [23] R. Tong, N. F. Chen, B. P. Lim, B. Ma, and H. Li, “Tokenizing fundamental frequency variation for mandarin tone error detection,” in *Proc. ICASSP*, 2015, pp. 5361–5365.
- [24] Y. Zhang, M. Chu, C. Huang, and M. Liang, “Detecting tone errors in continuous mandarin speech,” in *Proc. ICASSP*, 2008, pp. 5065–5068.
- [25] S. Wei, H. Wang, Q. Liu, and R. Wang, “Cdf-matching for automatic tone error detection in mandarin call system,” in *Proc. ICASSP*, 2007, pp. 205–208.
- [26] J. Lin, Y. Xie, and J. Zhang, “Automatic pronunciation evaluation of non-native mandarin tone by using multi-level confidence measures,” in *Proc. INTERSPEECH*, 2016, pp. 2666–2670.
- [27] H. Liao, J. Chen, S. Chang, Y. Guan, and C. Lee, “Decision tree based tone modeling with corrective feedbacks for automatic mandarin tone assessment,” in *Proc. INTERSPEECH*, 2010, pp. 602–605.
- [28] R. Tong, N. F. Chen, B. Ma, and H. Li, “Goodness of tone (GOT) for non-native mandarin tone recognition,” in *Proc. INTERSPEECH*, 2015, pp. 801–805.
- [29] W. Hu, Y. Qian, and F. K. Soong, “A dnn-based acoustic modeling of tonal language and its application to mandarin pronunciation training,” in *Proc. ICASSP*, 2014, pp. 3206–3210.
- [30] W. Li, N. F. Chen, S. M. Siniscalchi, and C. Lee, “Improving mandarin tone mispronunciation detection for non-native learners with soft-target tone labels and blstm-based deep models,” in *Proc. ICASSP*, 2018, pp. 6249–6253.
- [31] T. N. A. Ito and H. Ogasawara, “Automatic detection of english mispronunciation using speaker adaptation and automatic assessment of english intonation and rhythm,” *Educational technology research*, vol. 29, no. 1-2, pp. 13–23, 2006.
- [32] K. Li, X. Wu, and H. Meng, “Intonation classification for L2 english speech using multi-distribution deep neural networks,” *Computer Speech & Language*, vol. 43, pp. 18–33, 2017.
- [33] S. Yoon, M. Hasegawa-Johnson, and R. Sproat, “Landmark-based automated pronunciation error detection,” in *Proc. INTERSPEECH*, 2010, pp. 614–617.

- [34] R. Duan, J. Zhang, W. Cao, and Y. Xie, “A preliminary study on asr-based detection of chinese mispronunciation by japanese learners,” in *Proc. INTERSPEECH*, 2014, pp. 1478–1481.
- [35] W. Cao, D. Wang, J. Zhang, and Z. Xiong, “Developing a chinese L2 speech database of japanese learners with narrow-phonetic labels for computer assisted pronunciation training,” in *Proc. INTERSPEECH*, 2010, pp. 1922–1925.
- [36] C. Lee and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.
- [37] M. Peabody and S. Stephanie, “Annotation and features of non-native mandarin tone quality,” in *Proc. INTERSPEECH*, 2009.
- [38] C. Yang, “The acquisition of mandarin prosody by american learners of chinese as a foreign language (cfl),” PhD thesis, The Ohio State University, 2011.
- [39] W. Li, N. F. Chen, S. M. Siniscalchi, and C. Lee, “Improving mandarin tone mispronunciation detection for non-native learners with soft-target tone labels and blstm-based deep models,” *IEEE/ACM Trans. Audio, Speech & Language Processing (accepted)*, 2019.
- [40] J. Zhang and K. Hirose, “Tone nucleus modeling for chinese lexical tone recognition,” *Speech Communication*, vol. 42, no. 3-4, pp. 447–466, 2004.
- [41] M. C. D.W. Massaro and C. Tseng, “The evaluation and integration of pitch height and pitch contour in lexical tone perception in mandarin chinese,” *Journal of Chinese Linguistics*, pp. 267–289, 1985.
- [42] C. Wang, “Prosodic modeling for improved speech recognition and understanding,” PhD thesis, Massachusetts Institute of Technology, 2001.
- [43] J. Li, R. Zhao, J. Huang, and Y. Gong, “Learning small-size DNN with output-distribution-based criteria,” in *Proc. INTERSPEECH*, 2014, pp. 1910–1914.
- [44] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *CoRR*, vol. abs/1503.02531, 2015.
- [45] W. Chan, N. Ke, and I. Lane, “Transferring knowledge from a RNN to a DNN,” in *Proc. INTERSPEECH*, 2015, pp. 3264–3268.
- [46] J. Ba and R. Caruana, “Do deep nets really need to be deep?” In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information*

Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 2654–2662.

- [47] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.
- [48] R. A. L. S. Kullback, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [49] S. Wei, G. Hu, Y. Hu, and R. Wang, “A new method for mispronunciation detection using support vector machine based on pronunciation space models,” *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.
- [50] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [51] N. F. Chen and H. Li, “Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning,” in *Proc. APSIPA*, 2016, pp. 1–7.
- [52] S. M. Witt, “Automatic error detection in pronunciation training: Where we are and where we need to go,” *Proc. IS ADEPT*, vol. 6, 2012.
- [53] H. Meng, Y. Lo, L. Wang, and W. Lau, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *Proc. ASRU*, 2007, pp. 437–442.
- [54] H. Meng, “Developing speech recognition and synthesis technologies to support computer-aided pronunciation training for chinese learners of english,” in *Proc. PACLIC*, 2009, pp. 40–42.
- [55] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, “Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall,” *Speech Communication*, vol. 84, pp. 46–56, 2016.
- [56] M. A. Peabody, “Methods for pronunciation assessment in computer aided language learning,” PhD thesis, Massachusetts Institute of Technology, 2011.
- [57] D. Luo, X. Yang, and L. Wang, “Improvement of segmental mispronunciation detection with prior knowledge extracted from large l2 speech corpus,” in *Proc. INTERSPEECH*, 2011.

- [58] D. M. D. Kewley-Port C. Watson and D. Reed, “Speaker-dependent speech recognition as the basis for a speech training aid,” in *Proc. ICASSP*, vol. 12, 1987, pp. 372–375.
- [59] A. Lee and J. Glass, “A comparison-based approach to mispronunciation detection,” in *Proc. SLT*, 2012, pp. 382–387.
- [60] M. R. H. Franco L. Neumeyer and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,” in *Proc. INTERSPEECH*, 1999.
- [61] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.
- [62] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W. Ye, “Generalized segment posterior probability for automatic mandarin pronunciation evaluation,” in *Proc. ICASSP*, 2007, pp. 201–204.
- [63] A. Lee, Y. Zhang, and J. Glass, “Mispronunciation detection via dynamic time warping on deep belief network-based posteriograms,” in *Proc. ICASSP*, 2013, pp. 8227–8231.
- [64] Y. Hsu, M. Yang, H. Hung, and B. Chen, “Mispronunciation detection leveraging maximum performance criterion training of acoustic models and decision functions,” in *Proc. INTERSPEECH*, 2016, pp. 2646–2650.
- [65] H. Li, S. Wang, J. Liang, S. Huang, and B. Xu, “High performance automatic mispronunciation detection method based on neural network and TRAP features,” in *Proc. INTERSPEECH*, 2009, pp. 1911–1914.
- [66] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, “Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees,” in *Proc. INTERSPEECH*, 2016, pp. 3127–3131.
- [67] J. Tepperman and S. Narayanan, “Using articulatory representations to detect segmental errors in nonnative pronunciation,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 16, no. 1, pp. 8–22, 2008.
- [68] Y. Xie, M. Hasegawa-Johnson, L. Qu, and J. Zhang, “Landmark of mandarin nasal codas and its application in pronunciation error detection,” in *Proc. ICASSP*, 2016, pp. 5370–5374.
- [69] M. H. J. I. Amdal and E. Versvik, “Automatic evaluation of quantity contrast in non-native norwegian speech,” in *International Workshop on Speech and Language Technology in Education*, 2009.

- [70] F. D. W. H. Strik K. Truong and C. Cucchiarini, “Comparing different approaches for automatic pronunciation error detection,” *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.
- [71] S. Picard, G. Ananthakrishnan, P. O. Engwall, and S. M. Abdou, “Detection of specific mispronunciations using audiovisual features,” in *Proc. AVSP*, 2010, pp. 7–2.
- [72] H. Meng, Y. Lo, L. Wang, and W. Lau, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *Proc. ASRU*, 2007, pp. 437–442.
- [73] X. Qian, H. Meng, and F. K. Soong, “Capturing L2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (CAPT),” in *Proc. ISCSLP*, 2010, pp. 84–88.
- [74] W. Lo, S. Zhang, and H. Meng, “Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system,” in *Proc. INTERSPEECH*, 2010, pp. 765–768.
- [75] S. Mao, X. Li, K. Li, Z. Wu, X. Liu, and H. Meng, “Unsupervised discovery of an extended phoneme set in L2 english speech for mispronunciation detection and diagnosis,” in *Proc. ICASSP*, 2018, pp. 6244–6248.
- [76] M. S. N. Ryant and M. L. et al., “Highly accurate mandarin tone classification in the absence of pitch information,” in *Proc. Speech Prosody*, vol. 7, 2014.
- [77] J. Y. N. Ryant and M. Liberman, “Mandarin tone classification without pitch tracking,” in *Proc. ICASSP*, 2014, pp. 4868–4872.
- [78] C. Cao, Y. Xie, J. Lin, Q. Li, and J. Zhang, “The preliminary study of influence on tone perception from segments,” in *Proc. ISCSLP*, 2016, pp. 1–5.
- [79] C. Cao, Y. Xie, Q. Zhang, and J. Zhang, “The influence on realization and perception of lexical tones from affricate’s aspiration,” in *Proc. INTERSPEECH*, 2017, pp. 650–654.
- [80] A. Neri, C. Cucchiarini, and H. Strik, “Asr corrective feedback on pronunciation: Does it really work?,” 2006.
- [81] R. Ai, “Automatic pronunciation error detection and feedback generation for call applications,” in *International Conference on Learning and Collaboration Technologies*, Springer, 2015, pp. 175–186.

- [82] K. Yuen, W. Leung, and P. L. et al., “Enunciate: An internet-accessible computer-aided pronunciation training system and related user evaluations,” in *Proc. CO-COSDA*, IEEE, 2011, pp. 85–90.
- [83] O. Engwall, “Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher,” *Computer Assisted Language Learning*, vol. 25, no. 1, pp. 37–64, 2012.
- [84] D. M. Chun, “Teaching tone and intonation with microcomputers,” *CALICO Journal*, pp. 21–46, 1989.
- [85] D. M. Chun, Y. Jiang, and N. Ávila, “Visualization of tone for learning mandarin chinese,” in *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*, 2012, pp. 77–89.
- [86] M. Peabody and S. Seneff, “Towards automatic tone correction in non-native mandarin,” in *Proc. ISCSLP*, Springer, 2006, pp. 602–613.
- [87] A. Lee, “Language-independent methods for computer-assisted pronunciation training,” PhD thesis, Massachusetts Institute of Technology, 2016.
- [88] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,” *PloS one*, vol. 10, no. 3, e0118432, 2015.
- [89] S. Gao, B. Xu, H. Zhang, B. Zhao, C. Li, and T. Huang, “Update progress of sino-hear: Advanced mandarin lvcsr system at nlpr,” in *Proc. INTERSPEECH*, 2000, pp. 798–801.
- [90] D. Wang and X. Zhang, “Thchs-30: A free chinese speech corpus,” *arXiv preprint arXiv:1512.01882*, 2015.
- [91] J. Morris and E. Lussier, “Combining phonetic attributes using conditional random fields,” in *Proc. INTERSPEECH*, 2006.
- [92] J. Li and C. Lee, “On designing and evaluating speech event detectors,” in *Proc. INTERSPEECH*, 2005, pp. 3365–3368.
- [93] D. Yu, S. M. Siniscalchi, L. Deng, and C. Lee, “Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition,” in *Proc. ICASSP*, 2012, pp. 4169–4172.
- [94] J. Hou, “On the use of frame and segment-based methods for the detection and classification of speech sounds and features,” PhD thesis, Rutgers University-Graduate School-New Brunswick, 2009.

- [95] I. Chen, S. M. Siniscalchi, and C. Lee, “Attribute based lattice rescoring in spontaneous speech recognition,” in *Proc. ICASSP*, 2014, pp. 3325–3329.
- [96] S. M. Siniscalchi, J. Reed, T. Svendsen, and C. Lee, “Universal attribute characterization of spoken languages for automatic spoken language recognition,” *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.
- [97] S. M. Siniscalchi, T. Svendsen, and C. Lee, “A bottom-up modular search approach to large vocabulary continuous speech recognition,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 21, no. 4, pp. 786–797, 2013.
- [98] K. Kirchhoff, “Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments,” in *The 5th International Conference on Spoken Language Processing*, 1998.
- [99] F. Metze and A. Waibel, “A flexible stream architecture for ASR using articulatory features,” in *Proc. INTERSPEECH*, 2002.
- [100] A. Baker, *Ship or sheep? student’s book: an intermediate pronunciation course*. Ernst Klett Sprachen, 2006, vol. 1.
- [101] O. Bälter, O. Engwall, A. Öster, and H. Kjellström, “Wizard-of-oz test of artur: A computer-based speech training system with articulation correction,” in *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, ACM, 2005, pp. 36–43.
- [102] O. Engwall, “Pronunciation analysis by acoustic-to-articulatory feature inversion,” in *International Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012.
- [103] J. Tepperman and S. Narayanan, “Hidden-articulator markov models for pronunciation evaluation,” in *Proc. ASRU*, IEEE, 2005, pp. 174–179.
- [104] X. Xie and W. Abdulla, “Computer aided pronunciation learning systems,” PhD thesis, University of Auckland, 2010.
- [105] X. Wei, J. Chen, and W. W. et al., “A study of automatic annotation of pets with articulatory features,” in *Proc. APSIPA*, IEEE, 2017, pp. 1608–1612.
- [106] J. Zhang, W. Li, Y. Hou, W. Cao, and Z. Xiong, “A study on functional loads of phonetic contrasts under context based on mutual information of chinese text and phonemes,” in *Proc. ISCSLP*, 2010.
- [107] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

- [108] G. Fant, *Speech sounds and features*. MIT Press, 1973.
- [109] *Mandarin pinyin*, <https://en.wikipedia.org/wiki/Pinyin>.
- [110] S. Duanmu, *The phonology of standard Chinese*. Oxford University Press, 2007.
- [111] D. P. et al., “The kaldi speech recognition toolkit,” IEEE Signal Processing Society, Tech. Rep., 2011.
- [112] C. et al., “A preliminary study on corpus design for computer-assisted german and mandarin language learning,” in *Proc. COCOSDA*, 2009, pp. 154–159.
- [113] C. Wang, M. Peabody, S. Seneff, and J. Kim, “An interactive english pronunciation dictionary for korean learners,” in *Proc. INTERSPEECH*, 2004.
- [114] P. Liu, X. Qiu, and X. Huang, “Recurrent neural network for text classification with multi-task learning,” in *Proc. IJCAI*, 2016, pp. 2873–2879.
- [115] X. Wang, W. Jiang, and Z. Luo, “Combination of convolutional and recurrent neural network for sentiment analysis of short texts,” in *Proc. COLING*, 2016, pp. 2428–2437.
- [116] F. Chollet, “Keras: The python deep learning library,” *Astrophysics Source Code Library*, 2018.
- [117] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [118] D. Surendran and G. Levow, “The functional load of tone in mandarin is as high as that of vowels,” in *Proc. Speech Prosody*, 2004.
- [119] Y. Xu, “Production and perception of coarticulated tones,” *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2240–2253, 1994.
- [120] R. Tsai, “Teaching and learning the tones of mandarin chinese,” *Scottish Languages Review*, vol. 24, pp. 43–50, 2011.
- [121] Y. Hao, “Second language acquisition of mandarin chinese tones by tonal and non-tonal language speakers,” *Journal of phonetics*, vol. 40, no. 2, pp. 269–279, 2012.
- [122] Y. Wang, A. Jongman, and J. Sereno, “Acoustic and perceptual evaluation of mandarin tone productions before and after perceptual training,” *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1033–1043, 2003.

- [123] P. G. et al., “A pitch extraction algorithm tuned for automatic speech recognition,” in *Proc. ICASSP*, IEEE, 2014, pp. 2494–2498.
- [124] Y. Huang, Y. Wang, and Y. Gong, “Semi-supervised training in deep learning acoustic model.,” in *Proc. INTERSPEECH*, 2016, pp. 3848–3852.
- [125] M. Chen, *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press, 2000, vol. 92.
- [126] L. Breiman, *Classification and regression trees*. Routledge, 2017.
- [127] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, “The elements of statistical learning: Data mining, inference and prediction,” *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.
- [128] M. H. et al., “The weka data mining software: An update,” *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [129] S. M. Siniscalchi, D. C. Lyu, T. Svendsen, and C. Lee, “Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data,” *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 875–887, 2011.

VITA

Wei Li received his B.E. and M.S. degrees in 2011 and 2014 respectively, both major in computer science and technology from Beijing Language and Culture University, Beijing, China. Since 2014 fall, he has been working toward the Ph.D. degree under the supervision of Professor Chin-Hui Lee at the Georgia Institute of Technology, Atlanta, GA, USA, on the speech attribute modeling-based non-native mispronunciation detection, corrective feedback design, and Mandarin tone mispronunciation detection. He was a research intern with Microsoft Research Asia, Institute for Infocomm Research and Alibaba Research in 2011, 2016 and 2018 respectively. His major research interests lie in computer assisted language learning, cross-modal speech recognition, and speech enhancement.