2020

MISPRONUNCIATION DETECTION AND DIAGNOSIS IN MANDARIN ACCENTED ENGLISH SPEECH

Subash Khanal
University of Kentucky, subash.khanal33@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2020.340

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Subash Khanal, Student

Dr. Michael T. Johnson, Major Professor

Dr. Daniel Lau, Director of Graduate Studies

MISPRONUNCIATION DETECTION AND DIAGNOSIS IN MANDARIN ACCENTED ENGLISH SPEECH

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in Electrical Engineering in
the College of Engineering at the
University of Kentucky

By

Subash Khanal

Lexington, Kentucky

Director: Dr. Michael T. Johnson, Professor of Electrical Engineering

Lexington, Kentucky

2020

ABSTRACT OF THESIS

MISPRONUNCIATION DETECTION AND DIAGNOSIS IN MANDARIN ACCENTED ENGLISH SPEECH

This work presents the development, implementation, and evaluation of a Mispronunciation Detection and Diagnosis (MDD) system, with application to pronunciation evaluation of Mandarin-accented English speech. A comprehensive detection and diagnosis of errors in the Electromagnetic Articulography corpus of Mandarin-Accented English (EMA-MAE) was performed by using the expert phonetic transcripts and an Automatic Speech Recognition (ASR) system. Articulatory features derived from the parallel kinematic data available in the EMA-MAE corpus were used to identify the most significant articulatory error patterns seen in L2 speakers during common mispronunciations. Using both acoustic and articulatory information, an ASR based Mispronunciation Detection and Diagnosis (MDD) system was built and evaluated across different feature combinations and Deep Neural Network (DNN) architectures. The MDD system captured mispronunciation errors with a detection accuracy of 82.4%, a diagnostic accuracy of 75.8% and a false rejection rate of 17.2%. The results demonstrate the advantage of using articulatory features in revealing the significant contributors of mispronunciation as well as improving the performance of MDD systems.

KEYWORDS: Mispronunciation detection and diagnosis (MDD), Articulatory features, Automatic Speech Recognition, Deep Neural Network (DNN).

Subash Khanal

July 2020
Date

MISPRONUNCIATION DETECTION AND DIAGNOSIS IN MANDARIN ACCENTED ENGLISH SPEECH

By
Subash Khanal

Dr. Michael T. Johnson
Director of Thesis

Dr. Daniel Lau
Director of Graduate Studies

July 2020
Date

DEDICATION

To my Father and Mother

ACKNOWLEDGMENTS

I pay my sincere gratitude to my advisor, Dr. Michael T. Johnson, for his great support and invaluable guidance at every stage of my research at the University of Kentucky. I also would like to thank my graduate committee members, Dr. Kevin Donohue and Dr. Nathan Jacobs for their time and feedback which helped in shaping the finished product of my thesis.

I would like to thank Narjes Bozorg for her help in learning the speech recognition toolkits used for the research. I also would like to thank my other two friends in the Speech and Signal Processing Lab, Mohammad and Sofia. I also cannot miss to thank my friends Patrick Cordero and Aayush Karki. Their constant moral and emotional support kept me going.

Finally, I would like to thank all the teachers and my family in Nepal for their continuous love and support throughout my academic journey so far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. **INTRODUCTION**

1.1 Background and motivation

The research work presented here focuses on detection and diagnosis of mispronunciation in speech by Mandarin speakers of English. The initial work focuses on identification of common mispronunciation errors and their associated articulatory patterns as compared to the correct pronunciations. A Mispronunciation Detection and Diagnosis (MDD) system based on Automatic Speech Recognition (ASR) was designed and implemented and then used to compare automatically identified mispronunciations by MDD with mispronunciations based on expert transcription.

In the age of globalization, learning a second language can be useful to enhance understanding of other culture and exchange of trade and knowledge, and more than half the world's population speak at least two languages. While there are 34 languages that have 45 million or more speakers, a few languages such as English, Mandarin Chinese, Hindi and Spanish have more than 500 million total speakers each (Ethnologue 2019). These languages attract a large number of second language (L2) learners, making them some of the most popular second language choices for learners around the world. It is estimated that 750 million people speak English as their second language, ranking it as the most popular second language in the world (Ethnologue 2019). In particular, among the estimated 416 million Chinese foreign-language learners, 94% (390 million) of them are learning English (Wei and Su 2012). These numbers clearly indicate the popularity of English as second language around the globe, with a large population of English as second language (ESL) learners in China alone. Due to this large demand, the market size of the English language training (ELT) in China is projected to grow from 41.5 billion U.S.

dollars in 2017 to 75 billion U.S. dollars in 2022 (Statista 2019). This data on speakers and L2 learners around the world provides motivation for academia and computer technology based companies to research and develop effective tools for language learning.

Computer Assisted Language Learning (CALL) is the study of applications of the computer to language teaching and learning (Levy and Hubbard 2005). The history of CALL can be dated back to 1960s (ICT4LT), and a detailed overview of the history of CALL can be found in (Butler-Pascoe 2011). CALL is a broad field that encompasses a wide range of teaching-learning elements of language training, including vocabulary, grammar, phonetics and pronunciation. Whereas reading and writing skills in a language are acquired based on the understanding of language rules and vocabulary, speaking and listening require the physical skill of speech production as well as the ability and training to perceive speech. Pronunciation is considered as a key sub-skill of speaking. After a certain proficiency standard, the main factor that greatly hinders the communication process in L2 learners is pronunciation (Hinofotis and Bailey 1980). Some of the important linguistic factors affecting the learning of pronunciation are accent, stress, intonation, rhythm and mother tongue influence (Gilakjani and Ahmadi 2011). Moreover, various non-linguistic factors like age, personality, motivation and attitude of learner and instruction methodologies also have their effect on pronunciation learning (Gilakjani and Ahmadi 2011). Because of the many challenges associated with learning correct speech production in a new language, pronunciation errors are common in the speech of second language (L2) learners. Pronunciation being one of the important skills in learning a new language, for CALL to be effective, special focus has to be given in teaching learners the correct way of pronouncing words in L2. An important sub-field of CALL is Computer

Assisted Pronunciation Training (CAPT), which focuses on assisting learners with the pronunciation aspects of language. CAPT systems can offer realistic and contextualized spoken examples by means of videos and recordings that learners can use to mimic and learn word level pronunciations of the second language (L2) (Neri, Mich et al. 2008). In order to teach pronunciation to L2 learners, CAPT systems should be able to listen the learner's speech, detect mispronunciations and provide corrective feedback. Therefore Mispronunciation Detection and Diagnosis (MDD) plays a vital role in CAPT. MDD in most CAPT systems is based on Automatic Speech Recognition (ASR) technologies. In the context of pronunciation training using MDD, speech is typically solicited from the users, and ASR systems can be used to compare against the known prompts for the speech under test to detect mispronunciations and diagnose them. A detailed literature review of ASR based MDD systems is presented in Section 2.6 of this thesis.

Speech is a result of complex coordinated movement of articulators. Information about the movement of these articulators during mispronunciations can be essential in identifying the primary and secondary contributors for mispronunciations. A detailed literature review of the articulatory comparison between first language (L1) and second language (L2) pairs to diagnose the cause of mispronunciations in L2 speakers is presented in Section 2.4. These earlier works related to articulatory comparison between L1 and L2 speakers have primarily studied specific types of errors and regions of the vocal tract. An extensive data-driven study for a L1/L2 combination that identifies the most common mispronunciation errors and their corresponding articulatory differences between L1/L2 speaker groups across all the regions of vocal tract would greatly be beneficial. This study adds the research contribution in this area. This work compares the difference in

articulation patterns between English (L1) and Mandarin Chinese (L2) speakers of English, for the purpose of providing an understanding of mispronunciation behaviors of L2 learners. This study reveals insight into common substitution errors and the associated articulator movements that play a significant role in mispronunciation patterns.

ASR based MDD for CAPT systems is usually trained solely on acoustic data. However, kinematic data during speech production, represented in the form of articulatory features, can be useful in building a better MDD system. Apart from correctly identifying the types of mispronunciation, an MDD system that incorporates articulatory features can provide meaningful articulatory feedback to the learner, instead of the conventional score based feedback. Understanding the role of articulatory information in mispronunciation, in recent years various ways of incorporating articulatory information into MDD systems have been proposed. A literature review relating to use of articulatory information in design of MDD can be found in Section 2.6 of this document. The systems proposed so far have tried to include expert labeled articulatory state information for each phoneme in the transcript to train the MDD system. However, there is a lack of studies which use real articulatory features as obtained from kinematic recordings of speech and their usefulness in improving the performance of MDD systems. This study aims to fill this gap. This research utilizes the Electromagnetic Articulography Mandarin Accented English (EMA-MAE) database (Ji, Berry et al. 2014) , which is the largest of its kind with around 45 minutes of acoustic and kinematic data from 20 American and 20 Mandarin speakers of English. With the help of the articulatory information provided in the database, key articulatory contributors for mispronunciation were identified. Moreover, the articulatory

features derived from the EMA data were used to improve the performance of ASR based MDD system.

1.2 Contributions and Significance

The research work presented here is on the analysis of mispronunciation with an emphasis on the articulatory differences between the speech of Mandarin and American speakers of English. The key contributions of this work are described as follows.

With regard to comparing articulatory patterns associated with mispronunciation errors, a detailed study has been carried out to identify commonly occurring mispronunciation errors in Mandarin speakers of English and to identify the contributing articulatory error pattern for those errors. Speech recognition models are used to align and extract the articulatory feature frames for the mispronounced sound segments in L2 speech as well as that for corresponding correctly pronounced sound segments in L1 speech. Statistical comparison of these L1 and L2 speech segments in the articulatory feature space reveal the key articulators associated with the mispronunciation under study.

With regard to contributions in the area of Mispronunciation Detection and Diagnosis, an MDD system based on state-of-the-art ASR methods has been implemented and evaluated. Several ASR architectures were evaluated, and in addition the MDD system designed uses Articulatory and Acoustic features available in the EMA-MAE database and analyzes the effectiveness of these features individually and in combination.

Although in this study the L2 speakers were Mandarin speakers of English, the same approach could be used for any L1-L2 language speaker pairs.

1.3 Plan of thesis

The chapters in this thesis are organized as follows: After this introductory chapter, Chapter two provides relevant background concepts related to speech processing, statistical as well as neural network based speech recognition systems, phonological differences between Mandarin and English Language, literature review on analysis of articulatory differences between L1 and L2 speech, ASR based MDD systems and a brief introduction to the EMA-MAE database. Chapter three presents the experiments and results related to identification and diagnostic analysis of L2 mispronunciation errors in the articulatory feature space. Chapter four presents the experiments related to implementing an ASR based MDD system. Finally, the fifth chapter contains a conclusion of the overall study and directions for future work.

CHAPTER 2. **BACKGROUND AND RELATED WORKS**

2.1 Introduction

This chapter provides a general overview of background concepts and literature review of the works related to this study. This includes an overview of speech production, phonetic units of speech and their categories, spectral analysis and different acoustic feature extraction methods, statistical as well as neural network based acoustic modeling along with some techniques proven to improve their performance. The specific tasks associated with this work are analysis of articulatory differences between L1 and L2 speech and building automatic Mispronunciation Detection and Diagnosis (MDD) systems, and therefore literature review focused on these two topics is also provided.

The second section provides an overview of speech production in humans and introduces phonetic classes in English based on articulatory phonology. It also presents the fundamentals of frame-based speech processing, different types of acoustic features and extraction techniques.

The third section describes approaches to building an automatic speech recognition system. It provides the basic theory behind Hidden Markov Models, Gaussian Mixture Models and various Deep Neural Network (DNN) based architectures like Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and light GRU. The section also briefly discusses Speaker Adaptive Training (SAT), feature space Maximum Likelihood Linear Regression (fMLLR) techniques and different optimization algorithms used for training DNNs which have demonstrated to improve speech recognition performance.

The fourth section provides a literature review on articulatory differences between L1 and L2 speech for understanding causes of errors.

The fifth section of this chapter provides an overview of phonological differences between Mandarin and English languages. Commonly occurring mispronunciation errors as reported in the literature for Mandarin speakers of English are also discussed.

The sixth section provides a detailed literature review in the area of automatic Mispronunciation Detection and Diagnosis (MDD). It explores Goodness of Pronunciation based MDD systems, Extended Recognition networks (ERN) and various Deep learning based MDD systems.

The final section of this chapter briefly describes the EMA-MAE database used for this research.

2.2 Overview of speech processing and analysis

Human speech is the acoustic result of organized motions of the respiratory and masticatory apparatus (Flanagan 2013). The history of speech technology can be traced back to as early as 1779 when a Russian Professor Christian Kratzenstein designed an apparatus to produce five long vowels (/a/, /e/, /i/, /o/, /u/) artificially. The apparatus consisted of acoustic resonators similar to the human vocal tract. In 1791, Wolfgang von Kempelen developed his “Acoustic-Mechanical Speech Machine”, capable of producing single as well as some sound combinations. Some notable milestones in the history of speech technologies are as follows

- Invention of Telephone in 1876 by Alexander Graham Bell.
- Development of vocoder (voice coder) by Dudley in Bell labs in 1939.

- Creation of the first automatic speech recognizer called “Audrey” by researchers at Bell labs in 1952.
- In the 1960’s computers were introduced to lead the world towards the digital computing era.
- In the 1980’s Hidden Markov were successfully used for speech recognition tasks.
- 1990’s and 2000’s were the decades of high computing capability through hardware advancement and development of software tools. This brought ASR based products to mainstream market.

With the advent of Deep Neural Networks and their application in speech technologies, this field is growing rapidly. Riding on the wave of progress in Automatic Speech Recognition (ASR), applications like Computer Aided Language Learning (CALL), Computer Aided Pronunciation Training (CAPT) and speech accent conversion are showing promising results. Advances in speech processing and instrumentation have enabled researchers to look closely into the articulatory patterns in speech pronunciation, identify the articulatory cause of mispronunciation and hence provide meaningful diagnostic feedback to language learners.

The two main areas of speech research covered in this work are Articulatory comparisons between L1 and L2 speech for diagnostic analysis of mispronunciation seen in L2 speakers and design of ASR based Mispronunciation Detection and Diagnosis (MDD) systems.

2.2.1 Speech production

The organs involved in human vocal system are shown in Figure 1. The diagram represents a mid-sagittal view through the vocal tract of an adult. The vocal tract can be represented as an acoustical system of tubes with non-uniform cross-sectional area. Movement of the articulators; namely, the jaw, the lips, tongue and velum deforms the cross-sectional area of the vocal tract (Flanagan 2013). The velum is responsible for adjusting the acoustic coupling between the nasal and vocal tracts. For speech production, air is drawn into the lungs by lowering the diaphragm and enlarging the chest cavity and then expelled by contracting the rib cage and increasing the lung pressure (Flanagan 2013). As the forced air is expelled out of lungs, it passes through the trachea and then into the larynx. The Larynx, also called the voice box, houses two lips of ligament and muscle also called the vocal cords, and a slit between the cords called the glottis (Flanagan 2013). The shape of the vocal tract and the movement of the articulators filter the excitation signal, producing different types of sounds. Depending on the vibratory status of the excitation signals, human produce two types of sounds, voiced and unvoiced. Voiced sounds of speech are the result of vibratory action of vocal cords. In contrast, unvoiced sounds of speech do not involve vibration of vocal cords but are due to turbulent flow of air created at a restriction in the vocal tract.

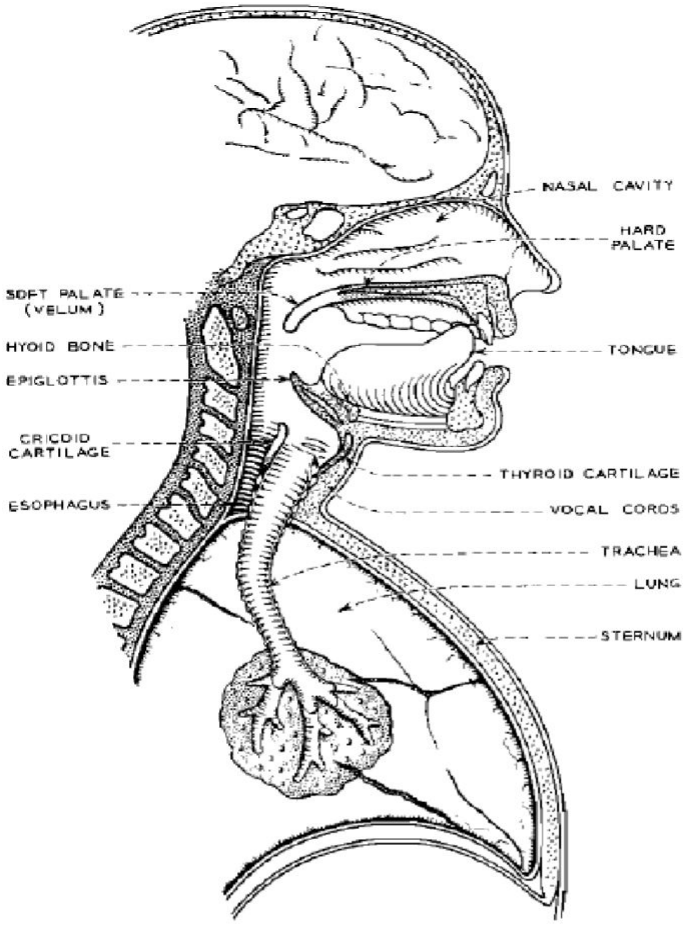

Figure 1 Systematic diagram of human vocal mechanism (Flanagan 2013)

2.2.2 Classification of sounds

Sounds are represented by linguistic symbols called phonemes. Phonemes are the fundamental linguistic units which differentiate meaning in a language. Two commonly used standard phonetic representations include IPA (International Phonetic Alphabet) and ARPABET (by Advanced Research Projects Agency). A list of the IPA and ARPABET phonetic representation is presented in Table 2-1.

In English, sounds are identified primarily by the resonance of the vocal tract, which is separate from the glottal excitation frequency (Flanagan 2013). Based on airflow restrictions in the vocal tract, sounds are broadly categorized into Vowels and Consonants.

The production of vowels does not involve major airflow restriction, whereas consonants are produced with airflow restriction(s) in the vocal tract.

Table 2-1 IPA and ARPABET phoneme representation adapted from (Rice April 1976)

IPA	ARPABET	Example	Translation
b	B	be	B IY
d	D	dee	D IY
e	EY	ate	EY T
f	F	fee	F IY
g	G	green	G R IY N
h	HH	he	HH IY
i	IY	eat	IY T
j	Y	yield	Y IY L D
k	K	key	K IY
l	L	lee	L IY
m	M	me	M IY
n	N	knee	N IY
oʊ	OW	oat	OW T
p	P	pee	P IY
r	R	read	R IY D
s	S	sea	S IY
t	T	tea	T IY
u	UW	two	T UW
v	V	vee	V IY
w	W	we	W IY
z	Z	zee	Z IY
æ	AE	at	AE T
ð	DH	thee	DH IY
ŋ	NG	ping	P IH NG
ɑ	AA	odd	AA D
ɔ	AO	ought	AO T
ə	AX	comma	K AA M AX
ɚ	AXR	letter	L EH T AXR
ɛ	EH	Ed	EH D
ɝ	ER	hurt	HH ER T
ɪ	IH	it	IH T
ʃ	SH	she	SH IY
ʊ	UH	hood	HH UH D
ʌ	AH	hut	HH AH T
ʒ	ZH	zee	ZH IY
ʤ	JH	gee	JH IY
ʧ	CH	cheese	CH IY Z
θ	TH	theta	TH EY T AH
aɪ	AY	hide	HH AY D
aʊ	AW	cow	K AW

2.2.2.1 Vowels

The filtering effect of the vocal tract creates resonances in the frequency spectrum called formants. Formants are numbered in order of increasing frequency with F_1, F_2, and F_3 being the most prominent. F_1 is related to the opened/closed position of the tongue and F_2 is related to the front/back position of the tongue. Based on the values of F_1 and F_2, vowels can be characterized using a vowel space representation as illustrated in Figure 2 for the 14 English vowels used in this work.

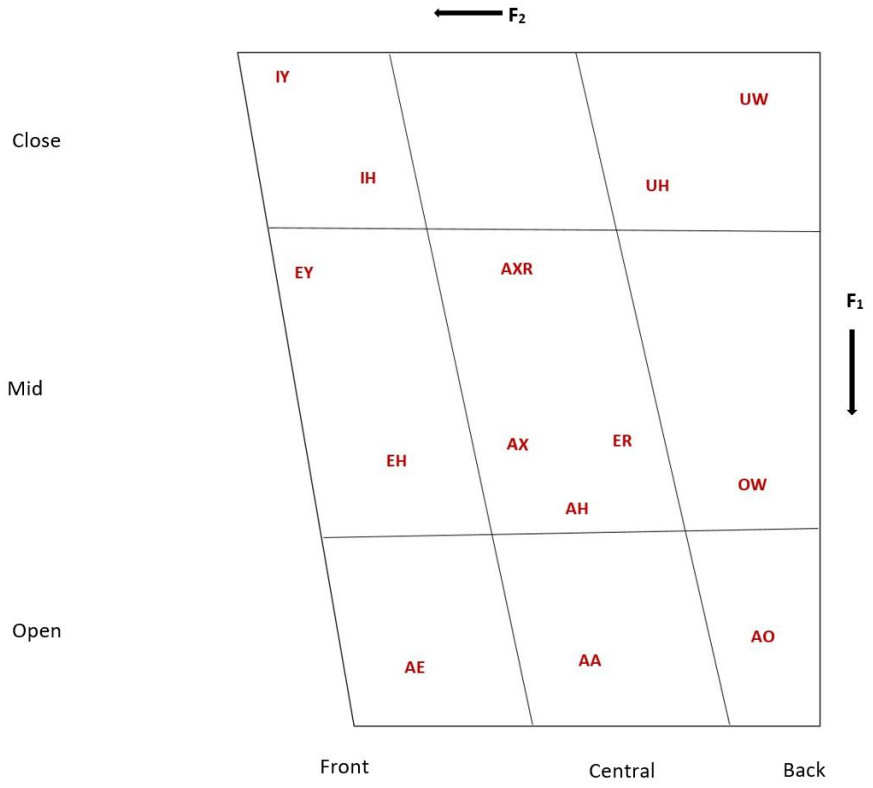

Figure 2 Vowel diagram for English vowels, with the horizontal axis representing front and back position of the tongue, related to F_1 and the vertical axis representing open and closed position of the tongue, related to F_2.

2.2.2.2 Consonants

Based on the place of articulation, consonants can be categorized as bilabial, labio-dental, dental, alveolar, palatal, velar and glottal. Similarly, based on the manner of articulation, consonants can be categorized as stop, fricative, affricate, nasal, liquid and glides. Figure 3 tabulates these categories for English consonants.

Place & Manner	Bilabial	Labio-dental	Dental	Alveolar	Palatal	Velar	Glottal
Stop	P B			T D		K G	
Fricative		F V	TH DH	S Z	SH		HH
Affricate				CH JH			
Nasal	M			N		NG	
Liquid				L	R		
Glide	W				Y		

Figure 3 Consonant chart for classification of consonants based on Place (columns) and Manner (rows) of articulation

2.2.3 Speech processing

Speech production can be represented by a source filter model, as shown in Figure 4. In this model, the signal excitation is represented by pseudo-periodic air flow though vocal folds which is filtered by the resonances of vocal tract to produce the speech signal. Since both the excitation source and vocal tract are time-varying in nature, speech is a non-stationary signal. Mathematically, it is modeled as the excitation signal ($e[n]$) convolved with the time-varying filter representing the vocal tract ($h[n]$) producing the speech signal ($s[n]$). The excitation signal can be modeled as a train of pulses for voiced speech and as white noise for unvoiced speech. Because of its non-stationarity characteristics, the spectral properties of speech are time varying and hence, short time processing using a sliding window called a frame is used as an analysis tool. Typical frame sizes range from 10-30ms, and speech within the frame is modeled as being stationary (Deller, Proakis et al. 2000). Choosing the frame size is a tradeoff between temporal and spectral resolution. Longer frames increase the spectral resolution but lose the stationarity of the speech signal. In contrast, narrower frames produce better temporal resolution but lower spectral resolution. Feature vectors extracted from each of these frames are used as input for building speech recognition or other speech processing systems. The most common feature representations for speech include Linear Predictive Coding (LPC), Cepstrum analysis, filter-banks and perceptual filter banks.

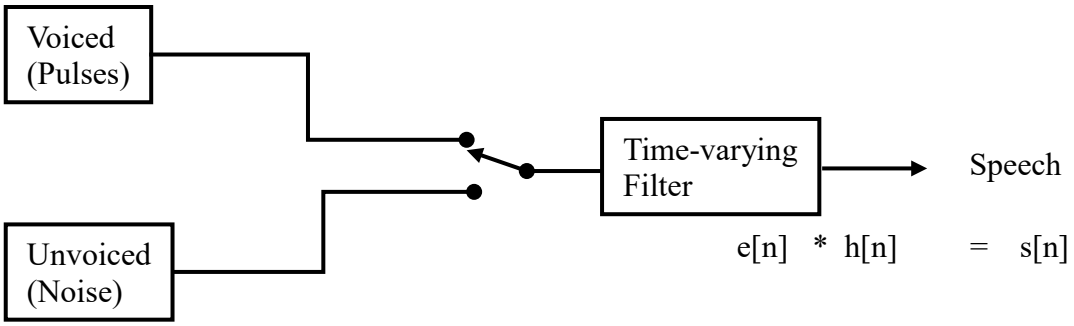

Figure 4 Source filter model for speech production

2.2.3.1 Cepstral Analysis

Cepstral analysis is a general type of Homomorphic Signal Processing, well suited to speech because it allows for separation of excitation and vocal tract filter characteristics and generates a set of largely uncorrelated features representing vocal tract characteristics. The real cepstrum of a signal $s[n]$ is the inverse Fourier Transform of the logarithm of the Fourier Transform Magnitude of the signal Cepstral analysis. This type of analysis separates excitation and vocal tract characteristics and also allows for the introduction on nonlinear frequency scaling to create more perceptually appropriate features. The Mel scale is a nonlinear frequency scale which represents the logarithmic sensitivity of the human auditory system (Huang, Acero et al. 2001). Mel Frequency Cepstral Coefficients (MFCCs) are cepstral coefficients warped on the Mel frequency scale. To compute MFCCs, the magnitude spectrum of each frame is computed and frequency warped on the Mel frequency scale then transformed to the cepstral domain using a Discrete Cosine transform to yield Mel frequency cepstral coefficients. Details of MFCC feature vector extraction for a frame is illustrated in Figure 5.

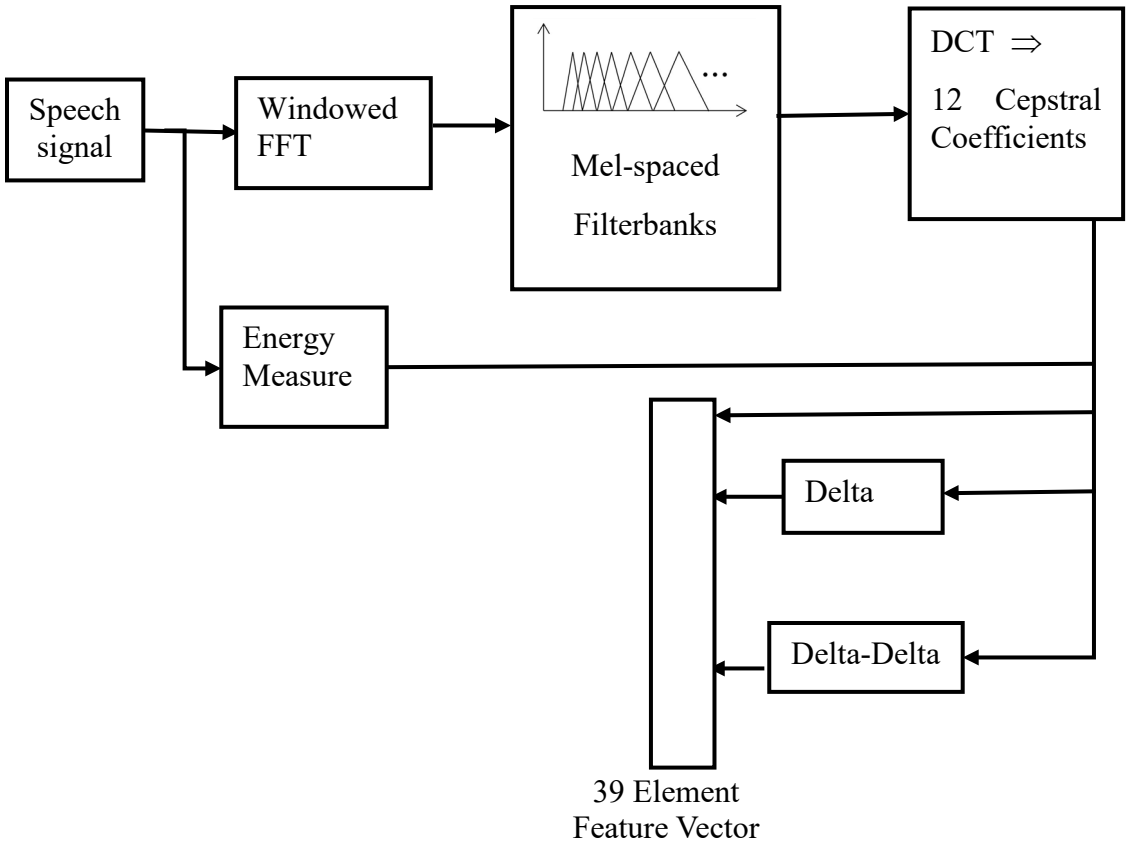

Figure 5 Block diagram of MFCCs feature extraction (Johnson 2018)

In order to obtain information about trajectories of MFCCs over time, the time derivatives of MFCCs called deltas (also called velocity) and delta-deltas (also called acceleration) can be computed and used along with the static MFCC features yielding better speech recognition performance (Yang, Soong et al. 2007). These dynamic MFCC features in the form of time derivatives are calculated using a linear regression.

2.3 Automatic speech recognition systems

In 1952, K. H. Davis, R. Biddulph and S. Balashek at Bell labs developed the first Automatic speech recognition system capable of recognizing spoken digits by a single speaker (Davis, Biddulph et al. 1952). In the 1980's, statistical methods like Hidden Markov Model (HMM) (Ferguson 1980, Levinson, Rabiner et al. 1983) began to be used for speech recognition systems. As a doubly stochastic process, the HMM framework was able to model the intrinsic temporal variations in speech as well as the spectral structure of spoken acoustics (Juang and Rabiner 2005). This framework was the foundation for ASR for over two decades. In the late 2000's, deep neural networks (DNNs) began to outpace HMMs for speech recognition accuracy. Even though Neural networks' theoretical history can be traced back to the 1950's, they could not be practically used (McCulloch and Pitts 1943) due to lack of computational resources as well as ineffective training methods for multiple layer networks. In the last decade, development of different software tools and deep neural network (DNN) variants well suited for sequential modeling task have opened a new direction of ASR technologies. DNN-HMM based frameworks and more recently end to end speech recognition systems are now realized purely based on DNNs.

The next section provides an introduction to acoustic modeling based on the GMM-HMM framework and some techniques that can improve speech recognition performance. The section also provides a brief overview of different DNN architectures used in this work to design ASR models used as the core engine in the Mispronunciation Detection and Diagnosis (MDD) system.

2.3.1 Acoustic Modeling

Features extracted from speech utterances are used to build the acoustic models for automatic speech recognition systems. Statistical models like Gaussian Mixture Models (GMMs) were previously the standard for acoustic modeling. In recent years, deep neural networks (DNNs) based acoustic modeling is increasingly becoming effective and popular.

2.3.1.1 Gaussian Mixture Model (GMM)

Gaussian Mixture Models model the statistical properties of speech signals in the form of a mixture of multiple Gaussians. With sufficient number of mixtures, GMM can model any probability distribution as a linear combination of multiple Gaussian distributions. Given the set of model parameters (λ) for a specific speech unit, probability of the n-dimensional feature vector (x) is modeled by a GMM with *M* n-variate Gaussian distributions as follows.

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i p_i(x), \quad (1)$$

where $p_i(x)$ is the i^{th} multivariate Gaussian distribution with means μ_i and covariance Σ_i weighted by factor w_i such that $\sum_{i=1}^{M} w_i = 1$. M is the number of Gaussian Mixtures. Values for these parameters are obtained by an iterative training algorithm called Expectation Maximization (EM) (Dempster, Laird et al. 1977)

2.3.1.2 Hidden Markov Model (HMM)

Hidden Markov Models have been the most widely used method for temporal sequencing in speech recognition. HMMs model the acoustic events in speech as the sequence of states such that the overall likelihood of the speech generated by the model is maximized. The probability distribution of states of HMM can be represented either by GMMs or DNNs. A conceptual diagram of HMM is presented below.

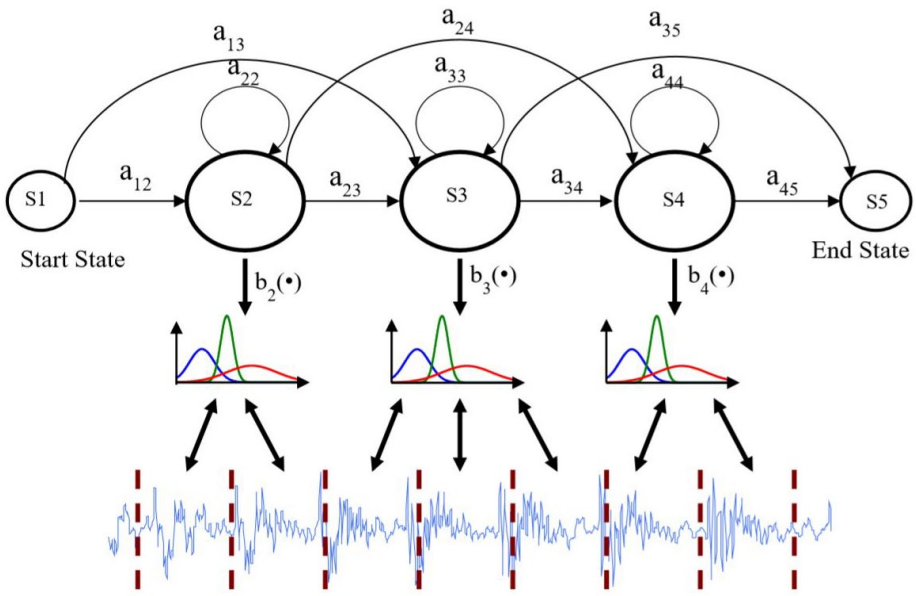

Figure 6 Conceptual diagram of HMM (Johnson 2018)

An HMM is represented by the set of the following parameters: initial state distribution π (the probability of one of the states being the first state of the sequence), the output probability matrix, $B = b_j(o_k)$, where $b_j(o_k)$ is the probability that the state s_i emits the observable symbol o_k, and the state transition probability matrix A containing the state transition probabilities a_{ij}, the likelihood of the transition from the current state (s_i) to the next state (s_j). The set of parameters includes the output probabilities for each state (B),

the transition probabilities (A) and initial state probabilities (π) are collectively called model parameters and can be denoted by $\lambda = (A,B, \pi)$

Fundamental problems and solutions of HMM

1) The Evaluation problem – Given a model λ, and an observation sequence $O = (o_1, o_2 \ldots\ldots o_T)$ compute $p(O \mid \lambda)$, the probability of observation given model. The efficient solution to this problem is a dynamic programming algorithm forward-backward algorithm (Baum, Petrie et al. 1970).
2) The Alignment problem – Given a model λ and an observation sequence O, compute the state sequence $Q = (q_1, q_2 \ldots\ldots\ldots q_T)$ such that $argmax_S \{p(Q|O, \lambda)$, the sequence which best matches with the observation sequence. This best sequence is obtained by a dynamic programming method called the Viterbi Algorithm (Forney 1973).
3) The Training problem – Given a group of observation sequences. Find an estimate of λ such that $\lambda_{ML} = argmax_\lambda \{p(O \mid \lambda)\}$. The solution to this problem is obtained using Baum Welch Algorithm (Baum, Petrie et al. 1970).

An excellent tutorial on HMMs and their application in speech recognition can be read in (Rabiner 1989).

2.3.1.3 Linear Discriminant Analysis-Maximum Likelihood Linear Transform (LDA-MLLT)

Linear Discriminant Analysis (LDA) (Fisher 1936) can be used to project features into a feature space with of lower dimension. Maximum Likelihood Linear Transform (MLLT) (Gales 1999) is a model based transform which clusters full covariance matrices over many distributions. This transform decomposes the covariance matrix for each component into two elements: a non-singular linear transformation matrix W^T shared over

a set of components, and the diagonal matrix Λ_j, yielding the inverse covariance matrix in the form as in (Psutka 2007)

$$\Sigma_j^{-1} \approx W \Lambda_j W^T = \sum_{k=1}^{n} \lambda_j^{\,k} w_k w_k^{\,T}, \tag{2}$$

where Λ_j is a diagonal matrix(diag(λ_j)) and w_k^T is the k^{th} row of the transformation matrix W^T. Analytically, the model parameters for this transform can be estimated using the maximum likelihood approach, however, in practice Expectation Maximization (EM) is used.

2.3.2 Source Variability in Acoustic Modeling

Sources of variability in speech include speaker related factors like gender, age and accent and environmental factors like noise, microphones and channel characteristics. This variability if not well represented in training data used in building acoustic models will cause poor recognition performance on test dataset. The mismatch between training and recognition can be reduced, either by adapting the model to better fit the features from different sources or by transforming features to better fit the model. Several techniques to reduce the effect of source variability are briefly discussed below.

2.3.2.1 Cepstral Mean and Variance Normalization (CMVN)

In order to remove the effect of source or channel variability in speech or speaker recognition task, cepstral mean and variance normalization (CMVN) can be applied (Viikki and Laurila 1998). In this method, once the cepstrum is calculated form the speech, the average of the cepstrum coefficients is subtracted from each coefficient. The mean

subtracted cepstrum coefficients are divided by the variance to obtain cepstral mean and variance normalized features.

2.3.2.2 Maximum Likelihood Linear Regression (MLLR)

Maximum Likelihood Linear Regression (MLLR) (Gales 1998) is a model-space speaker adaptation technique which adapts the parameters of Gaussians with the objective of maximizing the likelihood of the adaptation data for a particular speaker. Linear Model-space transformations can be unconstrained or constrained. In an unconstrained transformation, mean and variance transforms are independent of one another, whereas for a constrained transformation, the transform used for mean is used to transform the variance. The standard MLLR uses unconstrained transformation approach. The general form of a MLLR transform for the mean and variance is given as

$$\hat{\mu} = A\mu + b = W\xi \tag{3}$$

where ξ is the extended mean vector, $[1 \;\; \mu^T]^T$, and W is the extended transform, $[b^T \;\; A^T]^T$. The variance transformation with the transformation matrix H is obtained using

$$\hat{\Sigma} = H\Sigma H^T \tag{4}$$

The optimization of unconstrained MLLR is carried out in two steps. First the transformation of the mean is obtained given the current variance and its transform matrix. Second, the transformation of the variance is obtained given the current mean and its transform.

2.3.2.3 Feature space Maximum Likelihood Linear Regression (fMLLR)

fMLLR is a feature space transform where the features are transformed to better fit the model. It is a constrained MLLR (CMLLR) defined by

$$\hat{x}^{(t)} = W^{(s)}\xi^{(t)} \tag{5}$$

where $\xi^{(t)}$ is the extended feature vector, $[x^{(t)}\ 1]^{T}$ at time t, and $W^{(s)} = [A^{(s)}\ b^{(s)}]$ is the transformation matrix which contains the square matrix $A^{(s)}$ and the bias vector $b^{(s)}$. For the computation of fMLLR transform the sufficient statistics are stored including

$$k_i = \sum_{m=1}^{M} \frac{c^{(sm)}\mu_i^{(m)}\varepsilon\left(\xi\right)^{(sm)}}{\sigma_i^{2(m)}}, \tag{6}$$

$$G_i = \sum_{m=1}^{M} \frac{c^{(sm)}\varepsilon\left(\xi\xi^{T}\right)^{(sm)}}{\sigma_i^{2(m)}}, \tag{7}$$

where M is the total number of Gaussian mixtures, $c^{(sm)}$ is the soft count of Gaussian m from the current speaker s; $\mu_i^{(m)}$ and $\sigma_i^{2(m)}$ are the mean and the variance respectively for i^{th} dimension of mixture m respectively. For the frames x aligned to Gaussian m for speaker s, the quantities $\mathcal{E}(\xi)^{(sm)}$ and $\mathcal{E}(\xi\xi^{T})^{(sm)}$ (where $\mathcal{E}$ denote average) can be computed as in (Povey and Saon 2006) as follows

$$\varepsilon\left(\xi\right)^{(sm)} = \begin{bmatrix} \varepsilon\left(x\right)^{(sm)} \\ 1 \end{bmatrix} \tag{8}$$

$$\varepsilon\left(\xi\xi^{T}\right)^{(sm)} = \begin{bmatrix} \varepsilon\left(xx^{T}\right)^{(sm)} & \varepsilon\left(x\right)^{(sm)} \\ \varepsilon\left(x\right)^{(sm)^{T}} & 1 \end{bmatrix} \tag{9}$$

2.3.2.4 Speaker Adaptive Training (SAT)

Speaker Adaptive training (Anastasakos, McDonough et al. 1996) is an adaptation technique which alternates between the feature and the model space to adapt the speaker

independent model and reduce inter-speaker variation. The training procedure for SAT training can be described as follows:

1. Estimate the speaker independent(SI) model and initialize the transformation as an identity matrix;
2. For each speaker in the training set, compute fMLLR transforms given the current SI model and transform the features using the fMLLR transforms;
3. Using the speaker adapted fMLLR features, estimate a new model set using two iterations of Expectation Maximization algorithm;
4. If not converged, goto step 2.

During recognition, the speaker independent model is used to produce the first best output. This output is used as transcription to be used to estimate fMLLR transforms for the test speakers. The transformed features obtained by using the fMLLR transforms are used to run a second pass of recognition using the speaker adaptive model.

2.3.3 Artificial Neural Network (ANN)

GMM-HMM systems have been the state of the art for speech recognition for over two decades. Despite their strength and success, GMMs have their own limitations. The major one is that they are not efficient in modeling data that lie on or near a non-linear manifold in the data space (Hinton, Deng et al. 2012) . Artificial Neural Networks (ANN) trained with back-propagation algorithm (Rumelhart, Hinton et al. 1985) have proven to better model the non-linear surface of the data-space. Advancement in machine learning algorithms and hardware computational capability over the last two decades have enabled development of efficient training of artificial neural networks with multiple layers and a large number of nodes. Such networks are called Deep Neural Networks (DNNs). DNNs

in Automatic Speech Recognition (ASR) are characterized by many layers of non-linear hidden units. The last output layer of DNNs have the number of nodes equal to the number of context dependent phone states which are represented as the triphone states in the conventional HMM systems. Results on different datasets of different size have shown that DNNs can perform better than GMMs at acoustic modeling task for speech recognition. Different classes of DNNs successfully used in ASR systems are explained below.

2.3.3.1 Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) network is a type of feedforward ANN (Rosenblatt 1961) consisting of an input layer, at least one hidden layer and an output layer. The nodes in the hidden and output layers of an MLP use a nonlinear activation function. This characteristics of MLPs enables them to learn non-linearity in the data-space. An MLP is typically trained using the back-propagation algorithm in a supervised learning framework. An MLP is a fully connected network where the nodes in the current layer are connected to all the nodes in the following layer with some weight w_{ij}. These weights are updated with the goal of minimizing the error computed in the output layer, which is based on difference between the actual and predicted value by the perceptron.

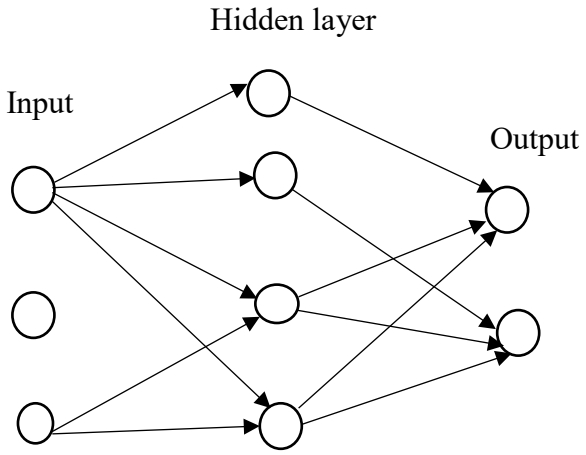

Figure 7 A simple Multilayer perceptron network with one hidden layer

As can be seen in Figure 7 some of the nodes of the MLP are not connected with each other. These missing connections indicate that the weight across them obtained after training was zero.

2.3.3.2 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a class of Neural Networks which capture the dynamic behavior of sequential data. Cells, the fundamental units of RNN, are connected in time by a set of shared weights.

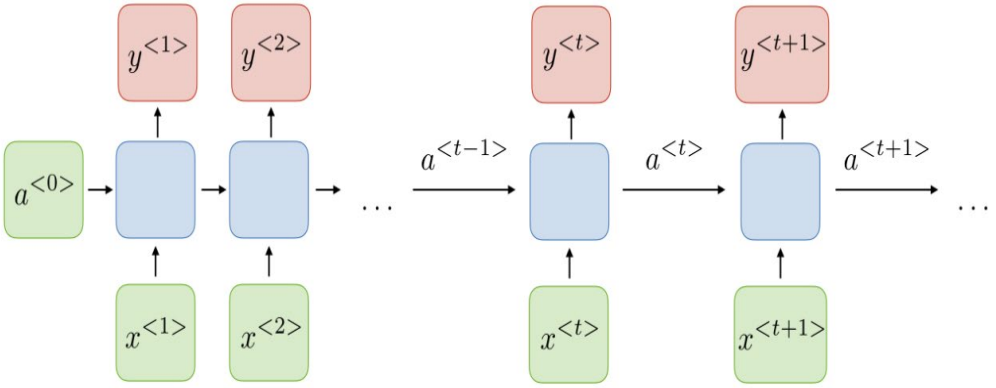

Figure 8 Unidirectional RNN unfolded in time (Afshine Amidi)

For unidirectional RNN, the current time step output ($y^{<t>}$) depends on the current time step input ($x^{<t>}$) and the previous time step activation ($a^{<t-1>}$), whereas for a bidirectional RNN, that depends on the current time step input ($x^{<t>}$) and both the previous time step activation ($a^{<t-1>}$) and the next time step activation ($a^{<t+1>}$). The activation $a^{<t>}$ and the output $y^{<t>}$ for time step t in unidirectional RNN are computed as follows

$$a^{\langle t \rangle} = g_1\left(W_{aa}a^{\langle t-1 \rangle} + W_{ax}x^{\langle t \rangle} + b_a\right) \tag{10}$$

$$y^{\langle t \rangle} = g_2\left(W_{ya}a^{\langle t \rangle} + b_y\right) \tag{11}$$

where $W_{ax}, W_{aa}, W_{ya}, b_a, b_y$ are the shared weights and biases between the RNN cells; g_1 and g_2 are the non-linear activation functions. The most commonly used activation functions in RNNs are: sigmoid, tanh and RELU (Rectified Linear Unit).

The loss function for RNN is computed as the summation of loss across the time steps for the given input sequence. It is defined as

$$L\left(\hat{y}, y\right) = \sum_{t=1}^{T_y} L\left(\hat{y}^{\langle t \rangle}, y^{\langle t \rangle}\right) \tag{12}$$

Once the loss is computed, backpropagation through time (BPT) is performed by partial differentiation of loss L with respect to the shared weights in the matrix W. BPT can be defined as

$$\frac{\partial L^{(T)}}{\partial W} = \sum_{t=1}^{T} \frac{\partial L^{(T)}}{\partial W}\bigg|_{(t)} \tag{13}$$

2.3.3.3 Long Short Term Memory (LSTM)

The fundamental mechanism of an RNN network is to pass the information present in previous time step(s) to the current time step so that the long term dependencies between the features are captured. However, as the gap between the current time step and the time step where the relevant information is present grows, RNNs are not able to effectively capture such dependencies. To address this, a special type of RNN called a Long Short Term Memory (LSTM) network was proposed by Hochreiter and Schmidhuber in (Hochreiter and Schmidhuber 1997). The fundamental element of an LSTM network is a cell. Each LSTM cell is implemented with a set of gates: Update gate (Γ_u), Relevance gate (Γ_r), Forget gate (Γ_f), and Output gate (Γ_o). Figure 9 followed by the corresponding set of equations illustrate a standard LSTM cell architecture.

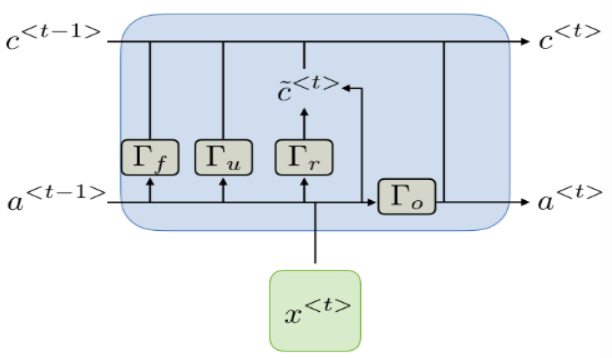

Figure 9 The LSTM cell (Afshine Amidi)

$$\tilde{c}^{\langle t\rangle} = \tanh\left(W_c\left[\Gamma_r * a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_c\right) \tag{14}$$

$$c^{\langle t\rangle} = \Gamma_u * \tilde{c}^{\langle t\rangle} + \Gamma_f * c^{\langle t-1\rangle} \tag{15}$$

$$a^{\langle t\rangle} = \Gamma_o * c^{\langle t\rangle} \tag{16}$$

2.3.3.4 Gated Recurrent Unit (GRU)

A Gated Recurrent Unit (GRU) (Cho, Van Merriënboer et al. 2014) is a variant of LSTM. Unlike the standard LSTM, GRU eliminates the need of a separate forget gate and the Output gate. It reduces the computational complexity and hence reduces the training time for each epoch.

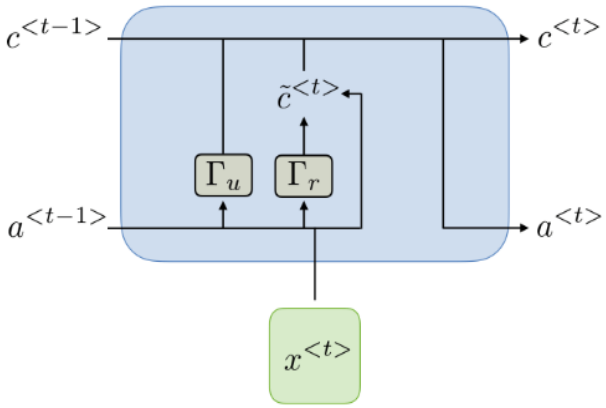

Figure 10 The GRU cell (Afshine Amidi)

$$\tilde{c}^{\langle t\rangle} = \tanh\left(W_c\left[\Gamma_r * a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_c\right) \tag{17}$$

$$c^{\langle t\rangle} = \Gamma_u * \tilde{c}^{\langle t\rangle} + (1-\Gamma_u) * c^{\langle t-1\rangle} \tag{18}$$

$$a^{\langle t\rangle} = c^{\langle t\rangle} \tag{19}$$

2.3.3.5 Light Gated Recurrent unit (liGRU)

For speech signals, the average activations produced by the update and reset gates in the standard GRU are shown to have strong temporal correlation (Ravanelli, Brakel et al. 2018). In other words, there is redundancy of information in activations produced by the update and reset gate. Therefore in (Ravanelli, Brakel et al. 2018) a variant of the

standard GRU where the reset gate is removed was proposed. This change reflected as modification of equation 17 as shown in the equation below

$$\tilde{c}^{\langle t \rangle} = \tanh\left(W_c\left[a^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_c\right) \tag{20}$$

This modification has proven to produce the best speech recognition results for the commonly experimented speech corpus like TIMIT, DIRHA, CHiME and LibriSpeech. Moreover, as reported in (Ravanelli, Brakel et al. 2018), the per-epoch training time was improved by 30% as compared to the standard GRU .

2.3.4 Optimization for Training Deep Neural Networks

Optimization in the simplest mathematical sense is a method of finding value of the argument for which the given objective function (*f(x)*) is maximized or minimized. Deep neural networks use an optimization algorithm to find weight and bias values of the network for which the loss function is minimized, which is the core goal of neural network training. The point where the overall loss function (*f(x)*) is minimum is called the global minimum. In practical training of deep neural networks, especially for features with multiple dimension, the global minimum is hard to achieve. The goal of optimization therefore is to reduce the loss function as well as the true error rate of the model being trained.

2.3.4.1 Gradient descent

Gradient descent, originally proposed by Cauchy in 1847, is the first order optimization algorithm used in DNN training. It iteratively updates the model parameters based on the gradient of the loss for the current model parameters are updated and the term $\Delta_\theta L(\theta_k)$ is the gradient (g_k) of the loss function for the model parameters θ_k . As

θ_k converges to a local minimum, the gradient approaches zero. The learning rate can be fixed or can be iteratively updated as follows

$$\alpha_k = \epsilon_0 - \left(k \big/ \Gamma\right)\left(\epsilon_0 - \epsilon_\Gamma\right) \tag{21}$$

where α_k is the learning rate for k^{th} epoch. This causes the learning rate to decrease from the initial learning rate (ϵ_0) to the final learning rate (ϵ_Γ).

Various advanced optimization algorithms with their own strength and weaknesses have been developed as a modification to Gradient descent. Some of the most commonly used optimization algorithms used for Training Deep Neural Networks are Stochastic Gradient Descent (Bottou 2010), AdaGrad (Duchi, Hazan et al. 2011), RMSprop (Hinton, Srivastava et al. 2012) , and Adam (Kingma and Ba 2014)

2.4 Articulatory comparison of L1 and L2 speech: Literature Review

According to speech learning models of L2 pronunciation, the phonetic system responsible for production and perception of phonetic units reconfigures itself while learning new sounds, through both addition of new phonetic categories and modification of phoneme paradigms from the L1 language (Flege 1995). The difference in phonology between the native language systems of L1 and L2 speakers has been proposed as a primary cause of negative language transfer effects (Meng, Zee et al. 2007). Since the actions of articulators are the phonological basis of pronunciation (Browman and Goldstein 1992), these negative language transfer effects can further be studied and mitigated by looking at the articulators from a speech production perspective during L2 pronunciation errors.

Many pronunciation errors made by second language learners are caused by incorrect articulatory patterns. In order to provide meaningful feedback to the learners, it is important to have a clear understanding of the kinds of pronunciation errors that are most common for a specific L1/L2 language combination and of what production or articulatory errors are associated with those pronunciation errors.

There have been several studies on articulatory phonology which have focused on modeling the primary articulators for different categories of sound (Stone and Lundberg 1996) (Sanguineti, Laboissiere et al. 1997) (Wang, Green et al. 2013). The understanding of the role of articulators in speech production can also be used to study the key vocal tract regions attributed to commonly occurring mispronunciations among L2 learners.

To investigate language-specific articulatory settings, an experiment was conducted using X-ray data of 5 English and 5 French speakers in (Gick, Wilson et al. 2004). The study revealed significant difference across languages in the positions of articulators during speech pauses at five locations in the vocal tract.

Using ultrasound imaging, the study in (Wilson 2013) looked into the anticipatory articulatory patterns during pauses in speech. Comparison of Inter-speech posture (ISP) articulators between Canadian English and Quebecois French monolinguals was performed. The significantly different ISP components across two groups of speakers were: upper and lower lip protrusion, tongue tip height and lip corner position.

The study in (Nissen, Dromey et al. 2007) compared tongue movement patterns for Spanish and Korean bilingual speakers when speaking L1 versus L2 (English). Investigation of Intraspeaker difference in Speed, duration and tongue strokes length

revealed that the speakers had slower but longer tongue movement durations for L2 as compared to L1.

For non-native speakers the most common pronunciation errors are often between sounds which have similar but not identical articulatory positions. While native speakers can perceive and produce the contrast between such sounds, non-native speakers struggle to differentiate. A study using Articulography was done (Wieling, Veenstra et al. 2015) on native and non-native (Dutch) speakers of English focusing on the anterior-posterior position of the tongue-tip during production of two pairs of sounds /s/-/ʃ/ and /t/-/θ/. In (Wieling, Veenstra et al. 2017) the authors compared articulatory trajectories between three groups of speakers – native English, German and Dutch, speaking English. Both studies revealed lower contrast in speech production among L2 speakers for such sound pairs compared to L1 speakers.

Difference in articulatory settings for two Dutch dialects was quantified in (Wieling, Tomaschek et al. 2016, Wieling and Tiede 2017). A comparison of tongue movement data recorded using EMA for 34 speakers during pauses in speech revealed significantly more frontal positions for the Ubbergen dialect speakers as compared to Ter Apel dialect speakers. Curves fitted to the data points for tongue trajectories for two groups of speakers revealed clear distinction between the dialects. Articulatory characteristics of frontal lingual consonants among Catalan dialects using electropalatography were studied in (Recasens 2010). The study revealed differences in location of constriction in anterior region of tongue among Catalan dialects.

In a related work, EMA articulatory data was compared between a Mandarin speaker of English and a native speaker of English in (Li and Wang 2012). For the English

phonemes not present in Mandarin phonemic inventory, pairwise Mahalanobis distance between the displacements of the articulators on tongue (3 points) and lips (3 points) was calculated between the two speakers. The dissimilarity information visualized from Hierarchical clustering analysis (HCA) and multi-dimensional scaling (MDS) clearly show significant difference in articulation between the Mandarin sounds and their English equivalents.

2.5 Phonological difference between Mandarin and English

Mispronunciation in second language learners is caused by a complex set of factors related to the differences between the L1 and L2 languages, the age at which second language acquisition began, and many issues related to individual, social, and cultural identity. One key aspect is transfer of phonological aspects from the native language (L1) to the second language (L2) (Edwards and Zampini 2008) .Various causes of interlanguage transfer proposed by (Weinreich 1953) as summarized in (Edwards and Zampini 2008) are discussed below.

1. Sound substitution: When the learner substitutes the closest L1 equivalent sound in the L2.
2. Phonological process: When the learner uses an allophonic variant (Allophone is one of the phonetically distinct contextual variants of a phoneme) in L1 that does not occur in the same environment in the L2.
3. Underdifferentiation: When two sounds are just allophones in the L1 but are separate phonemes in L2.

4. Overdifferentiation: When two sounds are separate phonemes in L1 but are just allophones in L2.
5. Reinterpretation of distinctions: When the learner reinterprets the distinctions in the L2 as a different form of distinction for them.
6. Phonotactic interference: When the learner applies the syllable structure in the L1 to that in L2.
7. Prosodic interference: Pronunciation error due to difference in prosodic features (for example, intonation) of language.

When second language learners speak, mispronunciation can present differently for the L2 phonemes not present in their native language and for the L2 phonemes which have similar but not identical equivalents in their native language. Fundamental differences between the native language and the second language can be studied to identify most likely pronunciation errors. In this work, the focus is on Mandarin speakers of English as a second language, therefore the phonological differences between Mandarin and English, and in particular the interlanguage transfer from Mandarin to English, are of central importance.

There have been many studies comparing Mandarin and English phonetics. A detailed contrastive study of English and Mandarin Chinese phonetics can be read in (Catford, Palmer et al. 1974). In the section below a summary of the key phonological differences between Mandarin Chinese and English based on the analysis in (Catford, Palmer et al. 1974) are discussed.

2.5.1 Consonants

Manner of Articulation \ Place of Articulation		Both Lips (bilabial)	Lower Lip and Upper Teeth (labiodental)	Tongue Tip and Teeth or Gum (apical)	Front of Tongue & Hard Palate (palatal)	Tongue Tip and Hard Palate (retroflex)	Back of Tonue & Soft Palate (velar)
Stops (all voiceless)	aspirated	p'		t'			k'
	unaspirated	(p)		(t)			(k)
Affricates (all voiceless)	aspirated			(ts')	(tɕ')	(tʂ')	
	unaspirated			(ts)	(tɕ)	(tʂ)	
Fricatives (all voiceless)			f	s	(ɕ)	(ʂ)	(x)
Nasals		m		n			ŋ
Lateral				l			
Continuants		(w)			(y)	(r)	

Figure 11 Consonant Phonemes of Mandarin Chinese. Circled phonemes are not shared with English , the apostrophe ' is used to indicate aspiration (Catford, Palmer et al. 1974)

Manner of Articulation \ Place of Articulation		Both Lips (bilabial)	Lower Lip and Upper Teeth (labiodental)	Tip of Tongue and Teeth (interdental)	Tip of Tongue and Tooth Ridge (apicoalveolar)	Front of Tongue and Hard Palate (laminopalatal)	Back of Tongue and Soft Palate (dorsovelar)	Throat (glottal)
Stops	voiceless	p			t		k	
	voiced	(b)			(d)		(g)	
Affricates	voiceless					(tʃ)		
	voiced					(dʒ)		
Fricatives	voiceless		f	(θ)	s	(ʃ)		(h)
	voiced		(v)	(ð)	(z)	(ʒ)		
Nasals		m			n			
Lateral					l			
Semivowels		w			(r)	y		

Figure 12 Consonant Phonemes of English. Circled phonemes are not shared with Mandarin (Catford, Palmer et al. 1974)

From observation of Figure 11 and Figure 12 we can notice several key differences. For Mandarin, all the stops, affricates and fricatives are voiceless. Unlike in English where the stops, affricates and fricatives are distinguished by voicing, for Mandarin the distinguishing manner of articulation is aspiration. The stops [b, d, g] of English are absent for Mandarin. The voiced stops of English [p, t, k] have both aspirated and unaspirated

versions but only voiceless versions of them are present in Mandarin. The thirteen consonant sounds present in English but not in Mandarin are [b, d, g, ʧ, ʤ, ɵ, ʃ, h, v, ð, z, ʒ and r]. The voiceless affricate sound ʧ found in English has a set of six similar affricates both with and without aspiration. English /r/ caries lip-rounding, whereas Mandarin /r/ is rounded only when immediately preceding a rounded vowel or semivowel (Catford, Palmer et al. 1974). Mandarin /x/ is similar to English /h/ except that the Mandarin sound is pronounced with some friction. Based on the consonants charts for English and Mandarin phonetics, it can be expected that the Mandarin learners of English might have difficulties in voicing and aspiration related pronunciation. Moreover, there are no final consonants except /n/ and /ŋ/ at the end of the syllables in Mandarin so the learners most likely omit the final consonants or add an extra vowel after the final consonants (Catford, Palmer et al. 1974). There are also no consonant clusters within a single syllable in Mandarin, therefore learners seem to have difficulty in pronouncing words containing consonant clusters.

Lastly, English is a stress language. Stress for English consonants decreases from initial position of the word to the final position (Catford, Palmer et al. 1974). In Mandarin however, there is lack of variation of stress based on position of consonants. This makes it difficult for Mandarin speakers to put right amount of stress in right position. Therefore, based on these phonological differences discussed above, difficulties in pronunciation of consonants for Mandarin learners of English can be broadly grouped into five categories: aspiration, voicing, final consonants, consonants cluster and positional variation of stress.

2.5.2 Vowels

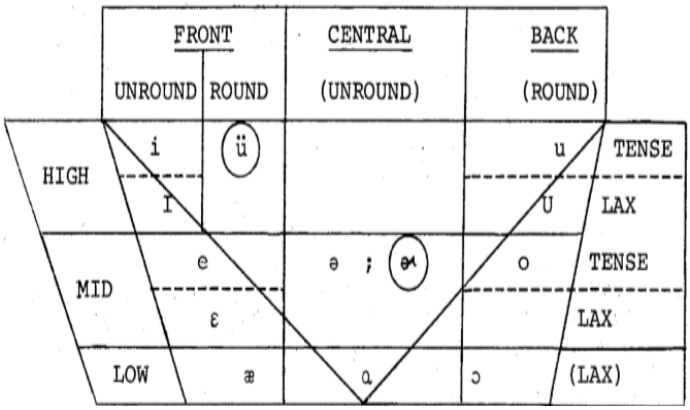

Figure 13 The eleven vowels in English in vowel quadrilateral along with two circled vowels only present in Mandarin (Catford, Palmer et al. 1974)

Figure 13 presents the English and Chinese vowel system for eleven vowels. The vowel quadrilateral here contains the eleven vowels in English [i, ɪ, e, ɛ, æ, ɑ, ə, ɔ, o, ʊ, u] whereas the inverted triangle within the quadrilateral represents the smaller vowel space for Mandarin phonetics containing six vowels [i, ə, ɑ, u, ü and ɚ]. The mandarin vowel /ü/ is pronounced with the same tongue position as /i/ but with the lips rounded and /ɚ/ is pronounced with the same tongue position as /ə/ but with the tongue tip raised behind the gum ridge (Catford, Palmer et al. 1974).Based on the length of pronunciation, in English, vowels can be classified into two groups: long vowels (also called tense) and short vowels (also called lax). However, all six vowels in Mandarin are lax. Therefore it is expected that Mandarin learners of English will have difficulty in distinguishing between lax and tense vowels of English. The glide sounds in English [j, y, w] for Mandarin speakers can be considered as consonantal allophones of high vowels [i, u, ü]. Therefore, these high vowels often get preceded by glides as [/ji/, / yu/, /w ü/] or themselves become glide if they are followed by or preceded by another vowel in the same syllable (Catford, Palmer et al.

1974) In summary, special difficulties regarding vowel sounds for the Mandarin learners of English can be attributed to three sources, small (triangular) vowel space, lack of tense-lax distinction and glides.

2.5.3 Commonly occurring Mispronunciation errors for Mandarin speakers of English

Mispronunciation errors for Mandarin speakers of English are primarily based on the difficulties due to phonological differences between Mandarin and English as discussed in Section 2.5. Mispronunciation can take one of the three general forms: substitution, deletion and insertion. Commonly occurring mispronunciation errors for Mandarin speakers of English (L2) as reported in (Eslan , Catford, Palmer et al. 1974, Deterding 2006, Zhang and Yin 2009, Deterding 2010, Huang and Pickering 2014) are summarized as follows:

1. Voiced stops substituted by voiceless stops

   /b/ → /p/; for example, [bill] is pronounced as [pill].

   /d/ → /t/; for example, [do] is pronounced as [to].

   /g/ → /k/; for example, [get] is pronounced as [ket].

2. Voiced fricatives substituted by voiceless fricatives

   /z/ → /s/; for example, [zip] is pronounced as [sip].

   /v/ → /f/; for example, [vicious] pronounced as [ficious].

3. Voiced affricate /ʤ/ pronounced as an unaspirated voiceless affricate /ts/; for example, [/ʤ/ive] (jive) is pronounced as [/ts/ive].

4. Difficulties with /ɵ/ and /ð/; /ɵ/ is often mispronounced as /s/, /ʃ/, /t/ or /ts/./ð/ can either be mispronounced as the same set or the voiced counter-parts of the same as /z/, /ʒ/, /d/ or /ʤ/.

5. Confusing /l/ for /n/; for example, night might be pronounced as light.
6. Confusion between the final consonants /n/ and /ŋ/; for example, sun might be pronounced as sung.
7. Substitution of /r/ with /l/ or sometimes with /w/; for example, [rice] might be pronounced as [lice].
8. Final consonants either omitted or pronounced with additional vowel in the end; for example, lab (læb) might be pronounced as [læ] or [læbə].
9. Difficulty with consonant clusters; For example, the cluster /zd/ in used might be pronounced as [yuz].
10. Vowel reduction and the schwa sound; the schwa sound denoted by /ə/ in English is a reduced unstressed vowel located at mid central location in vowel quadrilateral. L2 speakers find it difficult to use the idea of vowel reduction using the schwa sound.
11. Confusion between /ɑ/ and /ɛ, æ/;as seen in Figure 13, the range of tongue movement space for /ɑ/ in Mandarin vowel system can reach upto the location where / ɛ, æ/ are located in English vowel system, so it is expected for L2 speakers to have problems in distinguishing between /ɑ, ɛ, æ/.
12. Confusion between /e/, /ɛ/, and /æ/; because of the proximity of the vowels /e/, / ɛ/, and /æ/, all in the frontal part of the vowel space, there is often confusion between these vowels for L2 speakers.
13. Confusion between /i/ and /ɪ/; as mentioned before, all the Mandarin vowels are lax vowels, therefore it is difficult for L2 speakers to distinguish between /i/ (a tense vowel) and /ɪ/ (a lax vowel).

14. The rounded tense mid vowel /o/ may get substituted by rounded lax lower vowel /ɔ/.
15. The rounded lax lower vowel /ɔ/ may get substituted by an unrounded low vowel /ɑ/.
16. Confusion between the non-lower round vowels, /o, ʊ and u/ can happen for L2 speakers.

The earlier related works as described in Section 2.4 have primarily studied specific types of error and regions of the vocal tract. There is no extensive data-driven study for a L1/L2 combination that identifies the most common mispronunciation errors and their corresponding articulatory differences between L1/L2 speaker groups across all the regions of vocal tract. This study aims to fill this gap. In the first objective of this research, comparison of the difference in articulation patterns between native (L1) and non-native (Mandarin) (L2) speakers of English, for the purpose of providing an understanding of mispronunciation behaviors of L2 learners was performed. This study reveals insight into common substitution errors and the associated articulator movements that play a significant role in mispronunciation pattern.

2.6 ASR based Mispronunciation Detection and Diagnosis (MDD) systems: A literature Review

Automatic Speech Recognition (ASR) based Mispronunciation Detection and Diagnosis (MDD) systems are a central element of Computer Aided Pronunciation Training (CAPT). For over two decades, many systems using different features and speech

recognition models have been used for MDD. The research papers in this domain can be broadly organized into three groups.

The first group is research that has mainly focused on the Goodness of pronunciation based MDD systems which classified a phoneme as correct or mispronounced based on the "quality" measured in terms of acoustic probability generated by the ASR system. As an important part of Goodness of Pronunciation based MDD systems, there has been significant work in the area of including pronunciation variations during acquisition of second language (L2) in the form of extended recognition networks (ERNs). In ERNs, a decoding graph of possible pronunciations is formed , and if the path that the decoded phoneme sequence takes matches that of the reference sequence, the word is said to be correctly pronounced, otherwise it is labeled as mispronounced. These ERNs can be designed based on context-sensitive phonological rules of both the L1 and L2 language or can be purely data-driven.

The second group is in the direction of exploring different neural network architectures. Multi-distribution neural networks, neural networks trained in a different fashion and different targets all have shown promising results especially for the detection aspect of MDD systems. The third line of research has been towards incorporation of articulatory features both for detection and diagnosis of mispronunciation, based on the idea that pronunciation is rooted in articulatory phonology. This section will now present a brief review of the papers related to or falling in any of the above-mentioned categories.

Goodness of pronunciation (GOP) metrics quantify the quality of each phone of an utterance on the basis of acoustic likelihood. GOP is defined as log of the posterior probability $P(p|O^{(p)})$, which is the probability that the speaker pronounced phone p given

the corresponding acoustic segment $O^{(p)}$ normalized by the number of frames in the acoustic segment under consideration.

$$GOP(p) \equiv \frac{\left|\log\left(P\left(p \mid O^{(p)}\right)\right)\right|}{NF(p)}, \\ = \frac{\left|\log\left(\frac{P(O^{(p)} \mid p)P(p)}{\sum_{q \in Q} P(O^{(p)} \mid q)P(q)}\right)\right|}{NF(p)} \tag{22}$$

An HMM is used to determine the likelihood $p(O^{(q)} \mid q)$ of the acoustic segment $O^{(q)}$ corresponding to each phone q. Q denotes the set of all HMM phone models and $NF(p)$ represent the number of frames in the acoustic segment $O^{(p)}$. Assuming equal prior likelihoods of all phones and with the sum in the denominator approximated by the max operator, the equation 22 becomes

$$GOP(p) = \frac{\left|\log\left(\frac{P(O^{(p)} \mid p)}{\max_{q \in Q} P(O^{(p)} \mid q)}\right)\right|}{NF(p)} \tag{23}$$

The required likelihood and segment duration is obtained from Viterbi alignments. The numerator of equation 23 is computed by running HMM phone models in forced alignment model. During forced alignment, the sequence of phone models is fixed by the known transcription against which the acoustic feature frames are aligned. The denominator of the equation is determined by using an unconstrained phone loop (Witt and Young 2000). In (Witt and Young 2000), the authors have studied the automatic scoring method based on GOP for mispronunciation detection. The score computed is supplemented by the phone-dependent thresholds to identify the phone as mispronounced or not.

More recently, the GMM-HMM based GOP approach has been replaced by DNN-HMM based GOP for MDD in (Hu, Qian et al. 2015). Moreover, Extended Recognition Networks (ERNs) were implemented in MDD systems for Chinese learners of English in (Harrison, Lo et al. 2009). The ERN incorporated not only the native English speaker's correct pronunciations but also the common mispronunciations by the target L2 learners. ERNs were usually designed based on phonological rules. However, knowledge based phonological rules may not be sufficient or efficient in capturing all the variants of mispronunciations. Therefore automatic derivation of such phonological rules from L2 speech was proposed in (Lo, Zhang et al. 2010). In (Lo, Zhang et al. 2010), a database with 100 speakers was used to derive 2,320 context-dependent phonological rules to include all the types of mispronunciation in the training set. These basic rules were ranked based on the count of occurrence in descending order. Using the F1-score of the MDD system as the deciding metric, 216 phonological rules were selected to design the ERN. As compared to manually designed rules, these data-driven rules resulted in improved diagnostic accuracy of the system.

The knowledge of negative language transfer effect during L2 acquisition can aid better diagnosis of mispronunciation errors. In (Lo, Harrison et al. 2008), fusion of ERNs incorporating knowledge of negative language transfer effect with purely scoring based approach was demonstrated. This method of decision fusion led to 30% reduction in number of decision errors.

There are some limitations to the Extended Recognition Network based approach. It is not feasible to include all possible errors in the network. Therefore the errors not included in the ERNs are never recognized no matter how better the acoustic models are.

Moreover, the phonological rules derived from L2 speech and the acoustic models are trained independently hence contextual information is not preserved. To address these limitations, various ERN free approaches are being sought for.

In (Li, Qian et al. 2016), the authors used multi-distribution deep neural networks for MDD. An Acoustic-graphemic-phonemic model (AGPM) in a multidistribution DNN framework was built for the MDD task. The input to the network was a concatenation of acoustic features, corresponding graphemes and canonical transcription represented in the form of binary encoding. This proposed AGPM model performed better in all the MDD metrics, as compared to DNN acoustic models in conjunction with ERN. Because the number of mispronounced phones is much less than the number of correctly pronounced phones, there is a high data imbalance problem. As a solution to this, a multi-task learning Acoustic-Phonemic model (MT-APM) was proposed in (Mao, Wu et al. 2018). The correctly pronounced and mispronounced phonemes were dealt separately in two different tasks but were trained and decoded jointly.

Also as an alternative to ERN based MDD systems, a two-pass framework in discriminative training mode was proposed in (Qian, Meng et al. 2016). In the first pass, mispronunciation detection was done during which insertion, deletion and substitution was detected. In the free-phone recognition type second pass, mispronunciation diagnosis was carried out on the segments where the errors were detected during the first pass.

A CNN-RNN-CTC based MDD system was proposed in (Leung, Liu et al. 2019). This approach avoids the need for phonemic or graphemic information as well as for forced alignment as required for previously discussed models. Due to the lack of phonemic and graphemic information embedded in input, the proposed model slightly underperformed

the APM, AGM and AGPM approaches; however, it still was better than the conventional ERN based MDD models.

With a focus on providing corrective articulatory feedback, a decision tree based MDD framework has been proposed in (Li, Li et al. 2016). Knowledge guided and data-driven decision trees were constructed to represent articulatory characteristics of correct pronunciations and mispronunciations. Speech attributes were extracted from attribute classifiers of the following categories: Place, Manner, Aspiration, Voicing and silence. Frame level attribute posteriors from the attribute classifiers were passed through a pronunciation attribute scoring module, like in Goodness of Pronunciation scoring. The frame level attribute scores were appended together and fed as input to the MDD decision tree constructed for each phone. The advantage of this approach was that by traversing the decision tree, meaningful diagnostic feedback can be provided to the learner in the case of mispronunciation.

A novel approach of Hidden Articulator Markov models was explored for pronunciation evaluation and was reformulated in (Tepperman and Narayanan 2007) to incorporate articulatory representations in the input features for application to the detection of phone-level mispronunciation errors. Including multidimensional articulatory confidence scores to conventional phone-level confidence scores led to 3-4% absolute reduction of overall error rates as compared to the baseline using only phone-level acoustic features. Towards enhancing Arabic pronunciation, authors in (Abdou, Rashwan et al. 2012) designed an HMM based classifier which classifies each phoneme into manner of articulation features. The articulation based confidence score in helped in reduction of false rejection rate by 25 % as compared to the acoustic model based confidence scoring.

An Articulatory Goodness of Pronunciation approach was explored in (Ryu and Chung 2017). For a given segment, trained acoustic as well as separate articulatory models were used to generate a GOP and 24 aGOP (articulatory GOPs) for the associated articulatory attributes as predictors for diagnosis modeling respectively. If the force aligned segment was recognized as one of the consonants; voicing, place and manner based diagnostic models were used to binary classify it as correct or incorrect. However, if the segment was vowel; Rounding, Height and Backness based diagnostic models were used. Articulatory feature based pronunciation modeling was also proposed in (Livescu, Jyothi et al. 2016). In the proposed model, based on vocal tract variables in Articulatory Phonology, articulatory feature set was derived. When compared with phone-based models, the proposed articulatory feature based models showed significant improvement in frame perplexity as well as lexical access accuracy. Use of speech articulatory attributes for MDD task has also been proposed in (Li, Siniscalchi et al. 2016). A bank of speech attribute classifiers were trained to get frame level posterior probabilities for the attribute under consideration. Authors in (Yuan, Zhao et al. 2012) used articulatory feature based tandem features to improve the performance of such low-resource acoustic models.

Authors in (Mao, Wu et al. 2018) proposed three models: Articulatory-acoustic-phonemic model (AAPM) which includes articulatory features obtained from a phoneme-to-articulatory features map directly into input features, AAPM with feature representation (R-AAPM) which re-represents original input features and articulatory multi-task acoustic phonemic model and (A-MT-APM) where phoneme recognizer and classifiers for articulatory feature classification were trained in multi-task manner. As compared to their

APM baseline, the A-MT-APM approach gained 5.6 % and 7.0 % improvement in F1-measure and diagnostic accuracy respectively.

In summary, the research in ASR based MDD systems has been moving from a pure HMM based Goodness of Pronunciation approach to a Deep Neural Networks based approach. Various types of features and ASR architectures for MDD task have been proposed. In recent years there has been increasing interest towards analyzing and incorporating articulatory features with different representations for Mispronunciation detection and diagnosis tasks. In all of these work related to incorporating articulatory features, the articulatory features on which the models were trained on or sometimes trained for, are only the approximations of the vocal tract state. However, there has not been enough work in the area of including real kinematic data for MDD task. By using the kinematic as well as acoustic features available in Electromagnetic Articulography Mandarin Accented English (EMA-MAE) database (Ji, Berry et al. 2014), this work performs experiments with different DNN architectures and feature combinations to see the contribution of kinematic data in improving the MDD systems.

2.7 EMA-MAE Database

There have been several experimental techniques developed to study the articulatory movements with regards to sound production. One such technique is Electromagnetic Articulography (EMA) (Schönle, Gräbe et al. 1987). EMA is based on the principle of electromagnetic induction. A stationary magnetic field is created, and electromagnetic sensors moving within the field induce a current that can be used to track both position and orientation of the sensor. Sensors are placed on different articulatory

locations for a speaker. EMA therefore provides information about the position and movement patterns of articulators associated with production of sounds. This kinematic data obtained from EMA has been used in study of articulatory phonetics of a language. Information about the articulatory position can play significant role in analysis of mispronunciations in L2 speaker's speech.

This research utilizes Electromagnetic Articulography Mandarin Accented English (EMA-MAE) database (Ji, Berry et al. 2014). The EMA-MAE database consists of kinematic and acoustic data from 39 gender and dialect balanced speakers representing 20 Midwestern standard American English L1 speakers and 19 Mandarin Accented English (MAE) L2 speakers (Ji, Berry et al. 2014). Mandarin speakers include half Northern (Beijing) region and half Southern (Shanghai) region accents. Content includes word-pairs, sentences, and paragraphs totaling about 45 minutes per speaker. Along with audio data recorded with sampling rate of 22.05 KHz, three dimensional EMA data was collected at a 400 Hz sampling rate using the NDI Wave system. The system uses small toroidal electromagnetic sensors within a static electromagnetic field. A reference sensor with 6 Degree of Freedom (DOF) was mounted on the Nose Bridge of subjects to establish a base coordinate system. Other sensors with 5 DOF were attached to the articulators to collect orientation and position data. These articulatory sensors were: lower lip (LL), upper lip (UL), the jaw (MI) lower front incisor), tongue tip (TT), and tongue body (TD) all placed in the midsagittal plane. In addition, there were two lateral sensors, one (LC) at the left corner of the mouth to help indicate lip rounding and one (LT) in the left central midpoint of the tongue body to help indicate lateral tongue curvature.

EMA-MAE data is head-movement-corrected by the system using a stationary reference sensor and then calibrated using bite-plate data to form an articulatory working space based on the midsagittal and maxillary occlusal planes of individual speakers. Palate trace data was collected and used to convert vertical sensor positions into vertical opening measures, and dental measurements were used to normalize horizontal distances across speakers. Details about the collection, calibration and correction of the database can be read in (Ji, Berry et al. 2014).

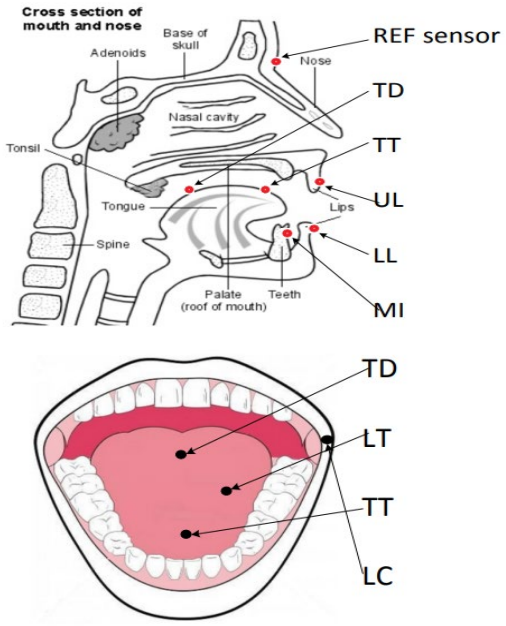

Figure 14 Sensor Placement for collecting kinematic data in EMA-MAE corpus

2.7.1 Articulatory Features

Kinematic data was converted into ten static articulatory features representing the state of the vocal tract during pronunciation. Features and equations are listed in Table 2-2 . The horizontal normalization constant k_x represents the ratio between the incisors to back molar distance of that subject divided by the average of this distance across all subjects.

Units of all features except LC are in millimeters. LC is a proportion relative to the baseline lip corner lateral distance.

Table 2-2 Equations for Articulatory features (Bozorg and Johnson 2018)

Features	Description	Formula
TD_H	Tongue Dorsum normalized horizontal position	$\frac{TD_x}{k_x}$
TD_V	Tongue Dorsum vertical height to hard palate	$palate_y\,(TD_X, TD_z) - TD_y$
TL_H	Tongue lateral normalized horizontal position	$\frac{TL_x}{k_x}$
TL_V	Tongue lateral vertical height to hard palate	$palate_y\,(TL_X, TL_z) - TL_y$
TA_H	Tongue Apex normalized horizontal position	$\frac{TT_x}{k_x}$
TA_V	Tongue Apex vertical height to hard palate	$palate_y(TT_X, TT_z) - TT_y$
LP_H	Normalized horizontal lip protrusion	$\frac{UL_x - \lvert mean(UL_x)_{Biteplanedata} \rvert}{K_x}$
LS_V	Normalized vertical lip separation	$(UL_y - LL_y) - 0.1percentile(UL_y - LL_y)_{CaterpillarPassage}$
LC	Normalized Lateral Lip rounding (Lip corner sensor)	$\frac{LC_z}{mean(LC_z)_{Biteplanedata}}$
JW_V	Vertical middle incisor (jaw)	MI_y

CHAPTER 3. **DIAGNOSTIC ANALYSIS OF L2 MISPRONUNCIATION ERRORS**

3.1 Overview

This chapter outlines the experiments performed with the human annotated phonetic transcripts available in the EMA-MAE corpus. The common phonemic substitution errors occurring in L2 speech were identified by alignment of human annotated phonetic transcripts with the correct standard phonetic prompts. Alignment of acoustic and articulatory feature frames with human labeled transcript was then performed to extract and compare articulatory features from correct pronunciation examples by L1 speakers with the matching incorrect pronunciations by L2 speakers. Statistical comparison of common phoneme substitution errors was done in order to reveal the nature of the associated articulatory errors among L2 speakers. The frames aligned to the correct and incorrect pronunciations were obtained by using speech recognition models for L1 and L2 speaker groups.

3.2 Analysis of Human Annotated Transcripts

The EMA-MAE corpus includes multiple individually annotated transcripts along with a consensus transcript for L2 speakers' speech. For L1 speakers a single trained annotator labeled the audio with its phonetic transcript. The availability of phoneme level transcripts allows us to identify mispronunciation errors within the dataset, which in turn enables evaluation of automatic Mispronunciation Detection and Diagnosis results obtained using an Automatic Speech Recognition (ASR) system. As a first step in this direction, commonly occurring mispronunciation errors from the corpus were identified.

Given the expert labeled phonetic transcript and known standard pronunciation for any word, a simple alignment of the transcript with the standard pronunciation can reveal the mispronunciation. For this task, human annotated transcripts were aligned with standard phonetic prompts, both available in the EMA-MAE database. This alignment provided information on mispronunciation errors. The Levenshtein minimum distance algorithm (I. 1966) was used to perform alignment between the standard prompt and human annotated phonetic transcripts.

Mispronunciation errors can be of three types: substitution, insertion and deletion. Substitution errors occur when an L2 learner substitutes an incorrect phoneme for the target phoneme during speech. This erroneous phoneme may come from the learner's native language phoneme inventory, from the learner's non-native language phoneme inventory, or from a combination or modification of those. Insertion errors occur when an L2 learner inserts an additional phoneme during speech. Deletion errors occur when an L2 learner omits the target phoneme during speech. For Mandarin speakers of English there are some well-established pronunciation errors as discussed in 2.5.3.

3.2.1 Phoneme set for corpus

Table 3-1 shows the phoneme set used in the EMA-MAE corpus. The table includes the IPA and ARPABET symbols for each phoneme in the set along with an illustrative example word and corresponding ARPABET phonetic transcript and the associated place and manner of articulation for consonants and location within the vowel quadrilateral for vowels.

Table 3-1 Phoneme set used in EMA-MAE corpus

IPA	ARPABET	Example	Translation	Articulatory Label
b	B	be	B IY	Bilabial Stop
d	D	dee	D IY	Alvelolar Stop
e	EY	ate	EY T	Front mid
f	F	fee	F IY	Labiodental Friative
g	G	green	G R IY N	Velar stop
h	HH	he	HH IY	Glottal Fricative
i	IY	eat	IY T	Front high
j	Y	yield	Y IY L D	Palatal glide
k	K	key	K IY	Velar Stop
l	L	lee	L IY	Alvelolar liquid
m	M	me	M IY	Bilabial nasal
n	N	knee	N IY	Alveolar nasal
o	OW	oat	OW T	Back mid
p	P	pee	P IY	Bilabial stop
r	R	read	R IY D	Palatal liquid
s	S	sea	S IY	Alveolar Fricative
t	T	tea	T IY	Alveolar Stop
u	UW	two	T UW	Back high
v	V	vee	V IY	Labiodental Fricative
w	W	we	W IY	Bilabial glide
z	Z	zee	Z IY	Alveolar Fricative
æ	AE	at	AE T	Front low
ð	DH	thee	DH IY	Dental Fricative
ŋ	NG	ping	P IH NG	Velar nasal
ɑ	AA	odd	AA D	Mid low
ɔ	AO	ought	AO T	Back low
ə	AX	comma	K AA M AX	Mid mid
ɚ	AXR	letter	L EH T AXR	Mid mid
ɛ	EH	Ed	EH D	Front mid
ɝ	ER	hurt	HH ER T	Mid mid
ɪ	IH	it	IH T	Front high
ʃ	SH	she	SH IY	Palatal Fricative
ʊ	UH	hood	HH UH D	Back high
ʌ	AH	hut	HH AH T	Mid mid
ʒ	Z	zee	Z IY	Alveolar Fricative
ʤ	JH	gee	JH IY	Alveolar affricative
ʧ	CH	cheese	CH IY Z	Alveolar affricative
θ	TH	theta	TH EY T AH	Dental Fricative
aɪ	AY	hide	HH AY D	Mid low
aʊ	AW	cow	K AW	Mid low

3.2.2 Distribution of Phonemes across corpus

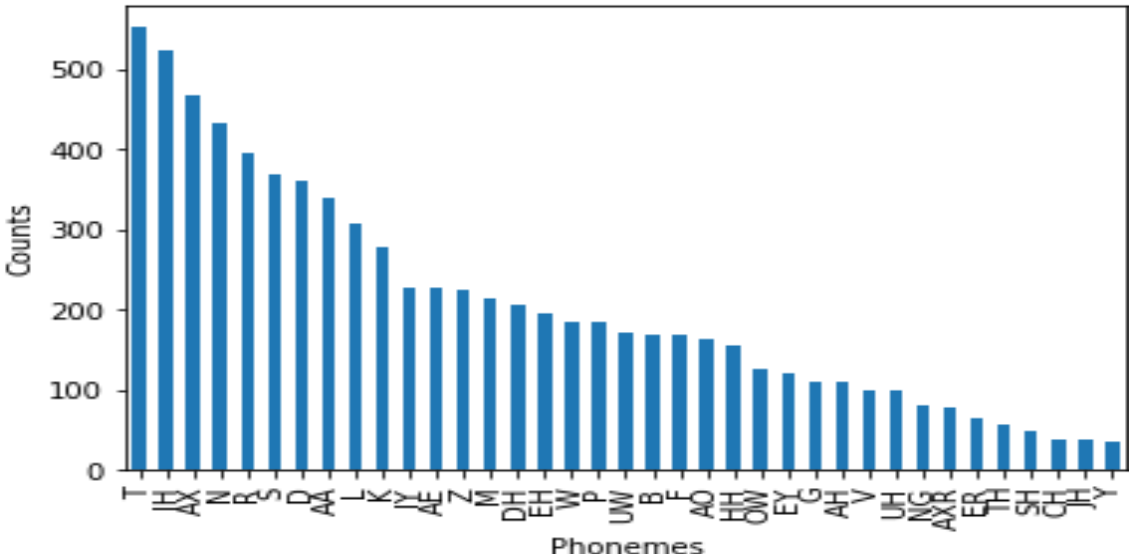

Figure 15 Phoneme count in prompts in the EMA-MAE corpus

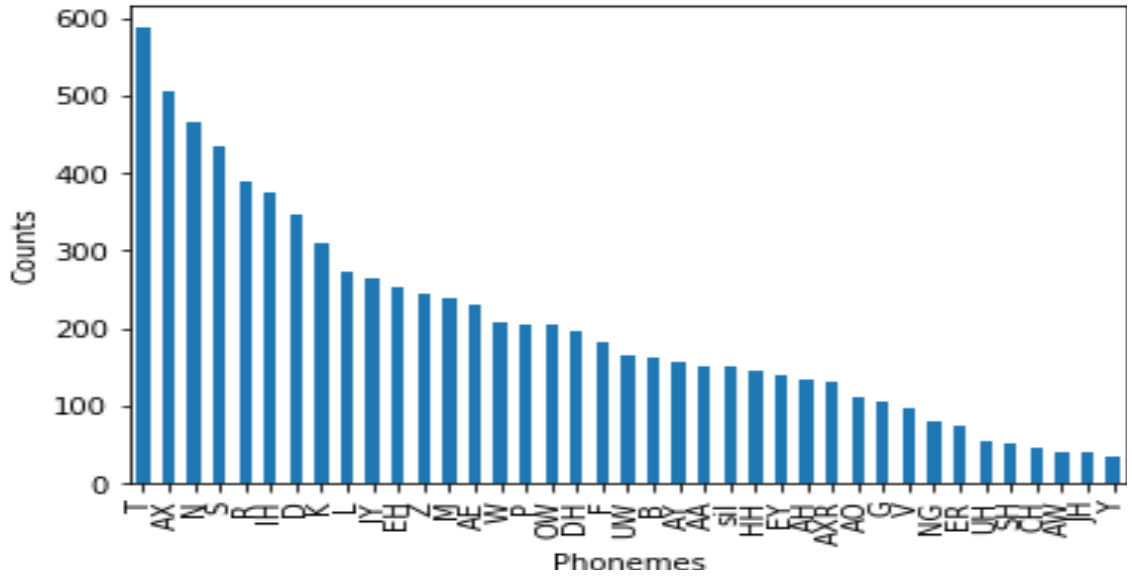

Figure 16 Phoneme count in annotated transcript for Mandarin speaker: 01MBF

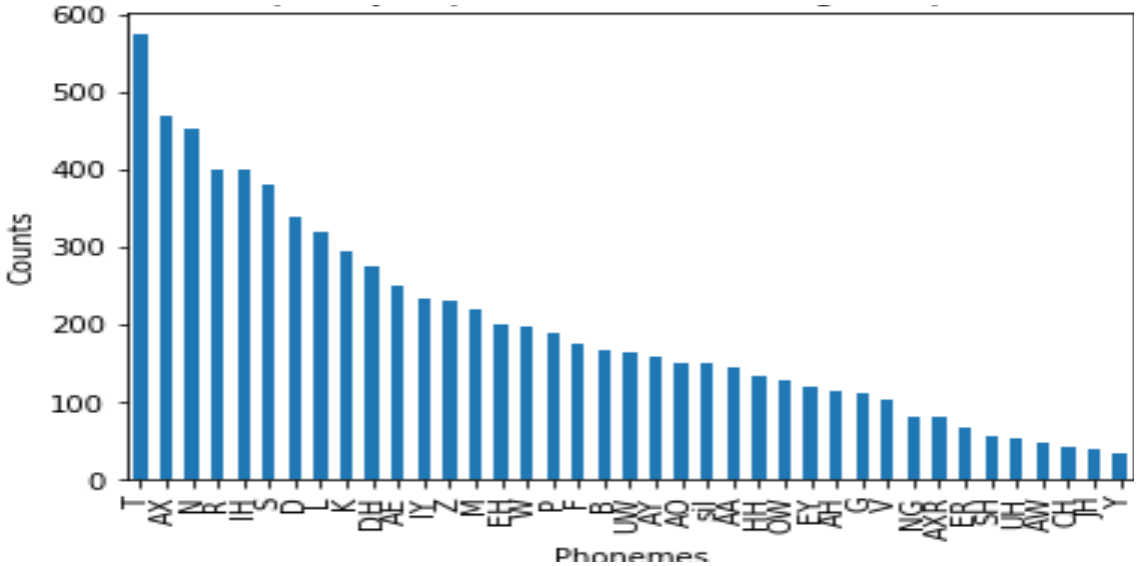

Figure 17 Phoneme count in annotated transcript for L1 speaker: 40ENF

Figure 15, Figure 16 and, Figure 17 provide information on the distribution of phonemes across the corpus. The phoneme distributions for individual L1 and L2 speakers show that there is a noticeable difference in the phoneme count in these two groups, indicating the likelihood of finding patterns of mispronunciation errors for the L2 speaker group.

3.2.3 Results and Discussion

This section discusses the results obtained from human labeled transcripts for L1 and L2 speakers. The goal is to identify common errors and assess the underlying articulatory patterns associated with these errors for different speaker groups.

3.2.3.1 Confusion Matrices: Prompt versus Human annotated Transcript

Alignment of standard phonetic prompts with the corresponding human annotated consensus transcript was performed using the Levenshtein distance algorithm (I. 1966). Correct pronunciations as well as substitution errors were found by analyzing the alignment from this algorithm, focusing explicitly on correct pronunciations and stand-

along substitution errors where the left and right context of the substitution were both correct pronunciations. Insertions and deletions and multi-error sequences were not included.

Results are presented in the form of confusion matrix, with rows representing phonemes that occur in prompts and columns representing the corresponding phoneme that occurred in the transcript. Separate confusion matrices are presented for vowels and for consonants, for readability since nearly all substitutions fall within these categories.

Table 3-2 Vowel Confusion Matrix (Prompt vs. Expert Transcript) for L1 speaker group

Phoneme in expert transcript

Phoneme in Prompt	AE	EY	IY	IH	AX	UW	AXR	EH	AH	OW	AA	AO	UH	ER
AE	3843	20	0	2	11	0	0	3	0	0	0	0	0	0
EY	41	1917	0	1	1	0	0	12	0	0	0	0	1	0
IY	0	0	3681	73	29	0	0	1	0	0	0	0	0	0
IH	0	3	126	5511	71	0	0	36	0	0	2	0	1	0
AX	0	88	26	352	7044	3	57	5	0	0	8	0	0	0
UW	0	0	0	17	83	2659	0	0	2	1	0	0	0	0
AXR	0	0	0	0	2	0	1102	0	0	0	0	0	0	0
EH	0	0	0	1	3	0	0	3251	1	0	0	0	0	0
AH	0	1	0	0	25	0	0	0	1961	0	0	2	2	0
OW	0	0	0	0	15	15	0	0	0	1983	23	11	0	0
AA	5	0	0	2	0	1	0	6	0	6	2402	3	0	0
AO	0	0	0	0	1	0	0	2	0	2	248	2423	0	1
UH	0	1	0	1	0	2	0	0	1	1	1	2	875	0
ER	0	0	1	0	0	0	0	0	0	0	0	0	0	980

Table 3-3 Consonant Confusion Matrix (Prompt vs. Expert Transcript) for L1 speaker group

Phoneme in expert transcript

Phoneme in Prompt	R	P	B	F	HH	W	NG	N	M	K	V	D	DH	S	Z	SH	T	L	CH	G	JH	Y
R	6017	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	2832	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	2371	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	2519	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
HH	0	0	0	0	1929	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	1	0	0	0	0	2909	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0
NG	0	0	0	0	0	0	1346	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0
N	0	0	0	0	0	0	1	6653	1	0	0	1	0	0	0	0	0	0	0	1	0	0
M	0	0	0	0	0	0	0	0	3014	0	1	0	0	1	0	0	0	0	0	0	0	0
K	0	1	0	0	0	0	0	1	0	4398	0	0	0	3	0	0	1	0	0	0	0	0
V	0	0	0	0	0	0	0	1	3	0	1429	1	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	1	0	0	0	1	1	6	5042	0	10	1	0	1	0	0	0	0	0
DH	0	0	0	0	0	0	0	0	0	0	0	2	3274	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	8	5524	19	4	0	0	0	0	0	0
Z	0	0	0	0	0	0	0	1	0	0	0	1	1	22	3282	72	1	0	0	0	0	0
SH	0	0	0	0	1	0	0	0	0	0	0	0	0	5	0	789	0	0	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	8311	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4537	0	0	0	0
CH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	602	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1674	0	0
JH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	608	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	536

Table 3-2 and Table 3-3 show the confusion matrices for vowels and for consonants, respectively, for the L1 speaker group. It is interesting to note that there are some noticeable substitution errors, especially for vowels, even among native speakers. There seem to be some noticeable confusion between /S/ and /Z/ consonant sounds. For vowel sounds, confusion between the sounds [/AE/ and /EY/], [/IY/, /IH/, /AX/] was observed. Other key noticeable vowel substitution errors were /AO/ substituted by /AA/, /AX/ substituted by /EY/, /UW/ substituted by /AX/ and /AX/ substituted by /AXR/. The sound /AX/ which is located in the mid central position of the vowel quadrilateral as presented in Figure 2, seems to have diverse set of errors in both the directions, i.e. /AX/ is substituted by different vowel sounds and for different vowel sounds across the vowel quadrilateral. The confusion between the sounds, /IY/ and /IH/, substitution of /AX/ by

/AXR/ that lie within the same region of vowel quadrilateral can be considered less surprising. In overall summary, it is interesting to see these vowel substitution errors among the L1 speakers. This suggests lack of one standard pronunciation for vowel sounds within a given word for native speakers of English, even when limited to speakers of standard Midwestern American English.

Table 3-4 Vowel Confusion Matrix (Prompt vs. Expert Transcript) for L2 speaker group

Phoneme in expert transcript

Phoneme in Prompt	IY	IH	EY	EH	AX	AE	AXR	AH	AA	AO	OW	ER	UW	UH
IY	2188	362	77	15	5	4	0	0	0	0	2	0	0	0
IH	284	3269	50	48	8	6	0	1	1	3	1	0	0	0
EY	33	11	1266	92	1	30	1	2	3	2	0	0	0	0
EH	56	53	36	1970	5	91	0	10	7	1	2	0	0	0
AX	56	127	17	51	4031	14	47	43	8	9	22	0	8	1
AE	2	21	19	92	13	2145	0	20	78	4	0	1	1	1
AXR	0	1	1	1	120	0	544	0	0	5	12	55	0	1
AH	0	1	3	3	1	13	2	1187	92	6	1	0	3	0
AA	0	0	0	0	9	1	0	11	1303	124	93	1	6	5
AO	0	0	3	3	4	2	0	31	282	1026	55	0	5	2
OW	0	1	0	0	19	1	0	7	24	49	1145	0	56	2
ER	0	2	1	1	13	0	0	17	4	5	11	563	2	40
UW	4	0	0	0	4	0	3	27	12	18	30	0	1755	11
UH	0	0	0	1	0	0	0	1	5	7	12	0	151	356

Table 3-4 presents the confusion matrices for vowel sounds for the L2 speaker group. When compared with the confusion matrix for L1 speakers, it is not surprising that there are a significantly larger number of errors for L2 speakers and these errors are distributed across different types of phoneme substitutions. An interesting observation is that vowel substitutions often happen in symmetric pairs rather than in isolation. This suggests that there was confusion between a pair of phonemes. From the confusion matrix for vowels it can be seen that there are two distinct groups of sounds that are confused with the sound within the group. A group of four vowels [/IY/, /IH/, /EY/, and /EH/] located in

upper-left corner and a group of three vowels [/AA/, /AO/ and /OW/] located in the lower-right region of the vowel quadrilateral are seen in Figure 18. With its location at the center of the vowel quadrilateral, the vowel sound /AX/ seems to have been substituted by the most diverse set of vowel sounds. The patterns of the most common substitution errors for vowels can also be visualized in the vowel quadrilateral presented as follows.

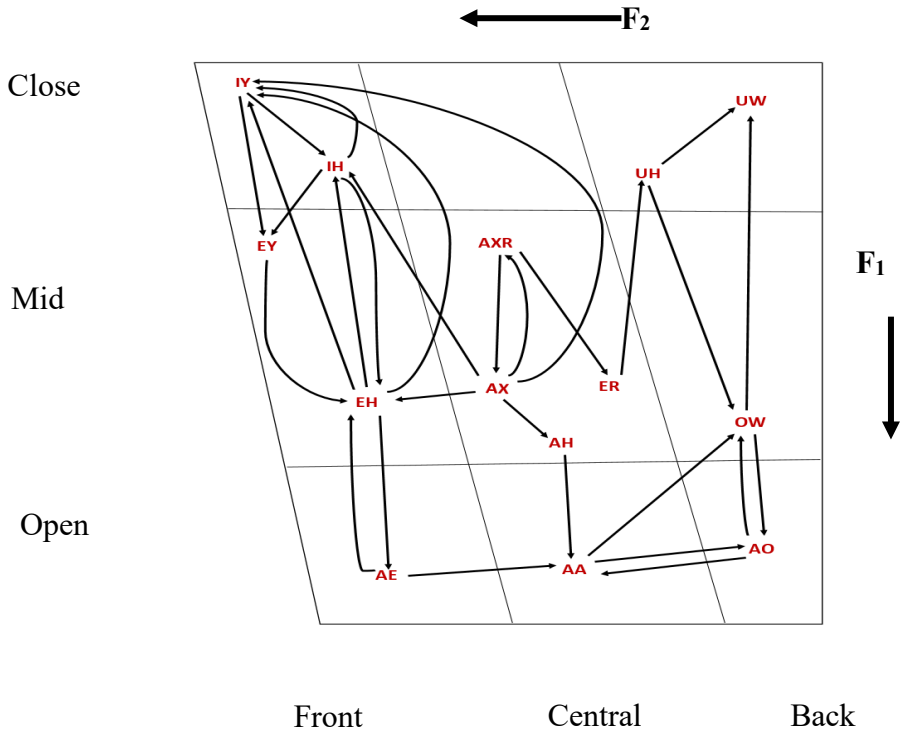

Figure 18 Vowel quadrilateral locations of common (40 occurrences or more) L2 vowel substitution errors

In terms of occurrence count, common confusable vowel pairs include: (/IY/, /IH/), (/AA/, /AO), (/EH/, /AE/), (/OW/ /AO/), (/AXR/, /AX/), and (/EH/, /IH/). The phonemes in confusable pairs (/IY/, /IH/) are both in the front-close category, (/AXR/, /AX/) are in the central-mid category. For the confusable pairs (/EH/, /AE/), (/IH/, /EH/), (/AA/, /AO/), (/OW/, /AO) it can be seen that the confusion happened between the phonemes that are in the adjacent region in the vowel diagram. Apart from the confusable

pairs, most of the isolated vowel substitution errors happen within or between the adjacent regions in the vowel quadrilateral. For example, this includes /UH/ substituted by /UW/, /AX/ substituted by /AH/, /AH/ substituted by /AA/, /AE/ substituted by /AA/, /AX/ substituted by /EH/, and /AXR/ substituted by /ER/. However there are some errors that happen between the regions that are relatively far in the vowel space. Examples of this include /ER/ substituted by /UH/, /AE/ substituted by /IH/, /AX/ substituted by /IY/, and /AX/ substituted by /IH/.

Table 3-5 Consonant Confusion Matrix (Prompt vs. Expert Transcript) for L2 speaker group

Phoneme in expert transcript

Phoneme in Prompt	D	DH	NG	N	Z	S	T	M	SH	L	R	W	V	HH	F	CH	K	G	JH	B	P	Y
D	2511	3	0	1	8	7	37	0	0	1	1	0	3	0	0	1	1	1	4	4	1	0
DH	302	1639	0	4	92	6	3	0	1	3	1	0	0	0	0	0	0	0	2	0	0	1
NG	0	0	772	127	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	2	1	236	3648	0	0	1	49	0	3	2	0	1	1	0	0	0	0	0	0	0	0
Z	2	20	0	1	1818	150	0	0	34	0	0	0	0	0	0	1	0	0	1	0	0	1
S	0	29	0	3	40	3297	0	0	10	0	1	0	0	0	0	1	2	0	0	0	0	0
T	47	2	0	3	1	8	4828	0	1	4	0	0	0	0	2	9	11	0	0	0	0	0
M	0	0	3	31	0	0	0	1689	0	0	1	0	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	0	0	5	0	0	473	0	0	0	0	15	0	6	0	0	0	0	0	0
L	3	0	0	37	0	0	1	1	0	2337	44	13	0	0	0	1	1	0	0	0	0	0
R	1	1	1	2	0	1	0	0	0	31	3046	83	1	1	1	0	0	0	0	0	0	2
W	0	0	0	0	0	0	0	0	0	0	7	1637	7	0	0	0	0	0	0	0	0	0
V	1	0	0	1	0	1	1	1	0	0	0	31	799	0	20	0	0	0	0	1	1	0
HH	0	1	0	0	0	0	0	0	1	0	0	0	0	986	0	0	1	0	0	0	0	0
F	0	0	0	0	0	1	0	0	0	0	0	1	3	1	1431	0	0	0	0	0	3	0
CH	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	351	1	0	0	0	0	0
K	0	0	0	0	0	0	4	0	0	0	0	0	0	1	0	0	2737	4	0	0	2	0
G	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	25	1149	2	0	0	0
JH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	344	0	0	0
B	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1520	30	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	6	1893	0
Y	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	272

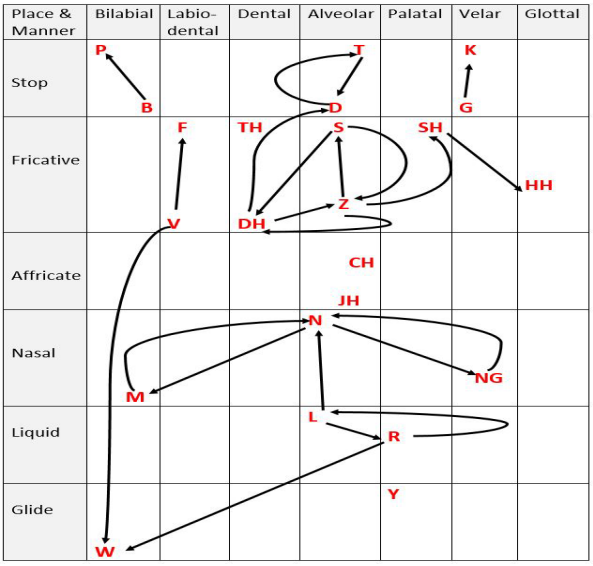

Figure 19 Place and manner of articulation for common (15 occurrences or more) L2 consonant substitution errors

Table 3-5 presents the confusion matrix for the consonants for the L2 speaker group. With the columns representing phonemes present in the expert labeled phonetic transcript of the utterances and rows representing those in the standard prompt for the utterances, the table reveals the common substitution errors related to consonant sounds in L2 speakers. Based on Table 3-5 , Figure 19 provides insights on the consonant pronunciation errors and the place and manner of articulation of the target and erroneously substituted phoneme. From Table 3-5 and Figure 19 it can be seen that the common confusable consonant pairs include: (/M/, /N/), (/N/, /NG/), (/S/, /Z/), (/D/, /T/), (/DH/, /Z/), and (/R/, /L/). It is observed that the majority of substitution errors happen in consonants with either the same place of articulation or the same manner of articulation. This is illustrated by confusable pairs of nasals, (/M/, /N/) and (/N/, /NG/), fricatives (/S/, /Z/) and (/DH/, /Z/) stops (/D/, /T/), and liquids (/R/, /L/). Substitutions that have both the same place and the same manner of articulation include: /B/ substituted by /P/ (both bilabial stops), /G/ by /K/ (both velar stops), and /V/ by /F/ (both labio-dental fricatives), /S/ by

/DH/ (both fricatives), /DH/ by /Z/ (both fricatives), /Z/ by /SH/ (both fricatives), /SH/ by /HH/ (both fricatives) and /L/ by /N/ (both Alveolar). There are also a few interesting substitution errors where the target phoneme is substituted by a phoneme which has both a different place and a different manner of articulation. This includes the palatal liquid /R/ substituted by a bilabial glide /W/, and the labio-dental fricative /V/ substituted by a bilabial glide /W/.

From the confusion matrices for vowels and consonants, the most frequent substitution errors were noted and listed in Table 3-6 and Table 3-7. The information in these tables is essential in terms of identifying the region of interest in large possibilities of substitution errors. These errors are similar to the most common mispronunciation errors in Mandarin speakers of English as reported in the literature and discussed in Section 2.5.3 of this document. Once the common mispronunciations are identified, diagnostic analysis of the errors in articulatory feature space is carried out as discussed in Section 3.3 of this chapter.

Table 3-6 Common Vowel Substitution Errors for L2 speakers

Prompt	Transcript	Correct Pronunciation	Mispronunciation count
IY	IH	2188	362
IH	IY	3269	284
AO	AA	1026	282
UH	UW	356	151
AX	IH	4031	127
AA	AO	1303	124
AXR	AX	544	120
AA	OW	1303	93
AH	AA	1187	92
EY	EH	1266	92
AE	EH	2145	92
EH	AE	1970	91
AE	AA	2145	78
IY	EY	2188	77
AX	IY	4031	56
EH	IY	1970	56
OW	UW	1145	56
D	AX	2511	56
AO	OW	1026	55
AXR	ER	544	55
EH	IH	1970	53
AX	EH	4031	51
IH	EY	3269	50
OW	AO	1145	49
IH	EH	3269	48
AX	AXR	4031	47
AX	AH	4031	43
ER	UH	563	40
EH	EY	1970	36
EY	IY	1266	33
AO	AH	1026	31
UW	OW	1755	30
EY	AE	1266	30
UW	AH	1755	27
T	AX	4828	25
OW	AA	1145	24
AX	OW	4031	22
AE	IH	2145	21
AE	AH	2145	20

Table 3-7 Common Consonant substitution errors for L2 speakers

Prompt	Transcript	Correct Pronunciation count	Mispronunciation count
DH	D	1639	302
N	NG	3648	236
Z	S	1818	150
NG	N	772	127
DH	Z	1639	92
R	W	3046	83
N	M	3648	49
T	D	4828	47
L	R	2337	44
S	Z	3297	40
L	N	2337	37
D	T	2511	37
Z	SH	1818	34
R	L	3046	31
M	N	1689	31
V	W	799	31
B	P	1520	30
S	DH	3297	29
G	K	1149	25
V	F	799	20
Z	DH	1818	20
SH	HH	473	15
L	W	2337	13
T	K	4828	11

3.2.3.2 Words ending with consonants

Mandarin speakers of English often have difficulty with words ending with consonants because the Mandarin language does not include ending consonants. Therefore a comparative experiment for the words ending with consonants was carried out for L1 and L2 speakers, including both Shanghai and Beijing dialects.

Table 3-8 Averaged Error distribution for words ending with consonants

Speaker Group	End-consonants	Substitution	Deletion	Insertion
Shanghai dialect(L2)	1159	133(11%)	140(12%)	2(0%)
Beijing dialect (L2)	1153	112(9%)	110(9%)	7(0%)
L1	1196	9(0%)	13(1%)	1(0%)

As evident from the Table 3-8, the mispronunciation error for the words ending with consonants is around 20%. More than the error rate itself, it is interesting to see that the substitution and deletion are almost equally likely. This suggests that the Mandarin speakers can either substitute the end consonants with some erroneous phoneme or drop the end consonants. Moreover, the error rate and its distribution for words ending with consonants was found to be similar across the two dialects, Shanghai and Beijing.

3.2.3.3 Error distribution across the two dialects

Mispronunciation error analysis across L2 speakers (9 Shanghai dialect and 11 Beijing dialect) was conducted to see the variability in mispronunciations among speakers and difference in mispronunciation rate between the dialects. Table 3-9 and Table 3-10 show the error percentage for the common mispronunciation errors for each dialect groups. Error percentage as computed here was the percentage of mispronunciation errors per the total occurrence of the target prompt phoneme for the substitution error under consideration. The information of error percentage across the two Mandarin dialects for consonants and vowels is depicted in Figure 20 and Figure 21 respectively. In terms of error count and percentage, the most common consonant substitution errors (/DH/ substituted by /D/, /N/ by /NG/, /Z/ by /S/, /NG/ by /N/, and /DH/ by /Z/) reveal that the speakers with Shanghai dialect have a higher error rate as compared to that with Beijing dialect. /N/ substituted by /NG/ has a 6% higher error rate for the Shanghai dialect group over the Beijing group. Similarly, /DH/ substituted by /D/ and /DH/ substituted by /Z/

were both 5% higher for the Shanghai dialect group, /NG/ substituted by /N/ was 5% higher, and /R/ substituted by /W/ and /L/ substituted by /N/ were 3% higher. This suggests that Shanghai dialect speakers are more prone to make certain types of consonant mispronunciations.

For vowels, the two substitution errors for which there is significant error rate difference between the Mandarin dialects are, /AXR/ substituted by /AX/ (with 24% higher error rate for Shanghai dialect) and /ER/ substituted by /UH/ (with 8% higher error rate for Shanghai dialect). /UH/ substituted by /UW/ error had a 6% higher error rate. Unlike for consonant sounds, there is no general trend of vowel mispronunciations seen across the dialects.

Table 3-9 Consonant substitution errors with count greater than 10 for Mandarin speakers with Beijing and Shanghai dialects

Prompt phoneme	Labeled phoneme	Error notation	Mispronunciation Count	Mispronunciation count for Beijing	Mispronunciation count for Shanghai	Prompt phoneme count for Beijing	Prompt phoneme count for Shanghai	error% for Beijing dialect	error% for Shanghai dialect
DH	D	DH_D	302	137	165	1098	957	12	17
N	NG	N_NG	236	71	165	2095	1852	3	9
Z	S	Z_S	150	71	79	1072	960	7	8
NG	N	NG_N	127	56	71	468	436	12	16
DH	Z	DH_Z	92	24	68	1098	957	2	7
R	W	R_W	83	18	65	1701	1503	1	4
N	M	N_M	49	18	31	2095	1852	1	2
T	D	T_D	47	25	22	2607	2340	1	1
L	R	L_R	44	11	33	1324	1139	1	3
S	Z	S_Z	40	19	21	1778	1607	1	1
L	N	L_N	37	1	36	1324	1139	0	3
D	T	D_T	37	24	13	1405	1239	2	1
Z	SH	Z_SH	34	14	20	1072	960	1	2
V	W	V_W	31	20	11	448	413	4	3
M	N	M_N	31	12	19	855	869	1	2
R	L	R_L	31	5	26	1701	1503	0	2
B	P	B_P	30	18	12	813	740	2	2
S	DH	S_DH	29	14	15	1778	1607	1	1
G	K	G_K	25	14	11	634	543	2	2
V	F	V_F	20	10	10	448	413	2	2
Z	DH	Z_DH	20	15	5	1072	960	1	1
SH	HH	SH_HH	15	9	6	267	232	3	3
L	W	L_W	13	4	9	1324	1139	0	1
T	K	T_K	11	5	6	2607	2340	0	0

Table 3-10 Vowel substitution errors with count greater than 50 for Mandarin speakers with Beijing and Shanghai dialects

Prompt phoneme	Labeled phoneme	Error notation	Mispronunciation Count	Mispronunciation count for Beijing	Mispronunciation count for Shanghai	Prompt phoneme count for Beijing	Prompt phoneme count for Shanghai	error% for Beijing dialect	error% for Shanghai dialect
IY	IH	IY_IH	362	189	173	1405	1248	13	14
IH	IY	IH_IY	284	175	109	1955	1717	9	6
AO	AA	AO_AA	282	141	141	795	632	18	22
UH	UW	UH_UW	151	89	62	286	250	31	25
AX	IH	AX_IH	127	70	57	2360	2080	3	3
AA	AO	AA_AO	124	69	55	904	726	8	8
AXR	AX	AXR_AX	120	20	100	396	349	5	29
AA	OW	AA_OW	93	60	33	904	726	7	5
AH	AA	AH_AA	92	67	25	731	584	9	4
EY	EH	EY_EH	92	43	49	782	659	5	7
AE	EH	AE_EH	92	44	48	1274	1126	3	4
EH	AE	EH_AE	91	59	32	1172	1059	5	3
AE	AA	AE_AA	78	57	21	1274	1126	4	2
IY	EY	IY_EY	77	54	23	1405	1248	4	2
OW	UW	OW_UW	56	30	26	711	596	4	4
EH	IY	EH_IY	56	31	25	1172	1059	3	2
D	AX	D_AX	56	19	37	1405	1239	1	3
AX	IY	AX_IY	56	29	27	2360	2080	1	1
AXR	ER	AXR_ER	55	31	24	396	349	8	7
AO	OW	AO_OW	55	33	22	795	632	4	3
EH	IH	EH_IH	53	24	29	1172	1059	2	3
AX	EH	AX_EH	51	25	26	2360	2080	1	1
IH	EY	IH_EY	50	29	21	1955	1717	1	1

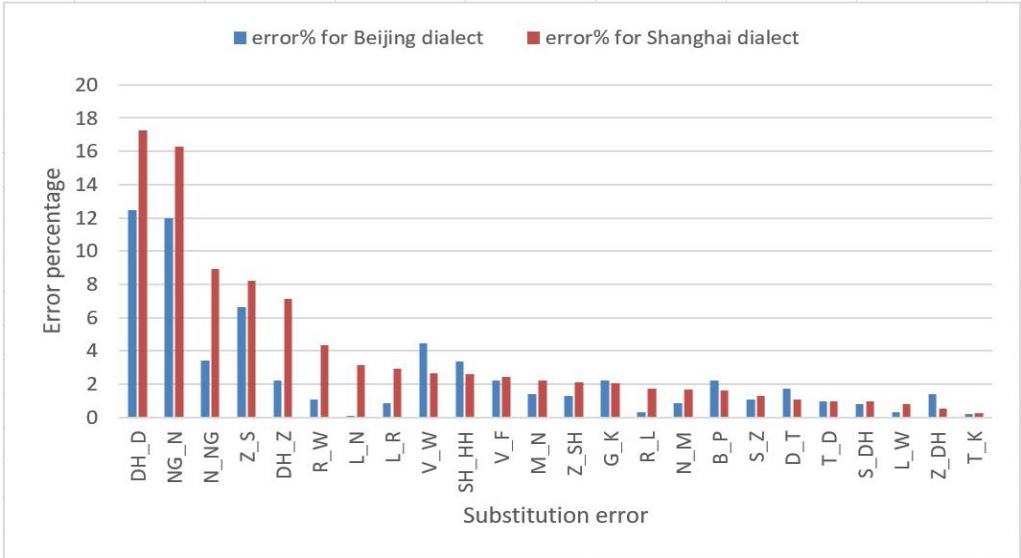

Figure 20 Consonant error percentage for Beijing and Shanghai dialect groups

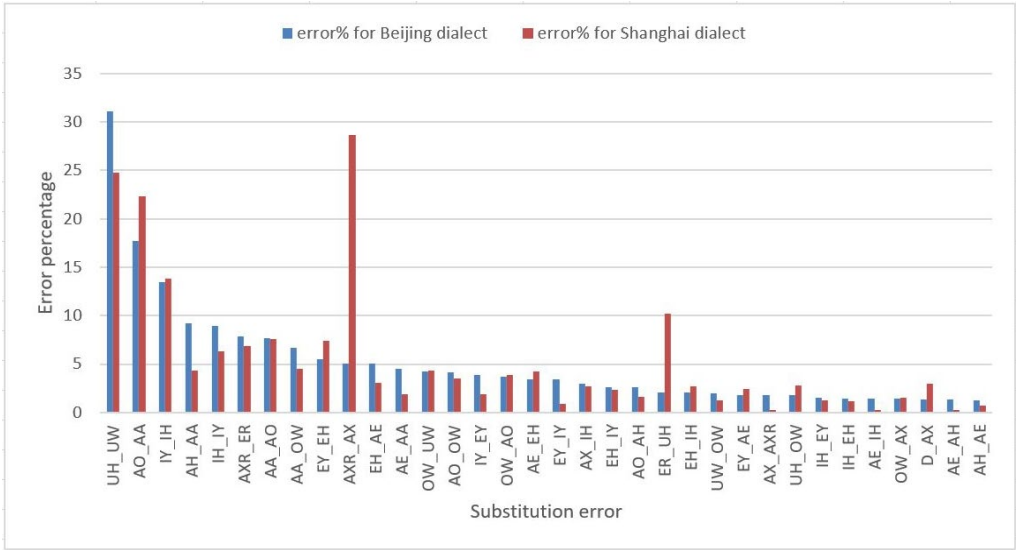

Figure 21 Vowel error percentage for Beijing and Shanghai dialect groups

3.3 Experimental Method for Diagnostic analysis of Mispronunciation Errors

The goal of this study is to increase our understanding of the relationship between articulatory patterns and pronunciation errors. In order to do this, we examine which articulatory features show the most significant differences between incorrect L2 pronunciations and the corresponding correct L1 pronunciation. As discussed in Section 3.2, the most frequently occurring mispronunciation errors for the EMAMAE corpus were identified by comparing the consensus human transcribed phoneme level transcripts with the English prompts. The focus is on substitution errors since they allow for a direct pronunciation comparison between L1 and L2 speakers. Only isolated phoneme substitution errors where left and right phonemes are correctly pronounced are considered. This allows for minimal context variability and a direct focus on the erroneous articulatory pattern of one specific error under consideration. Once the common mispronunciation errors are known, articulatory frames corresponding to the correctly pronounced target phoneme by L1 speakers and the frames where the target phoneme is incorrectly substituted by an erroneous phoneme by L2 speakers are extracted. Alignment of the frames of the utterance to its corresponding phonetic transcript was achieved by using speech recognition models in alignment mode.

3.3.1 Speech recognition model in Alignment mode

Separate GMM-HMM acoustic models were trained for L1 and L2 speakers using the speech recognition toolkit Kaldi (Povey, Ghoshal et al. 2011). Acoustic features consisted of 39 (MFCC-delta-delta) coefficients with cepstral mean normalization. As discussed in 2.3.2.1, in order to remove the effect of source or channel variability in speech recognition, cepstral mean and variance normalization (CMVN) was applied. After

training monophone models, LDA-MLLT (Linear Discriminant Analysis – Maximum Likelihood Linear Transform) based training was performed. LDA-MLLT, as discussed in Section 2.3.1.3, improves computational complexity and decreases storage requirements by reducing the dimension of the feature space and making the elements in the feature space mutually independent hence allowing the use of a diagonal covariance matrix. Tri-phone acoustic model training used a total of 15000 Gaussian states tied to 2500 leaf states by using decision tree clustering. Moreover, Speaker Adaptive Training (SAT), as discussed in 2.3.2.4, was implemented to reduce the effect on acoustic models due to inter-speaker variability in speech. Finally, fMLLR based estimation as discussed in 2.3.2.3 was used to generated alignments from the quasi-speaker-independent acoustic models.

3.3.2 Feature frames extraction and statistical comparison

The recognition system was operated in alignment mode using consensus transcripts for both L1 and L2 speakers. Based on the alignment, acoustic frames for the phonemes associated with all errors corresponding to each specific type of substitution made by L2 speakers were extracted, and acoustic frames taken from correctly pronounced target segments by L1 speakers were also extracted. Since the acoustic feature frames are synchronized with the articulatory feature frames, the articulatory feature frames corresponding to those acoustic frames were then obtained. In order to minimize co-articulation effects, only the middle 50 percent of the articulatory feature frames were used for computing articulatory configurations, with the beginning and ending 25 percent dropped. The selected articulatory samples were averaged in time to yield a 10-dimensional vector representing the articulatory position for every instance of error (eg. /B/ substituted by /P/ from an L2 speaker) and corresponding template (eg. Correct

pronunciation of /B/ by an L1 speaker). Finally, each correctly pronounced and mispronounced segment of the utterance is represented by a 10-dimensional vector in articulatory feature space. Samples of these 10-dimensional vectors for the correctly pronounced segments from the L1 speaker group and for the mispronounced segments from the L2 speaker group are used for statistical comparison, hence revealing the features that are significantly different between L1 and L2 sample groups.

Statistical comparison was accomplished by running a Welch -t test with critical value of significance $p< 0.001$. For the cases where there are multiple articulatory features having a p value below the 0.001 threshold, the most significant features were selected as the primary contributors, and all features within a pre-selected margin of the primary were considered as secondary contributors. The difference between the means of the L2 and L1 features was then computed for those cases, as a physical indicator of the direction and extent of the articulatory difference.

3.4 Results and Discussion

This section presents the results of diagnostic analysis of the common mispronunciation errors in Mandarin speakers of English in the articulatory space. The values in Table 3-11 and Table 3-12 are the difference in millimeters between the representative mean vector for the mispronounced instances by L2 speaker group ($L2_{mean}$) and for correctly pronounced instances by L1 speaker group ($L1_{mean}$). Only the values for which there is statistical difference between the L2 and L1 speaker group are noted. Among these values, the highlighted values represent the most statistically significant dimensions.

As described in 2.7 the reference origin for the 3-D articulatory space is at the lower mid-incisor, with the positive x-axis representing anterior, positive y-axis representing superior, and the positive z-axis representing right lateral directions. This means that positive values of ($L2_{mean}$-$L1_{mean}$) for features TD_H, TL_H, TA_H, and LP_H represent changes in the anterior direction, and positive values for TD_V, TL_V, and TA_V represent increased height of the respective tongue regions. A positive value for LS_V, and JW_V represent increased vertical lip separation and decreased jaw opening respectively. A negative value for LC represents a lateral widening of the lip corner position.

Based on the signs of the values in Table 3-11 and Table 3-12, Figure 22 shows consonant errors organized by place and manner of articulation, and Figure 23 shows vowel errors organized within the vowel space. Substitutions are shown by arrows, with the target sound at the arrow source and the substituted sound at the arrow end. Each diagram includes notations as to the primary articulatory differences associated with these errors, with direction symbols →, ←, ↑ and ↓ representing forward, backward, upward and downward movement of the articulators relative to the L1 speakers. For the purpose of these visualizations, only the most significant of the articulatory differences associated with substitution errors are included, based on the obtained p value. These error diagnosis diagrams provide an easier visualization of articulatory error movements correlated to the specific types of mispronunciations in L2 speaker group.

Table 3-11 (L2mean-L1mean) for L2 errors in consonants. Highlighted cells represent most statistically significant errors.

error_type	Count	Tongue						Lip			Jaw
		TD_H	TD_V	TL_H	TL_V	TA_H	TA_V	LP_H	LS_V	LC	JW_V
		(-)TD← (+)TD→	(-)TD↓ (+)TD↑	(-)TL← (+)TL→	(-)TL↓ (+)TL↑	(-)TA← (+)TA→	(-)TA↓ (+)TA↑	(-)LP← (+)LP→	(-)LS↓ (+)LS↑	(+)LC← (-)LC→	(-)JW↓ (+)JW↑
DH_subsby_D	302			-2.5	0.4	-3.5	-0.8		-2.1		1.0
N_subsby_NG	236		-1.5		2.6		2.8	-0.5	-1.6		0.8
Z_subsby_S	150				1.5	1.4	-0.8	-0.4	-1.7		
NG_subsby_N	127	-3.2	5.2	-2.5	1.6	-1.7	-2.4		-3.7		3.0
DH_subsby_Z	92	-4.5		-2.0	1.7	-4.3		-0.4	-2.8	0.0	4.1
R_subsby_W	83		4.7		6.5		4.8	1.3	-2.5		3.2
N_subsby_M	49		3.2		4.1			0.7	-7.0		3.1
T_subsby_D	47		3.2						-3.1		
L_subsby_R	44	-12.3	-8.8	-10.6	-2.8	-12.6			-3.7		
S_subsby_Z	40						-2.1		-1.9		
L_subsby_N	37	3.8	-4.6	1.8		5.5	-6.0		-2.6		4.2
D_subsby_T	37				1.9	2.6			-1.9		3.1
Z_subsby_SH	34		-3.6	-3.5			-1.8				
R_subsby_L	31				2.9				-4.1		4.2
M_subsby_N	31						-2.6	-0.7			
V_subsby_W	31	-8.2		-5.3	5.4	-5.9	5.9				
B_subsby_P	30				2.7						
S_subsby_DH	29		2.8		2.9	3.3					
G_subsby_K	25							-0.9	-2.7		
V_subsby_F	20				4.1				-2.0		4.1
Z_subsby_DH	20										
SH_subsby_HH	15				2.3	5.9	5.1	-1.6			-4.0
L_subsby_W	13										
T_subsby_K	11		-5.2								

Table 3-12 (L2mean-L1mean) for L2 errors in vowels. Highlighted cells represent most statistically significant errors.

error_type	Count	Tongue						Lip			Jaw
		TD_H	TD_V	TL_H	TL_V	TA_H	TA_V	LP_H	LS_V	LC	JW_V
		(-)TD← (+)TD→	(-)TD↓ (+)TD↑	(-)TL← (+)TL→	(-)TL↓ (+)TL↑	(-)TA← (+)TA→	(-)TA↓ (+)TA↑	(-)LP← (+)LP→	(-)LS↓ (+)LS↑	(+)LC← (-)LC→	(-)JW↓ (+)JW↑
IY_subsby_IH	362	-2.3	1.8	-1.2	0.9	-1.6			-2.6		1.4
IH_subsby_IY	284	1.1	-3.1	1.1		1.8	-1.2		-1.3		1.3
AO_subsby_AA	282	-2.4	3.6	-2.1	4.7	-2.1	3.2		-1.1		1.4
UH_subsby_UW	151	-2.3		-1.5	3.5		2.3	1.6	-4.4		1.7
AX_subsby_IH	127						-3.2	-0.5		0.0	1.5
AA_subsby_AO	124	-6.5		-4.1	6.7	-3.8	2.7	1.5	-6.4	0.0	3.0
AXR_subsby_AX	120		7.0		6.0	3.0	4.7		-1.2		2.5
AA_subsby_OW	93	-6.8		-5.3		-5.6		1.6	-8.6		6.2
AH_subsby_AA	92	-2.2	3.0	-1.9	4.5		2.5	-0.7	2.8		
EY_subsby_EH	92	-4.1	5.1	-2.5	4.4		3.1		-2.6		
AE_subsby_EH	92	-3.2		-2.1	1.7		-1.8		-4.0		2.7
EH_subsby_AE	91	2.4	2.1		2.4			-0.7			-2.2
AE_subsby_AA	78	-10.9		-6.8	5.7	-6.8			-3.5		
IY_subsby_EY	77	-5.0		-2.1	0.9	-2.7	1.9				
AX_subsby_IY	56		-2.9		-1.5	2.7	-3.6			0.0	1.5
EH_subsby_IY	56		-5.2		-1.4	3.4	-5.3	0.8	-5.1		4.5
OW_subsby_UW	56		-3.3		-3.0		-6.7		-3.5	0.0	4.7
AO_subsby_OW	55	-4.0			4.1				-3.8		2.4
AXR_subsby_ER	55	-5.7	3.6	-5.9		-6.2					
EH_subsby_IH	53		-2.9				-5.8	0.7	-6.3	0.0	5.2
AX_subsby_EH	51				2.7				1.9		
IH_subsby_EY	50										
OW_subsby_AO	49	-7.3		-5.9	4.9	-4.2					
IH_subsby_EH	48	-4.5	2.5	-2.4	3.4		3.7				
AX_subsby_AXR	47	-7.2		-9.1	-3.2	-9.9				0.0	2.5
AX_subsby_AH	43	-6.2	3.7	-5.6	6.7	-5.1	7.3				
ER_subsby_UH	40				6.6		7.8				2.2
EH_subsby_EY	36		-3.3				-3.9		-3.8		4.1
EY_subsby_IY	33								-4.9	0.0	2.9
AO_subsby_AH	31				5.6			-0.9			
UW_subsby_OW	30	-7.6	8.1	-6.2	9.8	-5.8	9.6				-2.2
EY_subsby_AE	30		4.8		3.8						
UW_subsby_AH	27	-7.4	7.0	-4.8	7.8	-4.0	5.7		3.0		-2.6
OW_subsby_AA	24							-1.3			
AX_subsby_OW	22	-8.7		-7.5	3.3	-7.5		1.3			
AE_subsby_IH	21			2.9		3.7	-6.9		-6.6		3.0
AE_subsby_AH	20	-7.1		-5.1					-4.0		

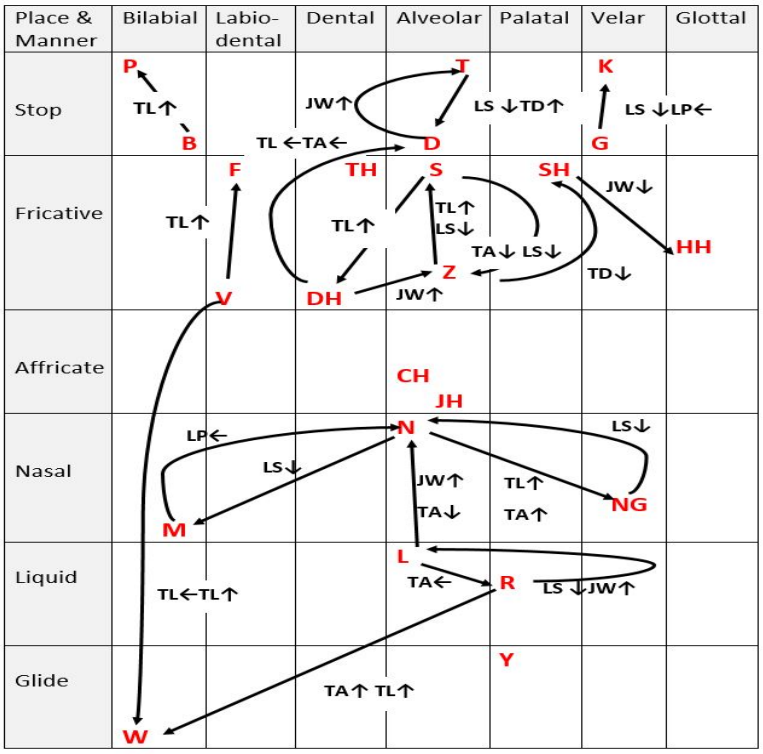

Figure 22 Diagnostic articulatory errors for consonant substitution errors occurring more than 15 times

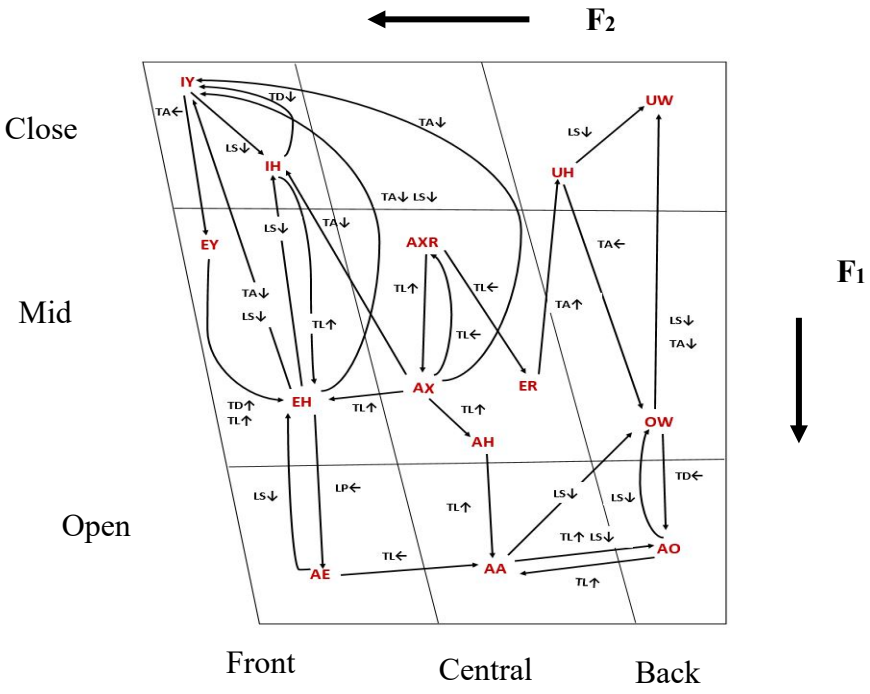

Figure 23 Diagnostic articulatory errors for vowel substitution errors occurring more than 40 times

The data in Table 3-11 and Table 3-12 show that there are a relatively small number of extremely common errors. For consonants this includes substituting DH with either D or Z, confusion between N and NG, and substituting Z with S. For vowels, this includes confusion between IH and IY, between AO and AA, substituting AXR with AX, substituting AX with IH, and substituting UH with UW.

From the numeric data, it can be seen that L2 speakers tend to have errors within a small group of articulatory features, often in the same direction. This includes too large height at tongue lateral ($TL_V\uparrow$), too far posterior position of tongue lateral ($TL_H\leftarrow$), errors in position of the tongue apex (TA_H and TA_V, all directions), too small of a vertical lip separation ($LS_V\downarrow$), and too small of a jaw opening ($JW_V\uparrow$). These erroneous articulatory patterns irrespective of the error type, suggest that L2 speakers have some generic erroneous articulatory movements that get translated to all types of mispronunciations.

This is an interesting observation and useful in the sense that it can be used towards providing some generalized feedback for improving pronunciations of Mandarin speakers of English.

The most significant finding from looking at the tables and associated diagnostic diagrams may be the lack of symmetry in terms of the articulatory differences. There are several pairs of consonants and vowels which are commonly substituted for each other (Consonants: T and D, S and Z, M and N, N and NG, L and R; Vowels: IY and IH, IH and EH, EH and AE, AO and AA, AX and AXR, OW and AO). It might be expected that articulatory differences in one direction would pair with an articulatory difference in the opposite direction for the opposite error, but this is not the case. In some cases completely different articulators are involved, and in some cases the same articulators show differences, but the differences are in the same direction in each case which is quite interesting. This suggests that the articulators with the most significant differences in positioning between L1 and L2 are not necessarily the only factors causing the associated pronunciation error.

3.5 Conclusion

This study has presented a detailed analysis of common substitution errors for Mandarin speakers of English, including a statistical comparison of articulatory features for these errors with respect to native speakers of English. The diagnostic errors as well as the associated difference in the mean articulatory configuration give information about the primary and secondary contributors and their extent for each specific mispronunciation type. It is found that there are specific articulatory differences, including increased tongue

lateral height, too far posterior position of tongue lateral, multiple tongue apex positioning errors, and reduced lip separation and jaw opening, that are seen across many types of mispronunciation for Mandarin L2 speakers. Diagnostic error charts indicate that the articulatory errors depend on each specific substitution error and are not consistent within consonant place or manner of articulation or within regions of the vowel space.

CHAPTER 4. AUTOMATIC MISPRONUNCIATION DETECTION AND DIAGNOSIS (MDD) SYSTEMS

4.1 Overview

This chapter outlines a series of experiments performed with Automatic Speech Recognition (ASR) based Mispronunciation Detection and Diagnosis (MDD) systems, across multiple system configurations and using acoustic and articulatory features. Input features for these MDD systems included acoustic, articulatory and concatenated acoustic and articulatory features. System architectures included GMM-HMM, DNN, and RNN based ASR engines for the MDD system. To evaluate the ability of the ASR systems to detect and diagnose pronunciation errors, the recognized sequence of phonemes generated by the ASR models were aligned with human labeled phonetic transcripts as well as with the original phoneme level prompts and used to determine MDD accuracy of the ASR system relative to the consensus human transcripts. The goal of this experiment was to assess the ability of speech recognition systems to detect and diagnose the common pronunciation errors seen in non-native speakers (L2) of English.

4.2 Experimental Method

4.2.1 Phoneme set

The EMA-MAE corpus uses 39 phonemes in its phonetic transcripts and prompts, including 24 consonants and 15 vowels. Due to insufficient counts for building acoustic models for the phonemes TH and ZH, the phoneme pairs TH/DH and ZH/Z were combined into DH and Z, respectively. This led to a reduced set of 37 phonemes, with 22 consonants and 15 vowels. The same set was used in the lexicon for building phoneme recognizer

models in Kaldi. Both IPA and ARPABET transcripts are available in the dataset, but for these experiments the alignment of the phonemic sequences was performed using IPA symbols because IPA's use of single-character representations made it easier to generate a parallel alignment between prompts, human labeled transcripts and recognized phoneme sequences.

An additional evaluation based on phonetic subgroups was also performed. To implement this, the 37 phonemes were grouped by place and manner of articulation for consonants and by region of the vowel space for vowels, into 24 phoneme groups as described in Table 4-1. In this evaluation, the original phonemes in the prompts, transcripts, and ASR generated phonetic transcript were replaced by a designated representative phoneme from the phonemes in that group. This converted aligned file with 24 unique phonemes was then used to calculate the same MDD metrics used for the original 37 phoneme set.

Table 4-1 37 Phonemes grouped into 24 articulatory groups based on place and manner of articulation for consonants and location in vowel space for vowels

ARPABET	Articulatory Label
B,P	Bilabial Stop
D,T	Alvelolar Stop
G,K	Velar Stop
V,F	Labiodental Friative
DH	Dental Fricative
Z,S	Alveolar Fricative
SH	Palatal Fricative
HH	Glottal Fricative
CH,JH	Alveolar Affricative
M	Bilabial Nasal
N	Alveolar Nasal
NG	Velar Nasal
L	Alveolar Liquid
R	Palatal liquid
W	Bilabial Glide
Y	Palatal Glide
IY,IH	Front High
EY,EH	Front Mid
AE	Front Low
AY,AA	Mid Low
AO	Back Low
OW	Back Mid
AX,AXR,ER,AH	Mid Mid
UH,UW	Back High

4.2.2 Train/Validation/Test split

For the purpose of training, validation and testing of ASR models, the EMA-MAE corpus was split into a Training set, Validation set and Test set in a 70:10:20 proportion. Out of 40 speakers in the overall corpus, data from 28 speakers (20 L1 speakers and 8 L2 speakers) were used for training, data from 4 L2 speakers was used for validation and data from 8 L2 speakers was used for testing. The models were trained on training data set, validated on validation dataset and finally tested on test dataset. The speaker groups chosen

within the splits were gender and dialect balanced. The details of the list of speakers split into these three set is presented as follows.

Training: 03MBM, 05ENF, 06ENM, 07ENF, 09ENF, 12MSF, 13MBF, 15ENM, 16ENM, 17ENF, 18ENF, 19ENM, 20MBF, 21ENF, 23MBM, 24MSF, 26MSM, 28ENF, 30MSM, 30MSM, 32ENM, 33ENM, 34ENM, 35ENM, 36ENF, 37ENF, 38ENM, 39ENM, 40ENF = 20 L1 speakers(10 Male: ENM and 10 Female: ENF) + 8 L2 speakers (2 Male Beijing dialect : MBM , 2 Female Beijing dialect : MBF, 2 Male Shanghai dialect : MSM, 2 Female Shanghai dialect : MSF)

Validation: 08MBM, 14MSF, 22MBF, 25MSM = 4 L2 speakers (1 Male Beijing dialect, 1 Female Beijing dialect, 1 Male Shanghai dialect, 1 Female Shanghai dialect)

Test: 01MBF, 02MBF, 04MSF, 10MSM, 11MBF, 27MSM, 29MBM, 31MBM (2 Male Beijing dialect, 3 Female Beijing dialect, 2 Male Shanghai dialect, 1 Female Shanghai dialect)

4.2.3 Input Features

To evaluate the performance of the ASR based MDD systems for different feature types, three different features were used for training and testing the ASR models: acoustic, articulatory and combined. Acoustic features were extracted at a frame rate of 10ms. The articulatory features were computed from the original kinematic data sampled at 400 Hz (2.5 ms) but then down-sampled by factor of 4 to a matching acoustic frame rate of 10ms. Therefore, the acoustic and articulatory features were frame-synchronized. Acoustic features consisted of 39 (MFCC; delta; delta-delta) coefficients. The frame-synchronized articulatory features included a 10-dimensional articulatory feature vector obtained from converting the raw articulatory sensor data to 10 articulatory features, as described in the

equations in Table 2-2. Along with 10 static articulatory features, their delta and delta-delta coefficients were obtained to get a 30-dimensional articulatory feature vector. The combined feature vector is the concatenation of acoustic and articulatory feature vector together to form a 69-dimensional feature vector.

4.2.4 GMM-HMM models

The open source Kaldi speech recognition toolkit (Povey, Ghoshal et al. 2011) was used to build baseline GMM-HMM based ASR models for implementing MDD. Separate models were built for each of the input feature types. Mean and variance normalization was carried out on the features to reduce differences in feature representation between speakers and to reduce the influence of background noise. After training monophone models, LDA-MLLT (Linear Discriminant Analysis – Maximum Likelihood Linear Transform and Speaker Adaptive Training (SAT) was performed. Triphone models were then created and trained with a total of 15000 Gaussian states tied to 2500 leaves states by using Decision tree based clustering technique. A Universal Background Model (UBM) was trained by using the Gaussians from the trained tri-phone HMM/GMM. The UBM model used a total of 400 final Gaussian states. The trained UBM model was then used to train Subspace Gaussian mixture model (SGMM) (Povey, Burget et al. 2011). In SGMM, all phonetic states share a common Gaussian mixture model but the means and the mixture weights differ in a subspace of the overall parameter space. SGMM here was implemented using 9000 Gaussian states tied to 7000 leave states.

To assess the performance of the ASR models for different feature combinations without a word based language model, a bigram phonetic language model trained only from the phonetic transcripts in the training set was used. Therefore the lexicon used only

contained phonemes. This bigram phone level language model was used for all experiments.

Phoneme to frame alignments for training, validation and test data were also generated from the trained Triphone model discussed in Section 4.2.4. Additionally, the feature space Maximum Likelihood Linear Regression (fMLLR) transform was applied to the features in the corpus. These alignments and features were used to train and test the DNN based models using the open-source toolkit Pytorch-Kaldi.

4.2.5 DNN based models

Pytorch-kaldi (Ravanelli, Parcollet et al. 2019) provides a very convenient framework to train ASR models using different DNN architectures. DNN training is accomplished using the alignments and decoding graph obtained from Kaldi, allowing users to experiment with different DNN networks to generate context dependent phone state posterior probabilities. Baseline configuration files provided in the Pytorch-kaldi repository for the common speech database like TIMIT, Librispeech and DIRHA were used as reference to experiment and choose appropriate configuration files for EMA-MAE database. Four different network types: Multi-layered Perception (MLP), Long short-term Memory (LSTM), Gated Recurrent Unit (GRU) and Light Gated Recurrent Unit (liGRU) were used to evaluate their individual performance for different input features. Among these different architectures, the light GRU based model trained on fMLLR transformed combined features produced the best results. The details of the configuration of this architecture is presented in Appendix 0 of this document.

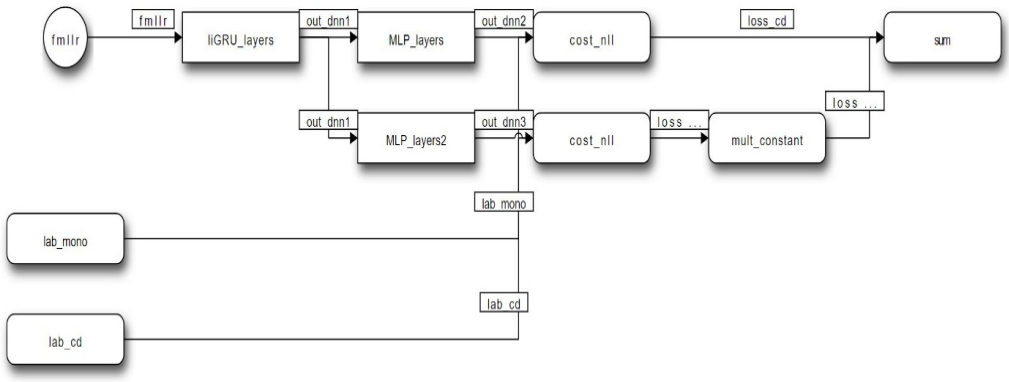

Figure 24 Pytorch-kaldi generated image of Architecture of the best model: light GRU with fmllr as input features

Figure 24 shows the layout of the light GRU based architecture with fMLLR transformed input features. The light GRU here was bidirectional with five layers containing 550 cells each. The activation function for all the layers in the light GRU unit was Relu. The architecture applies monophone regularization (Bell, Swietojanski et al. 2016). To implement this, a multi-task learning strategy was adopted by means of two softmax classifiers: the first one to estimate context-dependent states, while the second one to predict monophone targets.

As depicted in Figure 24 the output from the last layer of light GRU (out_dnn1) was fed to two different MLP layers: MLP_layers and MLP_layers2. Both of these single layered networks used softmax as the activation function. MLP_layers has the number of nodes set to the number of context dependent pdf-ids in the GMM-HMM based graph trained in Kaldi. This can be achieved by setting the variable "dnn_lay" of the MLP_layers architecture to the option "N_out_lab_cd" available in the configuration file for Pytorch-Kaldi. The number of nodes in MLP_layers was set to be equal to the number of monophone target phone states (i.e 166) by setting the variable " dnn_lay" of the MLP_layers2 architecture to the option "N_out_lab_mono". "lab_mono" and "lab_cd" as

shown in Figure 24 are the locations that store alignments and labels used for computing losses for monophone and context dependent phone state targets respectively. The loss function used here was Negative log-likelihood (NLL). Finally, the loss for context dependent targets and the loss for monophone targets multiplied by a multiplication factor (1.0) are summed together to get the final loss for the overall architecture. The section in the configuration file for the model, illustrating the aforementioned computations is presented as follows.

```
[model]
model_proto = proto/model.proto
model = out_dnn1= compute (liGRU_layers, fmllr)
        out_dnn2= compute (MLP_layers, out_dnn1)
        out_dnn3= compute (MLP_layers2, out_dnn1)
        loss_mono= cost_nll (out_dnn3, lab_mono)
        loss_mono_w=mult_constant (loss_mono, 1.0)
        loss_cd=cost_nll (out_dnn2, lab_cd)
        loss_final =sum (loss_cd, loss_mono_w)
        err_final =cost_err (out_dnn2, lab_cd)
```

4.2.6 MDD metrics calculation

For each utterance in the test dataset, alignment of the phonetic transcripts of the prompts against manually labeled consensus phonetic transcripts and ASR generated phonetic transcripts was performed using the Levenshtein minimum edit distance algorithm (I. 1966) . True Acceptance of correct pronunciation (TA), False Acceptance as correct pronunciation (FA), True Rejection as mispronunciation (TR), False Rejection as

mispronunciation (FR), Correct Diagnosis of mispronunciation (CD) and Diagnosis Error of mispronunciation (DE), as defined in more detail in Table 4-2 were counted based on the following rules:

If phoneme in [(prompt = = ground truth and (ASR = = prompt)]:

$$TA = TA + 1$$

If phoneme in [(prompt = = ground truth) and (ASR ≠ prompt)]:

$$FR = FR + 1$$

If phoneme in [(prompt ≠ ground truth) and (ASR = = prompt)]:

$$FA = FA + 1$$

If phoneme in [(prompt ≠ ground truth) and (ASR ≠ prompt)]:

$$TR = TR + 1$$

If phoneme in [(prompt ≠ ground truth) and (ASR ≠ prompt) and (ASR = = ground truth)]:

$$CD = CD + 1$$

If phoneme [(prompt ≠ ground truth) and (ASR ≠ prompt) and (ASR ≠ ground truth)]:

$$DE = DE + 1$$

Table 4-2 Definitions in the hierarchical evaluation used for MDD metrics calculation (Li, Qian et al. 2016)

		ASR Output	
		Correct Pronunciation	Mispronunciation
Expert transcription	Correct Pronunciation	TA	FR
	Mispronunciation	FA	TR (CD/DE)

Once the count for TA, FA, FR, TR, CD and DE is obtained, the MDD metrics used for evaluation of performance of different ASR models are calculated as below:

$$precision = \frac{TR}{TR + FR} \tag{24}$$

$$recall = \frac{TR}{TR + FA} \tag{25}$$

$$F_{measure} = \frac{2 * precision * recall}{precision + recall} \tag{26}$$

$$DetectionAccuracy(DetAcc) = \frac{TA + TR}{TA + FR + FA + TR} \tag{27}$$

$$DiagnosticAccuracy(DiagAcc) = \frac{CD}{CD + DE} \tag{28}$$

$$FalserejectionRate = \frac{FR}{TA + FR} \tag{29}$$

For MDD systems, Detection accuracy is a measure of whether the system correctly identified when there was a mispronunciation. Diagnostic Accuracy is the measure of the ability of the MDD system to correctly identify the type of mispronunciation error. Precision is the ratio of correctly detected mispronunciation errors to the total predicted mispronunciation errors. Higher precision would mean among the system predicted mispronunciation errors, there is a high probability that those errors are actual errors. Recall here is the ratio of correctly detected mispronunciation errors to the total actual mispronunciation errors. A high recall would mean that the system performed well in capturing most of the errors that are actually present. F score is the weighted average of precision and recall. This score takes into account both the precision and recall of the system. The False Rejection Rate (FRR) is the measure of how often the system falsely classified a phoneme as mispronounced among the total correctly pronounced ones.

4.3 Results and Discussion

This section discusses the results for the various experiments related to design of ASR based MDD system.

4.3.1 GMM-HMM based ASR

Table 4-3 ASR Phoneme Error Rate (PER) for the models built in Kaldi: triphone GMM-HMM with LDA +MLLT followed by SAT (tri3) and subspace GMM (SGMM)

Features	tri3	sgmm
Acoustic	33.1	30.41
Articulatory	56.49	54.71
Combined	29.13	26.5

The PER of the ASR model will clearly have a direct correlation with its MDD related performance metrics. The lower the PER the better the system is in terms of

correctly detecting and diagnosing the mispronunciation errors. It is evident from the performance of both tri3 and SGMM (Povey, Burget et al. 2011) models that the combined features, which incorporate the acoustic as well as articulatory features, produced better PER. It is also worth noting that the system utilizing only articulatory features does not perform as well as that using acoustic or combined features, which is typical of articulatory features. Moreover, it can also be noticed that the subspace GMM (SGMM) model, whose training starts with a pre-trained universal background model (UBM) (Povey, Chu et al. 2008) outperforms the triphone GMM-HMM model with LDA +MLLT followed by SAT.

4.3.2 Phoneme recognition Performance of the best DNN based model

System: /storage/subash/pytorch-exps/Combined/liGRU_fmllr/exps/0.1_0.0004/decode_EMAMAE_test_out_dnn2/score_4/ctm

Overall -> Overall
Female -> Female speakers
Male -> Male speakers

SPKR	Overall		Female		Male	
	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE
01mbf	[9056]	16.4	[9056]	16.4		
02mbf	[8775]	10.9	[8775]	10.9		
04msf	[8651]	11.0	[8651]	11.0		
10msm	[8605]	23.8			[8605]	23.8
11mbf	[8861]	16.7	[8861]	16.7		
27msm	[8840]	15.1			[8840]	15.1
29mbm	[8552]	24.9			[8552]	24.9
31mbm	[8860]	21.2			[8860]	21.2
Set Sum/Avg	[70200]	17.5	[35343]	13.8	[34857]	21.2
Mean	[8775]	17.5	[8835]	13.8	[8714]	21.3
StdDev	[165]	5.4	[170]	3.2	[158]	4.4
Median	[8807]	16.6	[8818]	13.7	[8722]	22.5

Figure 25 Details of PER for each speaker in test database for the best model using combined Features

Figure 25 presents the details of PER performance for each speaker present in the test dataset. The results presented are produced by the light GRU based model trained with combined features. The speaker ‘02MBF’ has the least PER of 10.9%, where the speaker ‘29MBM’ has the highest PER of 24.9%. Averaged PER performance of female and male speaker groups is 13.8% and 21.2% respectively.

```
SENTENCE RECOGNITION PERFORMANCE

 sentences                                      5456
 with errors                          47.0%   (2562)

   with substitions                   41.1%   (2244)
   with deletions                     21.0%   (1147)
   with insertions                    17.3%   ( 943)


WORD RECOGNITION PERFORMANCE

Percent Total Error       =   17.5%   (12273)

Percent Correct           =   85.2%   (59788)

Percent Substitution      =    9.6%   (6726)
Percent Deletions         =    5.3%   (3686)
Percent Insertions        =    2.7%   (1861)
Percent Word Accuracy     =   82.5%
```

Figure 26 Performance details for the liGRU based model using Combined Features

Figure 26 presents the breakdown of the recognition performance of the light GRU based model with PER of 17.5%. The substitution error, being the biggest contributor of error, had the most significant contribution of 9.6% error rate for the overall test corpus.

4.3.3 MDD performance results

This section presents the MDD performance results for different combinations of features, DNN based architectures and optimizers. Table 4-4 presents MDD performance metrics evaluated on ASR generated phonetic transcript containing all 37 phonemes. Table 4-5 presents MDD performance metrics evaluated on ASR generated phonetic transcript containing 24 representative phonemes based on groups as presented in Table 4-1.

Table 4-4 MDD performance metrics for different ASR models (Evaluation on transcript sets containing all 37 phonemes)

Feature Type	Architecture, Optimizer	PER	DetAcc	DiagAcc	F_measure	Precision	Recall	FRR
Articulatory_fmllr	MLP, sgd	56.3	47.3	53.3	35.6	22.3	87.7	60.7
Articulatory	liGRU,rmsprop	49.7	53.3	55.2	37.7	24.2	85.6	53.2
Articulatory_fmllr	liGRU,rmsprop	41.8	60.5	58.3	41.3	27.4	83.6	44.0
Articulatory_fmllr	LSTM,rmsprop	40.9	61.1	58.0	41.3	27.6	82.8	43.2
Articulatory_fmllr	LSTM,Adam	37.1	65.0	57.9	43.0	29.4	79.9	38.0
Articulatory_fmllr	liGRU,Adam	34.8	66.6	59.0	44.4	30.7	79.8	36.0
Acoustic_fmllr	MLP, sgd	32	68.9	69.5	48.8	33.7	**88.2**	35.0
Acoustic	liGRU,rmsprop	22.4	77.5	71.8	55.0	41.5	81.8	23.3
Acoustic_fmllr	LSTM,Adam	20.6	79.2	71.4	56.8	43.7	80.9	21.1
Acoustic_fmllr	liGRU,rmsprop	20.6	79.3	72.4	57.0	43.7	81.9	21.2
Acoustic_fmllr	LSTM,rmsprop	20.5	79.4	72.3	57.3	44.0	82.1	21.1
Acoustic_fmllr	liGRU,Adam	18.1	81.5	73.2	59.3	47.1	80.3	18.2
Combined_fmllr	MLP, sgd	29.8	71.3	70.5	50.4	35.4	87.4	31.9
Combined	liGRU,rmsprop	21.4	79.0	72.4	56.6	43.4	81.3	21.5
Combined_fmllr	LSTM,rmsprop	19.6	80.4	74.1	58.2	45.4	81.0	19.7
Combined_fmllr	liGRU,rmsprop	19.5	80.7	74.1	58.5	45.8	81.0	19.4
Combined_fmllr	liGRU, Adam	**17.5**	**82.4**	**75.8**	**60.5**	**48.6**	80.3	**17.2**

Table 4-4 presents the detailed performance results for different architectures and feature combinations. The first information to notice from the table is that the PER and MDD related performance metrics like Detection Accuracy (DetAcc), Diagnostic Accuracy (DiagAcc), F measure, Precision and False rejection rate (FRR) are all improved by inclusion of articulatory features with acoustic features. As for GMM-HMM models, the performance of DNN models trained and tested with articulatory features alone was poorer. The best performing models had PER of 34.8%, 18.1% and 17.5%, Detection Accuracy of 66.6%, 81.5% and 82.4%, Diagnostic Accuracy of 59.0%, 73.2% and 75.8%, and False Rejection Rate of 36.0%, 18.2% and 17.2%, for Articulatory, Acoustic and Combined input features respectively. This shows a 2.6% relative increment of Diagnostic Accuracy between the models using only Acoustic features to that of using Combined features. Similarly, relative increments of 0.9%, 1.5% and 1.2% were seen in Detection Accuracy, Precision and F_measure for the models using combined features over acoustic

features only. A relative decrement of 1.0% in False Rejection Rate (FRR) was seen while using combined features over acoustic features alone. However, one interesting observation was that the model with the highest PER in each feature type category, the MLP based model, produced the highest Recall rate for each of the feature types. With the best recall rates of 87.7%, 88.2% and 87.4% for Articulatory, Acoustic and Combined feature type respectively, the MLP based model trained on fmllr transformed Acoustic features produced the highest recall of 88.2%. It is also interesting to see that the worst performing model in terms of PER and other MDD related metrics with Articulatory features alone has the second best performance in terms of Recall rate. However, this comes with the cost of the highest False Rejection Rate of 60.7%. The model with best PER, DetAcc, DiagAcc and FRR for each of the feature types was the light GRU based architecture trained with Adam optimizer on fmllr transformed input features. The best relative PER improvements of 7%, 2.5%, and 2% for Articulatory, Acoustic and Combined features respectively was seen for the light GRU based architecture trained with Adam Optimizer as compared to the same architecture trained with Rmsprop optimizer. These results verify the noticeable difference in performance of models trained with different optimizers. Another distinct observation that can be drawn from Table 4-4 is that the fmllr based feature transformation significantly helped in improving PER and other MDD related metrics. Considering the light GRU based architecture trained with Rmsprop optimizer as the standard one, while using fmllr transformation, there is relative improvement in PER, DetAcc, DiagAcc, and FRR as follows : 7.9%, 7.2%, 3.1%, and 9.2% respectively for Articulatory features; 1.8%, 1.8%, 0.6%, and 2.1% respectively for Acoustic features and 1.9%, 1.7%, 1.7%, and 2.1% respectively for Combined features.

The key take away from the results presented in Table 4-4 are that the inclusion of articulatory features with acoustic features improves the MDD performance of the system. In addition, fMLLR feature transformation helps, and the light GRU trained with Adam optimizer produced the best overall results.

Table 4-5 MDD performance metrics for different ASR models (Evaluation on transcript sets containing grouped 24 phonemes)

Feature Type	Architecture, Optimizer	DetAcc	DiagAcc	F_measure	Precision	Recall	FRR
Articulatory_fmllr	MLP, sgd	50.7	59.3	33.7	20.8	87.7	55.4
Articulatory	liGRU,rmsprop	56.4	60.7	35.9	22.7	86.1	48.5
Articulatory_fmllr	liGRU,rmsprop	63.7	63.7	40.0	26.1	84.9	39.9
Articulatory_fmllr	LSTM,rmsprop	63.9	63.9	39.8	26.1	83.6	39.3
Articulatory_fmllr	LSTM,Adam	67.4	63.0	41.5	27.8	81.2	34.8
Articulatory_fmllr	liGRU,Adam	68.8	63.8	42.6	29.0	80.8	33.2
Acoustic_fmllr	MLP, sgd	71.9	74.7	47.7	32.6	**88.5**	30.9
Acoustic	liGRU,rmsprop	79.7	76.0	54.1	40.1	83.1	20.9
Acoustic_fmllr	LSTM,Adam	81.1	75.6	55.7	42.1	82.0	19.1
Acoustic_fmllr	liGRU,rmsprop	81.3	76.9	56.1	42.3	83.2	19.1
Acoustic_fmllr	LSTM,rmsprop	81.4	76.7	56.4	42.6	83.3	18.9
Acoustic_fmllr	liGRU,Adam	83.2	77.3	58.2	45.4	81.3	16.5
Combined_fmllr	MLP, sgd	74.8	76.0	50.0	34.9	87.9	27.4
Combined	liGRU,rmsprop	81.0	77.6	55.7	42.0	82.4	19.2
Combined_fmllr	LSTM,rmsprop	82.3	78.6	57.2	44.0	81.9	17.6
Combined_fmllr	liGRU,rmsprop	82.6	78.5	57.7	44.4	82.1	17.3
Combined_fmllr	liGRU, Adam	**83.8**	**79.7**	**59.1**	**46.6**	80.9	**15.7**

Table 4-5 shows the MDD metrics for the recognized phoneme sequence passed through a 24-phonetic group converter. It should be noted that the models presented in Table 4-5 are still the same as the ones in Table 4-4 and were trained for the overall set of original phonemes. The results presented here are simply to see how accurate the models were in terms of detecting and diagnosing the mispronunciation across the phonemic groups as presented in Table 4-1. The best performing light GRU based architecture with fMLLR transformed combined features had a Detection accuracy of 83.8%, Diagnostic accuracy of 79.7%, F-measure of 59.1% and FRR of 15.7%. This was a relative improvement of 0.6%, 2.4%, 0.9% and 0.8% in DetAcc, DiagAcc, F measure and FRR

respectively by using combined features over the Acoustic features only. The performance improvement trend remains similar as that was seen for the MDD metrics for the overall set of phonemes as described in Table 4-1.

However, the central finding here is that that the MDD results for the grouped set of phonemes is not significantly higher than that for the whole phoneme set. This indicates that mispronunciation errors as detected by the MDD system are not necessarily confined within phonetic groups based on place and manner of articulation or broad vowel category, but extend to substantially different phoneme targets.

4.3.4 Mispronunciations identified by the best performing model

This section explores the mispronunciation errors as detected by the best performing ASR model. Information obtained from these results is useful in understanding the types of errors that are over predicted or under-predicted by the MDD system. The confusion matrix creation and hence identification of key mispronunciation errors is done over the test corpus used in the experiment.

Table 4-6 Vowel Confusion matrix between the transcript for standard prompt and expert transcript for utterances in test corpus

Phoneme in expert transcript

Phoneme in Prompt	IY	IH	EY	EH	AE	AX	AXR	UH	UW	AA	AO	OW	AH	ER	AY
IY	951	152	46	9	3	1	0	0	0	1	0	1	0	0	1
IH	146	1590	26	30	1	3	0	0	0	1	2	0	1	0	6
EY	23	5	526	50	12	1	0	0	0	4	0	0	1	0	8
EH	30	22	22	799	36	2	0	0	0	3	0	1	8	0	8
AE	1	15	11	53	932	1	0	0	0	50	3	0	8	0	6
AX	19	33	7	24	3	1960	37	0	3	1	3	9	28	0	3
AXR	0	1	0	2	0	55	238	0	0	0	3	12	0	40	0
UH	0	0	0	0	0	0	0	301	70	3	4	5	1	0	0
UW	3	0	0	0	0	3	2	3	745	10	11	16	19	0	0
AA	0	0	0	0	1	7	0	2	5	681	75	56	7	1	0
AO	0	0	3	0	3	2	0	0	3	44	511	28	18	0	1
OW	0	0	0	0	0	5	0	2	20	9	29	494	4	0	0
AH	0	0	1	0	6	1	2	0	4	44	1	1	513	0	1
ER	0	2	1	0	0	3	0	21	1	3	2	6	11	228	0
AY	0	5	4	0	3	1	0	0	0	3	0	0	0	0	598

Table 4-7 List of isolated substitution errors (vowels) with counts >= 25 obtained by aligning prompt with human labeled transcript for the test corpus

Prompt	Human Label	Count
IY	IH	152
IH	IY	146
AA	AO	75
UH	UW	70
AA	OW	56
AXR	AX	55
AE	EH	53
EY	EH	50
AE	AA	50
AO	AA	44
AH	AA	44
AXR	ER	40
AX	AXR	37
EH	AE	36
AX	IH	33
EH	IY	30
IH	EH	30
OW	AO	29
AO	OW	28
AX	AH	28

Table 4-8 Consonant Confusion matrix between the transcript for standard prompt and expert transcript for utterances in test corpus

Phoneme in expert transcript

Phoneme in Prompt	NG	N	M	L	P	B	R	W	V	F	T	D	DH	S	Z	SH	HH	Y	CH	K	G	JH
NG	337	49	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	102	1791	21	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
M	2	12	915	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	42	1	1128	0	0	23	2	0	0	1	1	0	0	0	0	1	0	0	0	0	0
P	0	0	0	0	975	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	17	881	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
R	1	0	0	30	0	0	1419	59	0	0	0	1	0	1	0	0	0	4	0	0	0	0
W	0	0	0	0	0	0	7	964	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	32	394	8	1	1	0	1	0	0	0	0	0	0	0	0
F	0	0	0	0	2	0	0	1	2	796	0	0	0	0	0	0	1	0	0	0	0	0
T	0	1	0	2	0	0	0	0	0	0	2528	13	0	1	1	0	0	0	2	1	0	0
D	0	0	0	1	1	2	0	0	3	0	21	1303	0	3	4	0	0	0	0	1	0	1
DH	0	0	0	3	0	0	0	0	0	0	2	63	1196	113	96	0	0	1	0	0	0	2
S	0	2	0	0	0	0	1	0	0	0	0	0	20	1863	15	9	0	0	0	2	0	0
Z	0	1	0	0	0	0	0	0	0	0	0	1	5	79	849	5	0	0	1	0	0	0
SH	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	211	11	0	1	0	0	0
HH	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	566	0	0	0	0	0
Y	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	112	0	0	0	0
CH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	116	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1517	3	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	628	0
JH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	176

Table 4-9 List of isolated substitution errors (consonants) with counts >= 10 obtained by aligning prompt with human labeled transcript for the test corpus

Prompt	Human Label	Count
DH	S	113
N	NG	102
DH	Z	96
Z	S	79
DH	D	63
R	W	59
NG	N	49
L	N	42
V	W	32
R	L	30
L	R	23
D	T	21
N	M	21
S	DH	20
B	P	17
S	Z	15
T	D	13
M	N	12
SH	HH	11

Table 4-6 and Table 4-8 are the confusion matrices for vowel and consonant sounds in the test corpus. It is seen that there are certain sounds with noticeable confusion in pronunciation. For vowels, there seems to be confusion between the sounds [/IY/, /IH/, /EY/, /EH/, /AE], [/AX/, /AXR/], [/UH/, /UW/], and [/AA/, /AO/, /OW/, /AH/]. For consonants, following sounds are often confused: [/NG/, /N/, /M/], [/B/, /P/], [/R/, /W/], [/T/, /D/], and [/DH/, /S/, /Z/]. Table 4-7 extracted from Table 4-6 and Table 4-9 extracted from Table 4-8 list the common substitution errors found in the test dataset. These errors serve as a reference to see how well the MDD system built is able to do in detecting and diagnosing them.

Table 4-10 Vowel Confusion matrix between the expert transcript and ASR generated transcript

Phoneme in ASR generated transcript

Phoneme in expert transcript	IY	IH	EY	EH	AE	AX	AH	AA	AO	OW	AY	AXR	UW	UH	ER
IY	987	79	26	2	0	4	0	0	0	0	0	0	2	0	0
IH	120	1454	30	7	3	21	0	0	0	0	1	2	1	0	0
EY	24	31	504	14	9	1	0	0	0	0	6	0	0	0	0
EH	3	17	15	749	86	11	5	2	0	0	7	0	0	1	0
AE	0	3	0	40	829	6	3	22	1	0	7	0	0	0	0
AX	1	42	0	4	4	1808	9	12	15	8	2	18	6	4	3
AH	0	1	0	16	4	11	419	89	6	2	4	1	1	2	1
AA	0	0	0	6	8	9	45	704	68	4	7	0	0	1	0
AO	0	0	0	0	0	1	12	53	453	40	2	0	0	0	1
OW	0	0	0	0	0	16	2	8	32	605	1	1	3	5	0
AY	0	0	0	3	3	4	2	4	3	0	544	0	0	0	0
AXR	0	1	0	0	0	9	0	0	0	0	0	214	0	1	10
UW	1	2	0	1	0	2	0	0	1	13	0	1	744	19	0
UH	0	0	0	0	0	6	2	2	0	9	0	0	4	267	2
ER	0	0	0	0	0	3	1	3	1	2	0	15	0	3	241

Table 4-11 Consonant Confusion matrix between the expert transcript and ASR generated transcript

Phoneme in ASR generated transcript

Phoneme in expert transcript	T	D	DH	Z	S	B	P	NG	N	M	L	R	W	CH	SH	K	G	V	F	JH	HH	Y
T	2227	37	9	6	8	0	0	0	3	0	0	1	0	8	1	7	0	1	0	3	0	1
D	102	1109	59	7	1	0	0	0	1	0	0	0	0	0	0	1	0	2	0	3	0	0
DH	3	25	1015	16	21	0	0	0	1	0	2	0	1	0	0	0	0	0	0	1	0	0
Z	7	4	31	828	63	0	0	0	0	0	0	2	1	0	7	0	0	0	0	1	0	0
S	6	1	30	72	1861	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2	0	0
B	1	0	0	0	0	821	26	0	0	3	0	0	0	0	0	0	0	3	2	0	0	0
P	0	0	1	0	0	22	1040	0	0	0	0	0	0	0	0	0	0	1	4	0	1	0
NG	0	0	0	0	0	0	0	364	44	3	0	2	0	0	0	0	0	0	0	0	0	0
N	3	5	5	2	1	0	0	49	1547	27	14	0	0	0	0	1	0	0	0	0	0	0
M	1	1	0	0	0	2	3	3	18	889	0	0	3	0	0	0	0	2	0	0	0	0
L	2	1	2	0	1	1	0	0	5	0	1021	45	0	0	0	0	0	0	0	0	0	0
R	0	1	0	1	1	0	0	0	2	2	19	1328	9	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	8	0	0	0	16	1	41	888	0	0	0	0	15	0	0	0	0
CH	6	0	0	1	1	0	0	0	0	0	0	0	0	110	2	0	0	0	0	0	0	0
SH	2	0	0	0	6	0	0	0	0	0	0	1	0	2	221	0	0	0	0	0	0	0
K	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1573	19	0	0	0	3	0
G	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	21	606	0	1	0	0	0
V	0	2	0	0	0	2	0	0	0	2	1	0	1	0	0	0	0	311	5	0	0	0
F	1	0	0	1	1	3	8	0	0	0	0	0	0	0	0	1	0	12	743	0	0	0
JH	6	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	145	0	0
HH	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	2	0	551	0
Y	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	113

Table 4-10 and Table 4-11 present the confusion matrix between the expert labeled transcripts versus the ASR generated phonetic sequences for utterances in test dataset. These tables reflect the performance of phoneme recognition built for MDD system. The diagonal numbers are the counts for correctly recognized phonemes while the non-diagonal counts are the misrecognized counts by the ASR. From Table 4-7 and Table 4-9 we can see the common substitution errors for vowels and consonants respectively obtained solely by transcription analysis. One interesting observation from Table 4-10 and Table 4-11 is that the phonemes that were misrecognized by the ASR system built for MDD often were among the list of errors noted in Table 4-7 and Table 4-9. For example, /DH/ was misrecognized as /D/ (25 times) and as /Z/ (16 times), /Z/ as /S/ (63 times), /NG/ as /N/ (44 times), /L/ as /R/ (45 times), /T/ as /D/ (37 times) and /S/ as /Z/ (72 times).

Similarly, /IY/ was misrecognized as /IH/ (79 times), /IH/ as /IY/ (120 times), /AO/ as /AA/ (53 times), /AA/ as /AO/ (68 times), and /EY/ as /EH/ (14 times).

Table 4-12 Vowel Confusion matrix between the standard prompt and ASR generated transcript

Phoneme in ASR generated transcript

Phoneme in Prompt	IY	IH	EY	EH	AX	AE	AH	AA	AO	OW	AXR	UW	UH	ER	AY
IY	869	77	41	4	4	4	0	0	0	0	1	0	0	0	1
IH	98	1295	24	16	2	4	0	2	0	0	1	2	1	0	5
EY	25	10	448	32	0	27	1	3	1	0	0	0	0	0	8
EH	20	25	11	689	5	25	3	2	1	0	0	1	0	0	7
AX	11	55	3	17	1598	6	12	9	6	4	11	0	0	2	1
AE	2	13	9	18	5	810	13	44	3	1	0	0	1	0	7
AH	0	0	1	10	2	5	377	69	3	0	1	3	0	1	2
AA	0	0	0	0	4	0	10	547	54	48	0	3	1	0	1
AO	0	0	0	2	1	0	20	31	369	27	0	0	2	1	1
OW	0	0	0	0	7	0	5	17	14	422	0	23	1	0	1
AXR	0	2	0	1	24	0	1	4	0	8	221	0	1	24	0
UW	2	3	0	0	1	0	13	15	9	16	2	636	1	0	1
UH	0	0	0	0	0	0	1	3	0	6	0	35	268	0	0
ER	0	2	0	1	6	0	5	1	3	1	1	1	10	186	0
AY	1	4	3	1	4	1	1	3	3	0	0	0	0	0	517

Table 4-13 Consonant Confusion matrix between the standard prompt and ASR generated transcript

Phoneme in ASR generated transcript

Phoneme in Prompt	B	P	T	D	NG	N	M	L	R	DH	S	Z	W	V	K	G	SH	F	CH	JH	HH	Y
B	700	39	2	0	0	0	0	0	0	0	1	0	0	1	0	0	0	2	0	0	0	0
P	11	914	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0	0	0	0
T	0	1	1989	22	0	3	0	0	1	1	9	5	0	0	7	1	0	0	4	2	1	1
D	0	1	58	907	0	0	1	0	0	5	1	4	0	3	0	2	0	0	0	3	0	0
NG	0	0	1	0	307	41	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	5	2	87	1433	26	4	0	4	0	2	0	0	1	0	1	0	0	0	0	0
M	4	0	1	1	1	15	809	0	1	0	0	0	3	1	0	0	0	0	0	0	0	0
L	1	0	1	1	0	21	0	920	36	0	1	0	0	0	0	0	0	0	0	0	0	0
R	0	0	1	1	0	4	1	28	1227	0	1	1	8	3	0	0	0	0	0	0	0	0
DH	0	0	10	29	0	0	0	2	1	1041	95	34	1	0	0	0	0	0	0	2	0	1
S	0	0	2	3	0	0	0	0	1	22	1516	21	0	0	1	0	7	0	0	2	0	0
Z	0	0	5	1	0	0	0	0	1	2	65	732	0	0	0	0	12	0	0	0	0	0
W	2	0	0	0	0	0	8	0	5	0	0	0	795	9	0	0	0	0	0	0	0	0
V	1	0	1	0	0	0	1	0	0	0	0	0	15	321	0	0	0	0	0	0	0	0
K	0	0	6	0	0	0	0	0	0	0	0	0	0	0	1346	10	0	0	0	0	2	0
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	20	547	0	0	0	1	0	0
SH	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	181	0	1	0	11	0
F	2	5	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	666	0	0	0	0
CH	0	0	4	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	147	0	0	0
JH	0	0	3	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	120	0	0
HH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	472	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108

Table 4-12 and Table 4-13 show the confusion matrices for vowels and consonants respectively, obtained by aligning the standard phonetic prompt with the ASR generated phonetic sequence. These confusion matrices give insights on the ability of MDD system to diagnose the actual mispronunciation errors when the results are compared with the counts obtained from the confusion matrices obtained by aligning the standard phonetic prompt with the human labeled transcript. The alignment of the prompt with the human labeled transcript reveals the actual mispronunciation errors assuming human annotators were capable of correctly detecting and diagnosing the mispronunciation errors in L2 speech. Whereas Table 4-12 and Table 4-13 reveal the information about mispronunciation errors as represented by misrecognized phonemes with non-zero counts in non-diagonal location in these matrices. In a real world scenario, where there is no availability of human

labeled phonetic transcript, the count of phoneme in the standard prompt (row) misrecognized as the phoneme in ASR generated transcript (column) indicate the error count for target phoneme substituted by the erroneous phoneme, thus revealing the count of mispronunciation errors in the L2 speech under testing. Table 4-14 presents the error counts obtained by human annotation and the counts generated by the proposed MDD system.

Table 4-14 Common isolated substitution error counts obtained by aligning the prompts with expert labeled transcripts (Manual error count) and by aligning the prompts with ASR generated transcript (MDD error count)

Target	Substituted	Manual count	MDD count
DH	S	113	95
N	NG	102	87
DH	Z	96	34
Z	S	79	65
DH	D	63	29
R	W	59	8
NG	N	49	41
L	N	42	21
V	W	32	15
R	L	30	28
L	R	23	36
D	T	21	58
N	M	21	26
S	DH	20	22
B	P	17	39
IY	IH	152	77
IH	IY	146	98
AA	AO	75	54
UH	UW	70	35
AA	OW	56	48
AXR	AX	55	24
AE	EH	53	18
AE	AA	50	44
EY	EH	50	32
IY	EY	46	41
AH	AA	44	69
AO	AA	44	31
AXR	ER	40	24
AX	AXR	37	11
EH	AE	36	25

The data in Table 4-14 show that there are some pairs of consonants and vowels which are commonly substituted for each other (Consonants: fricatives S and DH, nasals N and NG and liquids L and R; Vowels: AE and EH, IY and IH, AA and AO, AX and

AXR). The substitution of the voiced stops D and B by their unvoiced versions T and P was seen. For vowels, it can be seen that most of the substitution happens between the sounds that are closer in the vowel quadrilateral, whose extreme four corners represent extreme points of articulation. It can also be observed from Table 4-14 that the MDD system often under predicted the error counts. It can thus be said that the system might miss capturing a fraction of mispronunciations.

A high False Rejection Rate (FRR) system would have a higher number of cases where the system would incorrectly reject the correctly pronounced segments by labeling them as mispronounced. This would act as a demotivating factor for learners in CAPT. Therefore, it is desired for a MDD system to have a low FRR. The best performing system as reported in Table 4-4 has false rejection rate of 17.2% and recall of 80.3%. The system is a high recall but low precision system. It can be further observed from Table 4-14 that most of the times the MDD system under predicted the error counts. For example the error /IY/ substituted by /IH/ was found to have a manual count of 152 but only 77 instances of those errors were correctly detected by the system. This along with the observation from Table 4-14 it can be said that the system might miss capturing a fraction of mispronunciations. Therefore even with the system with lower precision score, if the FRR, Detection Accuracy and Diagnostic accuracy are reasonable, the system can be useful as a MDD system in CAPT.

4.4 Conclusion

This study has presented details of the implemented Automatic Mispronunciation and Detection system. Multiple different architectures and features were implemented in

constructing the MDD system, and hyper parameters were tuned to find the best performing parameter values. The MDD metrics computed for different combination of features and architectures reveal that incorporating articulatory features with acoustic features improves almost all of the MDD metrics of the system. Using fMLLR transformed input features and the appropriate optimizer (Adam in this case) both significantly help improving PER and hence the MDD performance, and the best performing model was the light GRU based architecture using a monophone regularization technique and a multi-task learning approach.

Using phonetic prompts coupled with expertly labeled phonetic transcripts considered as ground truth, the MDD system generated transcript was aligned against these references to analyze the performance of the proposed MDD system. The implemented MDD system captures mispronunciation errors with a Detection Accuracy of 82.4%, a Diagnostic Accuracy of 75.8% and a False Rejection Rate of 17.2%.

CHAPTER 5. **CONCLUSION AND FUTURE WORK**

5.1 Overview

This research has focused on identification of commonly occurring mispronunciation errors in Mandarin speakers of English and their articulatory error patterns. Expert labeled transcripts were sufficient to identify the common mispronunciation errors in L2 speech. An automatic speech recognition (ASR) based Mispronunciation detection and diagnosis (MDD) system was built using acoustic and articulatory features available in EMA-MAE database. Statistical comparison between speech produced by native speakers of English (L1) and Mandarin speakers of English (L2) in articulatory feature space revealed the most significant differences in positioning of articulators between L1 and L2 causing the associated pronunciation error.

5.2 Conclusion

The main contribution of this work is the application of speech recognition models to perform detection and diagnostic analysis of mispronunciation errors in Mandarin speakers of English. The advantage of availability of kinematic data as well as the expert labeled phonetic transcript in EMA-MAE database was well utilized in identifying and diagnosing the commonly occurring mispronunciation errors in Mandarin speakers of English. While building an automatic MDD system, different neural network architectures and feature combinations were experimented.

In the study using the expert labeled transcripts aligned against the prompts, the common mispronunciation errors were identified. These errors provide the region of

interest in a large set of phonetic substitutions that might happen in L2 speech. The common errors identified in this study match the errors reported in literature studying mispronunciation errors in Mandarin speakers of English. Most of the reported errors can be easily detected in the form of substitution error, where the target phoneme is substituted by an erroneous phoneme. With reference to the target phoneme, the erroneous phoneme is more likely to fall either in same place and/or manner of articulation for consonants and closer in the vowel diagram. Study of mispronunciation errors across the two dialects of Mandarin speakers: Beijing and Shanghai reveal that the averaged count of consonant errors for Shanghai dialect is higher than that of the speakers with Beijing dialect. Since the Mandarin language does not include word ending consonants, errors at the end of words are of particular interest. The most common error type reported in the literature for words ending with a consonant is a substitution error. However, this study revealed that the error can be either a substitution or a deletion error with almost equal likelihood. This means that the Mandarin speakers tend to either drop the final consonants or substitute by some erroneous sound.

For the commonly occurring substitution errors identified from expert transcripts for L2 speech, a statistical comparison between the L1 speech (template) versus the L2 speech in articulatory feature space can reveal the erroneous articulatory pattern. There were some articulatory movements common across all the sounds produced by L2 speakers. This includes too large height at tongue lateral, too far posterior position of tongue lateral, errors in position of the tongue apex in all directions, too small of a vertical lip separation, and too small of a jaw opening. The most significant conclusion that can be drawn from the study regarding articulatory error diagnosis for mispronunciations is the

lack of symmetry between the pair of sounds which substituted for each other. It might be expected that articulatory error pattern in one direction of substitution would be complementary to the pattern for the substitution in opposite direction, but this was not the case. In some cases completely different articulators were involved, and in some cases the same articulators with error pattern in same direction was seen. This suggests that the most significant articulatory error patterns are not necessarily the only factors causing the associated pronunciation error.

In the study building ASR based MDD systems, different neural network architectures and feature combinations were implemented. Articulatory features combined with acoustic features improved the ASR performance which in turn improved all of the MDD related metrics. fMLLR transformed features had better performance in all of the ASR architectures tested. The best performing network architecture was based on a light GRU model trained in multi-task learning fashion with monophone regularization. The system had the ability of detecting mispronunciation errors in speech (Detection Accuracy) of 82.4% and the ability to correctly identify the type of substitution error for the mispronunciations detected (Diagnostic Accuracy) of 75.8%. With reference to the actual substitution errors as noted in expert labeled transcript, the MDD system mostly under-predicted the substitution errors. However, there are also certain errors, for example, /AH/ substituted by /AA/, /AX/ substituted by /IH/, /L/ substituted by /R/, and /D/ substituted by /T/ which are over-predicted by the system.

5.3 Future work

One potential experiments in terms of extending the work related to diagnostic analysis of mispronunciation in articulatory feature space could be to include the perspective of left and right context in the mispronounced sound for L2 speech. This way the articulatory error pattern for a specific mispronunciation for a given context can be analyzed. Through this approach, there will be too many combination of context dependent mispronunciation errors. Therefore, the context sounds can be grouped into some categories based on their manner of articulation. This way articulatory error patterns for L2 speakers can be studied while also keeping into consideration the co-articulation effect due to left and right context of the mispronounced sound segment.

From the observation of results of ASR based MDD systems for different feature combination and architectures, it can be noticed that the relative improvement of Detection and Diagnostic Accuracy for the best performing model using only acoustic features and the combined features is only 0.9% and 2.6% respectively. This suggests room for improvement. Future work in this regard includes identifying more meaningful representations of articulatory feature or better ways to incorporate those features into the ASR based MDD systems to optimally utilize the valuable information carried by the articulatory features.

APPENDICES

APPENDIX 1. BOX PLOTS FOR CORRECT AND MISPRONOUNCED ERRORS IN ARTICULATORY FEATURE SPACE

This section presents the box plots obtained between samples from two groups: Articulatory feature vectors for the correctly pronounced sound by native speakers of English (L1) versus that for the corresponding mispronunciation by Mandarin speakers of English (L2). As discussed in 3.3, these vectors are obtained by averaging the 10-dimensional articulatory feature frames across the time. Thus loosely speaking, they represent a snapshot of the articulatory state during the production of the sound of interest. Box plots between these averaged vectors for L1 and L2 sample group for a known type of mispronunciation error can reveal the information that which articulatory state is significantly different between the L1 and L2 speaker groups as visualized in the box plots presented as follows. Even though there are numerous mispronunciation errors detected, the box plots for consonant errors with count greater than 80 and that for vowels with count greater than 100 are only listed shown here.

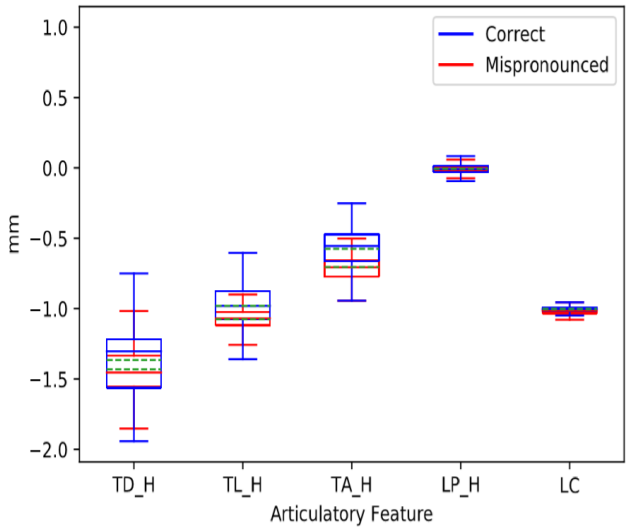

Figure 27 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /DH/ by L1 speakers versus /DH/ substituted by /D/ by L2 speakers

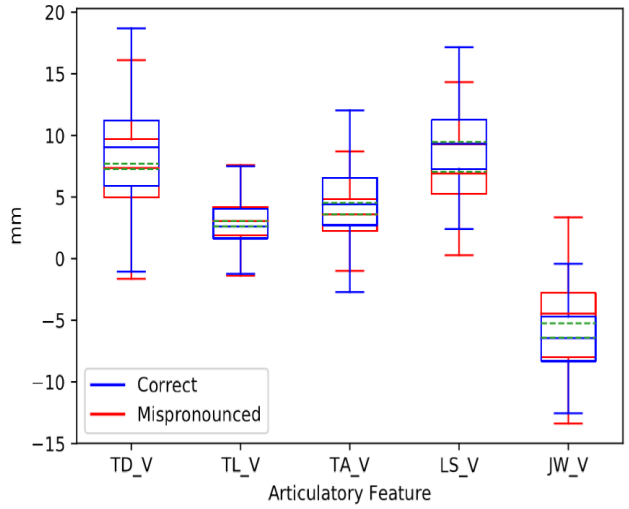

Figure 28 Box plots in vertical articulatory feature space for correctly pronounced /DH/ by L1 speakers versus /DH/ substituted by /D/ by L2 speakers

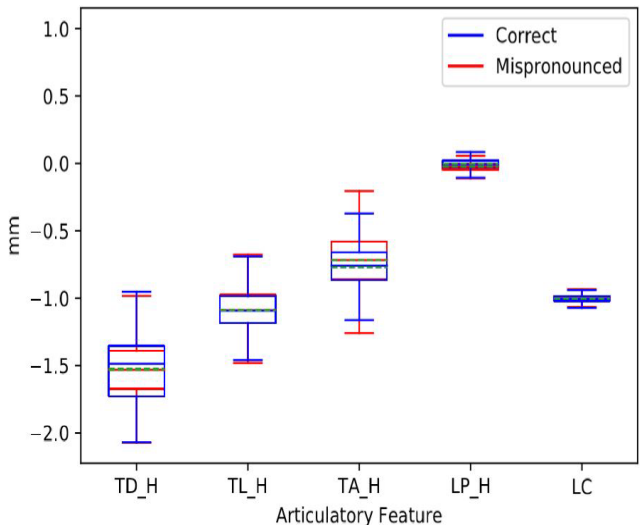

Figure 29 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /N/ by L1 speakers versus /N/ substituted by /NG/ by L2 speakers

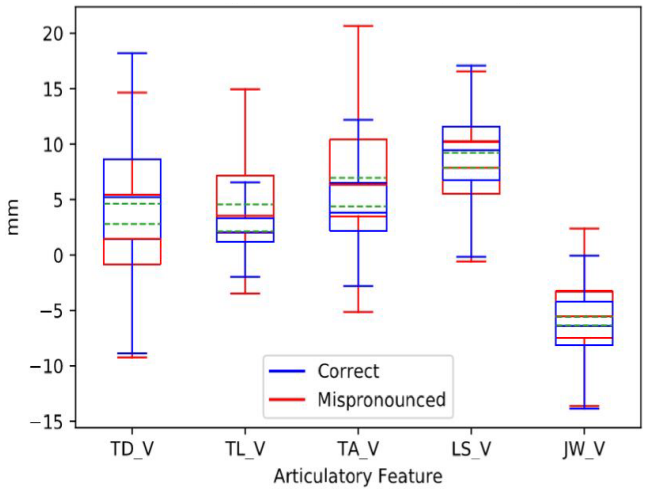

Figure 30 Box plots in vertical articulatory feature space for correctly pronounced /N/ by L1 speakers versus /N/ substituted by /NG/ by L2 speakers

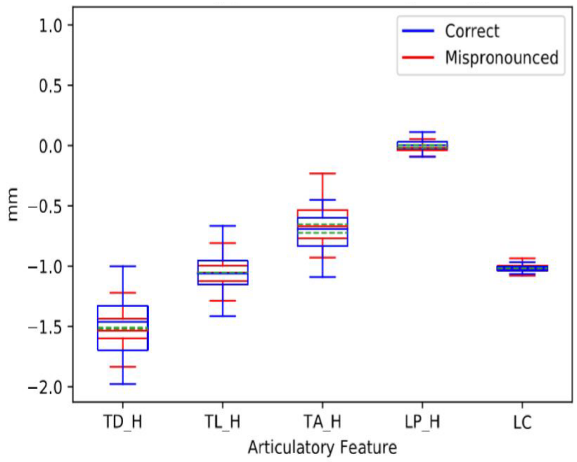

Figure 31 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /Z/ by L1 speakers versus /Z/ substituted by /S/ by L2 speakers

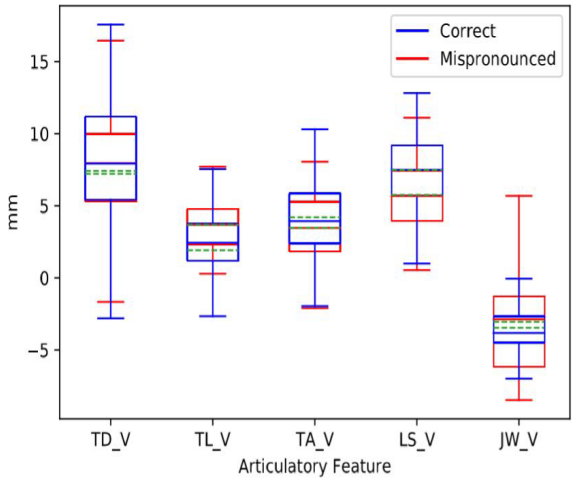

Figure 32 Box plots in vertical articulatory feature space for correctly pronounced /Z/ by L1 speakers versus /Z/ substituted by /S/ by L2 speakers

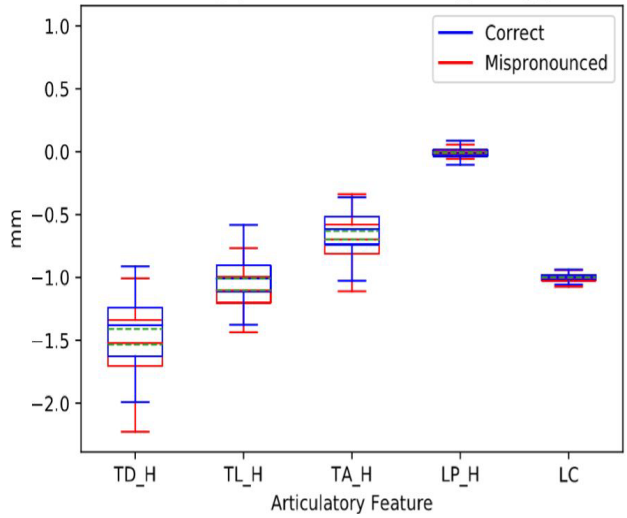

Figure 33 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /NG/ by L1 speakers versus /NG/ substituted by /N/ by L2 speakers

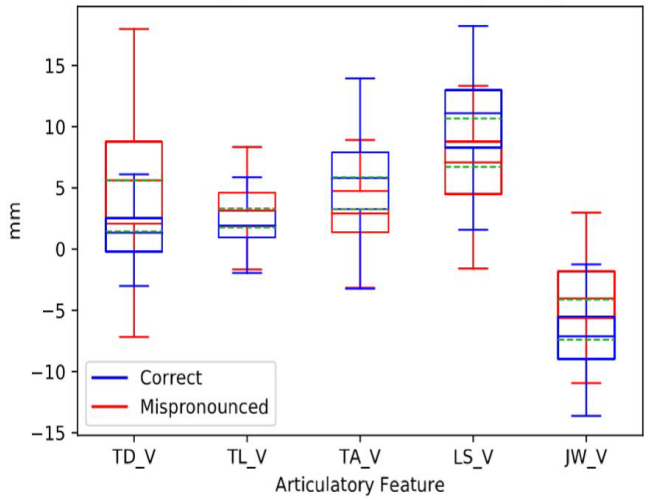

Figure 34 Box plots in vertical articulatory feature space for correctly pronounced /NG/ by L1 speakers versus /NG/ substituted by /N/ by L2 speakers

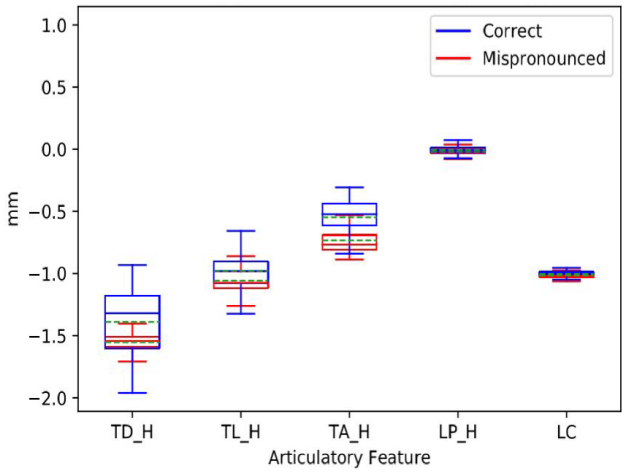

Figure 35 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /DH/ by L1 speakers versus /DH/ substituted by /Z/ by L2 speakers

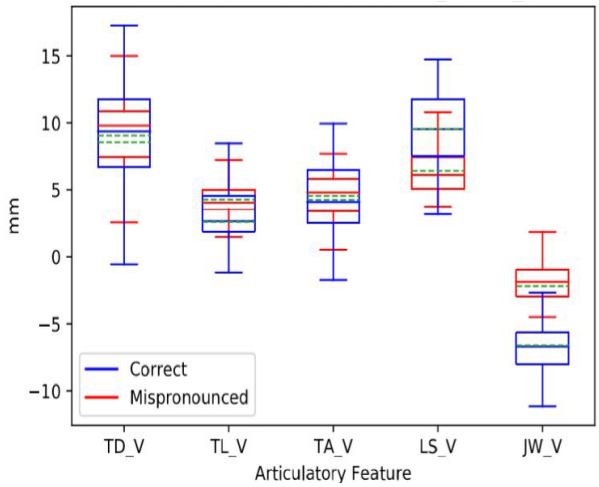

Figure 36 Box plots in vertical articulatory feature space for correctly pronounced /DH/ by L1 speakers versus /DH/ substituted by /Z/ by L2 speakers

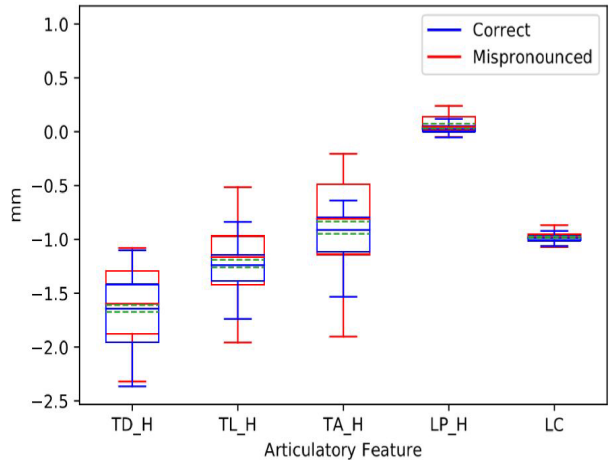

Figure 37 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /R/ by L1 speakers versus /R/ substituted by /W/ by L2 speakers

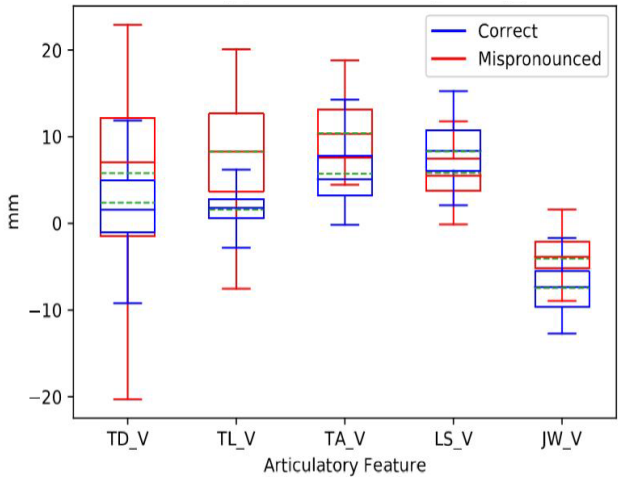

Figure 38 Box plots in vertical articulatory feature space for correctly pronounced /R/ by L1 speakers versus /R/ substituted by /W/ by L2 speakers

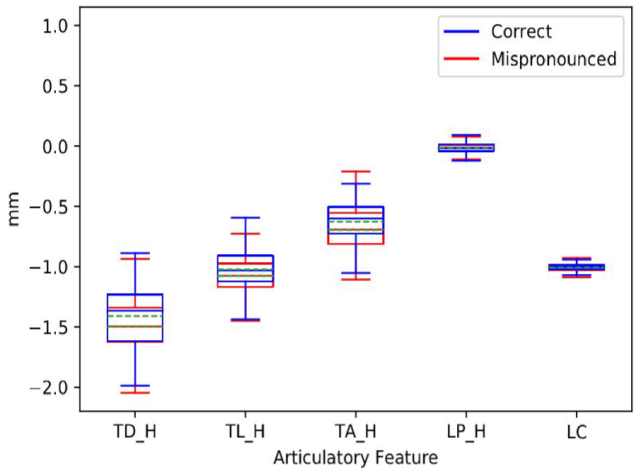

Figure 39 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /IY/ by L1 speakers versus /IY/ substituted by /IH/ by L2 speakers

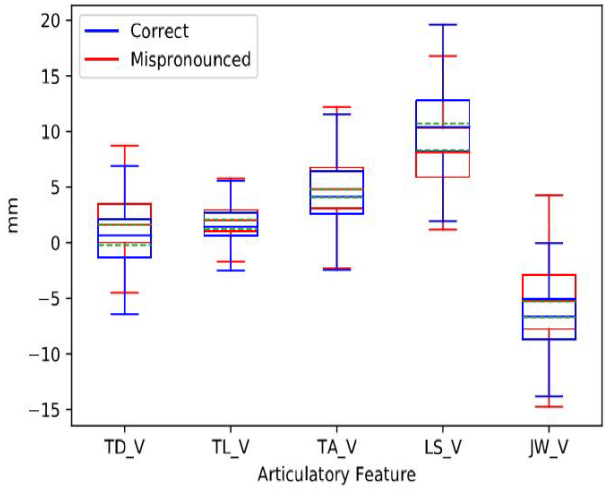

Figure 40 Box plots in vertical articulatory feature space for correctly pronounced /IY/ by L1 speakers versus /IY/ substituted by /IH/ by L2 speakers

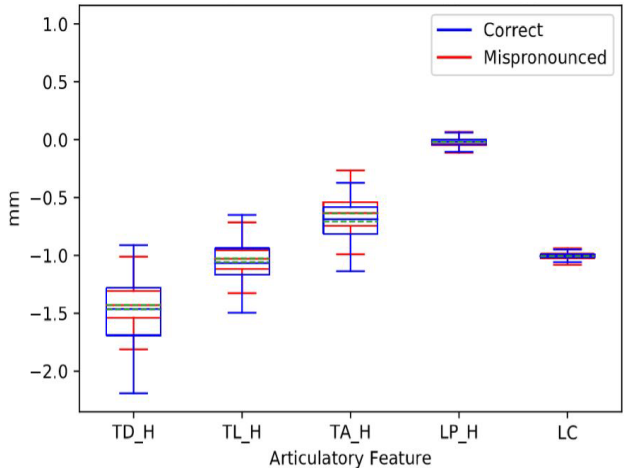

Figure 41 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /IH/ by L1 speakers versus /IH/ substituted by /IY/ by L2 speakers

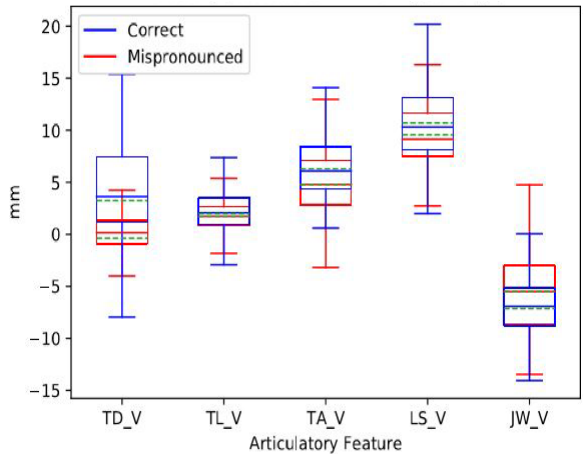

Figure 42 Box plots in vertical articulatory feature space for correctly pronounced /IH/ by L1 speakers versus /IH/ substituted by /IY/ by L2 speakers

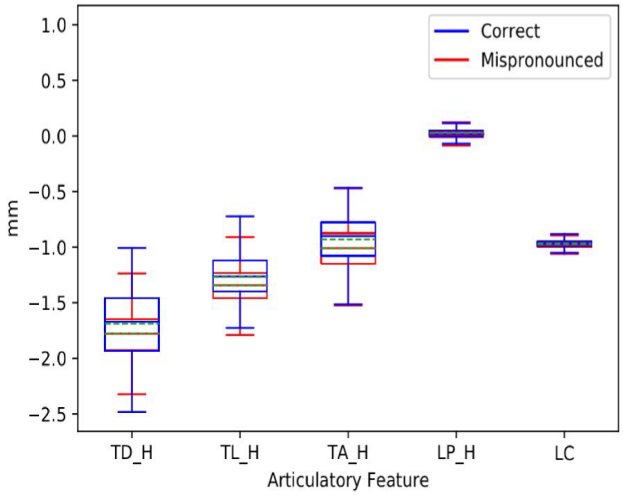

Figure 43 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /AO/ by L1 speakers versus /AO/ substituted by /AA/ by L2 speakers

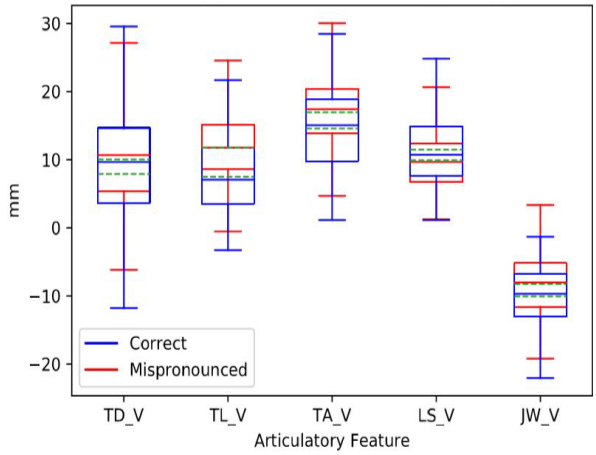

Figure 44 Box plots in vertical articulatory feature space for correctly pronounced /AO/ by L1 speakers versus /AO/ substituted by /AA/ by L2 speakers

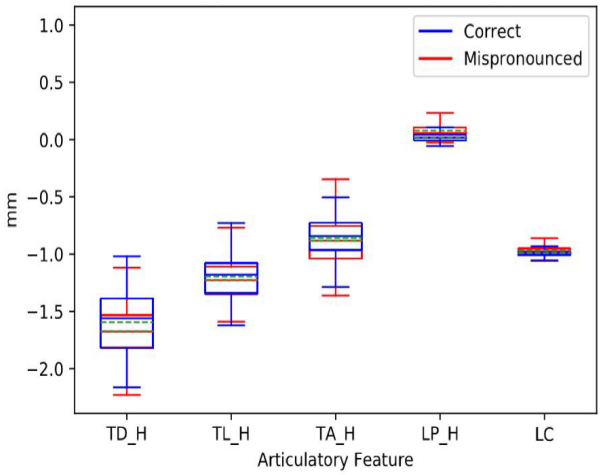

Figure 45 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /UH/ by L1 speakers versus /UH/ substituted by /UW/ by L2 speakers

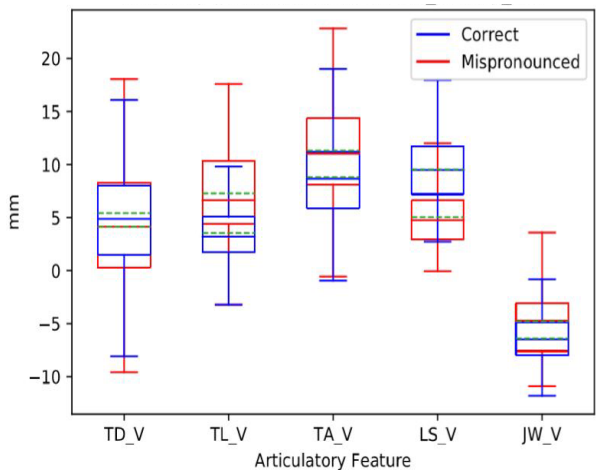

Figure 46 Box plots in vertical articulatory feature space for correctly pronounced /UH/ by L1 speakers versus /UH/ substituted by /UW/ by L2 speakers

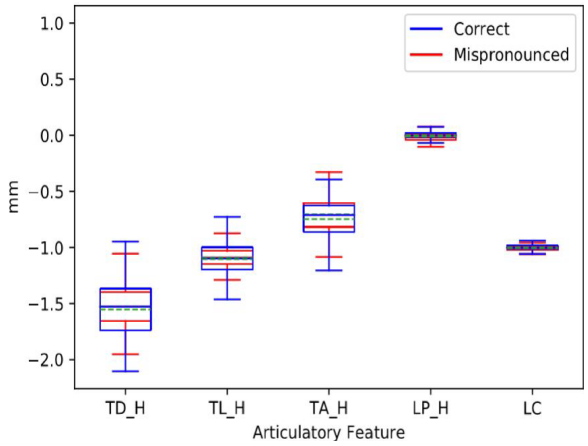

Figure 47 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /AX/ by L1 speakers versus /AX/ substituted by /IH/ by L2 speakers

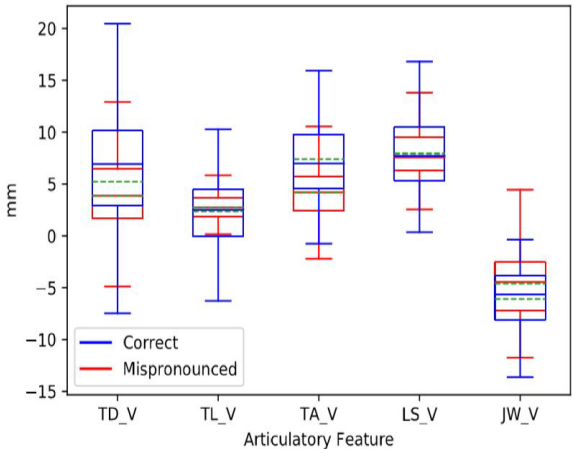

Figure 48 Box plots in vertical articulatory feature space for correctly pronounced /AX/ by L1 speakers versus /AX/ substituted by /IH/ by L2 speakers

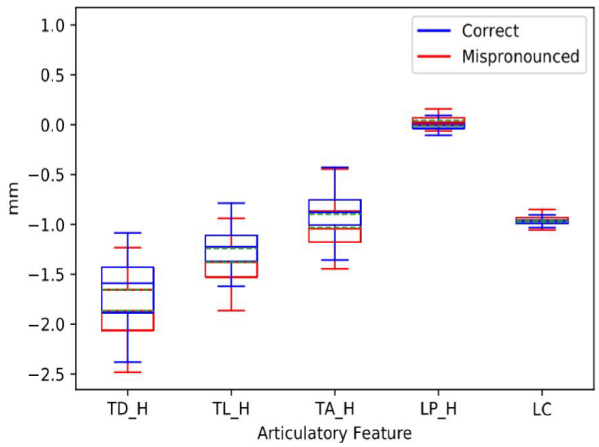

Figure 49 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /AA/ by L1 speakers versus /AA/ substituted by /AO/ by L2 speakers

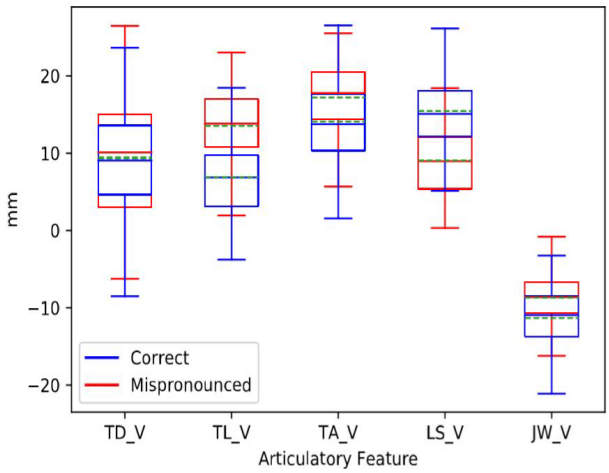

Figure 50 Box plots in vertical articulatory feature space for correctly pronounced /AA/ by L1 speakers versus /AA/ substituted by /AO/ by L2 speakers

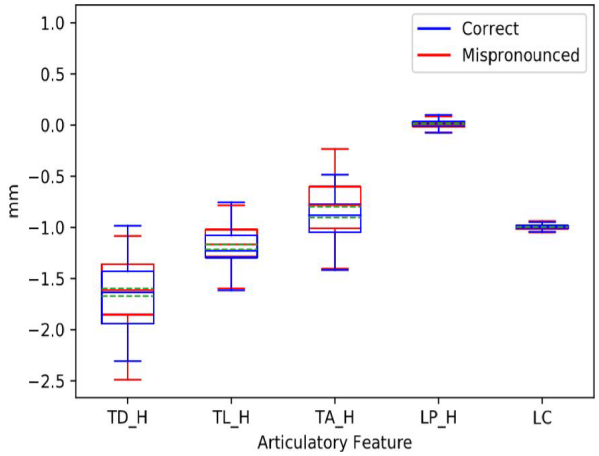

Figure 51 Box plots in horizontal and lateral articulatory feature space for correctly pronounced /AXR/ by L1 speakers versus /AXR/ substituted by /AX/ by L2 speakers

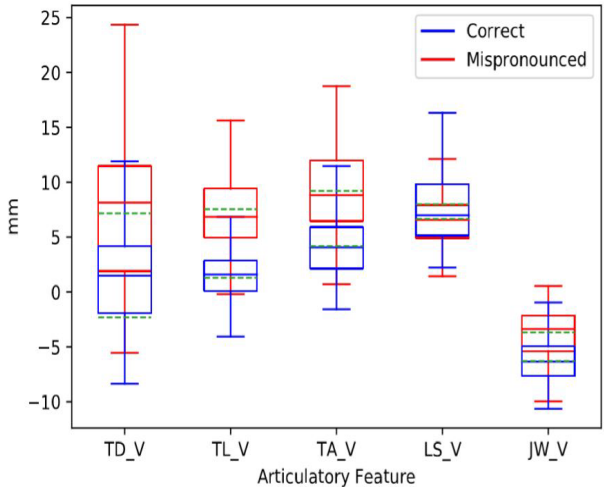

Figure 52 Box plots in vertical articulatory feature space for correctly pronounced /AXR/ by L1 speakers versus /AXR/ substituted by /AX/ by L2 speakers

APPENDIX 2. ERROR DISTRIBUTION ACROSS THE MANDARIN SPEAKERS

Table 5-1 Vowel substitution errors for Mandarin speakers with Beijing dialect

Prompt	Transcript	Error	08MBM	29MBM	13MBF	03MBM	31MBM	02MBF	20MBF	11MBF	22MBF	01MBF	23MBM	Beijing
IY	IH	362	22	18	14	16	11	14	19	17	19	17	22	189
IH	IY	284	5	7	14	5	29	9	16	22	24	30	14	175
AO	AA	282	11	8	15	13	11	17	15	10	15	10	16	141
UH	UW	151	11	3	5	8	19	7	8	12	4	3	9	89
AX	IH	127	9	4	6	9	5	2	10	1	6	5	13	70
AA	AO	124	1	6	7	4	9	2	6	14	2	5	13	69
AXR	AX	120	0	2	0	2	2	0	0	14	0	0	0	20
AA	OW	93	1	6	3	4	7	4	8	6	2	15	4	60
AH	AA	92	2	4	9	1	7	6	11	3	9	6	9	67
EY	EH	92	0	4	1	3	8	0	4	2	5	14	2	43
AE	EH	92	5	4	5	1	4	1	1	3	1	15	4	44
EH	AE	91	3	0	10	2	4	7	5	11	14	1	2	59
AE	AA	78	4	11	8	6	10	4	1	5	0	3	5	57
IY	EY	77	1	14	2	5	8	7	2	3	4	8	0	54
AX	IY	56	0	0	2	4	1	1	3	3	0	1	14	29
EH	IY	56	0	1	2	2	7	1	3	4	2	6	3	31
OW	UW	56	3	2	2	2	3	1	3	7	1	3	3	30
D	AX	56	0	5	5	1	1	0	4	1	0	0	2	19
AO	OW	55	2	3	0	1	3	4	3	2	3	6	6	33
AXR	ER	55	3	3	3	2	3	3	3	2	3	3	3	31
EH	IH	53	1	5	2	3	1	1	2	1	2	2	4	24
AX	EH	51	1	2	2	2	5	1	0	1	4	4	3	25
IH	EY	50	0	3	2	6	6	2	0	2	2	6	0	29
OW	AO	49	2	3	3	4	4	2	2	2	0	2	2	26
IH	EH	48	0	4	2	2	6	1	1	2	2	3	5	28
AX	AXR	47	0	1	6	0	2	0	0	3	0	27	3	42
AX	AH	43	2	2	2	2	4	2	3	3	2	2	2	26
ER	UH	40	0	1	0	0	4	0	1	0	0	1	0	7
EH	EY	36	0	2	2	0	1	0	1	1	0	3	0	10
EY	IY	33	2	3	0	2	2	0	0	11	3	0	4	27
AO	AH	31	6	4	2	2	2	1	0	2	0	2	0	21
UW	OW	30	0	3	3	3	3	0	0	2	1	2	2	19
EY	AE	30	1	0	1	2	1	0	2	2	2	2	1	14
UW	AH	27	1	2	2	0	1	1	0	1	0	3	1	12
T	AX	25	5	1	2	1	1	4	0	0	0	0	1	15
OW	AA	24	0	2	1	1	0	1	1	0	0	1	1	8
AX	OW	22	0	1	0	2	1	1	0	1	2	2	2	12
AE	IH	21	1	2	1	0	0	5	0	7	1	0	1	18
AE	AH	20	1	1	5	0	1	2	4	0	0	1	2	17
AE	EY	19	0	1	2	0	0	1	0	1	0	2	1	8

Table 5-2 Vowel substitution errors for Mandarin speakers with Shanghai dialect

Prompt	Transcript	Error	10MSM	24MSF	14MSF	04MSF	25MSM	27MSM	12MSF	30MSM	26MSM	Shanghai
IY	IH	362	17	16	26	23	16	17	24	18	16	173
IH	IY	284	11	8	19	7	11	12	16	11	14	109
AO	AA	282	11	23	10	18	22	12	12	20	13	141
UH	UW	151	12	3	6	5	7	8	4	12	5	62
AX	IH	127	6	1	9	7	5	6	4	9	10	57
AA	AO	124	4	4	9	5	3	17	8	4	1	55
AXR	AX	120	28	2	23	3	18	5	10	1	10	100
AA	OW	93	1	0	8	6	2	6	3	3	4	33
AH	AA	92	6	3	3	3	2	1	3	4	0	25
EY	EH	92	10	1	15	5	4	2	1	7	4	49
AE	EH	92	10	1	4	9	8	9	2	1	4	48
EH	AE	91	3	9	1	7	0	2	4	2	4	32
AE	AA	78	1	0	5	0	0	12	0	2	1	21
IY	EY	77	0	6	4	0	3	4	3	3	0	23
AX	IY	56	1	1	9	4	1	5	0	3	3	27
EH	IY	56	5	0	3	3	2	3	1	6	2	25
OW	UW	56	4	1	5	3	2	4	0	2	5	26
D	AX	56	0	15	13	1	1	3	0	2	2	37
AO	OW	55	2	0	5	2	4	4	0	4	1	22
AXR	ER	55	3	3	3	2	3	2	2	3	3	24
EH	IH	53	6	1	7	4	4	2	0	2	3	29
AX	EH	51	6	2	1	1	5	5	2	1	3	26
IH	EY	50	2	3	1	2	7	2	0	3	1	21
OW	AO	49	3	1	4	7	0	2	0	5	1	23
IH	EH	48	3	3	3	2	2	3	2	1	1	20
AX	AXR	47	0	0	0	0	1	2	2	0	0	5
AX	AH	43	2	2	2	2	2	1	2	2	2	17
ER	UH	40	7	1	14	1	2	3	2	0	3	33
EH	EY	36	6	1	4	6	3	2	0	1	3	26
EY	IY	33	4	0	0	1	0	0	0	0	1	6
AO	AH	31	2	0	0	0	1	3	1	0	3	10
UW	OW	30	2	3	1	2	0	1	0	1	1	11
EY	AE	30	3	0	2	1	2	0	0	1	7	16
UW	AH	27	3	0	4	2	1	4	0	1	0	15
T	AX	25	6	0	2	0	0	0	0	2	0	10
OW	AA	24	1	2	2	1	1	3	0	3	3	16
AX	OW	22	1	0	1	1	3	1	2	1	0	10
AE	IH	21	1	0	0	1	0	0	0	1	0	3
AE	AH	20	1	0	0	0	1	0	1	0	0	3
AE	EY	19	3	0	1	0	2	1	1	2	1	11

Table 5-3 Consonant substitution errors for Mandarin speakers with Beijing dialect

Prompt	Transcript	Error	08MBM	29MBM	13MBF	03MBM	31MBM	02MBF	20MBF	11MBF	01MBF	23MBM	Beijing
DH	D	302	53	2	9	18	0	3	1	42	5	4	137
N	NG	236	7	4	15	0	12	2	4	4	4	10	71
Z	S	150	3	3	4	2	15	8	8	7	8	10	71
NG	N	127	1	11	8	12	2	1	5	3	4	4	56
DH	Z	92	1	2	0	2	8	0	1	0	8	2	24
R	W	83	0	1	1	1	3	3	4	1	0	0	18
N	M	49	4	0	3	0	0	3	1	2	4	0	18
T	D	47	5	2	3	5	0	2	1	1	1	5	25
L	R	44	0	2	0	0	0	1	0	1	3	2	11
S	Z	40	3	2	4	1	1	2	0	2	0	2	19
L	N	37	0	0	0	0	0	0	0	0	0	1	1
D	T	37	3	2	0	1	0	2	5	1	2	1	24
Z	SH	34	1	1	1	1	3	2	3	1	1	0	14
R	L	31	0	1	0	0	0	0	1	0	0	0	5
M	N	31	0	2	0	1	2	0	0	0	5	2	12
V	W	31	1	4	0	6	1	0	0	3	1	1	20
B	P	30	1	4	1	0	1	2	1	1	4	3	18
S	DH	29	0	2	0	2	0	0	3	0	0	7	14
G	K	25	2	1	1	2	1	1	1	2	0	2	14
V	F	20	0	0	0	0	0	0	7	1	0	2	10
Z	DH	20	0	6	0	2	0	0	0	0	0	4	15
SH	HH	15	0	1	0	0	4	0	0	0	0	0	9
L	W	13	0	0	1	1	0	0	1	0	0	0	4

Table 5-4 Consonant substitution errors for Mandarin speakers with Shanghai dialects

Prompt	Transcript	Error	10MSM	24MSF	14MSF	04MSF	25MSM	27MSM	12MSF	30MSM	26MSM	Shanghai
DH	D	302	2	70	2	0	37	26	5	14	9	165
N	NG	236	21	7	31	8	17	38	15	20	8	165
Z	S	150	4	6	12	6	15	18	5	9	4	79
NG	N	127	4	8	8	17	4	3	5	10	12	71
DH	Z	92	2	0	36	5	3	16	5	0	1	68
R	W	83	36	1	16	0	4	0	1	1	6	65
N	M	49	6	4	3	4	2	1	4	6	1	31
T	D	47	1	3	0	2	2	1	1	4	8	22
L	R	44	0	1	7	3	7	10	1	3	1	33
S	Z	40	2	3	0	3	2	2	1	4	4	21
L	N	37	33	0	1	1	0	1	0	0	0	36
D	T	37	5	1	2	1	2	1	0	0	1	13
Z	SH	34	2	3	3	3	3	1	1	2	2	20
R	L	31	6	0	8	1	0	4	1	4	2	26
M	N	31	3	2	2	0	3	1	2	5	1	19
V	W	31	3	0	1	0	3	1	0	3	0	11
B	P	30	1	1	1	2	1	1	0	0	5	12
S	DH	29	1	0	0	9	2	0	1	2	0	15
G	K	25	1	1	0	2	2	1	1	1	2	11
V	F	20	1	1	0	0	0	5	0	0	3	10
Z	DH	20	0	1	2	0	1	0	0	1	0	5
SH	HH	15	1	0	2	0	2	1	0	0	0	6
L	W	13	3	1	2	0	0	0	0	1	2	9

APPENDIX 3. CONFIGURATION FILE FOR THE BEST PERFORMING MODEL WITH COMBINED FEATURES

```
[cfg_proto]
cfg_proto = proto/global.proto
cfg_proto_chunk = proto/global_chunk.proto

[exp]
cmd =
run_nn_script = run_nn
out_folder = /storage/subash/pytorch-exps/Combined/liGRU_fmllr/exps/0.1_0.0004
seed = 1234
use_cuda = True
multi_gpu = True
save_gpumem = False
n_epochs_tr = 25
production = False

[dataset1]
data_name = EMAMAE_tr
fea = fea_name=mfcc
        fea_lst=/home/subash/kaldi/egs/Combined/data/train/feats.scp
        fea_opts=apply-cmvn --
utt2spk=ark:/home/subash/kaldi/egs/Combined/data/train/utt2spk
ark:/home/subash/kaldi/egs/Combined/data/train/data/cmvn_train.ark ark:- ark:- |
        cw_left=0
        cw_right=0

        fea_name=fbank
        fea_lst=/home/subash/kaldi/egs/Combined/data/train2/feats.scp
```

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data/train2/utt2spk ark:/home/subash/kaldi/egs/Combined/data/train2/data/cmvn_train2.ark ark:- ark:- |

cw_left=0

cw_right=0

fea_name=fmllr

fea_lst=/home/subash/kaldi/egs/Combined/data-fmllr/train/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data-fmllr/train/utt2spk ark:/home/subash/kaldi/egs/Combined/data-fmllr/train/data/cmvn_speaker.ark ark:- ark:- |

cw_left=0

cw_right=0

lab = lab_name=lab_cd

lab_folder=/home/subash/kaldi/egs/Combined/exp/tri3_ali

lab_opts=ali-to-pdf

lab_count_file=auto

lab_data_folder=/home/subash/kaldi/egs/Combined/data/train/

lab_graph=/home/subash/kaldi/egs/Combined/exp/tri3/graph

lab_name=lab_mono

lab_folder=/home/subash/kaldi/egs/Combined/exp/tri3_ali

lab_opts=ali-to-phones --per-frame=true

lab_count_file=none

lab_data_folder=/home/subash/kaldi/egs/Combined/data/train/

lab_graph=/home/subash/kaldi/egs/Combined/exp/tri3/graph

n_chunks = 15

[dataset2]

data_name = EMAMAE_dev

fea = fea_name=mfcc

fea_lst=/home/subash/kaldi/egs/Combined/data/dev/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data/dev/utt2spk ark:/home/subash/kaldi/egs/Combined/data/dev/data/cmvn_dev.ark ark:- ark:- |

cw_left=0

cw_right=0

fea_name=fbank

fea_lst=/home/subash/kaldi/egs/Combined/data/dev2/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data/dev2/utt2spk ark:/home/subash/kaldi/egs/Combined/data/dev2/data/cmvn_dev2.ark ark:- ark:- |

cw_left=0

cw_right=0

fea_name=fmllr

fea_lst=/home/subash/kaldi/egs/Combined/data-fmllr/dev/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data-fmllr/dev/utt2spk ark:/home/subash/kaldi/egs/Combined/data-fmllr/dev/data/cmvn_speaker.ark ark:- ark:- |

cw_left=0

cw_right=0

lab = lab_name=lab_cd

lab_folder=/home/subash/kaldi/egs/Combined/exp/tri3_ali_dev

lab_opts=ali-to-pdf

lab_count_file=auto

lab_data_folder=/home/subash/kaldi/egs/Combined/data/dev/

lab_graph=/home/subash/kaldi/egs/Combined/exp/tri3/graph

lab_name=lab_mono

lab_folder=/home/subash/kaldi/egs/Combined/exp/tri3_ali_dev

lab_opts=ali-to-phones --per-frame=true

lab_count_file=none

lab_data_folder=/home/subash/kaldi/egs/Combined/data/dev/

lab_graph=/home/subash/kaldi/egs/Combined/exp/tri3/graph

n_chunks = 5

[dataset3]

data_name = EMAMAE_test

fea = fea_name=mfcc

fea_lst=/home/subash/kaldi/egs/Combined/data/test/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data/test/utt2spk ark:/home/subash/kaldi/egs/Combined/data/test/data/cmvn_test.ark ark:- ark:- |

cw_left=0

cw_right=0

fea_name=fbank

fea_lst=/home/subash/kaldi/egs/Combined/data/test2/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data/test2/utt2spk ark:/home/subash/kaldi/egs/Combined/data/test2/data/cmvn_test2.ark ark:- ark:- |

cw_left=0

cw_right=0

fea_name=fmllr

fea_lst=/home/subash/kaldi/egs/Combined/data-fmllr/test/feats.scp

fea_opts=apply-cmvn --utt2spk=ark:/home/subash/kaldi/egs/Combined/data-fmllr/test/utt2spk ark:/home/subash/kaldi/egs/Combined/data-fmllr/test/data/cmvn_speaker.ark ark:- ark:- |

cw_left=0

cw_right=0

lab = lab_name=lab_cd

lab_folder=/home/subash/kaldi/egs/Combined/exp/tri3_ali_test

```
        lab_opts=ali-to-pdf
        lab_count_file=auto
        lab_data_folder=/home/subash/kaldi/egs/Combined/data/test/
        lab_graph=/home/subash/kaldi/egs/Combined/exp/tri3/graph

        lab_name=lab_mono
        lab_folder=/home/subash/kaldi/egs/Combined/exp/tri3_ali_test
        lab_opts=ali-to-phones --per-frame=true
        lab_count_file=none
        lab_data_folder=/home/subash/kaldi/egs/Combined/data/test/
        lab_graph=/home/subash/kaldi/egs/Combined/exp/tri3/graph
n_chunks = 5

[data_use]
train_with = EMAMAE_tr
valid_with = EMAMAE_dev
forward_with = EMAMAE_test

[batches]
batch_size_train = 8
max_seq_length_train = 1000
increase_seq_length_train = True
start_seq_len_train = 100
multply_factor_seq_len_train = 2
batch_size_valid = 8
max_seq_length_valid = 1000

[architecture1]
arch_name = liGRU_layers
arch_proto = proto/liGRU.proto
```

arch_library = neural_networks

arch_class = liGRU

arch_pretrain_file = none

arch_freeze = False

arch_seq_model = True

ligru_lay = 550,550,550,550,550

ligru_drop = 0.1,0.1,0.1,0.1,0.1

ligru_use_laynorm_inp = False

ligru_use_batchnorm_inp = False

ligru_use_laynorm = False,False,False,False,False

ligru_use_batchnorm = True,True,True,True,True

ligru_bidir = True

ligru_act = relu,relu,relu,relu,relu

ligru_orthinit = True

arch_lr = 0.0004

arch_halving_factor = 0.5

arch_improvement_threshold = 0.001

arch_opt = adam

opt_momentum = 0.0

opt_alpha = 0.95

opt_eps = 1e-8

opt_centered = False

opt_weight_decay = 0.0

opt_betas = 0.9,0.999

opt_amsgrad = False

[architecture2]

arch_name = MLP_layers

arch_proto = proto/MLP.proto

arch_library = neural_networks

arch_class = MLP

arch_pretrain_file = none

arch_freeze = False

arch_seq_model = False

dnn_lay = 2064

dnn_drop = 0.0

dnn_use_laynorm_inp = False

dnn_use_batchnorm_inp = False

dnn_use_batchnorm = False

dnn_use_laynorm = False

dnn_act = softmax

arch_lr = 0.0004

arch_halving_factor = 0.5

arch_improvement_threshold = 0.001

arch_opt = adam

opt_momentum = 0.0

opt_alpha = 0.95

opt_eps = 1e-8

opt_centered = False

opt_weight_decay = 0.0

opt_betas = 0.9,0.999

opt_amsgrad = False

[architecture3]

arch_name = MLP_layers2

arch_proto = proto/MLP.proto

arch_library = neural_networks

arch_class = MLP

arch_pretrain_file = none

arch_freeze = False

```
arch_seq_model = False
dnn_lay = 166
dnn_drop = 0.0
dnn_use_laynorm_inp = False
dnn_use_batchnorm_inp = False
dnn_use_batchnorm = False
dnn_use_laynorm = False
dnn_act = softmax
arch_lr = 0.0004
arch_halving_factor = 0.5
arch_improvement_threshold = 0.001
arch_opt = adam
opt_momentum = 0.0
opt_alpha = 0.95
opt_eps = 1e-8
opt_centered = False
opt_weight_decay = 0.0
opt_betas = 0.9,0.999
opt_amsgrad = False

[model]
model_proto = proto/model.proto
model = out_dnn1=compute(liGRU_layers,fmllr)
        out_dnn2=compute(MLP_layers,out_dnn1)
        out_dnn3=compute(MLP_layers2,out_dnn1)
        loss_mono=cost_nll(out_dnn3,lab_mono)
        loss_mono_w=mult_constant(loss_mono,1.0)
        loss_cd=cost_nll(out_dnn2,lab_cd)
        loss_final=sum(loss_cd,loss_mono_w)
        err_final=cost_err(out_dnn2,lab_cd)
```

[forward]

forward_out = out_dnn2

normalize_posteriors = True

normalize_with_counts_from = /storage/subash/Pytorch-exps/Combined/liGRU_fmllr/exps/0.1_0.0004/exp_files/forward_out_dnn2_lab_cd.count

save_out_file = False

require_decoding = True

[decoding]

decoding_script_folder = kaldi_decoding_scripts/

decoding_script = decode_dnn.sh

decoding_proto = proto/decoding.proto

min_active = 200

max_active = 7000

max_mem = 50000000

beam = 13.0

latbeam = 8.0

acwt = 0.2

max_arcs = -1

skip_scoring = false

scoring_script = local/score_wsj.sh

scoring_opts = "--min-lmwt 1 --max-lmwt 10"

norm_vars = False

REFERENCES

"Mandarin Phonemic Inventory by ASHA." Retrieved 2/14/2020, from https://www.asha.org/uploadedFiles/practice/multicultural/MandarinPhonemicInventory.pdf.

Abdou, S., M. Rashwan, H. Al-Barhamtoshy, K. Jambi and W. Al-Jedaibi (2012). Enhancing the Confidence Measure for an Arabic Pronunciation Verification System. Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training June.

Afshine Amidi, S. A. Retrieved 2/13/2020, from https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks.

Anastasakos, T., J. McDonough, R. Schwartz and J. Makhoul (1996). A compact model for speaker-adaptive training. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, IEEE.

Baum, L. E., T. Petrie, G. Soules and N. Weiss (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." The annals of mathematical statistics **41**(1): 164-171.

Bell, P., P. Swietojanski and S. Renals (2016). "Multitask learning of context-dependent targets in deep neural network acoustic models." IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(2): 238-247.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010, Springer**:** 177-186.

Bozorg, N. and M. T. Johnson (2018). Comparing performance of acoustic-to-articulatory inversion for mandarin accented english and american english speakers. 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE.

Browman, C. P. and L. Goldstein (1992). "Articulatory phonology: An overview." Phonetica **49**(3-4): 155-180.

Butler-Pascoe, M. E. (2011). "The history of CALL: The intertwining paths of technology and second/foreign language teaching." International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT) **1**(1): 16-32.

Catford, J., J. Palmer, J. Dew, R. Barry, H. Cheng, V. Hsu and Y. Li (1974). "A contrastive study of English and Mandarin Chinese." Defense Language Institute.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078.

Davis, K. H., R. Biddulph and S. Balashek (1952). "Automatic recognition of spoken digits." The Journal of the Acoustical Society of America **24**(6): 637-642.

Deller, J. R., J. G. Proakis and J. H. Hansen (2000). Discrete-time processing of speech signals, Institute of Electrical and Electronics Engineers.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society: Series B (Methodological) **39**(1): 1-22.

Deterding, D. (2006). " The pronunciation of English by speakers from China." English World-Wide **27**(2): 175-198.

Deterding, D. (2010). "ELF-based Pronunciation Teaching in China." Chinese Journal of Applied Linguistics (Foreign Language Teaching & Research Press) **33**(6).
Duchi, J., E. Hazan and Y. Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization." Journal of machine learning research **12**(Jul): 2121-2159.
Edwards, J. G. H. and M. L. Zampini (2008). Phonology and second language acquisition, John Benjamins Publishing.
Eslan. "Chinese Pronunciation Problem in English." Retrieved 2/15/2020, from http://englishspeaklikenative.com/resources/common-pronunciation-problems/chinese-pronunciation-problems/.
Ethnologue (2019). "Summary by language size."
Ferguson, J. (1980). "Hidden Markov analysis: an introduction." Hidden Markov Models for Speech.
Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems." Annals of eugenics **7**(2): 179-188.
Flanagan, J. L. (2013). Speech analysis synthesis and perception, Springer Science & Business Media.
Flege, J. E. (1995). "Second language speech learning: Theory, findings, and problems." Speech perception and linguistic experience: Issues in cross-language research **92**: 233-277.
Forney, G. D. (1973). "The viterbi algorithm." Proceedings of the IEEE **61**(3): 268-278.
Gales, M. J. (1998). "Maximum likelihood linear transformations for HMM-based speech recognition." Computer speech & language **12**(2): 75-98.
Gales, M. J. (1999). "Semi-tied covariance matrices for hidden Markov models." IEEE transactions on speech and audio processing **7**(3): 272-281.
Gick, B., I. Wilson, K. Koch and C. Cook (2004). "Language-specific articulatory settings: Evidence from inter-utterance rest position." Phonetica **61**(4): 220-233.
Gilakjani, A. P. and M. R. Ahmadi (2011). "Why Is Pronunciation So Difficult to Learn?" English language teaching **4**(3): 74-83.
Harrison, A. M., W.-K. Lo, X.-j. Qian and H. Meng (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. International Workshop on Speech and Language Technology in Education.
Hinofotis, F. and K. Bailey (1980). "American undergraduates' reactions to the communication skills of foreign teaching assistants." On TESOL **80**: 120-133.
Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen and T. N. Sainath (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal processing magazine **29**(6): 82-97.
Hinton, G., N. Srivastava and K. Swersky (2012). "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent." Cited on **14**(8).
Hochreiter, S. and J. Schmidhuber (1997). "Long short-term memory." Neural computation **9**(8): 1735-1780.
Hu, W., Y. Qian and F. K. Soong (2015). An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech. SLaTE.

Huang, M. and L. Pickering (2014). Revisiting the pronunciation of english by speakers from mainland china. Pronunciation in second language learning and teaching conference (issn 2380-9566).
Huang, X., A. Acero, H.-W. Hon and R. Foreword By-Reddy (2001). Spoken language processing: A guide to theory, algorithm, and system development, Prentice hall PTR.
I., L. V. (1966). "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." Soviet Physics Doklady **10**: 707.
ICT4LT. "Introduction to Computer Assisted Language Learning (CALL)." Module 1.4 Retrieved 3/13/2020, from http://www.ict4lt.org/en/index.htm.
Ji, A., J. J. Berry and M. T. Johnson (2014). The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
Johnson, M. T. (2018). Speech Recognition system-HMM.
Juang, B.-H. and L. R. Rabiner (2005). "Automatic speech recognition–a brief history of the technology development." Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara **1**: 67.
Kingma, D. P. and J. Ba (2014). "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980.
Leung, W.-K., X. Liu and H. Meng (2019). CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
Levinson, S. E., L. R. Rabiner and M. M. Sondhi (1983). "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition." Bell System Technical Journal **62**(4): 1035-1074.
Levy, M. and P. Hubbard (2005). "Why call call "CALL"?" Computer Assisted Language Learning **18**(3): 143-149.
Li, K., X. Qian and H. Meng (2016). "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(1): 193-207.
Li, S. and L. Wang (2012). Cross linguistic comparison of Mandarin and English EMA articulatory data. Thirteenth Annual Conference of the International Speech Communication Association.
Li, W., K. Li, S. M. Siniscalchi, N. F. Chen and C.-H. Lee (2016). Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-Guided and Data-Driven Decision Trees. Interspeech.
Li, W., S. M. Siniscalchi, N. F. Chen and C.-H. Lee (2016). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
Livescu, K., P. Jyothi and E. Fosler-Lussier (2016). "Articulatory feature-based pronunciation modeling." Computer Speech & Language **36**: 212-232.
Lo, W.-K., S. Zhang and H. Meng (2010). Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. Eleventh Annual Conference of the International Speech Communication Association.

Lo, W. K., A. M. Harrison, H. Meng and L. Wang (2008). Decision fusion for improving mispronunciation detection using language transfer knowledge and phoneme-dependent pronunciation scoring. 2008 6th International Symposium on Chinese Spoken Language Processing, IEEE.
Mao, S., Z. Wu, R. Li, X. Li, H. Meng and L. Cai (2018). Applying Multitask Learning to Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
Mao, S., Z. Wu, X. Li, R. Li, X. Wu and H. Meng (2018). Integrating Articulatory Features into Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech. 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE.
McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." The bulletin of mathematical biophysics **5**(4): 115-133.
Meng, H., E. Zee and W. S. Lee (2007). "A contrastive phonetic study between Cantonese and English to predict salient mispronunciations by Cantonese learners of English." Unpublished article. The Chinese University of Hong Kong.
Neri, A., O. Mich, M. Gerosa and D. Giuliani (2008). "The effectiveness of computer assisted pronunciation training for foreign language learning by children." Computer Assisted Language Learning **21**(5): 393-408.
Nissen, S. L., C. Dromey and C. Wheeler (2007). "First and second language tongue movements in Spanish and Korean bilingual speakers." Phonetica **64**(4): 201-216.
Povey, D., L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát and A. Rastrow (2011). "The subspace Gaussian mixture model—A structured model for speech recognition." Computer Speech & Language **25**(2): 404-439.
Povey, D., S. M. Chu and B. Varadarajan (2008). Universal background model based speech recognition. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE.
Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian and P. Schwarz (2011). The Kaldi speech recognition toolkit. IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society.
Povey, D. and G. Saon (2006). Feature and model space speaker adaptation with full covariance Gaussians. Ninth International Conference on Spoken Language Processing.
Psutka, J. V. (2007). Benefit of maximum likelihood linear transform (MLLT) used at different levels of covariance matrices clustering in ASR systems. International Conference on Text, Speech and Dialogue, Springer.
Qian, X., H. Meng and F. Soong (2016). "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **24**(6): 1020-1028.
Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE **77**(2): 257-286.
Ravanelli, M., P. Brakel, M. Omologo and Y. Bengio (2018). "Light gated recurrent units for speech recognition." IEEE Transactions on Emerging Topics in Computational Intelligence **2**(2): 92-102.

Ravanelli, M., T. Parcollet and Y. Bengio (2019). The pytorch-kaldi speech recognition toolkit. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
Recasens, D. (2010). "Differences in base of articulation for consonants among Catalan dialects." Phonetica **67**(4): 201-218.
Rice, L. (April 1976). "Hardware & software for speech synthesis." Dr. Dobb's Journal of Computer Calisthenics & Orthodontia. **1**: 6–8.
Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms, Cornell Aeronautical Lab Inc Buffalo NY.
Rumelhart, D. E., G. E. Hinton and R. J. Williams (1985). Learning internal representations by error propagation, California Univ San Diego La Jolla Inst for Cognitive Science.
Ryu, H. and M. Chung (2017). Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features. SLaTE.
Sanguineti, V., R. Laboissiere and Y. Payan (1997). "A control model of human tongue movements in speech." Biological cybernetics **77**(1): 11-22.
Schönle, P. W., K. Gräbe, P. Wenig, J. Höhne, J. Schrader and B. Conrad (1987). "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract." Brain and Language **31**(1): 26-35.
Statista. (2019). "ELT market size in China 2017-2022." Retrieved 2/23/2020, from https://www.statista.com/statistics/967696/china-english-language-training-market-size/.
Stone, M. and A. Lundberg (1996). "Three-dimensional tongue surface shapes of English consonants and vowels." The Journal of the Acoustical Society of America **99**(6): 3728-3737.
Tepperman, J. and S. Narayanan (2007). "Using articulatory representations to detect segmental errors in nonnative pronunciation." IEEE transactions on audio, speech, and language processing **16**(1): 8-22.
Viikki, O. and K. Laurila (1998). "Cepstral domain segmental feature vector normalization for noise robust speech recognition." Speech Communication **25**(1-3): 133-147.
Wang, J., J. R. Green, A. Samal and Y. Yunusova (2013). "Articulatory distinctiveness of vowels and consonants: A data-driven approach." Journal of Speech, Language, and Hearing Research.
Wei, R. and J. Su (2012). "The statistics of English in China: An analysis of the best available data from government sources." English Today **28**(3): 10-14.
Weinreich, U. (1953). "Languages in Contact, New York." Publications of the Linguistic Circle of NY.
Wieling, M. and M. Tiede (2017). "Quantitative identification of dialect-specific articulatory settings." The Journal of the Acoustical Society of America **142**(1): 389-394.
Wieling, M., F. Tomaschek, D. Arnold, M. Tiede, F. Bröker, S. Thiele, S. N. Wood and R. H. Baayen (2016). "Investigating dialectal differences using articulography." Journal of Phonetics **59**: 122-143.
Wieling, M., P. Veenstra, P. Adank and M. Tiede (2017). Articulatory differences between L1 and L2 speakers of English. Proceedings of The 11th International Seminar on Speech Production, Tianjin, China, October.

Wieling, M., P. Veenstra, A. Weber, P. Adank and M. Tiede (2015). Comparing L1 and L2 speakers using articulography. Proceedings of the 18th International Congress of Phonetic Sciences, International Phonetic Association.
Wilson, I. (2013). "Articulatory settings of French and English monolinguals." Sophia University Working Papers in Phonetics: 39-58.
Witt, S. M. and S. J. Young (2000). "Phone-level pronunciation scoring and assessment for interactive language learning." Speech communication **30**(2-3): 95-108.
Yang, C., F. K. Soong and T. Lee (2007). "Static and dynamic spectral features: Their noise robustness and optimal weights for ASR." IEEE transactions on audio, speech, and language processing **15**(3): 1087-1097.
Yuan, H., J. Zhao and J. Liu (2012). Improve mispronunciation detection with Tandem feature. 2012 8th International Symposium on Chinese Spoken Language Processing, IEEE.
Zhang, F. and P. Yin (2009). "A study of pronunciation problems of English learners in China." Asian social science **5**(6): 141-146.

VITA

Subash Khanal, Master student

Department of Electrical and Computer Engineering, University of Kentucky

EDUCATION

Master of Science- Electrical Engineering, August 2020

University of Kentucky

Thesis: Mispronunciation Detection and Diagnosis in Mandarin Accented English Speech

Director of Thesis: Dr. Michael T. Johnson

Bachelor of Engineering- Electronics and Communication Engineering, August 2016

Visvesvaraya Technological University, Belgaum, India