DISSERTATION

Narjes Bozorg

The Graduate School
University of Kentucky
2020

ARTICULATORY-WAVENET: DEEP AUTOREGRESSIVE MODEL FOR ACOUSTIC-TO-ARTICULATORY INVERSION

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Narjes Bozorg
Lexington, Kentucky

Director: Dr. Michael T. Johnson, Professor of
Electrical and Computer Engineering
Lexington, Kentucky 2020

ABSTRACT OF DISSERTATION

ARTICULATORY-WAVENET: DEEP AUTOREGRESSIVE MODEL FOR ACOUSTIC-TO-ARTICULATORY INVERSION

Acoustic-to-Articulatory Inversion, the estimation of articulatory kinematics from speech, is an important problem which has received significant attention in recent years. Estimated articulatory movements from such models can be used for many applications, including speech synthesis, automatic speech recognition, and facial kinematics for talking-head animation devices. Knowledge about the position of the articulators can also be extremely useful in speech therapy systems and Computer-Aided Language Learning (CALL) and Computer-Aided Pronunciation Training (CAPT) systems for second language learners.

Acoustic-to-Articulatory Inversion is a challenging problem due to the complexity of articulation patterns and significant inter-speaker differences. This is even more challenging when applied to non-native speakers without any kinematic training data. This dissertation attempts to address these problems through the development of upgraded architectures for Articulatory Inversion. The proposed Articulatory-WaveNet architecture is based on a dilated causal convolutional layer structure improves the Acoustic-to-Articulatory Inversion estimated results for both speaker-dependent and speaker-independent scenarios.

The system has been evaluated on the ElectroMagnetic Articulography corpus of Mandarin Accented English (EMA-MAE) corpus, consisting of 39 speakers including both native English speakers and Mandarin accented English speakers. Results show that Articulatory-WaveNet improves the performance of the speaker-dependent and speaker-independent Acoustic-to-Articulatory Inversion systems significantly compared to the previously reported results.

KEYWORDS: acoustic-to-articulatory inversion, Electro-Magnetic Articulography, speaker-dependent, speaker-independent, WaveNet, deep autoregressive model

Author's signature: Narjes Bozorg

Date: November 3, 2020

ARTICULATORY-WAVENET: DEEP AUTOREGRESSIVE MODEL FOR ACOUSTIC-TO-ARTICULATORY INVERSION

By
Narjes Bozorg

Director of Dissertation: Michael T. Johnson

Director of Graduate Studies: Daniel Lau

Date: November 3, 2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Statement of The Problem and Motivations

Speech production is a highly complex task involving synchronized motor control of more than 100 different muscles. The problem of estimating articulatory characteristics (like tongue, lips, and jaw movements) from an acoustic signal is known as Acoustic-to-Articulatory Inversion (AAI).

The study of AAI plays an important role in many fields of study related to pronunciation training and language understanding [1] [2]. While intuitively a simple function of one's ability to perceive and produce native-like speech sounds, pronunciation is in fact one of the most complicated characteristics of human speech, resulting from a complex interaction across multiple linguistic levels that reflects structural and functional differences between native and nonnative speakers.

The successful outcome of pronunciation assessments can be useful for helping second language (L2) learners to elevate their social and professional communications with native speakers and to improve their conversational effectiveness with native English (L1) speakers who serve as students, colleagues and patrons to L2 speakers of English.

One way to learn how we can improve L2 pronunciation is to compare it with L1 pronunciation. In order to develop more effective methods for pronunciation error correction, it is essential to understand variations between native and second language speaker articulatory patterns for different phonetic conditions. For instance, the structural and functional differences between Mandarin accented English speakers and native English speakers have predicted some of the known characteristics of L2 speakers and provided a basis for generating hypotheses regarding other factors that influence the extent of perceived accent and the optimal foci for pronunciation modification. AAI results analysis can help us understand these differences between the characteristics of L1 and L2 speakers.

In addition to using AAI to compare L1 and L2 speakers, AAI has direct application to instructional technologies for developing the L2 pronunciation such as Computer-Aided Language Learning (CALL) [3, 4, 5], and Computer-Aided Pronunciation Training (CAPT) [3, 4, 5, 6]. In second language acquisition, learners continue to have difficulty attaining pronunciation like that of native language speakers, even given massive individual pronunciation training. The quintessential issue is that learners lack specific unique expertise on how to map the installed speech generated patterns of their native language onto a different set of phonemes. Improvement in this expertise can be gained using AAI for CALL and CAPT which helps L2 speakers to improve their pronunciation. CALL systems use AAI to create precise visualizations that enable second language learners to adjust their articulators for better pronunciation.

For many applications like CALL and CAPT, inversion needs to be implemented

on new unknown speakers, for whom a small amount of acoustic adaptation data can easily be collected but obtaining kinematic data is infeasible. Many traditional approaches to AAI are Speaker Dependent (SD), modeling parallel acoustic and articulatory spaces by using the information from a known target speaker to create a mapping representation. Articulatory movement tracking and recording methods are also very time consuming and expensive, requiring kinematic tracking equipment and the proper environment to obtain accurate aligned acoustic-articulatory data [7, 8, 9], which limits the applications where this is possible. To address these problems, this work also includes AAI models for unknown target speakers without articulatory training data, which we refer to as Speaker Independent-AAI (SI-AAI).

1.2 Challenges of Acoustic-to-Articulatory Inversion

Acoustic to articulatory inversion is a challenging task. AAI mappings are nonlinear and non-unique [10, 11], which means a given articulatory state has always only one acoustic realization, but an acoustic signal can be the outcome of more than one articulator state.

Current AAI methods have been mostly focused on estimating articulatory parameters using matched acoustic-kinematic data for a specific speaker, but there has been less success with the broader task of speaker-independent modeling for estimating articulatory trajectories for speakers with no kinematic training data and only a small quantity of acoustic adaptation records. This is true for applications like Automatic Speech Recognition (ASR) [7, 8, 9, 12, 13] , audio-visual speech synthesis [7, 12, 13], and animated virtual talking head applications (which utilizes augmented reality to displays the motions of speech articulators) [13], in which only acoustic information is available, and there is no access to articulatory trajectories.

Inspired by the success of deep learning methods in different applications including kinematic inversion and to address the aforementioned problems, this dissertation introduces a new method using deep autoregressive Articulatory-WaveNet for learning the nonlinear mapping between acoustic waveforms and kinematic trajectories for both SD-AAI and SI-AAI scenarios.

WaveNet [14] introduced a novel approach to speech synthesis based on the point-to-point prediction of the raw audio signals. Inspired by the success of WaveNet architectures in different speech synthesis tasks [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24], we decided to deploy the stacked dilated convolutional layers for AAI. This dissertation shows that applying Articulatory-WaveNet for speaker-dependent and speaker-independent AAI will not only promote the previous results, it also provides higher consistency for estimating articulatory trajectories for the different subgroup of speakers including Male/Female, L1/L2, and even for unknown speakers.

1.3 Contribution of The Work

This dissertation includes several different contributions in the articulatory-speech analysis domain. The main goal of the work is to improve the performance of AAI

using state of the art approaches. Articulatory-WaveNet has been presented here for this purpose and improves AAI for both speaker-dependent and speaker-independent frameworks. Specific contributions include the following:

1.3.1 L1 Vs. L2 Analysis

This work studies different aspects of articulatory-acoustic differences and similarities between L1 and L2 speakers. Language learners struggle with both perceiving and producing vowels that do not exist in L1 [25]. On the other hand, L2 vowels which are similar to L1 vowels are easy for L2 learners to pronounce but are often considered as identical so that small differences in pronunciation or co-articulation with other sounds may never be developed at all.

This dissertation compares the articulatory configurations of native English speakers and Mandarin accented English speakers. The comparison is made between English vowels that have corresponding vowels in Mandarin, versus those that do not, with results supporting the idea that variability of articulator positioning in L2 speakers is larger for vowels that are unique to English than for those that have corresponding vowels in the native language.

1.3.2 Comparing The Performance of SD-AAI For L1 and L2 Speakers

This work has studied different aspects of SD-AAI using the EMA-MAE corpus. The performance of SD-AAI for both L1 and L2 speakers of English, as a function of the number of Gaussian Mixtures used in the inversion model, has been investigated. The inversion system is based on a Hidden Markov Model-Gaussian Mixture Model (HMM-GMM)approach and is implemented on different native English and native Mandarin speakers of English.

The consistency and predictability of articulatory trajectory patterns between L1 and L2 speakers of English have been also compared for the GMM-HMM AAI model. Results indicate that Mandarin accented English speakers have more explicit, i.e. lower Root-Mean-Squared error (RMSE), estimated articulatory trajectories than native American English speakers for vertical motion of the primary articulators in the mid-sagittal plane, but less precise estimated trajectories for horizontal, lateral, and non-midsagittal vertical directions.

Based on these results, we hypothesize that the Mandarin L2 speakers focus intently on the movement of a small number of primary articulators in specific directions, to the detriment of other parts of their articulatory patterns including horizontal tongue positioning and lateral tongue curvature.

1.3.3 Comparing The Performance of SI-AAI For L1 and L2 Speakers

We have considered different architectures for SI-AAI using the EMA-MAE dataset, including introducing a new speaker adaptation method based on using different adaptation approaches for the acoustic model and the weighted articulatory model. In this approach, the acoustic model for the unknown target speaker is adapted by the Maximum Likelihood Linear Regression (MLLR) model, while the articulatory

model has adapted with the Parallel Reference Speaker Weighting (PRSW) method. Results show a combination of the MLLR and PRSW outperforms parallel-PRSW and previous SI-AAI frameworks with results very close to the SD-AAI methods.

In addition, an investigation of the characteristics of the most effective reference speaker set for the Parallel Reference Speaker Weighting (PRSW) algorithm for kinematic- independent acoustic-to-articulatory inversion has been implemented. To obtain the adaptation weights for estimating the articulatory model, different reference speaker accent types and quantities have been acquired. The reference speaker sets have been selected not only based on their performance in speaker-dependent kinematic-inversion but also based on the type of accent. A comparison is made between different types of target speakers and reference speakers with results indicating that the accuracy of the adapted model increases when we select balanced distributed accents of English and lower number of speakers.

1.3.4 SD-AWN

The main objective of this dissertation is to find an efficient framework for AAI that improves on previous methods. Articulatory-WaveNet is presented as a new model for acoustic-to-articulator inversion. The system uses the WaveNet speech synthesis architecture, with dilated causal convolutional layers using previous values of the predicted articulatory trajectories conditioned on acoustic features.

Results show significant improvement in both correlation and RMSE between the generated and true articulatory trajectories for the new method, with an average correlation of 0.83, representing a 36% relative improvement over the 0.61 correlation obtained with a baseline HMM-GMM AAI framework. To the best of our knowledge, this work presents the first application of a point-by-point waveform synthesis approach to the problem of AAI and the results show improved performance compared to previous methods for SD-AAI.

1.3.5 SI-AWN

This work also introduces a new speaker-independent method for AAI. The proposed architecture, Speaker Independent-Articulatory WaveNet (SI-AWN), models the relationship between acoustic and articulatory features by conditioning the articulatory trajectories on acoustic features and then utilizing the structure for unseen target speakers. The proposed SI-AWN is evaluated on the Electro-Magnetic Articulography corpus of L1 and L2 speakers, using the pool of acoustic-articulatory information from 35 reference speakers and testing on target speakers that include male, female, native and non-native speakers. The results suggest that SI-AWN improves the performance of the AAI process compared to the baseline Maximum Likelihood Regression-Parallel Reference Speaker Weighting (MLLR-PRSW) method by 21 percent. This is the first application of a WaveNet synthesis approach to the problem of SI-AAI, and results are comparable to or better than the best currently published systems.

Chapter 2 Background and Related Works

Acoustic-to-Articulatory Inversion (AAI) maps from acoustic to articulatory space to estimate articulatory movements from acoustic data. The accurate and robust estimation of AAI can be used for many applications and speech technologies.

This chapter aims to provide a general background needed for articulatory and speech domain analysis. This includes an overview of articulatory datasets such as the bilingual multi-speaker corpus of parallel acoustic and EMA kinematic data, EMA-MAE, used throughout this dissertation. A literature review of different Machine Learning methods for SD/SI-AAI including classic Machine Learning and Deep Learning methods is included, as well as an overview of the WaveNet speech synthesis architecture which is used in this dissertation for application to AAI.

2.1 Articulatory Data Acquisition and Electromagnetic Articulography

Speech signals can be characterized by two parallel and interconnected representations: acoustic and articulatory spaces. Acoustic data is to the speech signal transmitted by the speaker to the listener, represented by an acoustic feature space that captures the frequency information within this signal, while the articulatory space is the kinematic motion of the underlying speech production system that generates and forms a speech signal. Although most research in speech processing focuses on the acoustic domain, there are many ways in which characterizing the articulatory domain can be utilized for tackling speech processing problems.

Many attempts have been made to use articulatory characteristics for representing speech signals, for technologies like automatic speech recognition and speech synthesis. For example, Mcdermott and Nakamura [26] have used articulatory data for automatic speech recognition and King et al. [27] uses an articulatory HMM-based framework for speech synthesis.

There are a number of articulography methods for extracting articulatory features and modeling the kinematic system. Methods like ElectroPalatoGraphy (EPG), ElectroMagnetic Articulography (EMA), X-radiation (X-ray) cinematography, ultrasound and Magnetic Resonance Imaging (MRI) have been used previously to track the articulatory movements that are involved in speech production.

Corpora based on these modalities include the Wisconsin X-Ray Micro Beam (XRM) [28], AIMSL Chinese [29, 30], MOCHA [31], MNGU0 [32], EUR-ACCOR multi-language articulatory [33], Edinburgh speech production facility Double Talk [34], XRMB [35], and EMA-IEEE [36]. These datasets have been used by many researchers for different investigations and articulatory analysis.

The main problem with most of these corpora is the lack of diversity among speakers. For applications like AAI, many different types of speakers in the training model stage are needed to provide a robust and correct estimation of predictions.

Therefore, this dissertation has selected the EMA-MAE dataset, a bilingual multi-speaker corpus, for AAI and pronunciation learning analysis. The EMA-MAE corpus [37] has significantly more speaker variability compared to other common datasets like MNGU0 and MOCHA which contain records from just one or two speakers.

The next section explains the differences and similarities between the two different groups of accented (L2) and native English (L1) speakers. Later it will be shown how the EMA dataset can be used for pronunciation learning and improving the pronunciation assessment techniques for the L2 group of speakers.

2.2 EMA-MAE Corpus

The EMA-MAE corpus [37] provides high temporal and spatial resolution parallel acoustic and articulatory data. To collect the articulatory dataset, a Northern Digital NDI Wave Speech Research system was utilized with five degrees of freedom sensors (three-dimensional position plus two-dimensional sensor plane orientation) at a 400Hz sampling rate. Data were collected in a sound-attenuating acoustic booth, with time-synced acoustic data. The sampling rate for collecting acoustic is set at 22050 Hz. The system works through the use of small toroidal electromagnetic sensors within a static electromagnetic field.

A reference sensor is mounted in such a way that it moves with the subject's head without changing position or relative orientation, and this sensor is used to establish a base coordinate system. Other sensors are then attached to the articulators to collect both position and orientation data.

The reference sensor is a slightly larger 6 Degree of Freedom (DOF) sensor which captures the 3-dimensional position as well as the full 3-dimensional orientation of the sensor relative to a known base orientation. The articulator sensors capture 5 DOF information, including 3-dimensional position information plus the 2-dimensional orientation of the sensor plane.

40 speakers in the EMA-MAE corpus have been divided into two subject categories L1 and L2, each of which consists of 10 women and 10 men. The L1 group included native English speakers with the upper Midwest American English dialect. The L2 group consists of Mandarin Chinese speakers who speak English as a second language. This group is further divided into Northern Beijing-region and Southern Shanghai dialect region, with 5 females and 5 male speakers from each.

These speakers are between the ages of 18 and 40 and they have no history of using anti-convulsion, anti-psychotic, or anti-anxiety medicines. They also have no history of speech, language and hearing disorder, history of orofacial surgery (other than typical dental extractions), or other influential parameters on EMA results.

Following the data collection, one subject (Female Mandarin speaker) was found to have errors in the majority of the collected kinematic data and is therefore unusable for use with acoustic articulatory inversion studies or other studies related to articulatory kinematics. This leaves a set of 39 speakers usable for the present investigation.

For each individual speaker, about 45 minutes of acoustic and articulatory data have been collected, including word, sentence, and paragraph-level speech samples.

The record has been phonetically transcribed with particular interest to the distinctive features of MAE, with the aid of a team of transcribers with a common upper-Midwestern dialect base and practical training and experience in extensive and narrative phonetic transcription.

2.3 Sensors Specifications and Articulatory Space

Every speaker wore a pair of plastic glass to which a reference sensor was attached, to compensate for head movements. In addition, a bite plate was used to locate the maxillary occlusal plane and the midsagittal plane of each speaker. Head correction is built into the NDI Wave software, while bite plate calibration was implemented in a post-processing step by rotating the original head-calibrated coordinate space to a new Cartesian articulatory space defined relative to the individual speaker's midsagittal and maxillary occlusal planes. Based on this configuration, the anterior-posterior movements form the x-axis, superior-inferior movements form the y-axis and the lateral movements are represented by the z axis. Accordingly, xy plane represents the speaker's mid-sagittal workspace, the xz plane is the maxillary occlusal plane. Figure 2.1 illustrates this articulatory referenced coordinate system and indicates the Cartesian origin location which is the central point of the upper maxillary incisors.

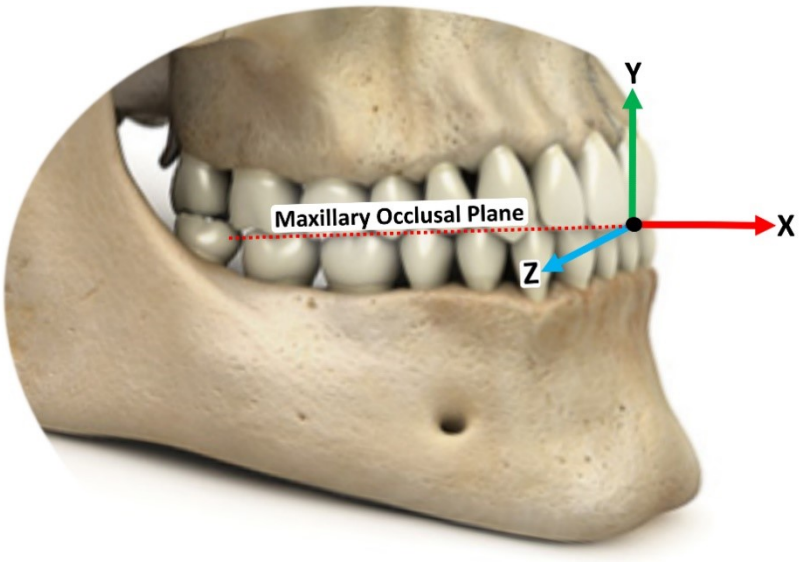


Figure 2.1: Articulatory Referenced Coordinate System

The articulatory sensors include the reference sensor (REF), jaw sensor at the lower Middle Incisor (MI), Lower Lip (LL), Upper Lip (UL), Tongue Dorsum (TD), and Tongue Apex (TA), all placed in the mid-sagittal plane. In addition, there were two lateral sensors, one at the Lip Corner (LC) of the mouth to help indicate lip rounding and one in the left central midpoint of the tongue body to help indicate Lateral Tongue curvature (LT). Figure 2.2 represents the location of the applied articulatory sensors.

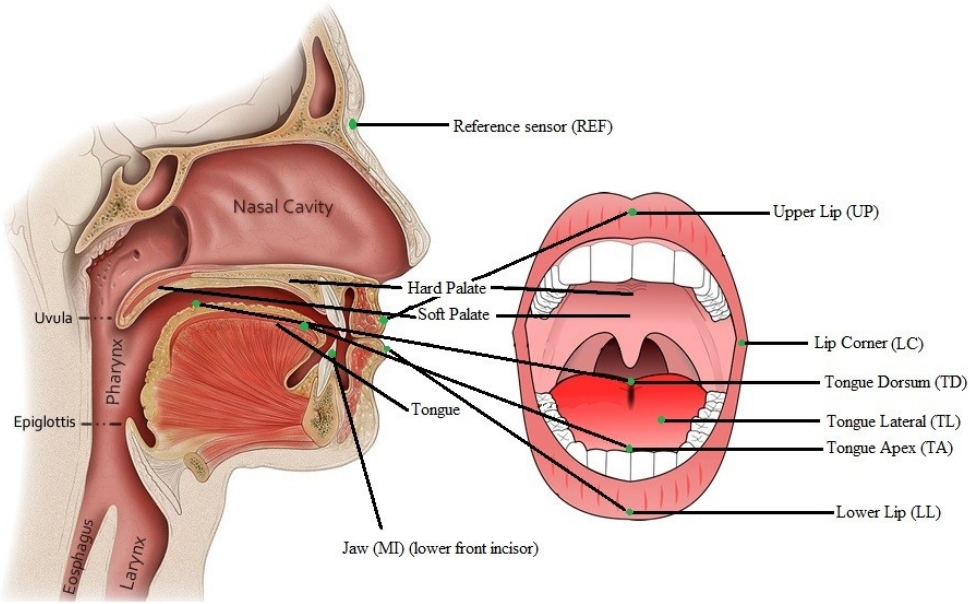


Figure 2.2: EMA-MAE sensor layout. The sensor locations and articulators have been shown on the cross-section of the human head and oral cavity.

2.4 Articulatory Features

While sensor position data provides a simple representation of articulatory motion, there are several reasons that this may not be the optimal representation for use in acoustic-to-articulatory inversion. Among these is the fact that raw sensor position is not itself demonstrative of the shape of the vocal tract during speech production.

The acoustics of speech is largely driven by the cross-section of the vocal tract. Given that sensor position data only provides information about a small number of locations in the vocal tract, this measure cannot provide meaningful information about the corresponding acoustics without any reference to the surrounding vocal tract parameters.

A normalized and palate-referenced set of articulatory features are more informative than direct sensor coordinates. This is based on the idea that acoustics are primarily driven by the cross-section of the vocal tract opening and therefore sensor-to-palate vertical distances are more related to acoustic properties than simple sensor position.

The reliable and meaningful set of articulatory features for characterizing the vocal tract from EMA measurements have been presented in this work.

2.4.1 Horizontal Normalization For VT1, VT3, VT5 VT7

The distance between the center incisors and the middle point of the back molar has been calculated from the biteplate score of every speaker, denoted K_x and used as the normalization scale. Therefore, the horizontal articulatory characteristics like Horizontal Lip Protrusion (VT7) and Horizontal Tongue Apex, Lateral, and Dorsum (VT1, 3 and 5) are described in Table 2.1 are all calculated directly from sensor position divided by this normalization constant.

The normalization process is implemented through following equation:

$$NormalizedArticulatoryFeature = \frac{RawSensorData}{K_x} \tag{2.1}$$

Where K_x measured horizontal distance from front incisor to back molar in mm.

2.4.2 Palate-to-Sensor Distance for VT2, VT4, VT6

For each speaker, the trace of the mid-sagittal palate line, a series of horizontal traces across the palate, and both inner perimeter and outer perimeter dental traces at the gum line were recorded. Together with the biteplate record, this information provides reference data that can be used to calculate physiologically referenced vocal tract measures.

The vertical (y-axis) variables VT2, VT4, and VT6 are computed directly from the vertical distance between the sensor position and the palate. This represents the vocal tract opening at the sensor positions, including two mid-sagittal positions and one lateral position.

Assuming that palate (VTx, VTz) be the thin plate spline interpolation of the mesh at the (x , z) location for a particular sensor, the aforementioned vertical articulatory features can be computed as follows:

$$VT = palate(VT_x, VT_z) - VT_y \tag{2.2}$$

2.4.3 Lip Protrusion and Separation VT7, VT8

Lip protrusion VT7 is taken from the mean value of all biteplate Horizontal Upper Lip sensor values subtracted from the subject's Horizontal Upper Lip sensor value divided by this normalization constant.

$$VT7 = \frac{UL_x - |mean(UL_x)_{BiteplateData}|}{K_x} \tag{2.3}$$

Vertical lip separation VT8 is calculated via:

$$VT8 = (UL_y - LL_y) - 0.1Percentile(UL_y - LL_y)_{CaterpillarPassage} \tag{2.4}$$

The 0.1 percentile used here represents the threshold value under which 0.001 of the observations occur. This represents a soft minimum value which is robust to outliers. This formula uses the range of upper lip and lower lip distances across the popular "caterpillar passage", one of the readings in the EMA-MAE dataset, to compute the lower bound on lip separation for a particular subject. The final result for lip separation is a measure in mm of the lip opening relative to this.

2.4.4 Lateral Lip Rounding and Jaw Characterization VT9, VT10

Lateral Lip rounding VT9 is computed from the normalized Lip corner sensor. This represents the amount of laterally lip movement, which is an indicator of rounding. VT9 is presented by the following equation:

$$VT9 = \frac{LC_z}{|mean(LC_z)_{BiteplateData}|} \tag{2.5}$$

The jaw variable VT10 is computed from the lower middle incisor sensor, which is rigidly attached to the jaw. This represents the jaw vertical movements.

Overall, 6 of the articulatory features are used for modeling tongue movements, 3 features represent lip movement and 1 feature tracks jaw movement. This set of articulatory features has been summarized in Table 2.1.

VT Feature	Description	Formula
VT1	Tongue Dorsum Horizontal Position	$\frac{TD_x}{K_x}$
VT2	Tongue Dorsum Vertical Height to HardPalate	$palate_y(TD_x, TD_z) - TD_y$
VT3	Lateral Tongue Horizontal Position	$\frac{TL_x}{K_x}$
VT4	Lateral Tongue Vertical Height to HardPalate	$palate_y(TL_x, TL_z) - TL_y$
VT5	Tongue Tip Horizontal Position	$\frac{TB_x}{K_x}$
VT6	Tongue Tip Vertical Height to HardPalate	$palate_y(TB_x, TB_z) - TB_y$
VT7	Horizontal Lip Protrusion	$\frac{UL_x - \lvert mean(UL_x)_{BitePlateData}\rvert}{K_x}$
VT8	Vertical Lip Separation	$(UL_Y - LL_Y) - 0.1percentile(UL_Y - LL_Y)_{CaterpillarPassage}$
VT9	Lateral Lip Corner (Lip Corner Sensor)	$\frac{LC_z}{mean(LC_z)_{BitePlateData}}$
VT10	Vertical Middle Incisor (Jaw)	MI_y

Table 2.1: Articulatory Feature Set: Equations and Descriptions

The articulatory feature set includes the ten static features [VT1, VT2, VT3, ..., VT10] which are described above. In addition, the velocity (delta) and acceleration

(delta-delta) of each individual static articulatory feature have been computed to represent the dynamic changes of articulatory movements. These values supplement the articulatory feature set to provide both dynamic and static representations of kinematic data.

To calculate the delta and delta-delta values, the velocity is calculated from the first-order regression, and approximate estimation of acceleration from repeated first-order regression on the velocity coefficients.

To extract these dynamic characteristics the Hidden Markov Model Toolkit (HTK) [38] has been applied for computing velocity and acceleration. It has been shown that a 3-frame window for calculating velocity and a 5-frame window for acceleration, the configuration has the best outcome compared to the other frame sizes [4].

Based on the aforementioned, the overall articulatory feature set includes static features, plus velocity and acceleration, saved into kinematic feature matrices for each subject.

2.5 Comparative Study of L1 and L2 Pronunciation and Utilizing EMA For Pronunciation Assessments

Many challenges in getting to know a new language can be traced back to the structural variations between the first and second languages. Some of these factors relate to the degree to which an L1 accent transfers to speech in L2, however, the foremost impact lies in the sound system of the first language [39]. These effects of L1 are understood to compete or intervene with the production of L2 [40] and may cause wide variations in the articulatory movements. Prior investigation suggests that language learners tend to comprise extra challenge perceiving and producing L2 contrasts that contain strange phonetic features [25]. However, whilst the variations in phonetic context make a contribution to a severe interfere with L2 production, similarities between the two languages can also result in flawed pronunciation. Cases of language learners substituting L2 sounds with similar L1 sounds have been documented [41].

Dissimilar to English, Mandarin Chinese is a tonal language. That means a minor change in tone, like stress in English, can change the concept and implication of the articulated phrase into a different unintended phrase [42].

Therefore, it might be presumed that Mandarin speakers purposely attempt to control their articulators in a manner consistent with tonal formation, and consequently, they may also have greater complexity in their articulator trajectories.

In the English language, linguists typically identify 13 distinct vowels, /ɑ/, /æ/, /ɔ/, /eɪ/, /ɛ/, /ɪ/, /i/, /ɔɪ/, /ʊ/, /u/, /oʊ/, and /ɑɪ/. However, there are only 6 common vowels in Mandarin Chinese with close English equivalents:/i/, /eɪ/, /ɑɪ/, /u/, /oʊ/, /ɑ/. Other English vowels do not have equivalents in Mandarin Chinese. The most common learning theory is that L2 speakers substitute vowels with the most acoustically similar sound in their own native language to form a new pronunciation that sounds like L1 speakers as much as possible [40].

Articulation of unstressed vowels in English is normally much less distinguished than in Mandarin, with their formants shifting nearer to the neutral schwa [43]. This means that the vowels themselves can fluctuate with stress. In addition, it is difficult to figure out where stress should be positioned based on context, and therefore the replication of English stress is a challenging task for L2 speakers.

The idea of Mandarin accented English speakers supplanting English sounds with analogous native sounds applies to the consonants as well, and even some of the consonants divided via both languages motive confusion in English due to the dissimilarity in application across languages [44]. In Mandarin, phonemes normally end with a vowel sound (with the exclusive irregularities being the front and back nasals /n/ and /ŋ/). Many Mandarin speakers switch this pattern to English through either casting off the final consonant of the English syllable or including an extraneous vowel to the syllable [44]. These adjustments will cause inconsistent articulatory patterns and different trajectories than native speakers.

One of the most considerable dissimilarities between consonant application in English and Mandarin is voicing contrasts. Mandarin replaces voiced stops with the ambition to point out stop voicing contrasts, and consequently, Mandarin accented English speakers inclined to have feeble voicing for voiced English consonants.

The last remarkable distinction in consonant application between English and Mandarin is the behavior toward consonant clusters. Consonant clusters are frequent phenomena in English in many word locations, whereas initial and final clusters do not exist in Mandarin. Mandarin speakers of English have a tendency to either eliminate the last consonant from the cluster or to create an extra syllable by way of the attachment of a shortened vowel (such as the neutral schwa) [44]. Therefore, these variations will additionally cause the inconsistency of articulatory patterns between native English and Mandarin accented English speakers due to their different pronunciation.

Understanding the variations between L1 and L2 articulatory patterns will allow us to generate greater comprehensive and effective feedback mechanisms in such CALL structures for pronunciation adjustment. In recent years, many comparative studies have been conducted using Electromagnetic Articulography corpora for pronunciation assessments.

Felps et al. [45] compared two methods for selecting units in the context of concatenative synthesis, one based on acoustic similarity and a second one based on articulatory similarity as measured utilizing electromagnetic articulography (EMA). The study showed that articulatory trajectories provide a more accurate metric for linguistic similarity across speakers than acoustic characterizations.

Suemitsu et al. [46] demonstrated that visual training improved vowel pronunciation regardless of whether audio training was also included or not.

Wieling et al. [47] compared the articulatory trajectories from native English speakers and the Dutch and German-accented speakers who are speaking in English as the second language. They investigated particularly the articulatory differences for /t/-/θ/ and/s/-/ʃ/ sounds. They showed that Dutch speakers have more difficulties to discriminate against these sounds compared to the German speakers.

The study presented in this dissertation (in chapter 4) aims to use the bilingual corpus, EMA-MAE for specifically comparing the different individual articulatory feature patterns for L1 and L2 speakers and improve the AAI systems for different articulography-speech technologies.

2.6 Acoustic Features

Typically, in most speech processing tasks Cepstrum analysis with a perceptually warped frequency axis is used to generate a set of features, called Mel Frequency Cepstral Coefficients (MFCCs) [48]. These coefficients are a set of robust representation of vocal tract configuration information regardless of the source of excitation.

The MFCC of an acoustic signal is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal. The following equation represents the MFCC feature extraction:

$$SignalPowerCepstrum = |F^{-1}\{\log|F\{x(t)\}|^2\}|^2 \tag{2.6}$$

Where $x(t)$ is the acoustic signal at the time domain.

Figure 2.3 demonstrates the different stages of extracting MFCC features.

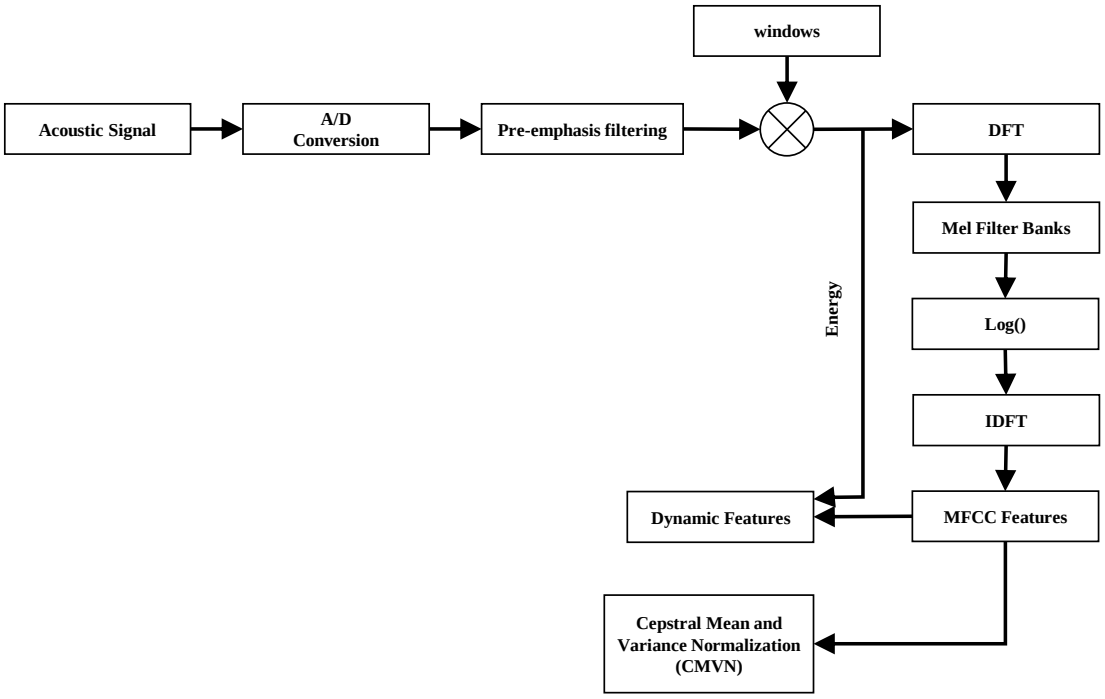


Figure 2.3: The Block Diagram of Acoustic Feature Extraction for MFCCs.

The two primary steps, A/D conversion and pre-emphasis filtering, prepare the analog acoustic signal for the feature extraction procedure. The Analog/Digital

(A/D) converter transfers the analog signal into a discrete domain by using the desired frequency sampling rate. Pre-emphasis filtering is for equalizing the energy contribution and balancing the distribution of the energy across the frequency spectrum.

Speech signal energy is not uniformly distributed across the frequency domain. For voiced speech signals like vowels, the lower frequencies have more energy compared to the higher frequencies. This is also known as spectral tilt. To improve the speech acoustic modeling and to promote phone detection, boosting higher frequency energy is essential. Therefore, using filters at the pre-emphasis stage helps us to compensate for the strength of energy at higher frequencies and to balance the spectrum of the energy signal for all frequencies [49].

In order to process the speech signal, we need to divide it into smaller pieces. Several different methods of windowing are used to segment speech signals into smaller pieces called frames. An efficient windowing method allows the amplitude of the speech signal to drop off gradually near the edges. This will reduce the noise at parts with high frequencies and makes the segments more robust against changes, especially for speech data. Hanning and Hamming windows are effective for framing the speech signals before MFCC feature extraction.

MFCC features model the human hearing system including the differential sensitivity of the human ear throughout the frequency domain. The human ear is less sensitive to the changes at a higher frequency compared to the changes at lower frequencies. Mel scaling models these differences.

The Discrete Fourier Transform (DFT) power spectrum is computed from squaring the output of DFT of the windowed speech signal. From this, the Mel-scale power spectrum is computed by using the Mel filter banks which does the Mel binning process. These scales have been derived by running a set of experiments on human subjects to be the best representative of human hearing specifications. The signal in the linear frequency scale (Hz) can be converted to the Mel scale with the following transformation:

$$Mel(X(f)) = 2595 \log(1 + X(f)/700) \tag{2.7}$$

Where $X(f)$ is the measured speech signal at the frequency domain.

The shape of the vocal tract specifies the characteristics of the different speech signals. In order to model the vocal tract shape, MFCC uses the envelope of the time power spectrum of the speech signal.

The human hearing system has logarithmic sensitivity, with logarithmic differential sensitivity to changes in amplitude intensity. To mimic this scale and provide features that are perceptually accurate, the logarithm of the Mel power spectrum is computed.

To derive an efficient acoustic model for speech recognition tasks, the formant information should be separated from the fundamental frequency. In addition, the MFCC features need to be independent and uncorrelated from each other in order to be more applicable for Machine Learning algorithms (like GMM-HMM).

To accomplish this, the Inverse-DFT (IDFT) transform separates the fundamental frequency from the formants. After this transformation, the leftmost side of the cepstral coefficients will only convey the information about formants of the speech signal and the fundamental frequency will be moved to the far right side. Figure 2.4 [49] shows the output of Mel spectrum after the IDFT.

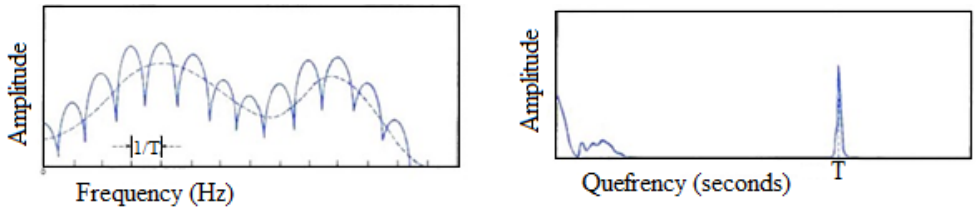


Figure 2.4: The right picture shows the output after the IDFT. The fundamental frequency (information related to the pitch) with the $\frac{1}{T}$ period is transformed to a peak near T at the right side.

The log power spectrum is real and symmetric, therefore in this case the IDFT is equivalent to Discrete Cosine Transform (DCT). The DCT transform maps the inputs to the orthogonal carriers. Consequently, after using this transformation function, the output MFCC channels will be approximately uncorrelated and independent, which is beneficial for training compact machine learning models.

The velocity and acceleration of the static MFCCs can be derived afterward to represent they dynamics of the speech signal over time. These features are known as delta and delta-delta and they are typically concatenated to the static MFCC features. For instance, if 13 MFCC channels are extracted from the acoustic signal, after computing delta and delta-delta the number of features will be 39.

In order to normalize the extracted features and compensate for the noise, statistical information from the MFCCs can be useful. The normalized features are computed by subtracting the means of the features and dividing them by the variance. For each of the utterance files X_t the mean and variance scores are calculated with the feature I across all the utterance frames X. This allows us to suppress noise and adjust scores to compensate the variants in each recording as follows:

$$\widehat{X_t}[i] = \frac{X_t[i] - \mu(X[i])}{\sigma(X[i])} \tag{2.8}$$

For small utterance files, we may choose to normalize the mean and variance for every individual speaker or across the entire training set. This stage reiterates the pre-emphasis step and effectively cancels it.

Uncorrelated acoustic features like MFCCs are very useful for classic Machine Learning algorithms like GMM. However, recently with the advance of deep learning

algorithms and strong classifiers like Convolutional Neural Networks (CNN), correlated features like Mel spectrograms can be more useful.

For speech data, the observations and samples are naturally correlated. Every incident relates to past and future events. Therefore by using DCT transform and whitening the features the valuable relationship within the feature set will be lost, reducing the accuracy of the system model. It has been shown [50] that using Mel filter bank features or Mel spectrograms improves the deep architecture performance compared to the MFCC based deep frameworks.

2.7 Classical Methods For Acoustic-to-Articulatory Inversion

Since the 1980s, there have been many investigations for speaker dependent acoustic-to-articulatory inversion. These efforts mainly focused on finding an efficient way to model the mappings between acoustic and articulatory domains. Some of these methods are based on traditional algorithms like the codebook approach [51, 52]. According to the codebook method, the articulatory trajectories are estimated by a greedy search among pairs of acoustic and articulatory features from the aligned parallel collection of the articulatory-acoustic codebook. The other classic SD-AAI approaches experimented with Kalman filtering [53], Gaussian Mixture Model (GMM) [54], and Hidden Markov Model (HMM) [55] for Acoustic-Articulatory mappings.

For example, Toda et al. [56] modeled the joint distribution of acoustic and articulatory features with a GMM. Zhang et al. [55] proposed an inversion method by implementing two parallel HMM model streams. In their approach, acoustic and articulatory HMMs are connected through a highly abstracted phoneme level representation. Huebert et al. [57] integrated voice conversion and AAI into a single GMM-based mapping framework. Their results were based on using a dataset that includes only two speakers. Richmond et.al [58, 59] used a Mixture Density Network (MDN) architecture and augmented static articulatory features with dynamic features.

Most of the aforementioned approaches do not generalize to new speakers without kinematic data. However, the acquisition and measurement of kinematic data are much more problematic than acoustic data, with higher equipment costs and significantly greater invasiveness and inconvenience to speakers. Some of the previous Machine Learning (ML) methods have been also designed to address this problem.

Hiroya [60] estimated articulatory movements by using a speaker normalized HMM-based speech production model. They assumed that the dynamical limitations of the unknown speaker are similar to the reference speakers, but their accuracy results were much lower than speaker dependent models.

Ghosh et al. [61], developed a subject-independent acoustic to articulatory inversion method considering the generalized smoothness criterion (GSC). Their method outcomes were very close to speaker dependent results on the MOCHA data set but have not been demonstrated across a more diverse range of speakers.

Ji et al. [3] introduced the PRSW method, which requires no kinematic data for the target speaker and a small amount of acoustic adaptation data. PRSW hypothe-

sizes that acoustic and kinematic similarities are correlated and uses speaker-adapted articulatory models derived from acoustically derived weights. Results demonstrate that by restricting the reference group to a subset consisting of speakers with strong individual speaker-dependent inversion performance, the PRSW method is able to attain kinematic-independent acoustic-to-articulatory inversion performance close to that of the GMM-HMM speaker-dependent AAI model.

In this work, two traditional ML methods for SD-AAI and SI-AAI, GMM-HMM and PRSW, have been considered as the baseline systems for experiments.

2.7.1 Evaluation Metrics

The primary metric for evaluation of the acoustic-to-articulatory inversion results is Root-Mean-Squared-Error (RMSE), using the known kinematic trajectory data from the EMA dataset as a reference. These values were compared across individuals and across speaker groups, in terms of average RMSE across individual utterances. RMSE is given as:

$$E_{rms} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2} \tag{2.9}$$

where y are the actual values of the articulatory data, $f(x)$ is the corresponding value of the estimated output and m is the number of test files.

Another useful evaluation metric for AAI is the correlation between actual and estimated articulatory trajectories. The correlation coefficients for AAI are computed as:

$$CC = \frac{\sum_{i=1}^{m}(f(x_i) - \overline{f(x)})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(f(x_i) - \overline{f(x)})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} \tag{2.10}$$

here y are the actual values of the articulatory data, $f(x)$ is the corresponding value of the estimated output, x_i is the articulatory input, m is the number of test files, $f(x)$ denotes the mean of the estimated trajectory, and $\bar{y}$ refers to the mean of actual articulatory values.

For SI-AAI, correlation needs to be used rather than RMSE because SI-AAI leads to an offset in terms of mean and dynamic range relative to the true unknown kinematics, even if accurately estimating the trajectory.

The final goal for an efficient AAI is having low RMSE and high correlation.

2.7.2 GMM-HMM For SD-AAI

Acoustic modeling of speech data is the process of capturing the relationship between sound units and acoustic feature vectors. Previously, HMMs were the most common approach for speech recognition tasks. HMM is a statistical state machine, which maps a discrete sequence of observation vectors onto a set of HMM states, based on models of the underlying observation probability distribution function associated

with each state as well as a transition probability structure that guides the likelihood of the underlying process moving from one state to another. Figure 2.5 illustrates a left-to-right 6-state HMM structure for acoustic modeling.

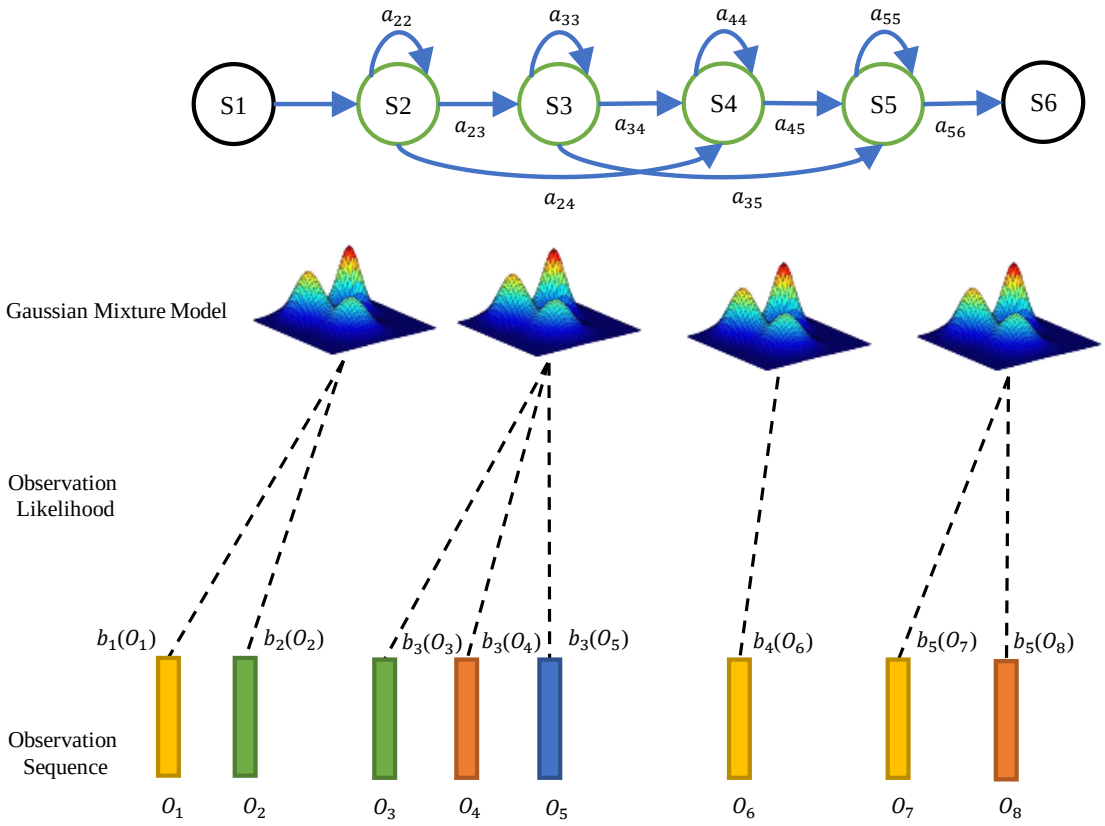


Figure 2.5: HMM Configuration for an Observation Series

In this picture the parameters needed to define the HMM are:

- Observation sequence be $[O_1, O_2, O_3, O_4, O_5, O_6]$.
- The observation likelihood probability is $B = b_i(O_t)$ each represents the probability of an observation O_t being generated from a state i.
- The transition probabilities are $A = [a_{11}a_{12}...a_{n1}...a_{nn}]$ each a_{ij} represents the probability of transition from i to j state.
- An initial probability distribution over the states, such that π_i is the probability that the HMM will start in the state i.
- At each time interval t within i the state, an observation feature sequence $[O_1, O_2, O_3, O_4, O_5, O_6]$ is generated by the probability density function $b_i(O_t)$.

The two special states of HMM (S_1 and S_6) are called non-emitting states. They allow for connecting multiple HMMs together in a longer sequence. All states generate observations except these two non-emitting states. In this model, the emission

probability is the probability of observing a possible internal state and the transition probability is the probability of transitioning from one internal state to another. The observation distribution $b_i(O_t)$ is typically represented by Gaussian mixture models (GMMs).

In this work, the traditional-ML, GMM-HMM framework is selected as a baseline modeling method for SD-AAI system. This model consists of parallel acoustic and articulatory HMMs, with dynamic smoothing to account for the presence of discrete rather than continuous state variables. Figure 2.6 represents the block diagram of GMM-HMM for SD-AAI.

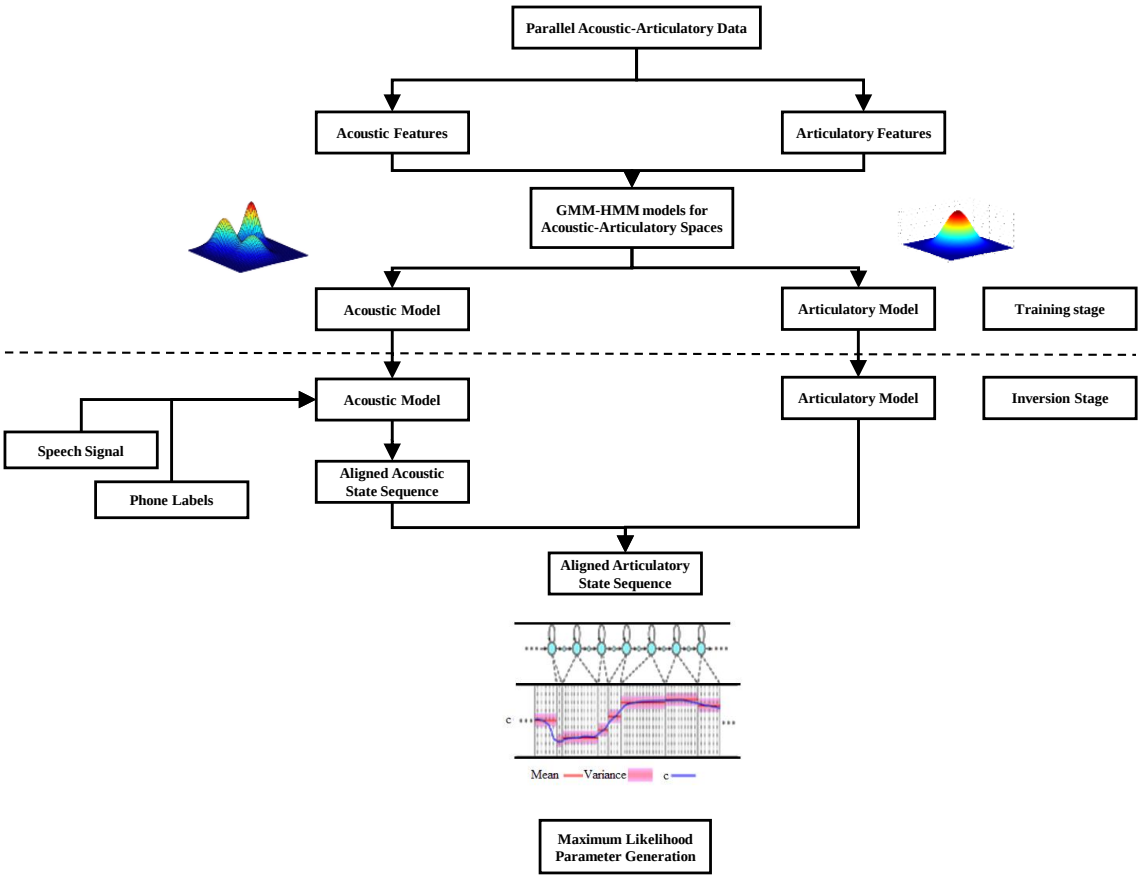


Figure 2.6: HMM Configuration for an Observation Series

In the training phase, parallel acoustic-articulatory data is trained separately for each individual speaker. In the inversion stage, the test speech is input to the trained acoustic HMMs to derive an optimal HMM state alignment, and then the corresponding aligned articulatory HMMs were used to recover the articulatory trajectory. Once

the alignment of articulatory states is completed, the recovery algorithm estimates a smooth articulatory trajectory from the HMM.

2.7.3 PRSW For SI-AAI

Mapping from acoustic speech space to the articulatory domain is not unique and varies across different speakers. Each subject has an exclusive physiological vocal tract anatomy and consequently a different speech production mechanism. Furthermore, kinematic sensors are not at the exact same locations for each subject. Therefore, a robust speaker independent articulatory synthesis model is needed to estimate an unknown speaker's articulatory information from the input acoustic data.

The goal is to develop a framework to take the reference acoustic-articulatory mappings and generate a new mapping that will correctly estimate the kinematic trajectories for a new speakers speech information. The PRSW method [4] uses the acoustic adaption techniques to identify different acoustic patterns and generates the parallel adapted acoustic and kinematic models.

This parallel model can be used for a new subject to identify its articulatory trajectory based on the parallel acoustic information. A group of speakers has been chosen as a reference speaker set and their weighted articulatory models have been used to recover the new speaker's kinematic trajectory.

If R is the set of reference speaker articulatory super vectors:

$$R = \{r_1, r_2, r_3, ..., r_K\} \tag{2.11}$$

Then the RSW estimates the articulatory super vector for the new speaker based on the following equation:

$$R_{unknown} \approx \sum_{k=1}^{K} w_k r_k = WR \tag{2.12}$$

In this equation W is the weight vector $W = [w_1, w_2, w_3, ..., w_K]$. Given the adaptation data $O = \{O_t, t = 1, .., T\}$ the Maximum Likelihood estimate w can be found by maximizing the 2.13 function.

$$Q(w) = -\sum_{g=1}^{R}\sum_{t=1}^{T} \gamma_t(g)(O_t - R_{unknown}(w))' C_g^{-1}(O_t - R_{unknown}(w) \tag{2.13}$$

Where $\gamma_t(g)$ is the posterior probability of observing O_t in the g^{th} Gaussian, and C_g is the covariance matrix of the g^{th} Gaussian. The optimal weight vector may be found by setting the statement 2.13 to zero.

$$\frac{\partial Q}{\partial w} = 2\sum_{g=1}^{R}\sum_{t=1}^{T} \gamma_t(g) R_g' C_g^{-1}(O_t - R_g(w)) = 0 \tag{2.14}$$

Thus, the weights w may be obtained by solving a system of K linear equations:

$$w = [\sum_{g=1}^{R}(\sum_{t=1}^{T}\gamma_t(g))R'_g C_g^{-1} R_g]^{-1}[\sum_{g=1}^{R} R'_g C_g^{-1}(\sum_{t=1}^{T}\gamma_t(g)O_t)] \tag{2.15}$$

Therefore, RSW uses the model parameters of the selected speakers to create a composite model. Figure 2.7 illustrates the PRSW algorithm.

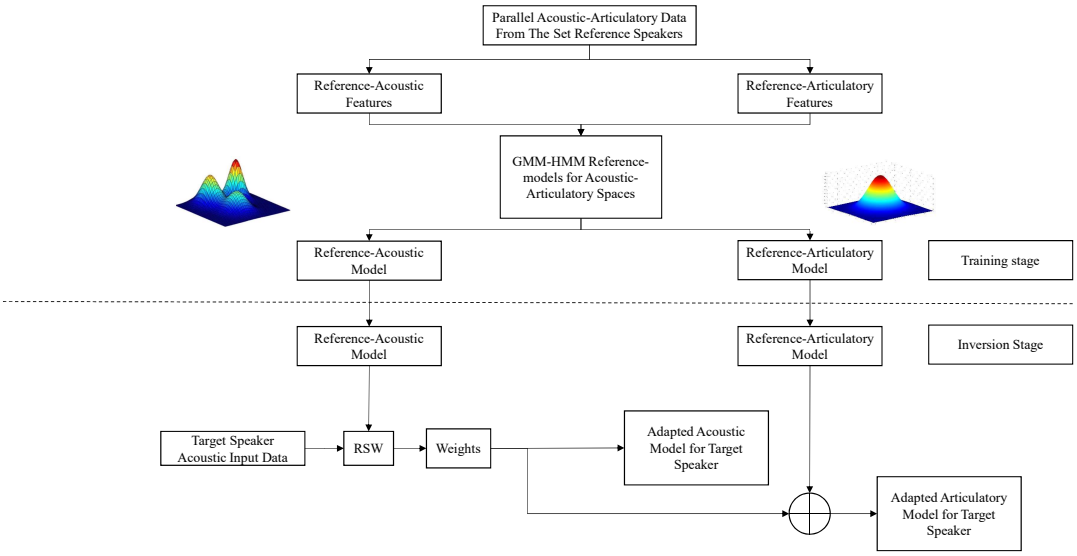


Figure 2.7: Parallel Reference Speaker Weighting Block Diagram

2.8 Deep Architectures for AAI

In recent years, there have been several notable works for using Deep Learning architectures for speech processing applications including speech recognition and speech synthesis. Due to its powerful predictive capability, studies have focused on using deep neural networks (DNN) for various audio processing domains. In this regard, recently the traditional models for acoustic-articulatory mapping have been also upgraded by various deep learning architectures. This section reviews some of these approaches.

Tobing et al. [62] presented the latent trajectory model in deep acoustic-to-articulatory inversion mapping systems. They used a latent trajectory model which

allows frame interdependency to be considered in training the model by utilizing a soft constraint between static and dynamic features in the latent space.

Mitra et al. [63] investigated the application of DNNs and convolutional neural networks (CNNs) for mapping speech data into its corresponding articulatory space. To effectively model the temporal articulatory features, they explored a joint modeling strategy to simultaneously learn both the acoustic and articulatory spaces. The results from multiple speech recognition tasks indicated that articulatory features can improve recognition performance when the acoustic and articulatory spaces are jointly learned with one common objective function.

Uria et al. [64] proposed a deep version of the trajectory mixture density network (TMDN) to invert acoustic input into articulatory motions. They used pre-defined fixed length frames to cover the most important speech context information as an input acoustic model. Their results showed the RMSE of 0.88mm for the MNGU0 test dataset by predicting kinematic trajectories from speech data.

Wei et al. [65] investigated the feasibility of using DNNs for articulatory movement prediction from the text. They also combined full-context features, state and phone information with stacked bottleneck features which provide wide linguistic context as network input, to improve the performance of articulatory movement prediction. They evaluated the DNN framework on the MNGU0 dataset and that resulted in an RMSE of 0.73 mm. In addition, they also applied stacked bottleneck features for acoustic space modeling and they showed that these sets of features can effectively capture the important contextual information from data.

There are several other attempts for applying deep learning architectures for acoustic-articulatory transformations. Sivarman et al.[66] applied Artificial Neural Networks (ANN), Cail et al.[67] and Seneviratne et al. [68] deployed a Deep Neural Network (DNN) architecture. Illa and Ghosh [69, 70, 71, 72] have proposed two different DNN approaches [69, 71], Bidirectional Long-Short Term Memory (BLSTM) [70] and Convolutional Neural Network (CNN) layer cascaded to the BLSTM network [72] for speaker dependent AAI. Mannem et al. [73] used a convolutional dense neural network. To capture dependencies between articulatory trajectories and corresponded past, current and future acoustic features Liu et al. [74] implemented BLSTM and deep recurrent MDN. Xie et al. [75] investigated different architectures such as DNN, Recurrent Neural Network (RNN), MDN, Time Delay DNN-MDN, RNN-MDN and RNN-MDN BLSTM, Biasutto et al. [76] applied bidirectional gated RNN and Maud et al. [77] make use of the BLSTM neural network with an additional convolutional layer, which acts as a low pass filter after the readout layer for AAI.

However, these approaches require a substantial amount of speaker-specific kinematic data to create inversion models and cannot be generalized to new speakers without kinematic data.

So far, there have been a few attempts for addressing the problem of estimating articulatory trajectories for unseen speakers. Ghosh et al. [61] applied generalized smoothness criterion, Huber et al. [13] experimented with different adaptation scenarios like Maximum Likelihood Linear Regression (MLLR), direct cross-speaker AAI Gaussian Mixture and Cascade-GMM and Siverman, et al.[8] Explored vocal tract

length normalization for SI-AAI.

2.9 WaveNet: Autoregressive model with Conditional Convolutional Neural Networks

In this dissertation, a new deep autoregressive AAI model Articulatory-WaveNet (AWN) has been introduced which uses a waveform-based speech synthesizer for the task of acoustic-to-articulatory inversion. This section reviews the original WaveNet architecture model for speech synthesis. In addition, details about architectural specifications, WaveNet applications and modified versions have been provided.

Googles WaveNet architecture [14] is a novel approach to the problem of speech waveform synthesis that has significantly improved intelligibility for text-to-speech applications. This architecture has been designed based on the point-to-point prediction of the raw audio signals. WaveNet is a fully probabilistic and autoregressive model that generates the time-series signal by using the causal conditional predictive distribution of samples. This architecture utilizes the stacked convolutional layers to model the conditional probability distribution.

The product of the sequential conditional probabilities over time is represented as a model of the joint probability of a time-series signal. The occurrence of each sample x_t from the time-series signal x can be conditioned incident on all previous samples $(x_1, \ldots, x_{t-1})$ [14, 23]:

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}) \tag{2.16}$$

In other words, by using the probabilistic chain-rule and product of the conditional distributions, the autoregressive WaveNet network models the joint distribution of high- dimensional data like acoustic speech signal for speech synthesis.

2.9.1 Dilated Causal Convolutions

WaveNet is a fully probabilistic and deep autoregressive architecture that tracks the samples from the time-series signal by applying the causal conditional predictive distribution of samples. According to the WaveNet the joint distribution probability of sequential samples $p(x_t \mid x_1, ..., x_{t-1})$ at time-series signals are modeled by using the probabilistic chain-rule and the product of the sequential conditional probability distributions. Using this method, the conditional probability distribution is modeled by stacked convolutional layers.

WaveNet utilizes masked or causal convolutional layers to eliminate the dependency of the future unseen samples and to predict the sample x_t based on the information from previous samples $x_1, ..., x_{t-1}$. This enables the system to generate all $p(x_t \mid x_1, ..., x_{t-1})$ in one forward pass.

The larger receptive field of the causal convolutional architecture requires more layers or a larger filter. However, dilated convolutional layers provide a vast receptive field by dilating the original filter with zeros. If the input values are skipped or masked, then the convolutional filter will operate more efficiently on a larger area

than its length. That would be the same as the pooling or shifted convolutions, except that the size of output would remain the same as the size of the input. This type of architecture would not only enlarge the receptive field but also keeps the computational costs and input resolution at the same value. The simple CNN model is a type of dilated convolutional architecture with dilation set to the 1 [14, 23].

Figure 2.8 represents the WaveNet, a fully convolutional neural network, where the convolutional layers have various dilations. By using this property, the deep architecture encompasses a massive amount of time steps and therefore it grows the receptive field exponentially.

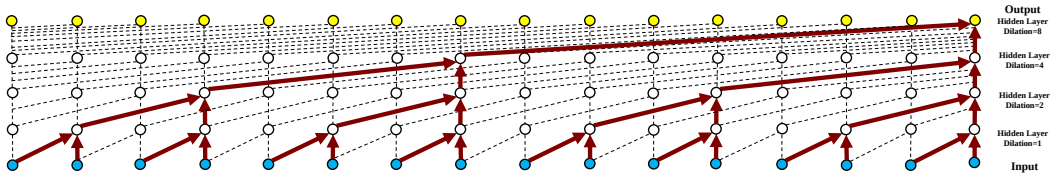


Figure 2.8: Visualization of WaveNet stacked causal convolutional layers

For this architecture model, the gates are the nonlinear activation units for modeling the time-series signal. Equation 2.17 describes the computation at these gates.

$$z = \tanh(w_{f,k} * X) \odot \sigma(w_{g,k} * X) \tag{2.17}$$

Where represents convolutional operator, $\odot$ is an element-wise multiplication operator, $\sigma(.)$ denotes a logistic sigmoid function, k is the layer index, f, and g are filter and gate, respectively, and W is convolutional filter weight matrix.

2.9.2 Conditioning

WaveNet has the capability to model a sequence of time-series samples which have been conditioned on a sequence of additional time-series inputs. By conditioning, the network, the generated outputs, and the predicted time samples will be based on the required specifications in the conditioning series [14]. The conditional probability distribution is represented as follows:

$$p(x|h_t) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}, h_t) \tag{2.18}$$

where time-invariant conditioning sequence has been represented by h_t. Therefore, considering the additional conditional input, the activation unit in Equation 2.17 becomes:

$$z = \tanh(w_{f,k} * x + V_{f,k} * h(t)) \odot \sigma(w_{g,k} * x + V_{g,k} * h(t)) \tag{2.19}$$

Where $*$ represents the convolutional operator, $\odot$ is an element-wise multiplication operator, $\sigma(.)$ denotes a logistic sigmoid function, k is the layer index, f and g are filter and gate indices, respectively, and w, V are the convolutional filter weight matrices for articulatory and acoustic features respectively. The time-invariant conditioning sequence has been represented by h_t.

2.9.3 Different Versions of WaveNet

WaveNet has been used as an autoregressive deep generative model for different type of corpuses other than speech like text [78], image [79, 80, 81, 82], video [83], handwriting [84], and music [85, 86].

Recently many publications like Nv-WaveNet [87, 88] waveRNN [89], Parallel WaveNet [23] WaveGlow [15], Clarinet [90], MCNN [91], FFTNert [92], LPCNet [93] have tried to improve WaveNet and use it for different applications. For instance, some of these studies tackle the problem of the slow synthesizing problem at the WaveNet sample generation stage by different parallelization and recurrent computation techniques.

The new proposed AAI model, Articulatory-WaveNet, like many other versions of WaveNet, uses acoustic features for conditioning instead of the fundamental frequency and linguistic features at original WaveNet. Some of the approaches that use acoustic features or bottleneck extracted abstractions from another deep architecture for the WaveNet conditioning have been mentioned below:

Kastner, et al.[22] used a WaveNet model conditioned on log Mel spectrograms for the TTS task. They used attention-based RNN to derive acoustic features from linguistic inputs. The combined linguistic characteristics are built from a mixture of the characters, phonemes, or mixed representations.

Tacotron and Tacotron 2 [17, 18] use attention-based recurrent sequence-to-sequence feature estimation architecture to predict the sequence of Mel-scale spectrograms features from character embedding inputs, with a modified WaveNet conditioned on Mel spectrograms to synthesize the speech waveforms. Moreover, they generated the speech signal at the frame level which helps the autoregressive model to generate faster compared to the sample-level architecture.

Prenger et al. [15] combined a flow-based approach, GLOW [94] architecture, with WaveNet to eliminate autoregression or long-term dependency on previous samples. They used a single loss function to maximize the likelihood of the training data. This model has been also conditioned on acoustic Mel spectrogram samples.

Maiti et al. [21] investigated WaveNet and WaveGlow architectures for parametric speech resynthesize which is conditioned on acoustic representations, log Mel-spectrogram for generating the noise suppressed clean speech. In addition, they compared those architectures objectively and subjectively with Chimera++ [95] and oracle Wiener mask methods. They showed that WaveNet outperforms the aforementioned methods, however, it is considerably slower for the speech generation step.

Tanaka et al. [16] proposed WaveGlowGAN2 which replaces dilated convolutions instead of the simple convolutions used in WaveGlowGAN, using linear projection and residual blocks. The model has been designed with no resampling modules to

eliminate aliasing. They also combined multi discriminators from the waveform and acoustic parameter space to avoid the vanishing gradient of the synthesizer.

Tamamori et al. [96] proposed a new method for speaker-dependent natural speech generation using WaveNet. They utilized the acoustic features from the existing vocoder as auxiliary features for WaveNet. It has been assumed that the network learns the correlation between speech waveforms and extracted acoustic features automatically during the speech synthesis process due to the physical limitations that have been imposed on the generation of the waveforms.

Spratley et al. [85] proposed a combined Generative Adversarial Network with WaveNet for Music Instrument Retrieval (MIR) and different instrumental audio applications. They conditioned WaveNet on the outputs of the Generative Adversarial Networks spectrogram translator to generate the final audio waveform.

Recently, many studies have been conducted to improve the WaveNet architecture and make the generation and synthesis of samples faster. For example, the Probability Density Distillation (PDD) method [97] combines two strategies of Inverse Autoregressive Flows (IAF) and WaveNet to make the system compatible with real-time processing and parallel computing [23]. The PDD utilizes a teacher trained WaveNet to train the parallel feed-forward IAF (student) network. This system is much faster than the vanilla WaveNet, therefore it can be used for variant languages and multiple speakers.

To eliminate unnecessary convolutional operations, Paine et al. [19] proposed a new method named Fast-WaveNet. This framework caches previous computations instead of recomputing them from scratch to predict the new samples. Compared to the naive WaveNet, Fast-WaveNet reduces the complexity of the operation from $O(2L)$ to $O(L)$, which L represents the number of layers in the neural network. In the work presented here, we have used this Fast-WaveNet approach with our Articulatory-WaveNet framework to generate articulatory trajectories faster.

2.10 Summary

This chapter has reviewed the technical background of pronunciation learning, articulatory data acquisition, acoustic to articulatory inversion, and WaveNet synthesis systems. In this dissertation, a novel acoustic-to-articulatory inversion approach, Articulatory-WaveNet, is introduced based on the WaveNet speech synthesis architecture. The proposed system uses the dilated causal convolutional layers with previous values of the predicted articulatory trajectories conditioned on acoustic features. The remainder of this dissertation will focus on methodology for baseline and proposed systems for SD and SI AAI, as well as a detailed examination of AAI for L1 and L2 speakers.

Chapter 3 Investigating Classic-ML Algorithms For AAI

This chapter describes methods of Classic-Machine Learning (ML) algorithms for Acoustic-to-Articulatory Inversion (AAI) and presents new experimental work showing the results of these methods on the EMA-MAE dataset, introducing a new method for selecting reference speakers, and comparing results between English L1 and Mandarin L2 speakers.

For Speaker Dependent-AAI (SD-AAI) different experiments are implemented to compare various classical Machine Learning methods on EMA-MAE. For Speaker Independent-AAI (SI-AAI), the PRSW method is evaluated in terms of the impact of different reference speaker sets, and a novel method Maximum Likelihood Linear Regression (MLLR)-PRSW approach is proposed to estimate articulatory trajectories for unseen target speakers. These systems are evaluated using the EMA-MAE corpus and the results of RMSE and correlation are reported for each of them.

3.1 SD-AAI framework

In this work, to evaluate the effectiveness and performance of the existing AAI classical Machine Learning (ML) methods, several experiments are implemented. The acoustic-articulatory data from EMA-MAE corpus are used in these sets of experiments to model the acoustic and articulatory spaces for Speaker Dependent-AAI using the classical-ML methods like Gaussian Mixture Model, Hidden Markov Model, Universal Background Model, and Maximum Likelihood Linear Regression.

For the first SD-AAI approach, a parallel GMM-HMM articulatory model is tied to the acoustic observation sequence, with dynamic smoothing to account for the presence of discrete rather than continuous state variables. The block diagram of the acoustic-articulatory model using GMM-HMM has been previously illustrated in Figure 2.6 . In this model, two synchronized acoustic-articulatory are trained separately for every speaker. Following this, in the inversion stage, the articulatory trajectories are predicted from input acoustic test data by using the aligned optimal HMM state from the trained acoustic GMM-HMM.

For the other classical-ML SD-AAI approaches, the Universal Background Model is adapted for modeling the acoustic/articulatory spaces. In this method, a Universal Background Model (UBM) is formed by using the information from all speakers other than the target speaker. Then by using the adaptation method, Maximum Likelihood Linear Regression (MLLR), the model for the target speaker is adapted individually.

During the MLLR process, statistical information is acquired from available adaptation acoustic/articulatory data from the target speaker and is applied to find a linear regression-based transformation for the mean scores. The MLLR performs the adaptation process for the target speaker distribution model which does not exist in training models (based on reference speakers) through tying the mappings among

several distributions that already exist in the training model (reference speakers).

Mathematically speaking, suppose the acoustic/articulatory records from K speakers have been collected and an HMM with K_g Gaussians has been acquired accordingly. If all of the K_g Gaussians in the speaker dependent model classified into L regression classes then the transformation function that maps the g^{th} Gaussian to its regression class would be defined by:

$$h = H(g) \tag{3.1}$$

Assuming $h = 1, \ldots, L$, $g = 1, \ldots, K_g$, and the g^{th} Gaussian mean of the k^{th} speaker is computed by:

$$\mu_g^{(k)} = Y_{H(g)}^{(K)'} \xi_g \tag{3.2}$$

Where $Y_{H(g)}^{(K)'}$ is the MLLR transformation for the $H(g)^{th}$ regression class of the K^{th} speaker and the extended mean of the corresponding Gaussian is represented by ξ_g.

By assigning $Y_{H(g)} = W$, the equation 3.2 will be reformatted to:

$$\mu_g = W \xi_g \tag{3.3}$$

Where W is $W \in \Re^{K \times (K+1)}$.

Given the adaptation data $O = \{O_t, t = 1, \ldots, T\}$ from the target speaker, the Eigen matrix of weights W can be found by maximizing the likelihood of O or equivalently $Q(W)$ function:

$$Q(w) = -\sum_{g=1}^{K_g} \sum_{t=1}^{T} \gamma_t(g)(O_t - \mu_g(W))' C_g^{-1} (O_t - \mu_g(w)) \tag{3.4}$$

Where $\gamma_t(g)$ is the posterior probability of observing O_t in the g^{th} Gaussian, and C_g is the covariance matrix of the g^{th} Gaussian. Consequently, the optimal weight may be found by finding the solution of the following equation:

$$\frac{\partial Q}{\partial w} = 2 \sum_{g=1}^{K_g} \sum_{t=1}^{T} \gamma_t(g)(O_t - \mu_g(W))' C_g^{-1} \frac{\partial \mu_g(W)}{\partial W} \tag{3.5}$$

By setting the derivative to zero, the optimal weights are obtained by solving a system of K linear equations [98]. UBM uses this method to adapt the individual model for every target speaker.

The UBM approach is usually useful for speaker verification tasks, where the target speaker can be verified by GMM-UBM from all other speakers [99]. We applied this strategy for SD-AAI to adapt models for Acoustic and Articulatory data of every speaker using the UBM that has been trained by GMM-HMMs of all other speakers. Figure 3.1 shows the block diagram of SD-AAI which uses GMM-HMM UBM for modeling the acoustic data.

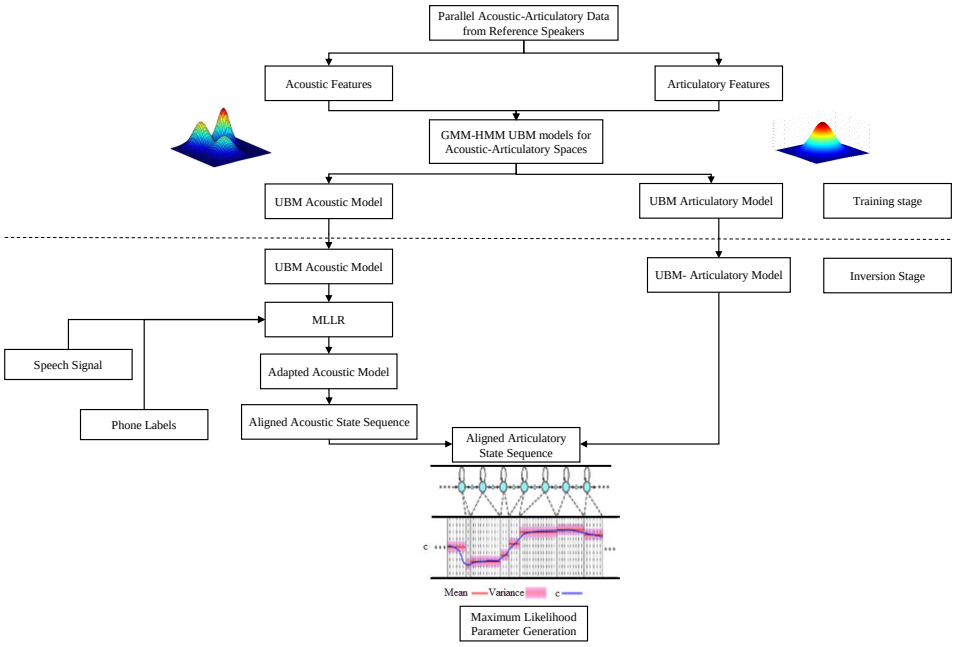


Figure 3.1: Speaker Dependent Acoustic-to-Articulatory Inversion with Gaussian Mixture Model, Hidden Markov Model, and Universal Background Model Structure

Therefore, for the second classical ML SD-AAI approach we used GMM-HMM UBM for modeling acoustic space and vanilla GMM-HMM for articulatory space. The third approach uses the reverse of this order for modeling the acoustic-articulatory spaces and the fourth approach, the GMM-HMM UBM is applied for modeling both acoustic and articulatory spaces.

The performance of these four methods discussed above has been evaluated for 39 speakers from the EMA-MAE corpus. The results are reported in Table 3.1.

VT	Metric	FrameWork			
		GMM-HMM (Acoustic) **GMM-HMM** (Articulatory) ***Method1***	**GMM-HMM** (Acoustic) **GMM-HMM UBM** (Articulatory) ***Method2***	**GMM-HMM UBM** (Acoustic) **GMM-HMM** (Articulatory) ***Method3***	**GMM-HMM UBM** (Acoustic) **GMM-HMM UBM** (Articulatory) ***Method4***
VT1	RMSE	3.41mm	2.74mm	3.27mm	3.75mm
	CC	0.64	0.66	0.69	0.65
VT2	RMSE	3.44mm	3.30mm	3.38mm	3.76mm
	CC	0.64	0.66	0.69	0.65
VT3	RMSE	2.62mm	1.89mm	2.65mm	2.89mm
	CC	0.61	0.58	0.67	0.60
VT4	RMSE	2.35mm	1.96mm	2.29mm	2.43mm
	CC	0.66	0.68	0.70	0.65
VT5	RMSE	3.28mm	2.45mm	3.15mm	3.42mm
	CC	0.62	0.60	0.68	0.62
VT6	RMSE	3.37mm	3.46mm	3.34mm	3.48mm
	CC	0.65	0.72	0.69	0.67
VT7	RMSE	3.37mm	1.01mm	0.99mm	1.03mm
	CC	0.55	0.65	0.59	0.57
VT8	RMSE	3.05mm	3.78mm	3.03mm	3.16mm
	CC	0.61	0.75	0.67	0.66
VT9	RMSE	0.83mm	0.86mm	0.85mm	0.86mm
	CC	0.50	0.63	0.55	0.54
VT10	RMSE	1.99mm	2.47mm	1.98mm	2.13mm
	CC	0.67	0.71	0.72	0.68
Mean	RMSE	2.83mm	2.39mm	2.50mm	2.70mm
	CC	0.61	0.66	0.67	0.62

Table 3.1: Performance Comparison of Classic-ML Methods Using GMM-HMM and UBM

The average RMSE for tracking the vocal tract height at the three tongue sensors, key variables for capturing physiological characteristics of tongue motion, are 3.05mm, 2.90mm, 3.00mm and 3.22mm, for methods 1, 2, 3 and 4 respectively. Speaker horizontal tongue sensor positions have an average RMSE of 3.10mm, 2.27mm, 3.02mm and 3.35mm for methods 1, 2, 3 and 4 respectively. Therefore, for RMSE the best-reported result is for method 2 in which articulatory space is modeled by GMM-HMM UBM and acoustic space is modeled by straight GMM-HMM.

Lip features (Vertical lip separation, Horizontal lip protrusion and Lateral lip Distance) have average RMSE of 2.41mm, 1.88mm, 1.62mm, 1.68mm for methods 1, 2, 3 and 4 respectively, and the middle incisor (jaw) sensor shows RMSE 1.99mm, 2.47mm, 1.98mm and 2.13mm for methods 1, 2, 3 and 4 respectively. Therefore the

best model for tracking the lip features and jaw movements is approach 3, which models acoustic space with GMM-HMM UBM and articulatory space with GMM-HMM.

Correlation result shows consistent scores across all approaches, with all of the articulatory feature trajectories having correlations around 60%, ranging from 61% to 67%. In a strict sense, the correlation results from approach 3 are slightly better than the others by average.

Generally, the results indicate that the optimum approaches for classic-ML SD-AAI are approaches 2 and 3. These systems have the highest average of correlation, 67%, and the lowest average RMSE, 2.39mm outcome from approaches 3 and 2 respectively. The worst system at Table 3.1. is approaching 1, which uses vanilla GMM-HMM for modeling both acoustic-articulatory spaces, with correlation 0.61 and RMSE 2.83mm. Consequently, using adapted GMM-HMM models from UBM for acoustic and articulatory spaces is always beneficial for SD-AAI system.

3.2 SI-AAI Framework

3.2.1 Evaluating SI-AAI Performance Using Different Reference Sets For PRSW

This work [100] investigates the most effective reference speaker set for the Parallel Reference Speaker Weighting (PRSW) algorithm for speaker-independent acoustic-to-articulatory inversion (SI-AAI). Previously PRSW [4] has been presented as a strategy for SI-AAI (SI-AAI). PRSW assumes that acoustic and kinematic similarities are correlated and uses speaker-adapted articulatory models derived from acoustically derived weights.

For the PRSW approach, the quality of an adapted acoustic model relies on a good selection of reference speakers. Speakers with the highest correlated results or lowest RMSE scores from speaker dependent acoustic-to-articulatory inversion experiments based on the GMM-HMM framework can be good candidates for PRSW reference speaker sets.

In addition, to compare speaker sets based on the best SD-AAI performance, we also evaluated the reference speaker set based on language background. In other words, we hypothesize that the speaker's accent may also have an influence on the SI-AAI results. For instance, we expected Mandarin reference speakers to be more suitable set for Mandarin target speakers and Native English reference speakers for Native English target speakers. Based on these assumptions we designed different reference sets with respect to L1 and L2 accents, numbers and high performances in speaker dependent inversion.

For this set of experiments parallel acoustic and articulatory information from the EMA-MAE corpus are used. In each experiment acoustic-articulatory information from the different number of speakers has been used to train the reference model and other speakers are selected as target speakers set to evaluate the PRSW model for SI-AAI task. For the first and second sets of experiments, the reference sets are selected based on the best performance for RMSE and correlation from SD-AAI

system respectively. For the third and fourth sets of experiments, we considered the balanced distribution of the L1 and L2 spears at reference set despite SD-AAI good performance. For the fifth and sixth sets of experiments reference sets are built from only well- performed SD-AAI native speakers, and the seventh and eight sets of experiments are considered only L2 well-performed speakers at reference sets. For every set of experiments, the number of reference speakers increases by a factor of 2 in every experiment which starts with 4 and reaches a maximum number of 18 (4, 6, 8, 10, 12, 14, 16 and 18 speakers at each set).

Table 3.2. reports the performance of the best number of the reference speaker set, for each different experiment set. The target test speaker sets are selected from different groups of L1 and L2 speakers to compare the performance of the PRSW system for different accents.

Number	**Reference Set**	**Mandarin Accented (L2) Target Speaker**		**Native English (L1) Target Speaker**	
		The Best Number of Reference Speakers	The Correlation Score	The Best Number of Reference Speakers	The Correlation Score
1	Selected from both L1 and L2 speakers based on the best RMSE performance	04	0.62	12	0.61
2	Selected from both L1 and L2 speakers based on the best Correlation performance	04	0.64	06	0.62
3	Selected from Equally Distributed L1 and L2 speakers based on the best RMSE performance	06	0.62	04	0.62
4	Selected from Equally Distributed L1 and L2 speakers based on the best Correlation performance	08	0.64	08	0.63
5	Selected from only L1 speakers based on the best RMSE performance	08	0.63	06	0.61
6	Selected from only L1 speakers based on the best Correlation performance	04	0.64	06	0.61
7	Selected from only L2 speakers based on the best RMSE performance	04	0.60	04	0.62
8	Selected from only L2 speakers based on the best Correlation performance	04	0.61	04	0.61

Table 3.2: Best Results for Different Reference Sets With Respect to L1 and L2 Accents, Numbers and High Performances in Speaker Dependent Inversion

Overall results indicate that the adaptation models generated using the well-performed SD-AAI speakers as the reference set for PRSW archive correlation above 60% for all different sets of SI-AAI experiments.

In addition, the results show that a smaller number of reference speakers have a positive impact on SI-AAI performance. In other words, the amount of adaptation data from various reference speakers influences the quality of SI-AAI using PRSW and reported results show that PRSW is able to predict articulatory trajectories for unseen speakers with a relatively small quantity of reference speakers.

By considering the performance of PRSW for different reference sets we can also infer that selecting the reference sets based on their correlation performance (best speakers at SD-AAI) is better than selecting them based on RMSE.

Not to mention that reference sets with both L1 and L2 speaker's accents have higher performance than reference sets with single accent L1 or L2 and finally equally distributed L1 and L2 reference speakers have the highest correlation results.

Figure 3.2 shows graphically the comparisons between different reference sets and target speakers (Mandarins and Native English speakers) based on their correlation results.

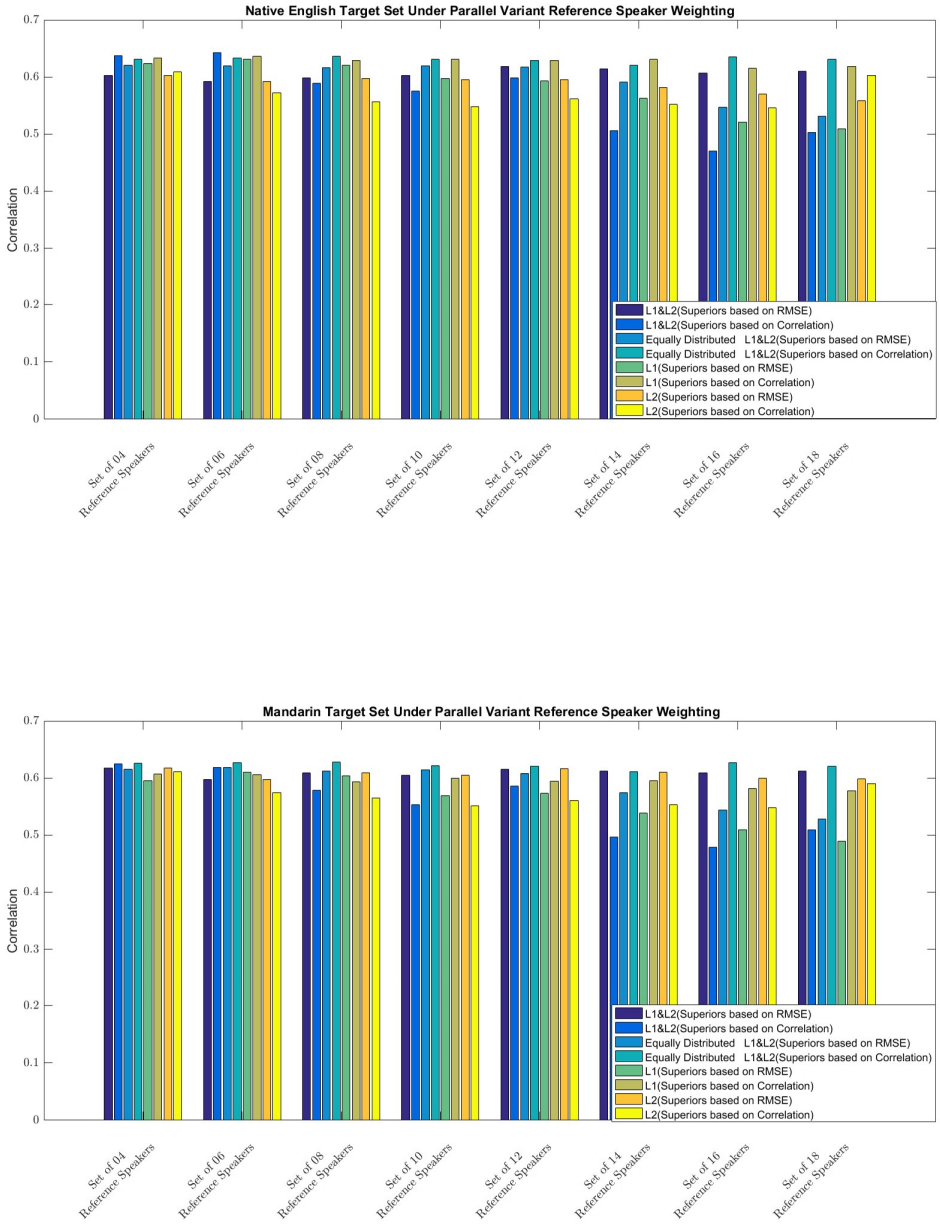


Figure 3.2: The Comparisons Between Different Reference Sets and Target Speakers

3.2.2 MLLR-PRSW For SI-AAI

In the previous approach, two parallel streams of Hidden Markov Models (HMMs) for aligned acoustic-articulatory data were used. For each individual reference speaker,

a parallel GMM-HMM including both acoustic and articulatory models was been estimated. In the PRSW approach, these models are separately and individually trained for each reference speaker. For a new target speaker, and RSW-type approach is used to estimate weightings for the reference speakers based on target speaker acoustic adaptation data, and these same weights are used to create an adapted articulatory model as well.

In this work, we propose the MLLR-PRSW approach [101] for SI-AAI. For this new system, a Uniform Background Model (UBM) [99] approach is used both for estimating the speaker dependent acoustic models for reference speakers and for estimating the target speaker acoustic model. According to this method, the maximum likelihood of the adaptation data is computed using acoustic data from a new target speaker, and MLLR is used to update mean values in the UBM model. RSW adaptation is implemented in parallel, and the PRSW approach is used to estimate the weights and the articulatory model from those.

Figure 3.3 shows the block diagram of the applied MLLR-PRSW structure that has been used in this approach.

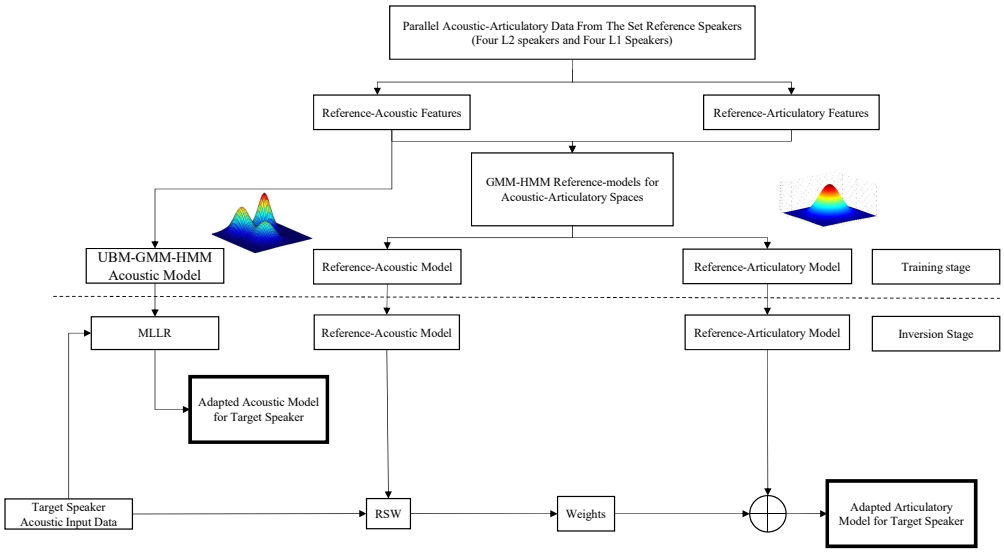


Figure 3.3: Maximum Likelihood Linear Regression Block Diagram

For this study aligned acoustic-articulatory records from EMA-MAE dataset are used for evaluating MLLR-PRSW. The articulatory sensors include all the sensors

from 2.3 REF, MI, LL, UL, TD, and TA, all placed in the mid-sagittal plane. In addition, there are two lateral sensors: LL and LT. Rather than using direct sensor coordinates as articulatory features, a normalized and palate-referenced set of articulatory features are used for modeling. Table 2.1 lists the selected articulatory features for this investigation. In addition to these base features, the velocity (delta) and acceleration (delta-delta) of each individual static articulatory feature have been included to encompass the dynamic characteristics.

For the acoustic data, Cepstrum analysis with a perceptually warped frequency axis is used to generate a set of Mel Frequency Cepstral Coefficient (MFCC) features. For this investigation, a set of 12 static MFCCs plus dynamic MFCCs (velocity and acceleration of static MFCCs) has been chosen to represent the acoustic features.

Table 3.3 shows the averaged correlation results of each individual articulatory feature for the two groups of target speakers including Mandarin accented English (L2) speakers and native English (L1) speakers. Since we used the 8 speakers as a reference set for training the models, the target speaker set for Mandarin speakers consists of 15 subjects and for native English speakers includes 16 subjects.

Average of Correlation Scores Across All the Native English Target Speakers											
Method	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8	VT9	VT10	Average
SD-AAI	0.63	0.69	0.64	0.70	0.65	0.72	0.63	0.69	0.58	0.74	0.67
PRSW-PRSW	0.60	0.63	0.60	0.68	0.58	0.70	0.59	0.70	0.57	0.70	0.63
MLLR-PRSW	0.68	0.69	0.62	0.71	0.60	0.73	0.62	0.73	0.56	0.71	0.66
Average of Correlation Scores Across All the Mandarin Accented English Target Speakers											
Method	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8	VT9	VT10	Average
SD-AAI	0.66	0.69	0.67	0.70	0.69	0.70	0.57	0.66	0.55	0.72	0.66
PRSW-PRSW	0.62	0.63	0.62	0.63	0.63	0.67	0.55	0.67	0.51	0.70	0.62
MLLR-PRSW	0.66	0.69	0.66	0.68	0.67	0.71	0.58	0.71	0.50	0.74	0.66

Table 3.3: The Averaged Correlation for Each Articulatory Feature

Based on these results, it can be seen that SD-AAI outperforms other methods in terms of average performance across all articulatory features. However, for some of these individual features including VT4 (vertical lateral tongue), VT6 (vertical tongue apex) and VT8 (lip separation in the vertical direction) for native English speakers and VT6, VT8 and VT10 (vertical midsagittal incisor) for Mandarin accented English speakers the MLLR-PRSW gives better results than the original speaker dependent model.

Figure 3.4 shows the correlation results among different inversion methods. We can see that the performance of the SI-AAI using MLLR-PRSW adaptation frameworks is very close to SD- AAI. Furthermore, the MLLR-PRSW outperforms the PRSW-PRSW (using PRSW adapted models for both acoustic and articulatory streams) for all the articulatory features, except for VT9 (lateral lip rounding).

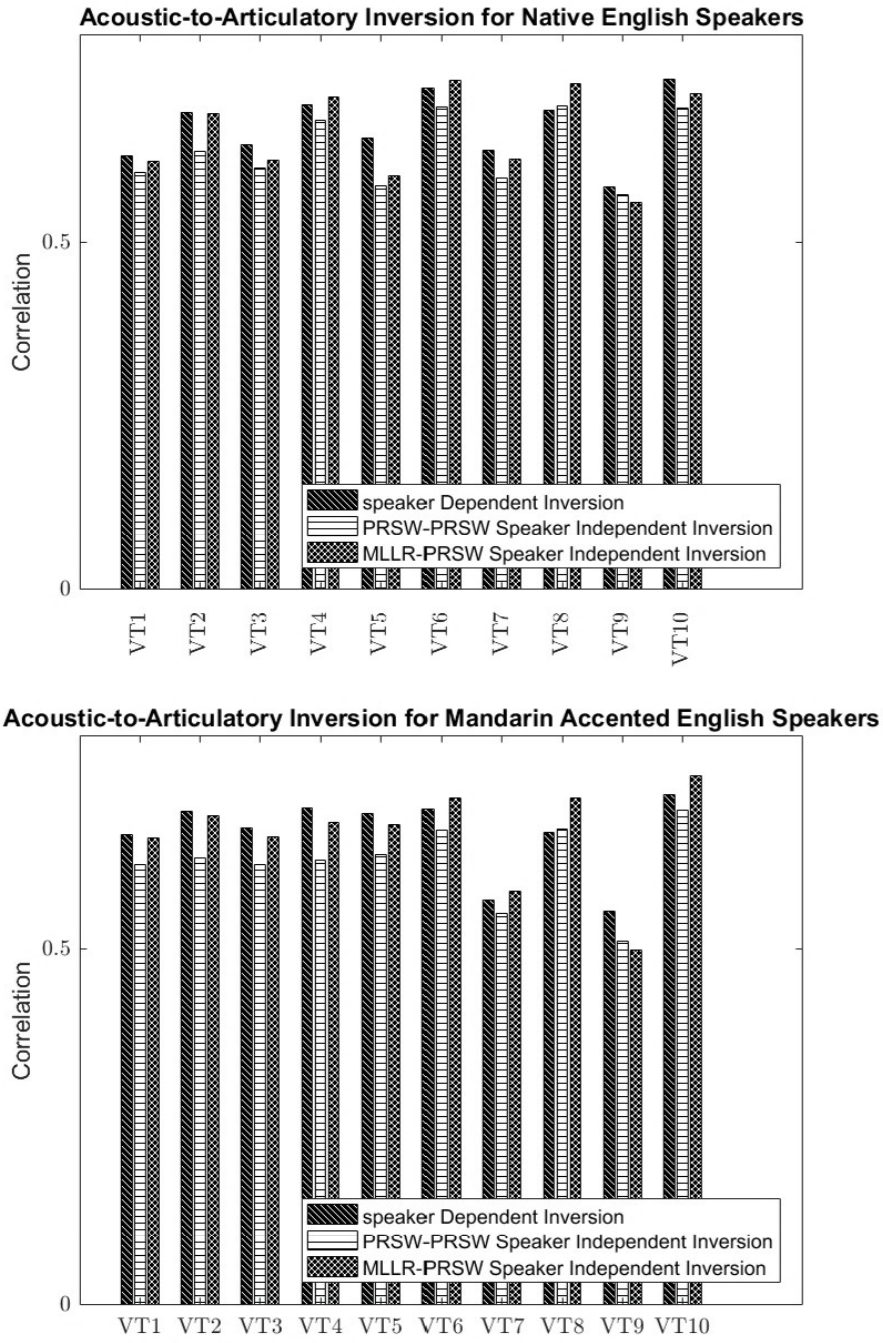


Figure 3.4: Correlation Performance of Inversion Methods for Each Feature

The overall results show that MLLR-PRSW provides consistent representation for acoustic data for L1 and L2 groups compared to the PRSW-PRSW structure. The MLLR-PRSW performance is very close to the SD-AAI. This is illustrated in Figure 3.5 which shows the true and estimated articulatory movements from an example

female native Mandarin speaker. It can be seen that the estimated patterns with the MLLR-PRSW method are very close to the SD-AAI patterns.

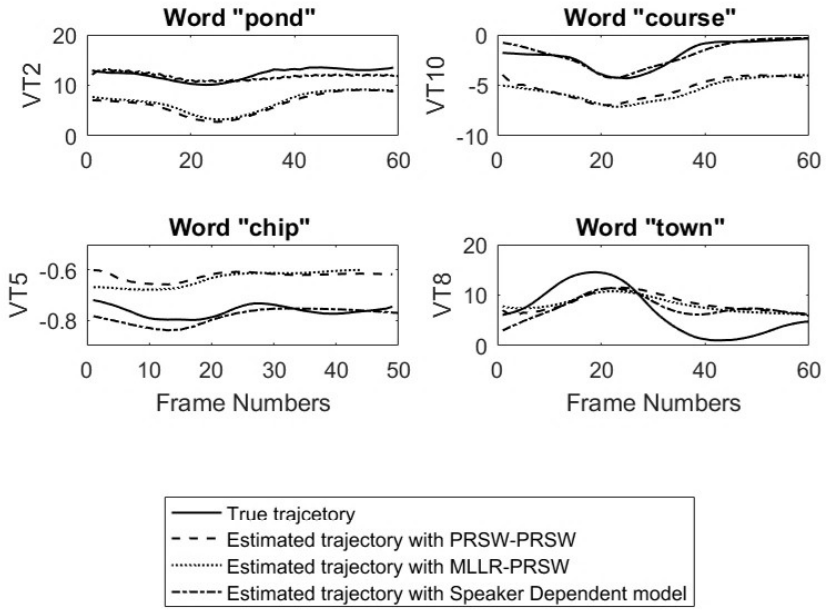


Figure 3.5: Example of Estimated and True Articulatory Trajectories for Different Inversion Approaches

3.3 Summary

This chapter has described different classic ML strategies for SD/SI AAI. The presented approaches SD-AAI include GMM-HMM and GMM-HMM UBM. For SI-AAI, the performance of the PRSW has been compared for different types of reference speaker sets and the new approach MLLR-PRSW introduced for SI-AAI. The next chapter presents different AAI analysis and pronunciation consistency between L1 and L2 speakers.

Chapter 4 Detailed Examination of AAI for L1 and L2 Speakers

The accurate and robust estimation of articulatory trajectories from acoustic data can be used for many applications such as Computer-Aided Language Learning (CALL) [2, 3, 4] Pronunciation Training (CAPT) [2, 3, 4, 5]. Any improvement of these systems can be extremely helpful for their users. The majority of the users of these systems are second language (L2) speakers who are applying these systems to have accurate feedback for pronunciation error correction. Such information can provide them with an efficient method to understand the source of pronunciation defects and correct them based on visualizations of correct articulatory trajectories. Since it is important to understand how the performance of AAI system differs across different language groups for aforementioned applications, in this work a comparative study is implemented to evaluate those differences.

This chapter describes the analysis and comparative studies that have been conducted on articulatory information from L1 and L2 speakers during the AAI process. The first section compares the performance of acoustic-to-articulatory inversion for both L1 and L2 speakers of English, as a function of the number of Gaussian Mixtures used in the GMM-HMM-based inversion model. The second section compares the consistency and predictability of articulatory trajectory patterns between L1 and L2 speakers of English. The last part of this chapter compares the consistency of articulatory positioning between L1 and L2 speakers using EMA data. This comparison was implemented for two different sets of vowels, those that exist in both English and Mandarin and those that exist only in English.

4.1 Comparing The Performance of GMM-HMM-Based SD-AAI For L1 and L2 Speakers

The accuracy of estimated articulatory trajectories from Acoustic-to-Articulatory Inversion relies upon the synchronization of Hidden Markov Model (HMM) states which are derived from the acoustic HMMs using a forced alignment phoneme label sequence. Because of this, the accuracy of acoustic models significantly affects the high quality of the derived HMM alignments.

Gaussian Mixture Models (GMMs) are often used for modeling the acoustic HMMs distribution. Increasing the number of Gaussian mixtures will improve the accuracy of the acoustic model but is limited based on the size of the training data.

The results from the SD-AAI on the EMA-MAE dataset [102] indicate that increasing the complexity of the acoustic models won't necessarily result in better estimation for the kinematic trajectories. It is not clear whether the structure and number of parameters needed for the English (L1) speakers and Mandarin (L2) speakers are the same.

In this work, GMM based systems have been compared to find the best number of mixtures for modeling the acoustic data for these two speaker groups. The inversion approach used in this work is based on a parallel HMM articulatory model tied to the acoustic observation sequence, with dynamic smoothing to account for the presence of discrete rather than continuous state variables. The diagram of the acoustic-articulatory model is illustrated in Figure 2.6.

In the training phase, parallel acoustic-articulatory data are trained separately for each individual speaker. In the inversion stage, the test speech from a new speaker is input to the trained acoustic HMMs to derive an optimal HMM state alignment, and then the corresponding aligned articulatory HMMs can recover the articulatory trajectory.

Using an HMM-based approach allows direct incorporation of the same pronunciation variability model used in the baseline acoustic system, allowing the system significant flexibility in terms of covering the wide range of pronunciation and corresponding articulator patterns present in the L2 speaker group. Once the alignment of articulatory states is complete, the recovery algorithm needs to estimate a smooth articulatory trajectory from the HMM.

In this regard, the average RMSE and Correlation results of the articulatory features and across the two groups of L1 and L2 speakers have been compared as a function of the number of Gaussian mixture components in the acoustic models for each.

Figure. 4.1 shows the average of RMSE and Correlation results for all of the articulatory features and across the two groups of L1 and L2 speakers as a function of the number of Gaussian mixture components in the acoustic models for each. The best performance points have been marked with stars.

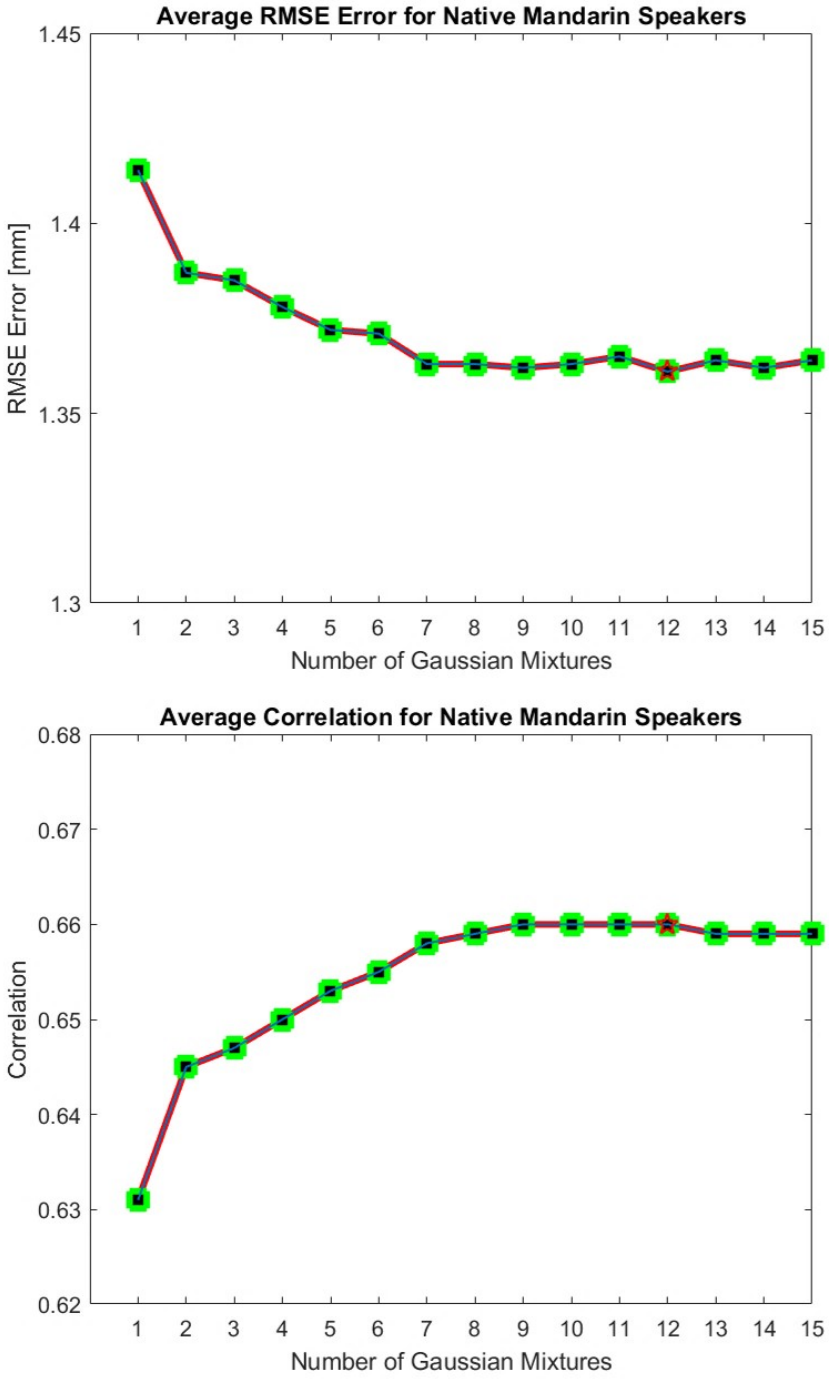


Figure 4.1: RMSE (mm) and Correlation for 10 Estimated Articulatory Features Across 19 Mandarin Accented Speakers Under Different Acoustic Models

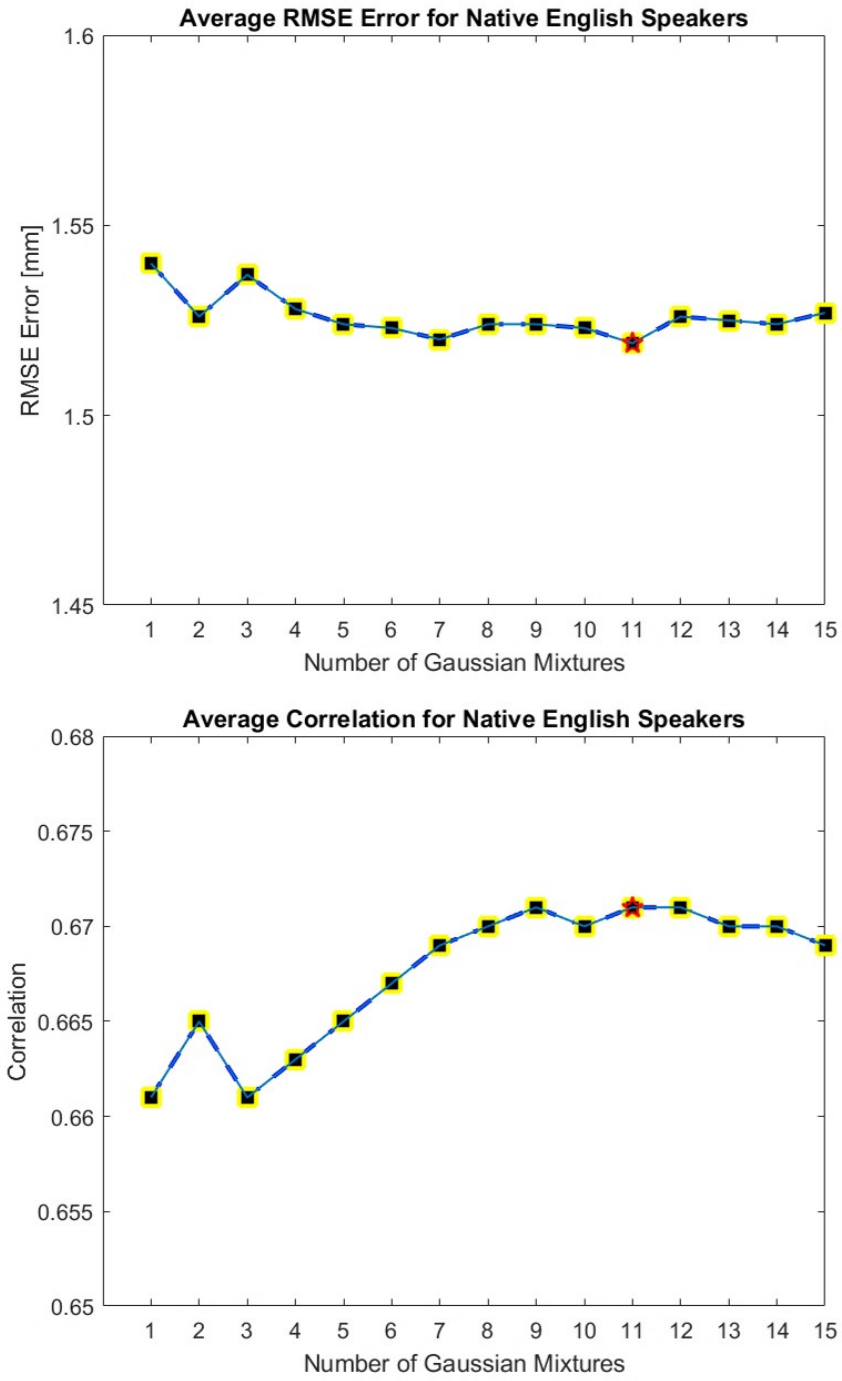


Figure 4.2: RMSE (mm) and Correlation for 10 Estimated Articulatory Features Across 20 Native English Speakers Under Different Acoustic Models

The results in Figure 4.1 show the RMSE and correlation across native Mandarin L2 speakers. This indicates that the best performance for L2 speakers is observed at 12 Gaussian mixture models which means at this point the highest correlation, lowest

RMSE and consequently the best performance is acquired. However, results in Figure 4.2 indicate that the best number of Gaussian mixtures is 11 for L1 speakers.

The result indicates that the Mandarin L2 speakers need more Gaussian mixtures for better performance despite having about the same amount of training data makes sense, since for the L2 acoustic data we expect more complexity and higher variance, and consequently we would need more Gaussian mixtures to represent this complexity.

Generally, for both L1 and L2 speakers, the performance of the number of Gaussian mixtures above the best value (12 and 11 for Mandarin accented English and American English speakers respectively), starts to decrease after that point.

Ordinarily, the upper limit on the number of mixtures, which is directly proportional to the total number of model parameters, is determined by the quantity of training data. To make certain that the model is sufficiently trained, and results will be generalizable to new unseen data, an adequate number of samples is required to estimate means and variances for each mixture in each state. If the number of parameters is incremented further away from this point, the model will begin to overfit to the training data, and test set accuracy will begin to decrease. In this case, using more than 11 and 12 mixtures for L1 and L2 indicates that the model is starting to overfit the training data.

4.2 Comparing The Accuracy of L1 and L2 Estimated Articulatory Features

In this work, we have evaluated the estimated trajectories for 39 speakers under speaker dependent acoustic-to-articulatory inversion. Each individual articulatory feature has been compared between the L1 and L2 speaker groups in different spatial directions.

Table 4.1 shows the RMSE results for American English Speakers and Mandarin accented English Speakers. The better/lower results in this comparison have been bolded.

Articulatory Features	**VT1**	**VT2**	**VT3**	**VT4**	**VT5**	**VT6**	**VT7**	**VT8**	**VT9**	**VT10**
L1 RMSE	**0.12**	3.78	**0.10**	2.36	**0.11**	3.59	0.04	3.35	**0.03**	2.19
L2 RMSE	0.14	**3.12**	0.12	**2.35**	0.14	**3.15**	**0.03**	**2.76**	0.03	**1.80**

Table 4.1: RMSE Results Across all L1 and L2 Speakers

Interestingly, the results comparing L1 and Mandarin L2 speakers indicate that L2 speakers have substantially lower averaged RMSE for the set of articulatory features [VT2, VT6, VT8, VT10], whereas [VT1, VT3, VT5,VT9] English speakers have lower RMSE and for [VT4, VT7] both L1 and L2 speaker groups are similar.

Overall, these differences balance out, with an average RMSE across 19 Mandarin accented English speakers of 1.37 and across 20 American English speakers of 1.57. This suggests that American English speakers have better results for their horizontal tongue and lateral lip position kinematic estimation, however for the horizontal lip

protrusion and vertical midsagittal articulatory motions including the tongue, lips and middle incisor, Mandarin accented English speakers are more consistent and predictable in terms of acoustic to articulatory inversion estimates.

The boxplot results in Figure 4.3 illustrates the degree of variation across articulatory individuals across all L1 and L2 speakers. These boxplots include the median, upper and lower quartile and dynamic range of each individual articulatory feature in different spatial directions. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. The results have been highlighted wherever Mandarin accented English speakers have better performance compared to American English speakers.

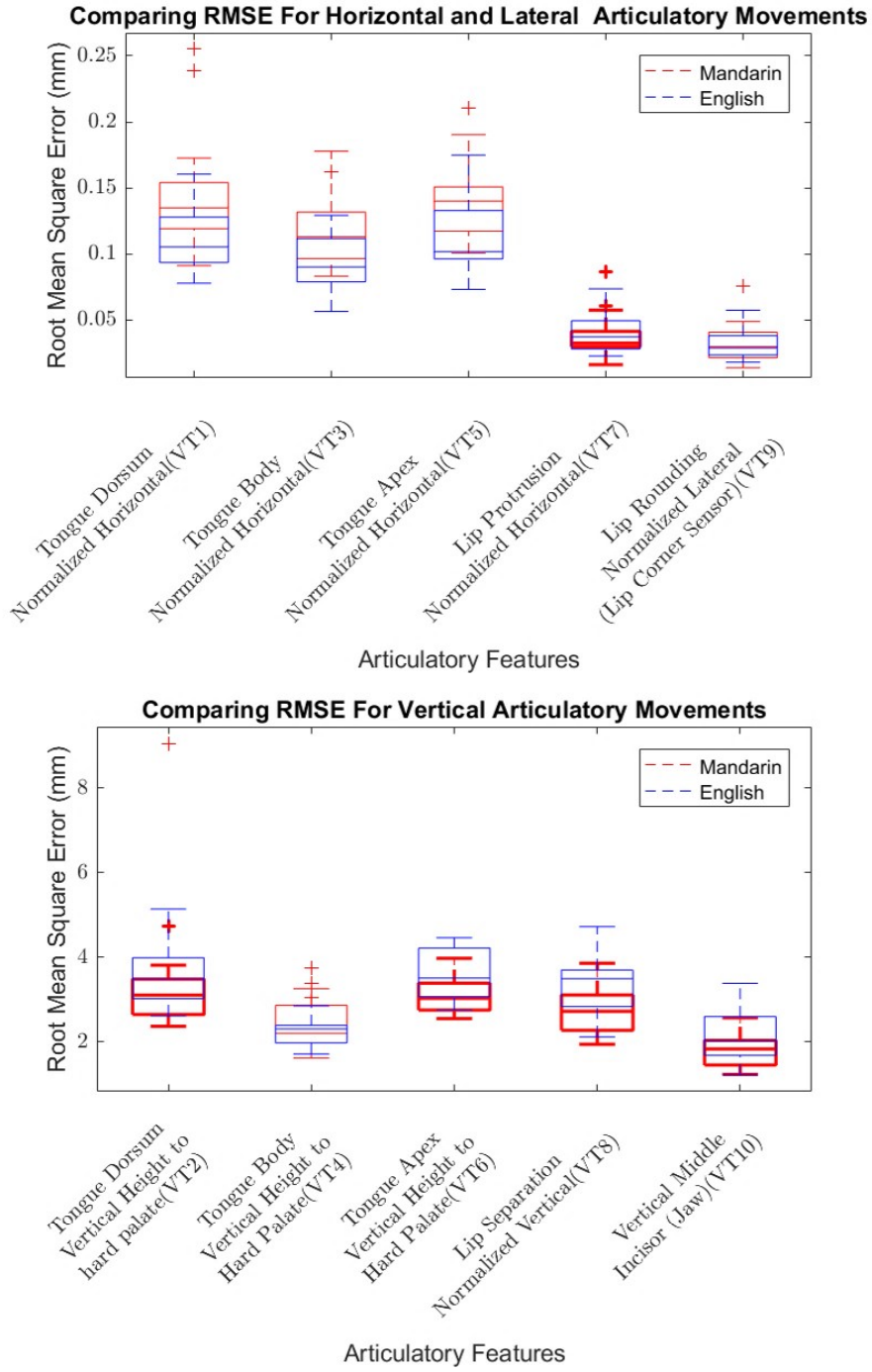


Figure 4.3: RMSE of articulators including tongue, lips and vertical middle incisor in different spatial directions across all the L1 and L2 speakers. The bolded boxplots belong to the Mandarin articulatory motions which have better results (lower RMSE) compared to American speakers. These articulatory motions included horizontal lip protrusion and central (mid-sagittal) vertical motions (including front and back tongue height, the extent of jaw opening, lip separation).

The results shown in Figure 4.3 indicate that Mandarin speakers have more predictable results for their midsagittal vertical articulatory movements and horizontal lip protrusion. It can also be seen that Mandarin speakers have more outliers than English speakers, as might be expected of second-language speakers.

Based on this idea, we hypothesize that the reason L2 speakers show lower RMSE in all the vertical midsagittal articulatory parameters is due to the careful attention to these specific articulators, which are the most important components related to acoustic-phonetic structure. However, controlling the horizontal, lateral and non-midsagittal vertical spatial directions are of secondary importance in contrast to vertical spatial directions, and L2 speakers are unable to maintain sufficient focus to correctly coordinate all articulators. This concept supports existing work in bilingual speech production, for example, the idea that Pronunciation is the only physical part of the language with complex neuromuscular demands (gestures and handwriting use simple movements compared with speech production), and correct pronunciation is strongly dependent on sensory feedback of how and where the articulators are moving, with specific timings and sequences " [103].

4.3 Comparing Articulatory Consistency Between L1 and L2 Speakers

This part of the dissertation presents a comparison between the articulatory configurations of native English speakers L1 and native Mandarin speakers L2 speaking English [104].

A comparison is made between English vowels that have corresponding vowels in Mandarin, versus those that do not, with results supporting the idea that variability of articulator positioning in L2 speakers is larger for vowels that are unique to English than for those that have corresponding vowels in the native language.

To evaluate articulatory consistency, the average articulatory feature values for each individual vowel example have been computed and these values have been applied as the base data for analysis. Then the variance of these average articulatory values has been calculated for each individual vowel under consideration, for each speaker in the dataset.

A higher variance for a given vowel and the speaker indicates that the speaker was not consistent in vowel positioning across different examples of the vowels, suggesting less certainty as to pronunciation.

To compare Mandarin and English speakers, a relative variance metric was implemented. Considering the L1 speakers as the reference of correct pronunciation the relative variance is computed by:

$$RelativeVariance = \frac{Variance_{MandarinSpeakers}}{Variance_{NativeEnglishSpeakers}} \tag{4.1}$$

where variance on the right-hand side specifically indicates the average of the variance values across all Mandarin speakers and all English speakers.

Based on this equation, the relative variance would be more than one if Mandarin speakers have less consistency of vowel pronunciation for that vowel, and less than one if Mandarin speakers have more consistency in articulatory positioning.

Relative variance significantly higher than one may indicate uncertainty in vowel positioning, while values significantly lower than one indicate over-exactness in positioning regardless of phonetic context. Results for the EMA-MAE data set for each vowel under consideration and each articulatory feature are shown in Tables 4.2 and 4.3.

Vowels that Do Not exist in Mandarin	Tongue Dorsum		Tongue Lateral		Tongue Apex		Lip Protrusion	Lip Separation	Lip Rounding
	Horizontal	Vertical	Horizontal	Vertical	Horizontal	vertical	Horizontal	vertical	Lateral
IH	0.48	0.34	1.11	0.75	1.86	0.61	1.39	0.89	0.5
EH	1.54	0.19	1.68	0.90	2.84	1.10	0.81	0.84	1.31
AE	1.10	0.17	0.34	0.01	0.49	0.01	2.29	0.01	2.87
AH	1.93	0.11	1.45	2.67	7.60	3.50	2.03	0.75	2.29
UH	2.55	0.95	2.55	0.95	2.55	2.17	2.55	1.80	2.54
AO	0.60	0.24	0.97	0.49	1.77	1.09	0.56	0.95	0.58
mean	1.37	0.34	1.35	0.96	2.85	1.41	1.61	0.87	1.69

Table 4.2: Comparing Relative Variances Across All the Speakers for Different Vowels That Do not Exist in Mandarin

Vowels that Do exist in Mandarin	Tongue Dorsum		Tongue Lateral		Tongue Apex		Lip Protrusion	Lip Separation	Lip Rounding
	Horizontal	Vertical	Horizontal	Vertical	Horizontal	vertical	Horizontal	vertical	Lateral
AA	0.74	0.88	0.86	1.70	1.59	0.59	1.65	0.76	0.98
AY	0.01	0.18	0.01	0.27	0.01	0.73	0.03	0.52	0.56
IY	0.63	0.24	1.17	0.52	0.66	0.94	0.62	0.42	1.76
OW	0.72	0.03	0.48	0.67	1.58	0.45	0.72	0.76	0.42
UW	0.03	0.28	0.72	1.53	1.00	1.85	0.41	1.28	0.89
mean	0.42	0.32	0.64	0.94	0.97	0.91	0.69	0.75	0.92

Table 4.3: Comparing Relative Variances Across All The Speakers for Different Vowels That Do Exist in Mandarin

The results show that the relative variance is substantially higher on average for vowels that do not exist in Mandarin, in keeping with our hypothesis. However, these relative variance values vary significantly, especially for vowels that do not exist in Mandarin, in some cases (such as vertical tongue dorsum positioning) being much lower than expected. Horizontal tongue position, lip protrusion, and lateral lip position are all substantially higher for these vowels.

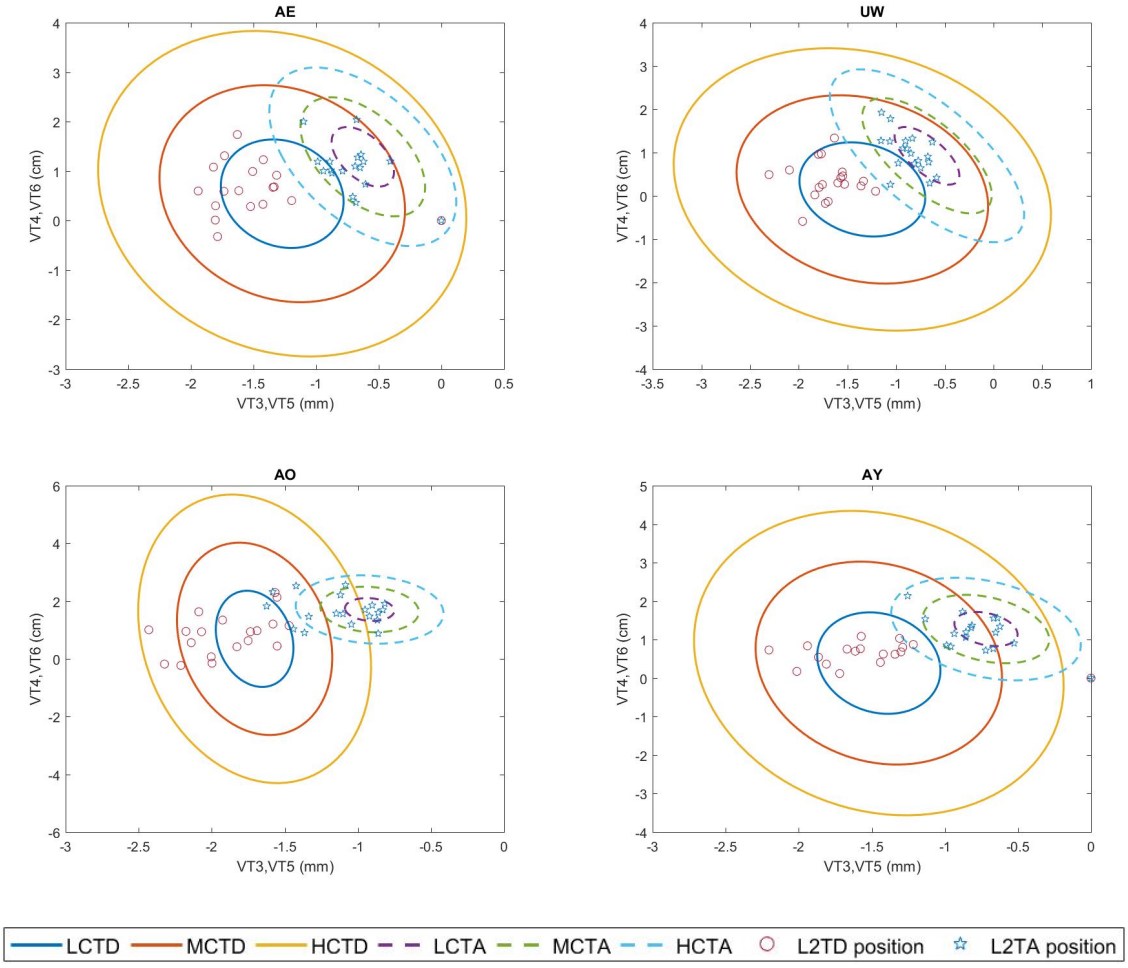
Interestingly, the relative variances for vowels that do exist in Mandarin are all less than 1, in five cases are less than 0.75 and in two cases less than 0.5. This suggests that Mandarin speakers are more consistent in articulatory placement for these vowels than English speakers.

The articulatory features showing the biggest difference between the two vowel sets are Horizontal Tongue Dorsum (1.37 vs. 0.42), Horizontal Tongue Lateral (1.35 vs. 0.64), Horizontal Tongue Apex (2.85 vs. 0.97), Lip Protrusion (1.61 vs. 0.69), and Lip Lateral position (1.69 vs. 0.92). It is interesting to note that these are all non-vertical measures. This suggests that there is much more confusion about horizontal and lateral positioning for Mandarin speakers than vertical articulatory configuration.

To visualize the results, a Standard Deviational Ellipse (SDE) is used to model the structures. Each ellipse is two dimensional using the horizontal and vertical measures for a single sensor, with a center located on the mean value of the data and ellipse based on the covariance matrix using concentric ellipses. The confidence levels were chosen to be 30%, 65%, and 95% (respectively).

According to this model, the center of these ellipses, which is the average of all the native English Speaker's data, would be the most accurate articulatory position or absolute target for pronunciation. In other words, the innermost ellipse conveys the ideal region for articulatory sensor positions, the middle ellipse and the area between it and the innermost ellipse would encompass the tolerable pronunciation and most likely correct area, and the outermost ellipse shows a needs improvement area and that may or may not be corresponded to accurate pronunciation. This efficient method provides Mandarin accented English speakers with the accurate visualizations of correct articulatory positions and a kinematic template for a given vowel. Figure 4.4 demonstrates some of the concentric confidence ellipses that have been provided for L2 speakers to visualize their articulatory positions.

Vowels that Do Not exist in MandarinVowels that Do exist in Mandarin



Vowels that Do Not exist in Mandarin **Vowels that Do exist in Mandarin**

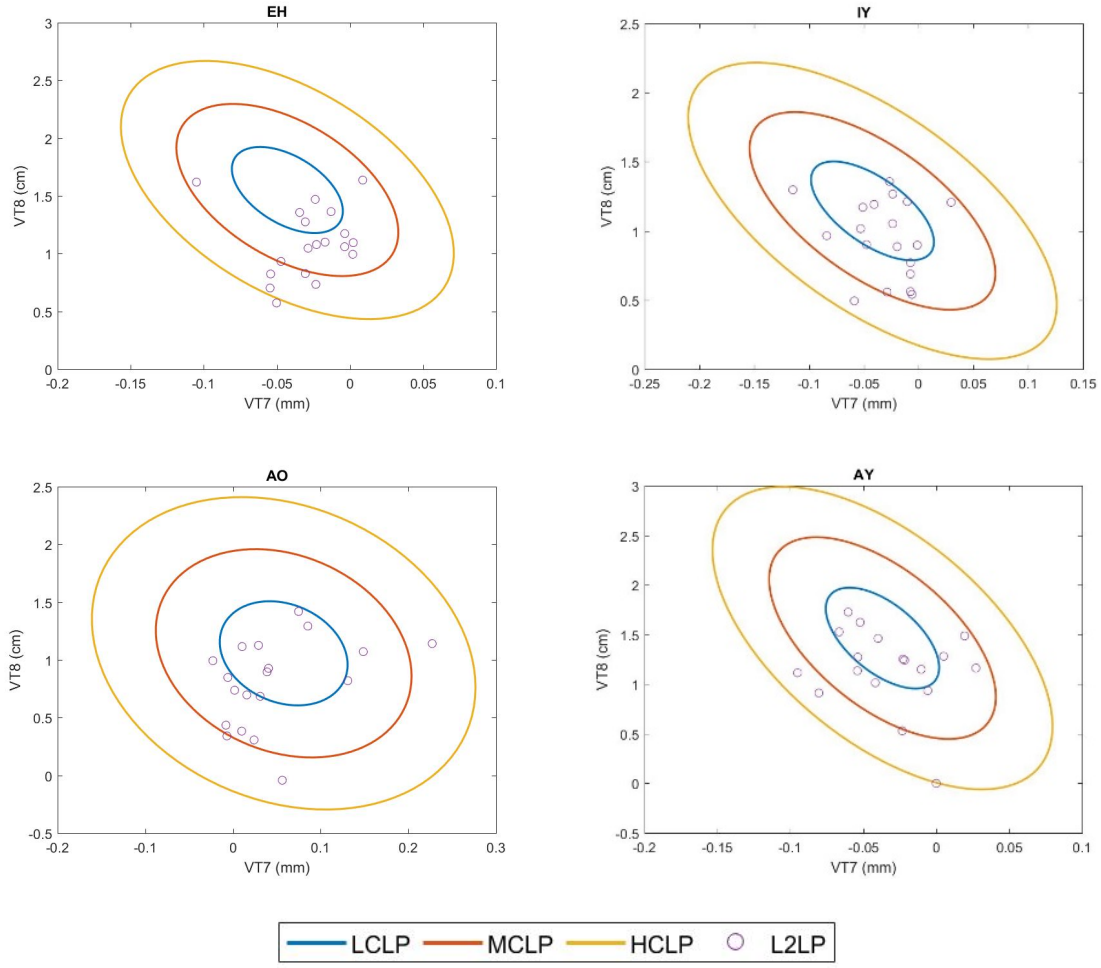


Figure 4.4: Illustrative Examples of The Kinematic Vowel Templates for Accurate Pronunciation, Selected From The Full Set of Vowels.

Figure 4.4 indicates that there is significantly more variability for Mandarin speakers for vowels that do not exist in Mandarin, and less variability compared to English speakers across those vowels that do exist in Mandarin.

4.4 Summary

This chapter has described different comparative studies between L1 and L2 groups of speakers using their articulatory characteristics and trajectory estimations from SD-AAI based on the GMM-HMM frameworks. In these comparative studies, the EMA-MAE dataset is used to compare articulatory consistency and predictability of trajectory patterns from different AAI systems. The next chapter will introduce a novel approach for AAI: Articulatory-WaveNet, which outperforms the previous systems for AAI.

Chapter 5 Deep Autoregressive Framework For AAI

This chapter presents Articulatory-WaveNet, a new approach for Speaker Dependent and Speaker Independent Acoustic-to-Articulatory Inversion. The proposed system uses the WaveNet speech synthesis architecture, with dilated causal convolutional layers using previous values of the predicted articulatory trajectories conditioned on acoustic features.

To evaluate the performance of Articulatory-WaveNet, the EMA-MAE corpus has been used for training the model and predicting articulatory trajectories. The overall results show significant improvement in both correlation and RMSE between the generated and true articulatory trajectories for the new method.

The remaining sections of the chapter are organized as follows:

Section 5.1 describes the Articulatory-WaveNet architecture for acoustic to articulatory inversion. The details about causal convolutional layers, activation functions, and conditioning processes have been provided in this section. Section 5.2 presents the Speaker Dependent-Articulatory WaveNet (SD-AWN) model. The proposed system has been evaluated with the EMA-MAE dataset from 39 speakers and the results have been compared for different subgroups of speakers including Male, Female, L1, and L2 speakers. Section 5.3 introduces Articulatory-WaveNet for Speaker Independent Acoustic-to-Articulatory inversion (SI-AWN). This framework models the relationship between acoustic and articulatory features by conditioning the articulatory trajectories on acoustic features and then utilizes the structure for unseen target speakers. In this set of experiments, different target speakers from the EMA-MAE corpus, including Female and Male from L1 and L2 group of speakers have been evaluated with the proposed framework. Each SD/SI AWN section includes information about data preparation, feature extraction, experimental setup, evaluation, and results. The last part of this chapter summarizes the findings and presents the overall conclusions for the AWN model.

5.1 Articulatory-WaveNet Architecture For AAI

Acoustic-to-Articulatory Inversion is an important speech technology task, allowing for accurate estimation of kinematic articulatory information given acoustic waveform data. AAI is a highly nonlinear and non-unique mapping since different combinations of articulatory movements can result in the same acoustic output. Traditional methods for AAI including Gaussian Mixture Model, Hidden Markov Model and Universal Background Model for SD-AAI, and PRSW and MLLR-PRSW for SI-AAI don't provide high precision estimation for predicting the articulatory trajectories.

Recently there has been significant progress on AAI, with several new approaches based on deep learning published in the last few years that have improved the state

of the art. Along similar lines, this dissertation works toward improving the results of our previous classic-ML approaches with a new deep learning strategy.

The approach introduced here is based on adapting a waveform-based speech synthesizer to the task of articulatory inversion, using the successful text-to-speech WaveNet [14] system and its derivatives.

In this work, we introduce Articulatory-WaveNet (AWN), which consists of stacked dilated convolutional layers to model the conditional probability distribution, providing an accurate estimation of articulatory trajectories from acoustic signals. This fully probabilistic autoregressive architecture predicts the articulatory trajectories from the given acoustic signal by utilizing the causal conditional predictive distribution of samples.

The core of this specific deep autoregressive architecture is the causal or masked convolutional layers. For causal convolutional operations, the occurrence of each sample x is conditioned on the previous samples $(x_1, \ldots, x_{t-1})$. By this assumption, the dependencies on future events or samples are eliminated, and all $P(X_t|x_{<t})$ can be generated in one forward pass. Therefore, AWN models the time series articulatory trajectories with the shifted convolutional results for the required time steps. The property of dilated casual convolution not only captures the long-term dependencies between samples but also significantly grows the receptive field of the network.

To provide a wide receptive field, either the number of neural network layers must add up or a bigger filter for spanning the space have to be acquired. Dilated convolutional layers use the masking convolution technique to dilate the original filter with zeros and grow the receptive field while saving the resolution of inputs and outputs at the same level. A vanilla CNN can be considered as a dilated convolutional architecture with dilation set at 1.

AWN has been built up from the stacked convolutional layers. Each stack contains non-linear activation unites for modeling the nonlinear acoustic-articulatory time-series signal and it follows up by residual and parametrized skip connections to speed up the convergence and enable us to design a deeper architecture.

The goal of AWN is to model the sequence of articulatory trajectories that have been conditioned on the sequence of time-series acoustic features. The predicted articulatory trajectories are synthesized from the fully trained network. Figure 5.1 illustrates AWN architecture.

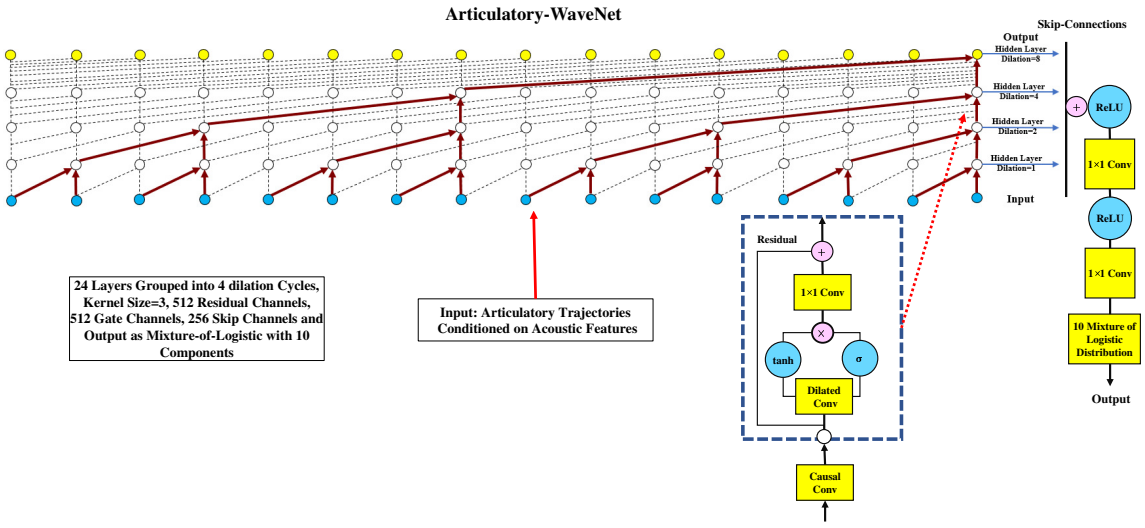


Figure 5.1: Visualization of Articulatory-WaveNet (AWN), Stacked Causal Convolutional Layers, With an Overview of The Residual Block and Overall Architecture.

The AWN architecture that is shown in Figure 5.1 uses *tanh* as the activation function which results in the outputs in the range of $[-1, 1]$. For speech synthesis applications this would properly work since the range of changes in the audio acoustic waveform is between -1 and 1. However, for articulatory sensor measurements, the mapping range would cause a loss of information. In order to avoid this problem, the primary articulatory inputs have been scaled to the range of $[-1, 1]$ using global dynamic range normalization:

$$(\text{Scaled Articulatory Feature})_i = 2\left(\frac{\text{Articulatory Feature} - Min_i}{Max_i - Min_i}\right) - 1$$

The dynamic range normalization is unique to each speaker and articulatory variable, with Max_i and Min_i representing the overall maximum and minimum of all articulatory trajectories for speaker i. This structure allows for easy conversion of predicted trajectories to the original feature space after the synthesizing step.

To speed up the synthesizing process, AWN uses the Fast-WaveNet Generation Algorithm [20] described in Section 2.9.3 Fast WaveNet caches previously computed information from the overlapping network states, called recurrent states, to eliminate redundant convolutions. To implement this, the network computes a new output sample using the caching information from the recurrent states. This is a significant computational improvement over WaveNet which re-computes all states at each time step.

The designed AWN used in the following experiments has 24 layers with 4 dilation stacks. The dilation rate increases by a factor of 2 in every layer at each stack. This

starts with no dilation (rate 1) and reaches a maximum dilation of 32. For this experiment we considered 4 stacks:

$1, 2, 4, 8, 16, 32; 1, 2, 4, 8, 16, 32; 1, 2, 4, 8, 16, 32; 1, 2, 4, 8, 16, 32.$

The stacking grows the receptive field size and increases the capacity of the network. The kernel size of the causal dilated convolutions is 3, with 512 units in the gating layers and residual connection channels and 256 hidden units at the skip connection channel and $1 * 1$ convolution before the output layer.

The output is modeled as a mixture of 10 logistic components for higher quality. To compute the logistic mixture distribution, the AWN stack output is passed through a ReLU activation followed by a linear projection to predict parameters $\theta = \{\{Mean\}\mu_i, \{LogScale\}S_i, \{MixtureWeight\}\pi_i\}$ for each mixture component. The loss is computed as the negative log-likelihood of the ground truth sample. The likelihood of sample x_t is:

$$P(x_t|\theta, h_t) = \sum_{i=1}^{k=10} \pi_i[\sigma(\frac{\tilde{x}_{ti} + 0.5}{S_i}) - \sigma(\frac{\tilde{x}_{ti} - 0.5}{S_i})] \tag{5.1}$$

Where $\tilde{x}_{ti} = x_t - \mu_i$ and $P(x_t|\theta, h_t)$ is the probability density function of the articulatory trajectory conditioned on mel-spectrogram h_t.

5.2 AWN for SD-AAI

In this work [105, 106], we have applied the Articulatory-WaveNet (AWN) architecture for SD-AAI, designated SD-AWN.

This experiment compares the performance of SD-AWN with our previous classical-ML SD-AAI framework that has been investigated in section 3.1. This baseline model consisted of parallel acoustic and articulatory HMMs, with dynamic smoothing to account for the presence of discrete rather than continuous state variables. In the training phase, parallel acoustic-articulatory data was trained separately for each individual speaker. In the inversion stage, the test speech as input to the trained acoustic HMMs to derive an optimal HMM state alignment, and then the corresponding aligned articulatory HMMs were used to recover the articulatory trajectory. Once the alignment of articulatory states is complete, the recovery algorithm estimates a smooth articulatory trajectory from the HMM.

The results on the EMA-MAE corpus, shown in Section 5.2.3, show significant improvement for RMSE and Correlation compared to the baseline GMM-HMM system across both L1 and L2 groups of speakers.

5.2.1 Data Preparation and Feature Extraction

The articulatory feature set used for SD-AWN experiments using the EMA-MAE corpus consists of 6 tongue-related features, 3 lip-related features, and a jaw feature. The tongue features include the 3 horizontal distances to the tip, dorsum, and lateral sensors and the 3 vertical distances between the sensors and the hard palate. Lip features include lip protrusion, lip separation, and lateral distance to the corner lip sensor, which is indicative of lip rounding. Table 5.1 represents the articulatory feature set applied for evaluating SD-AWN. Unlike the articulatory feature set at

Table 2.1 this set of features do not have normalization at the horizontal articulatory movements and lip characteristics for simplicity. Articulatory features were calculated point-by-point on the 400Hz EMA data, then downsampled by a factor of 4 to give one feature every 10ms.

VT Feature	Description
VT1	Tongue Dorsum Horizontal Position
VT2	Tongue Dorsum Vertical Height to HardPalate
VT3	Lateral Tongue Horizontal Position
VT4	Lateral Tongue Vertical Height to HardPalate
VT5	Tongue Tip Horizontal Position
VT6	Tongue Tip Vertical Height to HardPalate
VT7	Horizontal Lip Protrusion
VT8	Vertical Lip Separation
VT9	Lateral Lip Corner (Lip Corner Sensor)
VT10	Vertical Middle Incisor (Jaw)

Table 5.1: Vocal Tract Features for SD-AWN

For the acoustic data, Mel-Spectrograms features are used. Mel-Spectrograms are extracted through a Hanning windowed Short-Time Fourier Transform with 38.7ms frame size and 9.7ms frame hop. Log dynamic range compression is implemented using an 80 channel Mel filter bank spanning the range of 125Hz to 7.6kHz.

5.2.2 Experimental Setup and Evaluation

For this set of experiments, the utterances from EMA-MAE were used for training, development, and evaluating the performance of the SD-AWN framework. For the training set, 4000 utterances were randomly selected across all the speakers (102-103 utterances per speaker) and 0.04 percent of training utterances have been split for the validation set. For the test set, another 580 utterances (14-15 utterances per speaker) were randomly selected. The training and test sets were selected separately, and include both sentence and word speech samples from EMA-MAE.

The AWN network was trained for 20,000 epochs using the ADAM optimizer. There are 8 mini-batches with each mini-batch containing a maximum of 8000-time steps (roughly 302ms).

To measure the accuracy of the proposed SD-AWN system, two metrics, RMSE, and correlation, have been considered in this experiment. The performance of SD-AWN has been compared with the previous classic-ML framework, the GMM-HMM model. SD-AAI results with GMM-HMM method give an average correlation between

actual measured and estimated trajectories of 0.61 and an average RMSE of 2.83mm.

5.2.3 Results and Analysis

Table 5.2. shows the overall RMSE and correlation results of each individual articulatory feature averaged across all 39 speakers in the EMA-MAE dataset for two SD-AAI systems of using SD-AWN and GMM-HMM approaches.

ArticulatoryTrajectories	**CORRELATION**			**RMSE (Millimeters)**		
	HMM-GMM	**ART-WN**	**%increase**	**HMM-GMM**	**ART-WN**	**%decrease**
Horizontal Tongue Dorsum (VT1)	0.59	0.84	42.3	3.41	1.14	66.5
Tongue Dorsum Vertical Height to Hard Palate (VT2)	0.64	0.82	28.1	3.44	1.24	63.9
Horizontal Lateral Tongue (VT3)	0.61	0.83	36.1	2.62	1.40	46.5
Lateral Tongue Vertical Height to Hard Palate (VT4)	0.66	0.82	24.2	2.35	1.29	45.1
Horizontal Tongue Tip (VT5)	0.62	0.83	33.9	3.28	1.62	50.6
Tongue Tip Vertical Height to Hard Palate (VT6)	0.65	0.82	26.2	3.37	1.66	50.7
Horizontal Lip Protrusion (VT7)	0.55	0.82	49.1	3.37	0.26	92.2
Vertical Lip Separation (VT8)	0.61	0.84	37.7	3.05	1.65	45.9
Lateral Lip Corner (VT9)	0.50	0.82	64.0	0.83	0.18	78.3
Vertical Middle Incisor (Jaw) (VT10)	0.67	0.81	20.9	1.99	2.08	-4.3
MEAN	0.61	0.83	36.1	2.83	1.25	55.8

Table 5.2: Performance Comparison of The SD-AWN And HMM-GMM

Overall, the SD-AWN improved correlation from 0.61 to 0.83 (36% increase) and decreased RMSE from 2.83mm to 1.25mm (56% decrease) over the baseline GMM-HMM system, averaged across all speakers (both native English and native Mandarin) and articulatory features. Looking at RMSE specifically, the most significant improvements are for the horizontal Lip Protrusion, reduced from 3.37mm to 0.26mm (92.28), lateral Lip Corner, reduced from 0.83mm to 0.18mm (78.31), vertical and horizontal Tongue Dorsum, reduced from 3.44 and 3.41mm to 1.44 and 1.24mm (66% and 63.95% decrease), and vertical and horizontal Tongue Tip, reduced from 3.37 and 3.28mm to 1.66 and 1.62mm (51% decrease). The average RMSE for tracking the vocal tract height at the three tongue sensors, key variables for capturing physiological characteristics of tongue motion is 1.39mm, down from 3.05mm for the baseline method. Speaker horizontal tongue sensor positions have an average RMSE of 1.38mm, down from 3.10mm. Vertical lip separation had an RMSE of 1.65mm, down from 3.05mm. Horizontal lip protrusion and Lateral lip distance both show slightly lower RMSEs 0.26mm and 0.18mm, down from 3.37mm and 0.83mm respectively.

The middle incisor (jaw) sensor shows a slightly higher RMSE 2.08mm compared to baseline 1.99mm, which is interesting since it showed an improved correlation.

Correlation results show consistent improvement across all features, with all 10 of the articulatory feature trajectories having correlations above 80%, ranging from 81% to 84%.

Figure 5.2 shows the measured EMA and estimated articulatory movements for a selection of speakers and articulatory features, for visualization of the results.

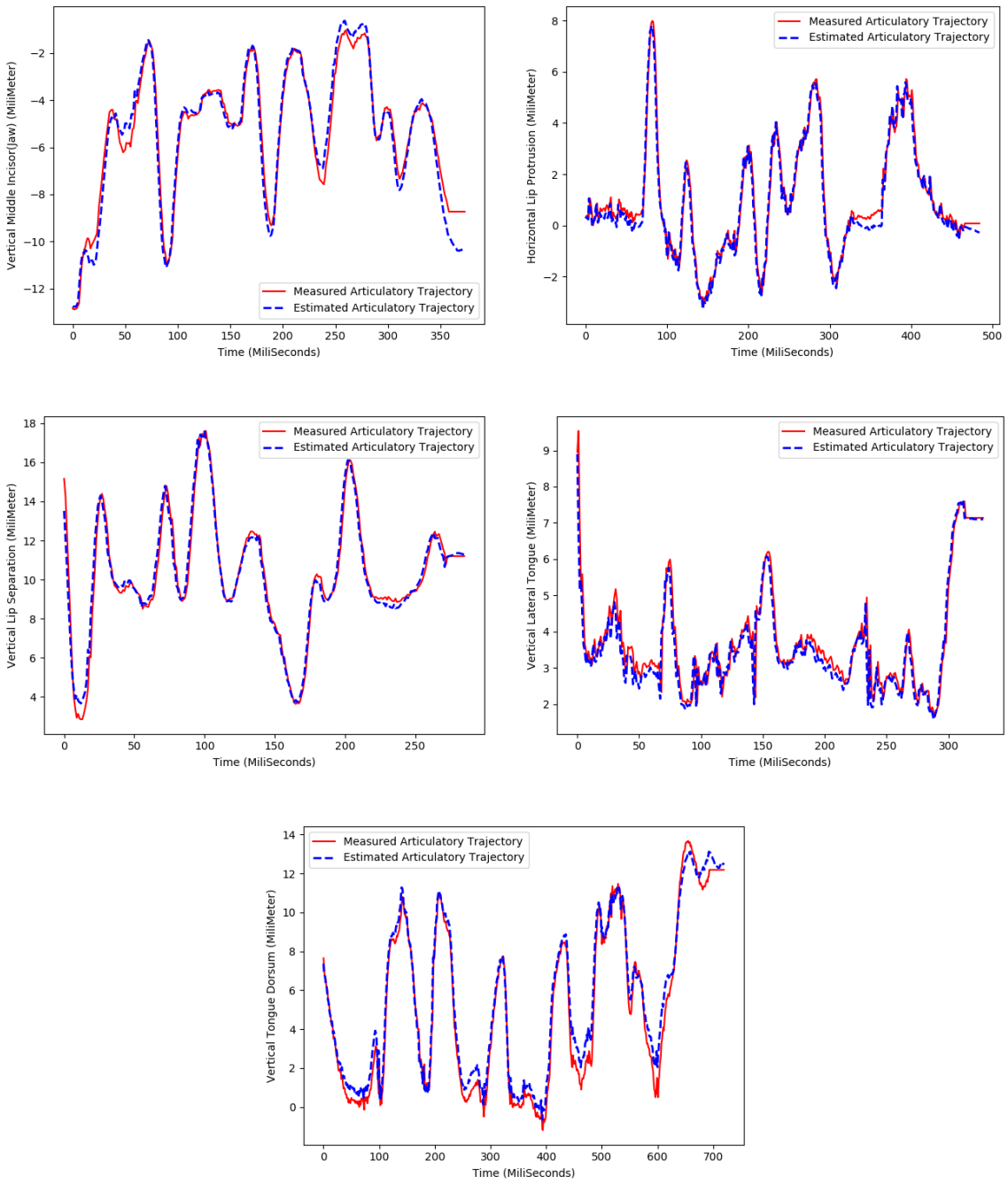


Figure 5.2: Trajectories of Selected Articulatory Features From a Typical Test Sentence Utterances. The plots show the trajectories that have been estimated by SD-AWN alongside the target actual articulatory trajectories.

In addition, we also compared the performance of SD-AWN for different subgroups of speakers. The results show the consistency of performance of the proposed architecture for predicting articulatory features from acoustic features across different subgroups of speakers L1, L2, and Male and Female speakers. Table 5.3. Shows the

RMSE and Correlation results for these different groups of speakers.

	RMSE (Millimeter) Results for Articulatory Trajectories										
Gender	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8	VT9	VT10	Average
Male	1.24	1.50	1.56	1.36	1.81	1.88	0.23	1.67	0.18	1.84	1.33
Female	1.04	0.97	1.23	1.21	1.42	1.43	0.30	1.63	0.19	2.34	1.18
	Correlation Results for Articulatory Trajectories										
Gender	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8	VT9	VT10	Average
Male	0.84	0.81	0.82	0.82	0.81	0.80	0.83	0.82	0.80	0.80	0.82
Female	0.83	0.82	0.82	0.81	0.82	0.81	0.82	0.83	0.82	0.80	0.82
	RMSE (Millimeter) Results for Articulatory Trajectories										
L1/L2	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8	VT9	VT10	Average
MN	1.16	1.28	1.91	1.62	2.35	1.71	0.33	1.81	0.16	2.06	1.44
EN	1.13	1.21	0.91	0.98	0.93	1.62	0.20	1.50	0.20	2.11	1.08
	Correlation Results for Articulatory Trajectories										
L1/L2	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8	VT9	VT10	Average
MN	0.83	0.82	0.82	0.81	0.81	0.80	0.83	0.84	0.80	0.81	0.82
EN	0.84	0.81	0.82	0.82	0.82	0.81	0.83	0.81	0.81	0.79	0.82

Table 5.3: Performance Comparison of The Articulatory-WaveNet for The Different L1/L2 and Male/Female subgroups.

Results indicate that the correlation is consistent across different types of L1, L2, male and female groups of speakers and it remains around 82%. However, RMSE results differ across speakers. A comparison of RMSE results for L1 and L2 speakers shows that for L1 results are more accurate, with L2 speakers having 0.36mm higher RMSE. This is consistent with what might be expected for L1 vs. L2 speaker groups in terms of pronunciation and articulatory consistency.

Looking at results for L1 English speakers in particular, several articulatory features including lips, tongue, and incisor show improvement with SD-AWN compared to the previous best-reported approaches. The average RMSE from Latent Trajectory DNN [62] approach for the vertical tongue (tip, body and dorsum) is around 1.80mm while for SD-AWN the vertical tongue (tip, lateral and dorsum) RMSE for English speakers is 1.27mm. The best-reported results with the CNN+BLSTM approach in [72] for 12 articulatory features including lip, jaw and tongue are reported around 0.84 correlation and 1.4mm RMSE. The SD-AWN approach for English speakers has a similar correlation, at 0.82, but an RMSE of only 1.08mm.

In addition, female speakers have slightly better results for RMSE compared to male speakers, 0.15mm lower.

5.3 AWN For SI-AAI

In this work, a new speaker independent method for Acoustic-to-Articulatory Inversion is introduced. The proposed architecture, Speaker Independent-Articulatory

WaveNet (SI-AWN), models the relationship between acoustic and articulatory features by conditioning the articulatory trajectories on acoustic features and then utilizes the structure for unseen target speakers.

SI-AWN has 24 layers with 4 dilation stacks. The dilation rate increases by a factor of 2 in every layer at each stack which starts with no dilation (rate 1) and reaches a maximum dilation of 32. The stacking enlarges the receptive field size and increases the capacity of the network. The kernel size of the causal dilated convolutions is 3, with 512 units in the gating layers and residual connection channels and 256 hidden units at the skip connection channel and $1 * 1$ convolution before the output layer. The output is modeled as a mixture of 10 logistic components.

To compute the logistic mixture distribution, the AWN stack output is passed through a ReLU activation followed by a linear projection to predict parameters for each mixture component. Like SD-AWN, in this experiment, we considered loss as the negative log-likelihood of the ground truth sample. Equation 5.1 shows how the likelihood is computed.

SI-AWN like SD-AWN uses the Fast-WaveNet Generation Algorithm to increase the speed of articulatory synthesizing. Fast-WaveNet is implemented by computing the new output sample by calling the information from the recurrent states. Therefore, the unnecessary computational effort for evaluating all states at each time step will be omitted.

The SI-AWN trained for 20,000 epochs using the ADAM optimizer. There are 8 mini-batches with each mini-batch containing a maximum of 8000-time steps.

The former implemented framework, MLLR-PRSW, as described in Section 3.2.2, is considered as the baseline system for comparison. According to this method, the maximum likelihood of the adaptation data is computed by acoustic information from a new target speaker, and MLLR is used to update the mean values of a Uniform Background Model (UBM) model. During the MLLR process, the statistical specifications are provided from available adaptation acoustic data from the target speaker and are used to find the linear regression-based transformation for the parameters. PRSW [4] uses a parallel adapted kinematic model to estimate the weights for the reference speakers based on target speaker acoustic adaptation data. It then combines the weighted reference models to create the adapted articulatory model for the target speaker.

The next section presents the results of evaluating the proposed SI-AWN on the EMA-MAE, using the pool of acoustic-articulatory information from 35 reference speakers and testing on target speakers that include male, female, native and non-native speakers.

The results suggest that SI-AWN improves the performance of the acoustic-to-articulatory inversion process compared to the baseline MLLR-PRSW method.

5.3.1 Data Preparation and Feature Extraction

In this experiment, we have used a set of articulatory features that are more representative of actual tongue movement compared to the raw sensor positions. The key element of the features is the use of palate trace data to compute vertical articulatory features which provide actual physical vocal tract height as the vertical

distance between the sensor and hard palate. Table 5.4 lists the selected articulatory features used in this experiment. To avoid the complexity at Table 2.1 this set of features are not normalized in horizontal direction or at lip characterizations.

Tongue Dorsum Horizontal Position
Tongue Dorsum Vertical Height to Hard Palate
Lateral Tongue Horizontal Position
Lateral Tongue Vertical Height to Hard Palate
Tongue Tip Horizontal Position
Tongue Tip Vertical Height to Hard Palate

Table 5.4: Vocal Tract Features for Tongue Movement

For modeling the acoustic space, Mel Frequency Cepstral Coefficients are extracted through a short-time Fourier Transform STFT with 21.5ms (1024 samples) frame size, 441 samples for frame hop, and a Hanning window function. Then the log dynamic range compression of the 80 channel Mel filter bank spanning the range of 5.6kHz to 3.4kHz has been used to transfer the STFT magnitude.

In order to generate the same number of acoustic features compared to the samples from articulatory features, the acoustic raw signals which have a 22050 Hz Sampling Rate (SR) are upsampled by a factor of 8 before the feature extraction process. The upsampling factor is computed by the following equation:

$$\frac{Articulatory - Feature_{SR} \times FrameSize}{AcousticWaveForm_{SR}} = \frac{400 \times 441}{22050} = 8 \tag{5.2}$$

5.3.2 Experimental Setup and Evaluation

The parallel acoustic-articulatory information from 35 speakers from the EMA-MAE dataset has been used to train the reference model, and 4 other speakers are selected as target speakers set to evaluate the AWN model for the SI-AAI task.

We evaluated the performance of SI-AWN using the correlation between actual and estimated trajectories. Correlation needs to be used rather than RMSE because SI-AAI leads to an offset in terms of mean and dynamic range relative to the true unknown kinematics, even if accurately estimating the trajectory.

5.3.3 Results and Analysis

Table 5.5 shows the averaged correlation results of each individual articulatory feature across two different methods for SI-AAI: proposed SI-AWN framework and MLLR-PRSW baseline model.

Articulatory Trajectories	Direction	Model		Percent of Improvement
		SI-AWN	MLLR-PRSW	
Tongue Dorsum	Horizontal	0.80	0.68	17.6%
	Vertical	0.82	0.69	18.8%
Tongue Lateral	Horizontal	0.81	0.62	30.6%
	Vertical	0.84	0.71	18.3%
Tongue Tip	Horizontal	0.78	0.60	30.0%
	Vertical	0.82	0.73	12.3%
mean		0.81	0.67	20.9%

Table 5.5: Performance Comparison of The SI-AWN and MLLR-PRSW

Table 5.5 indicates that correlation results have been consistently improved by the new method across all features. On average correlation has increased from 0.67 to 0.81 (21%improvement) over the baseline MLLR-PRSW system, averaged across all speakers and articulatory features. The individual tongue articulatory feature trajectories have correlations ranging from 78% to 84%. The most significant improvements are for the horizontal Tongue Lateral and Tongue Dorsum, which increased from 0.62 to 0.81 and 0.60 to 0.78 respectively (30% improvement).

The average correlation across each articulatory feature for different target speakers is shown in Table 5.6.

Articulatory Features	L1-F	L2-F	L2-M	L1-M	Mean
Horizontal Tongue Dorsum	0.79	0.81	0.81	0.77	0.80
Vertical Height to Hard Palate Tongue Dorsum	0.80	0.84	0.81	0.83	0.82
Horizontal Lateral Tongue	0.81	0.81	0.81	0.82	0.81
Vertical Height to Hard Palate Lateral Tongue	0.84	0.84	0.82	0.86	0.84
Horizontal Tongue Tip	0.76	0.80	0.78	0.78	0.78
Vertical Height to Hard Palate Tongue Tip	0.86	0.83	0.82	0.77	0.82
Average Correlation Score	0.81	0.82	0.82	0.81	0.81

Table 5.6: Performance Comparison of The SI-AWN For The Different L1/L2 and Male/Female Subgroups.

35 speakers including 18 Male/Female from L1 and 17 Male/Female from L2 have been used as a reference set for training the SI-AWN model. The target speaker

set includes one male and one female from each of the L1 and L2 speaker sets, designated as L2-M, L2-F, L1-M, and L1-F respectively. The correlation results from all the speakers are above 80% and show similar performance regardless of gender or native language. The average correlation for tracking the vocal tract height at the three tongue sensors, key variables for capturing physiological characteristics of tongue motion, is 0.83. Speaker horizontal tongue sensor positions have an average correlation of 0.80.

Figure 5.3 demonstrates the true and estimated articulatory movements from different examples of word utterances. In this figure, selected estimated trajectories from the SI-AWN have been compared with the true articulatory patterns, showing a strong correlation.

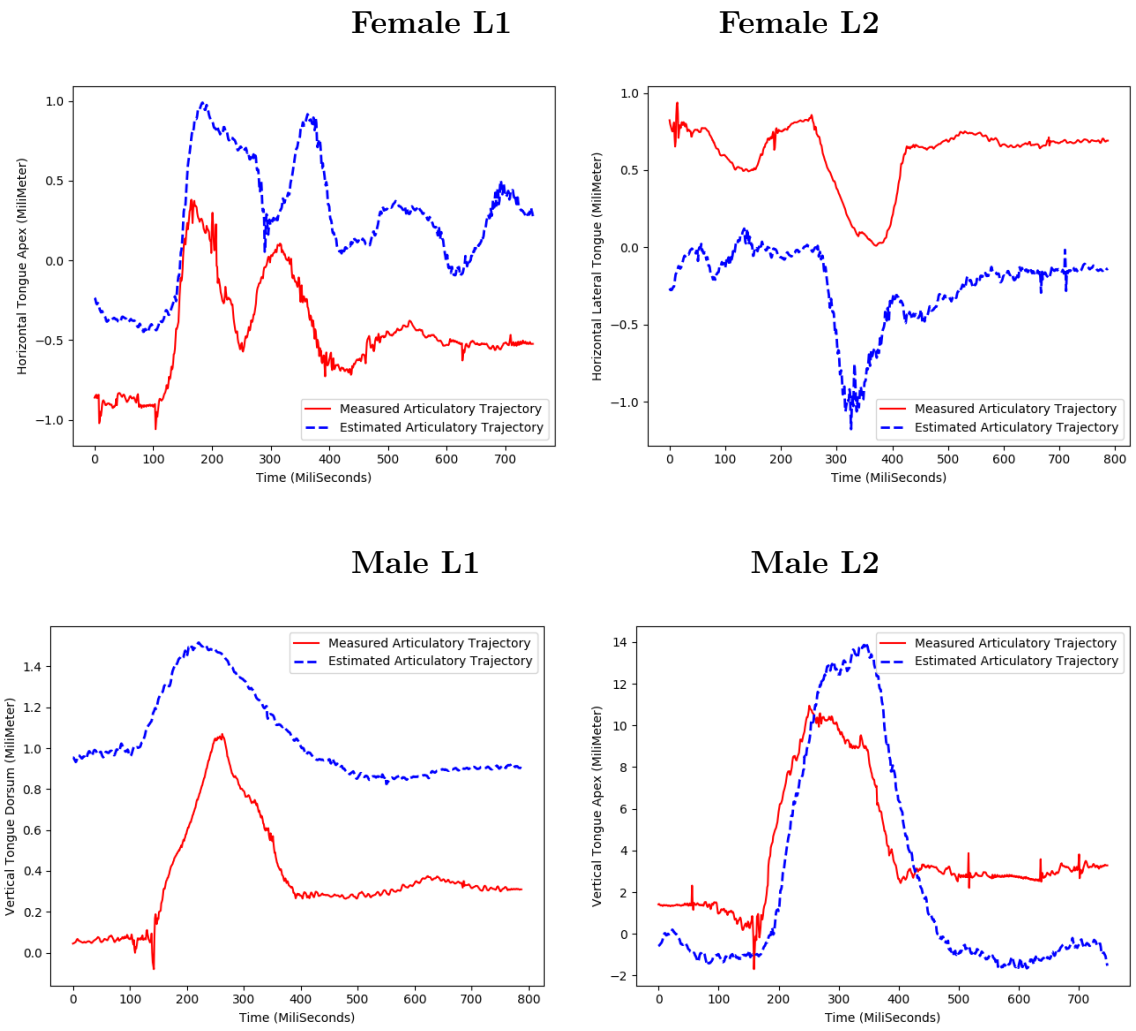


Figure 5.3: Trajectories of selected articulatory features from typical test sentence utterances. The plots show the trajectories that have been estimated by SI-AWN alongside the target actual articulatory trajectories.

Figure 5.3 shows that without articulatory information for the target speaker the estimated articulatory trajectories represent the correlated movement patterns but not necessarily the target speaker articulatory positions, because of a baseline shift. The estimated articulatory features are impacted by both physiological differences and sensor placement differences across subjects from reference and target speaker sets. Therefore, using the correlation metric is a more appropriate evaluation metric compared to the mean squared error for SI-AAI.

5.4 Summary

In this chapter, we have proposed a novel model, the Articulatory-WaveNet approach, based on stacked dilated convolutional layers for both speaker-dependent and speaker-independent acoustic-to-articulatory inversion. The new system has been compared with the performance of our previously proposed classic-ML frameworks, GMM-HMM and MLLR-PRSW, and evaluated on the EMA-MAE dataset including male and female English and Mandarin speakers. The overall results show that AWN provides more robust estimations for speaker-dependent and independent inversion compared to the classic-ML structure.

Chapter 6 Conclusions and Future Work

This chapter summarizes the original contributions and conclusions presented in this dissertation. Some future research subjects are also suggested that could improve and facilitate the progress of the important topics discussed in this work.

6.1 Original Contributions

Acoustic-to-Articulatory Inversion (AAI) is the non-linear regression problem of estimating articulatory trajectories from an acoustic signal. There are a wide variety of signal and speech processing applications of AAI, including automatic speech recognition and computer aided language learning. AAI is an ill-posed problem since it is highly non-linear and since multiple combinations of articulatory movements can generate similar speech acoustics. Without previous knowledge about articulatory specifications of the target speaker, the task of AAI becomes harder and more challenging to solve.

This dissertation has focused on addressing AAI through a novel deep autoregressive architecture, as well as presenting articulatory-based comparisons across L1 and L2 speakers and investigating the AAI performance of these diverse speaker groups. The contributions of this work are summarized as follows:

1. A comparative study has been conducted to evaluate the performance of several Classic-ML algorithms for SD-AAI. Established methods such as a GMM-HMM approach with and without UBM adaptation have been implemented to compare the effectiveness and performance of previous models for articulatory inversion. Results indicate that using adapted GMM-HMM models from UBM for both acoustic and articulatory spaces results in the best performance for a baseline SD-AAI framework. (chapter 3)

2. Reference speaker selection for PRSW based SI-AAI has been investigated. Reference speakers with different accents and quantities were used to determine the adaptation weights for estimating the articulatory model. The reference speaker sets were selected not only based on their performance in SD-AAI but also on the type of accent. A comparison has been made between different types of target and reference speakers, with results indicating that the accuracy of the adapted model increases when we select balanced distributed accents of English and a lower number of reference speakers. (chapter 3)

3. An improved version of classic-ML SI-AAI has been presented using acoustic adaptation to estimate weights for articulatory model creation from reference speakers. The new method applies different adaptation approaches for the acoustic model and the weighted articulatory model. The results show that the new approach of combining MLLR adaptive model for acoustic data and PRSW adaptive model for articulatory data is very accurate and provides results close to those of speaker-dependent GMM-HMM-based AAI model, outperforming

the speaker-dependent model for some articulatory features. The new MLLR-PRSW model gave an average correlation of 0.67 for Native English speakers and 0.67 for native Mandarin speakers, in comparison with the baseline PRSW-PRSW approach of 0.63 and 0.62, respectively. These results are on par with the baseline speaker dependent results for GMM-HMM with SD-AAI averaging 0.66. (chapter 3)

4. The trade-offs in performance and accuracy of the GMM-HMM-based SD-AAI have been investigated across a varying number of Gaussian Mixtures for the two different groups of speakers, including Mandarin accented English speakers (L2) and American English speakers (L1). Both RMSE and correlation metrics have been considered to identify the best number of Gaussian mixtures on the EMA-MAE dataset. It is shown that increasing the number of mixtures beyond this results in overfitting and lower performance. (chapter 4)

5. A comparative study has been presented to compare AAI accuracy between Mandarin accented English L2 speakers and American English L1 speakers. Since second language speakers have difficulties in achieving native-like production and control of articulatory movements, it was initially expected that their estimated results should be less accurate for AAI in all directions compared to the native speakers. However, the GMM-HMM-based SD-AAI experiments on EMA-MAE corpus show that this is not the case for several key articulatory variables, specifically including measurements related to midsagittal vocal tract height. For most other spatial directions the Mandarin speakers have less accurate articulatory prediction as expected. Our hypothesis from this observation is that the Mandarin speakers are more careful about control of their central (mid-sagittal) vertical motions (including front and back tongue height, extent of jaw opening, lip separation), as well as horizontal lip protrusion, to the detriment of other parts of their articulatory patterns including horizontal tongue positioning and lateral tongue curvature.(chapter 4)

6. The articulatory configurations of native English speakers and native Mandarin speakers speaking English has been investigated. The study was conducted between English vowels that have corresponding vowels in Mandarin, versus those that do not, with results supporting the idea that variability of articulator positioning in L2 speakers is larger for vowels that are unique to English than for those that have corresponding vowels in the native language. This is especially true for Tongue Apex, Lip Protrusion and Lip Rounding articulatory features. (chapter 4)

7. A new approach for AAI problem has been presented. The proposed system, Articulatory-WaveNet, uses the WaveNet speech synthesis architecture, with dilated causal convolutional layers to predict articulatory trajectories conditioned on acoustic features. The system was trained and evaluated on multiple groups of speakers including Female/Male and L1/L2 speakers from EMA-MAE corpus, and shows significant improvement for RMSE and Correlation compared

to the baseline GMM-HMM system, with correlations above 80% for all articulatory trajectory estimates and an average RMSE of 1.25mm across both L1 and L2 speaker groups. Within native English speakers, average RMSE across the set of ten articulatory features for the proposed method is 1.08mm, demonstrating state-of-the-art results on the AAI task. (chapter 5)

8. The new Articulatory-WaveNet approach has been extended to speaker independent AAI. The proposed architecture, Speaker Independent-Articulatory WaveNet (SI-AWN), models the relationship between acoustic and articulatory features by conditioning the articulatory trajectories on acoustic features and then utilizes the structure for unseen target speakers. The overall results show that SI-AWN provides more robust estimations for speaker-independent inversion compared to the MLLR-PRSW structure with an average correlation of 0.81. This is the first application of a WaveNet synthesis approach to the problem of SI-AAI, and results are comparable to or better than the best currently published systems.(chapter 5)

6.2 Recommendations for Future Work

Based on the results presented in this dissertation, recommended directions for future study include the following:

1. The current Articulatory-WaveNet has a fixed architectural structure. This could be extended to modify the dilations, layers, regularizations, optimizers and activation functions to compare robustness and accuracy across different architectural characteristics.

2. The new Articulatory-WaveNet method has been evaluated using the EMA-MAE bilingual multi-speaker corpus of parallel acoustic and EMA kinematic data. For conducting more detailed comparisons of articulatory features with other AAI approaches, the AWN architecture should be implemented on other common datasets in this field such as MOCHA and MNGU0.

3. The experiments with Articulatory-WaveNet has used raw original data from EMA-MAE corpus. There are recent augmentation methods for generating data that could potentially be used to improve the performance of the AWN. These methods can help the SI-AWN framework to achieve better consistency and more accurate estimations for predicting the articulatory trajectories from unseen target speakers.

4. In this dissertation, the AAI approaches have been implemented using a set of ten individual articulatory characteristics to model the articulatory space of human speech. This set of articulatory features have been assumed to be independent and the relationship between them has not been considered or modeled in any of the experiments presented here. However, the dependencies

between articulatory features can play an important role in AAI system. Therefore, future work to consider and represent the relationships among individual articulatory features may further improve overall AAI accuracy.

5. The EMA-MAE corpus also contains information about articulatory sensor orientation, but this information has not been incorporated into the feature representations used for AAI. Future studies could consider the orientation information along with the position to improve the articulatory feature representation for AAI and improve the accuracy of predictions through more comprehensive articulatory information.

6. For SI-AWN the system uses a single trained architecture for estimating the articulatory trajectories across different target speakers. This SI-AWN architecture could be customized through incorporation of speaker-specific information obtained from a small amount of acoustic adaptation data prior to inversion. Methods for this could include concepts such as eigen-voice weighting speaker embedding incorporated into the global conditioning stage of the AWN network.

6.3 Conclusion

In this dissertation, a new deep autoregressive method for acoustic-to-articulatory inversion has been introduced. This new model uses dilated causal convolutional layers to predict the articulatory trajectories from acoustic feature sequences. The novel propose method, Articulatory-WaveNet, has been implemented using the parallel acoustic-articulatory data of 39 speakers including both native English speakers and Mandarin accented English speakers from ElectroMagnetic Articulography-Mandarin Accented English corpus, EMA-MAE. Results demonstrate that the new system significantly out performs the baseline classic-machine learning algorithms both for both speaker dependent and speaker independent acoustic-to-articulatory inversion mappings.

In addition, this work has presented several comparative studies between articulatory patterns of native and Mandarin accented English speakers. These studies include a comparison of articulatory dynamics as well as AAI performance, across L1 and L2 speaker groups.

References

[1] B. E. Murdoch, *Speech and language disorders associated with subcortical pathology*. John Wiley & Sons, 2009.

[2] D. A. Rosenbaum, *Human motor control*. Academic press, 2009.

[3] A. Ji, M. T. Johnson, and J. J. Berry, "Parallel reference speaker weighting for kinematic-independent acoustic-to-articulatory inversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1865–1875, Oct 2016.

[4] A. Ji, "Speaker independent acoustic-to-articulatory inversion," Ph.D. Dissertation, Marquette University, Wisconsin, USA, 2014.

[5] D. K. Jones, "Development of kinematic templates for automatic pronunciation assessment using acoustic-to-articulatory inversion," M.S. thesis, Marquette University, Wisconsin, USA, 2017.

[6] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and blstm-based deep tone models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2012–2024, 2019.

[7] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory." in *INTERSPEECH*, 2018, pp. 3122–3126.

[8] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, 2019.

[9] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion." in *INTERSPEECH*, 2016, pp. 455–459.

[10] G. Fant, "Acoustic theory of speech production (mouton, the hague)(1970)," *The closely spaced horizontal lines shown in Fig. 1A are the harmonics of the fundamental frequency of phonation, and are typically revealed in narrowband spectrograms*, 1960.

[11] C. Qin and M. Á. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[12] M. Parrot, J. Millet, and E. Dunbar, "Independent and automatic evaluation of acoustic-to-articulatory inversion models," *arXiv preprint arXiv:1911.06573*, 2019.

[13] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded gaussian mixture regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2246–2259, 2015.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[15] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[16] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Wavecyclegan2: Time-domain neural post-filter for speech waveform generation," *arXiv preprint arXiv:1904.02892*, 2019.

[17] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[18] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, and Bengio, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[19] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," *arXiv preprint arXiv:1910.10909*, 2019.

[20] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, "Fast wavenet generation algorithm," *arXiv preprint arXiv:1611.09482*, 2016.

[21] S. Maiti and M. I. Mandel, "Parametric resynthesis with neural vocoders," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 303–307.

[22] K. Kastner, J. F. Santos, Y. Bengio, and A. Courville, "Representation mixing for tts synthesis," in *ICASSP 2019-2019 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5906–5910.

[23] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[24] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 7–12.

[25] J. E. Flege, "Production and perception of a novel, second-language phonetic contrast," *The Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1589–1608, 1993.

[26] E. Mcdermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE transactions on information and systems*, vol. 89, no. 3, pp. 1006–1014, 2006.

[27] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[28] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database users handbook," *University of Wisconsin*, 1994.

[29] D. Zhang, X. Liu, N. Yan, L. Wang, Y. Zhu, and H. Chen, "A multichannel/multi-speaker articulatory database in mandarin for speech visualization," in *The 9th International Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 299–303.

[30] X. Xie, X. Liu, and L. Wang, "Deep neural network based acoustic-to-articulatory inversion using phone sequence information." in *Interspeech*, 2016, pp. 1497–1501.

[31] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *In Proceedings 5 th Seminar of Speech Production*, 2000, pp. 305–308.

[32] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[33] W. J. Hardcastle and A. Marchal, "Eur-accor: A multi-lingual articulatory and acoustic database," in *First International Conference on Spoken Language Processing*, 1990.

[34] J. M. Scobbie, A. Turk, C. Geng, S. King, R. Lickley, and K. Richmond, "The edinburgh speech production facility doubletalk corpus," in *INTERSPEECH 2013: Proceedings of the 14th Annual Conference of the International Speech Communication Association (ISCA), 25-29 August 2013, Lyon, France.* International Speech Communication Association, 2013.

[35] P. MacNeilage, "Speech production mechanisms in aphasia," in *Speech motor control.* Elsevier, 1982, pp. 43–60.

[36] I. of Electrical and E. Engineers, "Ieee recommended practice for speech quality measurements," *IEEE transactions on audio and electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[37] A. Ji, J. J. Berry, and M. T. Johnson, "The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 7719–7723.

[38] S. Maiti and R. Smith, "Engineering department, cambridge university trumpington street, cambridge cb2 1pz, england tel: 0223 66466," *International Journal of Fracture*, vol. 27, p. R31, 1985.

[39] J. E. Flege, "The phonological basis of foreign accent: A hypothesis," *Tesol Quarterly*, vol. 15, no. 4, pp. 443–455, 1981.

[40] Y. Chen, M. Robb, H. Gilbert, and J. Lerman, "Vowel production by mandarin speakers of english," *Clinical Linguistics & Phonetics*, vol. 15, no. 6, pp. 427–440, 2001.

[41] L. Mi, S. Tao, W. Wang, Q. Dong, J. Guan, and C. Liu, "English vowel identification and vowel formant discrimination by native mandarin chinese-and native english-speaking listeners: The effect of vowel duration dependence," *Hearing research*, vol. 333, pp. 58–65, 2016.

[42] T. L. Gottfried and D. Riester, "Relation of pitch glide perception and mandarin tone identification," *Journal of the Acoustical Society of America*, vol. 108, no. 5, p. 2604, 2000.

[43] B. Lindblom, J. Lubker, and T. Gay, "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," *Journal of Phonetics*, vol. 7, no. 2, pp. 147–161, 1979.

[44] H. C. Chen and Q. Wang, "Development and application of a corpus-based online pronunciation learning system for chinese learners of english." *English Teaching & Learning*, vol. 40, no. 2, 2016.

[45] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2301–2312, 2012.

[46] A. Suemitsu, J. Dang, T. Ito, and M. Tiede, "A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning," *The Journal of the Acoustical Society of America*, vol. 138, no. 4, pp. EL382–EL387, 2015.

[47] M. Wieling, P. Veenstra, P. Adank, and M. Tiede, "Articulatory differences between l1 and l2 speakers of english," in *Proceedings paper Proceedings of the 11th International Seminar on Speech Production*, 2017.

[48] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[49] M. Picheny, B. Ramabhadran, S. F. Chen, and M. Nussbaum-Thom, "The big picture/language modeling," *https://www.ee.columbia.edu/ stanchen/spring16/e6870/slides/lecture5.pdf*, 2016.

[50] H. Fayek, "Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and whats in-between," *URL: https://haythamfayek. com/2016/04/21/speech-processingfor-machine-learning. html*, 2016.

[51] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[52] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[53] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 237–240.

[54] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[55] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory hmm," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[56] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[57] T. Hueber, L. Girin, L. Alameda-Pineda, and G. Bailly, "Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded gaussian mixture regressions," 2013.

[58] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Ninth International Conference on Spoken Language Processing*, 2006.

[59] ——, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *International Conference on Nonlinear Speech Processing*. Springer, 2007, pp. 263–272.

[60] S. Hiroya, "Acoustic-to-articulatory inversion using a speaker-normalized hmm-based speech production model," in *Proc. ISSP*. Citeseer, 2008, pp. 7–12.

[61] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4624–4627.

[62] P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 1274–1277.

[63] V. Mitra, G. Sivaraman, C. Bartels, H. Nam, W. Wang, C. Espy-Wilson, D. Vergyri, and H. Franco, "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5205–5209.

[64] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[65] Z. Wei, Z. Wu, and L. Xie, "Predicting articulatory movement from text using deep architecture with stacked bottleneck features," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[66] G. Sivaraman, C. Y. Espy-Wilson, and M. Wieling, "Analysis of acoustic-to-articulatory speech inversion across different accents and languages." in *INTERSPEECH*, 2017, pp. 974–978.

[67] Z. Cai, X. Qin, D. Cai, M. Li, X. Liu, and H. Zhong, "The dku-jnu-ema electromagnetic articulography database on mandarin and chinese dialects with tandem feature based acoustic-to-articulatory inversion," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 235–239.

[68] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, "Multi-corpus acoustic-to-articulatory speech inversion," *Proc. Interspeech 2019*, pp. 859–863, 2019.

[69] A. Illa and P. K. Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech & Language*, vol. 59, pp. 75–90, 2020.

[70] ——, "Closed-set speaker conditioned acoustic-to-articulatory inversion using bi-directional long short term memory network," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL171–EL176, 2020.

[71] A. Illa, P. K. Ghosh *et al.*, "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5075–5079.

[72] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5931–5935.

[73] R. Mannem, J. Mallela, A. Illa, and P. K. Ghosh, "Acoustic and articulatory feature based speech rate estimation using a convolutional dense neural network," *Proc. Interspeech 2019*, pp. 929–933, 2019.

[74] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4450–4454.

[75] X. Xie, X. Liu, T. Lee, and L. Wang, "Investigation of stacked deep neural networks and mixture density networks for acoustic-to-articulatory inversion," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Nov 2018, pp. 36–40.

[76] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated rnn," 2018.

[77] P. Maud, M. Juliette, and D. Ewan, "Independent and automatic evaluation of acoustic-to-articulatory inversion models," *arXiv preprint arXiv:1911.06573*, 2019.

[78] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition," in *Twelfth annual conference of the international speech communication association*, 2011.

[79] L. Theis and M. Bethge, "Generative image modeling using spatial lstms," in *Advances in Neural Information Processing Systems*, 2015, pp. 1927–1935.

[80] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 29–37.

[81] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[82] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems*, 2016, pp. 4790–4798.

[83] N. Kalchbrenner, A. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Video pixel networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1771–1779.

[84] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[85] S. Spratley, D. Beck, and T. Cohn, "A unified neural architecture for instrumental audio tasks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 461–465.

[86] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[87] B. Pharris, "Nv-wavenet: Better speech synthesis using gpu-enabled wavenet inference," *NVIDIA Developer Blog*, 2018.

[88] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," *arXiv preprint arXiv:1702.07825*, 2017.

[89] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[90] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[91] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2018.

[92] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "Fftnet: A real-time speaker-dependent neural vocoder," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2251–2255.

[93] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[94] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in neural information processing systems*, 2018, pp. 10 215–10 224.

[95] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[96] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *Interspeech*, vol. 2017, 2017, pp. 1118–1122.

[97] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[98] B. K.-W. Mak and R. W.-H. Hsiao, "Kernel eigenspace-based mllr adaptation," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 784–795, 2007.

[99] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[100] N. Bozorg and M. T. Johnson, "Reference speaker selection for kinematic-independent acoustic-to-articulatory-inversion," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1932–1932, 2019.

[101] ——, "Mllr-prsw for kinematic-independent acoustic-to-articulatory inversion," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2019, pp. 1–5.

[102] ——, "Comparing performance of acoustic-to-articulatory inversion for mandarin accented english and american english speakers," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 1–5.

[103] A. J. Simmonds, R. J. Wise, and R. Leech, "Two tongues, one brain: imaging bilingual speech production," *Frontiers in Psychology*, vol. 2, p. 166, 2011.

[104] N. Bozorg, M. T. Johnson, and J. J. Berry, "Comparing articulatory consistency between native and second language speakers," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2019, pp. 1–5.

[105] N. Bozorg and M. T. Johnson, "Articulatory-wavenet: Autoregressive model for acoustic-to-articulatory inversion," *arXiv preprint arXiv:2006.12594*, 2020.

[106] ——, "Acoustic-to-articulatory inversion with deep autoregressive articulatory-wavenet," *Networks (CNNs)*, vol. 22, p. 23.