Development of Kinematic Templates for Automatic Pronunciation Assessment Using Acoustic-to-Articulatory Inversion

Deriq K. Jones
Marquette University

DEVELOPMENT OF KINEMATIC TEMPLATES FOR AUTOMATIC PRONUNCIATION ASSESSMENT USING ACOUSTIC-TO-ARTICULATORY INVERSION

by

Deriq K. Jones, B.S.

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin
August 2017

ABSTRACT

DEVELOPMENT OF KINEMATIC TEMPLATES FOR AUTOMATIC PRONUNCIATION ASSESSMENT USING ACOUSTIC-TO-ARTICULATORY INVERSION

Deriq K. Jones, B.S.
Marquette University, 2017

Computer-aided pronunciation training (CAPT) is a subcategory of computer-aided language learning (CALL) that deals with the correction of mispronunciation during language learning. For a CAPT system to be effective, it must provide useful and informative feedback that is comprehensive, qualitative, quantitative, and corrective. While the majority of modern systems address the first 3 aspects of feedback, most of these systems do not provide corrective feedback. As part of the National Science Foundation (NSF) funded study "RI: Small: Speaker Independent Acoustic-Articulator Inversion for Pronunciation Assessment", the Marquette Speech and Swallowing Lab and Marquette Speech and Signal Processing Lab are conducting a pilot study on the feasibility of the use of acoustic-to-articulatory inversion for CAPT.

In order to evaluate the results of a speaker's acoustic-to-articulatory inversion to determine pronunciation accuracy, kinematic templates are required. The templates would represent the vowels, consonant clusters, and stress characteristics of a typical American English (AE) speaker in the midsagittal plane. The Marquette University electromagnetic articulography Mandarin-accented English (EMA-MAE) database, which contains acoustic and kinematic speech data for 40 speakers (20 of which are native AE speakers), provides the data used to form the kinematic templates. The objective of this work is the development and implementation of these templates.

The data provided in the EMA-MAE database is analyzed in detail, and the information obtained from the analysis is used to develop the kinematic templates. The vowel templates are designed as sets of concentric confidence ellipses, which specify (in the midsagittal plane) the ranges of tongue and lip positions corresponding to correct pronunciation. These ranges were defined using the typical articulator positioning of all English speakers of the EMA-MAE database. The data from these English speakers were also used to model the magnitude, speed history, movement pattern, and duration (MSTD) features of each consonant cluster in the EMA-MAE corpus. Cluster templates were designed as set of average MSTD parameters across English speakers for each cluster. Finally, English stress characteristics were similarly modeled as a set of average magnitude, speed, and duration parameters across English speakers.

The kinematic templates developed in this work, while still in early stages, form the groundwork for assessment of features returned by the acoustic-to-articulatory inversion system. This in turn allows for assessment of articulatory inversion as a pronunciation training tool.

ACKNOWLEDGEMENTS

Deriq K. Jones, B.S.

I have a number of people to thank for their assistance with both this work and all of my work at Marquette.

I thank my research advisor, Dr. Michael Johnson, for recruiting me to work on this project, and for his guidance during my time at Marquette. His instruction and advice increased both my quality of work and interest in speech processing applications.

I thank Dr. Jeffrey Berry, whose guidance and assistance were crucial to the completion of this project. His expertise laid the groundwork for a number of concepts introduced in this work, and played a significant role in my understanding of several speech analysis and language learning concepts.

I thank Dr. Richard Povinelli for serving as my co-advisor, and for helping develop my research and presentation skills through our regular meetings.

I thank my fellow lab members, who offered a great deal of assistance in the collection and processing of data required for the completion of this project.

Finally, I thank my friends and family, whose support and encouragement have brought me to where I am today.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 GENERAL BACKGROUND

Computer assisted language learning, or computer aided language learning (CALL) refers to the use of computers in the learning and teaching of foreign languages [1]. The development of CALL is essential in ensuring that people of various backgrounds are able to communicate and function effectively, despite language barriers. CALL has a long history that traces its roots to as early as the 1960's [1]. Early implementations typically presented a stimulus (usually in the form of on-screen text) to the learner, who would provide a response (usually via keyboard). The technological advances made as time progressed allowed for increased capabilities of CALL systems, including the incorporation of recorded voice and video, as well as speech recognition techniques for instruction and evaluation [1].

A key component of modern CALL implementations is computer assisted pronunciation training (CAPT), which deals in correction of mispronunciation during language learning. Useful and informative feedback is an important an important aspect of CAPT, as it plays a large role in developing a language learner's background in the new language. T. K. Hansen identified four essential aspects of feedback during CAPT: comprehensive, qualitative, quantitative, and corrective [2]. While the majority of current systems, which are typically based on automatic speech recognition (ASR) techniques [3] [4], can implement the first three aspects, most do not provide meaningful corrective feedback. The Marquette Speech and Signal Processing Lab and Speech and Swallowing

Lab have developed an acoustic-to-articulatory inversion system with the intention of meeting this final criteria [5].

Acoustic-to-articulatory inversion is the estimation of articulatory parameters from acoustic signals. In other words, this inversion accepts a speech signal then estimates and returns a set of articulatory features modeling the position and movement of articulators required to produce the speech from the input signal. In order to perform this inversion, a system must be trained using both acoustic and kinematic speech data.

As part of the National Science Foundation (NSF) funded study “RI: Small: Speaker Independent Acoustic-Articulator Inversion for Pronunciation Assessment”, the Marquette Speech and Swallowing Lab and Marquette Speech and Signal Processing Lab are conducting a pilot study on the use of acoustic-to-articulatory inversion for Computer Aided Pronunciation Training (CAPT). The acoustic-to-articulatory inversion system developed through this project analyzes English speech and predicts the motion of the articulators, including the jaw, lower lip, upper lip, and tongue, required to produce the corresponding sounds. In order to obtain train the inversion system, speech data was collected from several native American English (AE) and Mandarin accented English (MAE) speakers to form the Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus. In order to give the inversion system a frame of reference for pronunciation assessment, kinematic templates are needed for the system. These templates, through the modeling of EMA data, represent the positioning and movement of articulators (specifically, the tongue and lips) for correct pronunciation. Like the acoustic-to-articulatory inversion system, the kinematic templates were developed using the data from the EMA-MAE corpus.

1.2 Pilot Study Overview

This work focuses on the development and implementation of the kinematic templates described in section 1.1. Using these templates, detailed pronunciation assessment measures will be implemented for a small set of target pronunciation error categories for native Mandarin speakers of English. The pronunciation categories include vowels, consonant clusters, and contrastive stress. A consonant cluster is a set of two or more adjacent consonants in a word. Stress refers to the pattern of emphasis given to certain parts of words and sentences, and a contrastive stress is a stress on a syllable or word that is imposed contrary to its typical pronunciation in order to emphasize the word or syllable or to contrast it with another word or syllable [6]. The nature of clustering and stress, as well as the types of consonant clusters and stress used for data collection, are discussed in greater detail in chapter 2. The pilot study is being conducted in order to evaluate the effectiveness of these measures for providing meaningful corrective feedback.

The study participants consist of 10 Mandarin accented English speakers. Both undergraduate and graduate student clinicians will be trained in using results of the pronunciation assessment tool to generate and provide accent reduction to the participants using the features returned by the acoustic-to-articulatory inversion system and kinematic templates. Speech data collection and pronunciation feedback for the participants will be performed during multiple sessions over a 6 week period. Meanwhile, a control group of 10 participants will undergo conventional accent modification therapy using acoustic targets while receiving feedback regarding pronunciation accuracy.

Participants will be evaluated based on the amount of pronunciation improvement across sessions, with improvement being measured by how close the participants come to meeting the targets set by the kinematic templates compared to their pre-therapy pronunciation. Additionally, both the participants and clinicians will be surveyed regarding their opinion on the effectiveness of the proposed pronunciation assessment method in providing meaningful corrective feedback for accent modification.

1.3 Research Objectives

The research implemented and presented for this Master's thesis focuses on the development of the kinematic templates needed to support the upcoming pilot study. This includes the following research objectives:

- the extraction and analyzation of the speech data in the EMA-MAE corpus (specifically, a study of the differences in articulation between Mandarin accented English (MAE) speakers and native English (AE) speakers).
- determination of the feasibility of using the EMA-MAE speech data to create kinematic templates that model correct native English pronunciation along the midsagittal plane of the vowels, consonant clusters, and contrastive stresses used in the corpus prompts.
- the implementation of said kinematic templates, designed for use with the Marquette Speech and Signal Processing Lab's acoustic-to-articulatory inversion system for pronunciation assessment.

- the design of visualization plots that display the results of articulatory inversion to each speaker, as well as the relationship between those results and the targets provided by the kinematic templates.

While this research has a focus on the use of acoustic-to-articulatory inversion for pronunciation assessment, a great deal work also went into the study of differences between Mandarin Chinese and English, MAE speech production and the challenges of learning a second language, and the analysis and modeling of both MAE and AE speech. The research discussed in this thesis may be applied to several fields, CALL and CAPT being only a subset of the relevant applications.

1.4 THESIS ORGANIZATION

This thesis is organized into 6 chapters. Chapter 2 covers background information, including a general overview of speech production, differences between Mandarin Chinese and English and their effect on learning English as a native Mandarin Chinese speaker, the EMA-MAE corpus mentioned in section 1.1, and the Marquette Speech and Signal Processing Lab's acoustic-to-articulatory inversion system. Chapter 3 provides an analysis of the speech data contained in the EMA-MAE dataset, specifically in the context of vowel, consonant cluster, and contrastive stress production, and highlighting the differences between native English and Mandarin accented English speech data. Chapter 4 walks through the development of the kinematic templates to be used with the acoustic-to-articulatory inversion system for pronunciation assessment, as well as the design and implementation of visualization plots to be used to provide feedback to pilot study participants. This includes the representation of the results of acoustic-to-articulatory inversion, and those results compared to the targets set by the kinematic templates. Chapter 5 provides a brief summary of the thesis, as well as several future steps for the optimization of the kinematic templates.

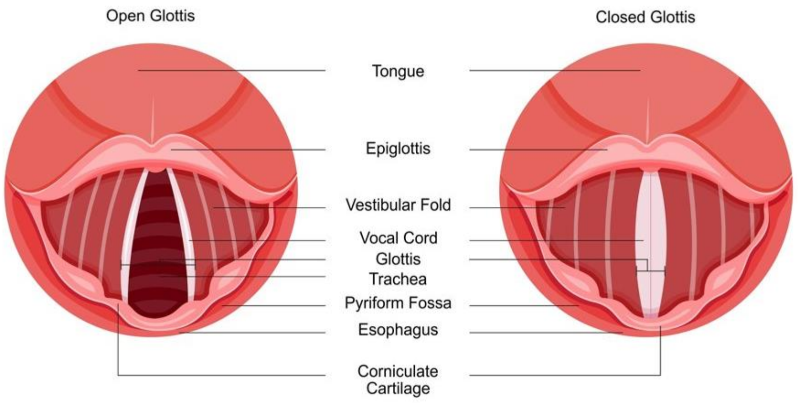
2 BACKGROUND

2.1 SPEECH PRODUCTION AND MANDARIN-ENGLISH OVERVIEW

2.1.1 HUMAN SPEECH PRODUCTION

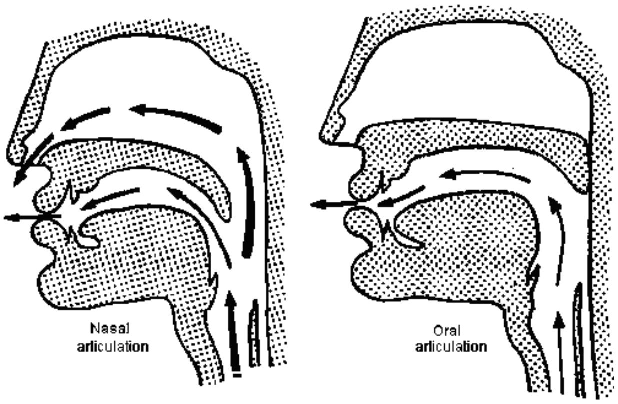
The physiological process of speech production starts with the lungs. The lungs provide a stream of air that passes through the trachea and oral and nasal cavities. This process involves four steps: initiation, phonation, oro-nasal processing, and articulation. Initiation occurs when the air is expelled from the lungs. The phonation step occurs at the larynx, which holds the vocal folds. The gap between the vocal folds is referred to as the *glottis*, and it can be closed, narrowly opened, or widely opened. When the glottis is closed, no air can pass through, meaning no speech can be produced. When the glottis is narrowly opened, the vocal folds vibrate when air passes through. This leads to the production of *voiced* sounds. When the glottis is widely opened, the vibration of the vocal folds is significantly reduced, leading to the production of *unvoiced* sounds. Figure 2.1 shows an example of an open and closed glottis:

Figure 2.1 – *Open and Closed Glottis* **[7]**



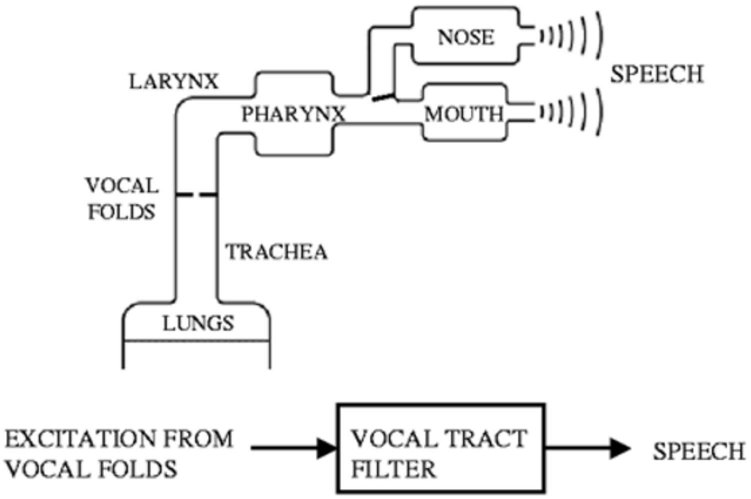
After passing through the larynx the air passes through the oral or nasal cavity, depending on the velum's position:

Figure 2.2 – *Nasal and oral air flow* **[8]**



In the articulation step, the oral cavity acts as a resonator and the articulators (tongue, teeth, lips) are used to determine the speech sound produced [9]. The basic sound unit for a language is known as a *phoneme.* The simplest and most common model of speech production combine all these steps into two elements: excitation and vocal tract filtering. This is known as the *source-filter model* [8]:

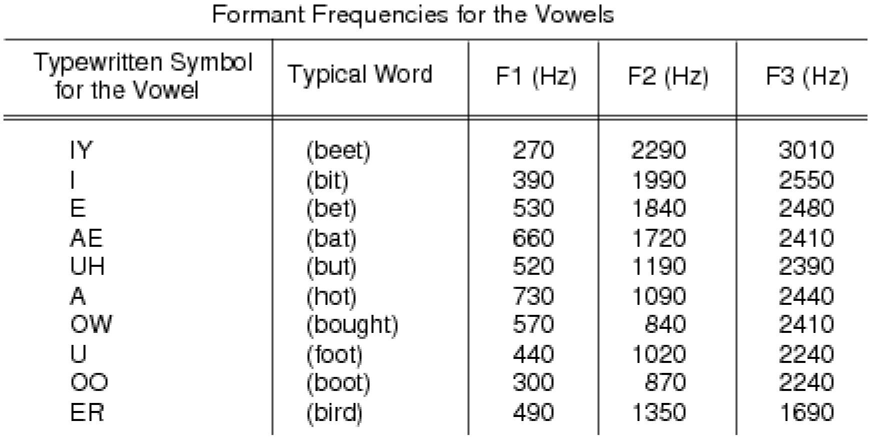
Figure 2.3 – *Source-filter model of speech production*



Resonances in the vocal tract lead to concentrations of acoustic energy at certain frequencies during speech. These resonant frequencies are known as *formants*, and each formant frequency depends on the shape and size of the cross section of the vocal tract during articulation. A larger vocal tract leads to lower formant frequencies (as the size of the resonator is increased) [10]. Adult men typically have larger vocal tracts than adult women, which leads to men having lower formant frequencies than women for the same speech sounds [10]. Formants are denoted as FX, where X is the index of the formant frequencies (i.e. F2 denotes the second formant frequency). Formants frequencies can be estimated through analyzation of the speech signal, typically through linear predictive coding (LPC) analysis [11].

Different speech sounds are formed by changing the shape of the vocal tract during articulation. In other words, the frequency spectrum of speech varies with vocal tract shape. As mentioned previously, formant frequencies are determined by the size and shape of the vocal tract. This means that different speech sounds each have their own sets of formant frequencies. This allows for certain phonemes to be estimated through the analysis of formant frequencies. Vowels are voiced (narrowly opened glottis) phonemes that are produced by maintaining a stationary vocal tract. The formants of vowels are fairly simple to detect in the frequency spectrum. Consonants are both voiced and unvoiced phonemes that are formed through a partial or complete closure of the vocal tract. This closure introduces different amounts of turbulence and anti-resonances into the frequency spectrum [12]. This significantly complicates the process of extracting formants from the speech signal; many consonants' formant frequencies cannot be reliably estimated. As a result, speech processing applications using formants will often focus on vowels. A table displaying average values of the first three formants for English vowels is shown in Table 2.1.

Table 2.1 - *English Vowel Formant Frequencies* **[11]**

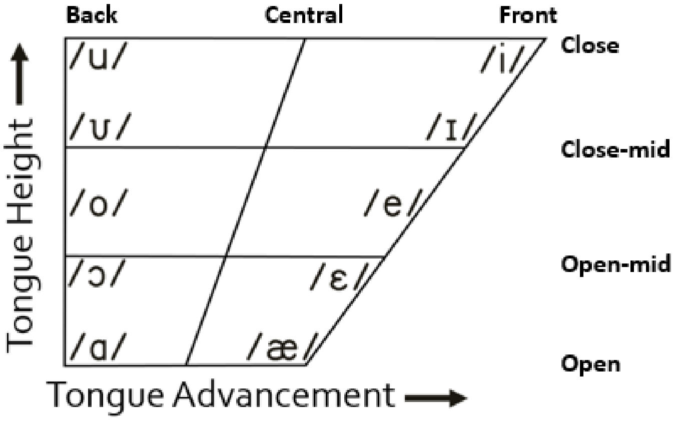


Formant Frequencies for the Vowels

Typewritten Symbol for the Vowel	Typical Word	F1 (Hz)	F2 (Hz)	F3 (Hz)
IY	(beet)	270	2290	3010
I	(bit)	390	1990	2550
E	(bet)	530	1840	2480
AE	(bat)	660	1720	2410
UH	(but)	520	1190	2390
A	(hot)	730	1090	2440
OW	(bought)	570	840	2410
U	(foot)	440	1020	2240
OO	(boot)	300	870	2240
ER	(bird)	490	1350	1690

Depending on the position of the tongue during articulation, the vowel produced may be classified as *open* or *closed* (*low* or *high* tongue position) and *front* or *back*. This lead to the creation of the *vowel quadrilateral* (or *triangle*), which is a diagram that displays the general positions of vowels in the context of these classifications [12]. The application of English vowels, written in International Phonetic Alphabet (IPA) format, to this diagram is shown in Figure 2.4.

Figure 2.4 - *English Vowel Quadrilateral*



Prior research suggests that the first two formants of a vowel are correlated to the speaker's tongue position during articulation. Specifically, F1 is said to be inversely related to the height of the tongue (F1 increases as the tongue body lowers) and F2 is said to be directly related to the anterior positioning of the tongue (F2 increases as the tongue body moves forward) [10]. This information is useful in building expectations for the positioning of a speaker's tongue, given the formants of the vowel being articulated. That being said, it is also generally believed that several different articulatory configurations can lead in the same acoustic result. This attests to the importance of the shape of the entire vocal tract, as opposed to only the articulator positions.

An important aspect of articulation is the concept of coarticulation. This occurs when speakers adjust their articulatory configurations based on preceding and following sounds in order to simplify the overall articulator motion. Coarticulation is generally defined as “the overlapping of adjacent articulations” or as two articulators “moving as the same time for different phonemes” [13]. Coarticulation occurs when different speech production processes (and the articulators involved) combine with different timing patterns [13]. An example of this can be observed when comparing the English words *pit* and *pin*. Both of these words contain the same vowel, but the pronunciation of the vowel (and therefore its formants) are changed towards the end of vowel pronunciation as the articulators begin forming the final consonant. Coarticulation significantly complicates speech processing applications, especially because each type of coarticulation is different depending on the phonetic context.

2.1.2 MANDARIN ACCENTED ENGLISH

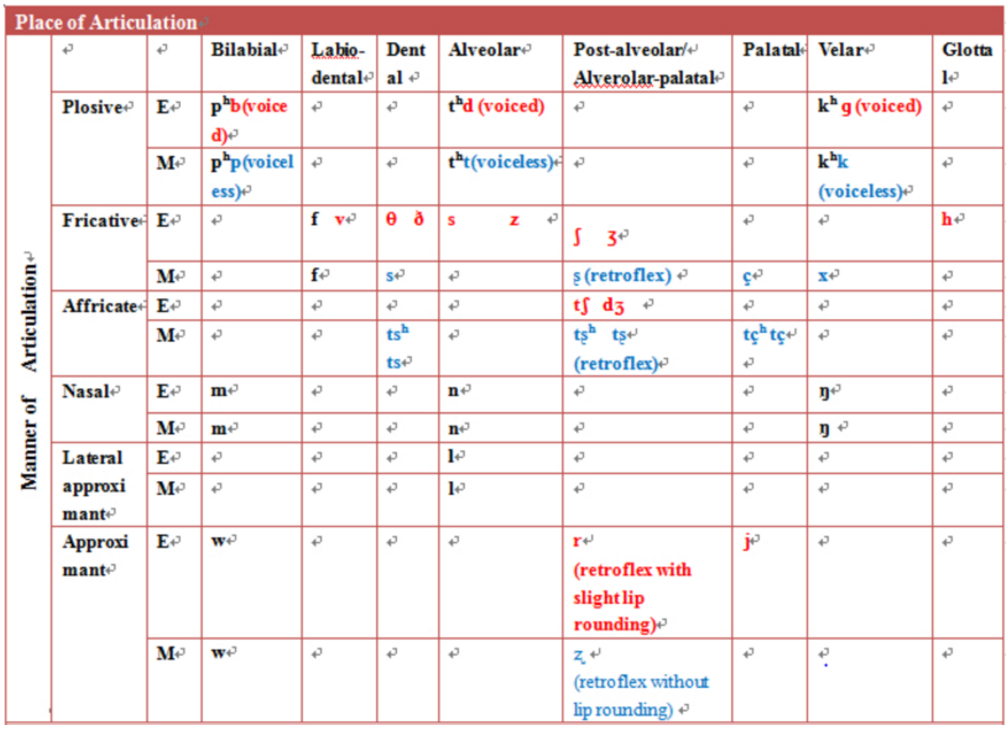
Many sources of difficulty in learning a new language can be traced back to the fundamental differences between the first and second languages (L1 and L2, respectively). Many factors contribute to the degree to which an L1 accent transfers to speech in L2, but the primary effect lies in the sound system of the first language [14]. These effects of L1 are assumed to compete or interfere with the production of L2 [15]. Specifically, prior research suggests that language learners tend to have more difficulty perceiving and producing L2 contrasts that involve non-familiar phonetic features [16]. However, while differences in phonetic context contribute a great deal to inaccurate L2 production, similarities between the two languages can also lead to incorrect pronunciation. Cases of language learners replacing L2 sounds that are similar to a native sounds with the L1 sounds themselves have been documented [17].

In general, phonetic inaccuracy of Asian L1 speakers when speaking English has been well documented [3] [15] [17] [18]. This is especially true for an L1 of Mandarin Chinese and L2 of English. As previously mentioned, many sources of difficulty in learning a new language stems from fundamental differences between L1 and L2. Unlike English, Mandarin Chinese is a tonal language. This means that tone, similar to stress in English, can change the meaning of a word, regardless of phonetic segmentation [18]. Mandarin Chinese has 4 lexical tones: high-level (1), high-rising (2), dipping (3), and high-falling (4). Studies suggest that fundamental frequency (f0) of speech is the primary acoustic indicator of tones in Mandarin [19]. Tones in Mandarin have also been shown to be distinguished by syllable duration, even when f0 information is not present [19]. In English, fundamental frequency and syllable duration (along with intensity and vowel quality) are known to be correlates of stress [20]. Given this fact, one might assume that L1 Mandarin Chinese speaker may apply the same acoustic properties used for tones to produce native English stress. However, prior research indicates that only a subset of Mandarin tones map to English intonation patterns [19]. Additionally, articulation of unstressed vowels in English are typically less prominent, with their formants moving closer to the neutral schwa [21]. This means that the vowels themselves can vary with stress. This fact, in addition to the challenge of actually determining where stress should be placed based on context, make replication of English stress a largely difficult task.

According to [15], the American English vowel system has 11 distinct monophthong vowels: /i, ɪ, e, ɛ, æ, ^, u, ʊ, o, ɔ, ɑ/. Meanwhile, while there are conflicting opinions concerning the exact size of the Mandarin vowel system [22], [15] reports 6 vowels in Mandarin Chinese: /i, e, y, u, o, ɑ/. Given this information, there are at least 5 vowels in Mandarin Chinese with close English equivalents (the vowels given in *beat, bait*, *boot, boat*, and *bot*). This information also indicates that there are several vowels in English that do not have a similar sound in Mandarin Chinese. This presents the opportunity for, as mentioned earlier in this section, an L2 learner to replace a sound that doesn't exist with the most acoustically similar sound in L1 (as opposed to working towards forming a new pronunciation). Additionally, English vowels contain a length contrast that doesn't exist in Mandarin (for example, the difference in length between *ship* and *sheep*) [15]. As a result, Mandarin speakers may not produce or perceive the durational differences present in vowels. Finally, Mandarin Chinese contains several diphthongs and triphthongs (combinations of two and three directly adjacent vowels, respectively), while American English only contains 5.

Table 2.2 displays the consonants of English and Chinese, organized by both manner and location of articulation. *E* represents English, while *M* represents Mandarin Chinese. The consonants highlighted red are those that exist in English, but not Mandarin Chinese. Similarly, the consonants highlighted blue are those that exist in Mandarin Chinese, but not English. According to this table, there are 15 English consonants that do not have a similar sound in Mandarin Chinese.

Table 2.2 – *Mandarin Chinese and English Consonants* **[23]**



Place of Articulation										
			Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar/ Alverolar-palatal	Palatal	Velar	Glottal
Manner of Articulation	Plosive	E	p^hb(voiced)			t^hd (voiced)			k^h g (voiced)	
		M	p^hp(voiceless)			t^ht(voiceless)			k^hk (voiceless)	
	Fricative	E		f v	θ ð	s z	ʃ ʒ			h
		M		f	s		ʂ (retroflex)	ɕ	x	
	Affricate	E					tʃ dʒ			
		M			ts^h ts		$tʂ^h$ tʂ (retroflex)	$tɕ^h$ tɕ		
	Nasal	E	m			n			ŋ	
		M	m			n			ŋ	
	Lateral approximant	E				l				
		M				l				
	Approximant	E	w				r (retroflex with slight lip rounding)	j		
		M	w				ʐ (retroflex without lip rounding)			

The concept of Mandarin accented English speakers replacing English sounds with similar native sounds applies to consonants as well, and even some of the consonants shared by both language cause confusion in English due to the difference in usage across languages. For example, /l/ exists in both languages, but the consonant only appears in the beginning of syllables in Mandarin [23]. This leads to confusion for English syllables containing /l/ in the middle or end, resulting in either the realization of /l/ as its preceding vowel in a syllable (for example, *fool* becomes *foo-o*) or deletion of the consonant altogether [23]. Aside from /l/, final consonants in general also suffer from L1 effects. In Mandarin, phonemes typically end with a vowel sound (with the only exceptions being the front and back nasals/n/ and /ŋ/). Many Mandarin speakers transfer this pattern to English by either removing the final consonant of the English syllable or adding an extraneous vowel to the syllable [23]. One of the most significant differences between consonant usage in English and Mandarin Chinese is voicing contrasts. Mandarin replaces voiced stops with aspiration to indicate stop voicing contrasts (as shown in Table 2.2, /b/, /d/, and /g/ do not exist in Mandarin), and as a result, Mandarin speakers of English tend to have weak voicing for voiced English consonants [23]. The final noteworthy difference in consonant usage between English and Mandarin is the treatment of consonant clusters. Consonant clusters are common occurrences in English, in all possible positions of words. Meanwhile, initial and final clusters do not exist in Mandarin [23]. Mandarin speakers of English tend to either remove the final consonant from the cluster or to create an additional syllable via the attachment of a reduced vowel (such as the neutral schwa) [23].

With these cross-language effects in mind, the EMA-MAE corpus was designed and collected. The study aims specifically to reduce the interference of L1 effects on the production of vowels, consonant clusters, and contrastive stress and to train Mandarin accented English speakers to produce these sounds as native-like (to American English) as possible.

2.2 The Electromagnetic Articulography Database

2.2.1 Background

In recent years, electromagnetic articulography (EMA) has become an important modality for studying the relationship between speech production and acoustics. Applications of EMA include, but are not limited to, speech modeling, recognition, and synthesis [24] [25] [26]. Through EMA, kinematic information about the articulatory organs (lips, teeth, tongue, jaw, etc.) during speech production may be obtained. This information includes position, orientation, speed, and range of motion. This is accomplished by attaching sensors to the articulatory organs and having the subject speak within a small electromagnetic field surrounding their head. The movement of the sensors is tracked in this EM field as the subject speaks. The basic functionality of EMA systems are described in more detail by [27]. While early EMA systems were designed for use in the midsagittal plane, modern systems operate in 3D [27].

Commercially available modern EMA systems include the Carstens AG500/AG501 and the NDI Wave Speech Research System. Both of these systems record both position of their sensors in 3 dimensions and the rotation of the sensors about the transverse axis and anterior-posterior axis. The Carstens AG500 can record data for up to 12 sensors at once at 200 Hz [28]. Meanwhile, the standard NDI Wave unit can track up to 8 sensors at once at 100 Hz, but may be upgraded to sample as many as 16 sensors at 400 Hz [29]. The Marquette University Speech and Swallowing Lab uses the upgraded NDI Wave unit for its EMA applications.

The NDI Wave consists of a data collection unit and a box containing transmitter coils. According to its specifications, the NDI Wave's position tracking is accurate within 0.5 mm. This falls within the target range for meaningful analysis of kinematic speech data [30]. The Wave can be used with 5 or 6 degree of freedom (5 or 6 DOF) sensors, and can be configured to operate using one of two available electromagnetic field sizes: 300 mm^3 or 500 mm^3. The MU Speech and Swallowing Lab collected data in the 300 mm^3 configuration using 5 DOF sensors (with a 6 DOF sensor used for reference).

2.2.2 Database Overview

A number of EMA datasets have been collected and released, coming from a variety of speaker populations for a variety of reasons. The University of Southern California's EMA database was created and shared for the study of expressive speech, with a number of target emotions in mind [31]. [32] describes a database collected for the study of coarticulation across languages. The Marquette University Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus is one the most recently released databases. As discussed in chapter 1, this database was collected in the interest of pronunciation training, with a focus on vowels, consonant clusters, and contrastive stress pairs.

2.2.2.1 Speaker Sample and Database Composition

The EMA-MAE corpus consists of an L1 group of 10 male and 10 female native American English (AE) speakers and an L2 group of 10 male and 10 female native Mandarin Chinese (MAE) speakers. All of the L1 speakers had an upper-midwestern American English dialect. All of the L2 speakers were primary Modern Standard Mandarin speakers, and were evenly divided between Beijing and Shanghai dialects (5 male speakers of Beijing dialect, 5 male speakers of Shanghai dialect, etc.). All speakers fell between the ages of 18-40, and had no history of speech, language, or hearing pathology, no history of orofacial surgery, and no history of medications that would affect motor performance (such as antipsychotics or anti-anxiety medications).

The database contains roughly 30-45 minutes of speech from each subject. The subjects read from text prompts at the word, sentence, and paragraph level. The prompts come from several different sources, with a focus on probable acoustic-phonetic confusions typical of Mandarin-accented English speakers (discussed in section 2.1). The corpus contains acoustic data, kinematic data, phonetic transcriptions, and onset/offset time labels for all speech. This is accompanied by kinematic data required for each speaker's data calibration. This includes individual biteplate records and palate traces.

As described in section 1.2, the study is focuses on three pronunciation error categories: vowels, consonant clusters, and contrastive stress. The vowel and consonant cluster data were collected at the word level, with each word consisting of a C-V-C, C-V-C-V, or C-V-C-V-C format, where C is a consonant or consonant cluster, and V is a vowel. As an example, consider the word *hid*, which is among those used in data collection at the word level. This word consists of a C-V-C format, as "i" is a vowel, while "h" and "d" are consonants. The contrastive stress data was collected at the sentence level. As explained in section 1.2, a contrastive stress is a stress on a syllable or word that is imposed contrary to its typical pronunciation. In the EMA-MAE dataset, all contrastive stress prompts used two-syllable words that, without the differentiation provided by stress, would be identical words. An example of this is *desert* and *dessert*. In *desert*, the first syllable of the word is stressed, while in *dessert*, the second syllable is stressed. Each of these contrastive stress pairs are placed in sentence prompts for each pilot study participant to recite.

The database contains data collected using 8 vowels, 43 consonant clusters, and 9 contrastive stress pairs. The vowels used to create the EMA-MAE corpus are shown in Table 2.3. Note that the vowels that do not exist in Mandarin Chinese have been highlighted.

Table 2.3 – *EMA-MAE Database Vowels*

Vowel ID	IPA	ARPA	Typical Word
1	i	iy	beat
2	ı	ih	bit
3	e	ey	bait
4	æ	æ	bat
5	u	uw	boot
6	ʊ	uh	hood
7	o	ow	boat
8	ɑ	aa	bot

Table 2.4 lists the contrastive stress words used in the EMA-MAE dataset. Note that when a word has alternate spelling based on stress location, both spellings are included in the table.

Table 2.4 – *EMA-MAE Database Contrastive Stress Words*

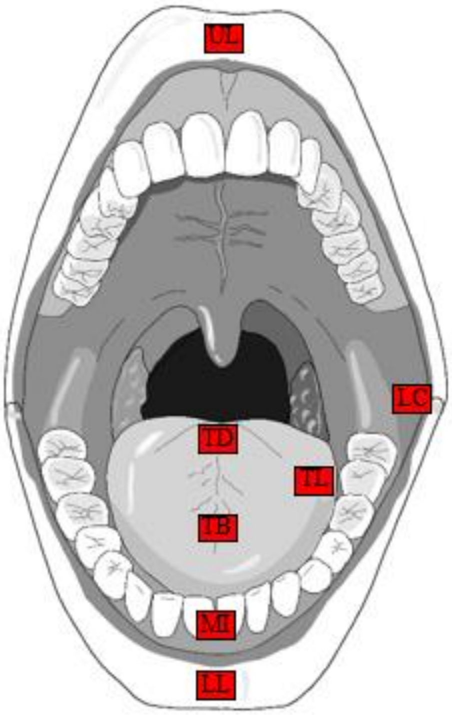
Stress ID	Stress Word
1	contest
2	desert/dessert
3	object
4	perfect
5	produce
6	project
7	rebel
8	record
9	subject

The consonant clusters and contrastive stress prompts are documented in [33].

2.2.2.2 Sensor Layout

Articulatory sensors were placed on the lower lip (LL), upper lip (UL), tongue dorsum (TD), tongue blade (TB), and middle incisors (MI) in the midsagittal plane. There were also two lateral sensors placed at the right corner of the speaker's mouth (LC) and the right central midpoint of the tongue blade (TL). The tongue blade sensor was placed about 1 cm posterior to the tip of the subject's tongue, and the tongue dorsum sensor was placed 3 cm posterior to the tongue blade sensor. This sensor configuration is displayed in Figure 2.5.

Figure 2.5 – *EMA-MAE Sensor Layout (Mouth Diagram:* **[34]***)*



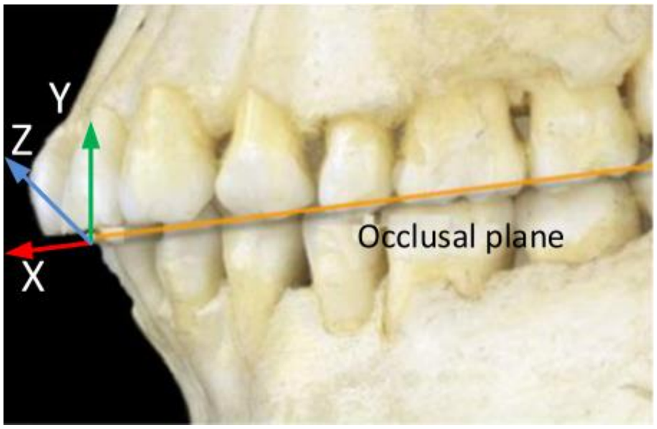
The 6 DOF reference sensor was attached to the midline of a pair of glasses that subjects wore during data collection. This sensor is needed to provide a rigid reference for

implementing the NDI Wave's head correction algorithm, which effectively factors out the speaker's head movement during EMA recording.

2.2.2.3 Speaker Calibration

As discussed in section 2.2.1, the EMA system tracks all movement of the sensors within the magnetic field. This presents an issue, as a speaker could easily move their head within the field while keeping their articulators stationary (relative to their head). To ensure that only movement of the articulators is captured, biteplate calibration is performed on each speaker. A biteplate was constructed for each speaker using dental impression wax. Two sensors were attached to the biteplate: sensor OS, placed one immediately anterior to the central maxillary incisors, and sensor MS, placed about 2-3 cm posterior to the central maxillary incisors. Prior to data collection, the speakers were required to hold the biteplate in their mouth for about 15 seconds in order to determine the spatial relations between the biteplate sensors and reference sensor. Biteplate correction (described in detail by [5]) was used to define the midsagittal and maxillary occlusal planes, and to define the local coordinate origin at the tips of the central maxillary incisors. After this correction, the coordinate system is as shown in Figure 2.6.

Figure 2.6 – *EMA-MAE Dataset Coordinate System* **[5]**



After the palate trace collection, the speaker read a specified paragraph to allow for them to adjust to speaking with EMA sensors placed on their articulators. The data obtained from this read-through was also used to adjust microphone levels and assess sensor adhesion.

2.2.2.4 Palate Trace

Section 2.1.1 discussed the importance of the vocal tract shape during speech production. While the tongue and lip sensors provide a great deal of information themselves, the palate forms the upper limit of the vocal tract at the tongue's location. It is crucial to extract a speaker's palate to form a detailed image of the speaker's vocal tract. The palate trace record was obtained using a probe with a 5 DOF sensor attached to the end. The wand was swept across each speaker's palate, both laterally from left to right, and along the midsagittal plane toward the uvula.

Upon inspection of the kinematic data for each of the subjects, it was noted that in many cases, the coordinates of the palate traces indicated that they were located under the tongue sensors ($y_{palate} < y_{tongue}$). This is physically impossible, and it is unknown how the palate traces came to be distorted in this way. As a result of this issue, the original palate traces were discarded, and the tongue sensor data and dental perimeter for each speaker were used to recreate the palate traces.

The palate recreation was completed using the convex hulls surrounding the position data given by the three tongue sensors (TD – tongue dorsum, TB – tongue blade, and TL – tongue lateral). Figure 2.7 demonstrates an example of the process. To start, all of the positional speech data for a given speaker was aggregated into a single set of data. The midsagittal palate trace was obtained by first limiting the data to the points that fall into the range of $-2.5 < z < 2.5$ mm (Figure 2.7-3), then calculating the convex hull of this new data (Figure 2.7-1). The upper section outline of this convex hull is a rough estimate of the new midsagittal palate trace (Figure 2.7-2). The data was then divided into 100 “slices” in the Y-Z plane along the midsagittal line, and for each slice, the convex hull of the data was calculated (Figure 2.7-4). As previously discussed, the tongue lateral sensor is located on the left side of the speaker’s tongue. To obtain a full convex hull (encompassing the left side of the mouth as well), this data was reflected across the midsagittal plane (Figure 2.7-5). The convex hull was edited to include the dental perimeter. With the general shape of the new palate trace established, Bezier smoothing [35] and upward shifting (to correct for downward shift cause by the shifting) in the X-Y plane were performed to form the general shape of the new palate (Figure 2.7-6).

Figure 2.7 – *Palate Trace Generation Process*

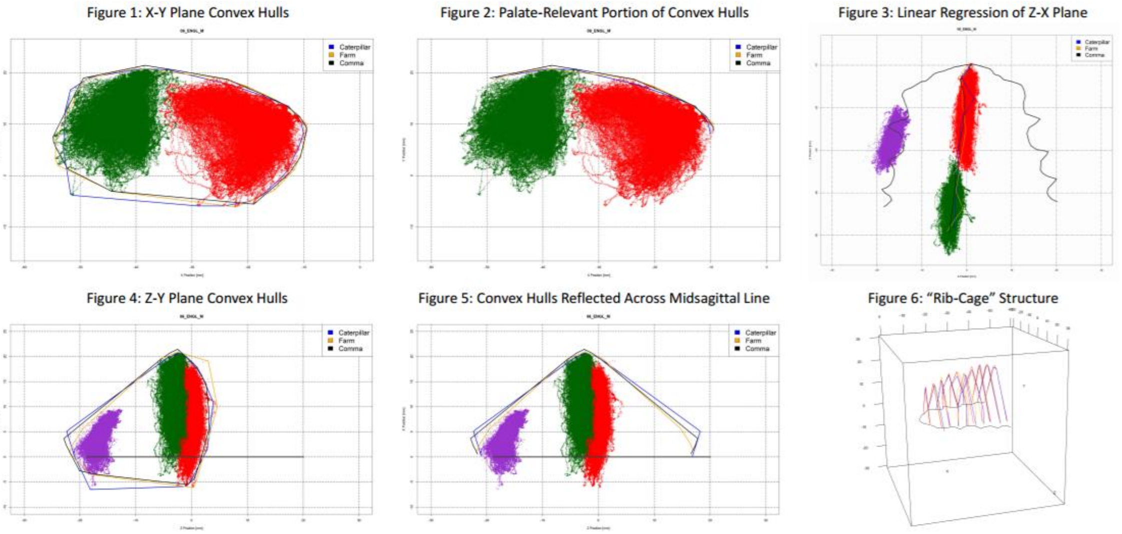


Figure 1: X-Y Plane Convex Hulls

Figure 2: Palate-Relevant Portion of Convex Hulls

Figure 3: Linear Regression of Z-X Plane

Figure 4: Z-Y Plane Convex Hulls

Figure 5: Convex Hulls Reflected Across Midsagittal Line

Figure 6: "Rib-Cage" Structure

2.2.2.5 Acoustic and Kinematic Data Format

As previously mentioned, speech samples were taken from each speaker at the word, sentence, and paragraph level. The acoustic and kinematic data were also stored accordingly. Each set of words, sentences, and paragraphs has a corresponding audio and kinematic data file.

Speech audio was collected using a cardioid pattern directional condenser microphone, placed 1 m from the center of the EM field generated by the EMA equipment. The audio was recorded in .wav format. Using linear predictive coding (LPC) analysis [11], the 1st-3rd formants of each speech sample were calculated. The kinematic data is organized as large table, with the different columns corresponding to each sensor's status, ID, positional dimension (x, y, z) value, and rotational orientation (q0, q1, q2, q3) value, as well as the time stamps (in seconds) during data collection. Full details of the methodology and data format are available in the EMA-MAE user manual [33].

Each kinematic and formant data file is accompanied by a label file that lists the onset and offset times of relevant acoustic occurrences. These labels were marked and index by trained students of the Marquette University Speech and Swallowing lab through the observation of speech audio signals and spectrographs of those signals. For vowels, which are collected at the word level, the onset and offset time of each vowel recited during recording is listed along with a vowel ID number. For consonant clusters, the onset and offset time of each consonant cluster occurrence during recording is listed along with a cluster ID number. For contrastive stress pairs, the onset and offset time of each of the two syllables of each word is listed, along with a stress ID number. These labeling files are used for the extraction of the vowels, clusters, and stress pairs used to generate templates.

2.2.3 Known Issues and Concerns

2.2.3.1 Missing Sensor Data

When a sensor is not within the NDI Wave sensor range, no position or orientation data is recorded for that sensor. This also occurs when the sensor is in range, but has too weak a signal for the NDI Wave to accurately capture its location. On average, less than 1% of sensor data is missing, but it is not evenly distributed throughout the database [33].

2.2.3.2 Unreliable Sensor Attachment

Some well-known disadvantages of using EMA for articulator motion tracking are the associated sensor placement issues. By placing sensors in or around a subject’s mouth, the chance of sensors breaking, detaching, or becoming misaligned increases. While the MU Speech and Swallowing lab did not encounter the issue of sensors breaking, detached and misaligned sensors are a very real concern. In some cases, the tongue sensors would end up closer to each other than anticipated, due to stretching and compression of the tongue. In other cases, a sensor may become completely detached from the tongue. Unless the speaker were to notify the experimenter, this issue could go undetected. These issues lead to misrepresentation of articulator locations, and therefore a corruption of the kinematic data. While these cases are likely limited and ideally singled out due to inconsistency with the rest of a given speaker's data, the issue is present and could have significant effects on the analysis of kinematic data.

2.2.3.3 Unreliable Formant Data

While the speech audio for the EMA-MAE dataset was collected in a sound-attenuating acoustic booth, noise was not completed removed from the system. The primary source of noise was interference between the EMA sensors and the microphone used for audio collection. In general, formants F3 and F4 are very difficult to capture reliably. The use of unreliable data threatens to corrupt the data analysis and template development process. Therefore, all formant operations and analyses in this thesis were performed using only formants F1 and F2. While this leads to a loss of information, it prevents potential sources of error in formant analysis.

2.3 The Acoustic-to-Articulatory Inversion System

2.3.1 Background

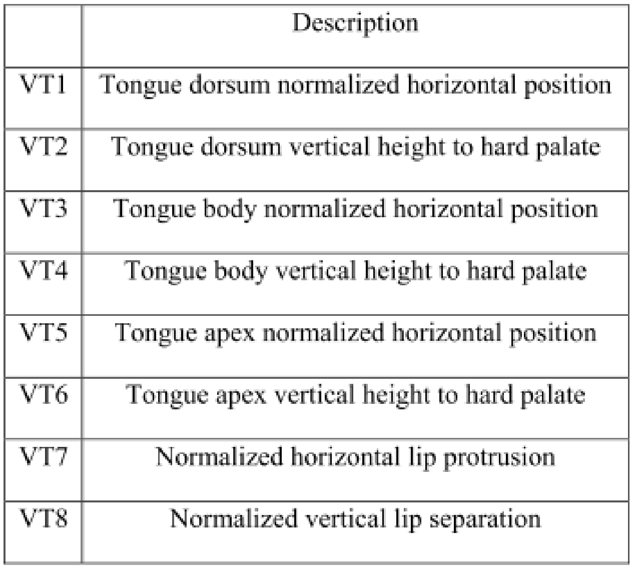
Acoustic-to-articulatory inversion is the estimation of a speaker's articulatory configuration from their speech data. Using the acoustic signal as input, the inversion system produces the articulatory positioning and movement. Acoustic-to-articulatory inversion has a variety of speech processing applications, including speech coding, automatic speech recognition (ASR), computer aided language learning (CALL), and computer aided pronunciation training (CAPT) [36] [37] [38]. There have been several successful implementations of speaker-dependent acoustic to articulatory inversion [39] [40], but most of these implementations must be trained on simultaneous acoustic and kinematic data from participating speakers. The Marquette University Speech Lab's acoustic-to-articulatory inversion system performs speaker-independent speech inversions without the use of a speaker's kinematic data, and uses the 20 native English speakers from the Speech and Swallowing Lab's EMA-MAE database (discussed in section 2.2) for training and analysis.

2.3.2 Articulatory Features

While sensor position data provides a simple representation of articulator motion, there are a number of reasons that it may not be the optimal representation for use in acoustic-to-articulatory inversion. Among these is the fact that raw sensor position provides barely any information about the shape of the vocal tract during speech production. As discussed in section 2.1, the acoustics of speech are 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. Also, there are times when information may be represented in a simpler format. For example, when concerned with lip separation, it would be much simpler to express the value as the difference in lip heights (UL_y-LL_y) than storing 3 dimensions of two different sensors (UL and LL).

The MU Speech Lab defined a conversion of EMA kinematic data to vocal tract parameters in order to represent the speaker's articulatory configurations in a more meaningful format. Table 2.5 lists these features.

Table 2.5 – *Acoustic-to-Articulatory Inversion System Features* **[5]**



	Description
VT1	Tongue dorsum normalized horizontal position
VT2	Tongue dorsum vertical height to hard palate
VT3	Tongue body normalized horizontal position
VT4	Tongue body vertical height to hard palate
VT5	Tongue apex normalized horizontal position
VT6	Tongue apex vertical height to hard palate
VT7	Normalized horizontal lip protrusion
VT8	Normalized vertical lip separation

Humans have unique vocal tract sizes and shapes, especially across gender (see section 2.1.1 for a discussion on this topic). Without horizontal normalization, the sensor x positions are relative to a given speaker and are therefore meaningless when comparing to or modeling a different speaker or speaker group. The distance from the central incisors to the middle of the back molar was used as a horizontal normalization scalar for each speaker. This distance is thought to be related to the speaker's vocal tract length, and is introduced to the system to reduce cross-speaker horizontal variance. The horizontal features (those along the X-axis), VT1, VT3, VT5, and VT7, were each calculated directly from the corresponding sensor position divided by this scalar. The vertical features (those along the Y-axis), VT2, VT4, and VT6, were calculated directly as the vertical distance between the sensor position and the palate height at the same x position. VT8 was calculated as the vertical lip separation, rescaled to a [0,1] working space. Equations (2.1)-(2.8) show the exact calculations required to convert sensor positions to articulatory features.

$$VT1 = \frac{TD_x}{H} \tag{2.1}$$

$$VT2 = P(x,z) - TD_y \tag{2.2}$$

$$VT3 = \frac{TL_x}{H} \tag{2.3}$$

$$VT4 = P(x,z) - TL_y \tag{2.4}$$

$$VT5 = \frac{TB_x}{H} \tag{2.5}$$

$$VT6 = P(x,z) - TB_y \tag{2.6}$$

$$VT5 = \frac{UL_x}{H} \tag{2.7}$$

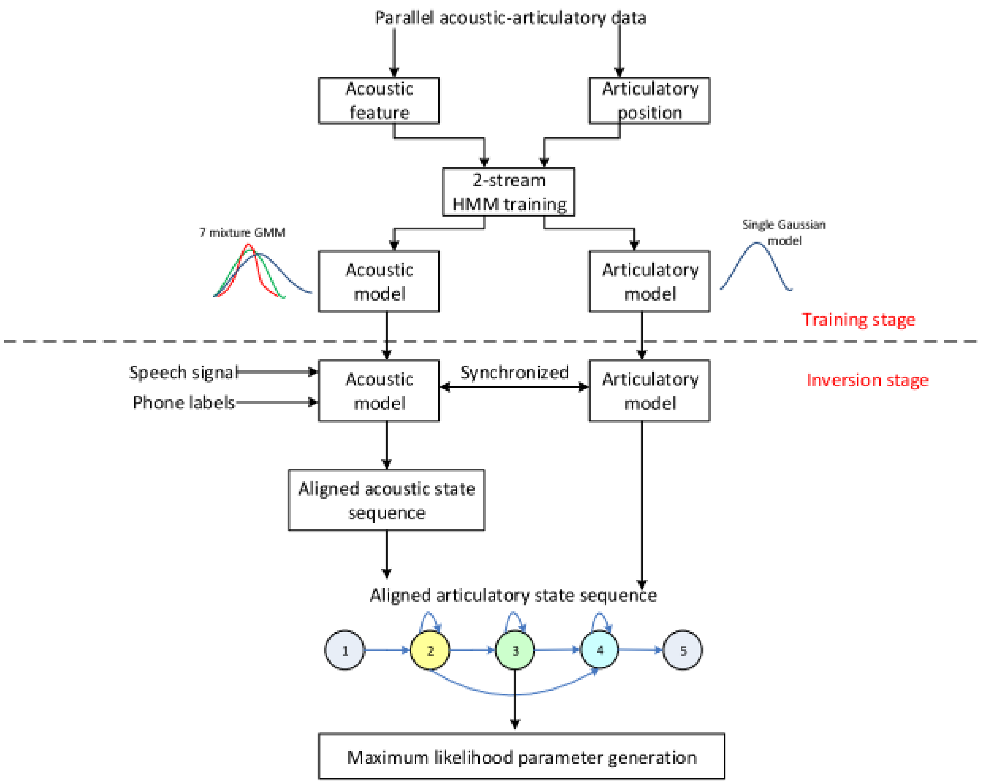
$$VT8 = \frac{(UL_y - LL_y) - (UL_y - LL_y)_{min}}{(UL_y - LL_y)_{max}} \tag{2.8}$$

P(x,z) represents the speaker's palate's y location corresponding to the x and z locations of the sensor being converted. For example, P(x,y) in equation **(2.2)** is the y value of the palate at the x and z locations of the TD sensor. H represents the horizontal normalization scalar. Note that in order to avoid outliers and measurement error, the minimum and maximum lip separation values were recorded as the 5th and 95th percentiles of all vertical lip distance measurements for each speaker.

2.3.3 Hidden Markov Model Based Inversion System

The Speech Lab acoustic-to-articulatory inversion system starts with a hidden Markov model (HMM) based inversion system. Parallel acoustic and articulatory data are used to train the acoustic and articulatory HMMs separately, and the HMMs are aligned by state sequences for each phonetic unit. During inversion, the speech signal is input to the acoustic HMM to derive an optimal HMM state sequence via the Viterbi algorithm [11]. The corresponding aligned articulatory HMMs are used to recover the articulatory motion. This is described in further detail by Ji [5]. Figure 2.8 displays a diagram of an HMM based inversion system.

Figure 2.8 – *HMM-Based Acoustic-to-Articulatory Inversion System* **[5]**



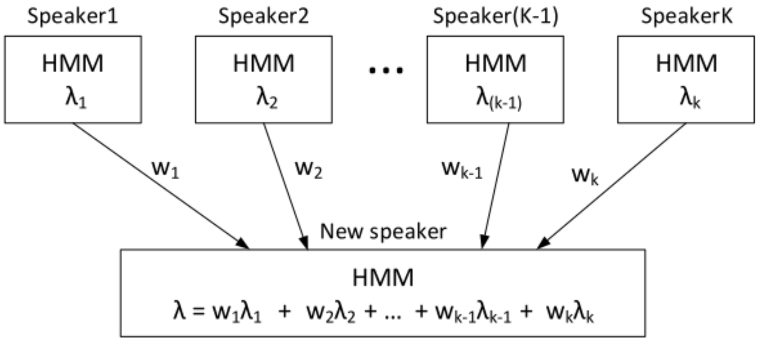
2.3.4 Speaker Adaptation

The baseline HMM based acoustic-to-articulatory inversion system is a speaker dependent system, meaning that parallel acoustic and articulatory training is implemented on data from a single subject. This acoustic-articulatory mapping varies from subject to subject, so this inversion method is unlikely to perform well without articulatory data from the target speaker. This is problematic, since many of the most important applications of acoustic-to-articulatory inversion would necessarily require inversion on subjects for whom no articulatory data is available to use for training. The HMM based inversion system was extended using the idea of speaker adaptation to create a new system that can accomplish inversion on a new speaker using only a small amount of acoustic data and no kinematic data [5].

2.3.4.1 Reference Speaker Weighting

Reference speaker weighting (RSW) is a rapid speaker adaptation approach that implements adaptation using about 5-10 seconds of speech [41]. RSW uses speaker-dependent models as a starting point towards estimation of the parameters of a new speaker. Specifically, RSW creates a model of a new speaker as a weighted combination of reference speakers, and the weights are determined using the adaptation data. A diagram describing the basic idea of RSW is shown in Figure 2.9.

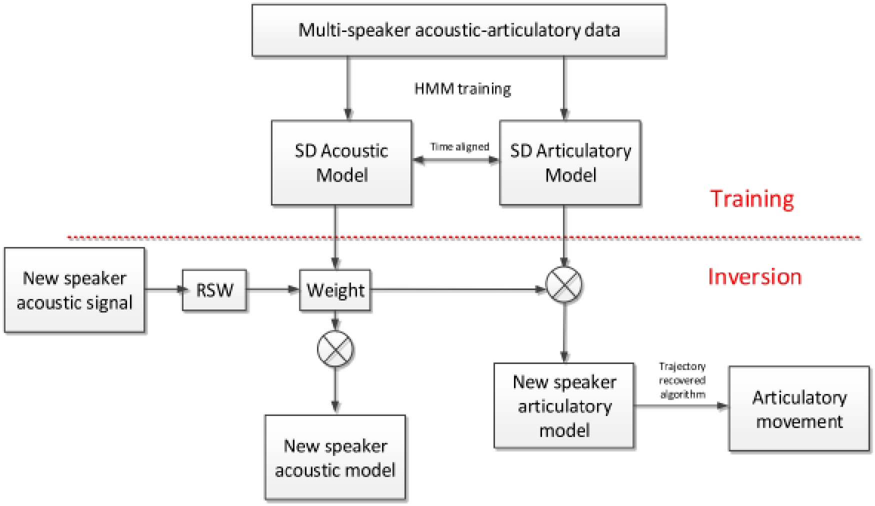
Figure 2.9 – *Reference Speaker Weighting* **[5]**



2.3.4.2 Parallel Reference Speaker Weighting

Parallel reference speaker weighting (PRSW) extends RSW used in the acoustic domain to estimate articulatory parameters. RSW is performed using the new speaker’s speech signal, and the derived weights are used in the articulatory domain during inversion. A diagram showing the operation of PRSW is shown in Figure 2.10.

Figure 2.10 – *Parallel Reference Speaker Weighting* **[5]**


Multi-speaker acoustic-articulatory data
HMM training
SD Acoustic Model
Time aligned
SD Articulatory Model
Training
Inversion
New speaker acoustic signal
RSW
Weight
New speaker acoustic model
New speaker articulatory model
Trajectory recovered algorithm
Articulatory movement

Note that PRSW assumes that the speaker combination used in the acoustic domain matches that of the articulatory domain. A detailed explanation of the implementation of PRSW can be found at [5].

During initial experiments, 13 of the 20 native English speakers' inversion results were shown to be superior to the RSW-based speaker independent model, and very close to the HMM based speaker-dependent model. Further experiments showed that implementations that used a subset of reference speakers based on acoustic model similarity, as well as implementations that used a subset of reference speakers based on speaker-dependent inversion performance both performed better than the baseline RSW system. This means that PRSW is capable of recovering a good articulatory configuration for a target speaker, provided that the set of reference speakers are selected according to acoustic and articulatory consistency. The results of the analyses are described in detail by Ji [5].

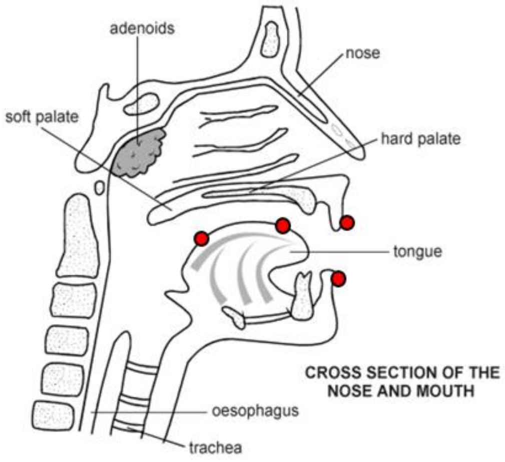
3 DATA EXTRACTION AND ANALYSIS

As explained in section 2.2, all kinematic and formant data is accompanied by labeling files that allow for the vowels, consonant clusters, and contrastive stress data to be extracted. The analysis in this thesis focuses on the midsagittal plane, so all kinematic data was extracted in two dimensions (x and y, which define the midsagittal plane, as shown in Figure 2.6). Each phonetic category has its own acoustic and articulatory characteristics, and were therefore extracted and analyzed in different ways. This chapter explains in detail how each of the three categories were handled.

3.1 RELEVANT ARTICULATORS AND SENSORS

As discussed in chapter 1, the kinematic templates designed for pronunciation training aim to model the movement and positioning of articulators in the midsagittal plane. Specifically, this refers to the articulators that have sensors along the midsagittal plane: tongue dorsum (TD), tongue blade (TB), upper lip (UL), and lower lip (LL). While there are other sensors that provide additional information, this research only uses these four midsagittal sensors. Figure 3.1 displays a midsagittal view of the articulators and the placement of the sensors (displayed as red circles) being used for template formation.

Figure 3.1 – *EMA Sensors Used For Template Creation*



For all phonetic categories (vowels, consonant clusters, and contrastive stress), these data from these four sensors was extracted for all analyses. In addition to information from the EMA sensors, the palate trace of the speaker is also crucial for analyzing and characterizing the speech data. As discussed in chapter 2, the shape of the vocal tract's cross section plays a large role in determining the sound produced during articulation. The palate forms the ceiling of the vocal tract during oral articulation, and therefore provides important information about the vocal tract shape.

3.2 Conversion to "Feature Space"

As discussed in section 2.3, the acoustic-to-articulatory inversion system returns the estimated articulatory parameters as a set of palate-referenced features. Table 2.5 displayed these features, as well as how they are defined. In order to compare the results of the acoustic-to-articulatory inversion to the kinematic templates, the two forms of data must be presented in the same format. To complete this requirement, all extracted kinematic data from the EMA-MAE corpus is re-referenced in the format of articulatory features using the conversion methods described in section 2.3.2. This conversion places all extracted EMA data in *feature space*. Only a subset of features from Table 2.5 are used for data extraction and analysis, as this analysis is only applied to the midsagittal plane. Table 3.1 lists the features specifically used for data analysis and template creation.

Table 3.1 – *Articulatory Features Used for Analysis and Template Creation*

Feature	Feature Description
VT1	Tongue Dorsum (TD) Normalized Horizontal Position
VT2	Tongue Dorsum (TD) Vertical Height to Hard Palate
VT5	Tongue Apex (TB) Normalized Horizontal Position
VT6	Tongue Apex (TB) Vertical Height to Hard Palate
VT7	Normalized Horizontal Lip (UL) Protrusion
VT8	Normalized Vertical Lip (UL,LL) Separation

As Table 3.1 displays, two of the features (VT2 and VT6) are referenced to the palate. This means that the palate traces of the speakers (described in section 2.2.2.4) are also needed to perform the conversion to feature space. Before extracting any speech data for a speaker, their palate data is extracted. The palate referenced tongue sensors are located along the midsagittal plane, so only the midline trace of the palate is needed. According to the coordinate system defined in Figure 2.6, the midline palate trace is located at z=0. For each speaker's palate trace, all data points located at $-0.3 < z < 0.3$ were extracted and resampled such that all speakers have the same number of x-y data points in their palate trace (arbitrarily chosen to be 200 points). Having the same sized palate trace for each speaker allowed for ease of calculating an average palate trace, which was used for template visualizations (see section 4.2.3 for details).

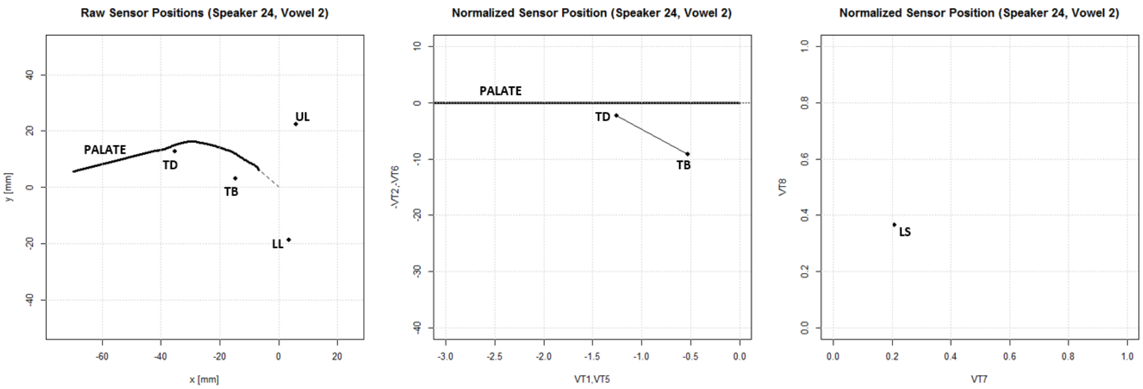
After the data is converted to feature space, the lips are expressed in terms of protrusion and separation. UL and LL become the two dimensional *LS*, whose x is normalized lip protrusion (taken from the x position of UL) and whose y is normalized lip separation (taken from the vertical distance between UL and LL, in accordance with equation (2.8)). Table 3.2 shows the relationship between the EMA sensor dimensions and the articulatory features of the inversion system.

Table 3.2 – *Relationship Between Sensors and Articulatory Features*

Sensor Dimension	**Feature Value**
TDx	VT1
TDy	-VT2
TBx	VT5
TBy	-VT6
LSx	VT7
LSy	VT8

Figure 3.2, which displays the sensor positions of a speaker for a given point in time, demonstrates the conversion from Euclidean space to feature space. The left plot shows the original data as extracted from the EMA-MAE corpus files, the middle plot shows the tongue sensors in feature space, and the right plot shows the aggregate LS sensor in feature space.

Figure 3.2 - *Conversion to Feature Space*



Note that through palate-referencing and expressing tongue heights as negative values, the palate's location is moved to y=0. Recall that the LS features use a [0,1] normalization. Given that this sensor uses a different scale than the tongue sensors, it needed to be plotted separately.

Note that while data must be presented in feature space for use with the acoustic-to-articulatory inversion system, the data analysis also benefits from this conversion (for the same reasons described in section 2.3.2). Feature space provides both a more compact and more intuitive data representation in the context of data analysis and comparison, and its normalization accounts for cross-speaker variability that threatens to corrupt any comparisons across speakers or speaker groups.

3.3 Statistical Analysis Overview

The wide variety of acoustic and articulatory parameters that factor into speech production (discussed in chapter 2), coupled with cross speaker variability, introduces a degree of unknown into the problem of analyzing and modeling those articulatory parameters. In order to account for this unpredictability, a probabilistic analysis is introduced to the study of speech data. Specifically, Student's t-tests and analysis of variance are used to assess the differences (or lack thereof) in articulation within and between speaker groups.

3.3.1 Student's T-Test

A *Student's t-test* (or *independent samples t-test*) is used to compare two means and determine if they are different from each other, as well as the significance of the difference between groups [42]. Specifically, Student's t-tests are used to determine if two means are different due to significant differences between the groups, or just by chance (in other words, whether the results can be re-produced through several repeated trials). While there are many variations of the t-test that differ based on the state of the sample data, t-tests are, in general, used for comparing groups of data with small sample sizes (typically less than 30) [42]. This work uses *Welch's t-test*, which should be used when the two populations being compared are not assumed to have equal sizes or variances [42].

In t-tests, the differences between groups is determined using the *t-value* (or *t-score*) [43]:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.1}$$

where x_i is the set of data values for group i, s_i is the unbiased estimator of the variance for group i, and n_i is the sample size of group i. The t-score is evaluated in order to determine whether the groups are significantly different using the Student's t-distribution, which is a family of curves in which the number of *degrees of freedom* determines the particular curve used for a calculation. In a normal Student's t-test (where the sample sizes and variances of the two groups are assumed to be equal), the number of degrees (DOF) is equal to one less than the sample size of either group [44]:

$$df_i = v_i = n_i - 1 \tag{3.2}$$

where df$_i$ (or v_i) is the number of DOF for group i. The degrees of freedom are indicative of the number of independent pieces of information in a sequence of numbers. Note that in order to perform these calculations, the means of the two groups must be known. For this reason, the sample size is decremented by 1 to form the number of DOF (because while N-1 pieces of information are allowed to vary freely, the mean is known). In a Welch's t-test, the number degrees of freedom is estimated using the Welch-Satterthwaite equation [43]:

$$v \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 v_1} + \frac{s_2^4}{n_2^2 v_2}} \tag{3.3}$$

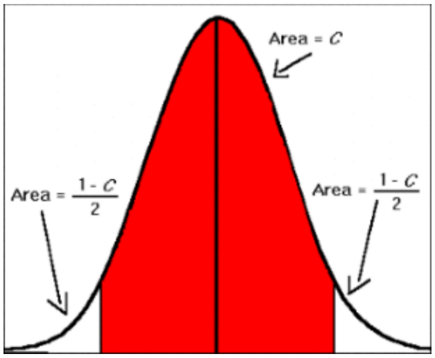
where v (or *df*) is the overall number of DOF, v$_i$ is the number of DOF for group i, and s$_i$ and N$_i$ are as defined in equation (3.1). The probability distribution function of the t-distribution is as follows [43]:

$$f_v(t) = \frac{\Gamma\left[\frac{1}{2}(v+1)\right]}{\sqrt{v\pi}\,\Gamma(\frac{1}{2}v)(1+\frac{t^2}{v})^{\frac{v+1}{2}}} \tag{3.4}$$

where Γ is the gamma function.

In order to form a metric for determining statistical significance, an *alpha value* (or *significance level*) must be chosen [45]. In hypothesis testing, the alpha level is the probability of rejecting the *null hypothesis* (the assumption that the means of the two groups are equal) when it is true [42]. In other words, the alpha level is the probability of a false positive. In choosing an alpha value, one also chooses the *confidence level*. The *confidence interval* of a set of sample data is a range of values that is likely to contain some population parameter (often, the mean) [45]. The probability that a confidence interval will contain the population parameter is known as the confidence level (C) [45]. The confidence interval is determined through the selection of C, and the confidence level corresponds to the percentage of the area under the probability distribution function of the normal distribution:

Figure 3.3 – *Confidence Level Demonstration* **[46]**



Note that the concept of confidence intervals and levels is founded on the Central Limit Theorem (CLT), which states that the distribution of a large sample size (typically, more than 30 samples) of a population tends to approach a normal distribution [46]. In confidence analysis, the sample data is treated as if it is normally distributed. For a given

C value, the probability of observing a value outside of the area under the curve is the alpha value:

$$\alpha = 1 - C \tag{3.5}$$

where α is the alpha level and C is the confidence level. A commonly used alpha level is 0.05, or 5% (corresponding to a confidence level of 0.95 or 95%) [42]. To determine whether there are significant differences between groups, the t-value must be compared against the *critical value* (c). For t-test operations, the critical value is obtained using a t-distribution table, which lists the critical value given the number of DOF and alpha level [43].

After calculating the t value and obtaining the critical value, the t-distribution may be used to calculate a *p-value* [42]:

$$p = \Pr(T > c) \tag{3.6}$$

where T is any given f(t) returned from the t-distribution of equation (3.4), and p is the p-value. When the p-value is less than the alpha value, the null hypothesis may be rejected [42]. In other words, a lower p-value than alpha value indicates that the means of the two groups being compared are indeed significantly different. One should note that this is only a convention, and that a low p value does not guarantee that the groups are different; it just confirms that they are very likely to be different.

3.3.2 Analysis of Variance

Analysis of variance (ANOVA), like t-tests, is a method of comparing the means of a variable across different groups. The main difference between ANOVA and t-tests is that while t-tests are used for comparisons of two groups, ANOVA is used for comparisons of 3 or more groups. Similar to t-tests, there are several variations of ANOVA. This study makes use of one-way (or one-factor) ANOVA, which compares 3 or more groups (or levels) for a single independent variable. For an ANOVA, the null hypothesis is that the means of all groups involved are equal.

In ANOVA, the t-value and t-distribution of t-tests are replaced with the F-value and F-distribution. The F-value is described as the ratio of variance between groups (*effect* or *treatment* variance) to the variance within groups (*error* variance) [44]:

$$F = \frac{Effect/Treatment\ Variance}{Error\ Variance} = \frac{Var_{between}}{Var_{within}} \tag{3.7}$$

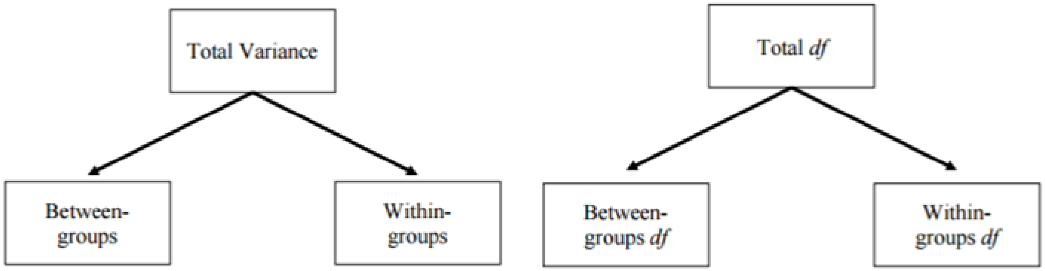
After observing this ratio, it is clear that the F value is a measure of the differences between different groups compared to the amount of variation within each group. By accounting for the individual speaker group variation (error), the analysis can focus on the differences between each group. Similar to the t-distribution, the F-distribution is a family of distributions in which the number of degrees of freedom determine which distribution is used. The probability density function of the F-distribution is as follows [44]:

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\, x)^{d_1}\ d_2^{d_2}}{(d_1\, x + d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \tag{3.8}$$

where d_1 is number of DOF between groups, d_2 is the number of DOF within groups, and B is the beta function.

In ANOVA, each group (or level) has its own number of degrees of freedom, following the same definition as equation (3.2). In ANOVA, the total DOF is made up of the DOF between groups and the DOF within groups. The same is true for the total variance:

Figure 3.4 – *ANOVA Variance and DOF Structure* **[44]**



Note that variance may be expressed as the ratio of the sum of squares (SS) to the number of degrees of freedom:

$$Var = \frac{SS}{df} = \frac{\sum(x - \bar{x})}{n - 1} = \frac{\sum(x^2) - \frac{(\sum x)^2}{n}}{n - 1} \tag{3.9}$$

To determine the variance estimate, the sum of squares and number of degrees of freedom must first be determined, both within and between groups. The procedure, as described by [44] is as follows:

For ease of calculations, define a *correction factor* (CF):

$$CF = \frac{(Sum\ of\ all\ x)^2}{Total\ length\ of\ all\ x} = \frac{\left(\sum_{i=1}^{l} \sum x_i\right)^2}{\sum_{i=1}^{l} n_i} \tag{3.10}$$

where x_i is the sequence of values for group i, n_i is the length of x_i, and l is the number of groups. Given the definition of the sum of sum of squares in equation (3.9), the correction factor may be used to calculate the total sum of squares:

$$SS_{total} = (Sum\ of\ all\ x)^2 - CF = \left(\sum_{i=1}^{l} \sum x_i\right)^2 - CF \tag{3.11}$$

Next, calculate the SS between groups, and use that to calculate the SS within groups:

$$SS_{between} = \sum_{i=1}^{l} \frac{\left(\sum x_i\right)^2}{n_i} - CF \tag{3.12}$$

$$SS_{within} = SS_{total} - SS_{between} \tag{3.13}$$

The total number of DOF is defined as one minus the total length of all data, and the number of DOF is defined as one minus the number of groups. These may be used to calculate the number of DOF within groups:

$$df_{total} = \sum_{i=1}^{l} n_i - 1 \tag{3.14}$$

$$df_{between} = l - 1 \tag{3.15}$$

$$df_{within} = df_{total} - df_{between} \tag{3.16}$$

With the between and within SS and df values calculated, the variance estimates may be calculated using equation (3.9):

$$Var_{between} = \frac{SS_{between}}{df_{between}} \tag{3.17}$$

$$Var_{within} = \frac{SS_{within}}{df_{within}} \tag{3.18}$$

Finally, the F value may be calculated using equation (3.7). Obtaining the p-value from the F-distribution is identical to the process of obtaining p-value from the t-distribution in a t-test (described in section 3.3.1). After choosing an alpha value, an F-distribution table is used to determine the critical value, and the p-value calculation becomes a slight variation of equation (3.6):

$$p = \Pr(F > c) \tag{3.19}$$

where F is any given f(t) returned from the F-distribution of equation (3.8), and c is the critical value obtained from the F-distribution table.

When the p-value is less than the alpha value, the null hypothesis is rejected, meaning all of the groups are not the same. Note that this does not mean that all groups are significantly different from each other, but instead that at least two of the total number of groups are different from each other. To determine which two groups are different from each other, t-tests must be performed on the groups. If the statistical relationship between every set of two groups is desired, the ANOVA may be bypassed, and t-tests may be performed immediately. Performing several t-tests complicates the process of determining statistical significance. This issue is discussed in section 3.3.3.

3.3.3 MULTIPLE COMPARISONS

When a large number of statistical tests are performed, some will result in p-values that are less than alpha values, purely by chance. Note that the p-value indicates the chance of obtaining the observed result when the null hypothesis is true (and not the probability that the null hypothesis is true), and each subsequent test presents another opportunity for a false positive to occur. In order to account for this inflated probability, the Bonferroni correction [47] is introduced to the analysis during multiple tests:

$$\alpha_N = \frac{\alpha}{\#\, tests} \tag{3.20}$$

where α_N is the new alpha value and α the original alpha value. When determining statistical significance in this case, the p-values must be less than α_N to reject the null hypothesis.

3.4 Vowel Extraction and Analysis

3.4.1 Vowel Extraction

All vowel data was extracted from the word-level prompts described in the EMA-MAE Dataset Manual [33]. As discussed in section 2.1.1, English vowels (aside from diphthongs) are formed by maintaining a stationary vocal tract during articulation. This means that the articulators (and by extension, the sensors placed on them) should barely move during vowel production. In other words, the formants and sensors positions should each have one data point for each vowel. In order to avoid the coarticulation effects described in section 2.1.1, the vowel data is extracted from the middle of the vowel pronunciation (as far away from either adjacent consonant as possible). In order to obtain a maximally stable representation of the vowel midpoint, the values over the middle 20 ms of the vowel pronunciation were averaged to obtain single data points for each sensor and formant.

As discussed in section 2.1.1, formant analysis is typically more reliable and suitable for vowels than for most consonants, and in this work, formants are only studied in the context of vowels. During the overview of the EMA-MAE corpus in section 2.2, the fact that interference led to significant suppression of formants greater than F2 was discussed. In order to avoid data corruption due to the reduced liability of these higher order formants, only F1 and F2 are studied in the formant analysis. For each occurrence of each vowel in the word-level prompts, the average F1 and F2 values are calculated from the middle 20 ms of the vowel pronunciation.

3.4.2 Vowel Formant Analysis

3.4.2.1 Formant Normalization

As discussed in section 2.1.1, the formants produced by a speaker are dependent on the size and shape of their vocal tract. No two humans have the exact same vocal tract shape and dimensions. This means that even if two people are articulating a sound in roughly the same way, the formants produced by these two people will still be different. Ideally, physiological differences like this would be removed from the analysis, allowing for a focus on the accuracy of articulation during cross-speaker comparisons. There are a number of formant normalization techniques that address this very issue. Popular normalization techniques include the Gerstman, Lobanov, Nordstrom, and Nearey methods [48] [49].

In a study conducted by [48], a multivariate analysis of variance (MANOVA) of a number of normalization techniques showed that when comparing across vowel, gender, and region, the Lobanov method proved to reduce anatomical/physiological differences most effectively while also preserving most of the sociolinguistic variation in acoustic measurements. For this reason, the Lobanov method was chosen as the formant normalization method for the data of the EMA-MAE corpus. The Lobanov normalization method calculation is as follows [49]:

$$F_i^N = \frac{F_i - \mu_i}{\sigma_i} \tag{3.21}$$

where F_i is the ith formant, F_i^N is the normalized value of F_i, μ_i is the mean value of the speaker's ith formant frequency, and σ_i is the standard deviation of the speaker's ith formant frequency. After each formant value was extracted as described in section 3.4.1, it was normalized using this method.

After normalization, the new F1 and F2 values are no longer expressed in Hertz. In order to display the values in the typical formant ranges, functions were used to scale them [50]:

$$F1' = 250 + 500\frac{F_1^N - F_{1-min}^N}{F_{1-max}^N - F_{1-min}^N} \tag{3.22}$$

$$F2' = 850 + 1400\frac{F_2^N - F_{2-min}^N}{F_{2-max}^N - F_{2-min}^N} \tag{3.23}$$

where F_i^N is the normalized value of the ith formant, Fi' is the scaled version of F_i^N, and $F_{i\text{-min(max)}}^N$ is the minimum (or maximum) value of the ith formant for *all* speakers. Note that these scaling functions do not maintain the exact relationships between the formant values (they become distorted) when transforming the data into a familiar format. Therefore, these functions are only used for plotting formant data (specifically, for comparison to the vowel quadrilateral), and are not involved in any analysis of the formants.

3.4.2.2 Data Analysis Overview

Several different calculations were made on the normalized formants obtained from each speaker's data. Among these, the calculations most relevant to this study are:

- the mean and standard deviation of the formant values (F1 and F2) of each vowel for each speaker.
- the convex hull surrounding all F1-F2 combinations for each speaker, as well as the area of this convex hull.
- ANOVAs and t-tests comparing both the formants of each vowel and the total convex hull areas across L1 and gender.

There were two different kinds of ANOVAs performed, both across the same speaker groups: native English-speaking males (ENGM), native English-speaking females (ENGF), native Mandarin-speaking males (MANM), and native Mandarin-speaking females (MANF). Each of these four groups contains 10 speakers. This corresponds to 4 different groups, and 40 total data points. Plugging these values into equations (3.14)-(3.16) yields df values of 3 DOF between and 36 DOF within. Given the df values and the formant data, the ANOVA results may be calculated by hand. However, R was used for the ANOVA and all other statistical analysis calculations. The first ANOVA performs vowel-by-vowel comparisons of F1 and F2 across the speaker groups above. The second ANOVA performs comparisons of the sizes of the F1-F2 for all vowels (effectively a comparison of the total vowel working spaces) across the same groups.

For all statistical analyses, a confidence level of 0.95 (or 95%) was chosen. By equation (3.5), the corresponding alpha value is 0.05 (or 5%). As explained in section 3.3, the alpha value sets the threshold for determining statistical significance, and 0.05 is the most commonly used value. As previously discussed, Bonferroni correction is used to account for the increased risk of type 1 error when performing multiple t-tests. With 4 speaker groups, 6 t-tests are required for individual comparisons (ENGM-ENGF, ENGM-MANM, ENGM-MANF, ENGF-MANM, ENGF-MANF, MANM-MANF). Using equation (3.20), the new alpha value may be calculated:

$$\alpha_N = \frac{\alpha}{\#\, tests} = \frac{0.05}{6} = \mathbf{0.0083} \tag{3.24}$$

For individual t-tests, the obtained p value must be less than 0.0083 to reject the null hypothesis.

The results of these calculations described above were analyzed with the primary intent of comparing native English (AE) speaker formant production to that of Mandarin accented English (MAE) speakers, with intra-language comparisons provided for frame of reference.

3.4.2.3 Comparison Expectations

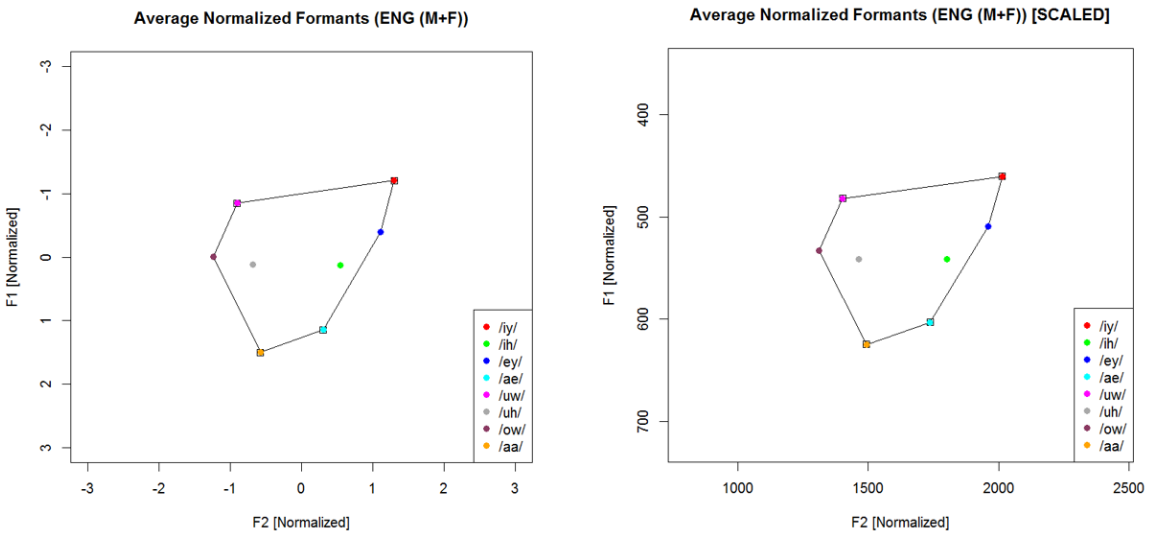
Given the background information covered in section 2.1, there were a number of expectations regarding the results of the analysis. Due to the fact that formants values are determined by the size and shape of the vocal tract, and because men typically have larger vocal tracts than women, the formants produced by men are usually lower than those of women. However, this is a physiological difference between speaker groups, and should be significantly reduced by the Lobanov formant normalization (leaving only sociolinguistic differences). Therefore, with AE speakers producing their native language (assumed to be spoken correctly at all times), comparisons between ENGM and ENGF were expected to yield no statistical significance (no significant differences in formant production expected).

Mandarin accented English (MAE) speakers are a different case. Three of the vowels contained within the EMA-MAE corpus (/ih/, /ae/, /uh/) do not exist in Mandarin Chinese. As explained in section 2.1, MAE speakers have been known to replace unfamiliar sounds with the most familiar sound in their native language. The mean formants of these vowels are expected to be farther from their English counterparts than the other vowels. There are also a number of factors that introduce a degree of unpredictability into the data. These include coarticulation effects and MAE speakers' distorted perception of English vowel duration (both discussed in section 2.1). Each MAE speaker may handle these complications differently, which would lead to a wide array of formant placements. Therefore, comparisons between MANM and MANF are not expected to yield any statistical significance, but comparisons across L1 are expected to vary, especially across vowels that don't exist in Mandarin Chinese.

3.4.2.4 RESULTS

Figure 3.5 displays the average normalized formant values of each EMA-MAE vowel across all AE speakers, both scaled and unscaled. The data points of the four corner vowels of the vowel quadrilateral are boxed. Note that the axes of the plots are changed to reflect the correlation between the vowel quadrilateral and formant frequencies.

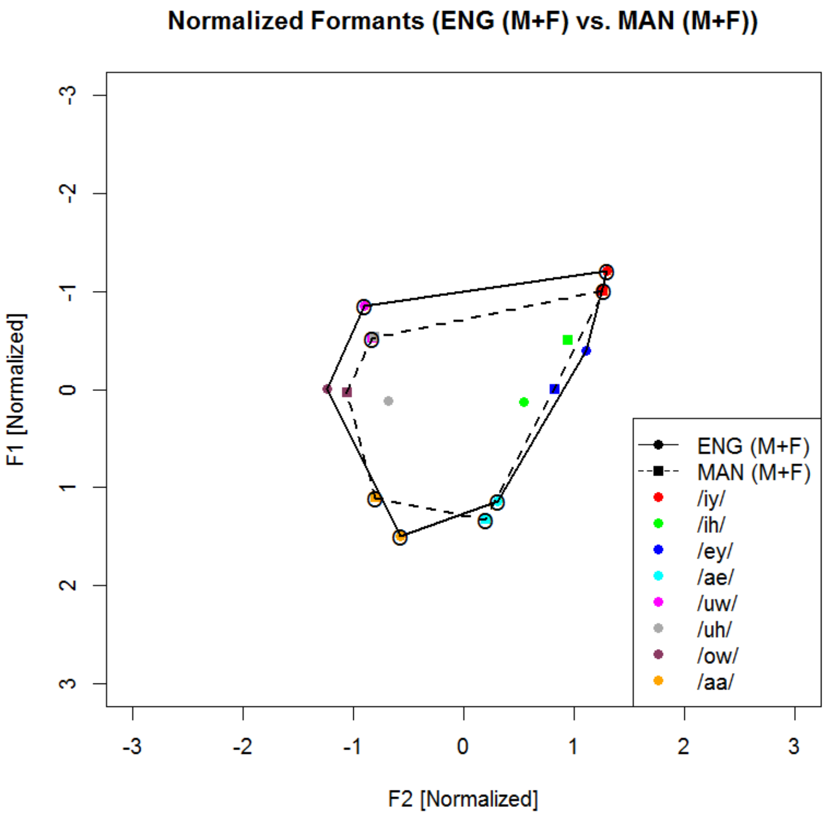
Figure 3.5 – *Average Normalized Vowel Formants: Native English*



Given the vowel quadrilateral of Figure 2.4 and the idea that formants are related to tongue positioning, as discussed in section 2.1, a few of the vowels land in slightly unexpected locations. Specifically, /ih/ and /uh/ were expected to have lower F1 values, and /ow/ was expected to have a greater F2. Different experimental data usually yields slight differences in formant measurements across English speakers [22] [21] , so small differences were expected. It is also worth noting these formants are the result of normalization designed to factor out physiological differences across speaker groups. The reduction of these differences may have resulted in formant shifts that result in a less similar formant distribution than un-normalized data. This can be checked via comparisons between normalized and un-normalized formant data, but that is not the focus of this study. In general, this formant distribution supports the idea that the vowel quadrilateral is related to formant frequencies.

Figure 3.6 displays the normalized average formant frequencies of each vowel for both AE speakers and MAE speakers in the same window. The corner vowels are marked with large circles, and groups are differentiated by point and line type.

Figure 3.6 - *Average Normalized Vowel Formants: English vs. Mandarin*



Treating the AE formants as targets for “correct” pronunciation, brief observation shows that most of the MAE formants fell within the general area of their targets. Two vowels, /ih/ and /uh/, landed fairly far from their English counterparts. The most interesting vowel is this figure is /uh/, which not only deviates from the AE target significantly, but also overlaps with the MAE formants of /uw/. This implies that MAE speakers pronounce these two vowels nearly identically. Both /ih/ and /uh/ are among the vowels that do not exist in Mandarin Chinese, so these may be cases of MAE speakers replacing unfamiliar sounds with nearby native-like sounds (as discussed in section 2.1).

Table 3.3 and Table 3.4, which display means and standard deviations (respectively) of the formants of each vowel for AE and MAE speakers, provides a more detailed comparison of the two groups. These tables include an *ENG-MAN* (English minus Mandarin) section, which display the difference in formant values between the AE and MAE speakers.

Table 3.3 – *Mean Normalized Vowel Formants: English vs. Mandarin*

	F1			**F2**		
Vowel	***ENG***	***MAN***	***ENG-MAN***	***ENG***	***MAN***	***ENG-MAN***
1 (/iy/)	-1.205	-1.006	-0.199	1.300	1.265	0.035
2 (/ih/)	0.131	-0.500	0.631	0.541	0.944	-0.403
3 (/ey/)	-0.396	-0.002	-0.393	1.112	0.822	0.290
4 (/ae/)	1.147	1.330	-0.182	0.305	0.199	0.107
5 (/uw/)	-0.850	-0.521	-0.329	-0.905	-0.831	-0.074
6 (/uh/)	0.122	-0.530	0.652	-0.684	-0.805	0.121
7 (/ow/)	-0.008	0.031	-0.039	-1.233	-1.057	-0.176
8 (/aa/)	1.502	1.106	0.395	-0.576	-0.800	0.223

Table 3.4 – *Standard Deviation of Normalized Vowel Formants: English vs. Mandarin*

	F1			**F2**		
Vowel	***ENG***	***MAN***	***ENG-MAN***	***ENG***	***MAN***	***ENG-MAN***
1 (/iy/)	0.196	0.140	0.056	0.075	0.100	-0.025
2 (/ih/)	0.192	0.219	-0.027	0.092	0.167	-0.075
3 (/ey/)	0.193	0.347	-0.153	0.087	0.115	-0.028
4 (/ae/)	0.152	0.161	-0.009	0.147	0.155	-0.007
5 (/uw/)	0.195	0.281	-0.086	0.148	0.130	0.018
6 (/uh/)	0.257	0.292	-0.035	0.117	0.170	-0.054
7 (/ow/)	0.198	0.236	-0.037	0.075	0.114	-0.039
8 (/aa/)	0.315	0.274	0.041	0.121	0.129	-0.009

Similar to the plot of Figure 3.6, Table 3.3 shows that the F1 value of /ow/, and the F2 values of /iy/ and /uw/ for MAE speakers are very close to meeting their AE counterparts. Meanwhile, the F1 values of /ih/ and /uh/ for MAE speakers are especially far from their AE counterparts. It can be observed that in general, there is greater difference in F1 values than F2 values across groups. Section 2.1 discussed the fact that first and second formants are thought to be related to tongue height and front/back-ness (respectively). F1 is said to increase as tongue height decreases. Applying this to the *ENG-MAN* section of the F1 values, this indicates that when the difference is positive, MAE speakers' tongues were placed higher than AE speakers on average (vice-versa when the differences are negative). Most of the vowels have negative differences, which may suggest that MAE speakers generally place their tongues higher during articulation of English vowels. Meanwhile, F2 is said to increase as the tongue moves forward. Applying this to the *ENG-MAN* section of the F2 values, this indicates that when the difference is positive, MAE speakers' tongues were placed further back than AE speakers on average. Overall, the smaller differences among F2 than F1 indicate that MAE speakers place their tongues closer to the target front/back-ness than to the target height during articulation.

In Table 3.4, the differences in standard deviation of normalized formant values between the two groups is almost always negative (-0.034 on average). This indicates that the formants produced by MAE speakers vary more widely than AE speakers on average. Extending this idea using the concept of correlation between tongue position and formant frequencies, this suggests that MAE use a wider range of tongue positions (both vertically and horizontally) than AE speakers to produce the same vowels. This makes sense, given that MAE are speaking English as their second language. Section 2.1 discussed the fact that L1 effects tend to interfere with L2 speech production (especially for languages as different as English and Mandarin Chinese). MAE speakers are aiming for a target pronunciation despite L1 speech habits when speaking English, while AE speakers place their articulators in a typical position naturally.

As described in section 3.4.2.2, two types of ANOVAs were performed on the formant data. The first ANOVA compared the means of the formants for each vowel across both L1 and gender (creating four groups: ENGM, ENGF, MANM, MANF). Table 3.5 displays the F values obtained from the ANOVA, and Table 3.6 displays the corresponding p values. Recall that the alpha value for all ANOVAs was chosen as 0.05. For the differences between groups to be statistically significant, the corresponding p value must be less than 0.05. In these tables, the comparisons with p values that fell below the alpha value (meaning that the group means have significant differences between them) are highlighted in orange.

Table 3.5 – *ANOVA: Cross-Language, Cross Gender Comparison of Vowel Formants [F-Values]*

	F-value	
Vowel	**F1**	**F2**
1 (/iy/)	4.394	0.634
2 (/ih/)	33.80	28.83
3 (/ey/)	8.446	33.03
4 (/ae/)	4.953	6.800
5 (/uw/)	6.254	2.574
6 (/uh/)	30.03	3.002
7 (/ow/)	0.360	12.63
8 (/aa/)	14.53	11.78

Table 3.6 - *ANOVA: Cross-Language, Cross Gender Comparison of Vowel Formants [p-Values]*

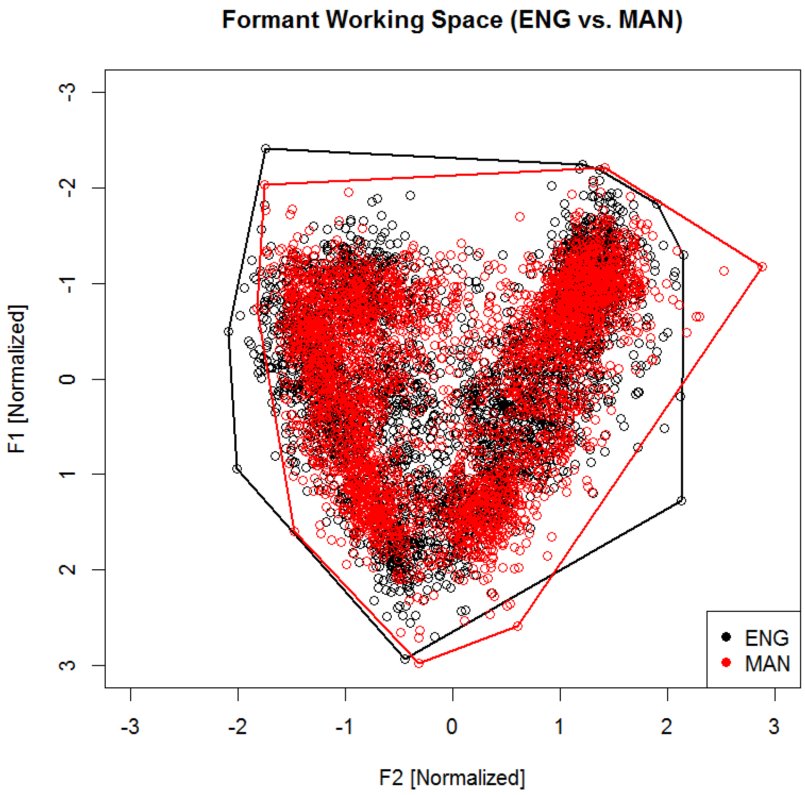
	p-Value	
Vowel	**F1**	**F2**
1 (/iy/)	0.0098	0.5980
2 (/ih/)	0.0000	0.0000
3 (/ey/)	0.0022	0.0000
4 (/ae/)	0.0056	0.0010
5 (/uw/)	0.0016	0.0691
6 (/uh/)	0.0000	0.0431
7 (/ow/)	0.7823	0.0000
8 (/aa/)	0.0000	0.0000

Although not the same comparisons, elements of these results support the conclusions made from Figure 2.1 and Table 3.3. Those results showed that the F1 values of /ow/ across L1 were very similar, as were the F2 values of /iy/ and /ih/. The fact that the groups yielded no statistical significance here means that the groups are also the same both within and across gender. Every other group yielded p values that fell below the alpha value, meaning at least two of the 4 groups have significant differences between each other.

T-tests were performed to examine the differences between individual groups (see section 3.3 for details on t-tests and multiple comparisons, and section 3.4.2.2 for specific details on the t-tests performed for the formant analysis). In cases where the ANOVA determined statistical significance from its comparisons, the t-tests determined statistical significance across most cross language comparisons (ENGM vs MANM, ENGM vs MANM, etc). The results generally supported the ANOVA, and no additional trends suggested by background information regarding Mandarin accented English production were observed. One interesting artifact of the analysis was that there were three formant comparisons that yielded statistical significance when comparing AE males to AE females. No differences among native English speakers were expected, so this may be due to inconsistent pronunciation by some of the AE participants or the failure of the Lobanov method to effectively factor out physiological differences between speakers from all formants. A chart displaying all of the statistical significance results of the t-tests is shown in Appendix A (Table 6.2).

Figure 3.7 displays a plot of all formants produced by all speakers in the AE and MAE speaker groups (essentially, the collective formant vowel working spaces of the two speaker groups).

Figure 3.7 – *Vowel Formant Working Space Comparison: English vs. Mandarin*



The figure makes it clear that the AE and MAE speakers of the EMA-MAE corpus generally occupy the same formant ranges. This result disproves any theory that suggests that differences in vowel formants across language are due to the different language groups occupying different formant spaces. In other words, differences in formants across groups is intrinsic to the vowel in question, and not because a given language group naturally produces different formant values. Table 3.7 provides the results of t-tests performed on the data shown in Figure 3.7.

Table 3.7 – *Formant Vowel Space T-Test Results*

F1		**F2**	
t-Value	***p-Value***	***t-Value***	***p-Value***
-0.152	0.879	0.435	0.664

The p values for both formants are well above 0.05, which confirms that there are no statistically significant differences between the overall formant spaces of AE and MAE speakers.

3.4.3 Vowel Kinematic Analysis

3.4.3.1 Data Analysis Overview

For vowels, the treatment of kinematic data is very similar to that of the formant data (see section 3.4.2). The main difference between the kinematic and formant data lies in the number of data points generated per vowel utterance. The F1-F2 extraction created one two dimensional data point. The kinematic analysis examines four EMA sensors: tongue dorsum (TD), tongue blade (TB), upper lip (UL), and lower lip (LL). Given that the analysis and template creation process both take place in the midsagittal plane, each of these sensors corresponds to a two-dimensional data point (x and y sensor positions). As discussed in section 3.2, the properties of UL and LL are combined to create the LS "sensor". And so, the kinematic data has 3 two dimensional data points, corresponding to x-y values of TD, TB, and LS.

For each sensor, the similar calculations to those performed on the formants were performed. These include:

- calculations of the mean and standard deviation of the each sensor position (x and y) of each vowel for each speaker.
- calculations of the convex hulls surrounding all x-y combinations for each sensor of each speaker, as well as the areas of these convex hulls.
- calculations of the convex hulls surrounding all replicates of each vowel for each sensor for all AE speakers, and again for all MAE speakers.
- ANOVAs and t-tests comparing both the sensor positions of each vowel and the total convex hull areas across L1 and gender.

The ANOVAs performed on the sensor positions have the same parameters and the formants ANOVAs. The only difference is that they are calculated three times (one set for each sensor). This means that, as equation (3.24) specifies, the threshold for statistical significance is also 0.0083 for multiple comparisons of sensor positions. Section 3.4.2.2 contains all the details regarding this analysis.

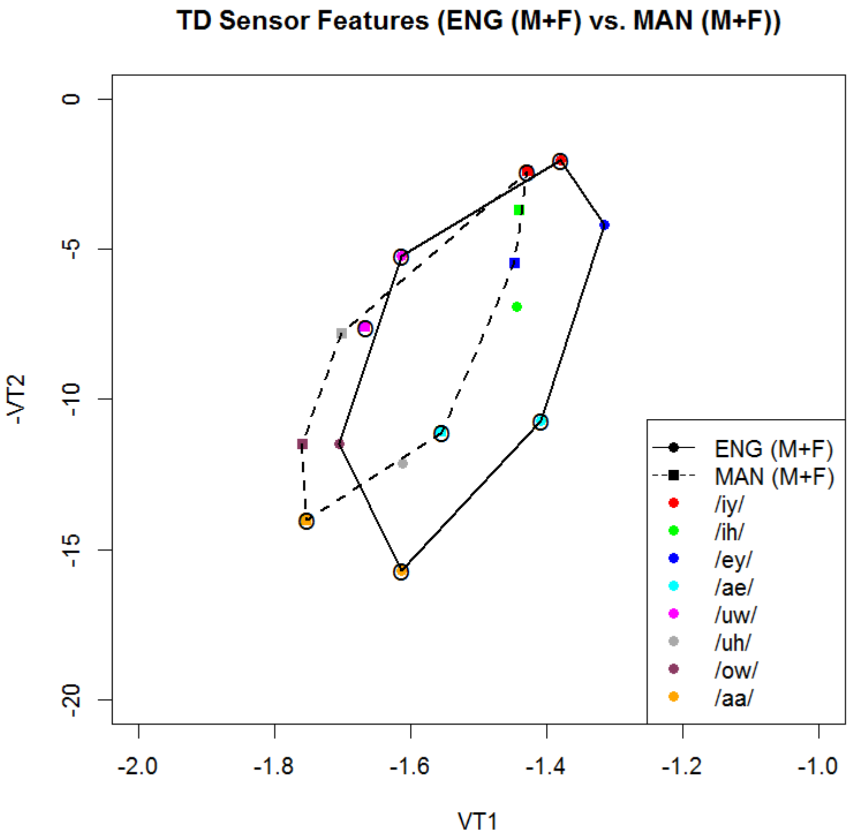
3.4.3.2 Comparison Expectations

The relationship between expectations based on prior research and the actual data shape the approach to vowel template creation. It is important to list and evaluate these expectations before transitioning into the template development stage. During the conversion to feature space, sensor x position data was normalized to factor our cross-speaker variance and sensor y data was re-referenced to be expressed relative to the speaker's palate. These operations (1) allow for meaningful comparisons of position across speakers and speaker groups, and (2) present the data in a way that better reflects the size and shape of the vocal tract. Given the relationship between tongue positioning, the vowel quadrilateral, and formant frequencies (see section 2.1 for details), expectations can be built for the outcome of the kinematic data analysis. As displayed in Figure 2.6, the origin of the kinematic data lies at the central maxillary incisors (CMI), the x-axis is pointed anterior to the CMI, and the y axis is directed straight up. This means that tongue sensor x values are directly related to the front/back-ness of the tongue and the second formant frequency (as the tongue moves forward, x and F2 increase). Meanwhile, tongue sensor y values are directly related to tongue height and inversely related to the first formant frequency (as tongue height increases, y increases and F1 decreases). Given this correlation between sensor position and formant frequency, the tongue sensor positions were expected to follow the same trends as the formant data.

3.4.3.3 Results

Figures Figure 3.8-Figure 3.10 display the normalized average sensor positions (in feature space) of each vowel for both AE speakers and MAE speakers in the same window for each EMA sensor. In each plot, the corner vowels are marked with large circles, and groups are differentiated by point and line type.

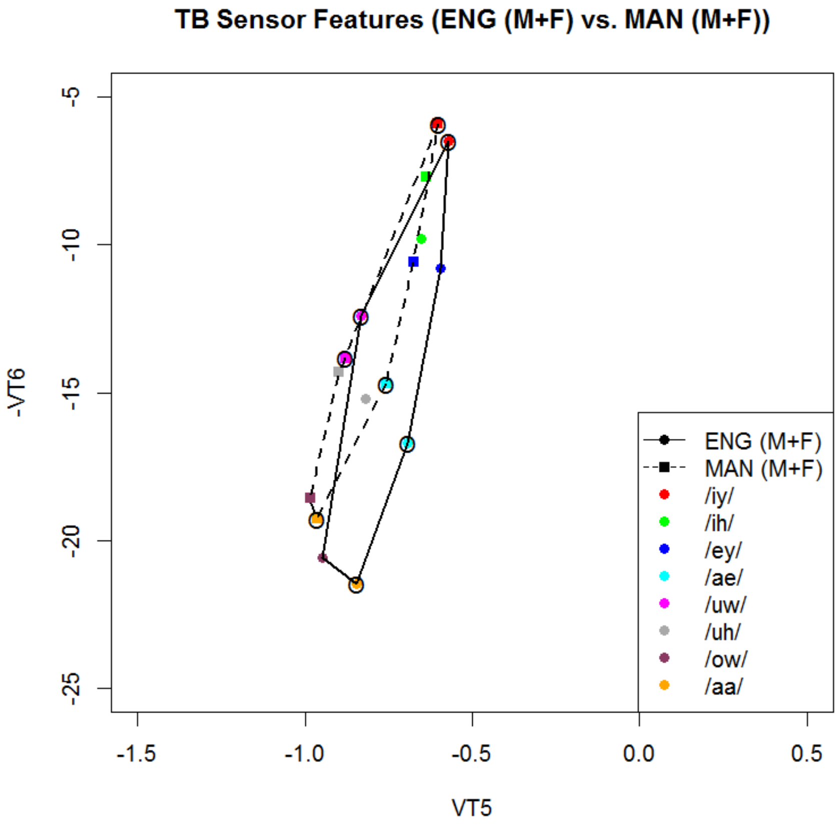
Figure 3.8 - *Average Normalized Vowel Sensor Positions: English vs. Mandarin [TD]*



Quick observation of Figure 3.8 makes it clear that for TD, the distribution of mean vowel positions of AE and MAE speakers have very similar shapes. However, the means for the groups are fairly far from each other for most vowels. Many of the greatest differences in location between AE and MAE for the TD sensor lie with the vowels that do not exist in Mandarin Chinese (/ih/, /ae/, /uh/). While /ih/ is located in roughly the same horizontal position across languages, there is a significant different in tongue-to-palate difference. Vowel /ae/ expresses the inverse of this relationship; there is a small vertical difference, but large horizontal difference between the two vowels. For vowel /uh/, there are significant differences between groups both vertically and horizontally.

Interestingly, for MAE speakers, the position of corner vowel /uw/ is not part of the convex hull of the vowel space. It can be seen that this is due to the fact that MAE speakers, on average, move their tongue dorsum into roughly the same position to produce /uh/. This vowel, which has now displayed the worst performance for both formant frequency and tongue dorsum, is likely a special case of MAE speakers attempting to produce a completely unfamiliar sound.

Figure 3.9 - *Average Normalized Vowel Sensor Positions: English vs. Mandarin [TB]*



The average positions of TB for the two language groups are much closer to each other than those of TD. There also seems to be much less anterior-posterior (VT5) movement of the tongue blade than the tongue dorsum for both AE and MAE speakers, according to the figure. Aside from /iy/, /ey/, and /uh/, the significant difference between vowels across L1 seems to be tongue-to-palate distance. While differences between AE and MAE speakers were generally smaller for TB, the overall distributions of the TB sensor positions were much more different from the formant distribution than the TD sensor positions. This implies that the tongue dorsum (TD) position is much more reflective of the relationship between tongue position and formant frequencies than the tongue blade (TB) position. As was the case for TD, MAE speakers occupy roughly the same position when pronouncing /uw/ and /uh/.

Figure 3.10 - *Average Normalized Vowel Sensor Positions: English vs. Mandarin [LS]*

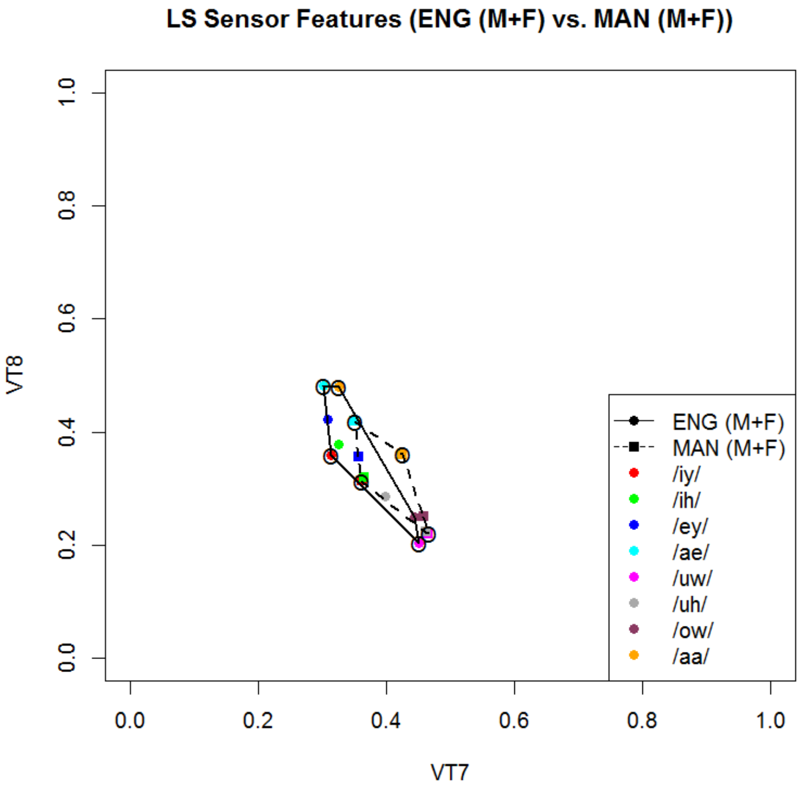


Figure 3.10 shows that for LS, the positions corresponding to each vowel are close to each other both within and across language groups. This indicates that for both AE and

MAE speakers, there is much less movement variation in the lips than tongue during articulation of vowels. The figure also indicates that on average, MAE speakers have a smaller range of lip motion than AE speakers.

WHEN COMPARING FIGURE 3.8 AND FIGURE 3.10 TO FIGURE 3.6, A SMALL

DISTRIBUTIONS BEWEEN TONGUE SENSOR POSITIONS AND FORMANT

THE FIGURES DO NOT REJECT THE IDEA THAT FORMANT FREQUENCIES AND

RELATED, BUT THE DISTRIBUTIONS OF SENSOR POSITIONS ARE STILL VERY

DISTRIBUTIONS. FOR MAE SPEAKERS, THE FORMANT FREQUENCIES, TD

SENSOR POSITIONS, AND LS SENSOR POSITIONS OF VOWELS /UH/ AND /UW/ ARE

TO EACH OTHER. WHILE THIS HIGHLIGHTS THE FACT THAT MAE SPEAKERS

THESE VOWELS, IT ALSO STRONGLY SUPPORTS THE IDEA THAT FORMANT

POSITIONING. THE FACT THAT THE SENSOR POSITIONS NEARLY OVERLAPPED

OVERLAPPED (ALONG WITH THE FACT THAT THE FORMANT AND SENSOR

DATA DISTRIBUTIONS) IMPLIES THAT FORMANT FREQUENCIES AND TONGUE

RELATED, BUT THE RELATIONSHIP IS NOT LINEAR. IT IS WORTH NOTING THAT

FEATURES ARE SIMPLY A SCALED VERSION OF THE ORIGINAL X COORDINATES,

VERTICAL POSITIONS OF THE TONGUE SENSORS MAY BE PARTIALLY

DIFFERENCES IN DISTRIBUTIONS BETWEEN FORMANT FREQUENCIES AND

FORMANT FREQUENCIES ARE DETERMINED BY THE SIZE AND SHAPE OF THE

DIMENSIONS, THE DIFFERENCES IN SENSOR POSITIONS BETWEEN GROUPS ARE

OBSERVING THE PLOTS IN FIGURES FIGURE 3.8-FIGURE 3.10. HOWEVER, A

MEANS OF THE SENSOR POSITIONS OF EACH VOWEL FOR AE AND MAE

SPEAKERS CAN BE FOUND IN

Appendix A (Table 6.3). Table 0.1 displays the standard deviations of the sensor positions of each vowel for AE and MAE speakers. These tables include *E-M* (English minus Mandarin) columns, which display the differences in standard deviations between the AE and MAE speakers.

Table 0.1 - *Standard Deviation of Normalized Vowel Sensor Positions: English vs. Mandarin*

	TDx			**TDy**		
Vowel	***ENG***	***MAN***	***E-M***	***ENG***	***MAN***	***E-M***
1 (/iy/)	0.231	0.222	0.009	0.978	0.972	0.006
2 (/ih/)	0.218	0.198	0.019	1.821	1.397	0.424
3 (/ey/)	0.224	0.225	-0.001	1.464	1.470	-0.007
4 (/ae/)	0.203	0.242	-0.039	2.545	2.293	0.252
5 (/uw/)	0.287	0.229	0.058	2.550	2.891	-0.341
6 (/uh/)	0.240	0.221	0.019	2.733	2.718	0.016
7 (/ow/)	0.264	0.234	0.030	3.243	3.945	-0.702
8 (/aa/)	0.248	0.238	0.010	4.097	3.422	0.674

	TBx			**TBy**		
Vowel	***ENG***	***MAN***	***E-M***	***ENG***	***MAN***	***E-M***
1 (/iy/)	0.128	0.123	0.005	2.022	2.544	-0.522
2 (/ih/)	0.125	0.133	-0.008	1.601	2.649	-1.048
3 (/ey/)	0.117	0.146	-0.029	2.851	2.670	0.180
4 (/ae/)	0.139	0.180	-0.041	2.417	2.980	-0.563
5 (/uw/)	0.162	0.160	0.002	3.025	3.000	0.026
6 (/uh/)	0.149	0.158	-0.009	2.688	3.025	-0.337
7 (/ow/)	0.150	0.174	-0.024	3.542	2.787	0.756
8 (/aa/)	0.154	0.166	-0.013	3.436	2.834	0.602

	LSx			**LSy**		
Vowel	***ENG***	***MAN***	***E-M***	***ENG***	***MAN***	***E-M***
1 (/iy/)	0.124	0.099	0.025	0.069	0.070	-0.001
2 (/ih/)	0.118	0.099	0.019	0.073	0.069	0.004
3 (/ey/)	0.115	0.100	0.015	0.072	0.072	0.000
4 (/ae/)	0.109	0.104	0.005	0.082	0.069	0.013
5 (/uw/)	0.104	0.102	0.002	0.065	0.065	-0.001
6 (/uh/)	0.109	0.106	0.003	0.064	0.071	-0.007
7 (/ow/)	0.099	0.105	-0.006	0.068	0.071	-0.003
8 (/aa/)	0.105	0.096	0.009	0.082	0.077	0.005

Table 0.1 indicates that on average, AE speakers typically have a larger horizontal variation of motion of the tongue dorsum (TD) and lips (LS), and smaller horizontal variation of motion of the tongue blade (TB). In most cases, these differences are very small, suggesting that there is not a significant difference in overall horizontal motion variability across L1. For vertical motion, the variability itself varies with each sensor. There are no identifiable trends in cases where AE speakers have a larger variation of motion than MAE speakers and vice-versa. One interesting observation is the fact that there is much less consistency in the amount of variation in vertical tongue movement across vowels than that of horizontal movement. For example, note that the horizontal standard deviation of TD movement for AE speakers across all vowels falls within the 0.2-0.3 range. Meanwhile vertical variation for the same sensor ranges from 0.978-4.097. Despite this significant amount of dispersion in sensor position variation, the variation in formant frequencies was shown to be much more consistent (see Table 0.1). This highlights the fact that several different articulatory configurations can result in the same acoustic result.

Similar to the formant analysis, an ANOVA comparing the means of the sensor positions for each vowel across both L1 and gender (corresponding to groups: ENGM, ENGF, MANM, MANF). Table 0.2 displays the F values obtained from the ANOVA, and Table 0.3 displays the corresponding p values.

Table 0.2 - *ANOVA: Cross-Language, Cross Gender Comparison of Vowel Sensor Positions [F-Values]*

	Sensor

Vowel	TDx	TDy	TBx	TBy	LSx	LSy
/iy/	2.518	0.941	0.271	0.956	5.542	2.410
/ih/	2.211	17.82	0.082	4.163	4.873	3.536
/ey/	4.243	5.668	1.327	0.488	5.234	3.617
/ae/	4.255	0.130	0.517	2.117	4.792	3.644
/uw/	2.907	9.002	0.590	5.709	2.471	0.656
/uh/	1.782	11.03	0.796	1.285	4.727	4.385
/ow/	2.100	3.908	0.176	2.813	2.370	0.516
/aa/	3.471	1.603	1.810	2.135	8.115	9.860

Table 0.3 - *ANOVA: Cross-Language, Cross Gender Comparison of Vowel Sensor Positions [p-Values]*

	Sensor					
Vowel	TDx	TDy	TBx	TBy	LSx	LSy
/iy/	0.0735	0.4312	0.8456	0.4240	0.0031	0.0829
/ih/	0.1036	0.0000	0.9694	0.0125	0.0060	0.0242
/ey/	0.0115	0.0028	0.2808	0.6926	0.0042	0.0221
/ae/	0.0113	0.9414	0.6731	0.1152	0.0066	0.0215
/uw/	0.0479	0.0001	0.6258	0.0027	0.0775	0.5846
/uh/	0.168	0.0000	0.5039	0.2941	0.0070	0.0100
/ow/	0.1173	0.0163	0.9118	0.0530	0.0867	0.6741
/aa/	0.0259	0.2058	0.1628	0.1129	0.0003	0.0001

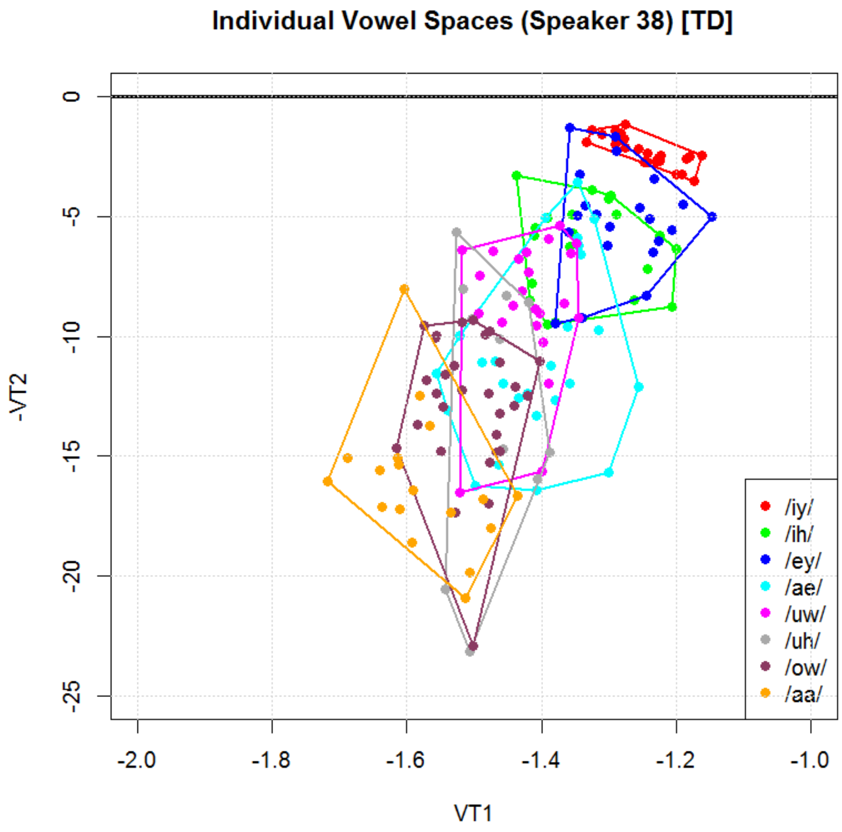
As explained in section 3.4.3.1, an alpha level of 0.05 was used for all ANOVAs. Note that all comparisons that yielded statistical significance ($p < 0.05$) were highlighted. The results of the LS comparisons supported the data shown in Figure 3.10. This agreement indicates that, for LS, the differences in means between language groups are reflective of the overall differences both within and across language and gender groups (in other words, no new trends among speaker groups were observed, and t-tests to study individual group differences are not needed).

The results of the TD comparisons, with the exception of /uh/, /ow/ and /aa/, matched the data shown in Figure 3.8. /uh/ yielded no significant differences in horizontal position, and /aa/ yielded no significant difference in vertical position across groups (despite the large differences in means). This indicates that the overall distributions of sensor positions for these vowels overlap to the point of removing any discernment between them. These results make sense, given the large amounts of deviation in these groups (see Table 0.1). ANOVA results for /ow/ yielded significant differences between in vertical position, despite the fact that the means shown in Figure 3.8 are nearly identical. This implies that there are significant differences among the four subgroups of the ANOVA, but t-tests yielded no significant differences between any of the individual groups. It is unclear why this inconsistency occurs, but it may be due to improper adjustment of the alpha value by the Bonferroni correction (see section 3.3.3 for details on the correction). It is worth noting that the p values returned from t-tests performed between AE females and MAE males and between AE females and MAE females both fell below 0.05. Before alpha value correction, these comparisons would have indicated statistical significance.

The comparisons performed for the TB sensor yielded statistical significance in only two cases: the y dimension for vowel /ih/, and the y dimension for vowel /uw/. T-tests revealed that for /ih/, the underlying comparisons that produced significant differences with each other were AE males vs. MAE males and AE females vs. MAE males. T-testing for vowel /uw/ revealed that the underlying comparisons that bore significant differences were AE females vs. MAE females, and interestingly, AE males vs. AE females. Given the notion that formant frequencies and tongue position are related, as well as the assumption that all AE speech is pronounced correctly, no differences within AE speakers were expected. This difference seems to further support the idea that different articulator positions can produce the same sound.

In general, individual t-tests showed that most subgroup sensor position differences that produced statistical significance were between AE females and MAE males. Even these differences occurred only a percentage of the time. The observation of means and ANOVA results showed that despite having significantly different means in some cases, there is enough variation in the data to blend it to the point of removing any ability to distinguish between groups in most cases. Figure 0.1, which plots the entire working space of each vowel of the TD sensor for one of the AE male speakers of the EMA-MAE corpus, demonstrates this fact.

Figure 0.1 – *TD Sensor Vowel Working Spaces [AE Male Speaker]*



NOTE THAT IN MOST CASES, A SINGLE SENSOR POSITION CANNOT BE TIED TO A GIVEN THE SIGNIFICANT OVERLAP IN SENSOR POSITIONS ACROSS VOWELS. POSITION CAN BARELY DISTINGUISH VOWELS FROM THE SAME SPEAKER, THE FOUND SUCH A SMALL AMOUNT OF SIGNIFICANT DIFFERENCES ACROSS SEVERAL MORE PLAUSIBLE. A TABLE DISPLAYING THE STATISTICAL SIGNIFICANCE PERFORMED ACROSS ALL INDIVIDUAL ANOVA GROUPS, VOWELS, AND SENSORS CAN BE FOUND IN

Appendix A (Table 6.4).

Figures Figure 0.1-Figure 0.3 display the total vowel working space for each sensor, for both AE and MAE speakers. Language groups are differentiated by color.

Figure 0.1 – *Total Sensor Position Working Space: English vs. Mandarin [TD]*

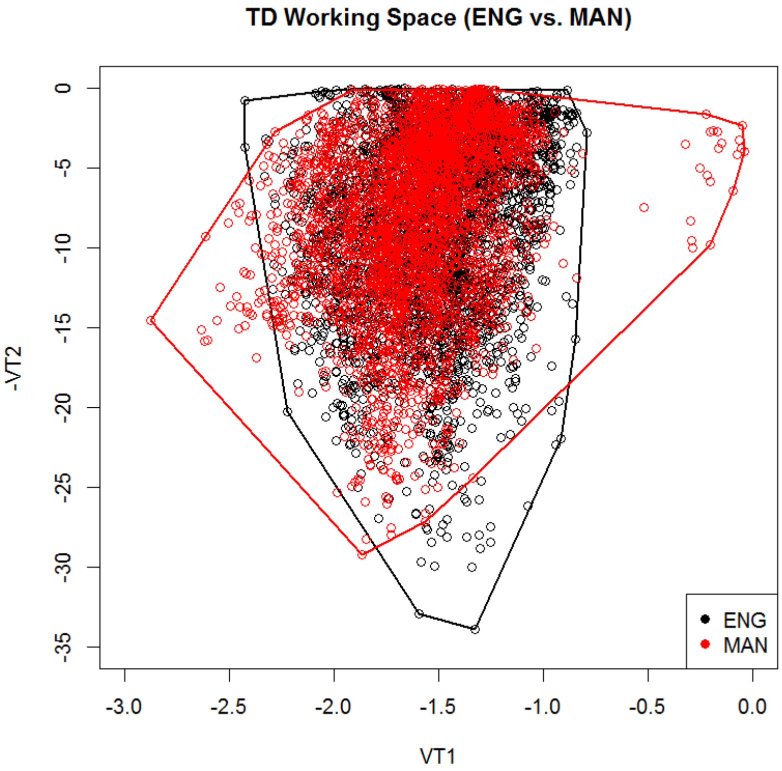


Figure 0.2 - *Total Sensor Position Working Space: English vs. Mandarin [TB]*

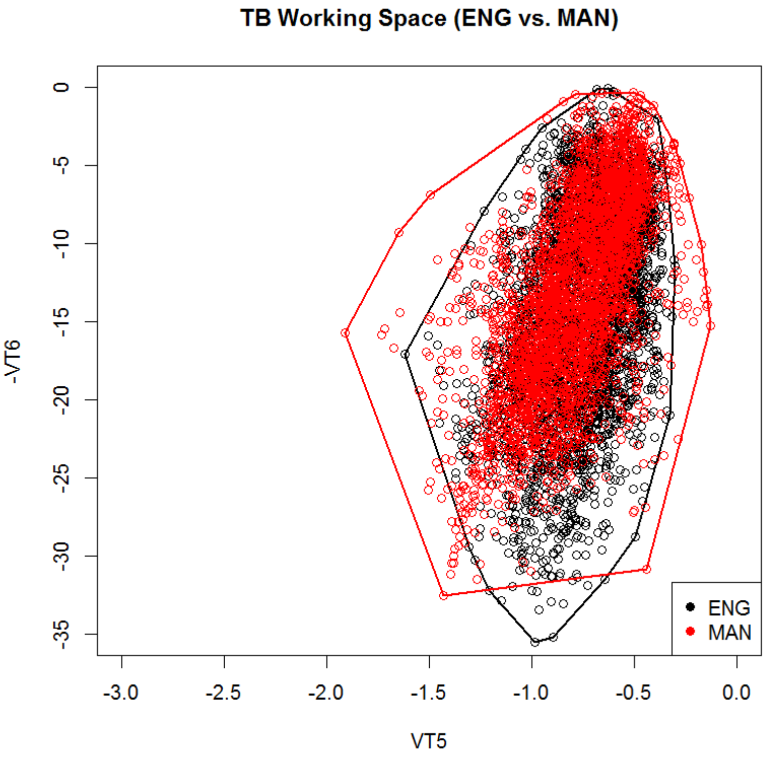
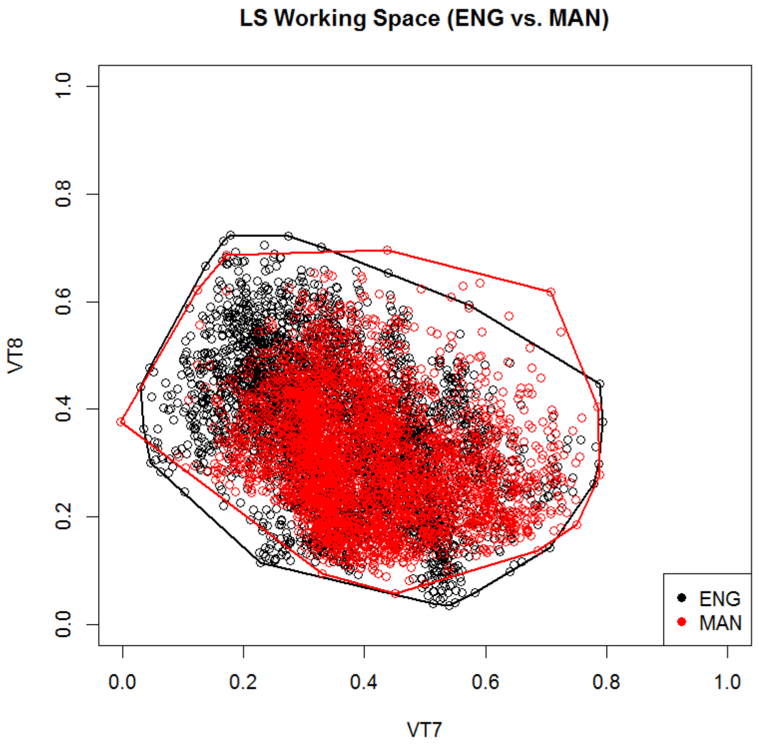


Figure 0.3 - *Total Sensor Position Working Space: English vs. Mandarin [LS]*



For all three sensors, MAE speakers seem to have less general dispersion, but more outliers. It can be observed that the sizes of the convex hulls for the MAE speakers is largely driven by the outliers of the data, which likely come from attempting to produce unfamiliar sounds. When observing the concentration of data point for each group, it seems that MAE speakers generally operate in a smaller range of motion than AE speakers when producing vowels. The working space of MAE speakers can nearly be considered a subset of the AE working space. Table 0.1 displays the results of t-tests performed on the data in these figures. Note that "-" is placed in cells with negligible values.

Table 0.1 - *Sensor Vowel Space T-Test Results*

	X		Y	
Sensor	***t-Value***	***p-Value***	***t-Value***	***p-Value***
TD	12.20	-	-2.386	0.017
TB	9.757	-	-7.546	-
LS	-12.47	-	14.87	-

All p values returned from the t-tests were well under 0.05, which confirm that the groups have significant differences in total working space for vowel production.

Overall, the data trends of the tongue sensor data did not match those of the formant frequencies. This lack of significant correlation between sensor position and formant frequency discredits kinematic data manipulation techniques founded in formant research. Therefore, the results and trends discovered through the formant analysis are largely uninvolved in the kinematic template creation process.

3.5 CONSONANT CLUSTER EXTRACTION AND ANALYSIS

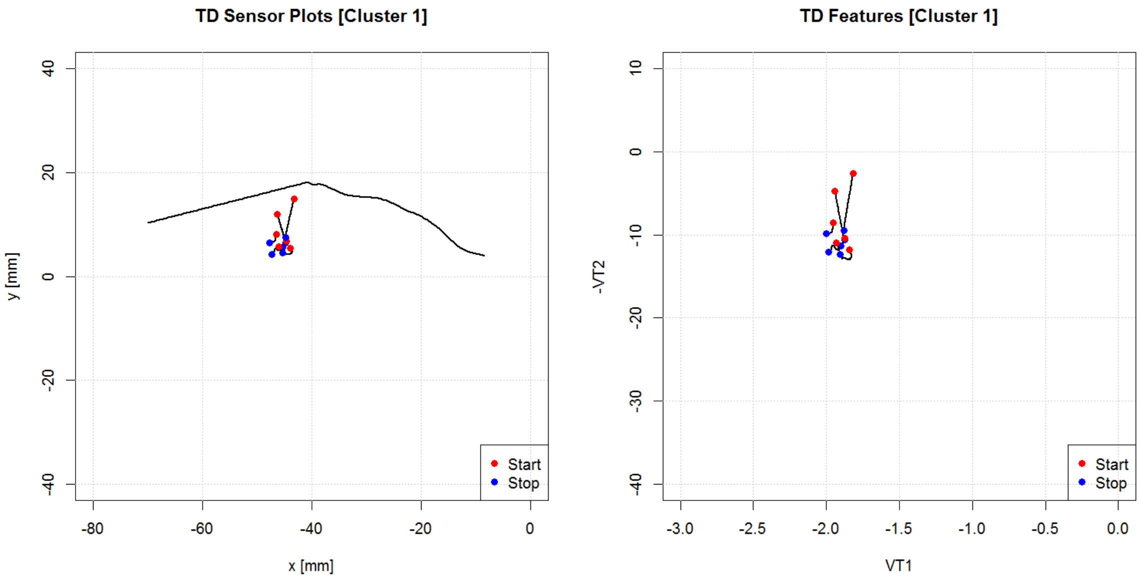
3.5.1 CLUSTER EXTRACTION

All consonant cluster data was extracted from the same word-level prompts that were used for vowel data extraction. Unlike vowels, which consist of stationary articulator positioning, consonant clusters are formed through continuous movement of the articulators (see section 2.1.1 for details on consonant cluster formation). This means that as opposed to a single data point, a cluster must be represented by a set of points that form a movement trajectory. Using labeling files, which contain the onset and offset times of each consonant cluster, every instance of each cluster was extracted as a set of x-y data points (along with the corresponding time stamps) for each EMA sensor for all speakers. In order to remove high frequency movement from each trajectory, the data sets were passed through a low pass digital filter with a cutoff frequency of 25 Hz. Each trajectory was converted to feature space, as described in section 3.2.

3.5.2 Kinematic Data Analysis

After extraction, plots were created to display the x-y trajectories of the consonant clusters. Figure 0.4 displays the TD sensor trajectories of all pronunciations of cluster */nd/* for a single speaker (specifically, a female native English speaker). The left plot displays the cluster repetitions in Euclidean space, while the right plot displays the same cluster repetitions in feature space. Note that for each plot, the start and end points of each cluster pronunciation are marked. Also note that in the Euclidean space representation, the speaker's palate is also shown.

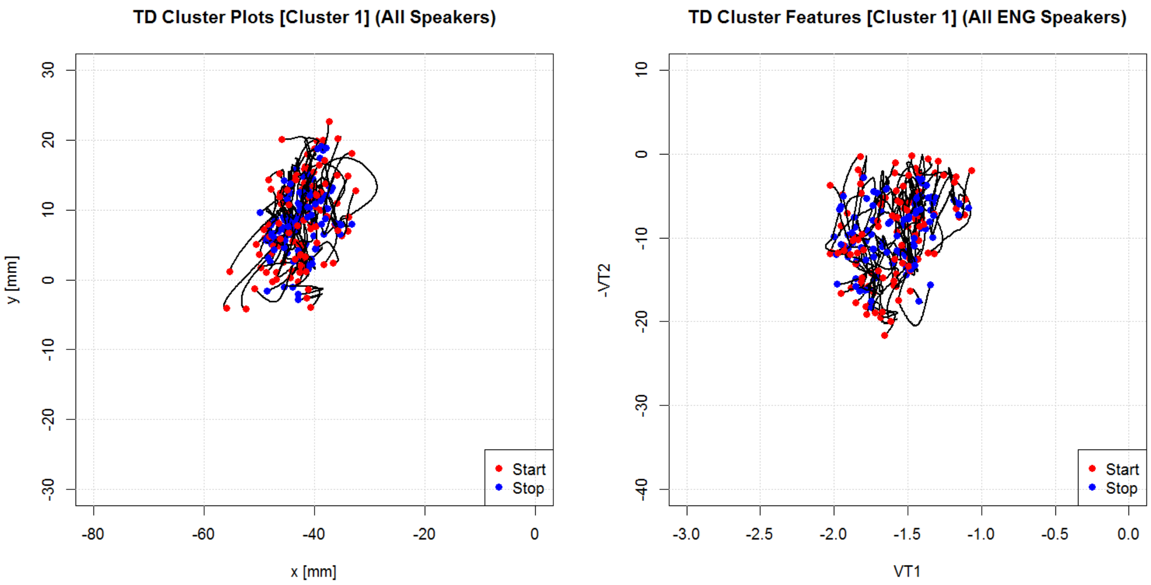
Figure 0.4 – *Consonant Cluster Euclidean and Feature Space Representations*



Recall that the conversion to feature space involves scaling the x sensor positions. Given a conversion by scaling, it makes sense that the horizontal locations of the trajectory points shown in Figure 0.4 relative to each other are maintained. Meanwhile, the vertical conversion of tongue sensor positions involves re-referencing the positions to the speaker's palate. This means that the conversion has the effect of vertically stretching and compressing the trajectories at different horizontal locations. In the example shown in Figure 0.4, the locations of the trajectories relative to each other are barely affected by the feature space conversion.

Figure 0.4 also displays an issue that is prevalent across all speakers, consonant clusters, and sensor positions: The different repetitions of the same cluster have very different movement trajectories in many cases. The figure shows trajectories of both small and large magnitudes, and multiple movement directions. Figure 0.5 shows the trajectories of the same cluster, but for all 20 English speakers.

Figure 0.5 – *Cluster /nd/ Euclidean and Feature Space Representation (All English Speakers)*



When the repetitions of the cluster are viewed for several speakers at once, the issue is further highlighted. While the plot is very busy, it can be observed that the trajectories start and end in varied locations. Figure 0.5 also displays the fact that the trajectories are not horizontally aligned. This means that despite horizontal normalization, the trajectories are still located at various horizontal positions. This complicates the process of modeling position of the trajectory when forming a kinematic template for the cluster.

Observation of trajectory plots of individual speakers suggests that in general, each speaker moves their articulators differently to produce the same consonant cluster. This was unexpected, but there are a number of reasons why this variation in the data may have occurred. One contributing factor may be lack of consistency in cluster labeling in the EMA-MAE dataset. As discussed in section 2.2, the cluster labels were created by trained staff of the Marquette Speech and Swallowing lab through observation of speech audio signals and their corresponding spectrographs. In many cases, the cluster boundaries may be difficult to identify (partially due to coarticulation effects, discussed in section 2.1). This can result in different repetitions of the same cluster having different start and end points in the trajectory.

Another possible contributor to the differences in cluster repetitions, (specifically, across speakers) is the feature space conversion (specifically, the vertical conversion). Each speaker has a different palate shape. This means that even if two speakers produced the exact same movement in Euclidean space, these trajectories in feature space would still be slightly different. However, while each speaker’s palate is different, the general shape of the palates are still fundamentally the same. In other words, the palate shapes are not different enough to significantly vary the trajectory shapes. Plotting of several speakers' cluster data produced similar results to those of Figure 0.4, supporting the idea that the feature space conversion does not significantly affect the overall trajectory of a cluster pronunciation. Meanwhile, the relationship between the tongue and palate positions remains important for the information it provides about the vocal tract shape (as discussed in section 2.3.2). For these reasons, the feature space conversion is maintained in the analysis of consonant clusters.

The significant variation in trajectories of consonant cluster pronunciations both within and across speakers makes it impossible to form an accurate model of the movement pattern to represent all English speakers. This is a serious setback, as a primary goal of this research is to form an English model of consonant clusters to evaluate results of acoustic-to-articulatory inversion, in addition to providing meaningful feedback to pilot study participants. However, while the overall movement trajectories are significantly varied, several individual characteristics of cluster pronunciation may be more consistent within and across speaker groups. These characteristics include magnitude (the length of the movement trajectory), the speed of the articulators, the movement pattern of the articulators, and the duration of the cluster pronunciation. Together (and individually), these features may provide insight into the similarities and differences in cluster production both within and between speaker groups. The study of these features in the interest of distinguishing AE and MAE speaker groups is henceforth called *MSTD (Magnitude, Speed, Trajectory Pattern, Duration) analysis.*

As previously mentioned, the magnitude of a cluster pronunciation is measured as the path length of the trajectory (Equation (0.1)). The duration of a cluster pronunciation is simply measured as the time between the onset and offset times of the cluster

(Equation (0.3)). The speed of a cluster pronunciation is a waveform calculated as the point-to-point distance between sensor positions, divided by the point-to-point time difference (Equation (0.2)). In order to remove high frequency variation and emphasize the overall shape of the curve, each speed waveform is then passed through a digital low pass filter with a cutoff frequency of 25 Hz. Given that this feature focuses specifically on the speed of the articulators, each speed waveform is also time normalized in order to factor out durational differences. The time normalization is performed through interpolation; each waveform was interpolated such that it consists of 100 points. This allows for point-by-point comparison of speed curves within and across speakers.

$$mag = \sum_{i=1}^{n} \sqrt{x_i^2 + y_i^2} \tag{0.1}$$

$$speed_i = \frac{\sqrt{(x_{i+1} - x_{i-1})^2 + (y_{i+1} + y_{i-1})^2}}{t_{i+1} - t_{i-1}} \tag{0.2}$$

$$dur = t_{Off} - t_{On} \tag{0.3}$$

In equations (0.1)-(0.3), *mag* is the trajectory magnitude, *dur* is the duration, *n* is the number of data points in the trajectory, *i* is the current index, *t* is the time, and “on” and “off” correspond to the onset and offset times of the cluster pronunciation (respectively).

The trajectory pattern calculation involves multiple steps. First, the trajectory is translated such that the minimum x and y values are both 0. Next, both the x and y dimensions of the curve are [0,1] normalized using the maximum x and y values of the trajectory. This serves as a magnitude normalization, providing each trajectory with the same scaling. Next, the trajectory is translated again in order to place the starting point at the origin. This defines a common starting point for all movement patterns, which allows for position-independent comparison across cluster repetitions. Finally, the trajectory is time normalized using the same interpolation applied to the speed curves. Figure 0.6, which shows plots of both the feature space representations (left) and extracted movement patterns of a cluster (right) for a single speaker, demonstrates the effects of this transformation.

Figure 0.6 – *Trajectory Pattern Extraction*

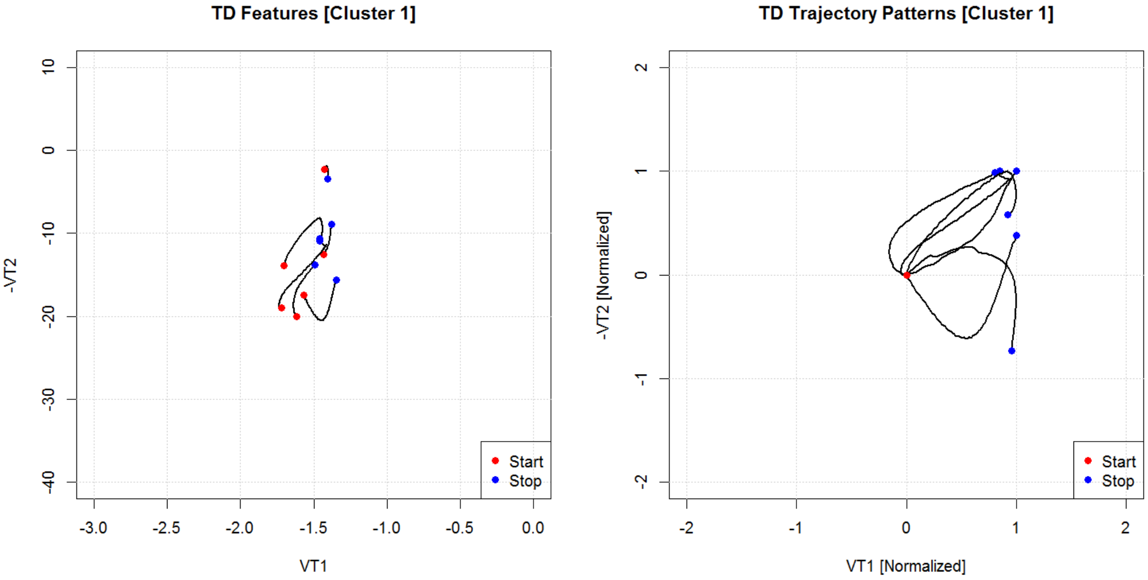
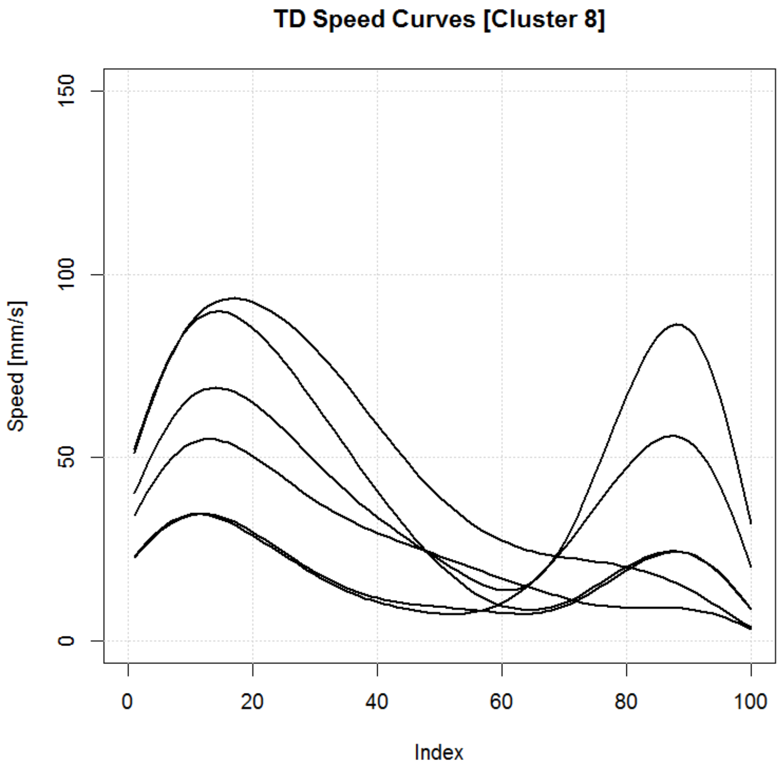


Figure 0.6 shows that the extraction process places the all of the trajectories on the same scale. Also note that each trajectory begins at the same point. These transformations disregard magnitude and locational differences in order to focus on the directionality of the articulator throughout the pronunciations of the cluster.

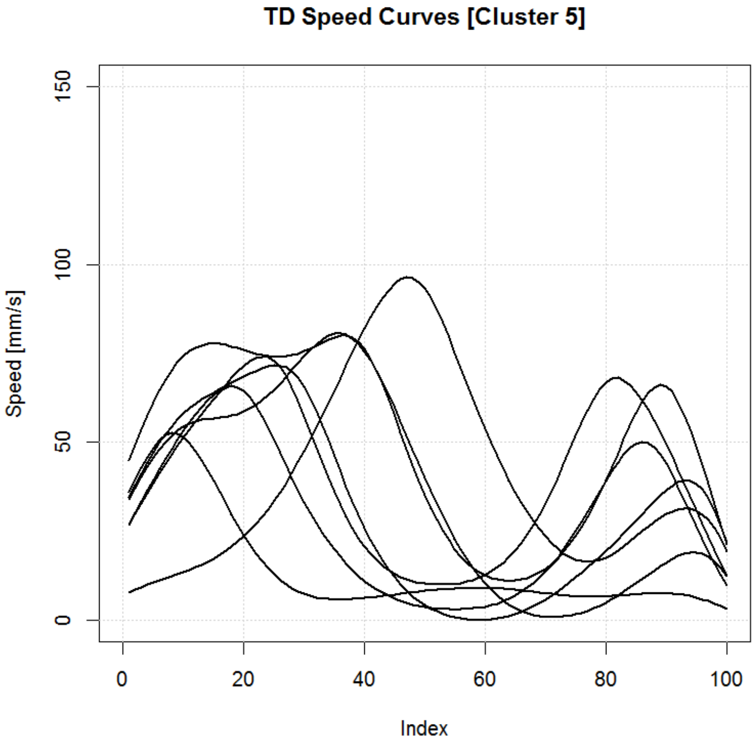
The methods of obtaining the speed and trajectory patterns have a number of strengths and weaknesses. As previously mentioned, interpolation of the curves allows for curves of different durations to be compared, with a focus on the feature of interest (whether it be speed or movement pattern). This method of normalization assumes that a speaker's speech characteristics scale linearly with duration of the pronunciation (for example, "this speaker always places the tip of their tongue on their palate in the middle of the pronunciation of cluster x, regardless of how long it takes them to pronounce the cluster."). Depending on the speaker and/or cluster, this may be either a strength or weakness of the method. For a speaker that generally speaks each part of the cluster at the same portions of pronunciation, regardless of duration, these curves will be horizontally aligned:

Figure 0.7 – *Speed Curves [Horizontally Aligned]*



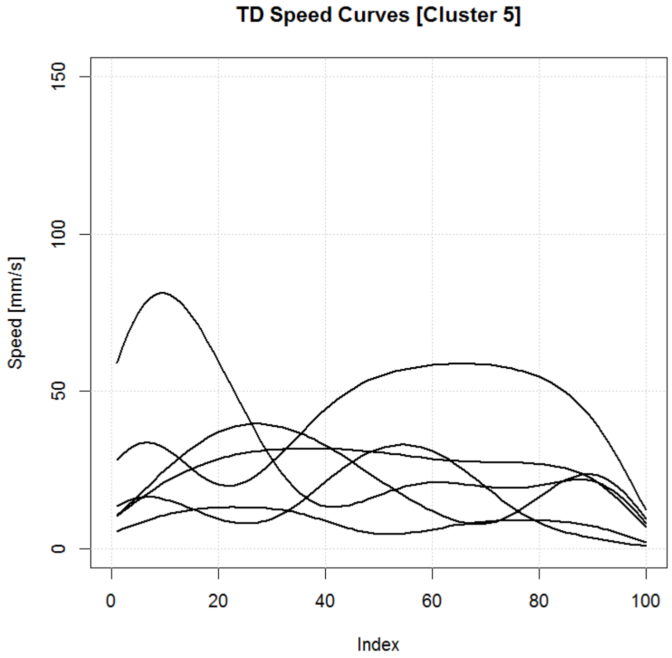
For speaker with varying pronunciation patterns regardless of duration, the speed curves will likely have horizontal shifts at some locations:

Figure 0.8 – *Speed Curves [Horizontally Misaligned]*



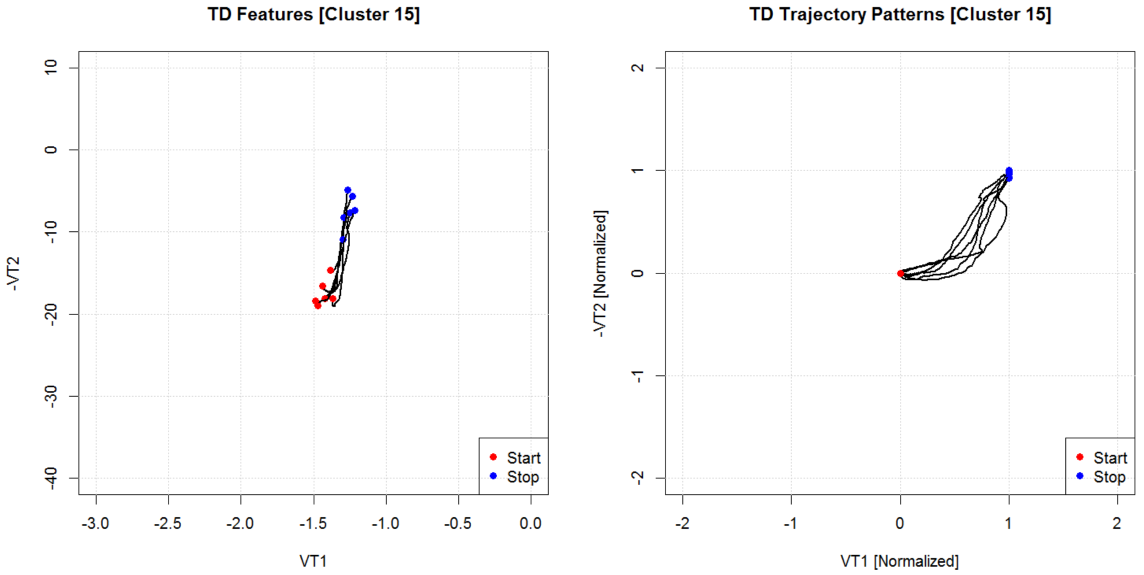
In the case of Figure 0.8, the results of a point-by-point comparison of the data would suggest that the curves are fairly different, despite the similar shapes. Meanwhile, if the curves were shifted for maximal alignment, information about the sections of the pronunciation where certain events (such as a peak in the curve) take place is lost. It is also worth noting that the plots of Figure 0.7 and Figure 0.8 correspond to very special cases where a speaker's cluster trajectories were fairly similar across repetitions. In most cases, the trajectories aren't similar, and the speed curves are both misaligned and shaped differently:

Figure 0.9 – *Speed Curves [Misaligned and Misshapen]*



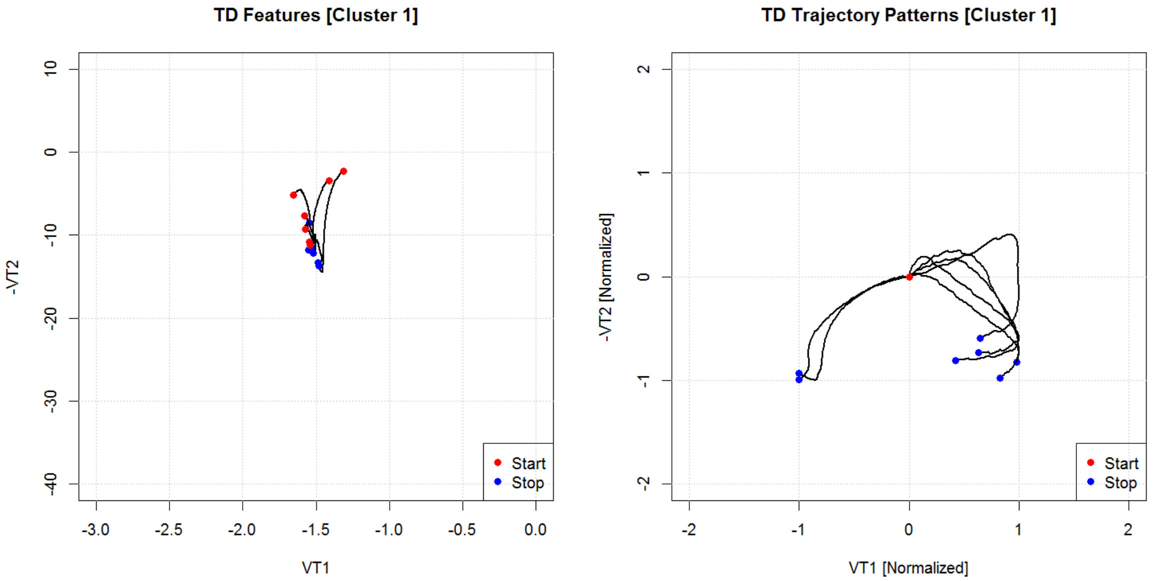
The [0,1] normalization scheme used for trajectory pattern extraction is also accompanied by number of strengths and weaknesses. Given that the method has the effect of “stretching”/”compressing” the trajectories such that they are all the same size, it highlights the similarity between curves of different magnitudes but similar shape and directionality:

Figure 0.10 – *Trajectory Extraction [Similar Shape, Different Magnitudes]*



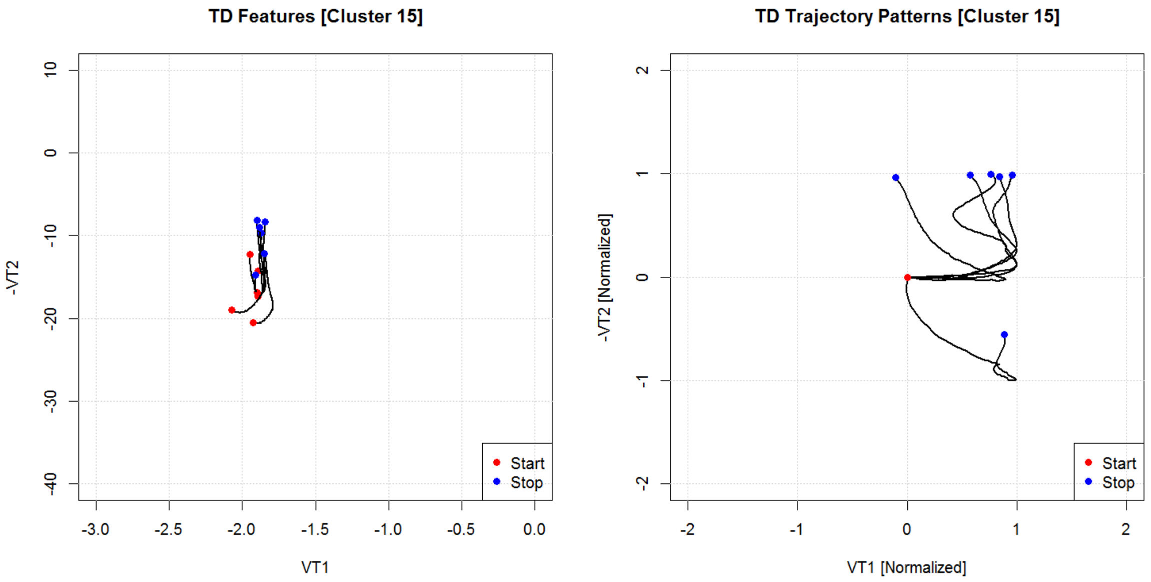
Moving the starting points of the trajectories to the same starting point further highlights similarity in directionality of cluster repetitions:

Figure 0.11 – *Trajectory Extraction [Similar Directionality, Different Locations]*



A weakness of the trajectory extraction method is the fact that it may accentuate the differences between trajectories in close proximity to each other:

Figure 0.12 – *Trajectory Extraction [Similar Locations]*



In the case shown in Figure 0.12, factoring out location and scaling the trajectories actually makes the overall trajectories less similar. However, the movement pattern is only one of four evaluation metrics. The combination of these metrics will form a more complete evaluation of the consonant cluster.

Note that location is not an included metric of the MSTD analysis. As shown in Figure 0.5, cluster location varies significantly across speakers and repetitions. With no consistent cluster placement to be identified, the analysis instead focuses on the other aspects of the cluster formation.

3.5.2.1 *MSTD Analysis Details*

After extraction of cluster data and conversion to feature space, the MSTD information was calculated. For each individual consonant cluster, the corresponding magnitude, speed curve, movement pattern, and duration were extracted as well (using the methods described in section 3.5.2.). Using this information, the following features were calculated:

- The mean magnitude of all repetitions of each cluster for each speaker and sensor, as well as the mean duration of each cluster production.
- A mean speed curve for each sensor of each cluster for every speaker, formed through the point-by-point averaging of the speed curves of all repetitions of each cluster.
- A mean trajectory pattern for each sensor of each cluster for every speaker, formed through the point-by-point averaging of the trajectory patterns of all repetitions of each cluster.
- Mean MSTD parameters for all AE speakers, as well as MAE speakers (formed through averaging across all speakers of the individual speaker groups).
- Deviation of MSTD parameters for each speaker, as well as the over deviation of MSTD parameters for each language and gender group (discussed later in this section).
- Mean deviation of MSTD parameters for all AE speakers, as well as MAE speakers (discussed later in this section).
- ANOVA comparing the deviation of MSTD parameters across language and gender groups (discussed later in this section).

For magnitude and duration, the deviation for each cluster is calculated as the standard deviation of the respective values across all repetitions (Equations (0.4) and (0.5)). For speed, the deviation is calculated as the sum of the point-by-point standard deviations across all repetitions of the cluster (Equation (0.6)). For trajectory pattern, the deviation is computed as the sum of the point-by-point standard deviations of both the x and y dimensions across all repetitions of the cluster.

$$mdev = SD(mvec) \tag{0.4}$$

$$ddev = SD(dvec) \tag{0.5}$$

$$sdev = \sum_{i=1}^{100} SD(svec_i) \tag{0.6}$$

$$tdev = \sum_{i=1}^{100} (SD(xvec_i) + SD(yvec_i)) \tag{0.7}$$

In equations (0.4**)**-(0.7**)**, the *dev* values correspond to the MSTD deviations (where the first letter indicates the corresponding MSTD feature). *mvec* and *dvec* correspond to the vectors containing the magnitude and duration (respectively) of each repetition of a given cluster for the current speaker. *svec$_i$*, *xvec$_i$*, and *yvec$_i$* correspond to vectors containing the speed, x trajectory position, and y trajectory position (respectively) of all repetitions of a given cluster at the *current index i*. Recall that the speed and trajectory features are curves consisting of 100 points. The index *i* specifies the current point, across all repetitions of the cluster.

In order to obtain deviational information about a cluster's features for a given speaker, the speaker will need to have been recorded producing the cluster multiple times. There are several clusters in the EMA-MAE dataset that speakers are recorded to have only produced once or twice. In these cases, there is not enough information to obtain a good estimate of deviation of production of the cluster. Therefore, when it comes to individual speakers, MSTD deviation calculations were only made for clusters that have at least 3 repetitions per speaker in the dataset. Table 0.2 shows the clusters that fall into this criteria.

Table 0.2 – *Consonant Clusters with 3+ Repetitions per Speaker*

Cluster ID	Cluster	Cluster Word
1	nd	find
3	kl	clone
5	ld	cold
7	kr	crick
8	kw	queen
15	lz	falls
21	ldz	fields
35	gr	green

The ANOVA that compares the MSTD deviation across language and gender groups (ENGM, ENGF, MANM, MANF) was performed for each of the clusters in Table 0.2. While only the clusters shown in this table had their MSTD deviational information studied on an individual speaker level, MSTD deviational information was also calculated for all speakers across each of the 4 language-gender groups (as well as the two language groups: ENG and MAN).

3.5.2.2 Comparison Expectations

As previously discussed, it is important to analyze the relationship between expectations and actual data in order to determine an effective approach to template creation. However, attempting to predict the behavior of the consonant cluster data is a largely difficult task. This is especially true given the highly variable nature of the data extracted from the EMA-MAE dataset, as discovered in the upper level of section 3.5.2. To start, the variability of MSTD parameters of the AE speakers (both individually and across speakers) were expected to be lower than those of MAE speakers. This is due to the fact that AE speakers are producing sounds that come naturally to them, while MAE speakers are attempting to produce sounds that, to some degree, are still new to them.

While the conversion to feature space performs a horizontal normalization, the vertical sensor positions are simply re-referenced. This means that while the feature space y sensor positions give more detailed information about the vocal tract shape, there is no vocal tract size normalization in the y direction. As discussed in section 2.1.1, men usually have larger vocal tracts than their female counterparts. Given this fact, male speakers may have larger magnitudes of movement to produce the same clusters as female speakers. The significant variety observed among the consonant cluster plots prevents meaningful expectations for cluster speeds and trajectories from being formed.

As discussed in section 2.1.2, initial and final consonant clusters do not

Mandarin Chinese. As a result, MAE speakers may tend to either

consonant of the cluster or create an extra syllable through the

vowel. While the removal of a consonant in the cluster would

magnitude and duration of the cluster, the attachment of a reduced

effectively be factored out through the cluster labeling process.

possible for coarticulation to occur between the final consonant

attached reduced vowel. This coarticulation could possibly be

OF CLUSTER, AND LABELED AS SUCH. INITIAL AND FINAL CONSONANT CLUSTERS

THE CLUSTERS ANALYZED IN THIS WORK (SEE

Appendix A for a listing of all clusters). Therefore, these characteristics of Mandarin-accented English shaped expectations regarding MAE speaker data. The information above creates four primary cluster pronunciation expectations for MAE speakers, all within reason:

- The speaker produces the cluster without removing any consonants or attaching any reduced vowel sound. The magnitude and duration of the cluster are similar to those of AE speakers.
- The speaker produces the cluster while removing the final consonant. The magnitude and duration of the cluster are both reduced.
- The speaker produces the cluster, but also attaches an additional reduced vowel. The cluster labels effectively factor out the attached vowel. The magnitude and duration of the cluster are similar to those of AE speakers.
- The speaker produces the cluster, but also attaches an additional reduced vowel. The cluster labels capture part of this attached vowel as a result of coarticulation effects. The magnitude and duration of the cluster are greater than those of AE speakers.

A special case is introduced to the discussion above when the consonant cluster in question includes a voiced stop. As discussed in section 2.1.2, stop voicing contracts in Mandarin Chinese are indicated by aspiration, as opposed to the voiced stops of American English. This tendency to aspirate leads to weak voicing of voiced stops among MAE speakers. This may result in reduced cluster magnitude and duration.

3.5.2.3 Results

After the mean cluster magnitudes and durations were captured for each speaker, they were observed to identify trends between (1) male and female AE speakers, (2) male and female (MAE) speakers, and (3) AE and MAE (cross-gender) speakers. Table 0.1 shows, for these three group comparisons and each sensor, the number of clusters where the first of the two groups (male AE, male MAE, and AE) had a greater average magnitude and the average (unsigned) percent error in magnitude between the two groups. These calculations were repeated for the duration parameter as well. The first metric provides insight into how often male speakers have greater cluster magnitudes than female speaker (and consequently, vice versa) across all clusters for both language groups and across language. Meanwhile, the second metric provides the average amount of difference between the two groups, as a percentage of the first group's size. Note that the percentage of clusters that had greater magnitudes for the first group is also shown in Table 0.1. Table 0.2 shows, for the group comparisons shown in Table 0.1, the number of clusters where the first group's magnitudes were greater than the second group, across *all* sensors (how many clusters maintained the relationship across sensors).

Table 0.1 – *Consonant Clusters Magnitude Comparisons*

TD		
Groups (G1 vs G2)	**# Clusters w/ G1>G2**	**% Error**
ENGM vs ENGF	23 (52%)	14.8
MANM vs MANF	33 (75%)	20.2
ENG vs MAN	36 (82%)	22.8

TB		
Groups (G1 vs G2)	**# Clusters w/ G1>G2**	**% Error**
ENGM vs ENGF	24 (55%)	18.5

MANM vs MANF	36 (82%)	20.7
ENG vs MAN	22 (50%)	23.4

LS		
Groups (G1 vs G2)	**# Clusters w/ G1>G2**	**% Error**
ENGM vs ENGF	16 (36%)	17.4
MANM vs MANF	22 (50%)	14.7
ENG vs MAN	32 (72%)	25.5

Table 0.2 – *Consonant Cluster Magnitude Comparisons: Common Clusters*

Groups (G1 vs G2)	**# Clusters In Common**
ENGM vs ENGF	5 (11%)
MANM vs MANF	15 (34%)
ENG vs MAN	19 (43%)

Table 0.1 displays a trend among the cluster magnitudes for the tongue sensors: The percent error between AE males and females is lower than the error between MAE males and females, which in turn is lower than the error between AE and MAE speakers. In theory, these results make sense. If AE speakers are producing the same sounds (as they are assumed to produce English correctly), it makes sense that their magnitude differences would be lower than those of non-native speakers. Meanwhile, a degree of unpredictability is introduced to the cluster production of MAE speakers due to L1 effects (as discussed in section 3.5.2.2). The fact that the final comparison (ENG vs MAN), which is a cross-language comparison, produced the greatest percent error is also within expectations, for the same reasons mentioned above.

Note that, in regards to the number of clusters where a single group had greater average magnitudes than the other, the results are largely varied. For the lip sensors, the number of clusters where AE males and AE females had greater magnitudes is roughly split in half. Meanwhile, AE males had greater lip magnitudes than AE females for only 36% of the clusters. These results reject the idea that males may have greater articulator movement magnitudes than females due to vocal tract size differences. Table 0.2 shows that only 5 consonant clusters maintained the magnitude relationships between AE males and females that were displayed in Table 0.1. When removing the lip sensor from consideration (as it is a different articulator), this number increases to 13. While the tongue sensor results maintain their relationships in a few more cases, the results suggest that in most cases, there are no magnitude trends between sensors. Overall, this data suggests that for AE speakers, cluster magnitudes vary across consonant clusters and sensors, possibly independently of gender.

Across both tongue sensors, MAE males had greater cluster magnitudes than MAE females for a majority of clusters (75+% of the clusters for both sensors). Given a 20+% error on average between these speaker groups (which suggests that, on average, the magnitude differences were not trivial), this suggests that MAE males typically produce greater cluster magnitudes than MAE females for most clusters. For the lip sensor, the cases where MAE males and females produced greater magnitudes was split in half, and the percent error between the groups was significantly lower than those of the tongue sensors. This suggests the lip cluster magnitudes among MAE speakers are more similar to each other than tongue cluster magnitudes. When removing the lip sensor from consideration, the number of clusters that maintained these relationships increases from 15 (34% of all clusters) to 28 (64% of all clusters). This emphasizes the idea of greater similarity between tongue sensors magnitudes than between all sensors.

Table 0.3 shows the same group comparisons, but for cluster duration instead of magnitude.

Table 0.3 – *Consonant Cluster Duration Comparisons*

Groups (G1 vs G2)	# Clusters w/ G1>G2	% Error
ENGM vs ENGF	16 (36%)	10.0
MANM vs MANF	6 (14%)	12.1
ENG vs MAN	26 (59%)	8.2

For AE speakers, the durational relationship between males and females did not match the magnitude relationships for any of the 3 sensors. While the number of clusters where AE males had greater lip sensor magnitude and duration than AE females were both 16, only 11 of the clusters were the same in these cases. This rejects the idea of a proportionality between magnitude and duration across clusters. MAE speakers also displayed no trends between cluster magnitude and duration. Given the challenges associated with MAE production of consonant clusters, this could be due to a number of factors (including a lack of correlation between cluster magnitude and duration in general).

The same group comparisons performed for cluster magnitude and duration were applied to cluster speed as well, but with a slight variation. The speed is represented as a time-normalized curve (as opposed to a single point). In order to represent these relationships as a single value, the speed comparisons represent the *average* difference between groups across all points of the curves:

$$Difference = \sum_{i=1}^{n} spd_1[i] - spd_2[i] \tag{0.1}$$

Where “1” and “2” correspond to the group number (for example, in the ENGM vs ENGF comparison, ENGM is 1 and ENGF is 2), *i* is the current index, and *n* is the total number of points (set to 100, as described in the upper level of section 3.5.2). Using these group difference values, the group relationship calculations were made. The results are shown in Table 0.4 and Table 0.5.

Table 0.4 – *Consonant Cluster Speed Comparisons*

TD		
Groups (G1 vs G2)	**# Clusters w/ G1>G2**	**% Error**
ENGM vs ENGF	29 (66%)	12.4
MANM vs MANF	40 (91%)	22.1
ENG vs MAN	26 (59%)	25.1

TB		
Groups (G1 vs G2)	**# Clusters w/ G1>G2**	**% Error**
ENGM vs ENGF	25 (57%)	17.2
MANM vs MANF	36 (82%)	25.8
ENG vs MAN	20 (45%)	23.4

LS		
Groups (G1 vs G2)	**# Clusters w/ G1>G2**	**% Error**
ENGM vs ENGF	25 (57%)	19.4
MANM vs MANF	36 (82%)	20.2
ENG vs MAN	27 (61%)	17.9

Table 0.5 – *Consonant Cluster Speed Comparisons: Common Clusters*

Groups (G1 vs G2)	**# Clusters In Common**

ENGM vs ENGF	8 (18%)
MANM vs MANF	31 (70%)
ENG vs MAN	11 (25%)

For both inner language group comparisons, the males of the groups, on average, had greater movement speed than females across all sensors for a majority of clusters. This could have indicated that male speakers tend to move their articulators more quickly to produce the same sounds as their female counterparts. However, there is also a significant amount of average percent error in each case. This may imply that in cases where females produced greater movement speeds than males, the differences between groups were not trivial. Furthermore, for AE speakers, this relationship is maintained across sensors for only 8 of the 44 clusters. This rejects any notion of male speakers generally having greater articulatory movement speeds than females.

WHILE A NUMBER OF ADDITIONAL CALCULATIONS WERE MADE DURING THE CONSONANT CLUSTERS, NONE OF THEM PROVIDED ADDITIONAL SIGNIFICANT SIMILARITIES AND DIFFERENCES IN CLUSTER PRODUCTION ACROSS GENDER, DETAILS REGARDING SOME OF THESE ADDITIONAL ANALYSES, INCLUDING THE PERFORMED ON THE DEVIATION OF MSTD PARAMETERS OF THE CLUSTERS FROM TABLE 0.2, MAY BE FOUND IN

Appendix A. The mean MSTD values of clusters are discussed in further detail in section 4.3, which covers the consonant cluster template creation process.

Overall, the consonant cluster analysis identified no significant trends across speaker groups, sensors, or clusters. The MSTD parameters generally varied with each combination of speaker, sensor, and cluster without no easily identifiable patterns. Given these results, the approach to building kinematic templates for consonant clusters should be chosen such that the template for each sensor and consonant cluster is built independently.

3.6 Contrastive Stress Extraction and Analysis

3.6.1 Contrastive Stress Extraction

All contrastive stress data was extracted from the sentence-level

in the EMA-MAE Dataset Manual [33]. Like consonant clusters, the

contrastive stress word is a dynamic movement is must be

There are 9 total contrastive stress words (see

Appendix A for the list of stress words), and each speaker spoke each word exactly two times. In the first pronunciation, the stress is placed on the first syllable, and the stress is placed on the second syllable during the second pronunciation. For each syllable, the x-y trajectory of the TD, TB, and LS sensors were extracted. This creates a working data set of 4 syllables (2 stress/un-stress pairings) per speaker.

3.6.2 Kinematic Data Analysis

While each contrastive stress word is pronounced twice by each speaker, the word is pronounced differently due to stress placement (meaning they are different words). This means that there is only one recording of each syllable of each word per speaker. Given such a small amount of data, it is impossible to form meaningful models of each contrastive stress word. This is not a problem, given that the focus of the analysis is the determination of the similarities and differences between AE and MAE speakers when producing stress. Given that fact, this analysis examines the characteristics of stress at a high level (across all words and speaker groups), as opposed to individual word analyses.

As described during the consonant cluster analysis (section 3.5.2), magnitude, speed, trajectory, and duration (MSTD) analysis can be an effective method of analyzing dynamic articulatory movement. Applying this analysis method to contrastive stress would involve defining these parameters for entire syllables instead of consonant clusters. However, as explained above, this analysis does not seek to model the pronunciation of stress words or syllables, but instead the characteristics of stress themselves. This eliminates any need for a trajectory (T) analysis. The need for a speed curve is also eliminated, but the general speed of a stress/un-stress pair provides significant information about articulatory speed when producing stress. The speed (S) parameter, for contrastive stress, replaces the speed curve of the consonant cluster analysis with a recording of average speed across a syllable's duration. This new MSTD analysis, with the removal of trajectory investigation and a focus on syllable level data, is called *MSD analysis.*

The primary focus of the MSD analysis is the determination of the effect of stress on the magnitude, speed, and duration of a contrastive stress syllable. Given that fact, the analysis is not focused on individual MSD values, but instead the relationships between these values in their stressed and un-stressed cases:

$$MR = Magnitude\ Ratio = \frac{Stressed\ M}{Unstressed\ M} = \frac{\sum_{i=1}^{n}\sqrt{x_{s_i}^2 + y_{s_i}^2}}{\sum_{i=1}^{n}\sqrt{x_{us_i}^2 + y_{us_i}^2}} \tag{0.1}$$

$$SR = Spd\ Ratio = \frac{Stressed\ S}{Unstressed\ S} = \frac{mean\left(\frac{\sqrt{\left(x_{s_{i+1}} - x_{s_{i-1}}\right)^2 + \left(y_{s_{i+1}} + y_{s_{i-1}}\right)^2}}{t_{i+1} - t_{i-1}}\right)}{mean\left(\frac{\sqrt{\left(x_{us_{i+1}} - x_{us_{i-1}}\right)^2 + \left(y_{us_{i+1}} + y_{us_{i-1}}\right)^2}}{t_{i+1} - t_{i-1}}\right)} \tag{0.2}$$

$$DR = Duration\ Ratio = \frac{Stressed\ D}{Unstressed\ D} = \frac{t_{s_{off}} - t_{s_{on}}}{t_{us_{off}} - t_{us_{on}}} \tag{0.3}$$

Where an *s* subscript denotes the stressed value (from the stressed version of the syllable being analyzed) of a parameter, and a *us* subscript denotes the unstressed value of a parameter. *t* refers to time, and the *on* and *off* subscripts refer to the onset of offset times (respectively) of the syllable.

3.6.2.1 MSD Analysis Details

After each syllable pronunciation was extracted from the dataset, the MSD parameters were calculated for each speaker. Using these values, the following features were calculated:

- For each speaker, an average stress/un-stress ratio for each MSD parameter for each contrastive stress paring, calculated from the two individual ratios from the pairings.
- The mean magnitude, speed, and duration ratios across the two language groups (ENG and MAN)
- ANOVAs comparing the average MSD parameters across the four language-gender groups (ENGM, ENGF, MANM, MANF).

3.6.2.2 Comparison Expectations

As previously discussed, it is important to evaluate expectations based on prior knowledge and research before attempting to model any data. The results of this analysis will potentially reform or update any previously considered approach to template creation. Stress is a very complicated topic in the context of Mandarin-accented English. The challenges associated with MAE production of English stress patterns are detailed in section 2.1.1. These challenges include the new concept of using stress to differentiate word meaning, the fact that unstressed English vowels are typically reduced (moving closer to the neutral schwa), and the difficulty of identifying where the stress should be placed in a word based on context.

Aside from the cross-language challenges, the actual contrastive stress words can present additional challenges in some cases. While unstressed vowels are often reduced in comparison to their stressed versions, some contrastive stress words in the EMA-MAE dataset have different vowels depending on the stress (see Table 2.4 for the list of all contrastive stress words). An example of this is the word *project*, which has different pronunciations of the vowel *o* (either /ah/ or /ow/) depending on where stress is placed. In this case, magnitude differences between the stressed and unstressed versions of the syllable could be purely due to the difference in vowel being produced (as opposed to a difference in speaking intensity, or other correlates of English stress). That being said, the /ow/ pronunciation of the vowel is still unstressed in this case, and therefore reduced and moved closer to the schwa by typical English speakers. Given this reduction, possible magnitude differences due to vowel pronunciation are likely significantly reduced.

Given all the factors discussed above, MSD parameters are expected to be similar across AE speakers, but varied across MAE speakers. Given the typical correlates of English stress (discussed in section 2.1.2), all 3 MSD parameters are expected to be greater in stressed syllables than unstressed syllables for AE speakers. The difficulty in identifying and replicating stress patterns may lead to reduced MSD ratios for MAE speakers.

3.6.2.3 RESULTS

Table 0.1 displays a table of the average MSD parameters across all AE speakers, and Table 0.2 displays the same for MAE speakers.

Table 0.1 – *Contrastive Stress MSD Parameters: English*

	M			***S***			***D***
Stress ID	**TD**	**TB**	**LS**	**TD**	**TB**	**LS**	
1	1.467	1.765	1.938	1.556	1.875	2.057	0.960
2	0.873	1.610	3.005	0.817	1.502	2.412	1.157
3	2.752	3.194	2.243	1.572	1.920	1.372	1.672
4	2.278	3.182	2.161	1.600	2.442	1.497	1.513
5	1.639	2.060	1.831	1.254	1.445	1.381	1.392
6	1.598	2.666	1.978	1.207	2.131	1.545	1.327
7	2.792	3.098	3.765	1.210	1.176	1.404	2.815
8	1.697	2.207	1.831	1.736	2.358	1.935	1.038

9	1.177	1.969	1.719	1.017	1.667	1.514	1.164

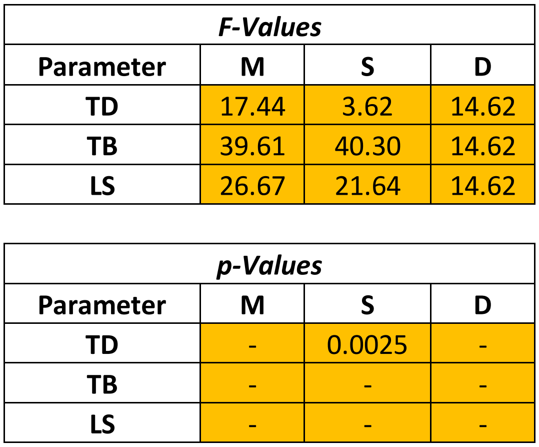
Table 0.2 - *Contrastive Stress MSD Parameters: Mandarin*

	M			***S***			***D***
Stress ID	**TD**	**TB**	**LS**	**TD**	**TB**	**LS**	
1	1.174	1.216	1.008	1.241	1.268	1.058	0.993
2	1.195	1.117	1.288	1.172	1.052	1.235	1.113
3	1.170	1.205	1.227	1.109	1.014	1.030	1.280
4	1.269	1.194	1.066	1.187	1.213	1.027	1.064
5	1.326	1.135	1.393	1.104	1.017	1.168	1.234
6	1.103	1.100	0.989	0.998	1.026	0.943	1.082
7	2.056	1.681	2.406	1.330	0.982	1.412	2.068
8	1.166	1.147	1.332	1.388	1.417	1.527	0.925
9	0.908	1.021	1.123	0.973	1.121	1.258	0.917

Table 0.1 shows that, as expected, magnitude, average speed, and duration of stressed syllables are almost always greater than those of unstressed syllables for AE speakers (shown by the fact that the stressed/unstressed ratios are usually greater than 1). This table shows different MSD results for each contrastive stress word. This implies that for AE speakers, the degree increase of magnitude, speed, and duration from unstressed to stressed syllables varies with each contrastive stress word. Meanwhile, Table 0.2 shows that MSD parameters for MAE speakers are much more similar then those of AE speakers across stress words. This implies that while AE speakers treat each stress word differently, MAE speakers tend to treat them more similarly.

Table 0.3 displays the results of the ANOVAs performed across the language-gender groups. Note that comparisons that yielded statistical significance (significant differences between at least two speaker groups) are highlighted. Also note that negligible values are marked with “-”.

Table 0.3 – *Contrastive Stress ANOVA Results*



F-Values			
Parameter	**M**	**S**	**D**
TD	17.44	3.62	14.62
TB	39.61	40.30	14.62
LS	26.67	21.64	14.62

p-Values			
Parameter	**M**	**S**	**D**
TD	-	0.0025	-
TB	-	-	-
LS	-	-	-

Note that the duration F-value is identical for all 3 sensors. Regardless of the sensor in question, pronunciation duration remains the same. Table 0.3 confirms the conclusion drawn from Table 0.1 and Table 0.2: AE speakers produce contrastive stress characteristics differently than MAE speakers. From those tables, it is known that this is due to variation in MSD parameters across clusters. While MAE speaker results are shown be different from those of AE speakers, MAE speakers are also shown to follow the trend of increasing magnitude, speed, and duration when producing stress in most cases. It must be stressed that these results are based on single stressed/unstressed pairs for each syllable per speaker. A significant increase in contrastive stress data would increase the reliability of these results.

Given such a small amount of data and large amount of native English speaker variation across contrastive stress words, any cross-word representation of MSD stress patterns may not provide meaningful information. Aside from the overall trend of increase, general MSD stress characteristics cannot be reliably modeled.

3.7 SUMMARY

Chapter 3 has covered the analysis of EMA-MAE data. This includes the investigation of formant and kinematic data for vowels, as well as kinematic data for consonant clusters and contrastive stress pairs. The primary purpose of this Prior to data processing, the subset of EMA sensors used for analysis in this work (section 3.1) and the conversion from Euclidean space to articulator feature space (section 3.2) were discussed in detail. This was followed by a walkthrough of the statistical analysis techniques used to implement the data analyses (section 3.3).

Vowels were the first of the three phonetic categories to be studied (section 3.4). This started with an analysis of the relationship between formant frequencies and sensor positions. While it is clear that a relationship exists between these formant frequencies and sensor positions, the nature of the relationship could not be quantified in a meaningful way. As a result, no formant analysis techniques are applied to the vowel template creation process. The spread of sensor positions for a single vowel seemed to confirm that several articulatory configurations can result in the same vowel pronunciation. This suggests that vowel kinematic templates should specify a range of possible positions, as opposed to a single configuration.

The vowel analysis was followed by the analysis of consonant clusters. Plotting of several consonant cluster repetitions indicated large amounts of variation in size and directionality of cluster productions both within and across speaker groups. In order to characterize the individual aspects of each cluster, the MSTD (magnitude, speed, trajectory, and duration) analysis was introduced. This analysis revealed that across all MSTD parameters, there were no noticeable trends across speaker groups, consonant clusters, or sensor positions. The fact that the MSTD parameters for a given cluster and sensor were seemingly independent of others in most cases indicates that each cluster template should be built independently.

Finally, the contrastive stress data of the speakers was analyzed. Given such a small amount of data, meaningful models of each contrastive stress word could not be reliably created. This analysis instead focused on the general correlates of stress, and how these correlates varied across speaker groups. The started with the introduction of the MSD analysis, a variation of MSTD analysis that examines the relationship between the stressed and unstressed magnitude (M), average speed (S), and duration (D) of each word pronunciation. In general, the MSD parameters were shown to increase when applying stress. However, while the amount of increase was fairly consistent across MAE speakers, the amount of increase in parameters varied across stress words for AE speakers. This significantly complicates the kinematic modeling process for contrastive stress pairs.

4 KINEMATIC TEMPLATE CREATION

This chapter addresses the primary purpose of the research discussed in this thesis: The creation of kinematic templates modeling English vowels, consonant clusters, and stress characteristics. As discussed in chapter 1, the Marquette Speech and Signal Processing and Marquette Speech and Swallowing labs are conducting a pilot study in order to determine the feasibility of using acoustic-to-articulatory inversion for pronunciation training. The kinematic templates are needed in order to evaluate the results of acoustic-to-articulatory inversion of a pilot study participant's speech.

In chapter 3, the formant and kinematic data for all speakers of the EMA-MAE corpus (see section 2.2 for details) was extracted and analyzed. This analysis:

- established the relationship between formant frequencies and corresponding articulator positioning.
- studied the similarities and differences in articulation both within and between native American English (AE) and Mandarin-accented English (MAE) speaker groups.
- evaluated expectations for articulation based on prior research and knowledge of speech production and language learning.

With the detailed analysis completed, the information obtained may be applied to the creation of the kinematic templates.

4.1 DEFINING THE *STANDARD ENGLISH SPEAKER*

Before kinematic templates can be created, the best way to model English speech must be determined. Given the available data, this model must be based on the English kinematic speech samples from the EMA-MAE corpus. As previously discussed in section 2.2, the EMA-MAE corpus contains data for 20 American English speakers (10 male speakers and 10 female speakers). This presents two primary methods of defining English kinematic models: Selecting a single speaker that best represents a typical English speaker for modeling, or forming models through some combination of multiple speakers. In the data analysis of chapter 3, all English speakers are assumed to be speaking English correctly throughout the EMA recording process. Without a method of determining if a single speaker represents a typical English speaker better than any other speaker, the decision was made to develop the kinematic templates using data from all 20 English speakers.

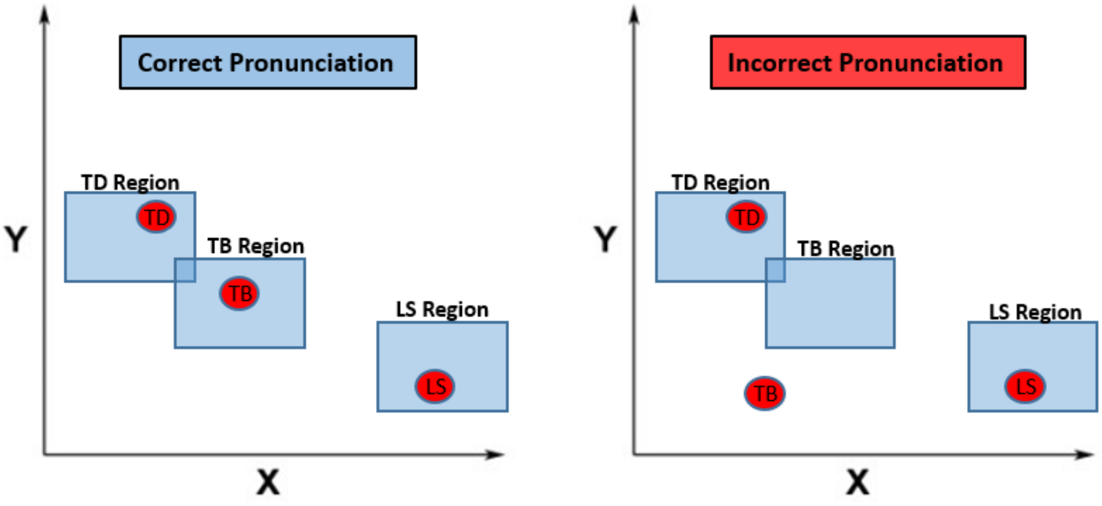
4.2 Vowel Templates

4.2.1 Handling the "One-to-Many" Problem

As shown by the results of the kinematic data analysis in section 3.4.3, across AE speakers, the repetitions of each vowel cover a wide range of sensor positions. This analysis confirmed that several different articulatory configurations can produce the same acoustic result. It is clear that the kinematic template for a single vowel must model not only a single combination of articulator locations, but a wide array of location combinations that all produce the same vowel. However, the sensor positions for a single vowel vary to the point of disagreeing with previously established concepts (namely, the relationship between formant frequencies and tongue positioning) and overlapping with the range of adjacent sensors. Prior to this discovery, a solution to the template creation problem under consideration was the establishment of mutually exclusive regions of the vocal tract (in the midsagittal plane) that could be used to distinguish vowels being produced (essentially creating a sensor variation of the vowel quadrilateral of Figure 2.4). This would lead to a classification system that could be used to provide meaningful feedback to pilot study participants.

A similar potential solution to the “sensor quadrilateral” may be formed using the overall range of each of the three sensors (TD, TB, and LS) for each vowel, but the different sections of the vocal tract would no longer be mutually exclusive. With sensor positions varying significantly, there would be a great deal of overlap between these sections. However, with three sensors in consideration, the kinematic model for a single vowel becomes more specific in that there are additional constraints on articulator positioning. In other words, for a pronunciation to be considered correct, all 3 sensors would need to fall within their respective regions:

Figure 4.1 – *Region-Based Vowel Template*



Given a structure similar to that of Figure 4.1, single point targets (likely to be the average locations of each sensor for AE speakers) may be established while also providing a range of tolerance for acceptable pronunciation. The primary concern then becomes the best way to define this “tolerance region” for each sensor.

Given that the kinematic templates are designed to model native English speech, it makes sense that the regions be defined using only data from the AE speakers of the EMA-MAE corpus. All AE speech data is assumed to be of correct native English pronunciation, so all of the vowel data for all AE speakers are considered in the development of the regions. There are a number of ways to enclose a set of data, but the enclosure must also model an entire population. One potential definition of the template region is the convex hull of each sensor’s data across all AE speakers. The idea of using convex hulls is rooted in the assumption that the average native English speaker’s data would fall within the same region as the EMA-MAE corpus members. While convex hulls can accurately enclose the AE data to define a “valid pronunciation region”, they do not account for spatial trends in the data (for example, a higher concentration of data points in a certain location) and are highly sensitive to outliers (with all AE data assumed to be correct, any data point that should be considered an “outlier” would be the result of measurement error).

Another potential solution is the standard deviational ellipse (SDE). The SDE is an ellipse whose dimensions are based on two dimensional deviation of the data, is centered on the mean, and is directed in the orientation of the data [51]. Given that the ellipse is created using overall deviational data, it both takes spatial trends into account and is much less sensitive to outliers. The main weakness of the SDE for this application is that while it represents the statistical parameters of the AE data, it does not have a consistent scaling mechanism. An ellipse whose dimensions match the two dimensional standard deviation of the data will always cut off a number of data points in the AE distribution (because several points always fall outside the standard deviation of a set of data). Given the assumption that these points correspond to correctly pronounced speech, this ellipse would cut off a section of the acceptable region of pronunciation. Even if the ellipse is scaled (for example, by a factor of 2) for one vowel/sensor, this scaling factor will not enclose amount of data for all vowels and sensor data distributions (in other words, a scaling factor that works well for one vowel may not be appropriate for another). Also, there is the fact that not all data distributions take the shape of an ellipse. Meaning, each ellipse would have sections that are unoccupied by any data points, but are identified as acceptable regions for correct pronunciation. However, it is also true that with only 20 AE speakers, the templates will be based on a very small set of data. If the EMA-MAE corpus consisted of 100 or even 1000 speakers, it may be discovered that some AE speakers typically occupy these areas during articulation as well.

Modifications to the SDE method of defining template regions that aid in addressing its weaknesses include (1) the creation of region “tiers” that provide additional information on the proximity to the target articulator positioning, and (2) consistent ellipse scaling across different sensors and vowels. The aforementioned tiers can be established via concentric ellipses. The center of these ellipses (which would be equal to the mean of the AE speaker data for a given vowel and sensor) would represent the absolute target for correct pronunciation, while the ellipses provide additional information regarding the proximity to the target. The innermost ellipse would enclose the area with the primary concentration of sensor positions (the region of "accurate" articulatory positioning) across vowel repetitions, while the outermost ellipse represents the absolute boundary of acceptable positioning.

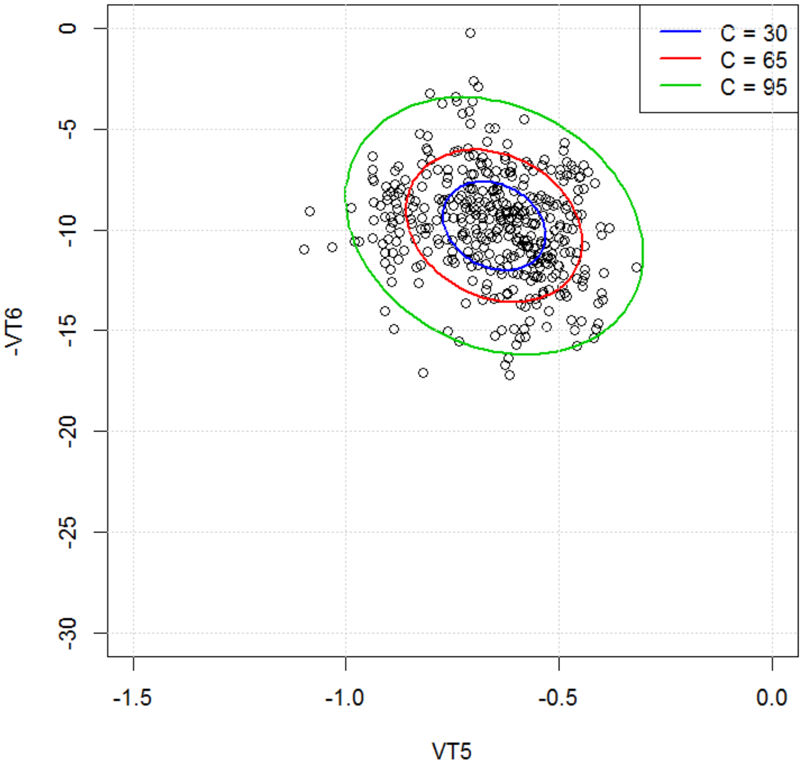
In order to implement these tiers with consistent scaling across vowels, a proper metric is needed. Possible metrics include standard deviation (for example, ellipses sized at one, two, and three standard deviations from the mean) and percentage of population enclosed (for example, ellipses that enclose 10%, 50%, and 90% of all data). While standard deviations provide a consistent scaling scheme, they do not provide specific information regarding the actual locations of the AE speaker data used to form the ellipses. Using percentage of population for ellipse scaling accounts for location of sensors, but is more sensitive to outliers and can correspond to several ellipse sizes at once (further complicating the process of proper scaling). As a semi-compromise between these two scaling methods, the metric chosen for template creation was confidence level. In using this metric, the SDE was replaced with the *confidence ellipse* as the method of enclosing the data.

4.2.2 TEMPLATE ELLIPSES

The vowel templates are created using 3 concentric ellipses. Each ellipse corresponds to a certain tier of proximity to the target articulator position. The innermost ellipse and the region enclosed by it correspond to the *correct* region. This describes the area that, in being in close proximity to the average location occupied by AE speakers, is considered to correspond to correct pronunciation. Again, the notion of using an entire region of values to represent the correct position is founded on the idea that several articulatory configurations can produce the same acoustic result. The middle ellipse and the region between it and the innermost ellipse correspond to the *most likely correct* region. This region corresponds to the areas that start to deviate from the average AE speaker position, but given the dispersion of the data, still falls within an acceptable range. In some cases, this region may be considered an extension of the correct region, rather than a separate tier. The outermost ellipse and the area between it and the most likely correct region correspond to the *needs improvement* region. This describes areas that may or may not correspond to correct pronunciation, and should be improved on before being considered acceptable.

The R package *ellipse* contains several functions for calculating and plotting ellipses, including a function that calculates the points of a confidence ellipse using input data and a specified confidence level. Through repeated plot tests, the confidence levels to represent the *correct*, *mostly correct*, and *needs improvement* regions were chosen to be 30%, 65%, and 95% (respectively). Figure 4.2 displays these concentric ellipses for the (feature space) TB position of all repetitions of the vowel /ae/ for all AE speakers.

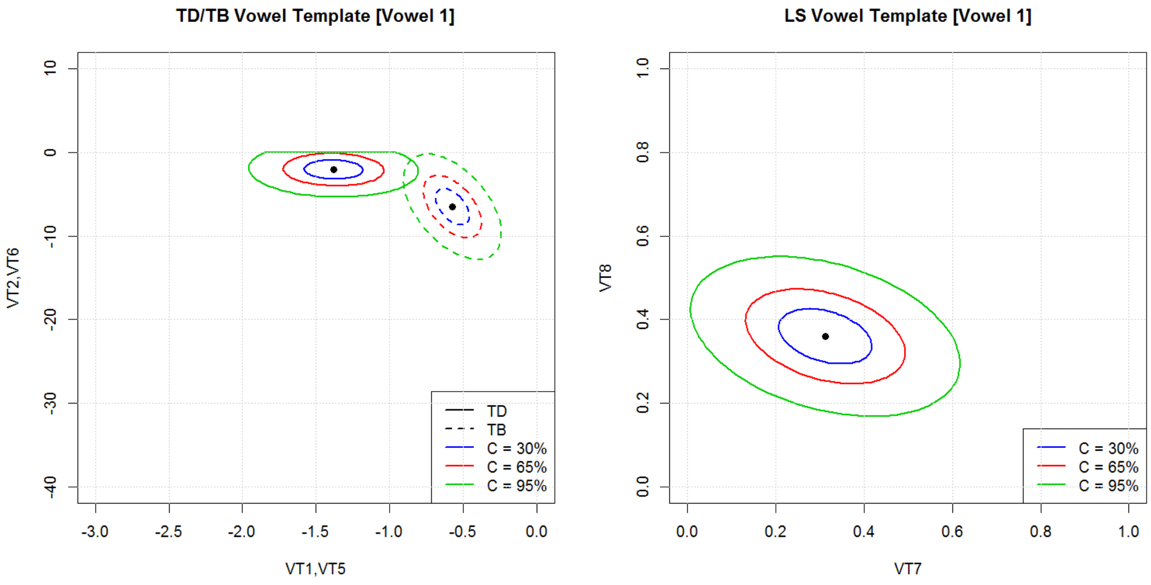
Figure 4.2 – *Vowel Template Ellipses*



Note that the outermost ellipse does not completely enclose the AE speaker data. Any data point that falls outside the 95% confidence ellipse is assumed to correspond to a "bad" pronunciation. While all AE speaker data is assumed to be correct, these points are not representative of a typical tongue blade positioning according to the data of the EMA-MAE corpus.

For each sensor, the positions of all native English repetitions of each vowel were used to form concentric confidence ellipses at levels of 30%, 65%, and 95%. The three sets of ellipses for each sensor represent the kinematic template for a given vowel. Recall that the conversion to feature space moved the palate trace to y=0 (see section 3.2). This means that any data point with a greater y value than 0 is impossible. Any confidence ellipse for the tongue sensors that extended into the y > 0 region had the corresponding y values set to 0. Figure 4.3, which shows the kinematic vowel template for /iy/, displays the effect of this truncation.

Figure 4.3 – *Kinematic Vowel Template – [/iy/, Feature Space]*



Note that the outermost TD ellipse had its y > 0 values set to 0. For a pronunciation to be considered correct, the articulators should fall within all three sets of ellipses. Feedback will provide information about the "pronunciation tiers" that the input data landed in for each sensor, as well as information about the distance from the centers of these ellipses. This is discussed in further detail in section 4.2.3

4.2.3 Visualization and Feedback

In order to be used to provide pronunciation feedback, the kinematic templates must be presented in an intuitive way for training participants. The first step towards this goal is the conversion of the templates from feature space to Euclidean space. While feature space (discussed in section 3.2) is especially useful for analysis and interpretation of results of acoustic-to-articulatory inversion, visualization plots presented in feature space would not be especially helpful to pilot study participants. However, there are challenges associated with converting back to Euclidean space. To convert to feature space, a speaker's palatal outline, distance between central maxillary incisor and back molar, maximum lip separation, and minimum lip separation are required. The acoustic-to-articulatory inversion system, which performs inversions without articulatory information and returns the results in feature space, provides no method of converting back to Euclidean space.

When providing pronunciation feedback to a speaker, a visualization plot does not need to meet the exact dimensions of the speaker's vocal tract. As long as the results and corrections shown in the plot are interpretable to the speaker, the plot's ability to assist in correcting pronunciation should be unaffected. With this in mind, the missing vocal tract parameters may be estimated and applied to all articulatory features when converting back to Euclidean space. A general midsagittal palate trace was formed through the point-by-point averaging of all 40 EMA-MAE speakers' palate traces. This process was repeated for CMI-to-back molar distance to form a general horizontal normalization scalar. Using the average palate trace and normalization scalar, the Euclidean tongue dimensions (TDx, TDy, TBx, TBy) are calculated from the corresponding articulatory features (VT1, VT2, VT5, VT6) using equations (2.1), (2.2), (2.5) and (2.6).

In the visualization plots, the lips' positions re-expressed in Euclidean space, but are still represented by the protrusion of the lips and distance between lips. These features are still represented by the aggregate LS sensor, as opposed to converting back to UL and LL. This is done by treating the upper lip as a stationary articulator in the y dimension (located at y = 0). Equation (2.7) is used to convert the lip protrusion (VT7) to Euclidean space, and this feature is applied to both lips. The lip separation, which is represented by a [0,1] normalization in feature space, was converted to Euclidean space using a *lip scalar*. The lip scalar represents the maximum lip separation in Euclidean space. This value was calculated as the average of the maximum vertical distance between the lips for all 40 EMA-MAE speakers, equal to 32.57. Given a stationary upper lip, the converted feature values may be thought of as describing the location of the lower lip. Equations (4.1) and (4.2) show the calculations for the dimensions of LS in Euclidean space.

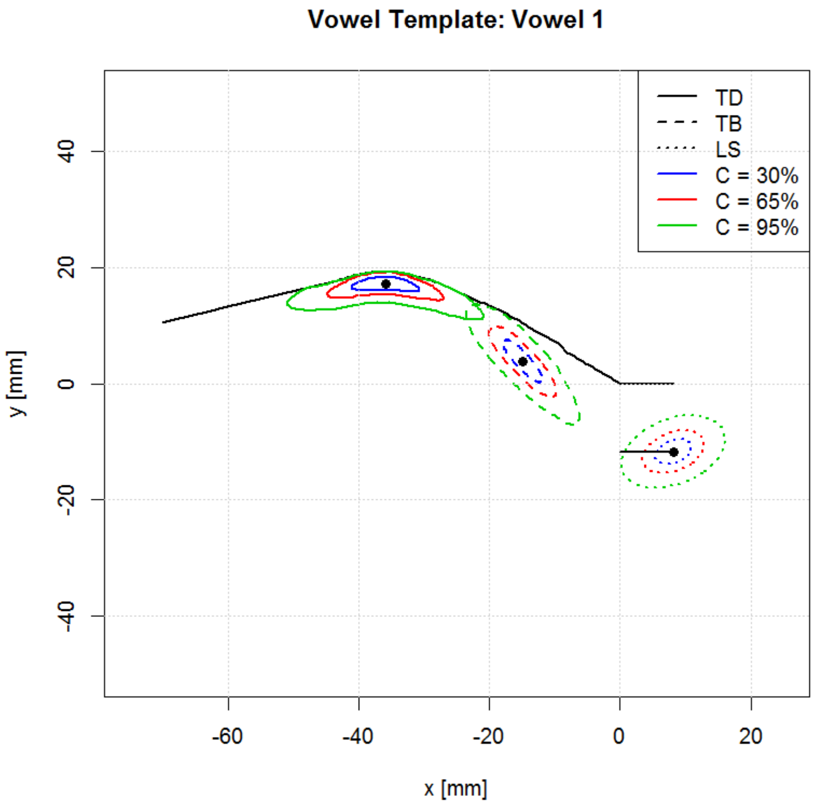
$$LS_x' = VT7 * H \tag{4.1}$$

$$LS_y' = -VT8 * 32.57 \tag{4.2}$$

LS' is the converted values of LS to Euclidean space, and H is the horizontal normalization scalar. Note that the new LS y dimension is expressed as a negative value. With the upper lip located at y=0 and LS_y' representing the both the lower lip's vertical location, the lower lip's height must be located at the negative value of the (scaled) distance between the lips. In the visualization plots, the lips are represented by straight lines drawn at the corresponding heights (y=0 for the upper lip and y= LS_y' for the lower lip) from x=0 to x= LS_x'.

Figure 4.4 displays the kinematic template for vowel /iy/ in Euclidean space.

Figure 4.4 - *Kinematic Vowel Template – [/iy/, Euclidean Space]*



Note that compared to the feature space template from Figure 4.3, the tongue template ellipses are warped. Recall that the ellipses are defined in feature space, which use palate referenced tongue positions. Converting back to Euclidean space removes this palate reference, and displays the corresponding Euclidean space representation. Also note that compared to Figure 4.3, the LS ellipses have been reflected across the x-axis. This is due to the fact that the lip separation was converted to a negative value when referencing the lower lip to a stationary upper lip (see equation (4.2)).

After forming the Euclidean representation of the vowel templates, functions were written to compare a set of articulator positions to the template. Both a feature space (VT1-VT8) and Euclidean space (TD, TB, LS) version of the template comparison function was created. A function that converts a set of feature space values to Euclidean space values was also written. This allows for data presented in feature space format to be compared to Euclidean templates. These functions plot the input data points over the templates, with arrows pointing from the points to the centers of the corresponding articulator ellipses. These arrows indicate both the distance and direction of movement required to correct the articulator positon. Figures display plots showing a sample comparison of points against the template of /aa/ in both feature space and Euclidean space.

Figure 4.5 – *Sample Template Comparison [Feature Space]*

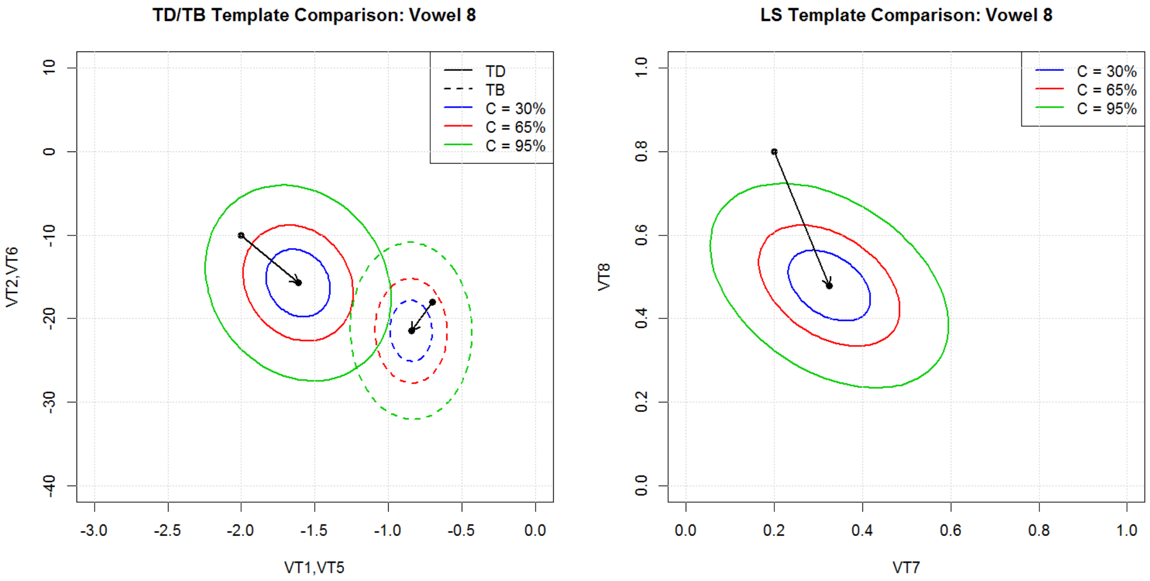
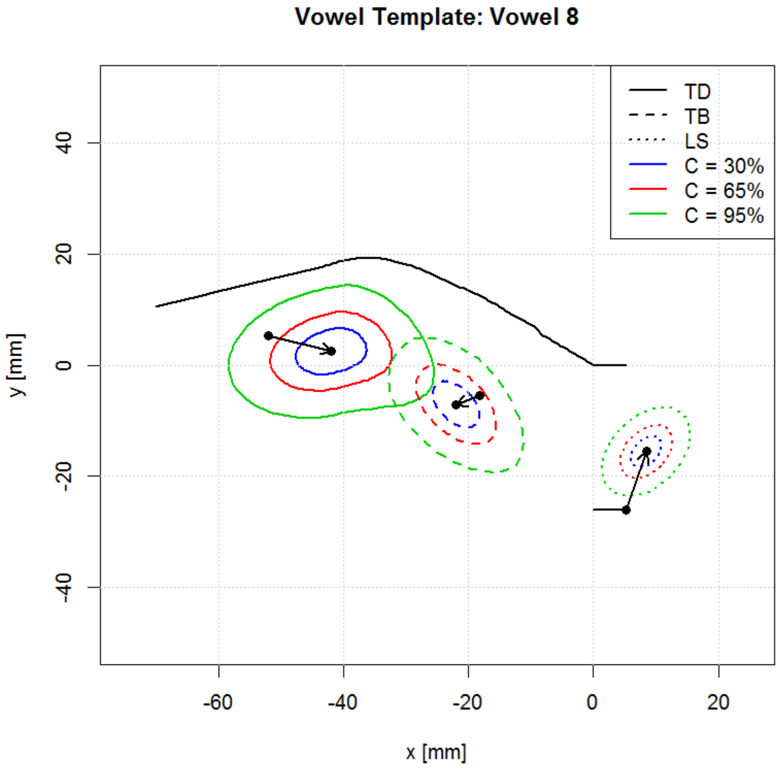


Figure 4.6 - *Sample Template Comparison [Euclidean Space]*



In addition to plotting the data points over the templates, the comparison functions also calculate and return numerical values expressing the relationship between the input data points and the template targets. These values include the straight line distance from the point to the target, the angle (with respect to the positive x axis) of the vector pointing from the point to the target, and the confidence level of the outermost ellipse that each point falls within (and -1 if the point is outside all three ellipses). To demonstrate this functionality, Table 4.1 displays the values returned from the comparison of Figure 4.6.

Table 4.1 - *Sample Template Comparison [Correction Information]*

Articulator	Distance [mm]	Angle [°]	Confidence
TD	10.48	-15.95	95
TB	4.111	-155.7	65
LS	10.95	72.71	-1

4.3 Consonant Cluster Templates

4.3.1 MSTD Modeling

During the EMA-MAE consonant cluster analysis (section 3.5), the highly variable nature of the consonant cluster data was observed. There were no noticeable trends among speaker groups across magnitude/speed/trajectory/duration (MSTD) parameters, consonant clusters, or EMA sensors. Given this fact, each MSTD parameter of each sensor for each cluster is modeled independently for AE speakers. The combination of these MSTD models for a given consonant cluster forms that cluster’s kinematic template.

As discussed in section 4.1, the kinematic templates are defined as a combination of the data from all AE speakers. For the consonant cluster templates, this is implemented as a combination of the MSTD parameters. Extending the concept of the confidence level from vowels to the consonant cluster template formation, the magnitude (M) parameter template is defined as the 95% confidence interval of the magnitudes of all AE replicates of a given cluster. The duration (M) parameter template is defined by the same calculation. The speed (S) parameter template for a given cluster is calculated as the mean speed curve for the cluster across all AE replicates. Similar to the speed template, the trajectory (T) parameter template is calculated as the mean trajectory of the cluster across all AE replicates (in both the x and y directions).

The magnitude parameter template intervals of the TB sensor for cluster 1, as well as the corresponding duration intervals, are shown in Table 4.2.

Table 4.2 – *Consonant Cluster Template: MD Parameters*

Parameter	Start	Stop
Magnitude [mm]	8.54	10.3
Duration [s]	0.212	0.235

Figure 4.7 displays the speed parameter template of the TB sensor for cluster 1 (/nd/), and Figure 4.8 displays the trajectory parameter template for the same sensor and cluster.

Figure 4.7 – *Consonant Cluster Template: S Parameter*

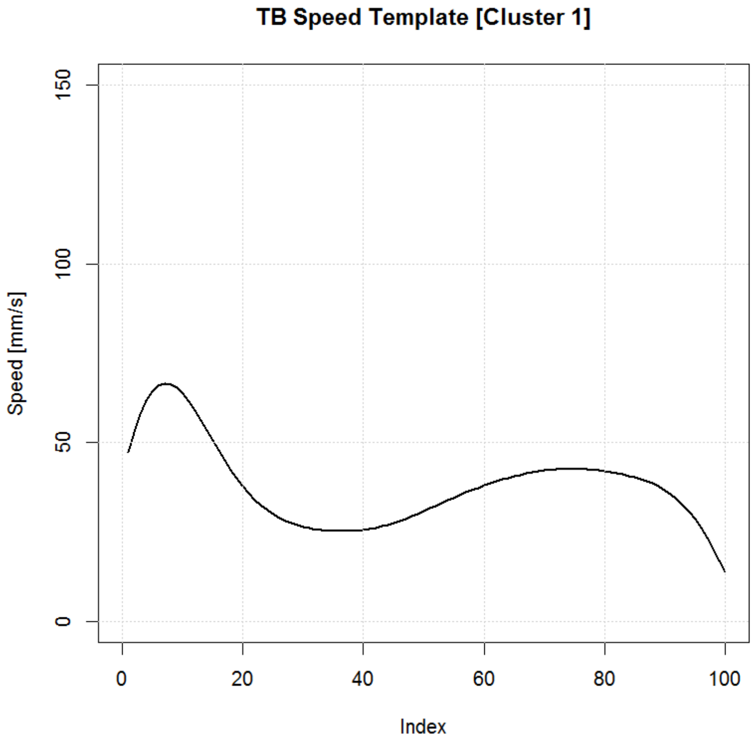
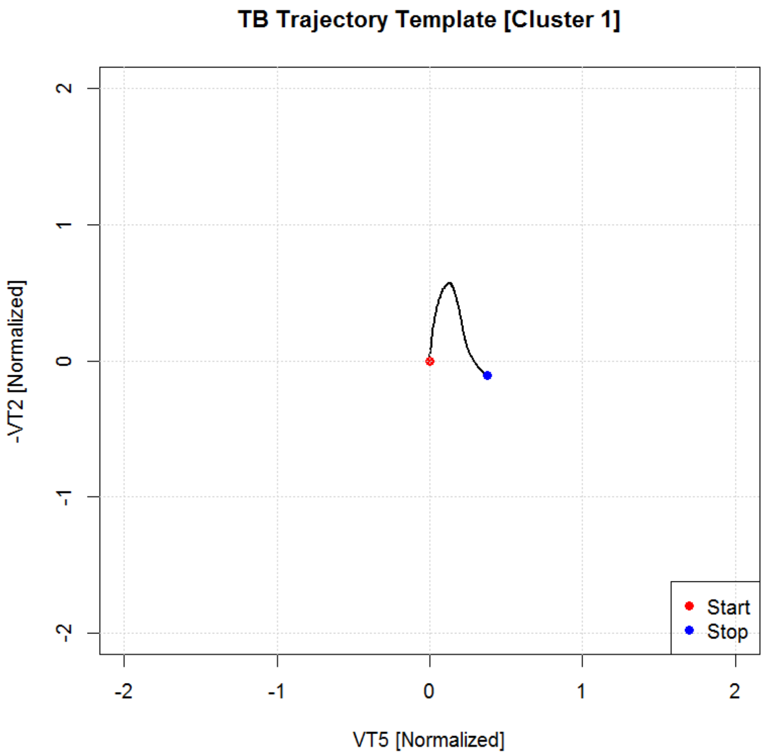


Figure 4.8 - *Consonant Cluster Template: T Parameter*



Together, the magnitude and duration intervals of Table 4.2, along with the speed and trajectory data of Figure 4.7 and Figure 4.8, form the English kinematic template for consonant cluster /nd/ (for TB; note that there are also corresponding templates for TD and LS).

4.3.2 Visualization and Feedback

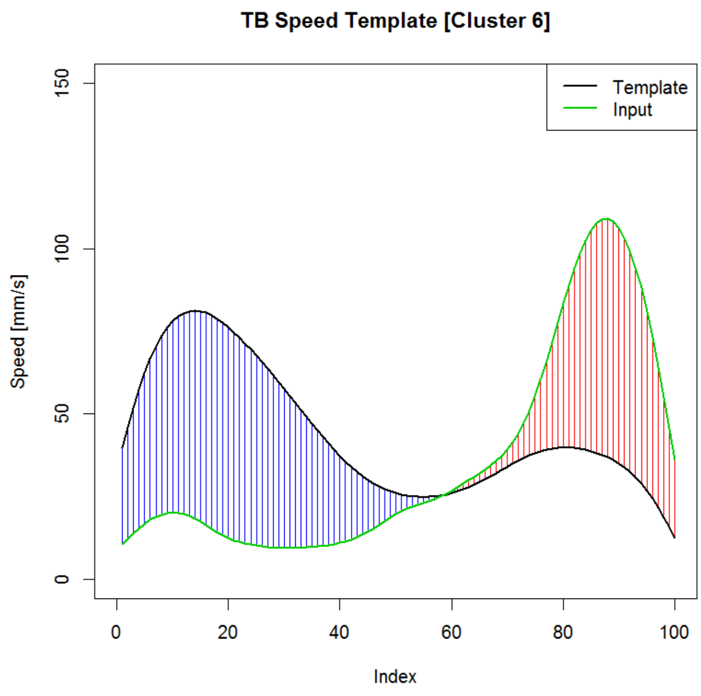
As previously discussed, the kinematic templates must be presented in an intuitive way to pilot study participants in order to provide meaningful pronunciation feedback. Of the 4 MSTD parameters, only two of them (speed and trajectory) require visualization plots. While the speed curves are already expressed in a presentable (and interpretable by clinicians and pilot study participants) format, the trajectory templates need to be converted to a familiar format. This is done by performing an approximate conversion back to Euclidean space (similarly to the vowel templates in section 4.2.3).

The conversion of each cluster trajectory to Euclidean space involves multiple steps. Like the vowel templates, cluster templates require a reference palate, normalization scalar, and lip scaling factor to convert back to Euclidean space. The same estimates of these values that were applied to the vowel templates were used for cluster templates (see section 4.2.3 for details on definition and derivation of these values). With these values obtained, the same un-normalization functions used for vowel templates (based on equations (2.1), (2.2), (2.5), and (2.6)) may be used to convert a trajectory to Euclidean space. However, recall that the trajectory extraction process moved all scaled all movement patterns and moved their starting points to the origin. Before the trajectories can be converted back to Euclidean space, they must first be converted back to feature space.

In order to place the trajectory on the proper scale and location for a feature space representation, estimates of the appropriate scaling and translation values are required. Each trajectory was [0,1] normalized during extraction, so the x and y dimensions were multiplied by the average width (x) and length (y) (respectively) of each cluster across all feature space AE repetitions of the cluster in order to scale it back to feature space dimensions. The trajectories were also translated during extraction such that their starting points were all located at the origin. To translate these trajectories back to an appropriate feature space location, the average starting location for each cluster across all AE repetitions was calculated. The trajectories were translated using these values. With the scaling and translation complete, the cluster template now has a complete feature space approximation. From here, each point in the trajectory may be converted back to Euclidean space using the same un-normalization functions used for vowel templates (section 4.2.3).

With all 4 MSTD parameter templates represented in an interpretable format, the results may be presented and compared against a speaker's input data. Functions were written for both the magnitude and duration parameters that check if a given input cluster falls within the template intervals. These functions return a Boolean indicating the result of the test. A function was also written to compare an input speed curve to a template speed curve. First, this function plots both curves in the same window. The differences in speed are highlighted in the plot through vertical lines (red when the input speed is higher than the template speed, and blue when the opposite occurs). Figure 4.9 displays sample speed curve comparison results.

Figure 4.9 – *Sample Cluster Template Comparison: S Parameter*

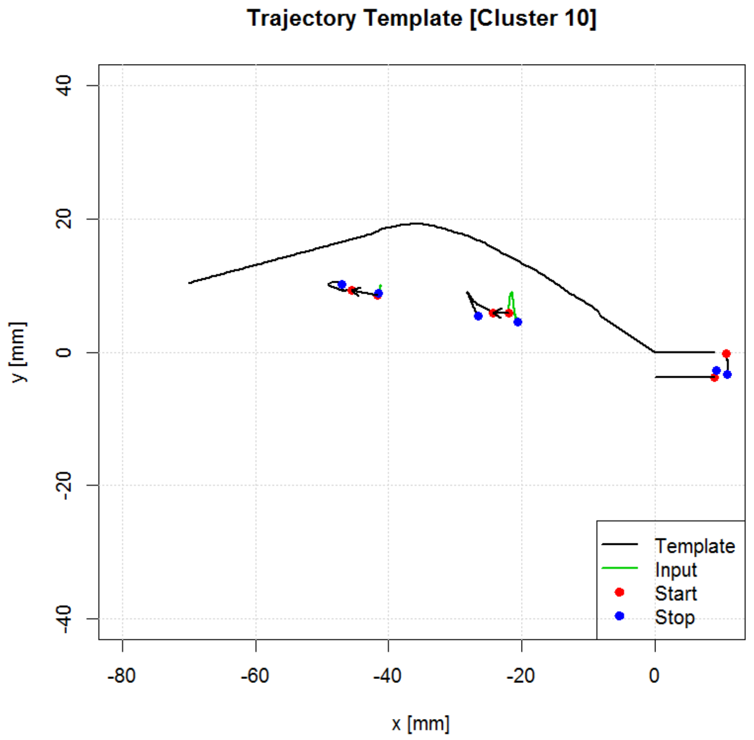


This function also returns a "difference curve", which is simply a vector containing the point-by-point differences between the template speed and input speed:

$$diff[i] = speed_{template}[i] - speed_{input}[i] \tag{4.3}$$

Two implementations of a trajectory template comparison function were created. One implementation performs the comparison in [0,1] normalized trajectory space, while the other performs the comparison in Euclidean space. Similar to the vowel template plots, the Euclidean space implementation plots an average palate, and represents the lips with horizontal lines starting at x=0 and extending in the positive x direction. The "lip lines" are referenced to the starting point of the input LS trajectory template. Figure 4.10 displays a sample Euclidean space template comparison.

Figure 4.10 – *Sample Cluster Template Comparison: T Parameter [Euclidean]*



Similar to the speed template comparison function, the trajectory functions return a "difference trajectory", which is a vector containing the x and y point-by-point differences between the template trajectory and input trajectory:

$$diff_x[i] = x_{template}[i] - x_{input}[i] \tag{4.4}$$

$$diff_y[i] = y_{template}[i] - y_{input}[i] \tag{4.5}$$

4.4 Contrastive Stress Templates

4.4.1 MSD Modeling and Feedback

During the EMA-MAE contrastive stress analysis (section 3.6), speakers were shown to increase magnitude, speed, and duration (MSD) when applying stress.

However, AE speakers were also shown to having varying amounts of MSD increase with each contrastive stress word. Given such a small amount of contrastive stress data to work with, it is unclear if this variation is a population trend. Without an answer to this question, and using the currently available data, the contrastive stress templates are built using the MSD data from all AE speakers.

The confidence interval based approach used for consonant cluster templates (section 4.3) was also applied to the contrastive stress data. For contrastive stress pairs, confidence intervals were calculated for all 3 MSD parameters. These intervals are built from all M, S, or D ratios across all AE speakers for each sensor. The combination of the MSD intervals forms the contrastive stress English template for a given sensor. Table 4.3 shows the MSD intervals for the TD sensor.

Table 4.3 – *Contrastive Stress Template Intervals*

	TD		***TB***		***LS***	
Parameter	**Start**	**Stop**	**Start**	**Stop**	**Start**	**Stop**
Magnitude [mm/mm]	1.660	1.956	2.222	2.612	2.073	2.476
Speed [(mm/s)/(mm/s)]	1.244	1.416	1.693	1.977	1.578	1.781
Duration [s/s]	1.357	1.540	1.357	1.540	1.357	1.540

Note that the duration interval is identical for all 3 sensors. Regardless of the sensor in question, pronunciation duration remains the same.

In order to provide feedback to speakers regarding their MSD parameters, functions were written to determine the proximity of input MSD parameters to the template intervals. These functions, for M, S, and D, return a Boolean value indicating whether or not the input parameter falls within the template interval, as well as the distance between the input parameter and template interval.

4.5 Summary

The English kinematic templates for vowels, consonant clusters, and contrastive stress pairs were built as combinations of data from all 20 native English speakers. For vowels, the templates were implemented as sets of concentric confidence ellipses that specify the proper articulator locations to produce the corresponding vowel. For consonant clusters, these templates were implemented as a set of *MSTD* parameters, which model the magnitude, speed history, movement trajectory, and duration of articulator for each cluster. Finally, for contrastive stress pairs, the templates were implemented as a set of *MSD* parameters, a variation of MSTD parameters which model the relationship between the stressed and unstressed magnitude, average speed, and duration characteristics for each articulator. While there are optimizations and extensions to be made, the combination of these features form the first version of the Marquette University Speech and Signal Processing and Speech and Swallowing labs' native English kinematic templates.

5 Summary and Conclusions

5.1 Summary

This thesis has presented a set of analyses of electromagnetic articulography (EMA) data, as well as implementations of midsagittal kinematic models of American English (AE) vowels, consonant clusters, and stress characteristics. These kinematic models were developed in order to evaluate the results of acoustic-to-articulatory inversion in a pilot study to assess the feasibility of using said inversion as a method of pronunciation training for Mandarin-accented English (MAE) speakers. The development of these models started with an introduction to the fundamentals of both computer aided language learning (CALL) (section 1.1) and speech production in general (section 2.1.1). This also included an overview of the differences between American English and Mandarin Chinese speech production, and the challenges associated with learning American English as a speaker of Mandarin Chinese (section 2.1.2). This was followed by an introduction to the electromagnetic articulography Mandarin-accented English (EMA-MAE) database and acoustic-to-articulatory inversion system, both developed by Marquette University's Speech and Signal Processing and Speech and Swallowing laboratories (sections 2.2 and 2.3).

After the general overview, the data of the EMA-MAE corpus was analyzed in order to characterize the relationship between the acoustic and kinematic data and the relationship between English and Mandarin-accented English speech production. This started with an introduction to the EMA sensors used for this study, as well as an introduction to the feature space conversion that allows kinematic speech data to be analyzed in the same format as the features returned from the acoustic-to-articulatory inversion system (sections 3.1 and 3.2). Prior to the discussion of the speaker data, a number of statistical analysis techniques used to evaluate the data were introduced (section 3.3). Finally, the analysis started with an investigation of both the formant frequencies and EMA sensor position data produced by each of the 40 speakers when articulating vowels, including a comparison of these formants and sensor positions to the vowel quadrilateral and each other (section 3.4). Next, the analysis moved from vowels to consonant clusters and stress characteristics. This included an introduction to magnitude, speed, trajectory, and duration (MSTD) analysis as a method of evaluating dynamic speech movement (sections 3.5 and 3.6).

Using the information obtained during the acoustic and kinematic data analyses of chapter 3, kinematic models of American English vowels, consonant clusters, and stress characteristics were developed. This began with an introduction to the “vowel template region”, which specifies, in the midsagittal plane, the articulator positioning corresponding to the correct pronunciation of a given vowel. This was followed by the implementation of the template regions through the use of concentric confidence ellipses, as well as visualization plots that allow pilot study participants to observe their current articulatory positioning with suggestions for improvement (section 4.2). Finally, the results of the MSTD analysis of chapter 3 were used to develop models of English consonant cluster and stress characteristics through the combination of information regarding the magnitude, speed, movement pattern, and duration of the speech data (section 4.3).

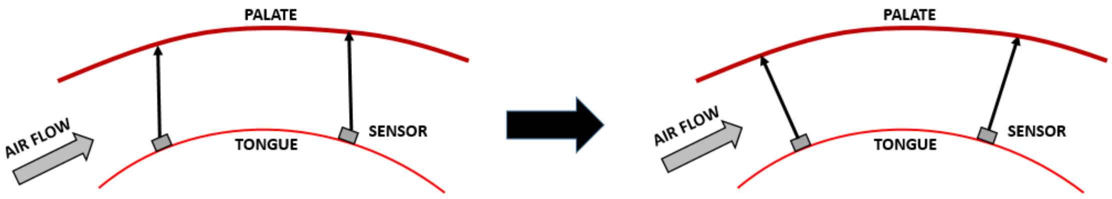
5.2 FUTURE WORK SUGGESTIONS

The work performed in this thesis is still very much in its early stages of development. This research and development may be extended to both improve and expand the capabilities of the kinematic templates and pronunciation training method as a whole. This section discusses a number of suggestions for next steps in the development of accurate templates and meaningful pronunciation feedback.

5.2.1 Adjusted Articulatory Features

The features that are currently used by the acoustic-to-articulatory inversion system (see section 2.3.2 for details) provide several advantages over unmodified sensor positions for analyzing and modeling human speech. While these features have laid the groundwork for Marquette's inversion system, they are only a subset of the several possibilities for articulatory features. Through extended research, the ideal features for the application of speech inversion may be discovered. In the case of Marquette's inversion system specifically, small changes to currently established features may improve the efficacy of the system. For example, the Marquette inversion system currently uses vertical distance between the tongue and palate as a means of representing the vocal tract shape. [52] describes an articulatory normalization that instead uses the shortest distance between a given tongue position and speaker's palate to represent the vocal tract shape. By considering the smallest distance at a given position instead of vertical distance, this method better accounts for the actual shape of the vocal tract in many cases:

Figure 5.1 – *Adjusted Vertical Tongue Features*



Note that given the shape of the tongue and palate, this “nearest neighbor” approach provides a better representation of the cross section of the vocal tract with respect to the air flow. Small adjustments such as these can lead to more accurate speech modeling.

5.2.2 Articulatory Inversion-Based Templates

The Speech and Signal Processing lab’s acoustic-to-articulatory system is not 100% accurate ([5] discusses the accuracy of this system). This means that if a speaker were to provide speech to the inversion system, the output features would not be identical to those given by the data obtained from EMA. In other words, even if the kinematic templates perfectly modeled English speakers, and a speaker produced perfect English, the inversion system results would still not match the templates. This introduces, in addition to the pronunciation error by the speaker, a second source of error: the error inherent in the acoustic-to-articulatory inversion system. At the moment, there is no way to distinguish one source of error from the other when inversion results don’t match the kinematic templates. A possible solution to this issue is to create the kinematic templates using inversion system data instead of EMA data. Theoretically, by producing both the inversion results and templates from the same source, the error introduced by the imperfection of the inversion system is eliminated. A potential problem with this solution is that while the pronunciation assessment might become more accurate, the kinematic templates become models of inaccurate models of actual speech. The templates would have no practical applications outside of use with Marquette’s inversion system.

5.2.3 Sensor Orientation Inclusion

As discussed in chapter 2, the EMA sensor data provides, in addition to a sensor's position in Euclidean space, the orientation of the sensor (in quaternion format). This orientation information provides more insight into a speaker's articulation at a given time, and could potentially be used in the development of articulatory features and kinematic templates. One possible application of this data is the use of the orientation information to estimate the tongue position at locations where sensors are not placed. [27] describes the use of EMA quaternion data to estimate the location of the tongue's surface at several locations. Information about the entire tongue surface could be useful in identifying additional differences between AE and MAE articulation, and could also be used to add detail to the visualization plots used for the pilot study. The inclusion of sensor orientation data has the potential to significantly improve the analysis of speech production.

5.2.4 Inclusion of Additional Tongue Sensors

Currently, the EMA-MAE dataset consists of data from three tongue sensors, two of which are located in the midsagittal plane. While sensor orientation data may potentially be used to estimate the tongue's position at various locations, additional tongue sensors provide more accurate information about the tongue's positioning. These additional sensors would also increase the accuracy of tongue surface position estimation at other locations. However, this increased resolution comes at a cost. Each additional sensor places on the speaker's tongue increases the likelihood of their speech becoming distorted. The decision of sensor quantity and placement becomes a tradeoff between resolution and speech quality.

5.2.5 Analysis of English Speech Audio data

The analyses performed in this thesis assume that all English speakers in the EMA-MAE data set spoke “perfect” English. That is to say, all speakers are assumed to produce all vowels, consonant clusters, and stress characteristics correctly. This led to the inclusion of data from all 20 English speakers in the development of kinematic templates. If any of the speakers were in fact producing English poorly, their data currently corrupts the kinematic templates in their current steps. While the template development process took steps to avoid outliers, it did not examine the audio quality of the speakers in any way. In order to assess the accuracy of the templates, one of the next steps should be the evaluation of English speaker’s audio data. One method of assessment of the speech quality would be the investigation of each speaker’s transcriptions in the EMA-MAE dataset. The dataset contains individual transcriptions of each speaker’s data from multiple transcribers, as well as a set of consensus transcriptions. The study and comparison of the consensus transcriptions would provide insight into the consistency in articulation of each speaker in comparison to the others.

5.2.6 Extension to Address Coarticulation Effects

Chapter 2 discussed introduced the concept of coarticulation, and mentioned the fact that all vowel and consonant cluster data analysis came from word level prompts. Currently, all identical consonant clusters, regardless of the adjacent speech sounds, are analyzed together as a single cluster type. Due to coarticulation effects, the production of a cluster (especially at the endpoints) will vary depending on the adjoining phonemes. A more accurate analysis would take these different speech contexts into account and attempt to model them along with the base consonant cluster. Additionally, this analysis would also model clusters in sentence and paragraph level speech segments (which introduce additional coarticulation effects due to nearby words). These are largely difficult tasks, but are a logical next step to the modeling of English speech for pronunciation training.

5.3 CONCLUSIONS

While the work performed in this thesis was done for the purpose of developing and implementing English kinematic templates, the presented information and methods may be applied to a number of applications (especially in the fields of speech modeling and language learning). The kinematic templates, though still in early stages, contribute to the larger goal of determining the feasibility of using acoustic-to-articulatory inversion for pronunciation training. In the upcoming pilot study, these templates and accompanying visualization plots will be evaluated on their ability to assess the features returned from the acoustic-to-articulatory inversion system and provide meaningful feedback to pilot study participants. After assessment of their ability to assist language learners, the templates can be further improved to become a formidable pronunciation assessment tool.

BIBLIOGRAPHY

[1] G. Davies, "CALL (computer assisted language learning)," LLAS Centre for Languages, Linguistics, and Area Studies, 2016. [Online]. Available: https://www.llas.ac.uk/resources/gpg/61. [Accessed 1 March 2017].

[2] T. K. Hansen, "Computer assisted pronunciation training: the four 'K's of feedback," in *International Conference on Multimedia and Information and Communication Technologies in Education*, Seville, 2006.

[3] H. Meng, "Developing speech recognition and synthesis technologies to support computer-aided pronunciation training for Chinese learners of English," The Chinese University of Hong Kong, 2009.

[4] A. K. Elimat and A. F. AbuSeileek, "Automatic speech recognition technology as an effective means for teaching pronunciation," *JALT CALL,* vol. 10, no. 1, pp. 21-47, 2014.

[5] A. Ji, "Speaker independent acoustic-to-articulatory inversion," Marquette Univ., Milwaukee, 2014.

[6] Dictionary.com, "Contrastive stress," 2017. [Online].

[7] P. H. School, "Digestive system," [Online]. Available: http://phs.psdr3.org/science/anatomy/digestive.html. [Accessed May 2017].

[8] R. Mannell, "Source-filter theory of speech production," Macquarie University, December 2008. [Online]. Available: http://clas.mq.edu.au/speech/acoustics/frequency/source_filter.html. [Accessed Jan 2017].

[9] F. Trujillo, "The production of speech sounds," June 2006. [Online]. Available: http://www.ugr.es/~ftsaez/fonetica/production_speech.pdf. [Accessed Jan 2017].

[10] G. Fant, "Acoustic theory of speech," in *Acoustic Theory of Speech Production*, The Hauge, Mouton Publishers, 1970, pp. 15-92.

[11] L. W. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Englewood Cliffs: Prentice-Hall, Inc., 1978.

[12] U. o. Lausanne, "Introduction to phonetics," University of Lausanne, [Online]. Available: https://www.unil.ch/sli/home/menuguid/ressources/cours-et-

livres-en-ligne/introduction-to-phonetics/introduction.html. [Accessed March 2017].

[13] J. Coleman, "Multiple articulation and coarticulation," University of Oxford, [Online]. Available: http://www.phon.ox.ac.uk/jcoleman/MULTART.htm. [Accessed December 2016].

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

[15] Y. Chen, M. Robb, H. Gilbert and J. Lerman, "Vowel proudction by Mandarin speakers of English," *Clinical Linguistics and Phonetics,* vol. 15, no. 6, pp. 427-440, 2001.

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

[17] L. Mi, S. Tao, W. Wang and C. Liu, "English vowel indetification and vowel formant discrimination by native Mandarin Chinese - and native English - speaking listeners: The effect of vowel duration dependence," *Elsevier,* pp. 58-65, 2016.

[18] T. L. Gottfried, A. Staby and D. Riester, "Relation of pitch glide perception and Mandarin tone identification," [Online]. Available: https://www2.lawrence.edu/fast/gottfrit/Mandmusic.html.

[19] Y. Zhang, S. L. Nissen and A. L. Francis, "Acoustic characteristics of English lexical stress produced by native Mandarin speakers," *Journal of the Acoustical Society of America,* vol. 123, no. 6, pp. 4498-4513, 2008.

[20] J. Morton and W. Jassem, "Acoustic correlates of stress," *SAGE Journals,* vol. 8, no. 3, pp. 159-181, 1965.

[21] B. Lindholm, J. Lubker and T. Gay, "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," *Journal of the Scoustical Society of America,* vol. 62, pp. 1115-1123, 1977.

[22] Y. Chen, "Acoustic characteristics of American English produced by native speakers of Mandarin," University of Connecticut, 1999.

[23] The Education University of Hong Kong, "Comparison of english and mandarin (segmentals)," The Education University of Hong Kong, 3 March 2014. [Online]. Available: http://ec-

concord.ied.edu.hk/phonetics_and_phonology/wordpress/?page_id=328. [Accessed May 2017].

[24] P. Birkholz, B. J. Kroger and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 5, pp. 1422-1433, 2011.

[25] H. Horn, G. Goz, M. Bacher, M. Mullauer and D. Axmann-Krcmar, "Reliability of electromagnetic articulography recording during speaking sequences," *European Journal of Orthodontics,* vol. 19, pp. 647-655, 1997.

[26] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," Edinburgh, 2000.

[27] A. Colb, "Software tools and analysis methods for the use of electromagnetic articulography data in speech research," Marquette Univ., Milwaukee, 2015.

[28] Y. Yunusova, J. G. Green and A. Mefferd, "Accuracy assessment for AG500, electromagnetic articulograph," *Journal of Speech, Language, and Hearing Research,* vol. 52, pp. 547-555, 2009.

[29] J. Berry, "Accuracy of the NDI wave speech research system," *Journal of Speech, Language, and Hearing Research,* vol. 54, pp. 1295-1301, 2011.

[30] V. L. Gracco, "Analysis of speech movements: practical considerations and clinical application," Haskins Laboratories, 1992.

[31] University of Southern California, "EMA Database," University of Southern California, 2010. [Online]. Available: http://sail.usc.edu/ema_web/. [Accessed March 2017].

[32] A. Marchal and W. J. Hardcastle, "ACCOR: instrumentation and database for the cross-language study of coarticulation," *Lang. Speech,* vol. 36, pp. 137-153, 1993.

[33] J. Berry, A. Ji and M. T. Johnson, "EMA-MAE corpus user's handbook," Milwaukee, 2014.

[34] H. A. Spanish, "The IPA chart for language learners," 9 April 2017. [Online]. Available: https://www.happyhourspanish.com/ipa-chart-language-learners/. [Accessed 2 May 2017].

[35] J. Queinec, "Smoothing algorithm using Bezier curves," 26 February 2005. [Online]. Available: http://www.efg2.com/Lab/Graphics/Jean-YvesQueinecBezierCurves.htm. [Accessed February 2017].

[36] B. S. Atal, J. J. Chang, M. V. Mathews and J. W. Turkey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *Journal of the Acoustical Society of America,* vol. 63, pp. 1535-1555, 1978.

[37] T. Hueber, A. B. Youssef, G. Bailly and F. Elisei, "Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory HMM for pronunciation training," in *Interspeech*, Portland, 2012.

[38] J. Wang, A. Samal and J. R. Green, "Across-speaker articulatory normalization for speaker-independent silent speech recognition," *Interspeech,* pp. 1179-1183, 2014.

[39] O. Engwall, "Pronunciation analysis by acoustic-to-articulatory feature inversion," in *International Symposium on Automatic Detection of Errors in Pronunciation Training*, Stockholm, 2012.

[40] S. Hiroya, "Acoustic-to-articulatory inversion using a speaker-normalized HMM-based speech production model," in *International Seminar on Speech Production*, Boston, 2008.

[41] F. Kubala, R. Schwartz and C. Barry, "Speaker adaptation using multiple reference speakers," in *Workshop on Speech and Natural Language*, Cape Cod, 1989.

[42] S. Creech, "Statistical data analysis," Statistically Significant Consulting, [Online]. Available: http://www.statisticallysignificantconsulting.com/StatisticalInference.htm. [Accessed February 2017].

[43] E. W. Weisstein, "Student's t-distribution," MathWorld, [Online]. Available: http://mathworld.wolfram.com/Studentst-Distribution.html. [Accessed March 2017].

[44] B. Winter, "The f distribution and the basic principle behind ANOVAs," Author, Birmingham, 2015.

[45] A. Baron, "Confidence intervals," Yale University, [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/confint.htm. [Accessed March 2017].

[46] Penn State University, "Confidence intervals and the Central Limit Theorem," Penn State University. [Online]. [Accessed March 2017].

[47] J. H. McDonald, "Multiple comparisons," February 2014. [Online]. Available: http://www.biostathandbook.com/multiplecomparisons.html. [Accessed 2017].

[48] P. Adank, R. Smits and R. V. Hout, "A comparison of vowel normalization procedures for language variation research," *Journal of the Acoustical Society of America,* vol. 116, no. 5, pp. 3099-3107, 2004.

[49] N. Flynn and P. Foulkes, "Comparing vowel formant normalization methods," in *International Congress of Phonetic Sciences*, Hong Kong, 2011.

[50] E. R. Thomas and T. Kendall, "The vowel normalization and plotting suite," University of Oregon, 18 November 2015. [Online]. Available: http://lingtools.uoregon.edu/norm/norm_methods.php. [Accessed January 2017].

[51] B. Wang, W. Shi and Z. Miao, "Confidence analysis of standard deviational ellipse and its extension into higher dimensional Euclidean space," *PLoS One,* vol. 10, no. 3, 2015.

[52] M. Hashi, J. R. Westbury and K. Honda, "Vowel posture normalization," *Acoustical Society of America,* pp. 2426-2437, 1998.

[53] A. Zierdt, P. Hoole and H. G. Tillmann, "Development of a system for three-dimensional fleshpoint measurement of speech movements," *International Conference of Phonetic Sciences,* pp. 73-75, 1999.

[54] U. o. Victoria, "IPA lab," University of Victoria, [Online]. Available: https://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm. [Accessed May 2017].

[55] C. Van Riper and J. V. Irwin, "Phonation," in *Voice and Articulation*, Englewood Cliffs, Prentice-Hall Inc., 1958, pp. 418-456.

[56] A. Toutios and K. Margaritis, "A rough guide to the acoustic-to-articulatory inversion of speech," 2003.

[57] V. Spruyt, "How to draw an error ellipse representing the covariance matrix," Computer Vision for Dummies, 14 April 2014. [Online]. Available: http://www.visiondummy.com/2014/04/draw-error-ellipse-representing-covariance-matrix/#Axis-aligned_confidence_ellipses. [Accessed March 2017].

[58] T. Smyth, "Some FM Instrument Examples," Simon Fraser University, 2012. [Online]. Available: https://www.cs.sfu.ca/~tamaras/freqMod/Some_FM_instrument.html. [Accessed 8 November 2015].

[59] R. W. Schafer and J. D. Markel, "Estimation of vocal tract parameters," in *Speech Analysis,* New York, IEEE Press, 1979, pp. 231-344.

[60] K. Richmond, P. Hoole and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," *Interspeech,* pp. 1505-1508, 2011.

[61] C. Qin, M. A. Carreira-Perpinan, K. Richmond, A. Wrench and S. Renals, "Predicting tongue shapes from a few landmark locations," *Interspeech,* pp. 2306-2309, 2008.

[62] S. Perkell, M. H. Cohen, M. Svirsky, M. L. Matthies, I. Garabieta and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America,* vol. 92, pp. 3078-3096, 1992.

[63] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim and S. Lee, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *Journal of the Acoustical Society of America,* vol. 136, no. 3, pp. 1307-1311, 2014.

[64] H. S. Magen, A. M. Kang, M. K. Tiede and D. H. Whalen, "Posterior pharyngeal wall position in the production of speech," *Journal of Speech, Language, and Hearing Research,* vol. 46, pp. 241-251, 2003.

[65] S. Maeda, Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, Boston: Kluwer Academic Publishers, 1990.

[66] A. Lyon, "Why are normal distributions normal?," *British Journal for the Philosophy of Science,* 2014.

[67] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLOS One,* 15 July 2015.

[68] N. D. Inc., *Wave user guide,* Author, 2016.

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

[70] R. A. Hoops, "Time and duration in sound," in *Speech Science,* Springfield, Charles C. Thomas Publishing, 1960, pp. 57-66.

[71] S. Gorard, "Revisiting a 90-year-old debate: the advantages of the mean deviation," University of New York, New York, 2004.

[72] H. A. Gleason, in *An Introduction to Descriptive Linguistics*, Chicago, Rinehart and Winston Hot Inc., 1961.

[73] G. Davies and F. Riley, "Glossary of ICT terminology," Information and Communications Technology for Language Learners, 2012. [Online]. Available: http://www.ict4lt.org/en/index.htm. [Accessed October 2016].

[74] P. Badin, A. B. Youssef, G. Bailly and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *Interspeech*, Mahukari, 2011.

[75] Wolfram MathWorld, "Chi-Squared Distribution," Wolfram MathWorld, [Online]. Available: http://mathworld.wolfram.com/Chi-SquaredDistribution.html. [Accessed March 2017].

6 APPENDIX A

This appendix contains tables and figures that provide more detailed information to support the thesis content, but did not lead to any additional discoveries or whose information was considered secondary to the topic discussed in the thesis.

6.1 BACKGROUND INFORMATION

Table 6.1 displays a list of all consonant clusters from the word level prompts of the EMA-MAE dataset.

Table 6.1 – *EMA-MAE Consonant Clusters*

Cluster ID	Cluster	Cluster Word
1	nd	find
2	sl	sled
3	kl	clone
4	nz	teens
5	ld	cold
6	lt	salt
7	kr	crick
8	kw	queen
9	tr	train
10	fr	frog
11	ʃr	shrine
12	st	stable
13	pθ	depth
14	nt	tent
15	lz	falls
16	ts	bits
17	ps	tops
18	ŋz	sings
19	skw	square
20	rdz	cords
21	ldz	fields

22	lvz	shelves
23	br	breathe
24	dr	drug
25	gl	glean
26	rmθ	warmth
27	rz	cores
28	rs	course
29	pl	please
30	dθ	breadth
31	bz	robes
32	rv	carve
33	ŋks	sinks
34	ns	sense
35	gr	green
36	tw	twin
37	pr	prize
38	lθ	wealth
39	dz	beads
40	nθ	tenth
41	ls	false
42	fl	fleas
43	sw	swell, sweet
44	sk	scare

6.2 Additional Analyses Results

Table 6.2 displays the results of the t-tests performed to determine the specific groups that contained significant differences in the vowel ANOVA results presented in Table 3.6 (section 3.4.2.4). This table marks an “X” on all comparisons that yielded statistical significance (p-values less than 0.0083).

Table 6.2 - *Table 3.6 Follow-Up T-Test Results*

		Groups Compared					
Vowel	**Formant**	**ENGM - ENGF**	**ENGM - MANM**	**ENGM - MANF**	**ENGF - MANM**	**ENGF - MANF**	**MANM - MANF**
	F1		X				

1 (/iy/)	F2						
2 (/ih/)	F1		X	X	X	X	
	F2		X	X	X	X	
3 (/ey/)	F1				X	X	
	F2		X	X	X	X	
4 (/ae/)	F1		X				
	F2	X	X	X			
5 (/uw/)	F1		X	X			
	F2						
6 (/uh/)	F1	X	X	X	X	X	
	F2						
7 (/ow/)	F1						
	F2		X	X	X	X	
8 (/aa/)	F1	X	X	X			
	F2		X	X		X	

Table 6.3 displays a feature-space comparison of average position of the EMA sensors for each vowel (provides numerical values corresponding to Figures Figure 3.8-Figure 3.10, section 3.4.3.3). In addition to reporting the average position of each vowel, the difference in position across L1 (English minus Mandarin) was calculated and recorded.

Table 6.3 - *Average Sensor Positions: English vs. Mandarin*

	TDx			TDy		
Vowel	***ENG***	***MAN***	***E-M***	***ENG***	***MAN***	***E-M***
1 (/iy/)	-1.380	-1.428	0.048	-2.047	-2.407	0.360
2 (/ih/)	-1.444	-1.439	-0.005	-6.894	-3.718	-3.175
3 (/ey/)	-1.315	-1.447	0.132	-4.179	-5.468	1.289
4 (/ae/)	-1.409	-1.555	0.146	-10.701	-11.091	0.389
5 (/uw/)	-1.614	-1.667	0.053	-5.218	-7.589	2.371
6 (/uh/)	-1.612	-1.701	0.089	-12.110	-7.814	-4.297
7 (/ow/)	-1.706	-1.760	0.054	-11.472	-11.492	0.021
8 (/aa/)	-1.614	-1.753	0.139	-15.700	-14.016	-1.685

	TBx	TBy

Vowel	*ENG*	*MAN*	*E-M*	*ENG*	*MAN*	*E-M*
1 (/iy/)	-0.572	-0.603	0.031	-6.503	-5.923	-0.579
2 (/ih/)	-0.652	-0.636	-0.016	-9.798	-7.729	-2.070
3 (/ey/)	-0.597	-0.677	0.080	-10.797	-10.594	-0.203
4 (/ae/)	-0.694	-0.758	0.064	-16.721	-14.709	-2.012
5 (/uw/)	-0.834	-0.879	0.044	-12.424	-13.831	1.407
6 (/uh/)	-0.821	-0.898	0.077	-15.200	-14.290	-0.909
7 (/ow/)	-0.949	-0.984	0.035	-20.594	-18.573	-2.021
8 (/aa/)	-0.845	-0.964	0.120	-21.468	-19.258	-2.211

	LSx			**LSy**		
Vowel	*ENG*	*MAN*	*E-M*	*ENG*	*MAN*	*E-M*
1 (/iy/)	0.312	0.362	-0.049	0.360	0.313	0.047
2 (/ih/)	0.325	0.365	-0.040	0.379	0.320	0.059
3 (/ey/)	0.308	0.356	-0.048	0.423	0.357	0.066
4 (/ae/)	0.301	0.350	-0.049	0.481	0.419	0.062
5 (/uw/)	0.451	0.466	-0.015	0.203	0.220	-0.017
6 (/uh/)	0.397	0.461	-0.064	0.286	0.223	0.064
7 (/ow/)	0.444	0.458	-0.014	0.250	0.250	0.000
8 (/aa/)	0.325	0.425	-0.100	0.480	0.360	0.120

Table 6.4 displays the results of the t-tests performed to determine the specific groups that contained significant differences in the ANOVA results presented in Table 0.3 (section 3.4.3.3). This table marks an "X" on all comparisons that yielded statistical significance (p-values less than 0.0083).

Table 6.4 – *Table 0.3 Follow-Up T-Test Results*

		Groups Compared					
Vowel	**Sensor**	**ENGM - ENGF**	**ENGM - MANM**	**ENGM - MANF**	**ENGF - MANM**	**ENGF - MANF**	**MANM - MANF**
1 (/iy/)	TDx						
	TDy						
	TBx						
	TBy						
	LSx				X		
	LSy						
	TDx						

2 (/ih/)	TDy		X	X	X		
	TBx						
	TBy			X	X		
	LSx				X		
	LSy				X		
3 (/ey/)	TDx				X		
	TDy	X			X	X	
	TBx						
	TBy						
	LSx				X		
	LSy				X		
4 (/ae/)	TDx						
	TDy						
	TBx						
	TBy						
	LSx				X		
	LSy				X		
5 (/uw/)	TDx						
	TDy	X			X	X	
	TBx						
	TBy	X				X	
	LSx						
	LSy						
6 (/uh/)	TDx						
	TDy		X	X		X	
	TBx						
	TBy						
	LSx				X		
	LSy				X	X	
7 (/ow/)	TDx						
	TDy						
	TBx						
	TBy						
	LSx						
	LSy						
8 (/aa/)	TDx						
	TDy						
	TBx						
	TBy						
	LSx				X	X	
	LSy				X	X	

Tables Table 6.5-Table 6.12 display the results of the MSTD analysis ANOVA performed on the consonant clusters of Table 0.2. This data supplements the results presented in section 3.5.2.3. All comparisons that yielded statistical significance (identified significant differences between groups) are highlighted in orange.

Table 6.5 – *Consonant Cluster ANOVA: Magnitude F-Values*

Cluster	TD	TB	LS
1 (/nd)	0.602	1.067	0.475
3 (/kl/)	2.447	1.319	2.667
5 (/ldl)	2.467	4.627	2.475
7 (/kr/)	2.236	2.658	0.575
8 (/kw/)	0.535	0.956	1.400
15 (/ls/)	0.350	4.086	1.149
21 (/ldz/)	0.828	2.816	0.398
35 (/gr/)	2.487	2.293	3.057

Table 6.6 - *Consonant Cluster ANOVA: Magnitude p-Values*

Cluster	TD	TB	LS
1 (/nd)	0.6180	0.3756	0.7014
3 (/kl/)	0.0800	0.2837	0.0628
5 (/ldl)	0.0783	0.0079	0.0776
7 (/kr/)	0.1013	0.0634	0.6351
8 (/kw/)	0.6614	0.4242	0.2592
15 (/ls/)	0.7891	0.0138	0.3431
21 (/ldz/)	0.4875	0.0533	0.7555
35 (/gr/)	0.0766	0.0950	0.0410

Table 6.7 - *Consonant Cluster ANOVA: Speed F-Values*

Cluster	TD	TB	LS
1 (/nd)	2.193	6.318	0.496
3 (/kl/)	2.006	0.817	3.889

5 (/ldl)	2.199	5.779	2.636
7 (/kr/)	3.971	4.638	1.299
8 (/kw/)	4.056	1.904	1.383
15 (/ls/)	3.177	7.384	0.705
21 (/ldz/)	1.283	8.898	0.834
35 (/gr/)	3.325	4.394	2.123

Table 6.8 - *Consonant Cluster ANOVA: Speed p-Values*

Cluster	**TD**	**TB**	**LS**
1 (/nd)	0.1062	0.0015	0.6871
3 (/kl/)	0.1311	0.4932	0.0169
5 (/ldl)	0.1056	0.0026	0.0649
7 (/kr/)	0.0155	0.0078	0.2903
8 (/kw/)	0.0142	0.1469	0.2640
15 (/ls/)	0.0360	0.0006	0.5556
21 (/ldz/)	0.2953	0.0002	0.4841
35 (/gr/)	0.0307	0.0100	0.1149

Table 6.9 - *Consonant Cluster ANOVA: Trajectory F-Values*

Cluster	**TD**	**TB**	**LS**
1 (/nd)	6.002	4.612	4.710
3 (/kl/)	0.998	1.066	0.978
5 (/ldl)	2.421	1.336	0.477
7 (/kr/)	4.055	1.687	4.344
8 (/kw/)	5.325	9.076	1.478
15 (/ls/)	0.902	0.712	0.596
21 (/ldz/)	0.927	0.757	1.698
35 (/gr/)	3.478	1.168	3.321

Table 6.10 - *Consonant Cluster ANOVA: Trajectory p-Values*

Cluster	**TD**	**TB**	**LS**
1 (/nd)	0.0021	0.0080	0.0073

3 (/kl/)	0.4051	0.3759	0.4143
5 (/ldl)	0.0824	0.2784	0.7003
7 (/kr/)	0.0142	0.1875	0.0105
8 (/kw/)	0.0040	0.0001	0.2372
15 (/ls/)	0.4500	0.5514	0.6217
21 (/ldz/)	0.4381	0.5261	0.1852
35 (/gr/)	0.0261	0.3358	0.0308

Table 6.11 – *Consonant Cluster ANOVA: Duration F-Values*

Cluster	F
1 (/nd)	0.602
3 (/kl/)	2.447
5 (/ldl)	2.467
7 (/kr/)	2.236
8 (/kw/)	0.535
15 (/ls/)	0.350
21 (/ldz/)	0.828
35 (/gr/)	2.487

Table 6.12 - *Consonant Cluster ANOVA: Duration p-Values*

Cluster	p
1 (/nd)	0.6180
3 (/kl/)	0.0800
5 (/ldl)	0.0783
7 (/kr/)	0.1013
8 (/kw/)	0.6614
15 (/ls/)	0.7891
21 (/ldz/)	0.4875
35 (/gr/)	0.0766