Magnetic Resonance Imaging Movies for Multivariate Analysis of Speech

Katherine McRoberts

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Biomedical Engineering

Stephen M. LaConte, Chair
Alexander Leonessa
Bradley P. Sutton
William J. Tyler

June 3, 2013
Roanoke, VA

Keywords: MRI, speech, multivariate analysis, support vector machine, canonical correlation analysis

Magnetic Resonance Imaging Movies for Multivariate Analysis of Speech

Katherine McRoberts

Abstract

The complex human motor function of speech presents a scientifically interesting, yet relatively unexplored, means to study brain-behavior relationships. Fortunately, magnetic resonance imaging (MRI), which has been proven to characterize soft tissue excellently, has recently become a promising technique for the study of speech. MRI's contributions in speech research could lead to new and individualized treatment for speech disorders.

Although many studies have shown that MRI can capture information about speech, this project sought to determine what covert information could be disclosed from MRI movies through multivariate analysis. The articulation of phoneme pairs was imaged using a novel sequence, and simultaneously recorded. The data were then analyzed using support vector machine (SVM) analysis and canonical correlation analysis (CCA).

Determination of classification accuracy through SVM analysis revealed that phoneme pairs were distinguishable from one another consistently over 90% of the time using information found from MRI movie clips of the speech. Additionally, study of the SVM weights demonstrated that SVM could identify regions of the vocal tract that are used to form auditory distinctions between the phonemes. Finally, CCA revealed relationships between images and the frequencies in corresponding audio waveforms; once again, the speech articulators were identified as lending maximum correlation to the sound profile.

These promising results demonstrate that multivariate analysis can uncover information that is known to be true concerning speech production. These analyses may perhaps even contribute to existing knowledge and thus provide a platform from which to advance the treatment of speech dysfunction.

Acknowledgements

First of all, I would like to extend a sincere thank you to my advisor and committee chairman, Dr. Stephen LaConte, who has provided invaluable guidance through my research process, and the inspiration to accomplish this achievement. I would also like to thank Dr. Bradley Sutton; his collaborative efforts made all of this work possible. Additionally, I thank my other committee members, Dr. William Tyler and Dr. Alexander Leonessa, for their time and effort, and for modeling excellence in research.

I would also like to express my appreciation for my family and friends, whose support has proven invaluable throughout this process. I would like to thank my parents for their unwavering willingness to listen to my concerns and to help in any way possible. I would like to thank my human study partners, Kelly Donoughe and Ada Tsoi, for keeping me on track, as well my feline study buddies, Scout and Boo, for their supportive purring and snuggling through the late nights. I would also like to thank my roommate, Martha Hay, for her willingness to understand my concerns and her encouragement to maintain an honorable perspective through the difficult times; additionally, I appreciate her readiness to provide thorough and constructive feedback for my writings and presentations. Most of all, I would like to thank my fiancé, Daniel Heckman, who works alongside me, motivates me, cares enough to truly understand the nature of my work, teaches me to respect myself, and inspires me to be better each day.

Finally, I would like to thank my Lord and Savior, Jesus Christ, who has given me the foundation upon which to grow my character, the strength to endure difficulties, and the grace to start clean when I make mistakes. My abilities and dreams are gifts of His goodness; may I use them to serve Him ever more each day.

Attributions

In addition to all of the guidance from my committee, I would like to acknowledge the attributions of other researchers whose help has proven invaluable to the completion of this work.

I would like to thank Dr. Bradley Sutton and his group at the University of Illinois, Urbana-Champaign, for their collaboration that made this research possible. I am honored to have used Dr. Sutton's novel and incredibly powerful sequence for the imaging. I am also grateful for the time that he and his students spent brainstorming, collecting data, sharing resources, and reconstructing the images.

I would also like to thank everyone in my lab for their support and constructive feedback. I would especially like to thank Jonathan Lisinski for his willingness to help in every aspect of the project.

Table of Contents

Acknowledgements ..... **iii**

Attributions ..... **iv**

List of Figures ..... **vi**

List of Tables ..... **vii**

Chapter 1: Introduction **1**
1.1 Motivation *1*
1.2 Project Overview *1*

Chapter 2: Background **3**
2.1 The Basis of Speech *3*
2.2 Methods of Speech Quantification *6*
2.3 MRI Speech Research *8*
2.4 Multivariate Analysis *10*

Chapter 3: Materials and Methods **13**
3.1 Experimental Design *13*
3.2 Data Acquisition *14*
3.3 Image Reconstruction *14*
3.4 Image Analysis Preparation *15*
3.5 Clip Selection *16*
3.6 Support Vector Machine Image Analysis *20*
3.7 Canonical Correlation Analysis *24*

4. Results and Discussion **27**
4.1 Automated Clip Selection Algorithm *27*
4.2 Support Vector Machine Analysis *28*
4.3 Canonical Correlation Analysis *35*

5. Conclusion **39**
5.1 Summary of Work *39*
5.2 Future Directions *40*

References **42**

List of Figures

Figure 1. Major constituents of the vocal tract. 3
Figure 2. Organs of speech. 4
Figure 3. Highest tongue positions of the cardinal front vowels. 5
Figure 4. Highest tongue positions of the cardinal back vowels. 5
Figure 5. Cardinal vowel diagram. 5
Figure 6. Places of articulation. 6
Figure 7. Plot of example SVM data with hyperplane. 11
Figure 8. Example mid-sagittal image from data set. 14
Figure 9. Example audio signal selections. 16
Figure 10. Example cross-correlation waveform and peak identification. 17
Figure 11. Example user display for assisting with clip verification and modification. 18
Figure 12. Zoom area positioning within images. 20
Figure 13. Example movie frames. 21
Figure 14. Formation of a single observation. 22
Figure 15. Spectrograms of /ata/ and /ana/. 25
Figure 16. Spectrograms of /ata/ and /ati/. 25
Figure 17. Spectrograms of /bababa/ and /pataka/. 25
Figure 18. Spectrograms of /ita/ and /iti/. 26
Figure 19. Representative frames from the /ata/ versus /ana/ weight map movie. 30
Figure 20. Representative frames from the /ata/ versus /ati/ weight map movie. 30
Figure 21. Representative frames from the /bababa/ versus /pataka/ weight map movie. 31
Figure 22. Representative frames from the /ita/ versus /iti/ weight map movie. 32
Figure 23. Representative frames from the /ba/ versus /ta/ weight map movie. 33
Figure 24. Representative frames from the /ba/ versus /ka/ weight map movie. 34
Figure 25. Representative frames from the no speech versus /pa/ weight map movie. 35
Figure 26. A and B weights for /ata/ and /ana/. 36
Figure 27. A and B weights for /ata/ and /ati/. 36
Figure 28. A and B weights for /bababa/ and /pataka/. 37
Figure 29. A and B weights for /ita/ and /iti/. 37

List of Tables

Table 1. Original Phoneme Pairs. ____ *13*
Table 2. Additional Phoneme Pairs. ____ *13*
Table 3. Imaging Parameters ____ *14*
Table 4. Runs and corresponding phoneme pairs. ____ *15*
Table 5. Phoneme Clip Sample Size Information. ____ *19*
Table 6. Threshold and color choices. ____ *24*
Table 7. Automated clip selection algorithm results. ____ *27*
Table 8. Classification accuracy results. ____ *29*

Chapter 1: Introduction

1.1 Motivation

The average person takes for granted that speech is both a complex motor function and a skill fundamental to personal well-being. Although those with speech disorders can greatly benefit from surgical or therapeutic approaches, speech remains poorly quantified in terms of brain activity and vocal tract dynamics.

Indeed, in the words of the German psychologist Ebbinghaus, the study of speech production has a long past but a short history[1]. This refers to the fact that the scientific study of speech production has just begun in the past 50 years[1], but has only continued to grow since then; the advent of new imaging modalities has been one contributing factor to this growth.

Magnetic resonance imaging (MRI) remains a powerful and relatively untapped tool for studying speech, as it can dynamically image soft-tissue articulators in the vocal tract with high spatial and temporal resolution. Furthermore, unlike other existing techniques for speech visualization, MRI is non-ionizing, non-invasive, and non-interfering.

Speech formation via articulator motion has been well studied with many modalities, such as x-ray, ultrasound, and electropalatography, and thus is well understood. Many studies have verified that MRI can, likewise, capture information about speech formation. MRI studies of speech may even be advantageous in that speech could be quantified using computational methods as opposed to simple visual inspection of speech formation that is often coupled with other speech imaging modalities. Thus, if MRI can capture meaningful speech information that may be discerned quantitatively through statistical analysis, this could lead to faster and more personalized speech therapy.

Most unique, though, is that MRI can also dynamically track the neurocorrelates of speech in the brain, and thus exhibits the long-term potential to examine brain-articulator relationships during speech production. Understanding where and when the brain goes awry in the speech production process could enable novel developments for the diagnoses and treatment of communication disorders.

1.2 Project Overview

As a first step toward the aims mentioned above, this project sought to determine MRI's capability to capture meaningful information about speech that may not be discernable by visual inspection. To achieve this, data were collected at the University of Illinois, Urbana-Champaign, with the Sutton group; speech was imaged using a novel and extremely fast spiral-navigated Cartesian FLASH sequence, and the resultant auditory production was captured using an optical microphone. The speech consisted of four separate phoneme pairs that had been predetermined to be physiologically interesting based on collaboration with a linguist at UIUC named Ryan Shosted; these included /ata/ versus /ana/, /ata/ versus /ati/, /bababa/ versus /pataka/, and /ita/ versus /iti/.

After image reconstruction, the images required processing and analysis. This involved extracting clips from the MRI movies and the audio data that corresponded to each phoneme. In prior work, this had been accomplished manually, so a goal for this work was to come up with an algorithm that could extract audio selections in an automated fashion.

Once the audio and video clips were prepared, analysis could begin. The first type of analysis was support vector machine analysis (SVM). SVM is a classifier that takes information from training data to make a generalized rule for data classification, and thus can classify new

information. One metric of the SVM's ability to classify data from a data set correctly is called leave-one-out classification accuracy. For these purposes, classification accuracy analysis could provide an indication of the SVM's ability to distinguish between paired phonemes using spatiotemporal data. It was expected that the SVM would be able to extract information from the pixels of the movie (the features), and make mostly accurate classifications.

SVM could provide further insights as well. Technically, SVM weights are a measure of a support vector's distance from the hyperplane; colloquially, this means that a weight is a metric of a support vector's relative importance in determining the distinction between one class and another. For these purposes, the idea was that, if the weight vector could be reshaped into its original space, this would result in a movie of weights; when thresholded and overlaid onto an example movie clip, this information could indicate which pixels in the image were most critical for distinguishing between phonemes. It was expected that these pixels might correspond with speech articulators, as articulatory placement is responsible for physically distinguishing sounds from one another.

The final goal was to use canonical correlation analysis (CCA) to relate image data with auditory data. CCA is a statistical model that works to study interrelationships between multiple independent and dependent variables; it determines sets of weights that, when applied to the variables, lend the maximum correlation between the data. For these purposes, the auditory data were studied as frequency information, which was correlated with the spatiotemporal image data. It was expected that, on the independent variable side, the speech articulators would be the most heavily weighted, and that on the dependent variable side, fundamental frequencies and formants would be the most heavily weighted.

Chapter 2: Background

2.1 The Basis of Speech

The vocal tract. The vocal tract includes all structures through which air is drawn and modified to produce sound; these spaces include the lungs, trachea, larynx, pharynx, oral cavity, and nasal cavity[2], and are shown in Figure 1[3]. The respiratory system and the musculature that controls it are responsible for maintaining heightened subglottal pressure and elongated expiratory phase; this facilitates the continuance of an airstream which can then be modified by the speech articulators to produce speech sounds, or phones[1,2,4].

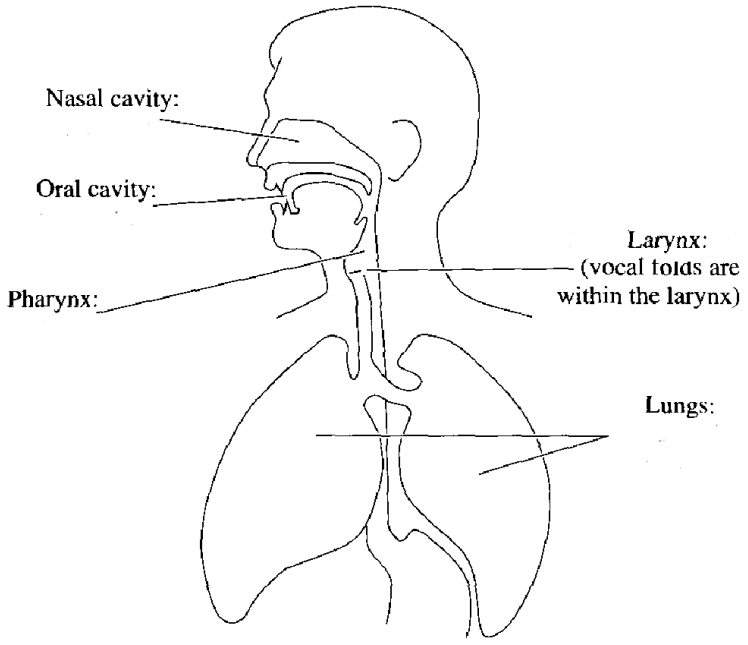

Figure 1. Major constituents of the vocal tract[3].

The larynx contains a group of cartilages that, for phonation purposes, suspend the vocal folds, and thereby control the space between the vocal folds, the glottis[2]. Motion of the arytenoid cartilages adjusts the tension and closeness of the vocal folds[2]. During voiced phonation, the combination of the vocal fold tension and pressure from the airstream cause the vocal folds to vibrate; this provides a resistance to the constant air stream generated from the lungs, changing it into small puffs of air[2]. The size and shape of the glottis, as well as its vibrating frequency, dictate certain aspects of the speaker's voice quality, and contribute to sound differentiation[2].

The supralaryngeal system consists of three cavities, the pharynx, oral cavity, and nasal cavity. The pharynx extends from the top of the larynx up to the oral and nasal cavities; it acts as a resonating chamber, and its shape and size can be altered very little[2]. The nasal cavity is also an unalterable resonating chamber[2]. It is separated from the oral cavity by the velopharyngeal port, which can be open if the soft palate (velum) is lowered, or closed if the velum is raised[2]. In normal breathing, the velum is lowered, whereas in normal speech the velum is closed, with the exception of a few nasal phones such as [m] and [n][2]. The oral cavity is the most adaptable

resonating chamber; its shape and size can be altered by the articulators, which include the tongue, lips, lower jaw, and velum[2]. Some of these organs of speech are shown in Figure 2.

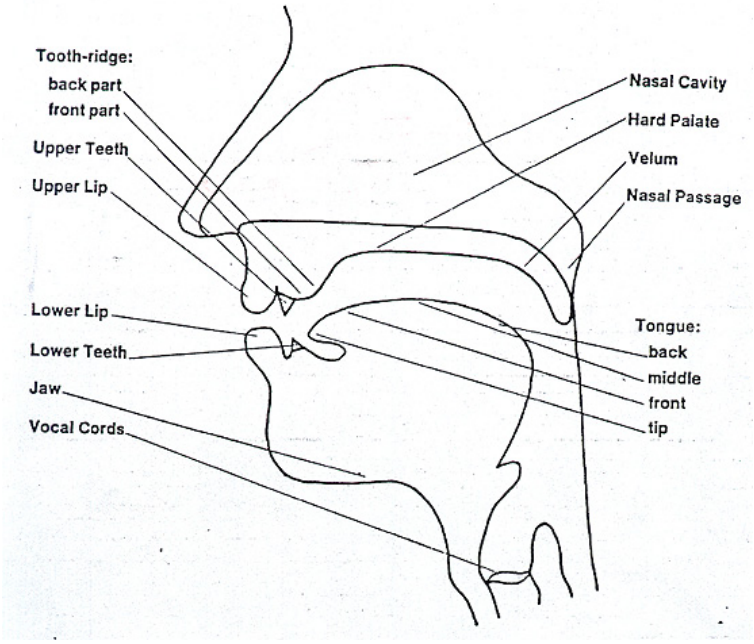

Figure 2. Organs of speech[5].

Speech formation. The primary classification of speech sounds is into the categories of vowel or consonant. Generally, these differ in that consonants involve near or full contact between articulators so as to temporarily halt or resist the air stream, while vowels leave wide spaces between articulators[2].

Distinct English vowels are produced by varying tongue position, tongue height, tongue shape, lip shape, and other factors[2] with the velopharyngeal port closed[6]; the exceptions to this rule are nasalized vowels, which Bae describes in this way: "when vowels occur adjacent to the nasal consonants… they become nasalized due to coupling between the oral and nasal cavities"[6]. Figure 3[7] shows the highest tongue positions of the cardinal front vowels and Figure 4[7] shows the highest tongue positions of the cardinal back vowels. The points on Figure 5[7] show the tongue position during the production of each of the cardinal vowels; diagrammed together, it is easy to see that some vowels require the tongue to be higher or lower, or more towards the front or back of the oral cavity – each vowel is produced in a distinct manner. It is important to note that other vowels exist and are produced from still different tongue positions; the ones relevant to this research, however, are numbered 1 and 5 on the cardinal vowel diagram. An example of the phone [i] is found in the word 'key', and an example of the phone [α] is found in the word 'pop'.

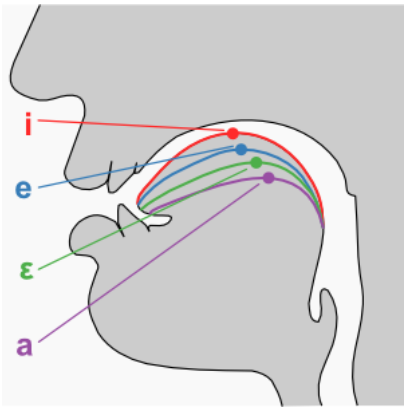

Figure 3. Highest tongue positions of the cardinal front vowels[7].

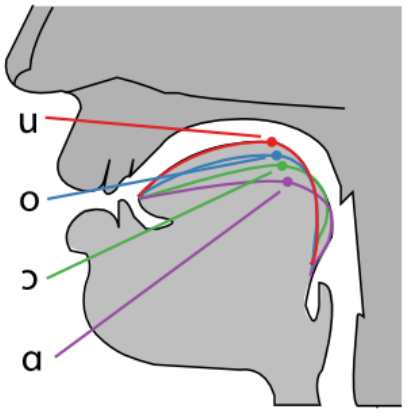

Figure 4. Highest tongue positions of the cardinal back vowels[7].

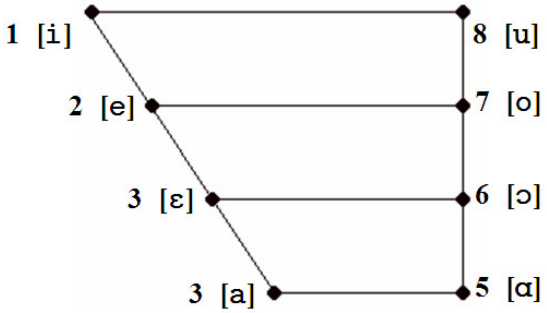

Figure 5. Cardinal vowel diagram[7].

Consonants may be described by the degree of closure between articulators, the place of articulation, as well as other factors. The consonants relevant to this research will be discussed.

Consonants may be distinguished by the degree of closure that occurs between articulators to produce them. Stops are created when the articulators are brought so closely together that an air-tight seal is formed; other types include fricatives, which leave some opening through which air flows turbulently, and approximants, which leave a wide space and result in non-turbulent airflow[2]. There are two types of stops, and each are utilized in this research. Oral stops, or plosives, are characterized by a buildup of pressure in the oral cavity (the velum is raised so that no air can escape through the oral cavity), then a release that results in a popping

quality to the phone[2]. The sounds [b], [p], and [k] are considered to be plosives. Nasal stops, or nasals, differ from plosives in that the velopharyngeal port remains open, so that air may flow through the nasal cavity and no pressure builds up[2]. Because the velum begins to lower in advance of the nasal, and only fully raises after the nasal, the vowels surrounding nasals, such as [m] and [n], will sound nasalized[2].

Consonants may also be distinguished by the place of articulation, that is, which specific articulators close together to form the sound. A diagram of the places of articulation is given in Figure 6[8]. For example, the [b] and [p] phones are considered to be bilabial plosives, because the airstream is stopped by the closure of the upper and lower lips together; the phone [k] is a velar stop, because the airstream is stopped by contact of the blade of the tongue with the velum. The phone [t] is an alveolar plosive, and [n] is an alveolar nasal, because they are formed by sealing off the oral cavity using the front of the tongue and the alveolar ridge.

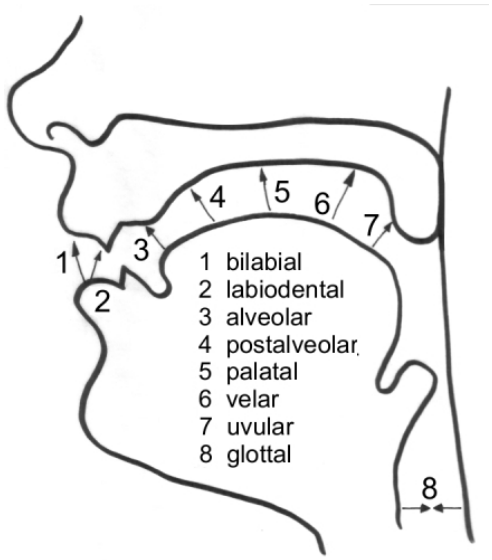

Figure 6. Places of articulation[8].

Spectrograms. So what of all this speech production? The results are auditory productions that may also be analyzed. Most often, sounds are studied using spectrograms, which plot time by frequency, and give intensity in the darkness of a given point.

Each vowel has a frequency signature; the most intense frequencies are described as fundamental frequencies, or formants. For the vowel [a], the first formant (F1) occurs around 720 Hz, the second formant (F2) occurs around 1200 Hz, and the third formant (F3) occurs around 2520 Hz[2]. For the vowel [i], F1 occurs around 360 Hz, F2 occurs around 2280 Hz, and F3 occurs around 3000 Hz[2]. Consonants often appear in spectrograms as blank gaps of no intensity, or as bands of aperiodicity.

2.2 Methods of Speech Quantification

The extensive effort put forth in the past to quantify speech has led to present day understanding of how speech is formed through the vocal tract, and modified by the articulators. Below is a summary of the techniques used to study speech; some methods involve imaging, while others do not.

Russel was the first to propose speech imaging in 1928 through cineradiography[9], which basically produces X-ray movies; this was the most common and reliable speech imaging technique until the 1970's[10]. This technique has been used extensively for the imaging of speech movements, using metal markers attached to the jaw, tongue, and lips; it has also been used to study dysphagia through the swallowing of contrast agent[9]. Unfortunately, this technique does not yield clear mid-sagittal slices, and it emits ionizing radiation[9].

X-ray microbeam imaging allows for reduced radiation exposure as compared to cineradiography; computer prediction facilitates the tracking of gold pellets attached to the articulators by narrow X-ray beams[11]. This technique has been used to examine tongue kinematics at up to 1000 gold pellet positions per second[12]; however, it does not enable full vocal tract imaging[9], and the equipment is available to few[11].

Ultrasound has also been used for vocal tract imaging, specifically for capturing two-dimensional contours of the dorsal surface of the tongue and tongue motion[11]. This is implemented by placing the transducer on the patient's neck[13]. This technique has been applied to swallowing studies and, occasionally, for biofeedback in speech therapy[12,13]. It is a relatively inexpensive modality that is noninvasive, as well as safe and comfortable for patients[9,13]. The drawbacks to this technique include that it is unable to capture the tongue tip and is limited by air-tissue interfaces[14], so that the palatal outline may not be discerned[15]. Future advances with this modality include three-dimensional ultrasound to more completely capture the tongue's shape; this has been implemented in a few feasibility studies to date[13].

Electromagnetic articulography (EMA) represents another technique that has been used to track speech movements and swallowing since the 1980's[11]. In this procedure, tiny transducer coils are attached at multiple locations to the midline of the tongue; reference coils can be attached to the upper and lower jaw, as well as the nose[11]. The subject is placed in an alternating magnetic field created by a few transmitters, which induces alternating voltage in the coils[11]. Because the amount of voltage induced is directly related to the distance between the transmitter and transducer, the position of the transducers can be traced by the articulograph[11]. This method provides an accurate means to measure tongue movement, but it cannot capture multiple articulators, cannot visualize tongue-palatal contact, and requires hardware that may interfere with speech production.

Electropalatography is useful for the identification of tongue-palate contact patterns[11], which is information that cannot be captured by ultrasound or EMA. This technique employs an artificial palate embedded with many electrodes, which register palatal contact[14]. The palatogram can then represent the spatial and temporal patterns of tongue-palate contact[16]. This technique is particularly useful for the studies of alveolar, palatal, and velar phones, but it cannot necessarily capture other phones; another drawback is that this technique is very invasive and may interfere with speech production[16].

Nasendoscopy and videofluoroscopy represent two techniques that are used for evaluation of velopharyngeal dysfunction, which is intrinsically linked with speech production[17]. Videofluoroscopy is a radiologic technique that is most commonly used for the evaluation of dysphagia (swallowing dysfunction)[18], but has also been used to study tongue movement[12], and can achieve a frame rate of 30 frames per second[12]. Nasendoscopy, a procedure which involves the insertion of a small camera into the nasal cavity for visualization of the velum, has been used successfully as a behavioral modification technique during speech therapy[19]. Its limitations include the invasive nature of the procedure, and the ability to visualize only the velum[20].

Computed tomography has also been used for visualizing laryngeal anatomy during phonation[21]. Like MRI, it is noninvasive and capable of capturing the entire vocal tract at once, but has only been used for a limited number of studies due to its irradiating nature[9].

Magnetic resonance imaging (MRI) has only recently been applied to speech imaging, but its popularity has grown quickly. Briefly, MRI works by measuring signal intensity based on the alignment of protons in a magnetic field. This technique's relation to speech research will be described further below.

2.3 MRI Speech Research

The following section will review MRI speech research as is presented in the literature, including a brief history of imaging approaches, research objectives, quantification approaches, and advantages and disadvantages of the technique.

Imaging approaches. MRI was first applied to speech in the late 1980's[10]. At that time, it was a promising and novel proof of concept, but only recently have technological advances pushed it to the forefront of speech research. As little as ten years ago, dynamic imaging of speech was not feasible; instead, subjects would perform sustained phonation by taking a deep breath and holding a vowel phone until an image could be acquired[22,23]. This technique has been used for the evaluation of velar function, the measurement of tongue shape and movement, and the study of vocal tract modeling; however, it is "considered not to reflect the reality of the velar portal during normal speech[20]."

As an alternative approach to imaging sustained phonation, a sequence would employ gating so as to reconstruct an MRI movie from data of multiple repetitions of a given word or sentence in order to achieve sufficient temporal resolution[6,12,14]. This approach is obviously not ideal because of inherent inaccuracies, such as the fact that speech is not periodic, unlike the cardiac motion that gating was developed for[12]. Additionally, the necessity of multiple repetitions poses a challenge to children or those with speech difficulties[6]. Nonetheless, this technique has been applied to many different types of speech research, including tongue motion, compensatory speech, and coarticulation studies[12]. Alternatively, sliding windows can be used to resample image data at higher frame rates, but this does not contribute any new information[6].

In the past ten years, MRI temporal resolution has progressed sufficiently so as to allow for dynamic, real-time MR imaging of speech[9]. Dynamic refers to the fact that the speaker is actively and naturally articulating, and not merely holding a sound or mouth position. Real-time refers to the fact that image acquisition occurs simultaneously with the speech production; images are not reconstructed after data acquisition from multiple repetitions of a word. These advances have enabled the application of MRI to a broad range of speech-related research, such as three-plane imaging of speech[24], and even the imaging of swallowing and singing[25]!

Because velar movement can occur within 100 to 150 ms, a frame rate of 20 images per second is required in order to sufficiently capture these speech movements[5]. Thus, fast low-angle shot (FLASH) imaging techniques are frequently used for rapid imaging[6,26], as are spiral data sampling approaches[14]. Nonetheless, research as recent as 2010 to 2012 achieves frame rates of only 2[20,26] to 17[27] frames per second, with leading frame rates between 21[28] and 30[6] fps. However, recent technological advances have demonstrated the capabilities to image at significantly higher frame rates; using parallel imaging, sparse sampling, and compressed sensing, one study was able to obtain a frame rate of 20 fps for a stack of five slices[29]; similar techniques were used for the data acquisition of this work, which achieved a frame rate of 102.2 fps.

Another sought-after objective has been to synchronize audio and imaging[26]. This has been performed by recording images first, and recording speech afterwards, although this approach is obviously prone to inconsistencies and speaker fatigue. Thus, a push has been made to simultaneously acquire image and audio information; barriers to this achievement included MRI compatibility with microphone equipment, and background scanner noise interfering with the speech signal. Now, however, MRI-compatible microphones have made this possible.

Advantages and disadvantages. MRI's unique strengths make it a powerful tool for imaging the vocal tract, and now that temporal resolution has exceeded the speed at which speech is typically produced, it possesses untapped potential for imaging speech in real time. MRI exhibits excellent soft tissue resolution, can visualize vocal tract musculature, and can acquire images in three dimensions[27]. Furthermore, it permits extended data collection without risk of significant biohazard, which gives this imaging method a distinct advantage over those that employ ionizing radiation, such as x-ray, videofluoroscopy, and CT[20,22,27]. It is noninvasive, unlike nasendoscopy and electromagnetic articulography[20,22]. In fact, it is the only imaging modality that yields dynamic images of the entire vocal tract, including every articulator[14].

The major drawback to this technique is its expense and inaccessibility; additionally, some would argue that the subject's supine position during imaging alters the position of the velum[20,21], and the noise of the scanner inhibits recording of the subject's auditory production[3,]. Others would contend that the lengthy data acquisition time presents another disadvantage to this technique[6], or that air-tissue interfaces can create susceptibility artifacts[32], although certain reconstruction techniques attempt to correct for this. Finally, as the bones and teeth contain small concentrations of mobile hydrogen, they produce little signal and are thus nearly invisible in MR images; however, this is mostly only problematic when implementing contour extraction of the oral cavity[9].

Research objectives. Despite any drawbacks to MR imaging of speech, clinically, there are many populations that could benefit from, and are of interest to, speech imaging research. One such group includes cleft lip and palate (CLP) patients, who can exhibit compensatory misarticulations, which are the articulatory substitutions of one sound for another related sound (such as interchanging plosives)[33]. Velopharyngeal closure is also of interest in this subpopulation[33]; many exhibit velopharyngeal insufficiency, or VPI, although VPI is not strictly limited to CLP patients. This is a condition characterized by the inability to completely close the port between the oral and nasal cavities[20], resulting in severe speech distortion[17], including hypernasality, nasal emissions, compensatory articulations, and aberrant facial grimacing[6,26].

Glossectomy patients, those who have undergone a partial or total surgical removal of the tongue, represent another population with speech difficulties. This population motivates much research regarding tongue shape and contours, as their speech is disturbed from the surgery and thus requires therapy[12,27]. Those who are still in need of surgical reconstruction of structure or musculature may also benefit from speech studies[20]; for example, the ability to correctly identify muscular dysfunction could indicate what type of surgical intervention may be most appropriate[26].

Additional populations of interest include dysarthritic patients, who suffer neurological speech trauma from a variety of causes such as stroke. Studies of dysphagia, or difficulty swallowing, are becoming more common with MRI[34], as are studies of those suffering from Parkinson's disease[35]. In the future, MRI could be used as a biofeedback tool for oral deaf speakers[13] or others.

From a scientific perspective, MRI has been employed extensively for vocal tract modeling studies; this has been done by calculation of vocal tract area and volume[9], or in combination with three-dimensional segmentation techniques[36]. Additionally, MRI may help to more fully explain the phenomenon of coarticulation[27], the situation in which conceptually isolated speech sounds are influenced by adjacent sounds.

Quantification techniques. There exists a lack of a standardized quantitative measure of speech images, so most research employs new and clever ways to quantify their work. Vocal tract area functions provide one mechanism for representing the vocal tract's shape and motion during speech[9]. These have been estimated using mid-sagittal images since cine x-ray has been used to image speech[14]. Some studies have looked at such physiologic features as lip, velar, and tongue positional changes[6,27], while others have considered the angle of the velum in relation to other physiologic features[20]. Certain acoustic features of interest have included peak amplitude and bandwidth of the first resonant frequency[6].

2.4 Multivariate Analysis

The two multivariate analysis techniques of interest to this research include support vector machine (SVM) analysis and canonical correlation analysis (CCA).

Support vector machines. Briefly, the support vector machine is a learning machine that classifies new data using rules derived from training data. The data set, X, is comprised of observations (for example, perhaps, x_1, x_2, x_3, might represent three flowers). Each observation consists of features that every observation possesses (for example, flower petal length and petal width). Each observation is also assigned a class. Often, this is a binary division (for example, a flower might belong to either the *virginica* or *versicolor* iris species, perhaps expressed as +1 or -1); these data are contained in Y.

The SVM's role, then, is to learn from the data pairs ((x_1, y_1), (x_2, y_2), etc.) so that it may predict the y value associated with an arbitrary x[37]. It does this by determining a discriminating hyperplane that geographically separates the data, so that new data for classification may be plotted, and identified based on what side of the hyperplane it falls on. An example image depicting plotted X data, along with the SVM-determined hyperplane, is shown in Figure 7. The equation for determination of the hyperplane is expressed in Equation 1.

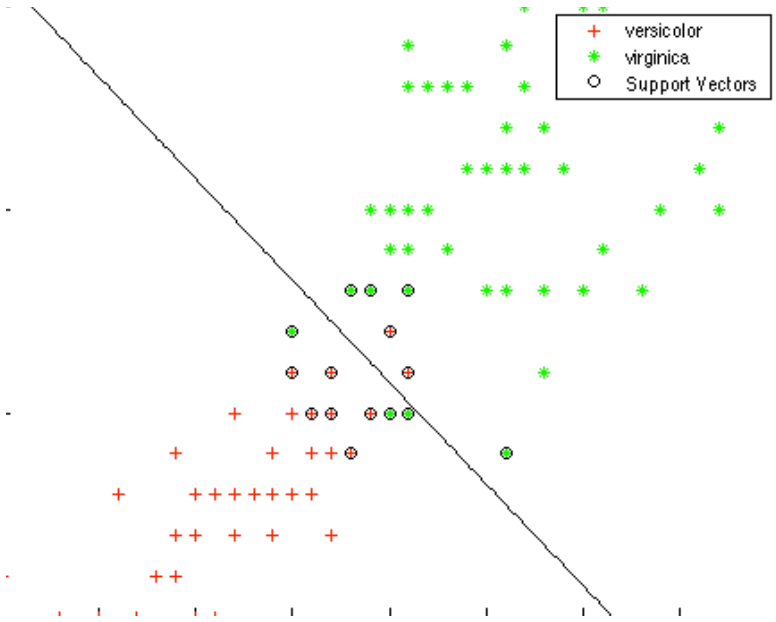

Figure 7. Plot of example SVM data with hyperplane.

$$\vec{w} \cdot \vec{x} - b = 0$$

Equation 1. Hyperplane determination.

The margin is considered to be a band around the hyperplane that SVM seeks to minimize. The x data that fall within the margin are considered to be the support vectors; these are the only data that contribute to hyperplane determination. The support vectors are circled in Figure 7. The equation for the margin is given in Equation 2.

$$\vec{w} \cdot \vec{x} - b = \pm 1$$

Equation 2. Margin determination.

The SVM also determines a set of alpha coefficients, one for each support vector, that indicates that support vector's relative importance to hyperplane determination. Alpha values are positive for one class and negative for another. Weights, or the support vectors' perpendicular distance to the hyperplane, are determined by multiplying the alpha vector by the support vectors to result in a vector that has one dimension equal to one, and another dimension equal to the number of features. This action effectively multiplies each support vector by its relative importance, then sums each feature across all of its observations; this yields information regarding the relative significance of each feature.

Classification accuracy is a parameter of interest to SVM analysis. Leave-one-out classification accuracy, in particular, is calculated by training an SVM on all but one pair of data (that is, one x from each class), then classifying that pair of data. This is repeated until each observation has been left out one time; then, the percentage of observations that were classified correctly is determined.

Canonical correlation analysis. CCA is a statistical model that studies interrelationships between variables. Given a set of independent data $X = (x_1, x_2, x_3 \ldots x_n)$ and a set of dependent data $Y = (y_1, y_2, y_3 \ldots y_n)$, the goal is to determine a and b vectors of coefficients so that x' and y'

maximize the correlation ρ[38], where x' and y' are determined as shown in Equations 3 and 4, and the equation for ρ is given in Equation 5.

$$x' = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

Equation 3. Weighted independent variable set.

$$y' = b_1y_1 + b_2y_2 + \cdots + b_ny_n$$

Equation 4. Weighted dependent variable set.

$$\rho = \frac{E[x'y'^T]}{E[x'x'^T]E[y'y'^T]} = \frac{a^T C_{xy} b}{\sqrt{a^T C_{xx} a b^T C_{yy} b}}$$

Equation 5. Correlation maximization.

Chapter 3: Materials and Methods

3.1 Experimental Design

The group had previously met with a speech expert, Ryan Shosted, from the University of Illinois, Urbana-Champaign, who suggested phoneme pairs that would be visually interesting for the subject to repeat while in the scanner. Phoneme pair information is listed in Table 1.

The first phoneme pair, /ata/ vs /ana/, was considered to be visually interesting because of the variation in consonant, as [t] is a voiceless alveolar stop, and [n] is an alveolar nasal. The second phoneme pair, /ata/ vs /ati/, was believed to be visually interesting because of the potential to view preparation differences; /ata/ and /ati/ both begin the same way, but it is possible that their beginnings are formed differently in the oral cavity because of the difference in phoneme ending.

The next pair, /bababa/ vs /pataka/, are phonemes that have been well studied in the literature. It was believed that, by studying these pairs, a 'roll' from the front to the back of the oral cavity could be visualized. This is because the bilabial stop [p] is formed by the lips, the [t] is formed further back in the oral cavity, by the tongue tip's contact with the alveolar ridge at the front of the hard palate, and the velar stop [k] is formed by the body of the tongue contacting the velum, which is even further back in the oral cavity; hence, a progression from the lips to the back of the oral cavity can be observed.

Finally, the phoneme pair /ita/ vs /iti/ was considered visually interesting for the same reason as the second word pair, for the potential preparation effect.

Table 1. Original Phoneme Pairs.

Phoneme 1	Phoneme 2
/ata/	/ana/
/ata/	/ati/
/bababa/	/pataka/
/ita/	/iti/

In addition to these four phoneme pairs, several other pairs were set up after data collection to strengthen and help understand the analysis. These were derived from sections of the existing data, and are listed in Table 2.

Table 2. Additional Phoneme Pairs.

Phoneme 1	Phoneme 2
/ba/	/pa/
/ba/	/ta/
/ba/	/ka/
no speech	/ba/
no speech	/pa/
no speech	/ta/
no speech	/ka/

3.2 Data Acquisition

All MRI scanning took place at the University of Illinois, Urbana-Champaign with the lab's collaborators, the Sutton group. A novel, custom spiral-navigated Cartesian FLASH sequence developed by Dr. Sutton was implemented on a 3 Tesla Siemens Trio MRI scanner; imaging parameters are outlined in Table 3. This sequence interleaves navigator acquisition with imaging acquisition, making for an effective repetition time (TR) of 9 ms.

Table 3. Imaging Parameters

Parameter	Value
Field of View	280 mm x 280 mm
In-plane Resolution	2.2 mm x 2.2 mm
Slice Thickness	6.5 mm
Echo Time	2.3 ms
Repetition Time	4.5 ms

Eight runs of data were acquired during the imaging session (two runs for each phoneme pair). The sequence described above was used to capture MRI movies of a single mid-sagittal slice of the subject's head and neck (an example of which is shown in Figure 8), and achieved a striking frame rate of 102.2 frames per second. While in the scanner, the subject was asked to clearly repeat a specified phoneme pair, alternating between the first and second phoneme, for just over two minutes. Initially, the subject was instructed to follow a metronome beat that was played into the subject's headphones, but this was found to be unhelpful over the scanner noise and the subject's own speech noise.

The subject's auditory production was recorded simultaneously using the Optoacoustics FOMRI fiber optic microphone. This microphone, which is specially designed for use in the MRI, records sound close to the user's mouth, and also farther from the user's mouth, so as to obtain the loud background scanner noise, which can then be subtracted from the user recording so as to more clearly hear the user's speech.

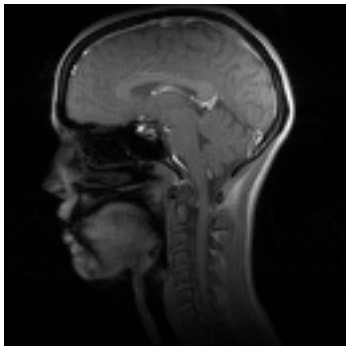

Figure 8. Example mid-sagittal image from data set.

3.3 Image Reconstruction

Image reconstruction took place at the University of Illinois, Urbana-Champaign, using GPUs and fast novel reconstruction methods developed by the Sutton group. Image reconstruction makes use of spatiotemporal correlations in the data using a partially separable model and compressed sensing[29,39,40]. The navigation data determines the temporal basis functions for the image time series, and the imaging data is used to fit spatial maps

corresponding to the temporal basis functions. The image reconstruction is such that the first 500 images from each run are thrown out and merely appear as zero-intensity images.

3.4 Image Analysis Preparation

The data received from the Sutton group for each of the eight runs included the audio recording of the subject's speech, the background scanner noise audio recording, 11452 Portable Network Graphics image files, and an Audio Video Interleave movie of the images. The audio was sampled at a sampling frequency of 8000 samples per second, and, as previously mentioned, the images were obtained at a frame rate of 102.2 frames per second. Table 4 shows the run number that corresponds with a given phoneme pair. Note that the additional phoneme pairs that were listed in Table 2 all come from the fifth run. Although the data included two runs of each phoneme pair, for the purpose of simplifying analysis, mostly only the second runs of each phoneme pair (runs 5-8) were used. For the rest of this document, it may be assumed that any reference to a phoneme pair refers to its second run (5, 6, 7, or 8, respectively). Any reference to runs 1-4 will be explicitly stated.

Table 4. Runs and corresponding phoneme pairs.

Scan Number	Phoneme Pair
1	/bababa/ vs /pataka/
2	/ata/ vs /ati/
3	/ita/ vs /iti/
4	/ata/ vs /ana/
5	/bababa/ vs /pataka/
6	/ata/ vs /ati/
7	/ita/ vs /iti/
8	/ata/ vs /ana/
5	/ba/ vs /pa/
5	/ba/ vs /ta/
5	/ba/ vs /ka/
5	no speech vs /ba/
5	no speech vs /pa/
5	no speech vs /ta/
5	no speech vs /ka/

All data processing and analysis were performed using MATLAB (R2012a, The Mathworks, Natick, MA). Audio data manipulation simply involved subtracting the background recording from the speech recording. Image analysis included support vector machine (SVM) analysis and canonical correlation analysis (CCA). Before these analyses could be performed, information needed to be extracted from the raw image data.

First, MATLAB structures were set up so that related information could be preserved together. One structure was created for each of the eight runs. The structure was set up to hold information such as the auditory waveform, the images, and other related information that would be added as the data were processed in order to perform the analyses. Then, the audio file, the audio frequency, and each image file were read in to MATLAB and properly stored in the structure.

3.5 Clip Selection

The next step was to select which sections of the audio track (audio clips) and which images (movie clips) corresponded with each of the phonemes. It was decided to perform the audio selections first because of the ease of both hearing the words and also visualizing them in the audio signal.

Template audio clip selection. First, an example audio clip of each phoneme was manually selected using custom code that would play a section of the audio track specified by the user and also display the corresponding audio waveform; viewing the waveform provided visual validation that the user was properly selecting a range of the auditory waveform that represented a phoneme. As shown in Figure 9, which shows the clip templates for "ata ana," "ata ati," "bababa pataka," and "ita iti," speech intensity is readily distinguishable from the low-level background noise for most of the phonemes, so the visualization of the waveform provided a more quantitative way to select the clip than by simply listening; for those phonemes for which speech intensity is not significantly different, waveform visualization was a helpful tool for attempting to distinguish between speech and noise. Note that low signal-to-noise ratio for the phone [i] has been reported in the literature[6], and is exhibited here as well.

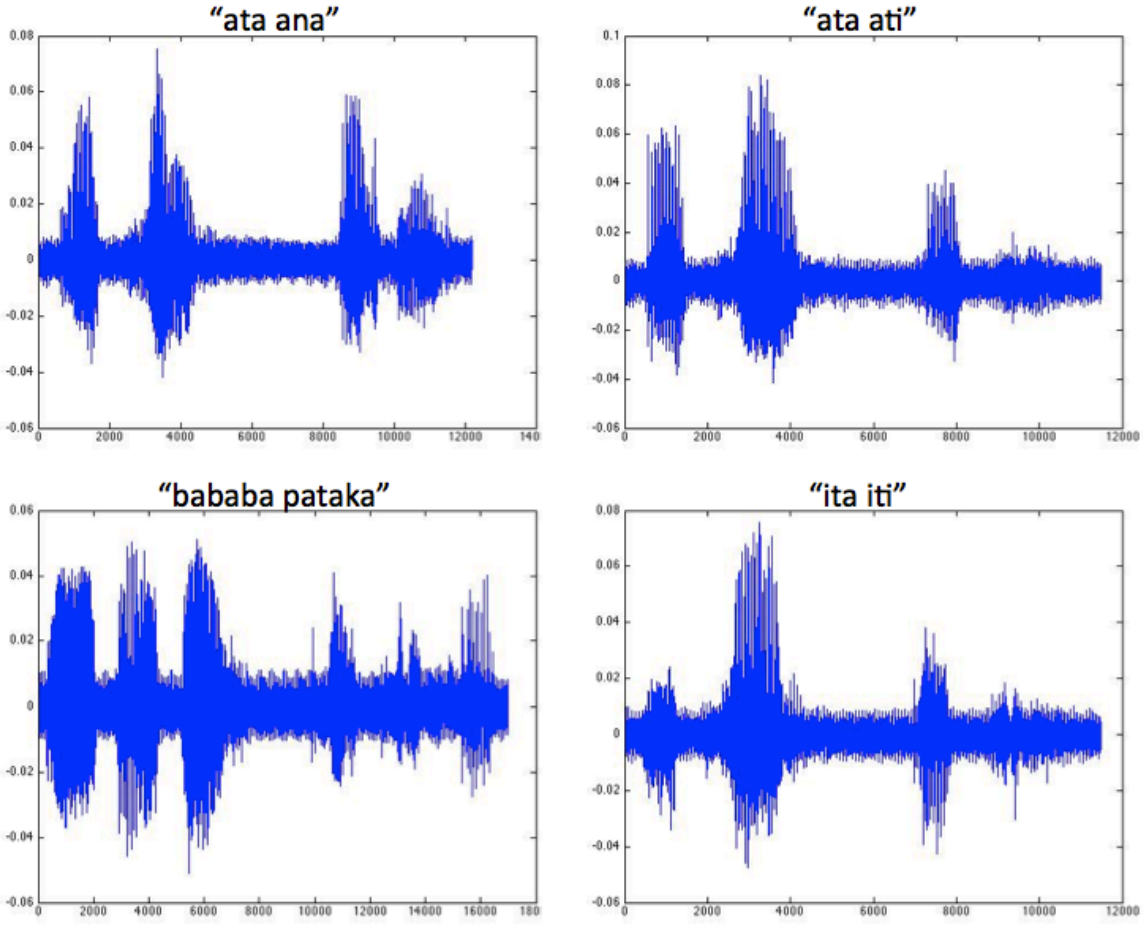

Figure 9. Example audio signal selections.

The code was set up so that a clip could be fine-tuned until determined to be satisfactory. At this stage, the length of each audio clip was determined; for this reason, occasionally this template selection process needed to be repeated if a different clip length was found to be better. It was important for the SVM analysis that the clip length of each phoneme within a run was the same; that is, if the first phoneme was 5000 samples long, the other phoneme needed to be 5000 samples long as well, and all subsequent phoneme instances needed to be 5000 samples long. In order to maintain this consistency, clips of shorter phonemes had a little more background noise incorporated at the end of the clip than their longer counterparts; however, the all of the phonemes were approximately the same length as their pair, so this is not believed to pose a problem. The finalized example clip would next serve as a template for further clip selection.

Automated audio clip selection. Once an example clip of a phoneme had been obtained, it was desired that an automated method could be used to help identify each of the rest of the clips of that phoneme. To achieve this, the clip waveform was cross-correlated with the audio waveform. This yielded a waveform that had peaks where the two signals were highly correlated, and thus, where another instance of the phoneme was likely to be.

Built-in MATLAB functionality was used to identify the peaks of this waveform; to use this functionality, the user may identify a number of features to be considered when the algorithm selects peaks. The two features used for this application were peak height and inter-peak distance. The user relied on visual inspection to select a height requirement for a peak to be considered a peak of interest, and knowledge of audio clip length, along with visual inspection, were used to estimate a required distance between peaks. These parameters were found to make sense in the context of the problem, and generally work well for this application of selecting audio clips. Figure 10 shows an example of the waveform cross-correlation (here, the "bababa" template with its full audio signal), along with the peak identification, as denoted by the red dots.

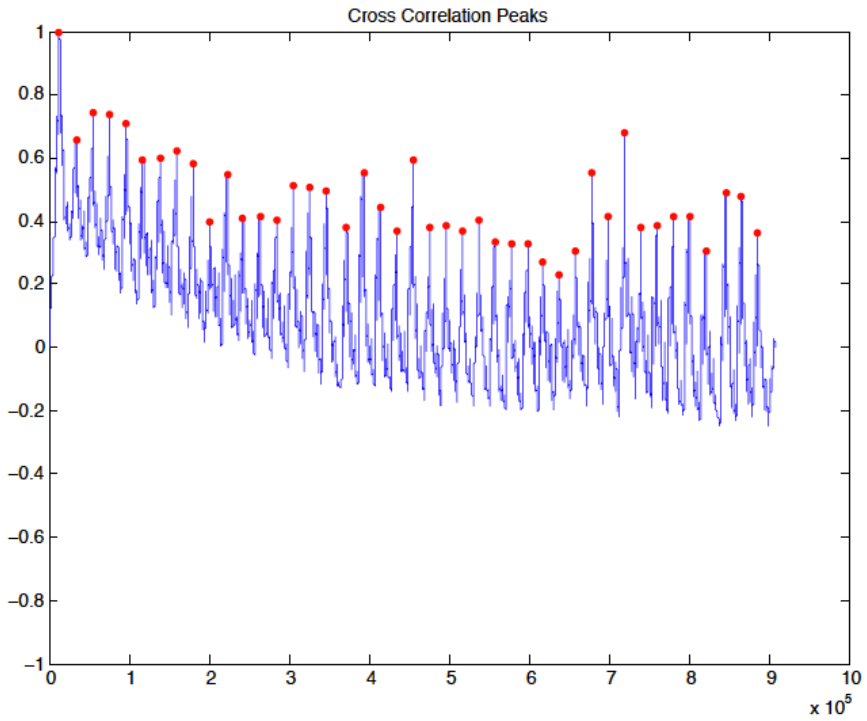

Figure 10. Example cross-correlation waveform and peak identification.

Once the algorithm had identified peaks of interest, custom code was written to translate this peak location information back into audio waveform range values that were determined to be highly correlated with the example phoneme audio clip.

Manual clip verification and modification. Next, each clip selection was manually verified, and modified as needed, so as to eliminate erroneous clip selections. Custom code was written to create a display that would aid the user in verifying satisfactory clips and modifying unsatisfactory ones. Figure 11 shows an example of this display for visualization of a "bababa" clip. The top graph shows the entire audio signal from the end of the last clip to the end of the current clip (here, this includes a "pataka" in the middle); this was useful for determining if the algorithm had missed a phoneme instance. The bottom graph shows the current clip; this was useful for determining if the algorithm had sufficiently captured the entire phoneme signal. It was desirous to have some background noise flanking the phoneme signal on each side so as to ensure that the phoneme was adequately captured. The result of this step of manual clip validation was knowledge of a phoneme's beginning and ending audio sample numbers; this was known for every instance of both phonemes in that run's audio track.

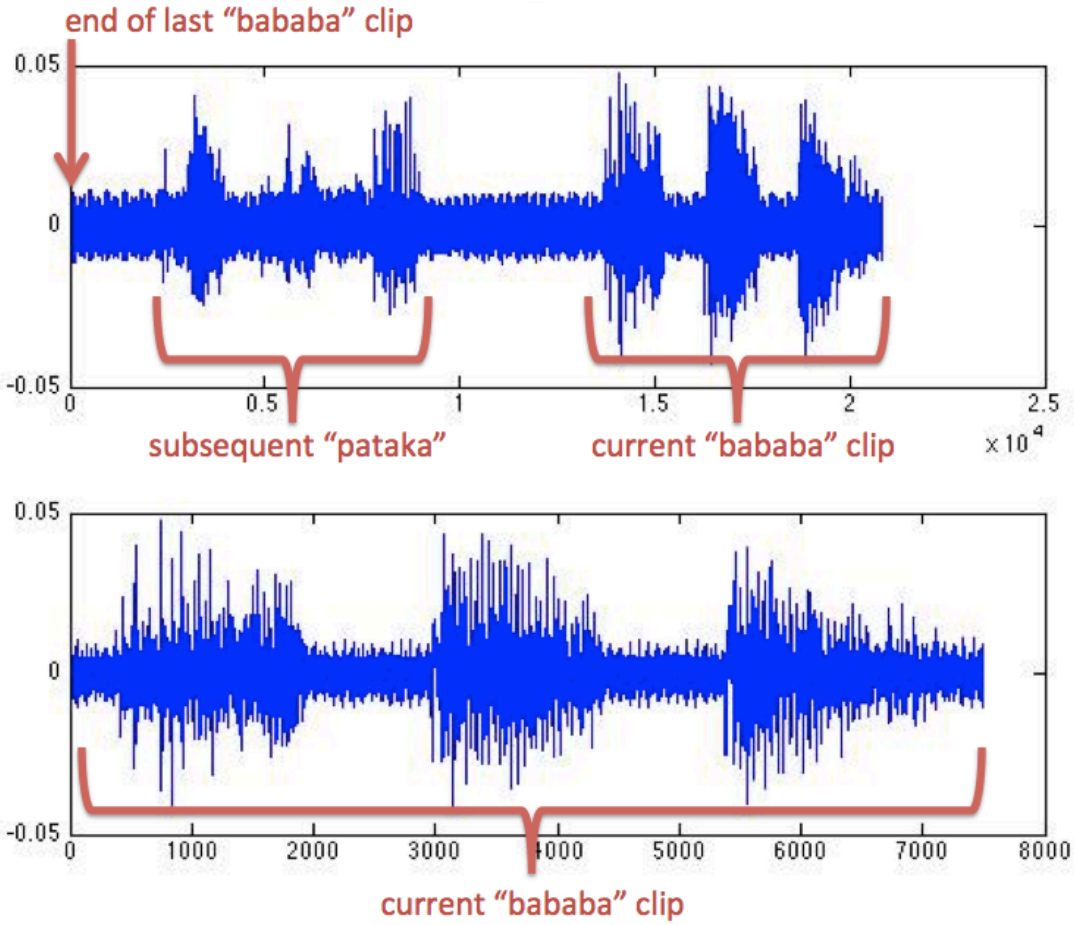

Figure 11. Example user display for assisting with clip verification and modification.

Movie clip selection. After determining all of the audio clips of each phoneme in a given run, the next step was to identify the images that corresponded with each audio clip. This was

approximated by taking the lower and upper audio sample limits of a given clip, and multiplying those numbers by a factor; this factor was determined by dividing the total number of images in the two-minute run (11,452 images) by the total number of audio samples in the two-minute run (900,001 samples). This gave the range of the image samples that corresponded to the audio sample that was representative of a given instance of a phoneme. For example, from run number 4, the first /ata/ vs /ana/ run, the last /ata/ is represented in the audio signal by audio samples 883,500 to 888,000. Multiplying these numbers by 11,452 and dividing by 900,001 (then rounding) results in the image numbers that correspond to that audio clip; here, 11242 to 11299.

Each audio clip did not have a movie clip analog, however. Because the image reconstruction requires the first 500 images to be thrown out and replaced with zero-intensity images, the number of complete phoneme pairs from the audio track differs from the number of complete phoneme pairs observed in the image data. Audio clips that did not have a corresponding image clip were not used for further data analysis; additionally, phonemes that were not appropriately paired were thrown out, because later analysis would require the same number of instances of each phoneme.

Some summarizing information, including the sample sizes of audio phoneme pairs and image phoneme pairs are given in Table 5. By multiplying by the number of samples per clip by the audio sampling frequency (8000 samples/second), one can determine the time length of each clip in seconds; this information is also listed in the table. Note that there are fewer complete pairs for the phonemes /ba/, /pa/, /ta/, and /ka/ versus no speech; this is because there were few clips of "no speech," which were chosen as audio or images that belonged to no phoneme, that were as long as the phonemes.

Table 5. Phoneme Clip Sample Size Information.

Scan #	Phoneme Pair	# Complete Audio Pairs	# Complete Image Pairs	Audio Clip Length (samples)	Image Clip Length (frames)	Clip Length (s)
1	/bababa/ vs /pataka/	36	35	11001	128	1.252
2	/ata/ vs /ati/	56	54	5001	64	0.625
3	/ita/ vs /iti/	54	51	4001	51	0.5
4	/ata/ vs /ana/	56	53	4501	58	0.5625
5	/bababa/ vs /pataka/	43	41	7501	96	0.9375
6	/ata/ vs /ati/	54	51	5001	64	0.625
7	/ita/ vs /iti/	53	50	5001	64	0.625
8	/ata/ vs /ana/	55	52	4501	58	0.5625
5	/ba/ vs /pa/	43	41	2001	26	0.25
5	/ba/ vs /ta/	43	41	2001	26	0.25
5	/ba/ vs /ka/	43	41	2001	26	0.25
5	no speech vs /ba/	42	41	2001	26	0.25
5	no speech vs /pa/	42	41	2001	26	0.25
5	no speech vs /ta/	42	41	2001	26	0.25
5	no speech vs /ka/	42	41	2001	26	0.25

3.6 Support Vector Machine Image Analysis

Once the audio and video clips had been determined for each instance of each phoneme of a given run, analysis could begin. The first analysis performed was support vector machine analysis, with a goal of seeing if the computer could distinguish images corresponding to a given phoneme from images that corresponded to the phoneme's pair – computerized "mouth reading." A complementary goal was to examine the resultant weight vectors and map them back into their original space (image-in-time space, or movie space).

SVM analysis preparation. First, to prepare the images for analysis, it was advantageous to zoom in on the vocal tract region of the images to eliminate noise. This was performed as a simpler alternative to masking the vocal tract region. A 46 x 46 pixel region was selected from each image (the same pixels for every image, across all runs) that was considered to capture the whole mouth and vocal tract, and the rest of each image was not analyzed using SVM. Figure 12 shows an example of the zoomed-in area's location within the whole image.

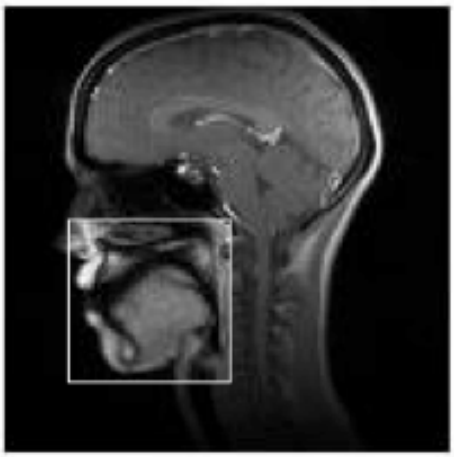

Figure 12. Zoom area positioning within images.

MATLAB's built-in support vector machine functionality was used for analysis. This function takes as inputs (1) a training data matrix (with each row as an observation, and each column a feature) and (2) a column vector containing class information.

The idea going in to SVM analysis was to use not individual images, but movie clips as observations; this would result in spatiotemporal data, as the movie clips possess two dimensions in space, plus the time dimension. Previous work indicated that this spatiotemporal method would be significantly more successful than using each image as an observation.

When the movie clips were initially selected, care was taken to ensure that a phoneme had the same number of movie clips as its pair, and that each movie clip was comprised of exactly the same number of images. It was essential for the phoneme pairs to have the same number of movie clips for leave-one-out classification, the technique upon which the classification accuracy is based. This method is only valid if there are equal numbers of observations of each class in the training data set; this technique also requires the classification of a pair (one observation of each class) after the SVM has been trained, for determining classification accuracy. It was essential that each movie clip contained the same number of images so that each observation would possess the same number of pixels, and therefore the same number of features. This is another requirement for SVM analysis.

The majority of the work for the SVM analysis consisted of transforming movie clips, which are three-dimensional data, into row-vector observations. This process will be portrayed in figures using example data that has been construed to demonstrate the concept. Figure 13 shows the three frames of an example movie that will be transformed into an observation. Example frames like these were used during the programming process for algorithm verification.

Figure 13. Example movie frames.

MATLAB's *reshape* function was used to first transform every image that was part of a movie clip into a column vector; the function works by reading pixels down the image, column-wise, and essentially appending the second column onto the end of the first, the third onto the end of the second, and so on. The results of transforming the example movie frame images in this manner are shown on the left in Figure 14.

Once every image was represented as a column vector, a movie clip matrix was formed by stacking each column-image from a given movie clip side-by-side, in order of appearance in the movie; set up like this, each column of the matrix was all the pixels from an image, and each row of the matrix was the time course of a given voxel over the movie clip. For example, the middle image of Figure 14 shows the movie clip matrix for the example data; note that the top row contains the top left pixel from each frame, the second row contains the pixel below that one from each frame, and so on.

Finally, the movie clip matrix was transformed in the same way that each image was transformed, using MATLAB's *reshape* function, to form a single column vector that contained every pixel from every frame of the whole movie clip, as shown on the far right in Figure 14. The last step was to transpose the column vector into a row vector, as this was the format required by MATLAB's SVM functionality.

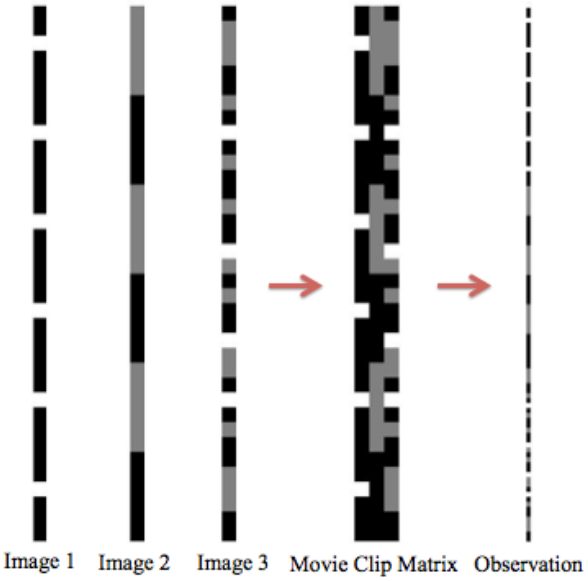

Figure 14. Formation of a single observation.

When every movie clip observation for both phonemes had been transformed into a row vector observation, the row vectors were put together into a training data matrix so that each row was still an observation, and each column was a voxel in time across observations; that is, the first column consisted the top left pixel from the first frame in each movie clip. A matching class column vector was made to indicate which class (which phoneme) the observation in the corresponding row belonged to.

Classification accuracy analysis. The first objective was to determine leave-one-out classification accuracy. To achieve this aim, the training data matrix was modified so that two observations, one from each class, were deleted from the matrix. The two corresponding rows of the class vector were deleted as well. MATLAB's *svmtrain* function was then used to train the support vector machine on this reduced training information.

The *svmtrain* function works to train an SVM classifier, and returns relevant information in a structure such as the support vectors, the support vector weights, and more. Other parameters that were specified included (1) a box constraint, or c value, of epsilon to the negative one-half power, as recommended based on previous work, (2) a linear kernel, which was the default option and was deemed to work sufficiently for these purposes, and (3) the option to have MATLAB normalize the data across columns (so, for these purposes, normalizing a given voxel across its time course) by subtracting the mean and dividing by the standard deviation of the column's values; analysis was performed both with and without data normalization.

Next, the deleted training matrix observations were used as inputs to the MATLAB function *svmclassify,* which takes information from the trained SVM structure (the result of *svmtrain*), and uses it to classify new data. This function outputs class names that correspond to how the SVM has classified the new data.

This procedure, training on a reduced data set then classifying the left-out data, was repeated once for every pair of observations (always one of each class). This is how leave one out classification accuracy was implemented.

Weight mapping analysis. The second objective was to map the support vector weights back into movie space; this would tell us what pixels in the movie were the most critical for distinguishing between the two classes or phonemes. The entire training data matrix (with every

observation of both phonemes) and class vector were used as arguments to MATLAB's *svmtrain* function, unlike for the classification accuracy determination. Other parameters that were specified included (1) a box constraint, or c value, of epsilon to the negative one-half power, as recommended based on previous work, (2) a linear kernel, which was the default option and was deemed to work sufficiently for these purposes, and (3) the option to have MATLAB normalize the data across columns (so, for these purposes, normalizing a given voxel across its time course) by subtracting the mean and dividing by the standard deviation of the column's values; analysis was performed without scaling.

It was decided to perform this analysis without scaling for the following reason. The original, unscaled support vector data are comprised of pixel intensities, which take on values between 0 and 255. These data are used to calculate the weight vector after SVM analysis by multiplying them by the alpha vector (which contains support vector weights). In MATLAB, the alpha vector values are positive for support vectors corresponding to the first phoneme, and negative for support vectors corresponding to the second phoneme. Thus, when the alpha vector is multiplied by the support vectors, the resultant weight vector has positive values for the pixels weighted towards the first phoneme, and negative for the second phoneme. However, when the data are normalized, the support vector data takes on both positive and negative values. Thus, the phonemes are indistinguishable by sign after multiplication by the alpha vector, because a positive weight could be generated by positive alpha values multiplied by positive data, or negative alpha values multiplied by negative data. For this reason, scaling was not used for the weight vector mapping analysis.

The relevant results of the SVM training were alpha values (a column vector of length equal to the number of support vectors) and support vectors (a matrix with the number of rows equal to the number of observations determined to be support vectors, and the number of columns equal to the total number of pixels in a movie clip for that phoneme). The SVM weights were determined by multiplying the transpose of the alpha vector by the support vector matrix, to result in a row vector with length equal to the number of pixels in a movie clip. This action effectively multiplies each support vector by its weight, then sums each feature across all of its observations; this yields information regarding the relative significance of each feature.

Once having obtained the row vector of weights, it was transformed back into movie frames, in exactly the reverse manner as was described above for forming the observations. Briefly, this involved (1) transposing the row into a column, (2) restructuring the column into a matrix with each column representing one image, and (3) restructuring each column into a two-dimensional image. The result of this was similar to a movie clip, but contained weight values instead of image data – a "weight movie."

After transformation back into movie space, the next step was to threshold and color-code the weight information so that colored weight information would overlay the black-and-white movie clips. Overlaying the weight information onto an example movie clip helped to visualize better what pixels, and therefore what regions in the mouth, had the highest and lowest weights.

To prepare this procedure, the weight information was scaled so that the weight values spanned a spectrum from zero to one. Threshold values and corresponding colors were chosen; this information is given in Table 6. Note that, because weights corresponding to the second phoneme were originally negative, after scaling, the highest weights for the second phoneme are represented by the weights closest to zero; for this reason, according to the table, the bottom 3% and 1% of weights are the highest 3% and 1% of weights for the second phoneme.

Next, an example movie clip upon which to overlay the weight information was selected and transformed into four dimensions – two spatial dimensions, one time dimension, and one color dimension, since RGB color values are specified by three numbers. Initially, the three color dimension values for a given pixel were identical, so that the movie clip remained grayscale.

Table 6. Threshold and color choices.

Threshold	Color
Top 0.1%	Red
Top 0.3%	Orange
Bottom 0.3%	Cyan
Bottom 0.1%	Royal Blue

Finally, a custom algorithm was used to iterate through every pixel location in the weight movie and determine what threshold band, if any, it belonged to. If a pixel location was determined to contain a weight value in the top threshold band, for example, the corresponding pixel location in the movie clip was transformed from grayscale to red; the same idea applied for each threshold band. In this way, the example movie clip was transformed from grayscale to combination grayscale and color, with the highest and lowest weight vector locations color-coded. In this way, the weight vectors were mapped back into the original movie space. In context, the weight maps showed which locations in the mouth were most heavily weighted, or most highly considered, for distinguishing between movies belonging to the first or second phoneme.

3.7 Canonical Correlation Analysis

Canonical correlation analysis of a given phoneme from a run was performed using custom code along with MATLAB's built-in *canoncorr* function, which takes X and Y data, and returns A and B matrices of coefficients. In this context, the image data would be formed into X, and audio information would be formed into Y.

To set up the CCA analysis, which required the preparation of both image and audio data, first a movie clip and its corresponding audio clip were arbitrarily selected (for this work, the first movie clip was used). Then, a subset of each frame was selected, just as with the SVM analysis, so as to "zoom in" on the vocal tract region of the images, as an alternative to masking. The frames were then transformed just as with the SVM analysis, by using the *reshape* function to turn each image into a column vector, and thus create a movie clip matrix. Finally, the movie clip matrix was transposed so that the number of rows equaled the number of frames in the movie clip, and the number of columns equaled the number of pixels in the zoomed-in area of each image. (See Figure 13 and Figure 14 for a depiction of this. Note that, unlike with the SVM analysis, the movie clip matrix was not transformed into a column vector observation.)

The corresponding audio clip was transformed using MATLAB's *spectrogram* function. The parameters given to this function included a window size and overlap amount (chosen to be zero). Because CCA would require the X and Y matrices to have one matching dimension, the time dimension was made to match. For X, the time dimension was the number of frames in the movie clip; to make the dimensions match, the spectrogram window size was set in such a way so that the number of windows would equal the number of movie frames. Example spectrograms from each phoneme pair are given in Figure 15, Figure 16, Figure 17, and Figure 18.

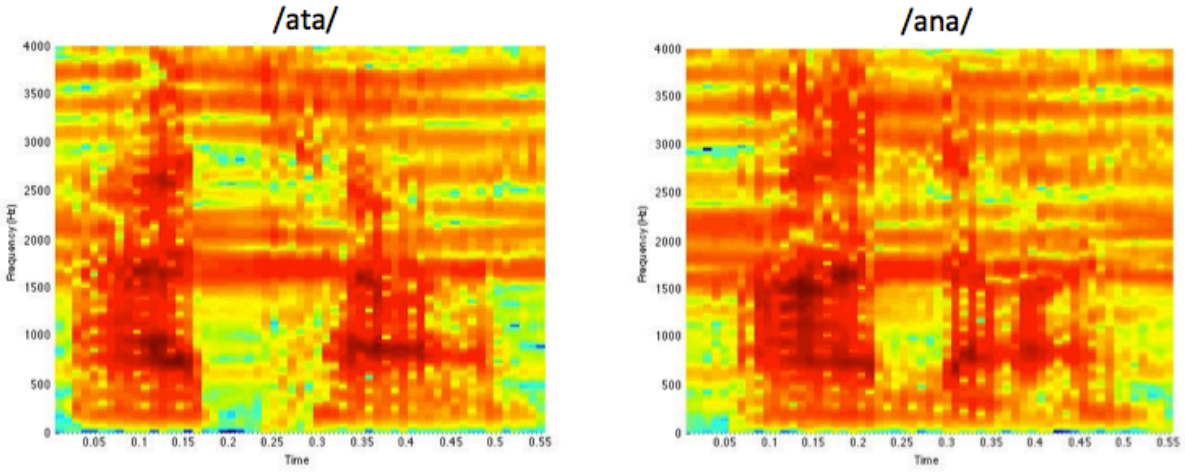

Figure 15. Spectrograms of /ata/ and /ana/.

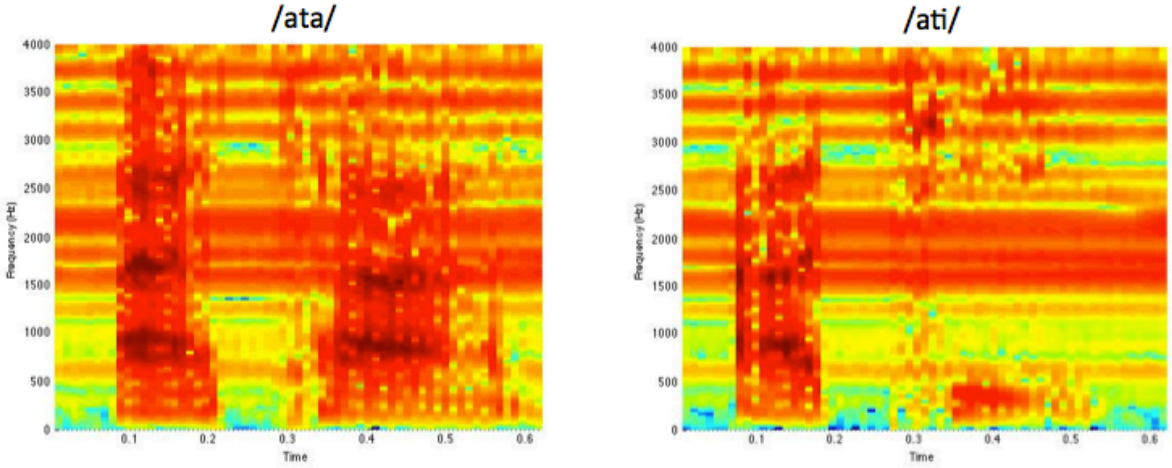

Figure 16. Spectrograms of /ata/ and /ati/.

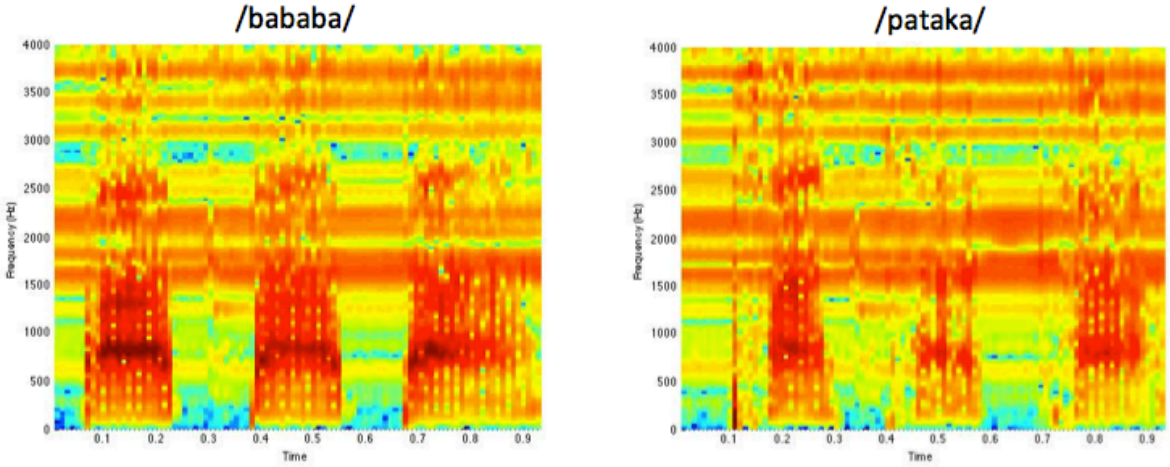

Figure 17. Spectrograms of /bababa/ and /pataka/.

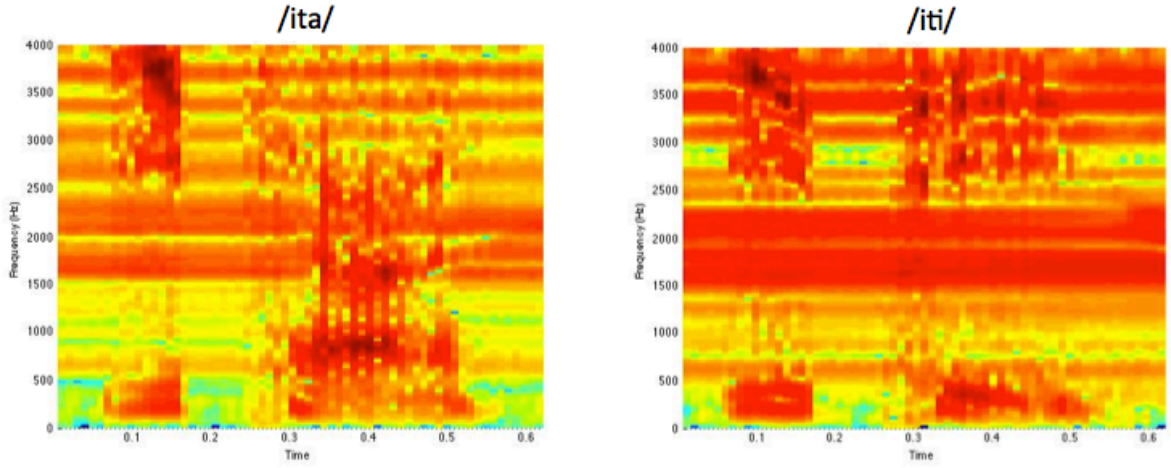

Figure 18. Spectrograms of /ita/ and /iti/.

Once X and Y were set up, they were used as arguments to the *canoncorr* function, which then returned A and B matrices. The goal was then to transform the weight information back into image space and overlay that information onto a sample image. To achieve this, first, the final frame of the movie clip was arbitrarily selected and then turned into a three-dimensional matrix by concatenating that image three times in the third dimension. This set up the image so that the pixels could be turned to color, because RGB values are expressed as three numbers.

Next, the first column of A, containing the first and most important set of weights, was reshaped into the dimensions of a movie frame. This "weight image" was then scaled so as to span the range from 0 to 1. Then, a custom algorithm was used to step through each pixel in the weight image and determine whether its value exceeded a chosen threshold (0.2, here). If so, the corresponding pixel location in the sample image was transformed from grayscale to blue. Finally, the first column of B, representing the first set of weights for the audio clip, was plotted as a bar graph, which is common with CCA.

4. Results and Discussion

4.1 Automated Clip Selection Algorithm

Results. Table 7 summarizes the results of this algorithm, listing the number of clips that the algorithm identified, as well as the number of those clips that were a satisfactory representation of the phoneme of interest (that is, the full waveform of the given phoneme clip was captured, as determined by visual inspection of the clip's waveform and the surrounding samples via the user display in Figure 11). This table also lists the true number of clips of each phoneme in its full audio track, for comparison; additionally, the table gives percent error (see Equation 6) using the number of satisfactory clips as the "measured" value and the true number of clips as the "actual" value. Finally, the table shows the most common erroneous sound that was found in the cases where a phoneme clip was not a satisfactory representation.

Table 7. Automated clip selection algorithm results.

Run Type	Phoneme	# Clips Found by Algorithm	# Satisfactory Clips	Actual # of Clips	% Error	Most Common Error
/ata/ vs /ana/	/ata/	52	27	56	51.79%	"ana"
/ata/ vs /ana/	/ana/	53	17	55	69.09%	"ata"
/ata/ vs /ati/	/ata/	55	55	55	0%	N/A
/ata/ vs /ati/	/ati/	52	47	54	12.96%	"at"
/bababa/ vs /pataka/	/bababa/	43	43	43	0%	N/A
/bababa/ vs /pataka/	/pataka/	43	43	43	0%	N/A
/ita/ vs /iti/	/ita/	51	47	53	11.32%	"it"
/ita/ vs /iti/	/iti/	51	9	53	83.02%	"i"

$$\%\ error = \frac{|measured - actual|}{actual} * 100$$

Equation 6. Percent Error.

Discussion. As seen from Table 7, this automated clip-selection method worked somewhat successfully, depending on the phoneme and the audio track. It was successful in identifying almost all of the clips from a given run, but it was only sometimes successful in *correctly* detecting these clips. It seemed that, often, consistent errors were made (these are given in Table 7 as well), which suggests that the algorithm could be greatly improved if these consistent errors could somehow be avoided.

Although about half of the percent errors are extremely high, optimizing this automated algorithm was not a priority; it was simply meant to be a tool for clip selection assistance. Because the number of datasets were limited, it was not advantageous to perfect this automated clip selection algorithm; it was quicker to simply correct erroneous clip selections manually. However, should this technique be applied to a large number of datasets in the future, it could be advantageous to improve the algorithm for better percent error.

Most likely, the results for the /ata/ vs /ana/ run could have been much better should the inter-peak distance have been optimized for those runs. This is evidenced by the fact that the most common error for /ata/ (from the /ata/ vs /ana/ run) was selection of the phoneme /ana/

instead. This suggests that the inter-peak distance was too large, so the algorithm would look past the /ata/ peak. As for the /ita/ vs /iti/ run, in many cases the /iti/ speech signal was not much stronger than the background scanner noise; because almost every clip consisted of background noise followed by a brief [i] sound (the beginning of the desired /iti/), it is plausible that low signal intensity could have caused the algorithm to correlate the /iti/ template clip with noise. Indeed, for both /ita/ and /iti/, the most common error was starting (and thus ending) the clip too early, so that the clip consisted of noise in the beginning, and the phoneme pronunciation was cut off. Alternatively, the algorithm could have been correlating the last [i] sound from the /iti/ template with the first [i] sound in the subsequent /iti/ instances.

An alternative approach to this would have been to create an algorithm that would detect all the phonemes in a given run; because the speaker alternated between saying two phonemes, every other detected clip then would have been assigned as a belonging to its respective phoneme. One way in which this all-phoneme detection model could have been implemented is that the algorithm would detect heightened signal over the background scanner noise as a means to detect the beginning of a phoneme. Then the clip length could be added to this starting point to determine the clip stopping point.

A potential future obstacle, should this technique be applied to new data, is keeping the clips of phoneme pairs to the same length. This was required for SVM performance later, because if the audio clips differed in length between one phoneme and its pair, then the movie clips would too, which means that the total number of pixels representing an observation would differ, so each phoneme would have a different number of features. Should this technique be applied to other phoneme, word, or even phrase pairs, this could pose a potential problem if the pair members are not similar in length. This could possibly be remedied by downsampling the longer expression. Alternatively, perhaps cutting off the longer phoneme early would not be a problem just for classification accuracy purposes, although no weight information could be obtained for the cut off images.

4.2 Support Vector Machine Analysis

There were two types of results for the SVM analysis. The first results were classification accuracy percentages, which indicated the computer's ability to "mouth read," meaning the ability to distinguish movies of one phoneme from movies of another phoneme. These accuracies were determined by the leave-one-out classification technique.

The other type of results were weight-mapped movies, which gave an indication of which pixels in the movie clips were most important for distinguishing between phonemes. The SVM was trained on all data for this part. The results of this analysis are combination grayscale and color movies, where the colored pixels represent weights, according to Table 6. To demonstrate these results, representative frames from the movies have been selected. Note that there is one set of weight map results per run, and these can be overlaid onto an example movie clip of either phoneme; however, just one phoneme backdrop will be shown for the sake of saving space.

Classification accuracy results. The classification accuracies are shown in Table 8; results for when MATLAB's automatic scaling was both on and off are shown.

Table 8. Classification accuracy results.

Phoneme Pair	Autoscale On Class. Acc.	Autoscale Off Class. Acc.
/bababa/ vs /pataka/ (run 1)	100%	52.86%
/ata/ vs /ati/ (run 2)	100%	57.41%
/ita/ vs /iti/ (run 3)	97.09%	56.86%
/ata/ vs /ana/ (run 4)	97.20%	82.08%
/bababa/ vs /pataka/ (run 5)	100%	51.22%
/ata/ vs /ati/ (run 6)	99.02%	50.00%
/ita/ vs /iti/ (run 7)	98.00%	80.00%
/ata/ vs /ana/ (run 8)	98.08%	50.96%
/ba/ vs /pa/	97.56%	68.29%
/ba/ vs /ta/	97.56%	76.83%
/ba/ vs /ka/	98.78%	92.68%
no speech vs /ba/	98.78%	68.29%
no speech vs /pa/	100%	65.85%
no speech vs /ta/	98.78%	85.37%
no speech vs /ka/	92.68%	76.83%

Selected results from the classification accuracy analysis are highly promising. As Table 8 shows, the classification accuracies were exceptional when the data were scaled, highly exceeding expectations. However, the classification accuracies ranged from good to poor when the data were not scaled. (Note that a classification accuracy of 50% is equivalent to random-chance guessing.) In this case, the original eight phoneme pairs exhibited the worst classification accuracies, with the exception of runs 4 and 7; the additional derived phoneme pairs performed much better, and more in line with the expected results. In particular, /ba/ vs /pa/, /ba/ vs /ta/, and /ba/ vs /ka/ each become more readily distinguishable. This is to be expected, as the [b] and [p] sounds are formed somewhat similarly, especially from just a mid-sagittal view; on the other hand, [b] and [t] are formed differently, and [b] and [k] are formed very differently.

Recall that the MATLAB autoscaling worked to scale the data for each feature by subtracting the mean and dividing by the standard deviation of the feature data. It is not believed that autoscaling should change the classification accuracies, but the researcher has come to understand how autoscaling affected the weight maps, so perhaps there is a reason that it affects the classification accuracies as well.

/Ata/ versus /ana/ weight maps. Representative frames from the weight-mapped /ata/ versus /ana/ run are shown below in Figure 19, overlaid onto example movie frames (58 frames long) from a clip of /ana/. Beginning frames, represented by frame 8 in the figure, are characterized by cyan/blue pixels in the gap between the velum and the back of the mouth, red/orange pixels lining the anterior side of the velum, and a few cyan pixels on the lower lip. After this, and up until about the halfway point through the frames, few colored pixels are found, as represented by frame 29; during these frames, the alveolar nasal [n] is formed. Then, as the tongue lowers to form the [a] sound, orange pixels begin to appear on the tip of the tongue and behind the velum, while cyan pixels appear on the lower lip; frame 46 characterizes these observations. Finally, the number of colored pixels increases from that point until frame 58. This frame shows the large red/orange cluster encompassing the velum, the blue line at the top of the soft palate, orange pixels on the tongue tip, and blue pixels on the lower lip. Additionally, for the first time, cyan and blue pixels are shown lining the bottom of the chin.

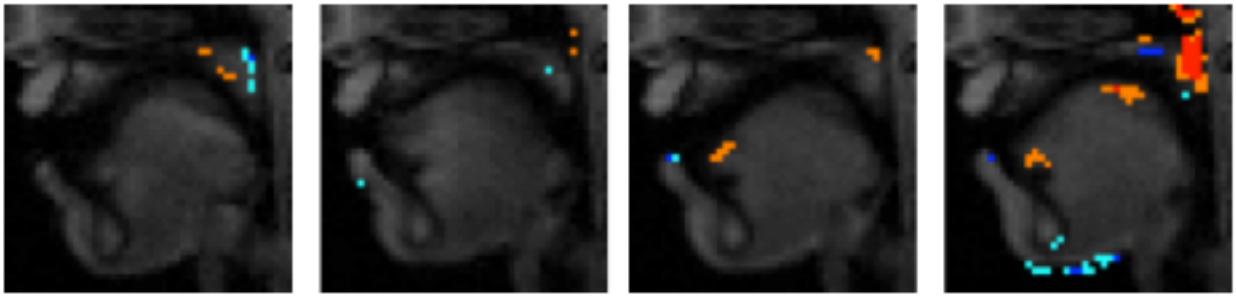

Figure 19. Representative frames from the /ata/ versus /ana/ weight map movie.

The weights target the speech articulators, including the lips, tongue, and velum very well in these results; little other noise is found. The fact that the weights follow the articulators and anatomical features of the vocal tract, is, in itself, an impressive result, indicating that physical positioning of the articulators indeed provides a distinction between phonemes. /Ata/ and /ana/ are formed similarly, as [t] is an alveolar plosive, and [n] is an alveolar nasal; in both cases, the tongue contacts the alveolar ridge. One key difference between their formation that can be viewed in a mid-sagittal slice is the position of the velum. The velum lowers for the production of the [n] sound so as to open the nasal chamber for additional resonance, and closes over the course of the next sound, here, [a]. Interestingly, the velum is highly weighted at the beginning and towards the end of the movie clip, but only slightly during the consonant production. This could be explained by the fact that vowels tend to become nasalized when they occur adjacent to a nasal consonant, so that coupling between the oral and nasal cavities occurs due to a lowered velum; thus, the [a] phones in /ana/ would be nasalized, while the [a] phones in /ata/ would not.

/Ata/ versus /ati/ weight maps. Representative frames from the 64-frame /ata/ versus /ati/ weight map movie are shown below in Figure 20, with weights overlaid onto an /ati/ movie clip. As shown from frame 1, the oral cavity and root of the tongue are strongly weighted, and the velum is weighted as well. These blocks of weights slowly diminish over the next several frames as the mouth moves to form the [t] sound. Frame 28 demonstrates that during [t] formation, the velum is highlighted slightly, but nothing else. Finally, frame 64 shows that, as the [i] sound is formed, clusters of weights form below the hard palate and at the root of the tongue; this pattern is very similar to that in frame 1, except with the colors switched.

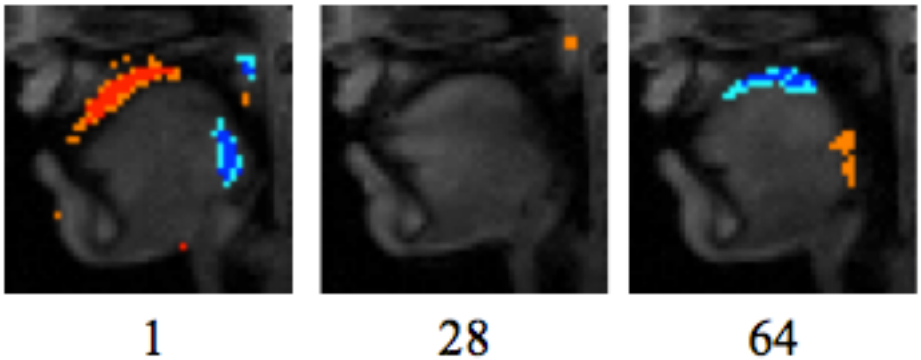

Figure 20. Representative frames from the /ata/ versus /ati/ weight map movie.

Once again, the weighting strongly correlates with anatomical features of the vocal tract, which indicates that physical positioning of the articulators indeed provides the difference between phonemes. The strong weighting in the first frame could be explained by positioning effects from the last spoken phoneme. Since [i] is a high, front vowel, and [a] is a low, back vowel, the tongue could start at a higher position when forming /ata/, because the [i] sound would have just been pronounced. Similarly, the tongue could start at a lower, back position when forming /ati/, because the [a] sound would have just been pronounced.

The strong weighting at the end of the movie makes sense, as the phonemes differ at their ends. The orange pixels at the root of the tongue are those pixels that are highly considered when deciding if the movie is of the phoneme /ata/, whereas the cyan/blue pixels at the top of the tongue indicate the pixels most highly considered when determining if a movie is /ati/. This makes sense, as the end of /ata/, [a], is formed with the tongue low and back in the oral cavity, whereas the end of /ati/, [i], is formed with the tongue high and towards the front of the oral cavity (as seen in frame 64). This further explains the theory of the residual effects seen in frame 1, as weights in the same positions are indicative of the next word. That is, the top of the tongue is indicative of the word /ati/ at its end, so if the tongue has been left in the [i] position for the start of /ata/, it would indicate that /ata/ was coming up; thus, the weights at the top of the tongue would be indicative of /ata/.

/Bababa/ versus /pataka/ weight maps. Figure 21 below shows representative images from the /bababa/ versus /pataka/ weight map movie; these weights are overlaid onto a 96-frame movie clip of /pataka/. Frame 5 is indicative of the first few frames, during [p] formation, which exhibit just a few weights on the nose, incidentally. Frames 26, 42, and 56 represent the [a] following the [p], the [t], and the subsequent [a]; none of these exhibited consistent weights. Frame 68 shows that the glottis was highlighted during [k] formation, and frame 93 shows that weights appear at the end of the movie. In this frame, the upper lip, the top of the tongue, the glottis, and the chin are heavily weighted.

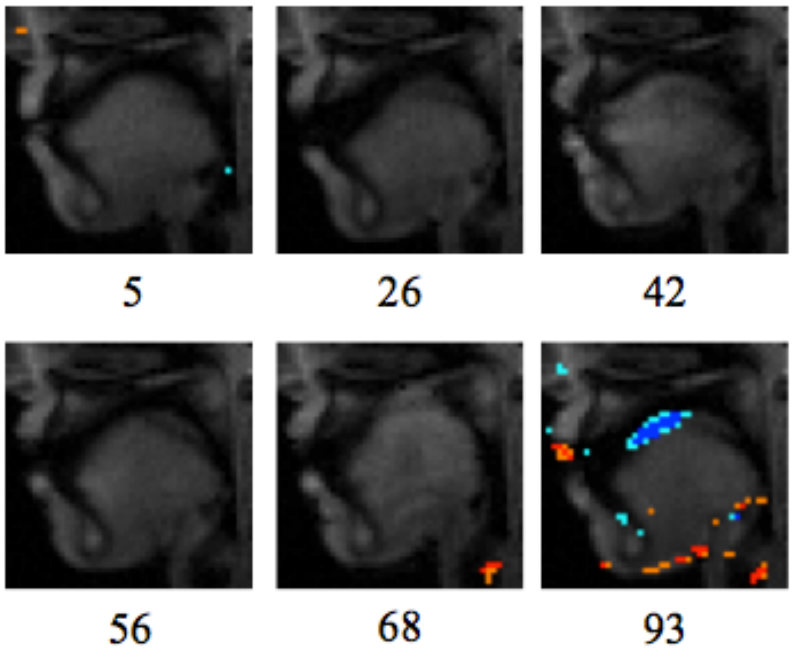

Figure 21. Representative frames from the /bababa/ versus /pataka/ weight map movie.

These results also generally follow lines and anatomical features of the vocal tract. However, these results were surprising, given the clear distinctions between the formations of /bababa/ and /pataka/. It is known that [b] and [p] are formed similarly, both being bilabial plosives. However, [t] is an alveolar plosive, which means that it is formed by the tongue's contact with the alveolar ridge, and [k] is a velar plosive, which means that it is formed by the tongue's contact with the velum. Thus, it was expected that the [b] vs [t] time points would exhibit weights on the lips and along the alveolar ridge, and that the [b] vs [k] time points would exhibit weights on the lips and on the velum.

To explain frame 5, it is possible that the nose is pulled differently due to the slight difference in the formations of [b] and [p], but it is more likely that this is noise, or a reconstruction artifact (reconstruction artifacts were found on the spine as well). One possible reason that weights barely show up through the [k] sound could be that the second [b] and the [t], or the third [b] and the [k], were not aligned in time so as to form a clear contrast.

It is very odd indeed that nearly all of the weights show up at the end of the movie, but these may perhaps be explained. The weights indicative of /bababa/ are found on the upper lip, and the weights indicative of /pataka/ are found on the top of the tongue. Perhaps there is a residual effect from the [b] being formed with the lips, and the [k] being formed with the tongue near the palate.

Ita versus iti weight maps. Representative results from the weight map movie of /ita/ versus /iti/ are shown below in Figure 22, overlaid onto a 64-frame movie clip of /ita/. Frame 1 shows clusters of weights along the lower lip, the top of the tongue and hard palate, the root of the tongue, and the velum during the production of the [i] sound. During the [t] sound, few weights are seen, as exhibited in frame 33. Finally, at the end of the phonemes, clusters of weights are found along the top of the tongue and hard palate, as well as at the root of the tongue; this is seen in frame 64.

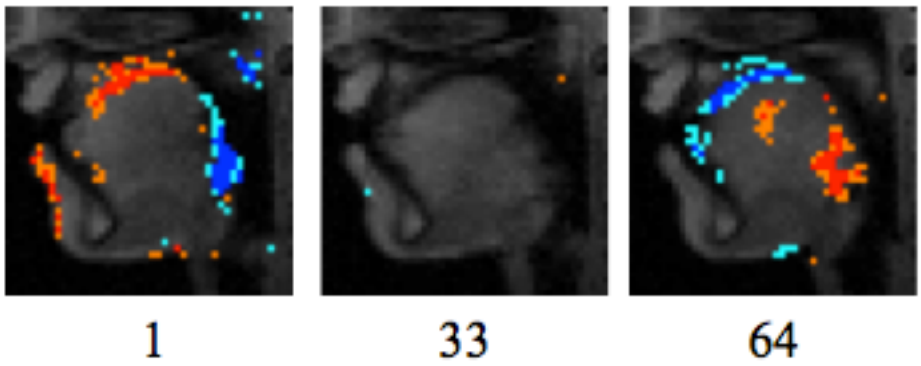

Figure 22. Representative frames from the /ita/ versus /iti/ weight map movie.

Once again, the weights in these results closely follow vocal tract anatomy and articulator position. The weights in frame 64 make the most sense, as the /iti/ weights show up high in the oral cavity, near where the tongue is located to make the [i] sound, and the /ita/ weights show up at the back of the oral cavity, where the tongue moves to form the [a] sound. Frame 1 can most likely be explained as the results from the /ata/ versus /ati/ run were explained. In this frame, the /ita/ weights show up near the hard palate, which would make sense if the tongue was still in the [i] position, from having just formed the phoneme /iti/; alternatively, the /iti/ weights show up at

the base of the tongue, which would make sense if the tongue was still in the [a] position from just having formed the phoneme /ata/.

Additional weight maps. Weight maps were also constructed for the additional phoneme pairs listed in Table 2. For these analyses, the thresholds were increased from 0.001 and 0.003 to 0.005 and 0.01; this allowed for more noise, but also easier identification of weight clusters. /Ba/ versus /pa/, /ba/ versus no speech, no speech versus /ta/, and no speech versus /ka/ were mainly noisy and showed no pattern to the weights. These results could be attributed to inconsistencies in the speech formation, in the movie clips, or in the temporal alignment of the phonemes. However, /ba/ versus /ta/, /ba/ versus /ka/, and no speech versus /pa/ demonstrated some interesting patterns and will be discussed here.

The representative results for /ba/ versus /ta/ are shown below in Figure 23, overlaid onto a 26-frame example /ba/ clip in the first row, and overlaid onto a 26-frame example /ta/ clip in the second row. In this figure, weights concentrate on the upper and lower lips, on the velum, and along the contour of the tongue. The weight maps are the same for both rows of Figure 23; only the underlying movie clip differs. The two underlays lend an interesting comparison. Relative to the position of the weights, it is clear that, in comparison to the bottom row frames, the upper lip is lower for the formation of [b] in the upper left frame, and remains lowered throughout the formation of the [a]. The tongue is clearly in a very different position in frame 6 for [b] versus [t], which would explain the cyan/blue weight cluster that contours the tongue in those frames. Finally, the velum is much lower in frame 24 of the /ta/ movie as compared to the /ba/ movie, which could explain the cyan/blue weight cluster on the velum.

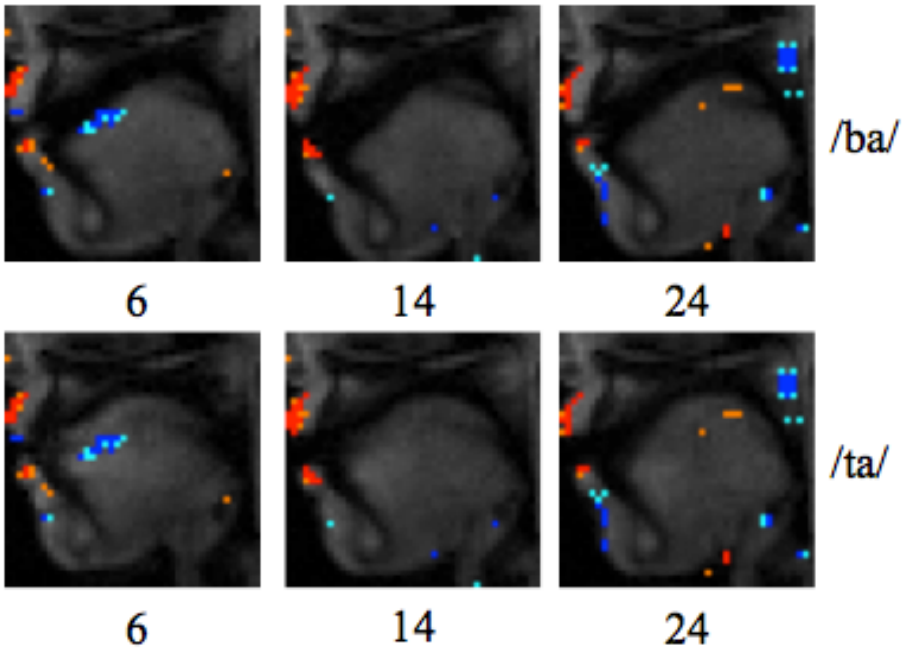

Figure 23. Representative frames from the /ba/ versus /ta/ weight map movie.

The representative results for /ba/ versus /ka/ are shown below in Figure 24, overlaid onto 26-frame example /ba/ and /ka/ movie clips. Frames 7 exhibit heavy red/orange weighting on the lips that is indicative of the /ba/ phoneme, as well as light cyan/blue weighting on the velum. As the movie progresses, the weighting on the velum increases, and cyan/blue weighting surfaces above the tongue blade. The heavy weighting on the lips makes sense, as the [b] sound is formed

with the lips, whereas the [k] sound is not. Two cyan pixels above the tongue in frames 7 are indicative of /ka/, which makes sense because the tongue raises to the palate during formation of the [k] sound.

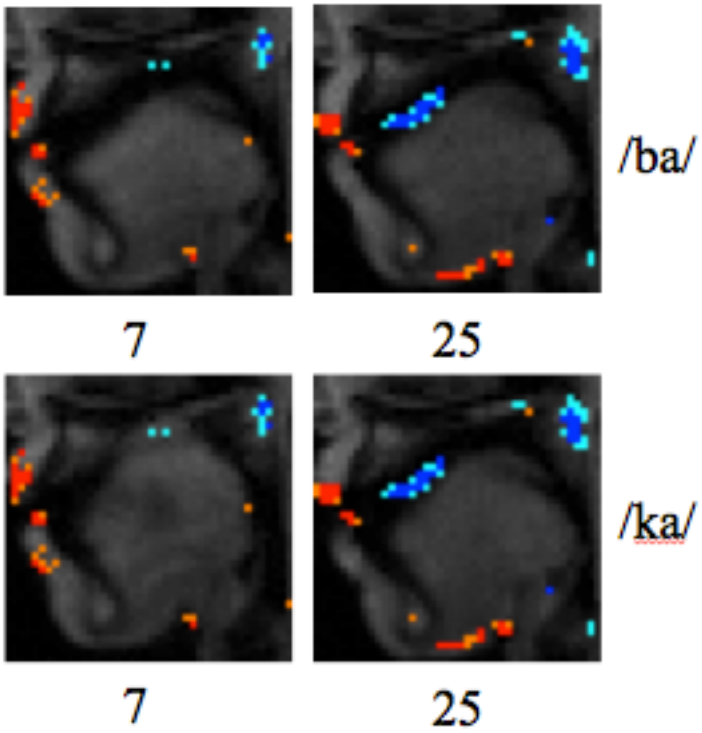

Figure 24. Representative frames from the /ba/ versus /ka/ weight map movie.

Finally, certain representative results for no speech versus /pa/ are shown below in Figure 25, overlaid onto an example movie clip of /pa/. Although these results are not as clean as some of the others, as mentioned before, the fact that the weights follow the articulators and anatomical features of the vocal tract, is, in itself, an impressive and indicative result. In frame 3, as in many of the beginning frames of this 26-frame movie during which the [p] sound is formed, the upper lip contains a concentration of cyan/blue weights; additionally, as with the /bababa/ versus /pataka/ maps, the nose strikes again! Frame 16 represents the strong velar weighting exhibited in the middle frames during [a] production, and frame 26 shows the strong weighting around the upper and lower lips that can be seen in the final frames as the mouth returns to a resting position. Weighting indicative of the phoneme /pa/ around the lips makes sense, as the [p] sound is a bilabial plosive. Weighting in the final frames could indicate a preparatory effect.

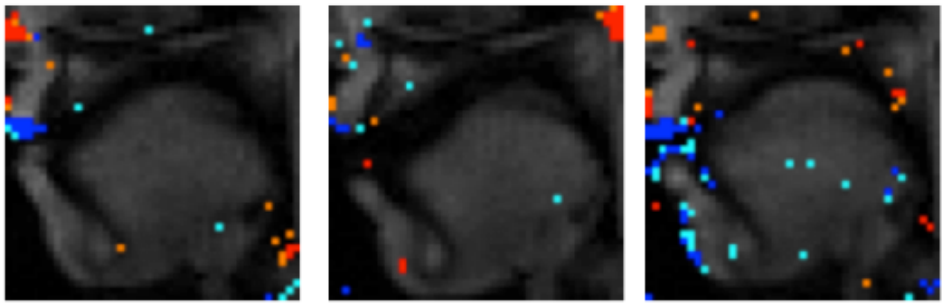

Figure 25. Representative frames from the no speech versus /pa/ weight map movie.

Summarizing Discussion. Generally, the results of the support vector machine analysis met or exceeded expectations. With scaling, the classification accuracy results are outstanding; without scaling, some results are still very good. It is uncertain as to how the scaling affects the classification accuracy, and it is uncertain as to which way is most appropriate for these purposes. Nonetheless, no new information is added to the data in the autoscaling process, which means that something intrinsic about the data enables it to be classified with extremely high accuracies.

The weight map results were excellent as well. The weights overlaid vocal tract features in a surprisingly consistent and noiseless manner; the fact that the weights follow vocal tract features is, in itself, an impressive result that indicates that articulators are used to distinguish between phonemes in the images; as it is well known that articulator positioning affects resultant sounds, these results are consistent with reality.

In addition to the fact that the weights highlight important anatomical features of the vocal tract, many of the results are interpretable in a manner that is consistent with reality. In particular, the weight maps for /ata/ versus /ati/ and /ita/ versus /iti/ were remarkably clean, and make sense according to current knowledge of how the [a] and [i] vowels are formed through the vocal tract. On the other hand, the /ata/ versus /ana/ and /bababa/ versus /pataka/ weight maps, while relatively clean, have less clear interpretations. Very well-defined expectations for the /bababa/ versus /pataka/ results had been developed, based on an understanding of how their constituent sounds are formed; however, none of these results were seen.

4.3 Canonical Correlation Analysis

The CCA results include the first set of weights from the A and B matrices translated back into their original spaces. The CCA results for the original phoneme pairs are given below in Figure 26, Figure 27, Figure 28, and Figure 29. The green pixels in the images represent all of the non-zero A coefficients, and represent the pixels that, when weighted, lend the greatest correlation with the weighted auditory information; similarly, the bar graphs show the values (and thus, relative strengths) of the coefficients corresponding to a given frequency (x axis), that, when weighted, lend the greatest correlation with the weighted image data. The frequencies were binned into 129 frequency bands (based on a default value) that spanned the range from 0 to 4000 Hz. The y-axis values for the B graphs are rather uninformative for these purposes; the important consideration is the relative weights of the frequency bands.

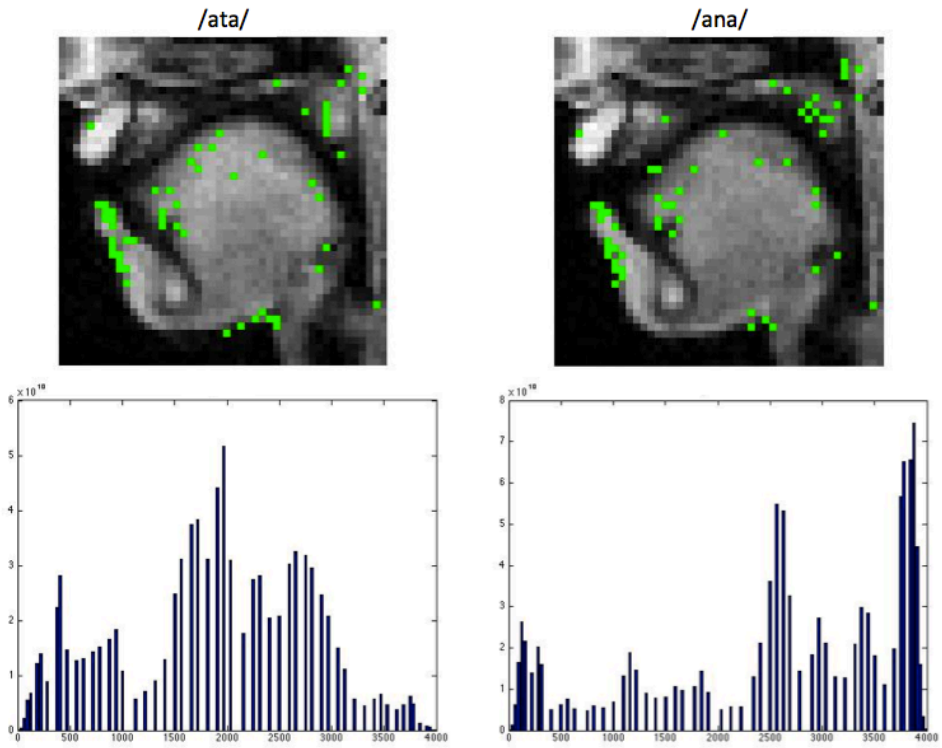

Figure 26. A and B weights for /ata/ and /ana/.

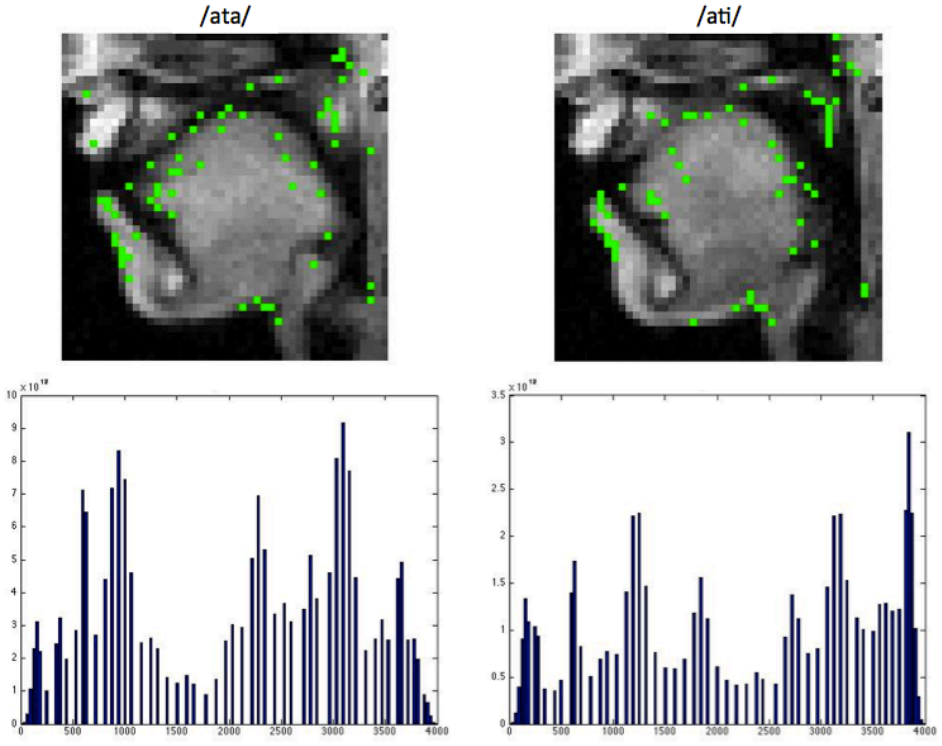

Figure 27. A and B weights for /ata/ and /ati/.

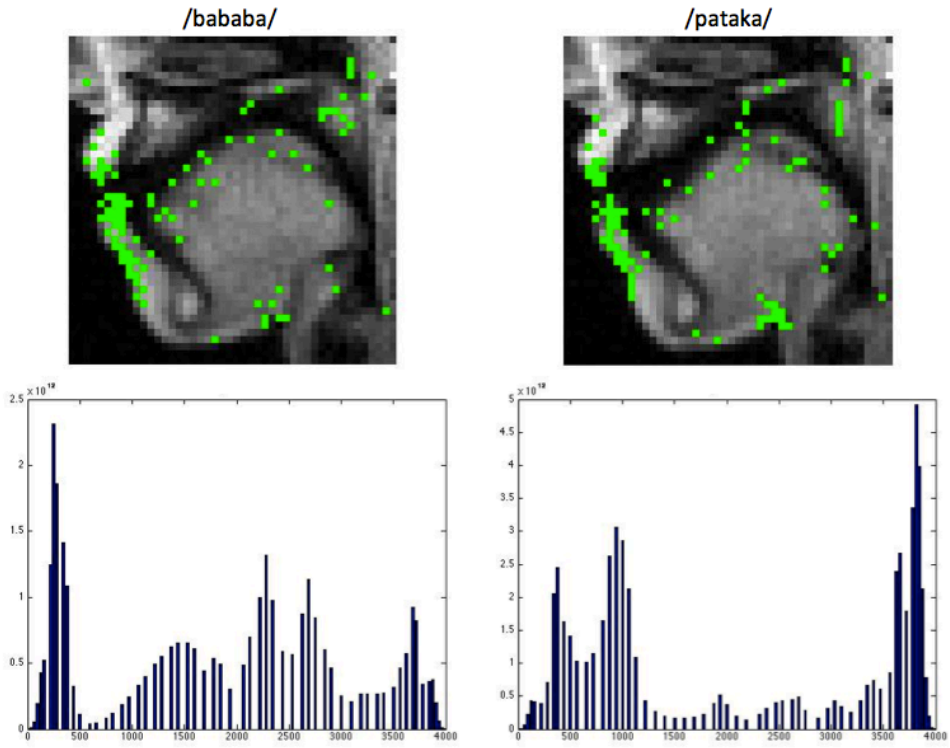

Figure 28. A and B weights for /bababa/ and /pataka/.

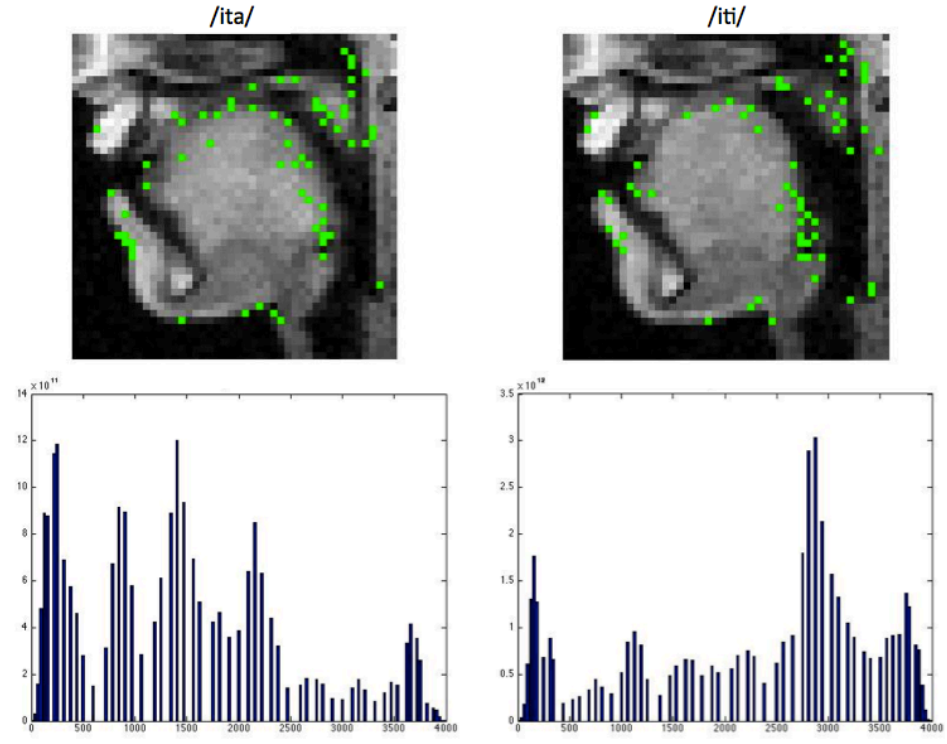

Figure 29. A and B weights for /ita/ and /iti/.

It is notable that the A coefficients correspond to regions of interest in the vocal tract. For example, in Figure 28, many of the weights are concentrated around the lips; in many of the figures, the weights concentrate on the velum or follow the contour of the tongue.

Additionally, to some extent, the B coefficients correspond to the formants of the vowels. For the vowel [a], the first formant (F1) occurs around 720 Hz, the second formant (F2) occurs around 1200 Hz, and the third formant (F3) occurs around 2520 Hz^2. For the vowel [i], F1 occurs around 360 Hz, F2 occurs around 2280 Hz, and F3 occurs around 3000 Hz^2. Certain spikes in coefficient values may be seen around these fundamental frequencies in the preceding figures; however, multiple phones contribute to the B values, as well as background scanner noise, which complicates these results.

These results are novel and complex, and thus require deeper analysis, particularly study of spectrograms and formants, before their full implications can be grasped. It is expected that fundamental frequencies and formants are the most heavily weighted.

5. Conclusion

5.1 Summary of Work

Although ample knowledge about speech production exists, and many modalities have been used to study speech, speech remains poorly quantified. Because a relationship exists between the physical production of speech and the resultant audio waveform, the idea for this project was that this relationship could be quantified using multivariate analysis, specifically canonical correlation analysis. Furthermore, a goal was to see if the computer could discern physiological information from speech movies; thus, support vector machines were used to determine weight information about movie pixels, and this information transformed back into its original spaces.

To achieve this, magnetic resonance images were taken using a novel pulse sequence to achieve an exceptionally high frame rate. Audio was simultaneously recorded. After image reconstruction, a key step before analysis could begin was to extract the audio and video clips of each phoneme from the data. This was partially performed with a custom automated clip selection algorithm that employed cross-correlation of one utterance of a phoneme with the rest of the audio track. This technique was highly successful for over half of the phonemes, but unsuccessful for the rest. This was considered to be acceptable, however, as this method was still an improvement over completely manual clip selection, which had been done in the past. After the audio clips were manually verified and corrected as needed, they were used to select corresponding movie clips.

Once audio and movie clips had been finalized, analysis could begin. Desired analyses included support vector machine analysis and canonical correlation analysis. Support vector machine analysis would provide information as to the SVM's ability to "mouth read," that is, to distinguish between movies of phonemes correctly, through the use of leave-one-out classification accuracy. Support vector machine analysis also would enable determination of, and remapping of, the SVM weights so as to understand just how the SVM was performing the mouth reading; the weights, mapped back into their original space, would result in a "weight movie" which, when overlaid onto an example movie clip, could show what parts of the frames were most important for distinguishing between phonemes. CCA, on the other hand, would demonstrate the relationship between the movie clips and their resultant audio waveforms by indicating which pixels and frequencies, when weighted, would lend the maximum correlation between the data sets.

In order to prepare the data for SVM analysis, it was necessary to transform each movie clip into a spatiotemporal vector containing every pixel from the movie. Each of these vectors was then considered to be an observation. This approach was considered to present advantages over simply using each image as an observation. The data were prepared similarly for CCA, except that the spatiotemporal data remained as a matrix instead of being transformed into a column.

As for the results of classification accuracy analysis, scaling the data resulted in exceptional classification accuracies. Although determining classification accuracy based off of unscaled data resulted in less excellent values, the fact of the matter is that no additional information is required to scale the data, so the data intrinsically are separable with these high classification accuracies. The difference only lies in how the data are presented to the SVM.

The results of the SVM weight mapping were highly encouraging. In most cases, the SVM identified the places of articulation in the vocal tract as those pixels of information that lent

the greatest distinction between phonemes, which is consistent with reality. Furthermore, for several of the phonemes, including /ata/ versus /ati/ and /ita/ versus /iti/, these results were consistent with what is known about how they are physically produced.

Finally, the results of CCA also demonstrated that places of articulation in the vocal tract are responsible for the resultant audio waveform. The frequency information that resulted from this analysis requires further expertise before analysis can take place.

In summary, this work demonstrates that multivariate analysis can be used to quantify speech, and that these analyses can confirm what is already known about speech production. This provides a first step towards computer-assisted speech therapy, as well as the study of brain-behavior systems.

5.2 Future Directions

This work provides the grounds upon which to build additional and more complex analyses, and has great potential for future application.

Improvements. This work has been a novel proof of concept that could be improved in many ways now that the researcher has gone through the efforts to complete the work and thus more fully understands the details of its implementation. First, the automated audio clip selection algorithm may be improved upon, which would provide great assistance in future work as more data are processed. It would be advantageous to study the consistent errors made by the algorithm so as to understand how to correct them. One thing that may make the speech more consistent would be a visual cue to speak while in the scanner, as the auditory cue was unsuccessful.

If this work was to be redone, the /ata/ versus /ati/ and /ita/ versus /iti/ phoneme pairs could be scanned with longer pauses between the phonemes; additionally, the subject could be instructed to return their tongue to a neutral position in between utterances. It would then be interesting to test the theory that the beginning weights come from residual effects of the end of the last phoneme.

Short term. Work that may be done in the immediate future includes further testing and expansion of the data and methodology.

Obviously, more subjects could be recruited for the further collection of data. This would provide a means of comparing how similarly or how differently people form their words; this study would prove interesting, as obviously there are fundamental similarities in how words are formed, despite the fact that speech is highly individualized. This would also give an indication of how much personalization speech treatment could require, and how much multivariate analysis may help with this personalization. Another interesting assessment from this work could include comparing the same phoneme from two different speakers, and determining if the SVM can distinguish between speakers.

Along the same lines of collecting more data, countless possibilities for phoneme pairs exist. Certainly, the list of physiologically interesting pairings is extensive; one such example would include the two nasal consonants, [m] and [n]. Furthermore, it would be of interest to study potential coarticulation effects. However, one consideration must be to ensure that phonemes or words are the same length in a given pair.

Finally, it would be interesting to move from studying fundamental sounds to more complex words or possibly even phrases. It is probable that more complex structures would be more difficult for the SVM to classify.

Long term. After extensive work, the hope is that this project could serve as a basis for the diagnosis and treatment of speech disorders. It could be helpful for speech therapists to view how exactly patients form their words in order to develop more individualized treatment plans. Also, visualization of speech could indicate the effectiveness of therapeutic exercises. Finally, perhaps visualization of speech could be helpful to surgeons before they perform reconstructive surgery, as in cases of adult cleft lip and palate.

This work could be potentially even more useful to surgeons and therapists if it could be implemented in three dimensions; for example, two-dimensional speech imaging is less useful for glossectomy patients, as the reconstruction is often asymmetrical. Some MRI speech studies have expanded beyond the mid-sagittal plane, and once this technique is refined, it would be relatively simple to implement it in three dimensions. Limitations to this advancement include increased complexity of data, as well as decreased spatial or temporal resolution.

Scientifically, this work is the first step to the interesting study of brain-behavior relationships. A technique called Simulscan[41] provides the capability to simultaneously perform functional MRI of the brain, and dynamic structural MRI. While studying brain-behavior relationships is logistically difficult in a scanner that requires a coil, because the vocal tract is so near the brain, it provides a unique system to study in conjunction with brain activity.

Certainly, even more improvements to, and applications of, this idea exist that simply need to be discovered. Hopefully this work ultimately may be used for the betterment of speech medicine.

References

1. MacNeilage, P. F. Speech production. *Language and Speech* **23**, 3–23 (1980).

2. Ball, M. J. & Muller, N. *Phonetics for Communication Disorders*. (2005).

3. Azu The Sound Producing System. (2006).at <http://www.azlifa.com/phonetics-phonology-lecture-2-notes/>

4. Honda, K. Physiological Processes of Speech Production. *Speech Production* 7–26 (2000).

5. Coleman, J. The Vocal Tract and Larynx. at <http://www.phon.ox.ac.uk/jcoleman/phonation.htm>

6. Bae, Y., Kuehn, D. P., Conway, C. a & Sutton, B. P. Real-time magnetic resonance imaging of velopharyngeal activities with simultaneous speech recordings. *The Cleft palate-craniofacial journal : official publication of the American Cleft Palate-Craniofacial Association* **48**, 695–707 (2011).

7. O'Connor, K. Vowels, Vowel Formants, and Vowel Modification. (2011).at <http://www.singwise.com/cgi-bin/main.pl?section=articles&doc=VowelsFormantsAndModifications>

8. <http://www.coli.uni-saarland.de/elaut/Sagittals/Sagittalschnitt_Glossar_2_kl.jpg>.

9. Ventura, S. R., Freitas, D. R. & Tavares, J. M. R. S. Application of MRI and biomedical engineering in speech production study. *Computer methods in biomechanics and biomedical engineering* **12**, 671–81 (2009).

10. Ball, M. J., Stone, M. & Gracco, V. L. A Comparison of Imaging Techniques for the Investigation of Normal and Disordered Speech Production. *Advances in Speech-Language Pathology* **3**, 13–24 (2001).

11. Steele, C. M. & Lieshout, P. H. H. M. Van Use of Electromagnetic Midsagittal Articulography in the Study of Swallowing. *Journal of speech, language, and hearing research* **47**, 342–352 (2004).

12. Quintero, J. Videofluoroscopic and Cine-MRI Examination of Tongue Movement during Partial Glossectomees ' Speech. (2010).

13. Bressmann, T., Heng, C. & Irish, J. C. Applications of 2D and 3D Ultrasound Imaging in Speech-Language Pathology. *Journal of speech-language pathology and audiology* **29**, 158–168 (2005).

14. Narayanan, S., Nayak, K., Lee, S., Sethy, A. & Byrd, D. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America* **115**, 1771 (2004).

15. Horn, H. *et al.* Reliability of electromagnetic articulography recording during speaking sequences. *European journal of orthodontics* **19**, 647–55 (1997).

16. Jesus, M. de S. V. & Reis, C. Phonetic description of alveolar phones using electropalatography. *J Soc Bras Fonoaudiol.* **24**, 255–261 (2012).

17. Dudas, J. R., Deleyiannis, F. W. B., Ford, M. D., Jiang, S. & Losee, J. E. Diagnosis and treatment of velopharyngeal insufficiency: clinical utility of speech evaluation and videofluoroscopy. *Annals of plastic surgery* **56**, 511–7 (2006).

18. Rugiu, M. G. Role of videofluoroscopy in evaluation of neurologic dysphagia. *Acta otorhinolaryngologica Italica* **27**, 306–16 (2007).

19. Pegoraro-Krook, M. I., Dutka-Souza, J. de C. R. & Marino, V. C. de C. Nasoendoscopy of Velopharynx Before and During Diagnostic Therapy. *Journal of Applied Oral Science* **16**, 181–188 (2008).

20. Drissi, C. *et al.* Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children. *European radiology* **21**, 1462–9 (2011).

21. Hiramatsu, H. *et al.* Analysis of high-pitched phonation using three-dimensional computed tomography. *Journal of voice : official journal of the Voice Foundation* **26**, 548–54 (2012).

22. Shinagawa, H. *et al.* Dynamic analysis of articulatory movement using magnetic resonance imaging movies: methods and implications in cleft lip and palate. *The Cleft palate-craniofacial journal : official publication of the American Cleft Palate-Craniofacial Association* **42**, 225–30 (2005).

23. Narayanan, S. & Alwan, A. Imaging applications in speech production research. **2709**, 120–131

24. Kim, Y.-C., Proctor, M. I., Narayanan, S. S. & Nayak, K. S. Improved imaging of lingual articulation using real-time multislice MRI. *Journal of magnetic resonance imaging : JMRI* **35**, 943–8 (2012).

25. Rua Ventura, S. M., Freitas, D. R. S., Ramos, I. M. a P. & Tavares, J. M. R. S. Morphologic differences in the vocal tract resonance cavities of voice professionals: an MRI-based study. *Journal of voice : official journal of the Voice Foundation* **27**, 132–40 (2013).

26. Maturo, S. *et al.* MRI with synchronized audio to evaluate velopharyngeal insufficiency. *The Cleft palate-craniofacial journal : official publication of the American Cleft Palate-Craniofacial Association* **49**, 761–3 (2012).

27. Ventura, S. M. R., Freitas, D. R. S. & Tavares, J. M. R. S. Toward dynamic magnetic resonance imaging of the vocal tract during speech production. *Journal of voice : official journal of the Voice Foundation* **25**, 511–8 (2011).

28. Sutton, B. P., Conway, C. a, Bae, Y., Seethamraju, R. & Kuehn, D. P. Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T. *Journal of magnetic resonance imaging : JMRI* **32**, 1228–37 (2010).

29. Fu, M. *et al.* High-frame-rate Multislice Speech Imaging with Sparse Samping of (k , t) - space. *Proc. Intl. Soc. Mag. Reson. Med.* 12 (2012).

30. Kitamura, T. *et al.* Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. *Acoustical Science and Technology* **26**, 465–468 (2005).

31. Perry, J. L. Variations in velopharyngeal structures between upright and supine positions using upright magnetic resonance imaging. *The Cleft palate-craniofacial journal : official publication of the American Cleft Palate-Craniofacial Association* **48**, 123–33 (2011).

32. Demolin, D., Hassid, S., Metens, T. & Soquet, A. Real-time MRI and articulatory coordination in speech. *Comptes rendus biologies* **325**, 547–56 (2002).

33. Sato-wakabayashi, M., Inoue-arai, M. S., Ph, D. & Ono, T. Combined fMRI and MRI Movie in the Evaluation of Articulation in Subjects With and Without Cleft Lip and Palate. (2007).doi:10.1597/07-070.1

34. Vijay Kumar, K. V, Shankar, V. & Santosham, R. Assessment of swallowing and its disorders-a dynamic MRI study. *European journal of radiology* **82**, 215–9 (2013).

35. Rusz, J. *et al.* Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test. *Movement disorders : official journal of the Movement Disorder Society* **26**, 1951–2 (2011).

36. Wismueller, A. *et al.* Human vocal tract analysis by in vivo 3D MRI during phonation: a complete system for imaging, quantitative modeling, and speech synthesis. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* **11**, 306–12 (2008).

37. Christmann, A. & Steinwart, I. Introduction. *Support Vector Machines* 1–20 (2008).doi:10.4324/9780203413494_chapter_1

38. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377 (2013).

39. Zhao, B., Haldar, J. P. & Liang, Z.-P. PSF model-based reconstruction with sparsity constraint: algorithm and application to real-time cardiac MRI. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* **2010**, 3390–3 (2010).

40. Haldar, J. P. & Liang, Z. Spatiotemporal Imaging with Partially Seperable Functions: A Matrix Recovery Approach. *IEEE ISBI* 716–719 (2010).

41. Paine, T. L., Conway, C. A., Malandraki, G. A. & Sutton, B. P. Simultaneous dynamic and functional MRI scanning (SimulScan) of natural swallows. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* **65**, 1247 (2011).