STUDY OF LOCUS EQUATIONS AS FEATURES FOR SPEECH CLASSIFICATION AND RECOGNITION

by

CLARICE VIRGINIA REID

(Under the Direction of MARGARET E. L. RENWICK)

ABSTRACT

Classification of speech is a difficult problem due to the continuous and variable nature of speech. A Locus Equation is a linear regression model that relates F2 at the start of a CV vowel transition to F2 in the middle of the vowel, where C is held constant and V is varied to cover the vowel space. The resulting equation, which takes the form $y = mx + b$, contains information about the consonant and the consonant transitions. In this thesis, Locus Equations are examined as a potential feature for speech classification and regression problems. The equations are automatically generated, and then used as features in classification of dialects and speech dysarthria. Although the equations did not perform well with the dialect classification, results were promising for classification of speakers with dysarthria.

INDEX WORDS: Locus Equations; Natural Language Processing; Computational Linguistics; Classification; Speech Disorders; Speech Dysarthria; Dialect Classification

STUDY OF LOCUS EQUATIONS AS FEATURES FOR SPEECH CLASSIFICATION AND RECOGNITION

by

CLARICE VIRGINIA REID

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2016

STUDY OF LOCUS EQUATIONS AS FEATURES FOR SPEECH CLASSIFICATION AND RECOGNITION

by

CLARICE VIRGINIA REID

Major Professor:	Margaret E. L. Renwick
Committee:	William A. Hollingsworth
	Walter D. Potter

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

DEDICATION

To my parents, my brother, and my wonderful roommates.

ACKNOWLEDGEMENTS

I would like to thank Dr. Hollingsworth for the incredible amount of time and effort he has put into helping me work my way through this amazing project. I could not have made it through without him. I would also like to thank Dr. Renwick for the advice she gave every time I had a question, and for always being happy and willing to help out. Finally, I would like to thank Dr. Potter for agreeing to serve on my committee, and also for the wonderful introduction to AI he gave me.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Speech recognition and classification poses a significant problem in the area of Natural Language Processing. The continuous nature and significant amount of variation in natural speech makes it hard to build consistent models that can generalize across multiple speakers, accents, or dialects. In response to this, finding new features that can be drawn from speech and used to represent the speech over time is important. One interesting phenomenon of natural speech is the consistent pattern of transitions from consonants into vowels. This pattern can be abstracted using a series of regression lines, called "Locus Equations," that can then be used to obtain information about the place of articulation and coarticulation of the preceding consonant. This thesis examines the relevance and usefulness of locus equations as a feature to be used in computational linguistics and speech modelling. Over the course of the paper, six research questions will be asked and then answered. (1) Can locus equations be automatically generated from conversational speech? Does this affect the integrity of the locus equation? (2) Are the generated locus equations accurate enough for recovery of place of articulation? Can machine learning methods be applied to recover place of articulation across more than one manner class? (3) Are the locus equation coefficients valuable features for classification of dialect? (4) Do locus equations still form for speakers with dysarthria? (5) Are the locus equation coefficients valuable features for recognition of speech disorders? (6) Do locus equations contain enough information to serve as valuable features in speech recognition and classification?

Each of these questions is designed to test the generalizability of and ease of access to locus equations, or to test the amount of information a locus equation can convey about speech. Question one will test whether locus equations can be drawn as features from everyday speech, or if they only work with laboratory speech produced specifically for locus equation analysis. It is expected that locus equations will form from the more conversational speech, although the fit of the regression lines may decrease. The second question is meant to both reconfirm the validity of locus equations generated for the first question, and to test the amount of articulation information contained in a locus equation mapping. The equations should be accurate enough for recovery of place of articulation, and an Artificial Neural Network (ANN) is expected to classify the points by place of articulation with the most accuracy. If classification in this step fails, there is evidence that the generated locus equations are not descriptive of place of articulation the way locus equations generated in the past have been. The third question is the first to examine locus equations as a feature indicative of speech characteristics. In the case of dialects, locus equations will only work as a classification feature if the different dialects have different patterns of coarticulation. The initial hypothesis for this question is that the locus equations will be a weak feature for classification—meaning they do provide the system with enough information to classify dialects with moderate results, but more accuracy would require additional information about the speech.

At this point, the focus of the study will shift towards disordered speech, specifically speakers with dysarthria. Dysarthria is a motor speech disorder characterized by difficulty controlling the muscles in the mouth. The fourth question asks whether locus equations can be generated for speakers with dysarthria. The expectation is that locus equations will be formed, but the fit of the regression lines will be significantly worse than it was for the dialect speakers without dysarthria. Question number five tests the usefulness of locus equations as a feature for classification of a speakers with dysarthria. Once again, locus equations will only provide useful information if speakers with dysarthria have systematically different coarticulation patterns than speakers without dysarthria. Since dysarthria is a motor speech disorder and movement of the tongue is affected, the coarticulation patterns should be unique. The equations are expected to perform well as features in this area. Finally, question six addresses the main point of investigation in this research. Are locus equations useful features for speech classification and prediction? Given the ability of locus equations to capture the patterns of vowel transitions from consonants in speech, they are expected to perform well as features. Accurate results from classification of either dialects or speech dysarthria would support this prediction.

CHAPTER 2

LITERATURE REVIEW

2.1 A FEW BASICS OF LINGUISTICS

There are a few basic areas of Linguistics which are particularly relevant to this thesis. These are place of articulation, manner of articulation, and formant values. A brief overview of each topic is included here, to aid in transparency of the following research. Articulation is a study of how articulators make sounds. There are a number of articulators present in the vocal tract that can be used to form a sound. These include the lips, teeth, tongue (tip, blade, and body), alveolar ridge, hard palate, soft palate, and glottis. This is not an extensive list, but it includes the articulators involved in the consonant sounds discussed in this thesis. Figure 1 below shows a mid-sagittal section of the vocal tract with each of these articulators labeled. Consonantal “place of articulation” refers to where the constriction in the vocal tract is happening to form a sound. It is defined in terms of two articulators. Relevant here are bilabial consonants, labio-dental consonants, alveolar consonants, and velar consonants. Bilabial and labio-dental consonants both include the lips as an articulator. They can collectively be referred to as labial consonants. Bilabial consonants (p, b, m) use both lips, and labio-dental consonants (f, v) use the lips and the teeth. Alveolar consonants are formed using the blade of the tongue and the alveolar ridge, which is the small ridge of bone located on top of the oral cavity directly behind the teeth. This grouping includes the consonants t, d, n, and s. Finally a velar place of articulation means the sound is produced using the body of the tongue (the dorsum) and the velum, also called the soft palate. These consonants are g, k, and ŋ (like the sound at the end of “sing”).

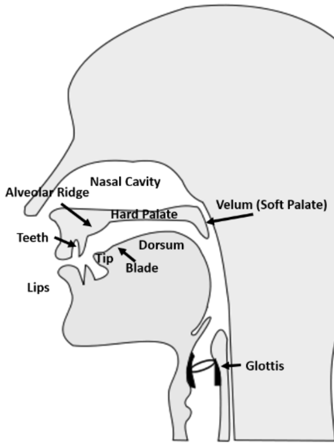


Figure 1: Places of Articulation

Manner of articulation refers to how a sound is made. The basic manners of articulation include stops, nasals, fricatives, affricates, flaps, trills, and approximants. In this research we examine plosives, fricatives, and nasals. A stop consonant is a sound where the flow of air through the vocal tract is completely stopped and then suddenly released. The consonant can either be voiced, where the vocal folds are vibrating, or unvoiced. Examples of stop consonants include p, t, and k. For a fricative consonant, the articulators constrict the airflow without completely stopping it, leading to a characteristic buzzing sound caused by turbulent air. Examples of fricatives would be f, s, or h. Finally, nasals are sounds created when the lips are closed, stopping any air from escaping the oral cavity, but the velum is lowered, opening a passageway to the nasal cavity where the air can resonate. Nasal sounds include m, n, and ŋ. Bringing place and manner of articulation together allows us to decrease the number of sounds being considered. For example, a bilabial stop refers to any sound that is created using both lips and that completely stops the flow of air through the vocal tract. /p/ and /b/ both fall into this category. The difference between them is voicing (the vocal tract is actively vibrating in the creation of /b/, but is still when producing /p/). /n/ can be described as an alveolar nasal. The place and manner of articulation for consonants can be found in the IPA chart (International Phonetics Association) included in Appendix A.

The third important linguistic concept for this research is the presence of formants and the use of spectrograms in inspecting speech. Speech sounds are combinations of vibrations at different frequencies, transferred through the air from the speaker to the listener. When air is expelled from the lungs it passes through the vocal tract, which serves as a filter. By the time the sound wave exits through the mouth, waves at certain frequencies have been cancelled out, and waves at other frequencies have been amplified. The amplified frequencies are what give the sound its pitch and quality, and they are characterized by their frequency in Hz and their bandwidth. These amplified bands of harmonics are called “formants.” The fundamental frequency, also called F0 (the zeroth formant), is the pitch of the sound. This formant will always be the lowest. The next highest formant is F1, and the next after that is F2, and so on. While F0 gives the pitch of a sound, the shape of the other formants gives a sound its quality. For example, /i/ and /a/ pronounced at the same levels of loudness and pitch, still sound different. This is due to the difference in the formants. F0 – F4 are most commonly examined in linguistic research. The locus equations examined here focus on F2 exclusively. Formant values can be seen using a spectrogram, a three dimensional graph of speech sounds with frequency in Hz along the *y*-axis, time along the *x*-axis, and intensity (measured in decibels) on the *z*-axis. Louder intensities are represented by darker lines. Figure 2 shows a sample spectrogram taken from the data set. The text below is a transcription of the sound, and the boundaries show where transitions from one sound to another take place.

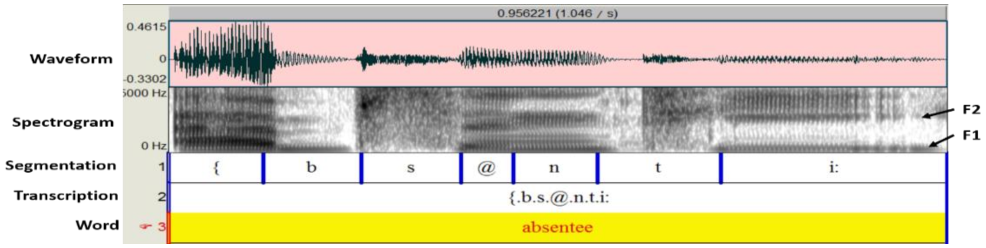


Figure 2: Formant Values

F1 and F2 are tied to vowel height and vowel backness, respectively. Vowels are partially characterized by the height and backness of the tongue. The vowel /i/ is produced by moving the tongue up towards the roof of the mouth and forwards toward the teeth. As such, it is defined as a high, front vowel. /a/ is produced by putting the tongue low in the mouth and back towards the throat, so it is low and back. The F1 and F2 values for the vowels reflect these characteristics. F1 is inversely tied to vowel height—the higher the position of the tongue, the lower the F1 values. In Figure 3, the formant values for American vowels can be seen. /i/ is the first vowel shown. F1 is very low, reflecting the height of the sound. The F1 value for /a/, also seen in Figure 3, has a higher F1 because it is produced with the tongue lower in the mouth. F2 is tied to the backness of a vowel in that front vowels have higher F2 values than back vowels. Figure 3 shows that /i/ has a much higher F2 than /a/ due to the difference in vowel backness.

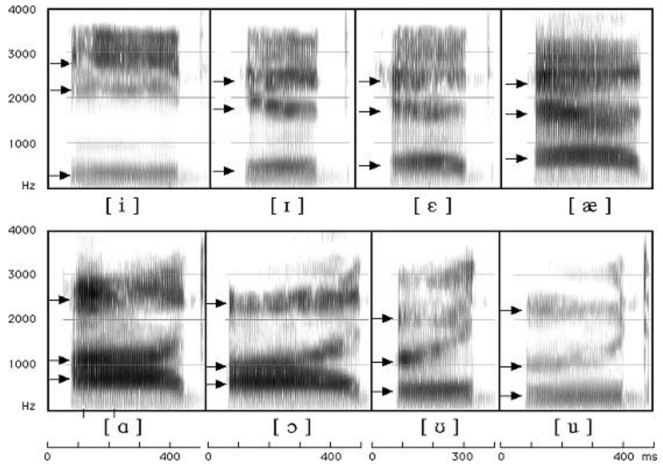


Figure 3: American Vowels
(Ladefoged, 2006)

2.2 SPEECH ALIGNMENT

As access to speech recordings increases, research done in phonetics is expanding to include ever larger amounts of data. Most modern phonetic analysis relies on aligning sound files with phonetic transcriptions, marking the places in the utterance where each phoneme begins and ends. While necessary, aligning speech data is an incredibly work intensive task, and properly aligning even a few minutes of speech can take hours. This has led to an increase in the need for automatic aligners—technology capable of taking in a sound file and outputting a file with the phonemic boundaries. SPPAS is one such aligner (Bigi, 2011). Implemented in Python, SPPAS ties together a variety of resources, like pronunciation dictionaries and acoustic models, to provide a tool for automatic segmentation. The program takes a sound file and a text transcription of the speech contained within that file as input. SPPAS then performs three steps: segmentation, phonetization, and alignment. The first step, segmentation, separates the sound file along inter-pausal units (IPU). This refers to macro units of sound, continuous productions of speech separated by quiet pauses. In the text transcription, these pauses in the speech must be marked with a "#". SPPAS searches through the sound file and automatically adjusts the sound threshold (in dB) in an attempt to match the pauses specified in the text transcription with the sound file.

The next task performed by SPPAS is phonetization, the transformation of English text into an appropriate phonetic transcription. First, the program does “tokenization,” which is not tokenization of words but a scan over the phrases in the text file. The phonetization itself is essentially a dictionary look-up. SPPAS takes the name of a phonetic dictionary as an input parameter—for English, the default dictionary is CMU’s pronunciation dictionary (CMU Pronouncing Dictionary). SPPAS searches through the dictionary for the word in question, and then pulls the given transcription. One strength of SPPAS is the ability to pull more than one transcription—if a dictionary has more than one entry for how a word can be pronounced, SPPAS will pull both options and place them into the transcription separated by a bar (|). Later analysis will decide on the final transcription. If a word is not found in the dictionary, SPPAS will scan the word from left to right, and then try and locate the longest match to that word in the dictionary to use as a substitution.

The third, most crucial step is alignment of the sound file with the generated phonemic transcription. This requires a both speech recognition engine (SRE) and an acoustic model. The SRE used by SPPAS is an external program called “Julius” (Nagoya Institute of Technology, 2010). Originally developed for Japanese, Julius is a “high-performance, two-pass, large vocabulary continuous speech recognition decoder software”. The program runs using 3-gram windows and context-dependent Hidden Markov Models (Rabiner, 1989). To run, it requires both a language model and an acoustic model. For English, the models were trained using the HTK toolkit taken from Voxforge (Voxforge, 2006-2011; Young et al., 1999). The HMM embedded in Julius uses the language and acoustic models as calculated probabilities to produce

TextGrid files in Praat (Boersma and Weenink, 2016) with a best calculated alignment for each of the sounds files (Bigi, 2011).

2.3 SOURCE FILTER THEORY

Source Filter theory is one of the basic theories of speech production. This theory explains speech as the result of a wave which begins at a source (the glottis) and then transforms as it passes through a filter (the vocal tract) (Stevens, 1998). The filter idea is also related to theory of the vocal tract as tubes, which simplifies the human vocal tract into tubes of various lengths and widths with constrictions at certain points. Sounds begin when air is pushed from the lungs up through the vocal folds. The folds rapidly vibrate open and closed, stopping the air and then releasing it in bursts and creating a compression wave. The fundamental frequency of the wave (measured in Hz) depends on the rate at which the vocal folds vibrate. As the wave traverses the vocal tract, the shape of the tract (simplified as a tube) creates standing waves at certain frequencies, leading to the creation of formant values above the more salient fundamental frequency. These formant values are what give a sound its quality and allows it to be identified as contrastively different from other sounds. The frequencies at which the standing waves resonate and the formant values are found depends on the shape of the vocal tract at that time. As the shape of the vocal tract shifts and the area of constriction changes, the formant frequencies will transition as well. When the shape of a vocal tract begins changing to accommodate an upcoming sound before the current sound is finished (ex: shifting the tongue to create [i] when the lips are still closed for [b]) also affects the formant values. This overlap in sound in called coarticulation.

Coarticulation has been defined as “the influence of one speech segment upon another…the influence of a phonetic context upon a given segment” (Daniloff and Hammarberg, 1976). Language is phonetically transcribed as a sequence of discrete segments, called “phonemes.” Although useful for the study of speech patterns, this abstract transcription of language has no proven basis in the acoustic reality of speech. When speech is produced, the individual sounds are not created in discrete segments. Instead, various sounds may overlap with one another—as the mouth is closed to form a stop, like “p” or “b”, the tongue may already be moving into position for a following vowel. The acoustics of the stop release will be affected by the changing position of the tongue. This effect is the basis of “the invariance dilemma,” which asks how speakers recognize the discrete phonemes of speech well enough to comprehend and transcribe them, even though acoustically the sound is never the same (Sussman, 1991).

Coarticulation is a type of phonological feature spreading, as the features of the surrounding phonemes encroach upon the observed phoneme due to the overlapping articulations. When coarticulation occurs, the “locus” of the influence is the segment observed, and influence itself is the articulatory feature (Daniloff and Hammarberg, 241). This influence is bidirectional, meaning that a feature can be influenced by segments preceding it in speech, or by the segments following. The locus equations discussed in this paper are a result of right-to-left coarticulation. The focus of a locus equation is on a vowel following a consonant in speech, and the effect that preceding consonant had upon the second formant values of the vowel.

2.4 THE ORIGINAL LOCUS

One of the earliest conceptualizations of a locus can be found in “Visible Speech” by Ralph Potter (Potter et al., 1947). In his observations regarding the new methods of visualizing speech waves, Potter notes that vowels pronounced as individual segments with no context have straight “second bars,” (second formants). When a consonant precedes a vowel, however, a curved transition from the consonant into the vowel can be seen. Potter wrote that the position of the second formant in isolation, which called a “hub,” shifted down following labial stops, and up or down for velar stops (pg. 38). This may seem to be a fairly mundane observation, but the effect of this trend on speech perception and synthesis was further explored a few years later (Liberman, 1954). This is when the idea of a “locus” point was introduced under that name. Liberman notes that the transitions from the second formant into the vowel steady state seem to provide cues for perception of stop consonants, and hypothesizes that consonants have a “locus point,” a second formant frequency which is set to a certain, unique value for each place of articulation. If this locus point were to exist, the transitions seen on the spectrograms represent the movement from the unseen locus point to the vowel steady state. In terms of tube theory, “…the transitions seen in spectrograms reflect the changes in cavity size and shape caused by the movements of the articulators” (Delattre et al., 1955). Figure 4 shows the plots created by Delattre et al. to demonstrate the transitions (Delattre et al., 1955).

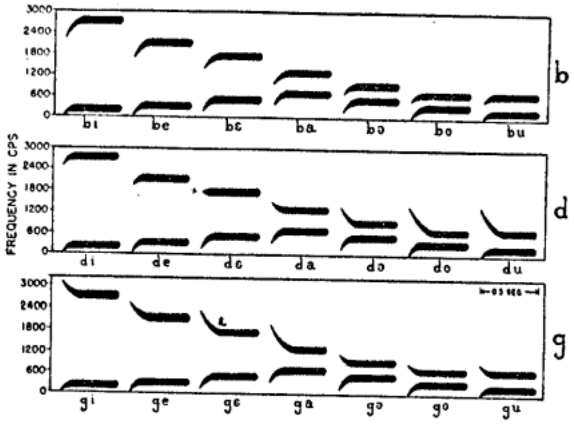


Figure 4: Synthetic spectrograms showing vowel transitions (Delattre et al., 1955, Figure 1)

After experimentation, it was hypothesized that [d] had a fixed locus at 1800 Hz, [b] at 720 Hz, and that [g] had a potential locus at 3000 Hz, but only when the vowel steady state was above 1200 Hz (Delattre et al., 1955). Despite these findings, [d] was the only consonant with a convincing locus point. [b] was agreed to originate from some low point, but the exact point could not be found. There was no comprehensive pattern found at all for [g], and Delattre notes that while a potential locus pattern was locatable for the front and mid vowels, the pattern did not apply when back vowels were added. He says “it is obvious that the same [g] locus cannot serve for all vowels” (Delattre et al., 1955). One more important observation from these experiments is that the second-formant transitions characteristic of /b, d, g/ also cued /p, t, k/ and /m, n, ŋ/ respectively, meaning the second-formant transitions provide information about place of articulation for more than just voiced stops. Although this early concept of a locus point has largely been disregarded, a few points remain relevant in more modern locus equations. These are the stability of [d] transitions, the unique break in pattern between +/- back vowels for [g], which is seen in later experiments, and the relation of locus equations to place of articulation as a phonetic feature.

In 1963 Stevens and House conducted an experiment examining the changes in vowel formants in a given dialect in a variety of consonant contexts (Stevens and House, 1963). The study supported the belief that F2 is the most sensitive to consonant contexts, and is therefore the best source of information regarding preceding consonants and coarticulation. Another study by Stevens and Blumstein in 1979 resumed the search for a “locus point” (Stevens and Blumstein, 1979). The study tested the use of the onset F2 values of vowels following stop and nasal consonants in CV tokens for classification of place of articulation. The tokens were classified onto three templates—diffuse-rising, diffuse-falling, and compact. These represented alveolar, labial, and velar places of articulation. Classification by template was 85% accurate for the voiced stop consonants, and 76% accurate for nasals (Stevens and House, 1979).

2.5 EARLY LOCUS EQUATIONS

"Locus Equations" as they are used in this research were originally introduced by Lindblom in his 1963 thesis. In it, he derived the following equation:

$$F2_{onset} = k * F2_{vowel} + c$$

Here, $F2_{onset}$ is the frequency of the second format at a vowel's first glottal pulse following a voiced stop consonant, and $F2_{vowel}$ is the frequency of the second formant at that same vowel's midpoint, where the frequency had reached a steady state (Lindblom, 1963). *k* and *c* are both constants, representing slope and *y*-intercept respectively. Lindblom performed tests using a single speaker, a Swedish male. The speaker produced CVC tokens for /b, d, g/, with the same consonant before and after a range of eight vowels (Lindblom, 1963). For each stop, Lindblom graphed the F2 values on a Cartesian plane as a series of ($F2_{vowel}$, $F2_{onset}$) points for each vowel. A regression line was fitted to these points to obtain the constants *k* and *c*. The slope values were 0.69, 0.28, and 0.95, and the *y*-intercept values were 410, 1225, and 360 Hz for /b, d, g/ respectively. Lindblom notes that the scatterplot for /g/ seemed to show a flatter slope for front vowels and a steeper slope for back vowels. Because of this, the linear regression was less well-fitted to the points.

Years later, the linear relationship between F2 at its onset and F2 at the vowel midpoint was studied by Nearey and Shammass in 1987. They had ten speakers of Canadian English, five male and five female, produce two repetitions each of CVd tokens, starting with consonants /b, d, g/ and covering eleven vowels picked to cover the vowel space. They pulled $F2_{onset}$ "as early as possible after stop release" and $F2_{vowel}$ at 60 milliseconds after the stop release (Nearey and Shammass, 1987). The tokens collected for each initial stop were graphed on scatterplots in the same manner as above, with $F2_{vowel}$ along the *x*-axis and $F2_{onset}$ on the *y*-axis. In their analysis of the data, they remarked that "all three plots indicate a strong positive correlation." They also noticed that /b/ and /g/ functions seemed more dependent on the vowels than /d/. The regression constant values from these experiments were 0.82, 0.49, and 0.99 for /b/, /d/, and /g/ slope values, and 192, 1041, and 214 for *y*-intercept. Following the creation of the regression lines, an algorithm was used to classify a spoken syllable as /b/, /d/, or /g/ based on its vertical distance from each regression line. This method classified the tokens with 73.9% accuracy.

Krull briefly investigated locus equations using Swedish vowels in her 1988 study of predictors for stop consonants (Krull, 1988). The pattern of bilabials (/b/) forming a much steeper slope and having a much lower *y*-intercept value than alveolars (/d/) held true in this experiment. Velars (/g/) were only investigated as palatal velars preceding front vowels. This resulted in a velar slope very similar to the alveolar slope, and a *y*-intercept higher than either of the other two categories. Krull made the important observation that the magnitude of the regression slope is directly related to the degree of coarticulation seen between the initial stop and the following vowel. Shallow slopes, like those seen for alveolars, imply a relatively fixed stop "locus" frequency that is largely unaffected by the vowel's steady midpoint frequency (Krull, 1998). This can be credited to the lack of coarticulation between alveolars and most vowels—the position of the tongue when producing alveolars does not shift much towards the other vowels being produced, perhaps because a small shift in the alveolar region could easily lead to the consonant being mistaken for something else. This relatively fixed $F2_{onset}$ translates to a shallow slope.

In 1991, Sussman et al. published an in-depth analysis of locus equations as a possible "source of relational invariance for stop place categorization" (Sussman 1991). They theorized that the systematic shift in F2 could serve as a cue that was invariant in its relation to place of articulation. This cue could then be used to help solve the problem of relational invariance. The study took twenty adult speakers, ten male and ten female, from multiple dialects of American English. The subjects produced CV/t/ syllables, with the three customary initial stops /b, d, g/ and ten medial vowels /i, I, e^y, ɛ, æ, a, o^w, ʌ, ɔ, u/. The $F2_{onset}$ value was taken from the first glottal pulse after release, and the $F2_{vowel}$ value was taken at the mid-vowel nucleus (Sussman et al, 1991). Regression lines were calculated for scatterplots created from each speaker. It was decided that a single line would be graphed for the velar scatterplot for two reasons. Firstly, the when the palatal /g/ (front vowels) and velar /g/ (back vowels) were given two separate regression lines, it was found that the slope for palatal /g/ overlapped with /d/, and the slope for velar /g/ overlapped with /b/. When /g/ was plotted as one scatterplot the composite line fell discernibly between the /b/ and /d/ lines. Secondly, the R^2 value was no higher for the separate regression lines than it was for the composite line. When averaged across all speakers for each place of articulation, /b/ had a slope of 0.89 with a *y*-intercept of 99 Hz, /d/ has a slope of 0.42 with a *y*-intercept of 1211 Hz, and /g/ had a slope of 0.71 with a *y*-intercept of 792 Hz. Slope analysis revealed that there was a significant difference in slope ($p < 0.05$) for each pair of consonants. *y*-intercept analysis showed that /b/ consistently had the lowest intercept, followed by /g/ and then /d/ (Sussman et al., 1991). Perhaps the most interesting aspect of this experiment was the attempt to classify place of articulation based on the F2 values. When classification was attempted using just the $F2_{onset}$, $F2_{vowel}$ pairs, classification rates for labial, alveolar, and velar place were 84%, 81%, and 69% accurate for males, and 82%, 78% , and 67% accurate for females. The next classification was performed on the higher-order slope/*y*-intercept constants pulled from the regression lines. This method led to 100% accuracy in classification across all three stop place categories (Sussman et al., 1991).

A year later Sussman et al. published a new paper examining the differences in locus equations cross linguistically (Sussman et al., 1992). The study addressed two issues of particular interest to this research. The first being "Do locus equations emerge successfully cross-linguistically?" and the second being the theory of phonetic "hot spots." The question is essentially this: do languages utilize the full phonetic space available to them? Or are certain regions typically preferred? The study sampled new data from native speakers of Thai, Cairene Arabic, and Urdu. Speakers of Thai were only sampled for bilabial and dental stops "because only bilabial and dental place contrasts exist." The tokens were of the form CV, with C being /b/ and /d/ and V including nine Thai vowels. The Cairene Arabic speakers were asked to pronounce stops /b/, /d/, pharyngealized /dʕ/ and /g/. The utterances were CV/t/ or CV/tt/ tokens, with each stop followed by one of eight medial vowels. Finally, the Urdu speakers used four stops: /b/, dental [d], retroflex [d], and /g/. The stops were followed by one of nine medial vowels, which varied with the language but always included the three vowels /i, a, u/. The tokens were CVC, with the final C varying to maximize the number of real tokens (Sussman et. al, 1992).

For the Thai speakers, the mean labial slope was 0.70 and the mean *y*-intercept was 228 Hz. The mean alveolar slope was 0.295 and the *y*-intercept was 1425 Hz. When these values were graphed in a scatterplot with slope along the *x* axis and *y*-intercept along the *y* axis, the two places of articulation showed clearly distinct clusters. For Arabic speakers, the stop places were labial, dental, dental pharyngeal, and velar. These demonstrated mean slopes of 0.77, 0.25, 0.21, and 0.92 and mean *y*-intercepts of 206 Hz, 1307 Hz, 933 Hz, and 220 Hz respectively. Graphing these values as points on a coordinate plane once again showed clear distinctions between each place of articulation. The Urdu stop place categories were labial, dental, retroflex, and velar. These had mean slopes of 0.81, 0.50, 0.44, and 0.97, and mean *y*-intercepts of 172 Hz, 857 Hz, 1070 Hz, and 212 Hz respectively. The categories were mostly distinct when graphed on a coordinate plane, although one speaker's retroflex /d/ overlapped with the dental cluster.

An overall cross-linguistic comparison of the slope, *y*-intercept mappings, including results from Swedish and English locus equations, reveals several consistent patterns. Alveolar/dental mappings are always significantly higher on the *y* axis than labial and velar, and they typically have a much lower position on the *x* axis. Labial and velar stops are usually found close together low on the *y* axis. They are typically separated only by slope, with labial stops found lower on the *x* axis than velar stops (Sussman et al., 1992). English is an outlier here, with velar stops appearing both significantly higher on the *y* axis than anywhere else, and also being lower than labials on the *x* axis. No clear phonetic "hot spots" could be found in the space—labial stops came the closest to forming a cluster in the coordinate space, while velars and dental/alveolar stops tended to move cross-linguistically. (Sussman et al., 1992).

One of the earliest studies to examine locus equations for more than just stop consonants came from Carol Fowler in 1994. Fowler claimed that Sussman et al. mislabeled locus equations when they called them invariant specifiers of place of articulation. Krull previously established that the slope of a locus equation reflects the degree of coarticulation demonstrated by the vowel transitions (Krull, 1988). Since a higher slope indicates a lower degree of coarticulation, Fowler argues that locus equations can actually be used as measurements of "coarticulation resistance." Using current notions of locus equations, /d, t, z, s/ should have the same locus equation because they have the same place of articulation, and likewise for all other sets of consonants with the same place. However, Fowler notes that utterances with different manners of articulation (fricatives, nasals, etc.) will display different degrees of coarticulation. She predicts that locus equations will differ slightly for sounds with the same place of articulation but a difference in manner. To test this theory, ten speakers (five male and five female) were recorded pronouncing CV/t/ tokens. Consonants included /b, v, ð, d, z, ʒ, g/, and they were followed by eight vowels similar to those used by Sussman et al. in 1991. The slope and *y*-intercept values for /b/, /d/, and /g/ when averaged across all speakers were 0.79, 0.47, and 0.71 for slope and 228, 1099, and 778 for y-intercept (Fowler, 1994). These values are comparable with the results of Sussman et al. 1991. The only difference found from previous papers was that /b/ and /g/ differed only marginally in slope (Fowler, 1994). Fowler theorizes that /g/ may be more resistant to coarticulation when followed by front vowels because front vowels may pull /g/ forward to be confused with a different stop, while there are no stops close behind /g/. When Fowler compared all seven consonant regression lines to one another, she found her assumption that slopes would differ for consonants with the same place of articulation but different manners held true—/z/ and /d/ locus equations had significantly different slopes. Also, the slopes of /g/ and /v/, which have different place and manners of articulation, overlapped slightly. Despite all this, Fowler writes that the *y*-intercept values pattern as they should, with /g/ distinct from /v/, and /d/ and /z/ having relatively the same intercept value (Fowler, 1994).

Two years later Sussman and Shore conducted an expanded version of Fowler’s study in an attempt to refute her claim that locus equations were not strong indicators of place of articulation (Sussman and Shore, 1996). They began by pointing out that Fowler based her findings only on the slope of the equations, and that for locus equations to work the slope and *y*-intercept should be treated as “codependent variables in a multivariate analysis.” (Sussman and

Shore, 1996). This expanded study used twenty-two male speakers of American English. Each speaker produced alveolar consonants /d, n, z, t, s/ followed by ten vowels. One important aspect of this experiment was Sussman and Shore's discussion of how to measure $F2_{onset}$ for voiceless stop /t/. Before a vowel, /t/ is often characterized by an extended period of aspiration before the vowel begins. This aspiration period gives time for the articulators to move, and so by the time the first glottal pulse appears F2 is already in position for the vowel midpoint. If $F2_{onset}$ were to be measured at this point the resulting regression line would have a significantly steeper slope that expected, indicating a high degree of coarticulation for /t/. This would be misleading, and so Sussman and Shore suggest that $F2_{onset}$ values from /t/ should be drawn from right after the stop burst, before aspiration and the first glottal pulse. In order to keep a fair comparison between /t/ and /d/, they generated a second locus equation for /d/, called /d/@burst, where $F2_{onset}$ values were taken at the stop burst rather than at the first glottal pulse. Multivariate discriminate analysis on the locus equations showed the differences in slope were non-significant for /d/ versus /n/, /d/ versus /z/, /n/ versus /s/, and /d/@burst versus /t/. /d/@burst and /t/ had the shallowest slopes, indicating the greatest degree of coarticulation resistance. The most relevant analysis done for the purpose of this paper was a comparison of the clusters formed when slope and *y*-intercept were plotted for all of the alveolar consonants, along with points representing velar and bilabial stops /g/ and /b/. This representation showed that the three groups (alveolar, labial, and velar) formed three distinct, non-overlapping clusters. Attempts to classify slope/*y*-intercept points based on the clusters were 87% accurate for alveolar consonants. All of the incorrectly classified alveolar tokens were marked as velars. Sussman and Shore concluded the experiment by claiming that alveolar obstruents, as a class, typically have lower slopes than velar and labial places of articulation. (Sussman and Shore, 1996).

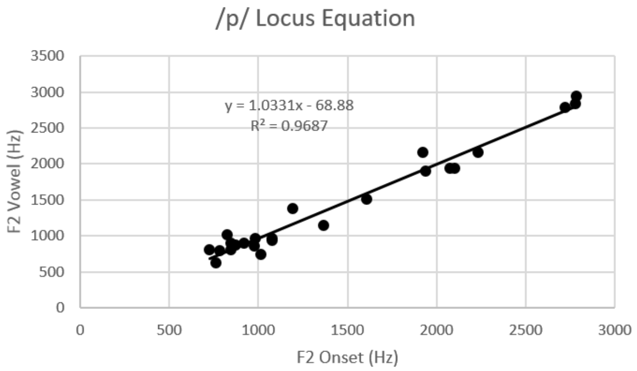


Figure 5: Example Locus Equation

2.6 MODERN LOCUS EQUATIONS

Following the initial exploration of locus equations and the replications of locus equation experiments in multiple languages, the debate began to stray into theoretical questions of the basis of the equations—acoustic or perceptual. There are a few papers, however, describing the use of locus equations as a phonetic tool for examining languages in a variety of populations. A 2007 study by Gibson and Ohde replicated previous experiments examining coarticulation in infants and toddlers using locus equations. The study recorded six girls and four boys, all aged 17-22 months, and pulled productions of CV tokens beginning with /b, d, g/. Keeping with pattern of previous findings, locus equations slopes descended from velar to bilabial to alveolar, with the slope for /g/ being significantly different than the slopes for /b/ and /d/, and the *y*-intercept for /d/ being significantly different from those of /b/ and /g/ (Gibson and Ohde, 2001). Gibson and Ohde used these results to try and support a particular theory of childhood coarticulation. The results supported the claim that a child begins with different patterns of coarticulation than an adult, and that as they grow the patterns change to match the speech patterns they are exposed to.

A year later Caleb Everett published a paper describing the use of locus equations for analysis of coarticulation in the Brazilian language Karitiâna (K) using locus equations (Everett, 2008). The data was taken from four native speakers of K, two male and two female. Each speaker produced 75 CV tokens for each of six initial consonants: /p, t, k, b, d, g/. The tokens were taken from carrier phrases, meaning they were not artificial isolated tokens, but existed in a larger phonetic frame. Analysis of the locus equations revealed the velar locus equation was very steep (nearing and even surpassing 1.0) suggesting that velar stops in K "can be considered back velars rather than front velars" (Everett, 2008). There is one particularly relevant observations made in this paper—although the actual $F2_{onset}$ and $F2_{vowel}$ values varied significantly for male versus female speakers, the locus equations and their coefficients were comparable. Comparison of locus equation coefficients produced for K with coefficients from previous studies revealed that the general trend of alveolar consonants displaying lower degrees of coarticulation than the other stops holds, but that K actually allows for more coarticulation of alveolar consonants than do other languages. This is evidence in the steeper slope seen for /t/ and /d/ in K when compared Thai, English, Swedish, etc. (Everett, 2008).

One of the best attempts at grounding locus equations in articulatory reality was produced by Iskarous, Fowler, and Whalen in 2010. They examine the definitions of coarticulation and coarticulation resistance, especially in relation to articulator positioning. A statistical analysis of various formulas from the theory of bivariate regression (which locus equations are based on) showed that locus equation slope is very close to representing coarticulation resistance (being the standard deviation of F2 at the vowel onset) normalized by deviations in the vowel (the standard deviation of F2 at the vowel midpoint). The only external factor is the correlation coefficient. Similarly, the *y*-intercept of the line is the average of F2 at the consonant release minus F2 at the vowel midpoint times the slope. These analyses drive Iskarous et al. to conclude “the locus equation intercept is therefore a complex measure affected by…C-to-V carryover coarticulation, and the average position of the tongue back and lips…” (Iskarous et al, 2010, pg. 2023). The following experiment compared the regression lines of locus equations to the actual position of the tongue body during speech. Electromagnetic Midsagittal Articulography Data (EMMA) was used to obtain the position of the tongue body at the times $F2_{onset}$ and $F2_{vowel}$ were drawn. Results from the experiment helped build an articulatory basis for locus equations, and led Iskarous et al. to the conclusion that locus equations could potentially be used to measure tongue body synergy.

Montgomery et al. published a study in 2014 describing the effects of incorrect F2 measurements and incomplete vowel sets on locus equation slopes. A typical study of locus equations is identified as focusing on CVC tokens with a voiced initial consonant followed by ten medial vowels. The study began with analysis of locus equation studies already existing in the literature for /b/, /d/, and /g/ onset consonants. These values were used to create statistical distributions of the slopes to be expected for each consonant. The next step was the creation of seventy-five sets of locus equations generated from an analysis of recordings. Each set had 20 F2 values, representing the onset and midpoint formant for each of the ten vowels. The third step was using a Monte Carlo technique to vary F2 values by up to 5%. Forty thousand unique numbers were generated, creating two thousand new simulated locus equation sets, each with 20 F2 values representing the same ten vowels. The distribution of the new slopes was normal, and so these new locus equations were accepted as error-free “samples of their populations” (Montgomery et al., 2014). From this point error was systematically applied to each of the locus equation sets. F2 values, either at onset or vowel midpoint, were randomly perturbed by 50 Hz, 100 Hz, or 200 Hz. The vowel set was also decreased to include only 8, 6, 4, or 3 vowels. The sets of vowels always included the cardinal corner vowels /i/, /a/, and /u/ to avoid locus equation error caused by lack of representation rather than incorrect measurement. A significant effect on the slopes was defined as the "mean absolute difference of the change in slope in corresponding [locus equations] in the sets of 2000" greater than 0.1 (Montgomery et al, 2014). Results showed that a vowel set of at least six vowels stayed within a 95% confidence interval for errors up to 50 Hz, but that more extreme error pushed the slopes for all three consonants over the cut-off point for error, into the next interval. The study concluded that locus equation slopes are "generally resistant to error and reduced number of vowels" (Montgomery et al., 2014).

2.7 LOCUS EQUATIONS AND SPEECH DISORDERS

So far the discussion of locus equations has been limited to standard speech across a variety of languages. There is significant evidence, however, that locus equations can be applied as a metric of measurement for disordered speech as well. Even though the equation coefficients may not fall within expected boundaries, the $F2_{onset}$ versus $F2_{vowel}$ mapping still shows a strong linear regression. One preliminary study into this area was conducted by Sussman, Marquardt, and Doyle in 2000. The experiment used locus equations as a tool to compare speech of children diagnosed with developmental apraxia of speech (DAS) to children of the same age with unaffected speech. DAS is a motor speech disorder, where children have trouble saying sounds or words. This leads to "a restricted phonemic repertoire, a predominance of omission errors, frequent vowel errors…abnormal prosodic patterns" (Sussman et al., 2000). Five children with DAS were compared to three children of similar ages who did not have DAS. The children were all between five and seven. The children produced CV tokens for /b, d, g/ by imitating an investigator. Locus equations were created for each child, and then the coefficients were plotted.

The Euclidean distance between each consonant point was calculated (/b-d/, /d-g/, /g-b/) and then those distances were totaled for the perimeter of a consonant triangle, called ED. This distance was a representation of the consonant space in the mouth. The three points mapped each consonant onto a plane. Consonants that were properly placed were further apart on the coordinate plane, while consonants that were improperly articulated would be misplaced or closer together. Children without DAS produced locus equations that closely matched those produced by adult speakers. Children with DAS produced locus equations with lower R^2 values, indicating significantly more variance in the vowels. The slopes for children with DAS were also all close to overlapping. This is reflected in the ED measure calculated for each child. Children without DAS had a mean ED value of 1.635, and children with DAS had an ED value of 0.465, a reduction of more than half. This difference was a reflection of the incorrectly articulated consonants. Although the study was not large enough to be statistically significant, this decrease suggests that DAS speakers show an "inability to refine coarticulation levels to maximally distinguish…stop place categories" (Sussman et al., 2000). More recently, Sussman et al. also conducted a study of locus equation production in adults who stutter, and found that even in such disjointed speech locus equations still emerge from the graph of F2 values (Sussman et al., 2010).

2.8 LOCUS EQUATION SUMMARY

There has been a significant amount of research done in the area of locus equations, but a few points are particularly relevant for the research discussed in this paper. (1) Locus equations work best when applied to voiced stops /b, d, g/ because it is easiest to find the $F2_{onset}$ point, and because aspiration is less of an issue. They still occur for voiceless stops and consonants with other manners, like nasals and fricatives, but there is some debate about where to measure $F2_{onset}$ in such cases for the most accurate equation. (2) Locus equations are relatively robust to measurement error, but F2 measurements should be kept as accurate as possible, and a vowel set of at least six vowels covering the three cardinal corner vowels must be used to ensure one side of the equation is not underrepresented. (3) The slope of locus equations is a reflection of coarticulation, where steeper slopes indicate more coarticulation. (4) The *y*-intercept of the locus equation is a function of several factors, but it tends to hold to certain ranges unique to place of articulation. (5) In English, slopes are expected to descend in the order labial > velar > alveolar, with *y*-intercepts increasing in the opposite order. (6) When the locus equation coefficients are plotted in a coordinate plane as (slope, intercept) points, the consonant space can be used as a tool for linguistic analysis of a language.

CHAPTER 3

SEMI-AUTOMATIC GENERATION OF LOCUS EQUATIONS

3.1 DATA

This experiment was done using the Nationwide Speech Project Corpus (NSP), collected by Dr. Cynthia Clopper (Clopper and Pisoni, 2006). This corpus was collected in an attempt to document the different dialects across America. The data includes recordings from 10 speakers (5 male and 5 female) from each of six dialects, for a total of sixty speakers. The six dialects are Mid-Atlantic (at), Midland (mi), New England (ne), North (no), South (so), and West (we). Each speaker was recorded producing ten different types of speech—from single syllable CVC tokens to short conversations. The sound files were labeled by accent, then by speaker number, then by experiment ID, and finally by token number. Speakers were numbered 0-9 for each accent, with 1-5 being male and the rest being female. For example, a file might be named “at0B0.wav,” with “at” for Mid-Atlantic, 0 as the speaker number (female), B as the token type (CVC) and 0 as the token number (utterance 0, “bean”).

3.2 CORPUS ALIGNMENT

The first step in the creation of locus equations was to align the speech. As discussed in the literature review above, $F2_{onset}$ and $F2_{vowel}$ values have to be pulled from very specific places in the vowel for the equations to work. $F2_{vowel}$ must be taken from the steady state of the vowel. The place of extraction for $F2_{onset}$ depends on the preceding consonant. For voiced stops, it should be taken from first glottal pulse following stop burst. For nasals and fricatives the value is drawn from the first visible glottal pulse of the vowel. The most controversial value is $F2_{onset}$ following voiceless stops. The period of aspiration that sometimes comes after the voiceless stops in syllable initial position allows the tongue time to move, meaning that by the appearance of the first glottal pulse, the vowel is already nearing its midpoint position. Arguments have been made for taking the $F2_{onset}$ value at the stop burst instead of at the glottal pulse, a value that should better reflect the actual frequency transition. In Sussman's paper the $F2_{onset}$ for /t/ is taken from right after the stop burst, with encouraging results (Sussman, 1996).

The first attempt at alignment was done using SPPAS, an automatic alignment program. SPPAS takes a WAV file and a text transcription, and then uses a phonetic dictionary and a model to automatically generate an aligned Praat TextGrid file (Boersma and Weenink, 2016). SPPAS was run using a python script, which systematically created a command line call for each sound file/text transcription pair, and then ran the appropriate SPPAS alignment commands. The end result was a .TextGrid file for each of the sounds files used. The automatic alignment results were mediocre at best, the boundaries often misplaced by anywhere from a few seconds to the length of the entire phrase. This occurred due to a few key issues. First was in the shortcomings of the phonetic dictionary. SPPAS generated a phonetic transcription for each recording using the CMU dictionary, which lists English words in alphabetic order and then provides a phonemic transcription for them (Bigi, 2012; CMU Pronouncing Dictionary). There can be more than one transcription for each word, to account for differences in pronunciation. Although the dictionary is fairly accurate for underlying phonemic representations of single word tokens, it quickly loses accuracy when faced with real speech, as shown in the following examples. When the phonemic underlying representation of a token is transformed into the surface form of real speech, it should change to accommodate speech characteristics like aspiration, deletion, nasalization, etc.

SPPAS, using the CMU dictionary, does not have this step. The speech models used were trained on data with allophonic variation, but the boundaries were still often misplaced. The previous paragraph discusses the importance of marking F2$_{onset}$ in exactly the right spot, especially when aspiration is involved. SPPAS didn't mark aspiration, so the vowel was often marked as starting a little ways into the vowel, or halfway through the aspiration where no reliable F2 could be found (see Figure 6 below). Another problem with the dictionary was its use of single word tokens. When SPPAS attempted to align sentences, it could not account for the ways words change in continuous speech. For instance, SPPAS correctly transcribed the flap phoneme when found in the middle of a word ("butter"), but failed to notice it at the end of a word ("but I"). Instead, the dictionary left the phoneme as a /t/.

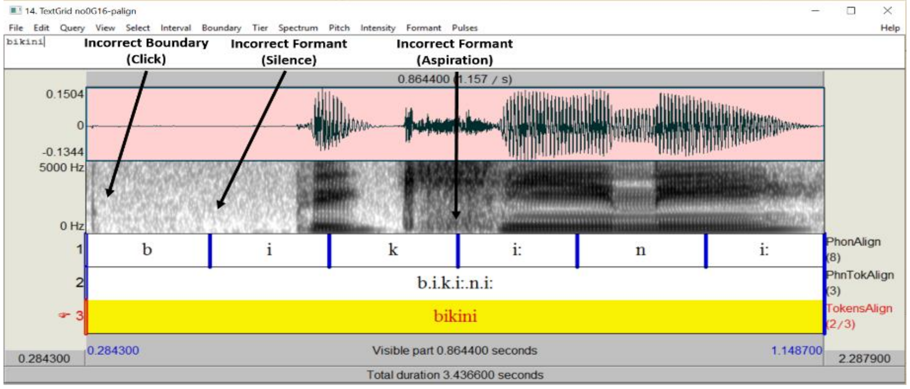

14. TextGrid no0G16-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
bikini
Incorrect Boundary (Click)
Incorrect Formant (Silence)
Incorrect Formant (Aspiration)
0.864400 (1.157 / s)
0.1504
0
-0.1344
5000 Hz
0 Hz
1
b
i
k
i:
n
i:
PhonAlign (8)
2
b.i.k.i:.n.i:
PhnTokAlign (3)
3
bikini
TokensAlign (2/3)
0.284300
0.284300
Visible part 0.864400 seconds
1.148700
2.287900
Total duration 3.436600 seconds

Figure 6: Alignment Error

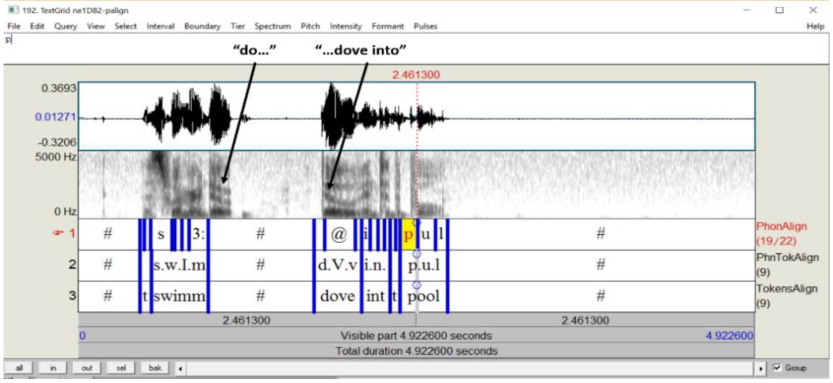


Figure 7: Speech Error

The second issue was an inability to handle outside noise in a WAV file. The recordings usually started with a "click" sound as the speaker clicked a mouse. Many of the recordings included a sigh or a yawn before the actual speech began. These non-speech sounds confused SPPAS. The alignment boundaries shifted out of place to try and include these sounds as part of the expected transcribed phonemes, and the whole alignment file would be thrown off as a result (see Figure 6). The third issue was with the speakers themselves. SPPAS expected the sound file to contain exactly what was provided in the text transcription. The transcription in the NSP included the tokens the speakers were supposed to produce, but the speakers often produced errors—missing their cues or stumbling over a word and starting again. For example, a recording that was meant to contain "The swimmer dove into the pool" might actually contain "The swimmer do…dove into the pool." (Figure 7) In these cases, SPPAS tried to align all of the speech, but the mismatch in the transcription and the speech could not be overcome.

SPPAS is a strong automatic alignment program, but it was incapable of providing the pinpoint accuracy needed to achieve results in this research. Valid locus equations require F2 values drawn from specific points in a sound, and the SPPAS alignment placed boundaries too far from where they needed to be. The only visible solution was to correct the alignments by hand. In the interest of time, the number of speakers was cut to 24—2 male and 2 female speakers from each of the six accents. Additionally, the number of files included in the analysis was cut to a more reasonable number. The original corpus contained 203 sentences and 195 single word tokens across 10 speakers, for a total of 3,980 tokens. All of the CVC tokens were kept, along with 43 of the high probability (HP) sentences and 25 of the multisyllabic words, leading to a corpus of 1,430 utterances. Many of the tokens had multiple transitions in one file—altogether, there were 10,880 vowel transitions included from all consonants. These tokens were chosen to satisfy the following: Consonants for each of the three examined places of articulation (labial, alveolar, velar) were represented as voiced stops, voiceless stops, fricatives, and nasals. This excludes the velar nasal /ŋ/, which did not occur often enough in the corpus for an accurate locus equation to be created. Each of these consonants had tokens transitioning into at least 6 different vowels, including the three cardinal vowel corners, /i/, /u/, and /a/. These transitions were seen at least once for each consonant, but preferably more often. These constraints were chosen to ensure that the F2 values taken from the aligned sound files would form accurate representative locus equations for each speaker, while cutting the number of hand aligned files down to a size fitting the time constraints. The choices were made to mimic previous experiments done with locus equations, and the inclusion of at least 6 vowels with the inclusion of /i, a, u/ is based on the results detailed in Montgomery (2014). The list of included files and their contents can be found in Appendix B.

The alignment choices were made to match the required $F2_{onset}$ values described above. The boundary for vowels following voiced stops was placed at the first glottal pulse following the stop burst. Following fricatives and nasals the vowel boundary was placed at the first discernable pulse of vowel, and vowel boundaries after voiceless stops were put at the first clear and stable appearance of F2. One example for each of these can be seen in Figure 8 below. When a phoneme transcribed by SPPAS was incorrect (for example, a /t/ where there should be a flap) the transcription was corrected. When the speaker misspoke, the incorrect speech was marked within its own segment, and the boundary was labeled with “#”. If a speaker paused for a significant period of time in the middle of a sentence, a “#” was added to mark a silent period. If the vowel occurred at the end of the utterance, the boundary was placed where the formant tracking stopped. Otherwise, the boundary was placed where the next sound began.

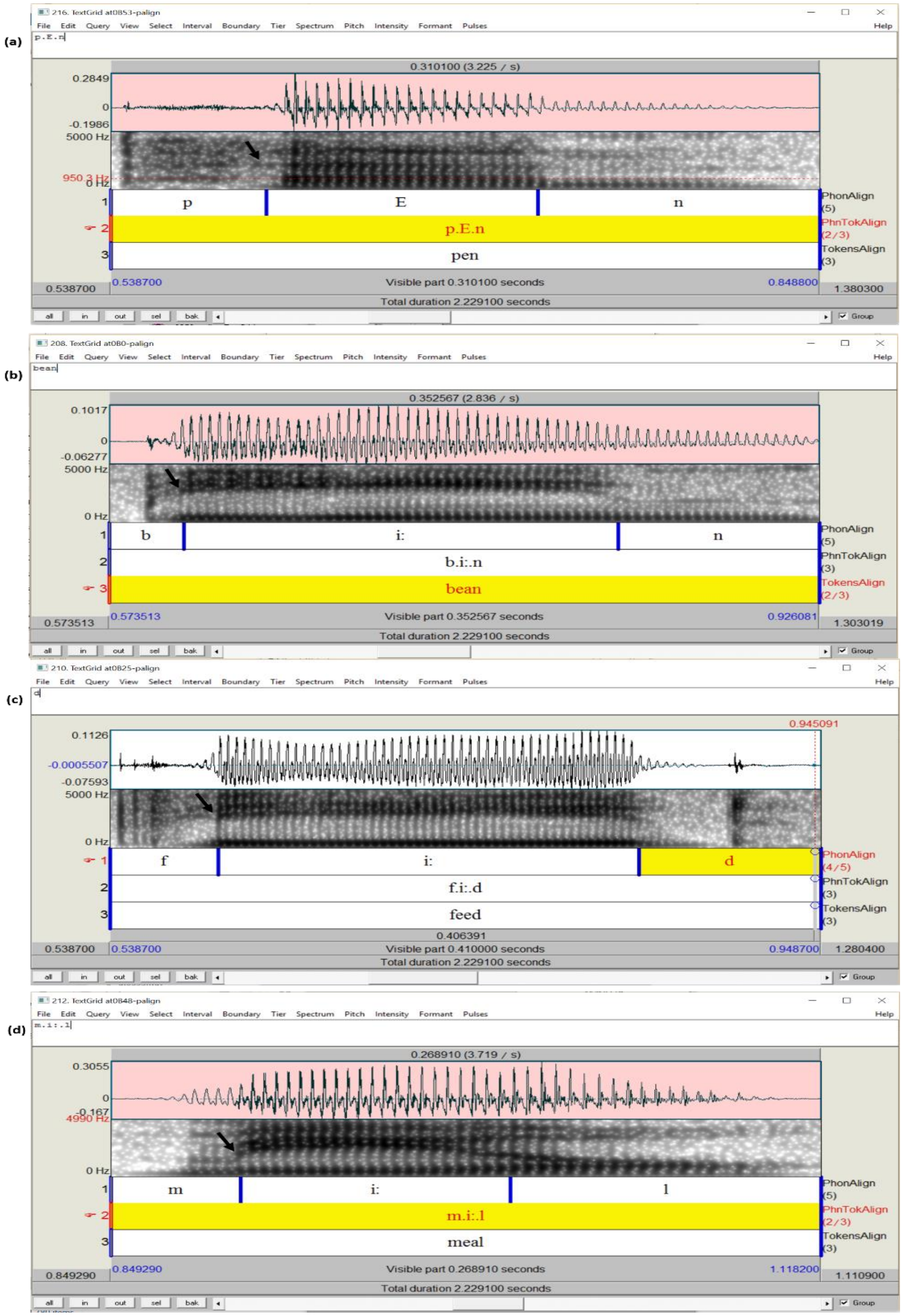

(a)
216. TextGrid at0853-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
p.E.n
0.310100 (3.225 / s)
0.2849
0
-0.1986
5000 Hz
950.3 Hz
0 Hz
1 p E n PhonAlign (5)
2 p.E.n PhnTokAlign (2/3)
3 pen TokensAlign (3)
0.538700
0.538700
Visible part 0.310100 seconds
0.848800
1.380300
Total duration 2.229100 seconds
all in out sel bak
☑ Group
(b)
208. TextGrid at080-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
bean
0.352567 (2.836 / s)
0.1017
0
-0.06277
5000 Hz
0 Hz
1 b i: n PhonAlign (5)
2 b.i:.n PhnTokAlign (3)
3 bean TokensAlign (2/3)
0.573513
0.573513
Visible part 0.352567 seconds
0.926081
1.303019
Total duration 2.229100 seconds
all in out sel bak
☑ Group
(c)
210. TextGrid at0825-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
d
0.945091
0.1126
-0.0005507
-0.07593
5000 Hz
0 Hz
1 f i: d PhonAlign (4/5)
2 f.i:.d PhnTokAlign (3)
3 feed TokensAlign (3)
0.406391
0.538700
0.538700
Visible part 0.410000 seconds
0.948700
1.280400
Total duration 2.229100 seconds
all in out sel bak
☑ Group
(d)
212. TextGrid at0848-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
m.i:.l
0.268910 (3.719 / s)
0.3055
0
-0.167
4990 Hz
0 Hz
1 m i: l PhonAlign (5)
2 m.i:.l PhnTokAlign (2/3)
3 meal TokensAlign (3)
0.849290
0.849290
Visible part 0.268910 seconds
1.118200
1.110900
Total duration 2.229100 seconds
all in out sel bak
☑ Group

Figure 8: Boundaries
(a) Voiceless Stop onset. (b) Voiced Stop onset.
(c) Voiceless Fricative onset. (d) Nasal onset.

3.3 LOCUS EQUATION CONSTRUCTION

Once the sound files were all properly aligned, the F2 data had to be pulled from the files and transformed into locus equations. This was done using a combination of a Locus Equation program implemented in java and a Praat script. The steps were as follows:

1. The Locus Equation program reads all of the files from specified folders, creates WAV/.TextGrid pairs and extracts demographic information for the pair.
2. The Locus Equation program runs the Praat script on the pair of files.
3. The Praat script pulls all relevant F2 values from the recording and writes them to a CSV file.
4. The Locus Equation program reads from the completed CSV file and creates "Speaker" objects with sets of "Consonants".
5. The Locus Equation program calculates the regression line for each consonant for each speaker.
6. The Locus Equation program identifies outliers in the data, deletes them, and recalculates the locus equation.
7. The Locus Equation program writes the locus equation slope, *y*-intercept, R^2 value, and Standard Error to a file as input for a classifier.

The details of this process and justification for each step are provided below.

In order to pull formant values at timestamps, Praat must be provided with both a sound file, for the recording, and a TextGrid file with the alignment. The hand-aligned files were all stored in a folder for each accent. The WAV file was labelled using the system seen in Table 1 below, and the TextGrid files shared the same name with the addition of "-palign.TextGrid". The Locus Equation program read through each of the accent files, and created pairs of WAV files and TextGrid files with matching names. It then used a regular expressions (regex) to pull information from the file name—accent, sex of the speaker, and token number. The program looped through the collection of file pairs, and fed each one into Praat using a system call to praatcon.exe, a console version of Praat. The Praat script was structured to take five arguments as input: The WAV file name, the TextGrid file name, the CSV output file name, the speaker ID (accent + speaker number), and the speaker sex (Table 1).

Input	**Value**
WAV File Name	“at0B0.wav”
TextGrid File Name	“at0B0-palign.TextGrid”
CSV Output File Name	“Formants.csv”
Speaker ID	at0
Speaker Sex	Female
Dialect	at

Table 1: SPPAS Input

The Praat script began by loading the sound file and the TextGrid file. Next, the script created a formant object for the sound file. The formant was created using the “Burg” method, which takes five arguments. The first, time step, specifies how often the formants will be sampled for the sound file. This was kept at the standard value, 0.0. The second is the maximum number of formants Praat will search for at each analysis step. The recommended setting is 5 for human speech, however, if a sound file had too much background noise searching for 5 formants sometimes led Praat to track an extra, non-existent formant, leading to incorrect readings. After looking through the sound files, it was decided that 5.5 would be the maximum number of formants for back and central vowels, and 4.5 would be the maximum number for front vowels. The effects of these settings are discussed later when outliers are addressed. The third formant setting is “Maxiumum formant,” which serves as the ceiling of the formant search. The standard settings here are 5500 Hz for females and 4500 Hz for males, who have lower voices and thus lower formant values overall. The fourth setting, “Window length” is measured in seconds. This dictates the Gaussian analysis window used in calculating the formants. The value was set to 0.20 seconds, for a total Gaussian window of 0.4 seconds (0.2 on each side). Finally, the last setting is “Pre-emphasis from,” which dictates the starting frequency Praat should search for formants in. The default setting of 50 Hz was used here.

Once the Formant objects were created, Praat looped through each phonetic segment in the TextGrid looking for vowels. When a vowel was located, the Formant settings were updated appropriately and then the F2 values were taken. $F2_{onset}$ was pulled from the boundary marking the beginning of that vowel segment. $F2_{vowel}$ was taken from the midpoint of the segment (duration * 0.5) for monophthongs (Figure 9), and the first quarter of the segment (duration * 0.25) for diphthongs (Figure 10). These equations for pulling the midpoint were decided upon after consideration of previous experiments and manually pulling formants to test the resulting locus equations. While some locus equations experiments were very careful to draw $F2_{vowel}$ from the best steady state available, others took $F2_{vowel}$ from 20 milliseconds in with comparable results. Additionally, visual inspection revealed that the midpoint of the vowel almost always fell within the steady state. After both F2 values were sampled for a vowel, they were written into the output file given in the arguments, along with the input file name, speaker ID, speaker sex, vowel label, preceding consonant label, and dialect.

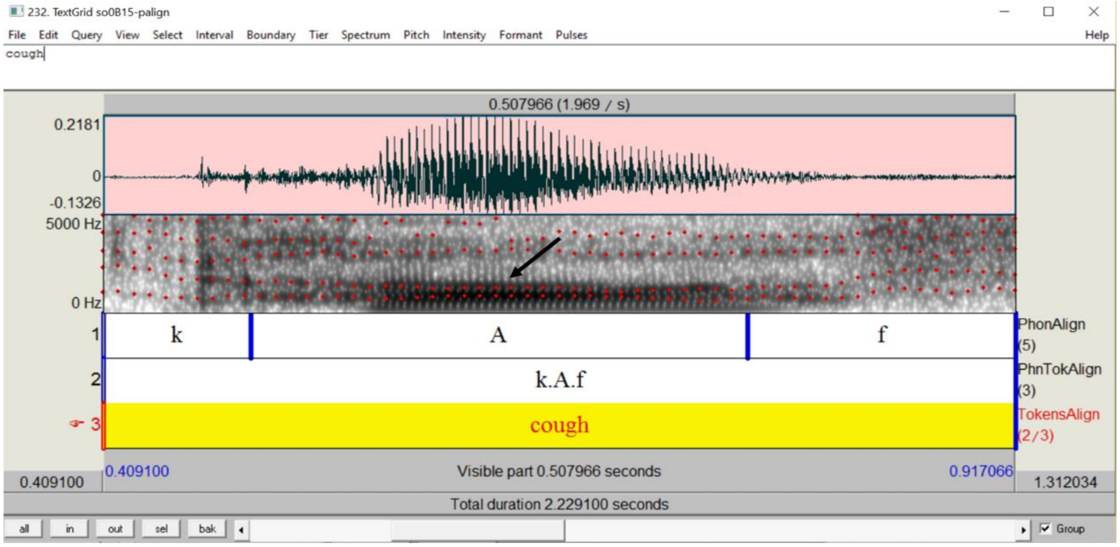

232. TextGrid so0B15-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
cough
0.507966 (1.969 / s)
0.2181
0
-0.1326
5000 Hz
0 Hz
1 k A f
2 k.A.f
3 cough
PhonAlign (5)
PhnTokAlign (3)
TokensAlign (2/3)
0.409100
0.409100
Visible part 0.507966 seconds
0.917066
1.312034
Total duration 2.229100 seconds
all in out sel bak
Group

Figure 9: Monophthong Midpoint

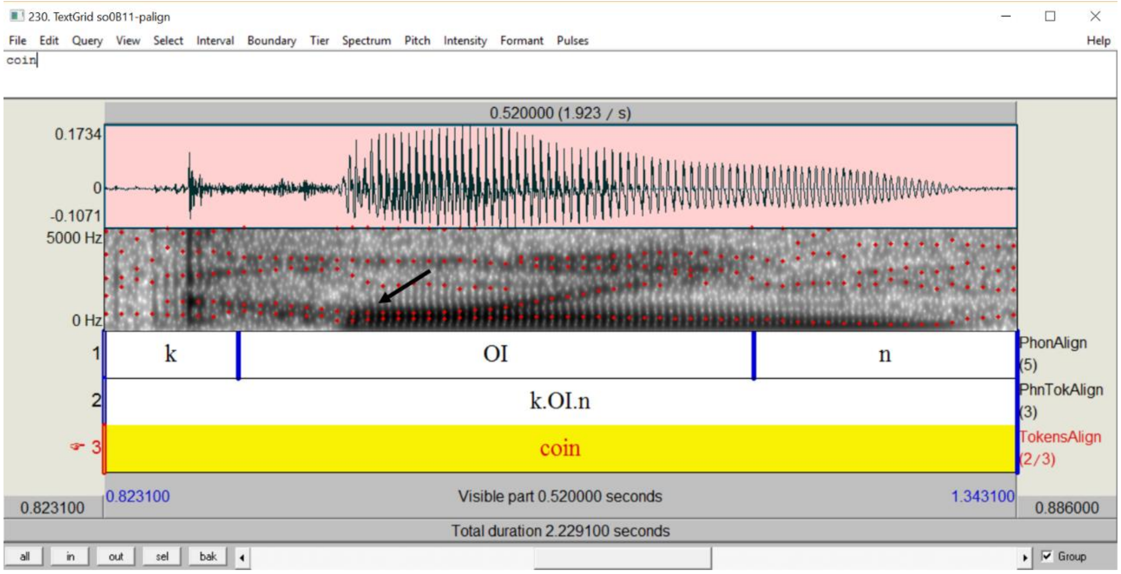

230. TextGrid so0B11-palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
coin
0.520000 (1.923 / s)
0.1734
0
-0.1071
5000 Hz
0 Hz
1 k OI n
2 k.OI.n
3 coin
PhonAlign (5)
PhnTokAlign (3)
TokensAlign (2/3)
0.823100
0.823100
Visible part 0.520000 seconds
1.343100
0.886000
Total duration 2.229100 seconds
all in out sel bak
Group

Figure 10: Diphthong Midpoint

Once the final call to a Praat script has finished, the Locus Equation program begins the next step of reading from the created CSV file and creating regression lines for each speaker. The program was structured in such a way that each speaker was its own object, represented by the speaker ID [accent] + [ID number]. Each speaker had a set of Consonant objects, and each Consonant had a set of Token objects. These Token objects contained the name of the WAV file they were taken from, the vowel they represented, the consonant they followed, and the F2 values Praat sampled for that vowel. The structure is visualized in Figure 11 below. Each line in the CSV file contained the $F2_{onset}$ and $F2_{vowel}$ values for one vowel transition from one consonant for one speaker. The Locus Equation program read in the CSV line by line and sorted the Tokens by Speaker and then by Consonant. When the entire CSV file had been read, the program moved on to the creation of locus equations for each speaker.

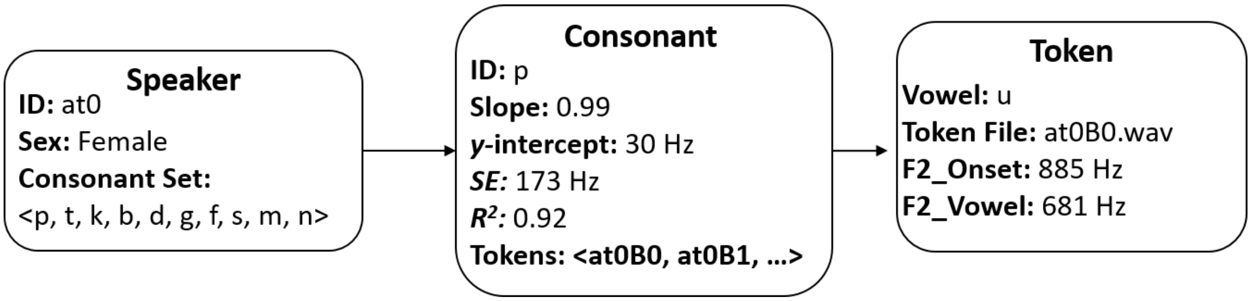


Figure 11: Java Classes

To calculate the locus equation lines, the Locus Equation program used a nested loop that iterated through every Consonant object of every Speaker. Every Consonant had a set of Token objects with $F2_{onset}$ and $F2_{vowel}$ values. Locus equations were created by mapping the line of best fit to these $F2_{onset}$, $F2_{vowel}$ pairs, with $F2_{vowel}$ on the *x*-axis and $F2_{onset}$ on the *y*-axis. The regression lines were represented by the equation $y = mx + b$, with *m* as the slope and *b* as the *y*-intercept. The coefficients *m* and *b* and the R^2 value were found with the following equations:

1. N = Number of $F2_{vowel}$, $F2_{onset}$ pairs.
2. $x_{sum} = \sum_{1 \text{ to } N} F2_{onset.}$
3. $y_{sum} = \sum_{1 \text{ to } N} F2_{vowel.}$

4. $xy_{sum} = \sum_{1 \text{ to } N} (F2_{vowel} * F2_{onset})$.

5. $x^2_{sum} = \sum_{1 \text{ to } N} (F2_{vowel} * F2_{vowel})$.

6. $y^2_{sum} = \sum_{1 \text{ to } N} (F2_{onset} * F2_{onset})$.

7. $m = \frac{(N*xy_{sum})-(x_{sum}*y_{sum})}{(N*x^2_{sum})-(x_{sum}*x_{sum})}$.

8. $b = \frac{(x^2_{sum}*y_{sum})-(x_{sum}*xy_{sum})}{(N*x^2_{sum})-(x_{sum}*x_{sum})}$.

9. $R^2 = \frac{((N*xy_{sum})-(x_{sum}*y_{sum}))^2}{((N*x^2_{sum})-(x_{sum}*x_{sum}))((N*y^2_{sum})-(y_{sum}*y_{sum}))}$.

10. $S.E. = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

These values were stored as variables in the Consonant object. The accuracy of the equations was checked by running the Token values through the Locus Equation program and then plotting the same F2 points in excel and plotting a trend line. Both approaches returned the same values.

3.4 OUTLIER DETECTION AND REMOVAL

The final step in the preparation of locus equations was finding and removing outliers and then replotting the locus equation lines. This step was necessary due to the use of an automated Praat script for pulling formant values, rather than pulling them by hand. Although Praat is fairly accurate, it can occasionally miss formants or pick up on extra “ghost formant” values, returning incorrect results. The typical approach in this situation would be to locate outliers and then go back and hand check them in Praat to ensure the formants were properly measured. While this method ensures the most accuracy, it requires further human interference into a process we are attempting to automate. The automatic detection and removal of outliers in this program is an attempt to replace the human error-checking. For this research, an outlier was described as any point outside the 98.5% prediction interval from the plotted regression line. In linear regression, a confidence interval marks the regression line with some level of confidence. In the case of outlier detection, the interval needed to be focused on the points rather than the line itself. A prediction interval is an interval around the regression line that marks where a newly plotted point might fall on the graph with some degree of confidence. This interval was deemed a better fit for the detection of outlier points since the calculation of the interval takes the points into account rather than the line itself. The equations used to calculate the prediction intervals were as follows:

The prediction interval for y_i at x_i is:

$$\hat{y}_i \pm t_{crit} * s.e.$$

11. $\hat{y}_i = mx_i + b.$

12. $s.e.(\hat{y}_i) = \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$.

13. $\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$

In the above equations, $\hat{y}_i$ is the predicted *y* value for any *x* value given the regression line. t_{crit} is the statistical t-value for a two-tailed distribution. This value was pulled from a table storing t values. *s.e.* is the standard error of prediction for the equation. This is different from the standard error of estimation *(SE),* given above and used as a method for measuring how well the plotted points fit the regression line. The third equation is recognizable as a close variant of the standard error of estimation (*SE).* The only difference lies in the division (*n* – 2) rather than by *N*. The plus/minus in the first equation accounts for the upper and lower bounds of the prediction intervals. A point was plotted at every *x* value ($F2_{vowel}$) using the above equations, and then these points were used to draw the interval around the regression line. For every coordinate point, the $F2_{vowel}$ value was used to calculate the upper and lower limits of a predicted $F2_{onset}$ value, using the t_{crit} value for 98.5% confidence. If the actual $F2_{onset}$ value was greater or lower than the limit, it was discarded as an outlier. Once every point on a line had been checked using this method, the locus equation line was recalculated using the methods described in the previous section, but with the reduced set of coordinate points. It should be noted that this process was done only once per regression line, with the intent of removing as few points as possible. Also, this method was only used on lines with more than twenty points plotted for the regression line. Based on observations of the line and a few tests that involved moving data points, it was decided that any fewer points would not be sufficiently robust to the removal of an outlier, and that every point was necessary for an accurate regression line. In the NSP data set, /s/ and /g/ were the two locus equation lines that were not evaluated using outlier removal. Instead, each of these points was checked by hand to ensure accuracy of the regression line.

Figures 12 and 13 show two examples incorrect formant measurements in Praat. Figure 12 shows a formant reading which was much too low. The “ghost” second formant which caused the incorrect reading is marked by an arrow, and can clearly be seen in the spectrogram. Figure 13 shows the opposite problem—Praat failed to track the first formant (F1) in the spectrogram, and so it treated F2 as the first formant and pulled an F3 value instead of an F2 value. The correct formant value and the value actually pulled are both marked by arrows. Outliers of this significance had an extreme influence on the locus equation line and coefficients. Figure 14 shows three examples of outlier removal for the locus equation lines, one for /p/, /t/, and /k/. The first image shows the plotted points and regression line prior to outlier removal. The second image is the same set of points, but the outliers were checked by hand and fixed rather than automatically detected and discarded. The third image depicts the plotted points and regression line after automatic outlier removal. Note the difference in the locus equation coefficients. The difference in both slope and *y*-intercept between the unchecked locus equation line and the two fixed lines is significant, but the coefficients are comparable between the line with fixed outliers and the line with discarded outliers.

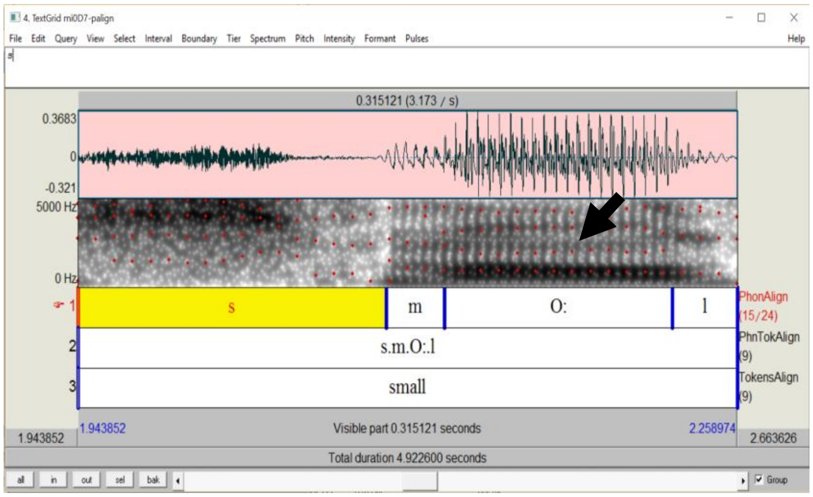


Figure 12: Formant Too High Error

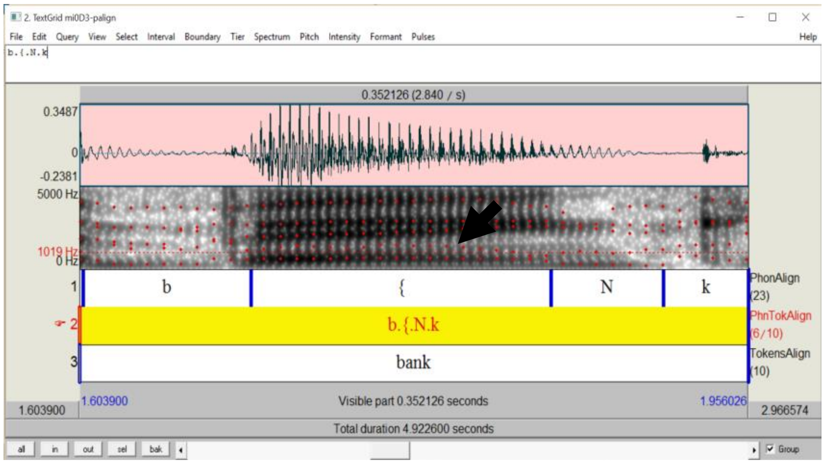


Figure 13: Formant Too Low Error

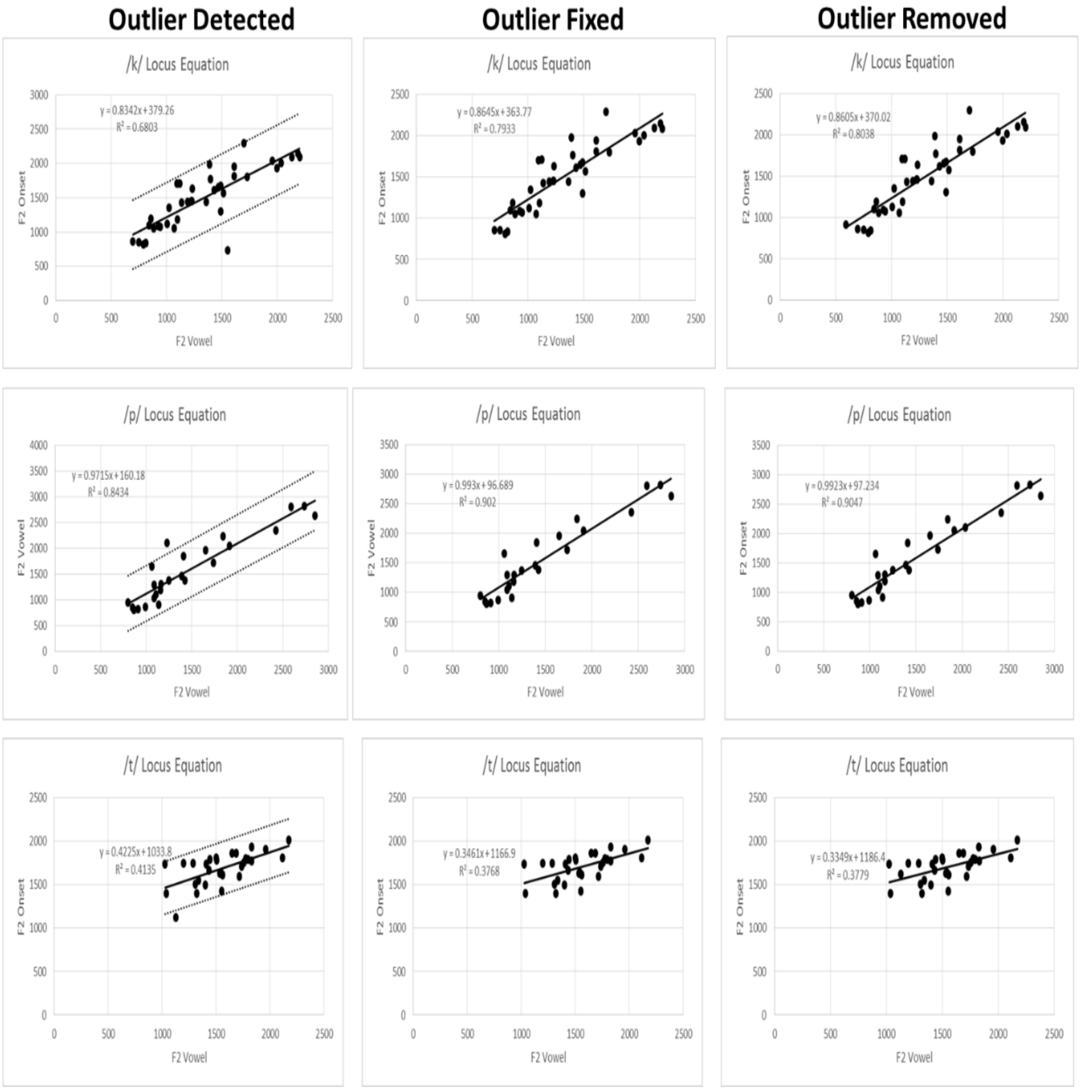


Figure 14: Outlier Removal Results

CHAPTER 4

LOCUS EQUATIONS AND DIALECTS

4.1 COMPARISON WITH PREVIOUS STUDIES

One way to check the validity of the generated locus equations was to compare them to previously attained results. Following the experiments of Sussman and Fowler, we can expect that the slopes of the locus equations will decrease in the order labial < velar < alveolar, and that the y-intercepts will pattern similarly (Sussman et al., 1991; Fowler, 1994). Table 2 below shows the Locus Equation slopes (*m*) and y-intercepts (*b*) for every consonant, for every speaker. The mean values for each coefficient pattern as expected: for every manner of articulation, labial > velar > alveolar. Each manner and place group is discussed individually below. The results obtained in this experiment are compared to the voiced stop locus equation coefficients from Sussman's 1991 paper (Sussman et al., 1991). All locus equations for speaker at0 are included in Appendix C for reference.

	Consonants																			
	Voiced Stops						Voiceless Stops						Nasals				Voiceless Fricatives			
	/b/		/d/		/g/		/p/		/t/		/k/		/m/		/n/		/f/		/s/	
Speaker	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***	***m***	***b***
at0	0.68	495	0.47	1201	0.83	510	0.99	97	0.58	957	0.84	440	0.5	651	0.32	1261	0.73	378	0.47	1000
at1	0.9	33	0.61	667	0.48	1150	0.97	57	0.52	833	0.94	241	0.82	236	0.58	635	0.82	191	0.48	848
at2	0.73	264	0.7	582	1	188	1.01	31	0.56	832	0.91	276	0.63	444	0.45	834	0.73	379	0.7	550
at6	0.77	332	0.51	1080	0.77	713	0.96	114	0.67	714	1	135	0.62	555	0.49	975	0.84	165	0.86	257
mi0	0.87	128	0.52	1018	0.6	1074	1.03	-55	0.84	367	1.01	83	0.76	383	0.65	673	0.98	85	0.59	739
mi1	0.85	111	0.6	653	0.65	727	1.01	32	0.64	642	0.96	173	0.67	362	0.59	582	0.81	182	0.66	544
mi2	0.78	231	0.51	871	0.57	1009	0.93	126	0.35	1167	0.87	363	0.74	250	0.6	593	0.85	178	0.71	481
mi6	0.7	413	0.52	951	0.84	455	1.04	-64	0.37	1238	1.04	126	0.76	349	0.52	817	0.88	137	0.86	291
ne0	0.77	289	0.41	1277	0.76	763	1.07	-94	0.47	1146	1.01	73	0.66	453	0.55	830	0.76	278	0.65	695
ne1	0.78	293	0.63	716	0.56	996	1.02	11	0.48	991	0.88	377	0.72	339	0.67	521	0.82	223	0.83	316
ne2	0.86	130	0.56	738	0.5	1038	0.91	147	0.47	886	0.85	329	0.81	187	0.44	799	0.91	113	0.62	530
ne6	0.87	175	0.59	870	0.66	938	1	23	0.55	957	0.93	257	0.74	385	0.44	1016	0.78	322	0.64	728
no0	0.64	538	0.57	920	0.68	883	0.97	111	0.53	1014	0.96	231	0.74	264	0.43	1051	0.78	242	0.65	616
no1	0.92	72	0.57	721	0.44	1160	0.9	162	0.41	1021	0.83	380	0.54	721	0.44	894	0.8	199	0.64	532
no2	0.82	164	0.64	620	0.42	1122	0.93	91	0.6	679	0.82	387	0.68	272	0.68	470	0.84	193	0.66	542
no6	0.75	394	0.49	1161	0.87	493	0.93	126	0.67	762	0.97	251	0.63	425	0.51	784	0.78	210	0.92	197
so0	0.56	698	0.61	839	0.68	872	0.82	274	0.64	796	1.03	107	0.39	1006	0.38	1137	0.69	452	0.91	158
so1	0.83	138	0.53	815	0.41	1222	0.99	20	0.41	1061	0.88	309	0.66	325	0.49	743	0.85	150	0.58	641
so2	0.78	165	0.46	859	0.51	1039	0.92	141	0.53	777	0.86	364	0.76	182	0.58	574	0.78	204	0.56	647
so6	0.84	161	0.51	984	0.81	596	0.96	102	0.56	1003	1.01	157	0.49	731	0.4	1139	0.83	170	0.65	544
we0	0.7	384	0.44	1253	0.78	785	0.99	40	0.59	968	0.89	413	0.54	669	0.4	1265	0.82	276	0.75	542
we1	0.81	193	0.43	947	0.57	867	0.99	61	0.55	783	0.86	340	0.73	333	0.67	543	0.87	198	0.65	601
we2	0.96	34	0.54	752	0.65	774	1.04	-7	0.48	862	0.96	188	0.82	143	0.53	663	0.76	264	0.73	458
we6	0.84	88	0.49	1023	0.85	623	1.1	-115	0.64	811	0.91	322	0.67	436	0.48	950	0.86	134	0.69	589
Mean	0.793	247	0.537	897	0.663	833	0.979	60	0.546	886	0.927	263	0.671	421	0.512	823	0.815	222	0.686	544

Table 2: Dialect Locus Equation Results

Voiced stops are the manner class most often examined in Locus Equation papers (Fowler 1994; Lindblom 1963; Sussman et al, 1991), and so we will examine these first. In the Sussman et al 1991 paper, the mean slope for /b/ was 0.89 with a y-intercept of 99 Hz, the mean slope for /d/ was 0.42 with a y-intercept of 99 Hz, and the mean slope for /g/ was 0.71 with a y-intercept of 792 Hz. Table 3 shows a direct comparison between the results of the 1991 paper and the results from this experiment using male, female, and total averages for the Voiced Stop Equations. The slope averages are comparable, with /b/ > /g/ > /d/. The /b/ slope for the current experiment is shallower than previous results. The breakdown by male and female shows that the male values are very similar, but the female slope is flatter than expected. This result is likely due to the conversational nature of the speech used to create the equations. In the 1991 experiment single tokens beginning with [b] were produced, meaning every token began with a total stop of air followed by a stop burst and then the vowel (Sussman et al., 1991). In the multisyllabic and sentence tokens of the NSP corpus, the /b/ transitions sometimes took place within a word, and [b] sound was not as well formed. It is worth mentioning that one female speaker in particular, so0, had a /b/ locus equation slope of 0.56, a value low enough to pattern more closely with an alveolar stop. The tokens for this locus equation were checked for outliers, but all of the vowel transitions proved to be correctly measured. Additionally, the y-intercept for so0’s /b/ locus equation falls at 698 Hz, which is very high for a /b/ locus equation but lower than it would be for an alveolar locus equation with the same slope (see speaker ne6). Re-examining the speakers showed that so0 had the most obvious Southern accented speech of all four speakers taken from that region, and so it is possible that the speaker’s dialect is responsible for the slope value. The different in the locus equation could be a reflection of the coarticulation patterns, or they could be a result of fronted vowels in the southern dialect. The y-intercept values for male /b/ locus equations are as expected, and the female values match the lower slope.

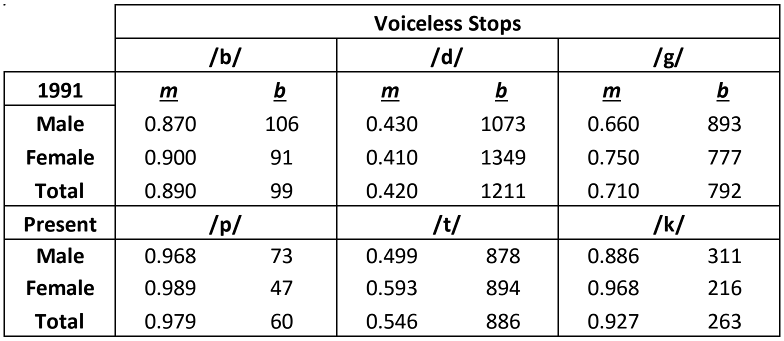
	Voiced Stops					
	/b/		**/d/**		**/g/**	
1991	***m***	***b***	***m***	***b***	***m***	***b***
Male	0.870	106	0.430	1073	0.660	893
Female	0.900	91	0.410	1349	0.750	777
Total	0.890	99	0.420	1211	0.710	792
Present						
Male	0.836	152	0.564	745	0.565	941
Female	0.749	341	0.511	1048	0.760	726
Total	0.789	252	0.543	885	0.624	911

Table 3: Voiced Stop Averages

The average /d/ locus equation slope, while still flat enough to be characteristic of an alveolar stop transition, was overall steeper than the values found in the 1991 experiment (Sussman et al, 1991). Both male and female speakers produced a steeper slope, leading in turn to slightly lower y-intercept values than expected. Once again, this can be accredited to the multisyllabic and sentence tokens in the corpus. A steeper slope indicates more coarticulation between the initial consonant and the following vowels. The continuous nature of the tokens in the present experiment may have led to heavier coarticulation between the [d] stops and the following vowels. While the /d/ slopes are steeper and the y-intercepts are lower than in previous experiments, the mapping of /d/ locus equation coefficients and /b/ locus equation coefficients still shows significant separation between the two clusters (see Figure 15).

Voiced velar stops were the furthest from their expected values. Unlike /b/ and /d/ locus equations, which patterned as expected with a few differences, /g/ locus equations were often significantly shallower than previous experiments would imply. This result is not due to any error in previous experiments, or due to any error in measurement of vowel transitions for the locus equation. The fault lies instead with the distribution of tokens with /g/ transitions available in the NSP corpus. It was previously stated that a valid locus equation must include at a minimum: transitions into six distinct vowels, where three of those vowels cover the cardinal corners of the vowel space /i, a, u/. There were enough tokens in the corpus for each speaker to have a /g/ locus equation that met these requirements, so it was included in the experimentation. Closer evaluation, however, revealed that the tokens are unbalanced between front vowels and back vowels. /i/ and /æ/ in particular had four or five tokens each, while /u/ was only represented once or twice. As seen in previous experiments (Sussman et al, 1991; Fowler, 1994; Krull, 1988) the /g/ transition in English is better characterized by two locus equations, an equation with a shallow slope similar to the /d/ locus equation marked by front vowels, and an equation with a steep slope, similar to a /b/ locus equation, following the back vowels. The imbalance between front and back vowels in the token set led to /g/ locus equations shallower than expected. The slope of the /g/ locus equations in this experiment was ultimately a result of the mid-vowel transitions. A few speakers, like at0 or so6, had lower F2 onset values for mid-vowels, which helped make the slope of the locus equation steeper. Other speakers, like at1 and no2, tended towards higher F2 onset values for mid-vowels, pushing the slope away from the back vowels and making it shallower. Females tended to have lower F2 onset values for mid-vowels, leading to steeper /g/ locus equation slopes. Although the average /g/ slope does fall between the slope for the /b/ locus equation and the /d/ locus equation, the average y-intercept for the /g/ locus equation is actually higher that the average y-intercept for the /d/ locus equation, which is not as expected. These inconsistencies with the /g/ locus equations led to a few classifier errors, and they highlight the importance of a well-balanced vowel set. These are further discussed in the next section, 4.2.

The next manner class is voiceless stops. The three voiceless counterparts to the voiced stops, /p, t, k/, were members of this class. Of every class examined in this set, voiceless stops had the largest number of tokens in the corpus, which led to the most accurate and well-balanced locus equations. They are not examined as often as voiced stops because the aspiration that often occurs after a voiceless stop gives the tongue time to move, which in turn skews the transitions from consonant to vowel. As described in Chapter 3, F2 onset was taken at the first visible steady F2 in an attempt to counteract the aspiration. /t/ and /k/ transitions have been examined before, with success (Sussman and Shore, 1996; Everett, 2008). The initial expectation was that each voiceless stop would have locus equation coefficients similar to those seen for voiced stops in the same place. That is, /b/ and /p/ would be similar, /d/ and /t/ would be similar, and /g/ and /k/ would be similar. This generally held true. Table 4 below shows the locus equation coefficients for /p, t, k/ in direct comparison to the locus equation coefficients for /b, d, g/ from the Sussman et al. 1991 study.



	Voiceless Stops					
	/b/		/d/		/g/	
1991	***m***	***b***	***m***	***b***	***m***	***b***
Male	0.870	106	0.430	1073	0.660	893
Female	0.900	91	0.410	1349	0.750	777
Total	0.890	99	0.420	1211	0.710	792
Present	/p/		/t/		/k/	
Male	0.968	73	0.499	878	0.886	311
Female	0.989	47	0.593	894	0.968	216
Total	0.979	60	0.546	886	0.927	263

Table 4: Voiceless Stop Averages

It is clear from the comparison above that voiceless stops do tend to pattern with voiced stops that have the same place of articulation. The /p/ locus equation slope is steep, with average values for both men and women nearing one. This matches the steep slope of the average /b/

locus equations from the 1991 paper, and the relatively steep slopes of the /b/ locus equations from this study. The /p/ locus equation slope may be greater than the /b/ locus equation slope because /p/ is often aspirated, so $F2_{onset}$ values were measured closer to the vowel than they were for /b/. The y-intercept averages are all under one hundred. Examination of individual locus equation coefficients (Table 2) shows that many speakers had /p/ locus equation slopes greater than 1, with negative y-intercept values. The /t/ locus equation coefficients also meet expectations. Although not as shallow as the /d/ locus equation slopes from the 1991 study, the /t/ locus equation slopes from this study are shallower than the /d/ locus equation slopes, and significantly different from both the /p/ locus equation slopes and the /k/ locus equation slopes. The y-intercept values for the /t/ locus equations are significantly higher than the y-intercept values for the other two voiceless stops. /t/, like /d/, also had a higher Standard Error and a lower R^2 value.

Like /g/, /k/ is the most dissimilar to the expectations set by the results from the Sussman paper. While the voiced /g/ locus equation slope from Sussman's paper was a moderate value between the slope of the /b/ locus equation and the /d/ locus equation, the voiceless /k/ locus equation slope is much steeper, closer to the slope of the /p/ locus equation than that of the /t/ locus equation. Unlike the /g/ locus equations from this study, the /k/ locus equation values are not caused by any error in vowel token distribution or formant measurement. Instead, it appears that /k/ locus equations lack the flatter slope seen with front vowels in /g/ locus equation slopes. In other words, /k/ does not seem to resist coarticulation with front vowels as strongly as /g/ does. The lack of a flatter slope makes the locus equation values for /k/ much steeper than those for /g/. Despite this, the average /k/ locus equation slope does still fall between the /p/ slope and the /t/ slope. Additionally, the /k/ locus equation y-intercept is distinctively much higher than the

/p/ locus equation y-intercepts, setting the two clusters apart. This separation is examined more closely in the next section.

The third manner class examined is voiceless fricatives, containing only /f/ and /s/, for labial and alveolar places of articulation. The closest consonant to a velar voiceless fricative would be /h/, but this sound is known for moving considerably with the vowels, and so it is not used in locus equation evaluations. Locus equations for fricative consonants are first discussed in depth in Fowler's response to the Sussman et al. paper (Fowler, 1994; Sussman et al., 1991). Fowler notes that fricatives are characterized by less constriction in the vocal tract, and this characteristic means they, /s/ in particular, will be more susceptible to coarticulation than a stop with the same place might be (Fowler, 1994). The locus equations for /f/ and /s/ were generated with the expectation that the locus equation coefficients would be similar to those of stops with the same place of articulation, but that the magnitude of the slope would increase with the new coarticulation allowed by frication. Sussman and Shore returned with the theory that slope decreases inversely as a function of constriction in the vocal tract, and that y-intercept increases in the opposite manner. Table 5 shows the locus equation coefficients for /f/ and /s/, compared with the values from the Sussman 1991 study. The place for a velar fricative from the current study is left blank.

	Voiceless Fricatives					
	/b/		**/d/**		**/g/**	
1991	***m***	***b***	***m***	***b***	***m***	***b***
Male	0.870	106	0.430	1073	0.660	893
Female	0.900	91	0.410	1349	0.750	777
Total	0.890	99	0.420	1211	0.710	792
Present	**/f/**		**/s/**		**N/A**	
Male	0.820	206	0.652	557	-	-
Female	0.810	238	0.720	530	-	-
Total	0.815	222	0.686	544	-	-

Table 5: Fricative Averages

As expected, the fricatives keep with the previously observed pattern, labial > alveolar for both slope and y-intercept. /f/ locus equations had slopes steeper than the /b/ locus equations from this study, but shallower than most /p/ locus equation slopes. The y-intercept values were generally higher than those for both /b/ and /p/, which was also expected, and fit with Sussman and Shore's theory (Sussman and Shore, 1996). The /f/ locus equations were very well-fitted, with data points that clustered closely around the regression lines. The /s/ locus equations, conversely, showed a significant amount of variance. The average values were as expected—steeper than the /d/ and /t/ locus equation slopes, but still shallower than the /f/ locus equation slope. The y-intercept values were much lower than those for alveolar voiced stops, but were still higher than those for the labial fricatives. The /s/ locus equations generated by Sussman and Shore had a mean slope of .57 and a mean y-intercept of 643. These values are once again shallower than the /s/ values achieved in this study, but they are similar in the /s/ locus equation slope is steeper than the /d/ locus equation slope, and the /s/ locus equation y-intercept is smaller than the /d/ locus equation y-intercept. The /s/ locus equations for individual speakers varied widely. Although a visual inspection of the equations was conducted, /s/ is the only consonant other than /g/ that had under 20 vowel tokens in the regression line, meaning the /s/ equations were particularly susceptible to outliers or imbalances in the vowel tokens.

The final manner class is the nasals, once again representing only the labial and alveolar places of articulation. There is a nasal velar, but it does not occur in English speech as often as the other consonants, and there were not enough tokens with vowels transitioning from the nasal velar for a valid locus equation. The alveolar nasal, /n/, was examined in the same Sussman study that first included the /t/ and /s/ locus equations (Sussman 1994). As with the fricatives, the nasal consonants /m/ and /n/ were expected to have locus equation coefficients similar to those of other consonants with the same place of articulation. /m/ and /n/ are somewhat unique because they do not face the issue seen with voiceless stops and fricatives—that is, the period of aspiration, frication, or silence where the tongue may be moving in transition but no formant can be seen. /m/ and /n/ are both fully voiced and sonorant, and so the F2 transitions are clearly visible the whole way through. Table 6 has the locus equation coefficients for /m/ and /n/ from this study, and the coefficients for /b, d, g/ from the Sussman et al. 1991 study. The velar place is left empty for the current study. The /m/ locus equation slope is shallower than the slopes for the /f, b, p/ locus equations, but still steeper than the slope for the /n/ locus equation. The /n/ locus equation slope is shallower than the /s/ locus equation slope. The /n/ locus equations generated by Sussman and Shore in 1996 had a mean slope of .48, with a mean y-intercept of 899. These values put the Sussman and Shore /n/ average in the same position: the slope was smaller than the /s/ locus equation slope, and the y-intercept was greater. In general, nasal consonants seem to have locus equation slopes that are shallower than fricative locus equation slopes in the same place of articulation. As slope is related to degree of coarticulation resistance, it can be surmised that nasals are more resistant to coarticulation than voiceless fricatives.

	Nasals					
	/b/		**/d/**		**/g/**	
1991	***m***	***b***	***m***	***b***	***m***	***b***
Male	0.870	106	0.430	1073	0.660	893
Female	0.900	91	0.410	1349	0.750	777
Total	0.890	99	0.420	1211	0.710	792
Present	**/m/**		**/n/**		**N/A**	
Male	0.717	316	0.560	654	-	-
Female	0.625	526	0.463	991	-	-
Total	0.671	421	0.512	823	-	-

Table 6: Nasal Averages

4.2 RECOVERY OF PLACE OF ARTICULATION

In the 1991 study of locus equations, the equation coefficients are plotted in a coordinate plane as slope/y-intercept pairs and then used to classify place of articulation (Sussman et al, 1991). Sussman reported that discriminant analyses applied to these higher-order mappings of locus equations classified place of articulation with 100% accuracy (Sussman, 1991). In later papers, attempts at classification were made using locus equations with more than one manner of articulation—for example, classifying both /s/ and /d/ as alveolar consonants based on the coefficient plot (Fowler, 1994; Sussman and Shore, 1996). It was seen that as manner of articulation changes, the coefficient clusters begin to overlap with one another, making classification more difficult. Figure 15 below shows the coefficient mapping of every locus equation generated from the NSP corpus. There are 10 locus equations for 24 speakers, equaling a total of 240 tokens. Labial consonants are marked by x's, alveolar consonants by dashes, and velar consonants by crosses. The locus equation slope is plotted along the x-axis, and the y-intercept is on the y-axis. The coefficients generally appear where expected—alveolar consonants are both higher on the y-axis and lower on the x-axis than the other two clusters, and labial and velar consonants are separated more by differences in slope than differences in y-intercept. There is also some clear overlap in the three clusters. Velar consonants and alveolar consonants in particular overlap. The velar instances that fall in the alveolar range in on the y-axis are mostly from /g/ locus equations that were too shallow. The labial consonants encroaching on the alveolar cluster are mostly from /m/ locus equations, and the lowest of the alveolar consonant points belong to /s/ locus equations.

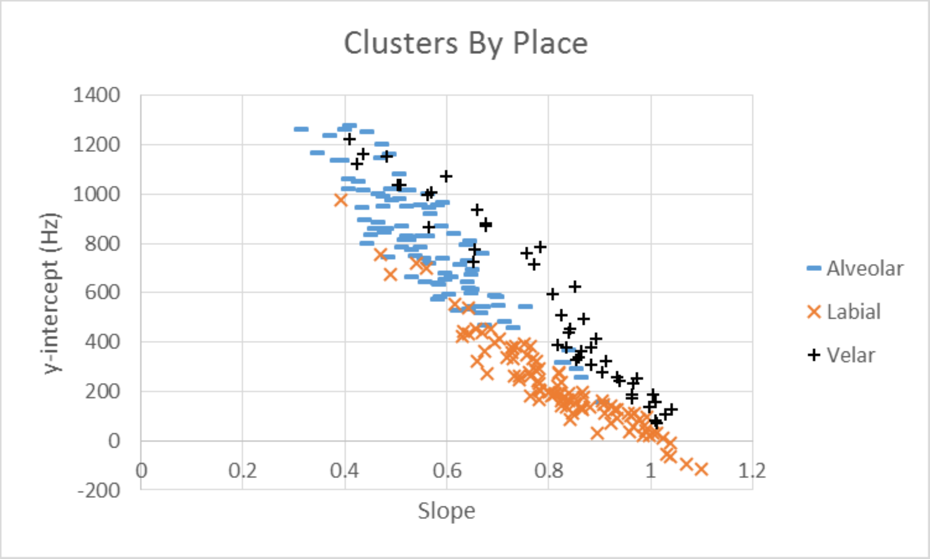


Figure 15: Dialect Clusters by Place

An obvious first step in the classification of place of articulation given such a wide variety of consonants is to include the manner of articulation as a feature in the classifier. The following describes the use of three different classification algorithms for recovery of place of articulation—a K-means clustering algorithm, an Artificial Neural Network (ANN), and a Classification Tree. The K-means algorithm was not expected to do well, both because of the overlap between clusters and because the coefficients do not group in the typical "circular" clustering pattern a centroid-based method like K-means is meant for. Additionally, this method did not take the manner as a feature for classification. This algorithm is used only as a starting point for classifier accuracy. The ANN and the Classification tree are both expected to have improved performance on classification of the data set. The K-means algorithm and the ANN were implemented both by hand using Java, and through the WEKA Machine Learning package, a free online package that takes sets of features as input and runs various machine learning algorithms on them (Hall et al., 2009).

K-means is one of the most basic clustering algorithms. The program works by randomly generating a pre-set number of "centroids," points which will serve as the center of the clusters. For this data set, the number of centroids was set to three—one for each place of articulation. Once the centroids are generated, each point in the space is evaluated and assigned to the closest node. The centroids are then moved to represent the center of this group of points. The distance metric used for this algorithm was a simple Euclidean distance. The first K-means algorithm was implemented in java and run on the set of 240 instances for 50 iterations, even though the centroids typically stopped moving after 5 to 10 iterations. The whole algorithm was run 1000 times, to check the consistency of the results. As expected, the classification results were poor, with an overall accuracy of 59%. Broken down by place of articulation, the "Labial" cluster was 71% accurate. The "Alveolar" cluster was 71 % accurate, and the "Velar" cluster was 18% accurate. The results of the algorithm are visualized in Figure 16. The color of the points represents the class predicted by the K-Means classifier. As in Figure 15, orange marks labial points, blue marks alveolar points, and black marks velar points. The shape of the marker symbol denotes the actual place that locus equation represents—x for labial consonants, - for alveolar consonants, and + for velar consonants. The performance of this K-Means algorithm was checked by running the program through the K-Means option in the WEKA machine learning package (Hall et al, 2009). The WEKA package K-Means clustering algorithm was 55.5% accurate, making it a little less accurate than the java implemented K-Means. The WEKA clusters are visualized in Figure 17. This figure is taken from WEKA, so it differs slightly from Figure 16. The color scheme is the same—orange marks labial, blue marks alveolar, and black marks velar. All correctly classified instances are marked by a cross (+) while incorrect instances are marked by a box. The slope is plotted along the x-axis, and the y-intercept in Hz is along the y-axis.

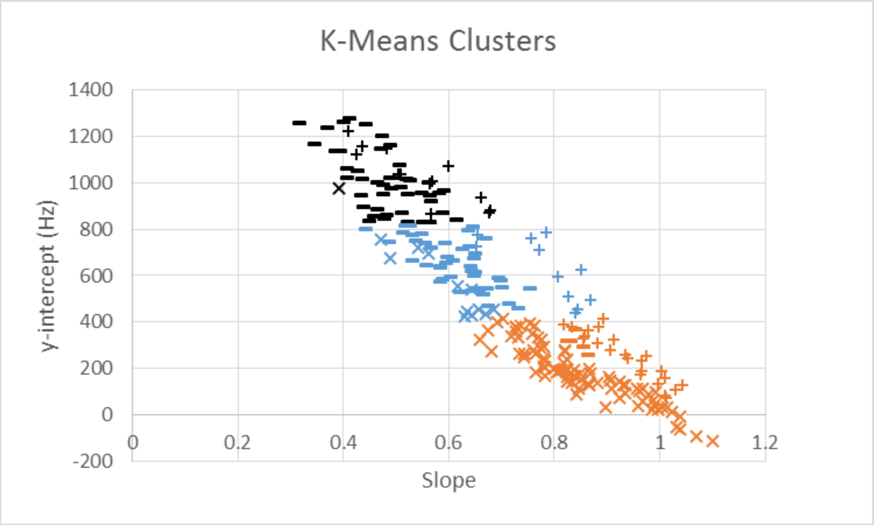


Figure 16: Dialect Java K-Means

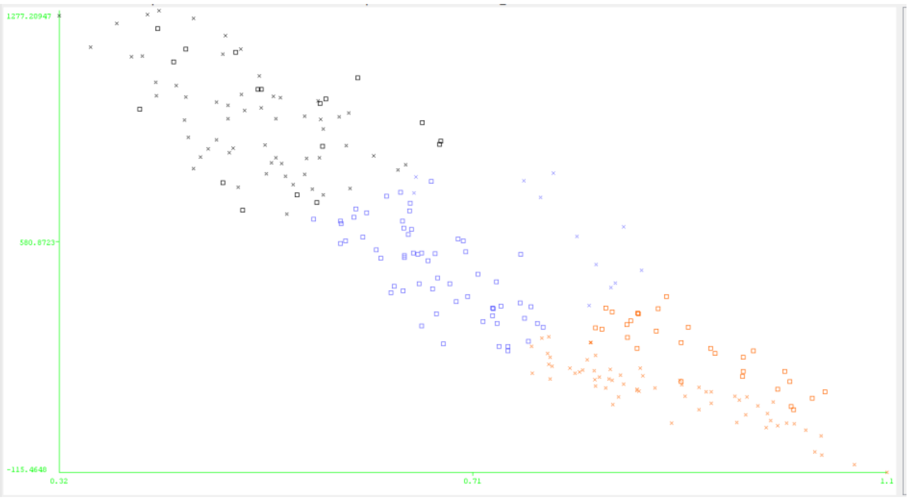


Figure 17: Dialect WEKA K-Means

The second classification technique applied to this data set was an Artificial Neural Network (ANN). The network was implemented in Java. The data set is relatively small, and so a 12 fold cross-validation technique was used to avoid overfitting the network. The data was subdivided into 12 sets, each containing two instances of each of the ten consonants. The ANN was then trained on 11 of the 12 sets, and tested on the twelfth set for accuracy. Each set was set aside and used as a testing set once. The network itself took three values as input—slope, y-intercept, and manner—and outputted one of three values as a classification of place. The hidden layer had five nodes. The learning rate was set at 0.85, and the momentum was 0.095. The network trained for a total of 500 epochs on each training set, and accuracy of classification on the testing set was measured as percentage of instances properly classified. The entire process was repeated 1000 times to check for consistency of results. The average accuracy of the ANN classifier was 84.13%, although the ANN did reach an accuracy of 88% at its best. The classifier error is plotted in Figure 18. Note that no labial instances were incorrectly classified as velar instances, or vice versa. The majority of the error is from the velar voiced stop /g/ instances misclassified as alveolar. There were also alveolar fricative /s/ equations misclassified as velar, and labial nasal /m/ equations misclassified as alveolar.

The data set was also trained and classified using the WEKA "Multilayer Perceptron" function, which is equivalent to an ANN. The WEKA network took six input values—the numeric x and y values for slope and y-intercept, and one binary input for each manner of articulation (Voiced Stop, Voiceless Stop, Fricative, Nasal). The hidden layer contained 4 nodes, and classification came from three output nodes—one for each place of articulation. The learning rate was 0.3, the momentum was 0.2, and the model was trained for 250 epochs. This model had a classification accuracy rate of 89.16%, meaning it correctly classified 214 of the 240 instances. This result was 5% more accurate than the Java network. The results could be taken as support of multiple output nodes in classification problems over one output node that returns different values. A visualization of the classification error can be seen in Figure 19 below. The error is plotted the same as in Figure 17.

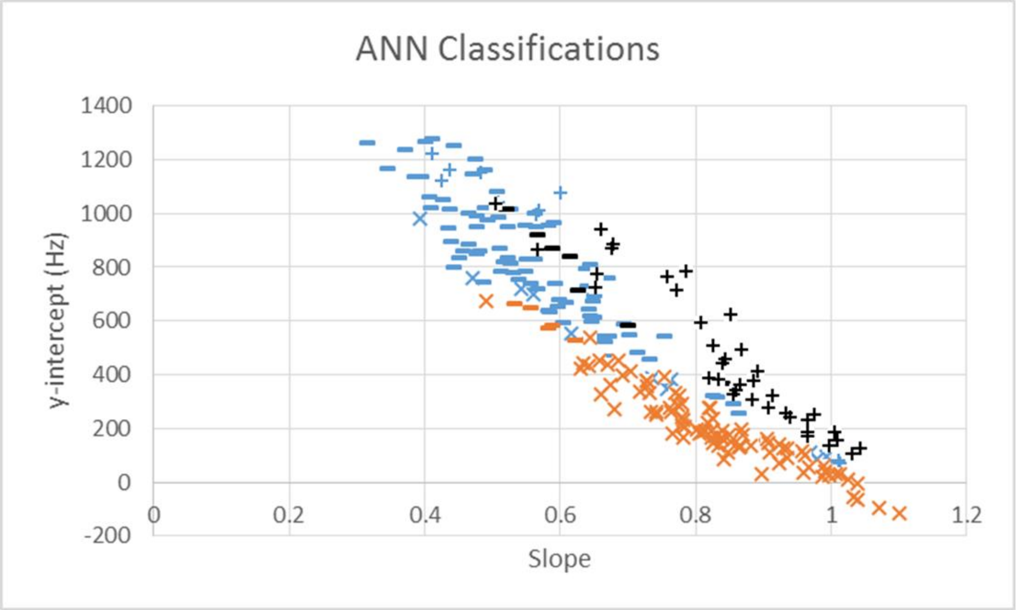


Figure 18: Dialect Java ANN

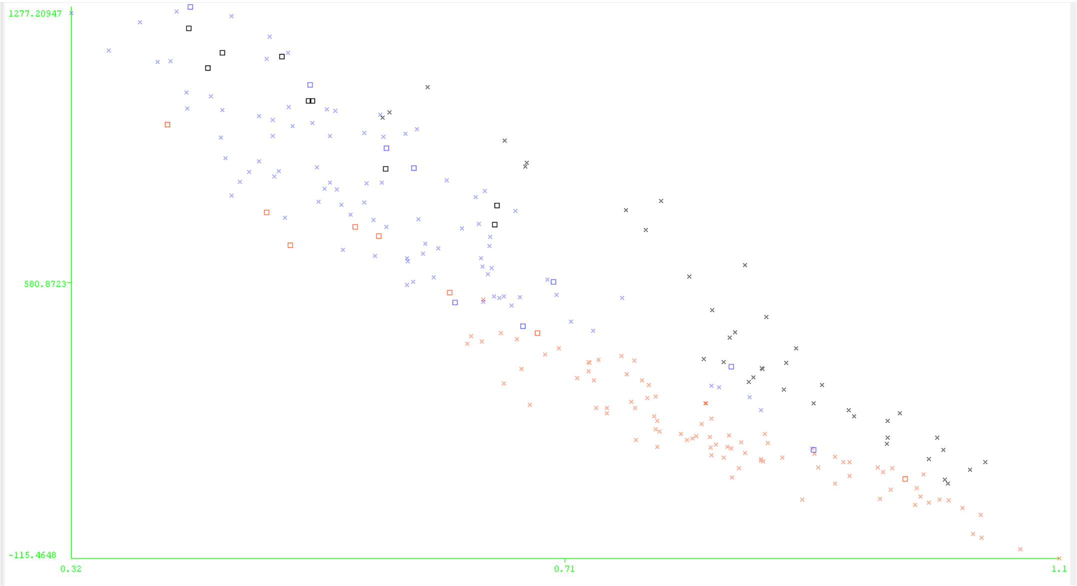


Figure 19: Dialect WEKA ANN

The final classification method used on this data set was a Classification and Regression Tree (CART). CART trees are sets of rules that divide data sets based on certain attributes, like manner of articulation or slope values, and eventually assign classification labels to the subsets. Decision tree algorithms use measures of entropy and information gain to decide which attribute to divide the dataset on at each point in the process. The attribute which lowers the resulting entropy of each sorted set the most is chosen as the next step in the process. The tree discussed here was generated using WEKA. The CART algorithm implemented by WEKA is the J48 algorithm, which uses reduced-error pruning and rule-ordered pruning to avoid overfitting the data. The tree was created and testing using 10-fold cross validation. The J48 algorithm consistently created a tree that classified the 240 instances with 87.5% accuracy, meaning 210 instances were correctly classified and 30 instances were not. The generated decision tree can be seen in Figure 20 below. Nodes labelled "Feature" split the data set based on that feature. The bolded values seen along the pathways show which feature values were sorted where. The nodes labelled "Class" are child nodes—all locus equation instances sorted into that child node are assigned the defined class. The numbers in parentheses show the accuracy of the set sorted to that node, (Correctly Labelled/Incorrectly Labelled).

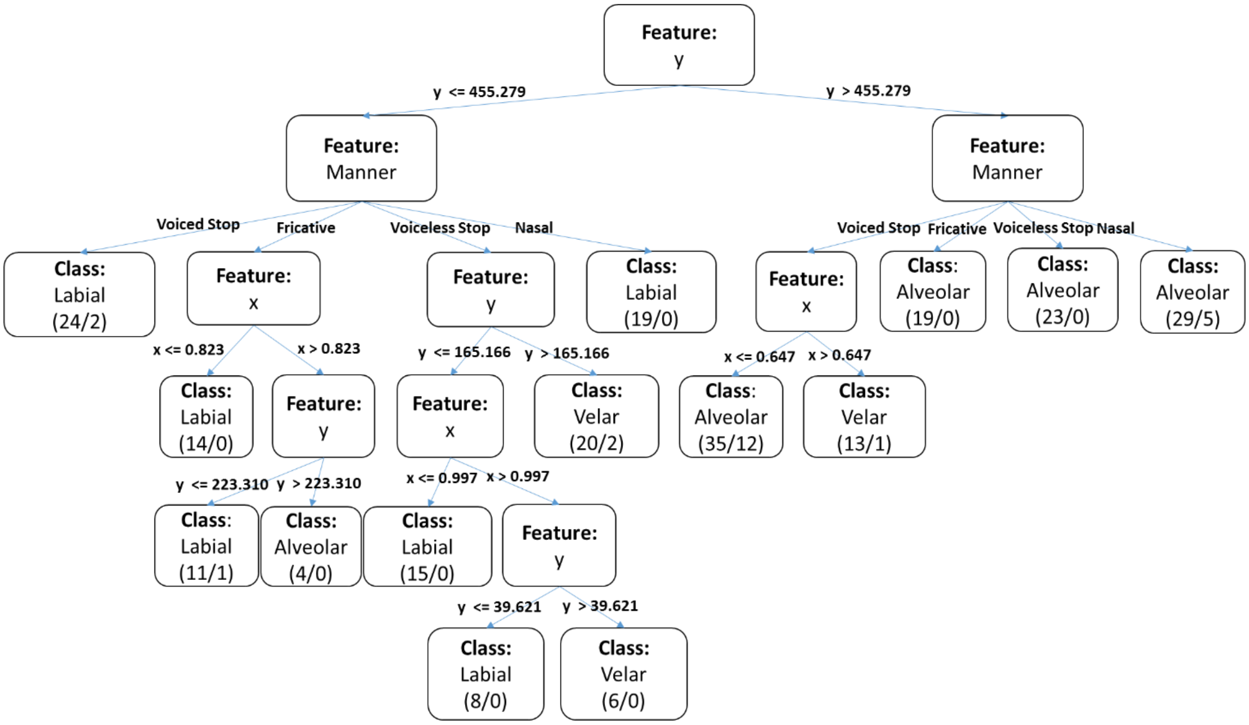


Figure 20: Dialect Decision Tree

The rules defined in a decision tree are good descriptors of the layout of the data set. Note, for instance, that the very first split made by the tree is between locus equations with y-intercepts above 455 Hz, and those with y-intercept values below 455 Hz. The equations with high y-intercept values are nearly all classified as alveolar. The exception is the subset of voiced stops, which is further divided by slope (greater or less than 0.647). Equations with higher slopes are classified as velar, and those with lower slopes are classified as alveolar. Having looked at the data set, we know that these rules are an attempt to handle the /g/ locus equations that are unusually flat and have higher y-intercept values that mix them with the /d/ locus equations. Looking at the number of instances correctly classified versus the number incorrectly classified also provides valuable information. The locus equations with y-intercepts over 455 Hz that fall into the Fricative and Voiceless Stop classes are classified as alveolar with 100%

accuracy. The Nasal and Voiced Stop instances are less accurate, with 5 Nasal and 12 Voiced Stop locus equations incorrectly classified as alveolar. This is a result of the flat /g/ locus equation slopes muddying the alveolar voiced stop cluster, and the few /m/ locus equation slopes that were flatter than expected mixing with the /n/ locus equation slopes. The Decision Tree classification errors are visualized in Figure 21.

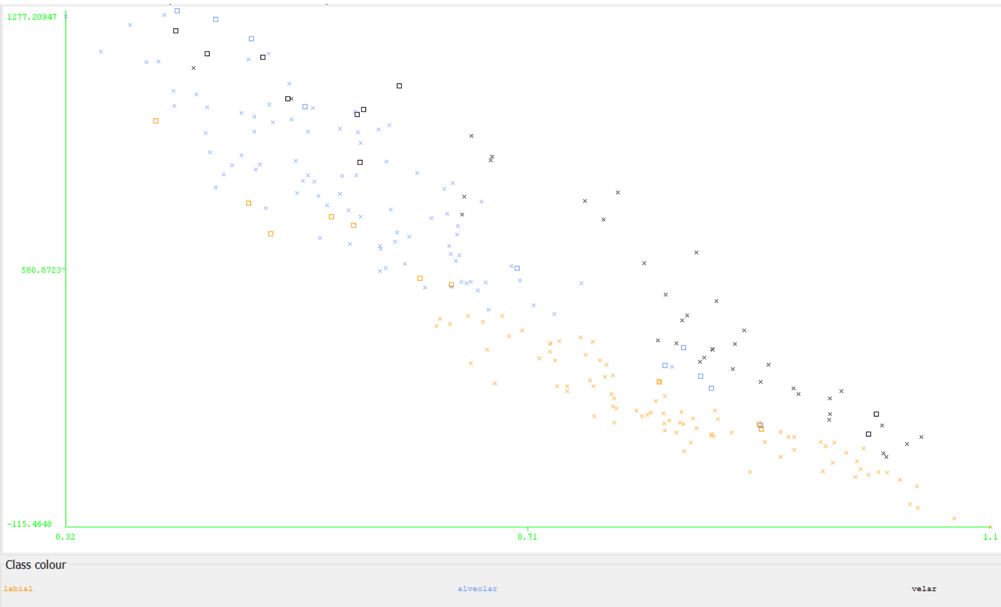


Figure 21: Dialect WEKA J48

The data set used above contains locus equations for consonants with multiple different manners of articulation—voiced stops, voiceless stops, fricatives, and nasals are all represented. This increases the difficulty of classification on the set, as it causes overlap in the clusters. Fowler notes in her 1994 paper that a labial stop might have the same slope as an alveolar fricative, not because they share a place of articulation, but because of the difference in manner (Fowler, 1994). The addition of the "manner" feature was meant to counteract this obstacle in the ANN and the Decision Tree described above. As an exploratory measure, the three WEKA classifier systems—K-Means clustering, Multilayer Perceptron, and Decision Tree—were run on subsets of the data divided by manner class. The results can be seen in Table 7 below. K-Means clustering results improved significantly on data sets with only two clusters, like Fricatives and Nasals. Both data sets with three places of articulation represented—Voiced and Voiceless Stops—had a higher accuracy using the multilayer perceptron than using the decision tree. Voiceless Stops had the highest overall classification accuracy rate—94.44% accurate when the multilayer perceptron was used. Voiced stops had the lowest accuracy percentages for both the multilayer perceptron and the decision tree, further confirming that the voiced stops were the hardest manner class to correctly classify. This is most likely caused by the irregular /g/ locus equations.

	Classification Method		
Manner	**K-Means**	**Multilayer Perceptron**	**Decision Tree**
Voiced Stops	59.72	80.56	72.22
Voiceless Stops	52.78	94.44	87.5
Fricatives	87.5	91.66	93.75
Nasals	75	85.41	87.5

Table 7: Dialect Classification

4.3 CLASSIFICATION OF DIALECTS

Having established that the automatically generated locus equations are valid, they can now be examined as features in a classifier system. These equations would only serve as useful features for classification if the dialects all had unique coarticulation patterns. The NSP data set contains six dialect classes—Mid-Atlantic (at), Midland (mi), New England (ne), North (no), South (so), and West (we). Locus equations were generated for four speakers from each dialect, two male and two female. The end result was 24 speaker instances available for classification. Two different types of feature sets were generated and tested. The first was the entire set of locus equation coefficients for each speaker. For example, speaker at0 would have the feature set {/b/ slope, /b/ y-intercept, /d/ slope, /d/ y-intercept…/s/ slope, /s/ y-intercept}. This feature set included all of the raw locus equation data. The second set was an abstraction to the distances between the various slope/y-intercept points of a speakers locus equations. These higher order mappings of the locus equation coefficients in dialect space serve as a method of representing the "consonant space" of a speaker. Figure 22 shows the locus equation mappings for every consonant of speaker at0. This approach has been used before, in Sussman et al.'s study of children with developmental apraxia of speech (Sussman, 2000). In that study, the locus equation coefficients for voiced stops /b, d, g/ were mapped into a coordinate plane, and the total distance between the three points was used as a method for analyzing the person's speech. Sussman found that the voiced stop points were collapsed in towards each other in speakers with apraxia, implying that the consonants were not coarticulated in a way that maximally distinguished them from one another. The purpose of the classification experiments described here was to determine if dialects have differences in coarticulation that are reflected in the consonant space, and, if so, if those differences could be used for classification.

For this study, the features used were:

1. The sum of distances between the voiced stops.
2. The sum of distances between the voiceless stops.
3. The distance between the two fricatives.
4. The distance between the two nasals.
5. The centroid point of the voiced stops.
6. The centroid point of the voiceless stops.
7. The centroid point of the fricatives.

8. The centroid point of the nasals.
9. The sum of distances between the labial, alveolar, and velar centroids.

Variations on the feature set included the distances between individual points and the centroid for that manner class, and distances as pairs rather than sums (b-d, d-g, g-b, instead of b-d-g). These features were chosen with the intent of fully capturing the consonant space of the speaker.

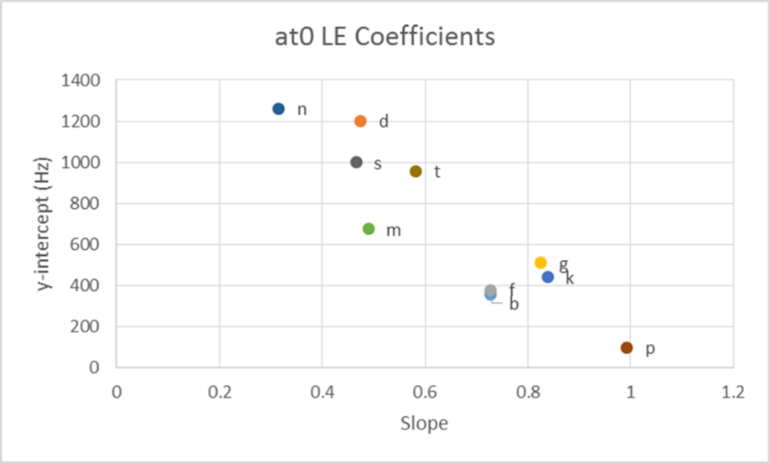


Figure 22: at0 Consonant Space

All five of the classification algorithms described in section 4.2—that is, two implementations of K-Means clustering, two implementations of an ANN, and a Decision Tree—were applied to both feature sets. The results were poor across the board. Both the set of all locus equation coefficients and the set of distances resulted in classification accuracies ranging from 10% accurate all the way down to 0% accurate. None of the classification algorithms had a better accuracy than random guessing. Figures 23-27 below are used to examine why the results might have been so poor. Each figure shows the plots of Locus Equation coefficients that have been averaged across a dialect. For instance, the three points in Figure 24 belonging to the "at" dialect are actually representations of the average slope and y-intercept of that consonant's locus equation for every speaker with that dialect. For an individual speaker, feature number 1 from above, “Sum of distances across voiced stops”, would have been the total distance between each of the “at” points plotted on the graph.

The circles surrounding the clusters of points on each graph group the consonant the coefficients represent for each dialect. These clusters further support the classifications described in section 4.2—for each manner class, the consonants with different places of articulation each clearly inhabit their own cluster. It is clear from the figures that, despite the differences heard in dialect, the coarticulation patterns are similar and so the locus equation coefficients for each consonant tend to cluster into the same space on a coordinate plane. This is useful for recovery of place of articulation, but it also explains why the classification algorithms performed so poorly on the data set. There is no systematic difference in the position of locus equation coefficients for the different dialects, and so the classifier has no reliable value on which it can divide the data set.

It is worth mentioning that the dialects of the NSP corpus, while present in the speech, are generally mild in their intensity. For example, of the speakers examined from the “South” dialect set, only speaker so0 had an immediately obvious accent in her speech. Speaker so0 also had consistently lower labial slopes than the other speakers, especially for /b/ and /m/. Her influence may be what pulled the “so” average to the highest position in the labial cluster for Voiced Stops, Voiceless Stops, and Nasals. One speaker is not sufficient evidence that dialectal differences may be reflected in a consonant space, but a further investigation into the effect of more extreme speech differences, like foreign accents, would provide interesting insight into the matter. Based on the data present, it can be concluded that locus equation coefficient mappings into a consonant space are resilient to at least mild differences in speaker dialect.

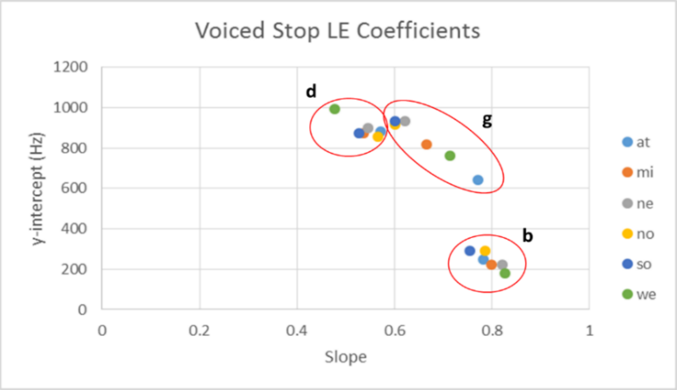


Figure 23: Voiced Stop Consonant Space

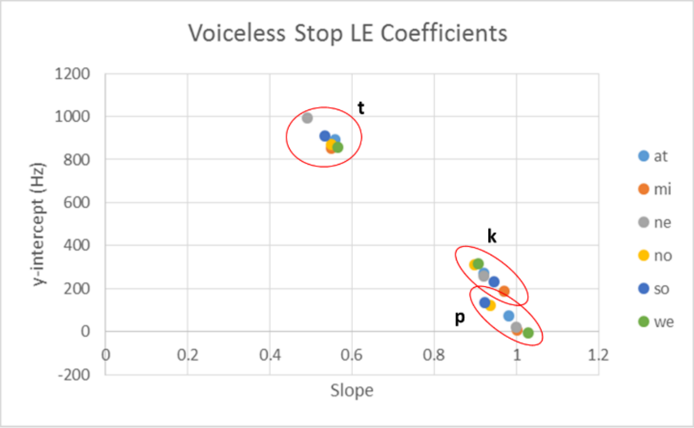


Figure 24: Voiceless Stop Consonant Space

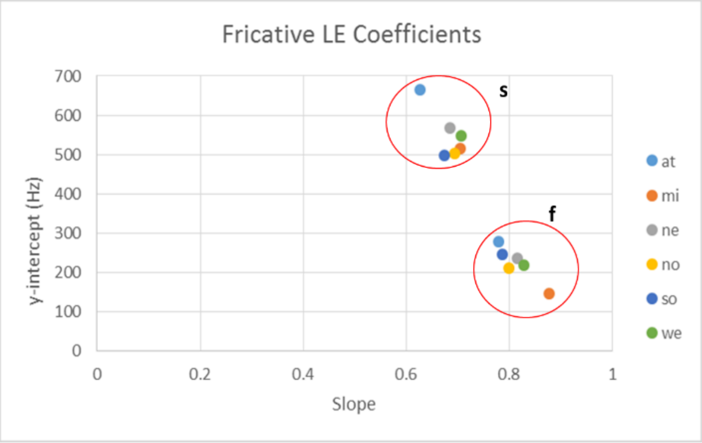


Figure 25: Fricative Consonant Space

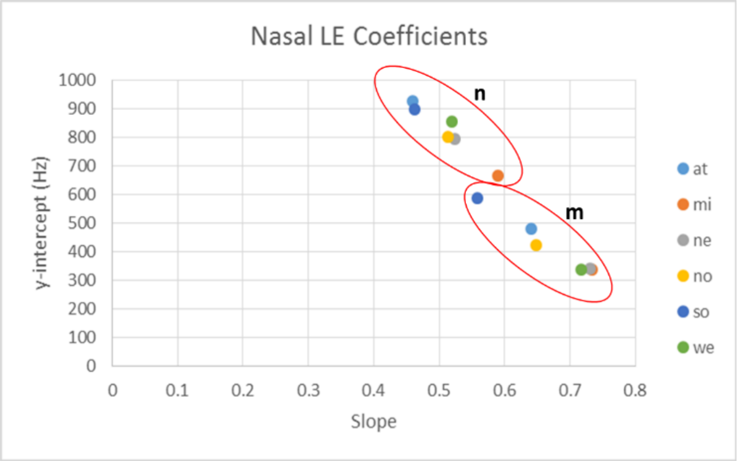


Figure 26: Nasal Consonant Space

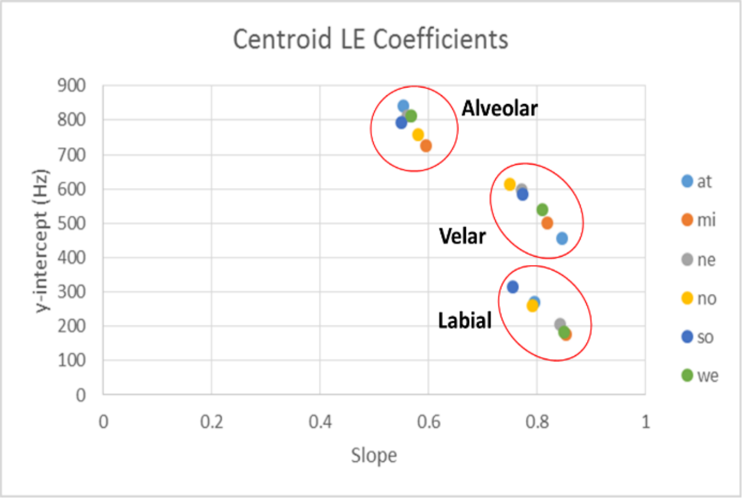
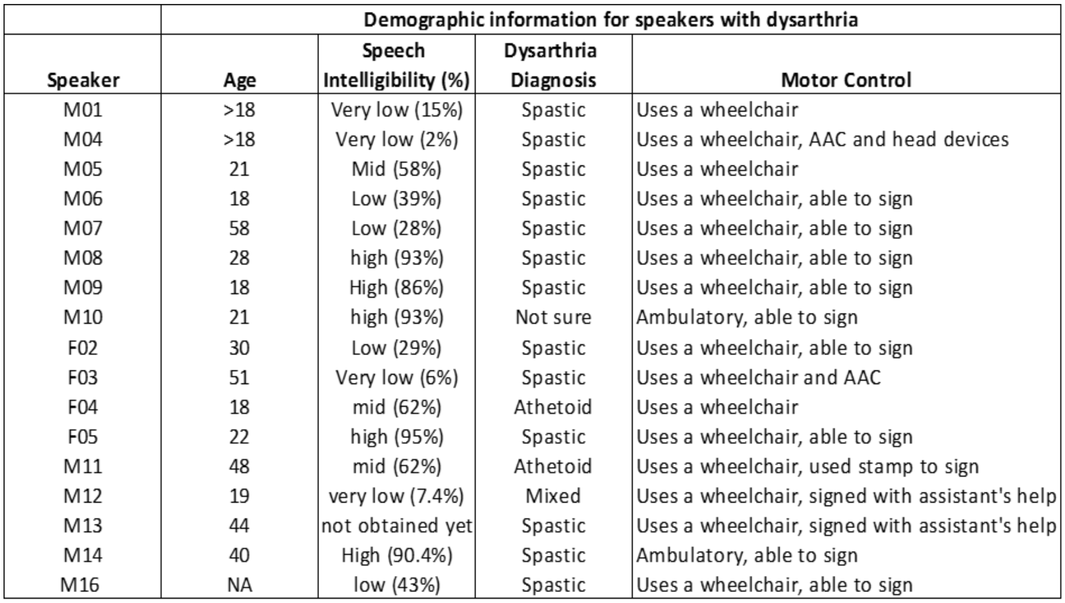


Figure 27: Centroid Consonant Space

CHAPTER 5

LOCUS EQUATIONS AND SPEECH DYSARTHRIA

The third experiment used the University of Illinois UA (Universal Access) Speech database (Kim et al., 2008). This collection of files was from speakers with dysarthria, a motor speech disorder. Dysarthria is a term for speech where the speaker has trouble with movement of the articulators. (“Dysarthria”, ASHA) It can be caused by a number of things, including stroke, tumors, Parkinson’s disease, Lou Gehrig’s disease, and multiple sclerosis. The type of dysarthria is determined by the cause, and by the extent of damage to the nervous system. A few characteristic signs of dysarthria are slurred or choppy speech, a slow speech rate, limited tongue, lip, and jaw movement, or changes in voice quality. The UA speech data base included recordings from seventeen different speakers. Of these, thirteen speakers were evaluated in this study. M04 was cut from the speakers due to his extremely low rate of intelligibility—2%. This was low enough that there was no discernable way to place alignment boundaries, and most of the time half of the word was missing from the token. M05 and M06 were both eliminated because in the updated audio, which was the file used, they were missing either half or all of their data.



	Demographic information for speakers with dysarthria			
		Speech	Dysarthria	
Speaker	Age	Intelligibility (%)	Diagnosis	Motor Control
M01	>18	Very low (15%)	Spastic	Uses a wheelchair
M04	>18	Very low (2%)	Spastic	Uses a wheelchair, AAC and head devices
M05	21	Mid (58%)	Spastic	Uses a wheelchair
M06	18	Low (39%)	Spastic	Uses a wheelchair, able to sign
M07	58	Low (28%)	Spastic	Uses a wheelchair, able to sign
M08	28	high (93%)	Spastic	Uses a wheelchair, able to sign
M09	18	High (86%)	Spastic	Uses a wheelchair, able to sign
M10	21	high (93%)	Not sure	Ambulatory, able to sign
F02	30	Low (29%)	Spastic	Uses a wheelchair, able to sign
F03	51	Very low (6%)	Spastic	Uses a wheelchair and AAC
F04	18	mid (62%)	Athetoid	Uses a wheelchair
F05	22	high (95%)	Spastic	Uses a wheelchair, able to sign
M11	48	mid (62%)	Athetoid	Uses a wheelchair, used stamp to sign
M12	19	very low (7.4%)	Mixed	Uses a wheelchair, signed with assistant's help
M13	44	not obtained yet	Spastic	Uses a wheelchair, signed with assistant's help
M14	40	High (90.4%)	Spastic	Ambulatory, able to sign
M16	NA	low (43%)	Spastic	Uses a wheelchair, able to sign

Table 8: UA Demographic Information

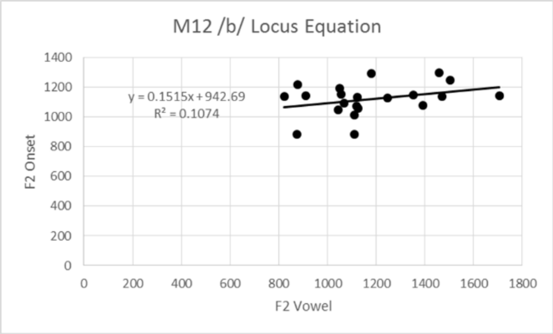
For the remaining speakers, locus equations for the three voiced stops were generated using the same procedure described in Chapter 3. 49 tokens were identified as containing the necessary vowel transitions to create well-formed locus equations, and these were then hand aligned. Only voiced stops were evaluated because the speech with dysarthria was often mumbled or unclear, making boundary points for sounds difficult to accurately pinpoint. Voiced stop boundaries are the easiest to locate, and so for an initial study only these were evaluated. After all of the tokens had been aligned, the Java and Praat scripts were run to automatically generate the locus equations. The only difference was the lack of outlier detection. The speech was expected to contain outliers caused by the dysarthria, and so when all of the equations were generated, the equations were hand checked for outliers instead. All of the token words can be found in Appendix D, and speaker F02's locus equations are included in Appendix E for reference.

5.1 COMPARISON WITH PREVIOUS STUDIES

Table 9 below gives the slope and y-intercept values for the three stop consonant locus equations generated from the corpus. Immediately obvious is the large amount of variation in the slopes. The *m* coefficient for the /b/ locus equations varies from 0.151 as a minimum to 1.029 as a maximum, taken from speakers M12 and F03, respectively. These equations can be seen plotted in Figures 28 and 29. Both speakers have a very low rate of intelligibility, and yet the equation slopes are on opposite ends of the spectrum, one being much lower than the expected value and the other being much higher. Although the /b/ locus equations for speaker F03 are better fitted to the line, all of the data points for both equations have been hand checked for validity, and so both equations are kept as data points. A similar variance in slope can be seen in /d/ locus equations, although the differences are less extreme. The lowest slope for a /d/ locus equation is 0.161, and the highest slope is 0.811. The /g/ locus equations have the most stable slope values, with the all but two speakers producing slopes between 0.8 and 1.2.

	Consonants					
	Voiced Stops					
	/b/		**/d/**		**/g/**	
Speaker	***m***	***b***	***m***	***b***	***m***	***b***
F02	0.696	288.4	0.379	1311	0.859	360.3
F03	1.029	21.89	0.476	919.7	1.038	-8.38
F04	0.671	331.1	0.603	909.5	0.809	418.3
F05	0.689	357.4	0.396	1280	0.992	207.2
M01	0.764	278.4	0.811	424.6	0.932	132.7
M07	0.801	183	0.161	1393	0.692	538.4
M08	0.68	349.2	0.283	1419	0.808	458.3
M09	0.826	196.3	0.297	1111	0.813	249.8
M10	0.651	400.9	0.449	1094	0.516	1018
M11	0.657	317.9	0.212	1419	0.754	404.4
M12	0.151	942.7	0.313	935.2	1.119	-162
M14	0.605	476	0.731	648.2	0.882	412
M16	0.692	378.2	0.329	1125	1.026	101.7
Mean	0.685	348	0.418	1076	0.865	318

Table 9: Dysarthria Voiced Stops



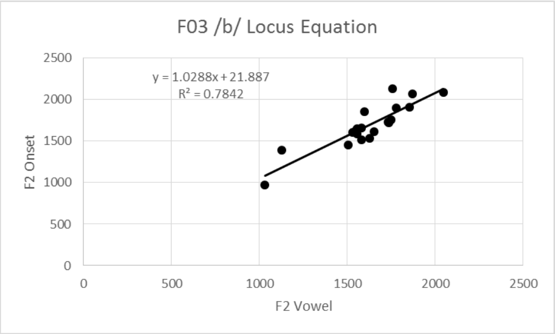
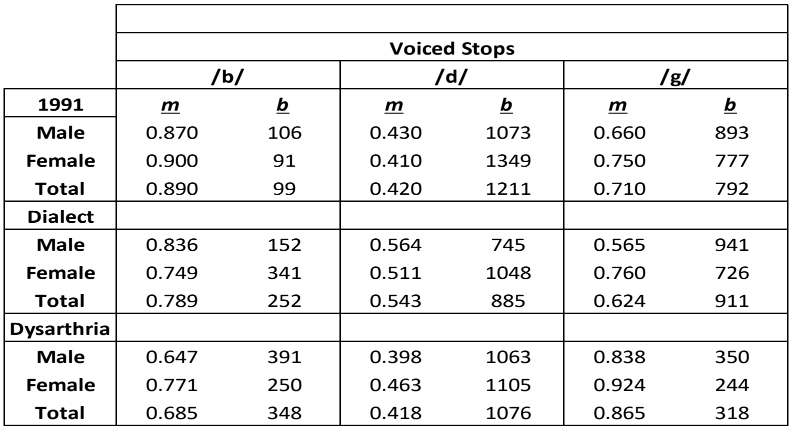


Figure 28: M12 /b/ Locus Equation

Figure 29: F03 /b/ Locus Equation

Having noted the variability of the slope values in the locus equations, the following comparisons of average values should be considered with caution as they are likely to overlook the individual variation. The mean values will be used as a starting point for comparison of locus equations across studies, and then a more in-depth consideration of the individual slopes will be made for each place of articulation. In past experiments, the slope values for locus equations have been found to decrease in the order labial is greater than velar is greater than alveolar, while y-intercepts increase inversely (Fowler 1994; Lindblom, 1963; Sussman, 1991). The averages in Table 9, however, flip the velar and labial consonants, so the velar > labial > alveolar. As seen in Table 10, the average slope for the /b/ locus equation is shallower for speakers with dysarthria than it was for the dialect speakers from this study and the speakers from Sussman et al, 1991. In fact, the average /b/ locus equation slope values for speech with dysarthria are most similar to the average /g/ locus equation slope values from Sussman et al. (1991). The average /d/ locus equation values for the speakers with dysarthria closely match the /d/ locus equation values from Sussman et al, 1991. Keeping with the velar/labial flip, the /g/ locus equation coefficients from speakers with dysarthria most closely match the /b/ locus equation values from the 1991 study. This implies that, for many speakers with dysarthria, /g/ is more heavily coarticulated than /b/. A stronger implication would be that /g/ is often articulated with the wrong place or manner entirely, in a way that encourages coarticulation.



	Voiced Stops					
	/b/		/d/		/g/	
1991	*m*	*b*	*m*	*b*	*m*	*b*
Male	0.870	106	0.430	1073	0.660	893
Female	0.900	91	0.410	1349	0.750	777
Total	0.890	99	0.420	1211	0.710	792
Dialect						
Male	0.836	152	0.564	745	0.565	941
Female	0.749	341	0.511	1048	0.760	726
Total	0.789	252	0.543	885	0.624	911
Dysarthria						
Male	0.647	391	0.398	1063	0.838	350
Female	0.771	250	0.463	1105	0.924	244
Total	0.685	348	0.418	1076	0.865	318

Table 10: Dysarthria Locus Equation Averages

The variances in the individual locus equations from Table 9 are best explained by looking back at the nature of speech dysarthria and the theory of locus equations. As mentioned above, dysarthria is a type of motor speech disorder that occurs when the movement of muscles required for speech production is impaired. This leads to abnormal patterns in the person’s speech. Locus equations are tools for the examination of consonant production in speech. The locus equation slope is directly related to degree of coarticulation—a consonant that is coarticulated heavily will have a locus equation with a larger slope, while a consonant with minimal coarticulation will have a locus equation slope approaching 0. With this in mind, the locus equation coefficients from Table 9 can be reconsidered.

The locus equation slopes for each speaker vary widely across all three stop consonants. Every speaker has at least one locus equation slope value that falls significantly outside of the expected range. There are a few patterns that appear in the coefficients. Most speakers have /b/ locus equation slopes that are smaller than the expected range, indicating a lesser degree of coarticulation than expected. This could be caused by the reduced mobility of tongue and lips that sometimes accompanies speech dysarthria—an inability to move the tongue quickly would keep the vocal tract from shifting towards a vowel shape while the /b/ is still being formed. /g/ locus equation slopes are mostly much steeper than expected. This may be because /g/ is the hardest of the three consonants to form, and so many speakers do not fully move their tongue into position for a "g" before moving onwards to the next vowel (Ferguson and Farwell, 1975). This incomplete transition would increase coarticulation. The consonants that each individual speaker struggles with, and the way that they struggle with them, is heavily dependent on the underlying cause for the speech dysarthria, and the symptoms that affect that particular person. More data could potentially help to predict which stops a speaker with speech dysarthria will fail to coarticulate and which stops a speaker will over coarticulate. Given the current corpus, the only consistent pattern is that each speaker with dysarthria does have at least one stop consonant with a locus equation slope that stands out as abnormal.

5.2 RECOVERY OF PLACE OF ARTICULATION

The previous section discussed the validity of the locus equations generated from the speakers with dysarthria. It has been established that reasonably well-fitted lines can be generated for each of the three voiced stops, and that the slope coefficients of the locus equations are different from the expected values, which in turn throws the y-intercept values off by a margin. In this section, the same classification algorithms used for recovery of Place of Articulation on the NSP dialect data are used on the UA locus equations. Since the algorithms were discussed in detail previously, this section is limited to classification results. For more detail on the algorithms, see section 4.2. Results are expected to be lower than they were in chapter 4, both due to the variance in slope and y-intercept values, and the smaller number of examples for the set to train on. Figure 30 shows a representation of the clusters when all tokens are properly classified. Note that the alveolar consonants remained mostly separate, but the labial and velar clusters overlap with one another, especially as slope grows.

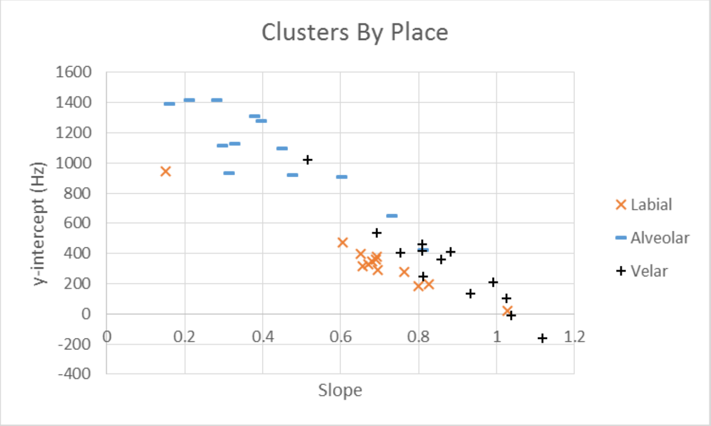


Figure 30: Dysarthria Clusters by Place

The first classification algorithm used was the simple K-Means clustering algorithm. As with the dialect data locus equations, this method was used only as a baseline and was not expected to do well. The algorithm was run with a preset number of three clusters—one for each place of articulation. Both the hand-written Java algorithm and the WEKA K-Means classifier returned a classification accuracy of about 70%, which was achieved mostly by minimizing the “velar” cluster to only six instances, and having particularly low accuracy for that place of articulation. Figure 31 shows the WEKA K-Means clustering output.

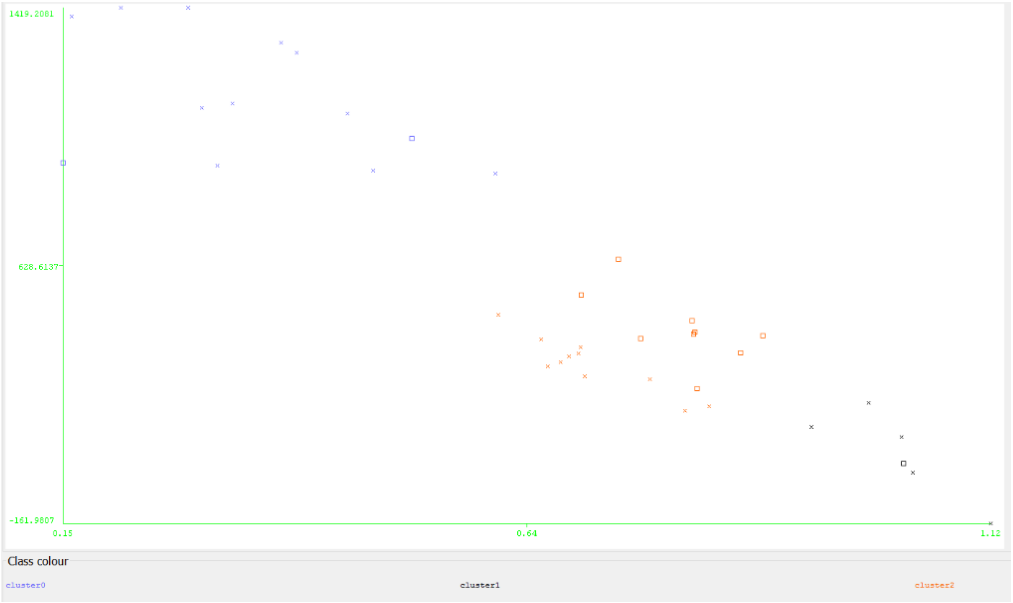


Figure 31: Dysarthria WEKA K-Means

The next approach to recovery of place of articulation was the Artificial Neural Network. Both the hand-implemented Java ANN and the WEKA multilayer perceptron algorithm were run on the data set. Both methods were run using 12-fold cross validation in an attempt to prevent overfitting of the data. The networks were run using the same settings used for classification of the dialect data in section 4.2. For the network implemented in Java, the ANN was trained for 500 epochs on the data set, and the classification accuracy was measured as percentage of instances properly classified. The hidden layer had five nodes. The learning rate was set at 0.85, and the momentum was 0.095. With these settings, the average accuracy of the Java classifier was 77.77%. The classifier error is plotted in Figure 32. The WEKA model was run for 250 epochs using the default settings—learning rate was 0.3, momentum was 0.2, number of hidden nodes was 4, and number output nodes was 3. This method had a classification accuracy of 79.48%, meaning 31 of the 39 instances were properly classified. The WEKA classification output is plotted in Figure 33.

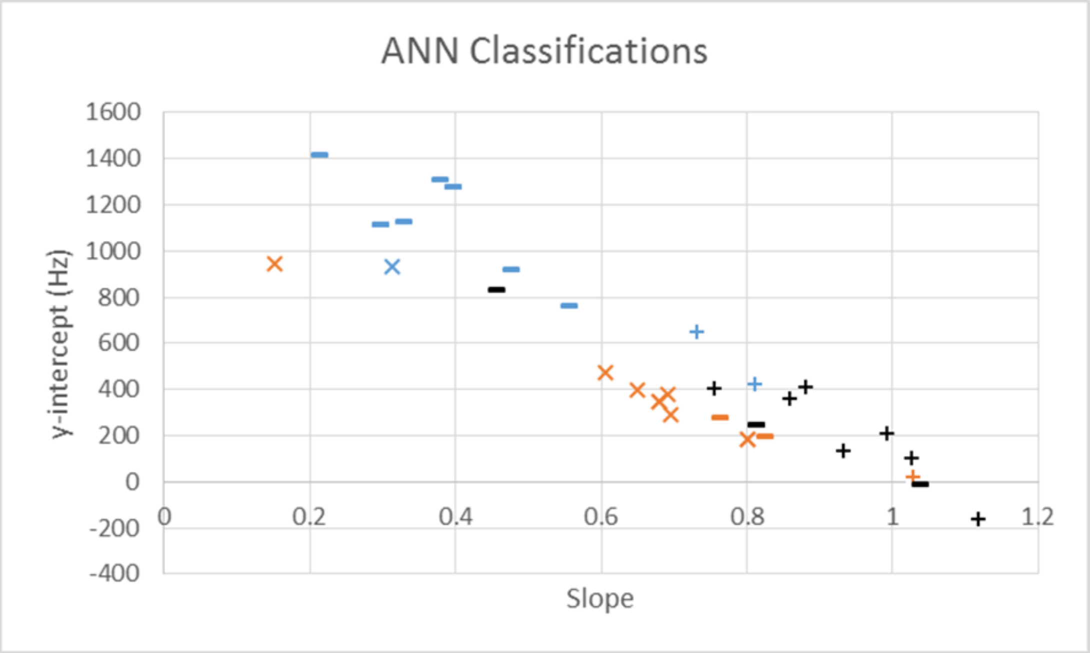


Figure 32: Dysarthria Java ANN

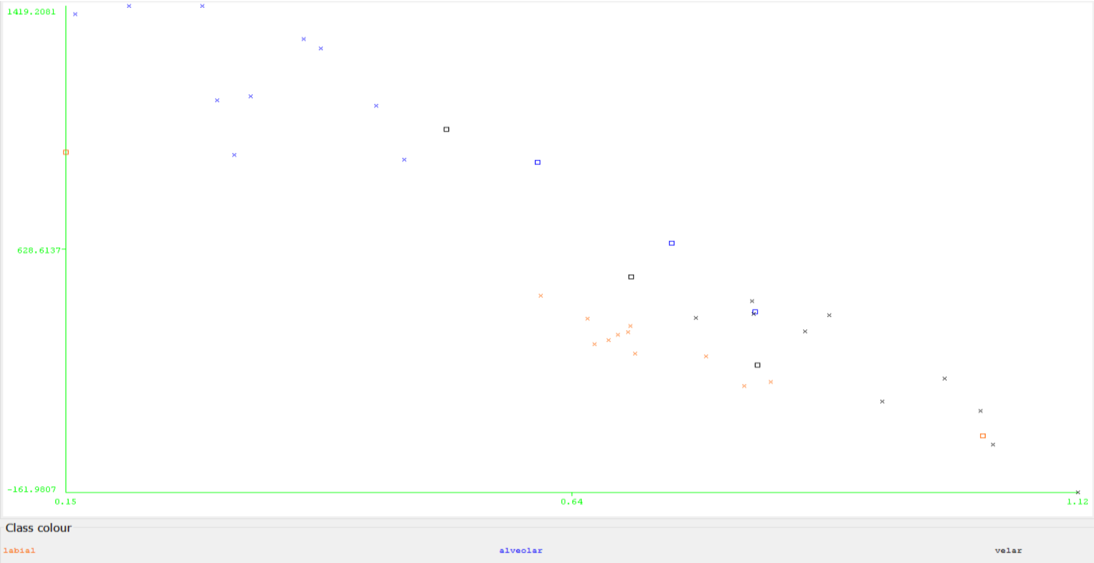


Figure 33: Dysarthria WEKA ANN

The final classification algorithm examined was WEKA's built-in J48 decision tree, which was created and tested using 10-fold cross validation. The tree generated for the UA data set is significantly smaller than the one generated for the dialect data set in section 4.2. The reason for this is the lack of a manner feature—only locus equations for voiced stops were generated from the UA dataset, so the inclusion of a manner feature would be redundant. The WEKA decision tree had classification accuracy of 71%, meaning 28 of the 39 instances were classified correctly by the tree. Figure 34 shows the rules generated by the J48 algorithm as a tree. The rules of the tree are dedicated mostly to separating the velar /g/ from the labial /b/. This is a reflection of the overlapping y-intercept values of the two voiced stops, caused by the unusually high /g/ locus equation slope average and the unusually low /b/ locus equation slope average. Overall, the classification algorithms did reasonably well considering the overlap of tokens in the data set. As with the dialect locus equations, WEKA ANN had the highest performance accuracy, followed by the Java ANN and then by the WEKA decision tree. The results can be interpreted as a sign that locus equations are not entirely robust to extreme variance in speech. We know the positions of the consonants are changing enough to overlap with one another, which provides support in favor of their use as a feature in classification of different types of speech.

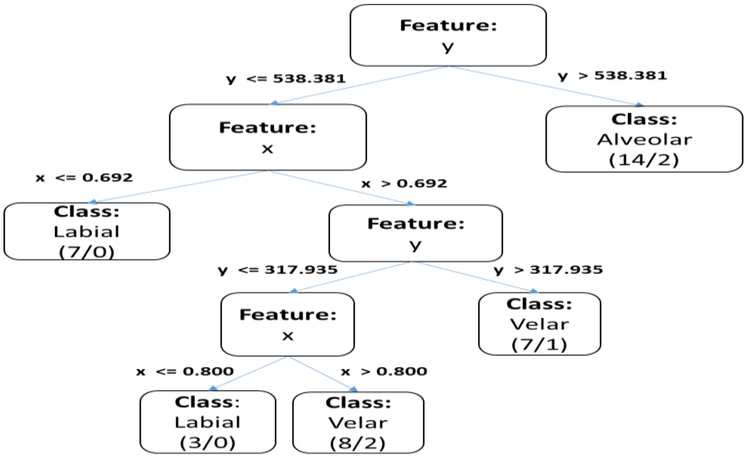


Figure 34: Dysarthria WEKA Decision Tree

5.3 CLASSIFICATION OF SPEAKERS WITH DYSARTHRIA

This section discusses the potential application of locus equations as features for classification speech dysarthria. Section 4.3 in the previous chapter examined locus equations as features for classification of dialects, and found that the equations were generally robust to the smaller dialectal differences in speech, making them ill-suited as features in that area. This is not the case for the locus equations generated from the UA corpus, representing speakers with dysarthria. Section 5.1 discusses the variance in the locus equation slopes for each of the voiced stops, noting that the speakers with dysarthria tend to have slopes that vary from the expected value for a particular place and manner of articulation. Given this trend, it seems reasonable to hypothesize that a classifier should be able to distinguish between speakers with dysarthria and speakers without dysarthria given the locus equation coefficients for all three voiced stops. This hypothesis is tested using a combination of the two data sets. The /b/, /d/, and /g/ locus equation coefficients for every speaker from the UA corpus and every speaker from the NSP corpus were put together as one data set. The 24 speakers from the NSP corpus were classified as being “Negative” for dysarthria, and the 13 speakers from the UA corpus were labelled as “Positive”. Following this, two classification algorithms—an ANN and a J48 decision tree—were run on the data set.

The features described above were chosen mostly as a reaction to the specific type of disordered speech contained in the UA corpus. Speech dysarthria is a motor speech disorder, and there are a number of different causes and symptoms that can appear. The generality of the disorder means that individual speakers can all have speech dysarthria, but each struggle with different sounds in different ways. The variety here means that it would be difficult to classify more specific aspects of the data—like level of intelligibility—without more information about the speaker and the particulars of their speech. Choosing to classify in a broader sense, just "Positive" or "Negative" for speech dysarthria, ensures that the classifier has the information necessary to find patterns within the data. This is also why the "raw" locus equation coefficients were used as input to the classifiers, rather than some feature like distance in the consonant space or position of centroid values. The speakers examined are not guaranteed to show any stable pattern beyond "variance from the norm in speech," and so such abstractions were more likely to erase relevant data than they were to increase classifier accuracy.

The WEKA multilayer perceptron algorithm has consistently outperformed the Java ANN algorithm in classification accuracy, and so it was chosen as the first tool for classifying speech as positive or negative for dysarthria. The network had six input nodes, these being the slope and y-intercept values for each voiced stop. There were four hidden nodes, and two output nodes—one representing "Positive" and one for "Negative". The network was run for 200 epochs, with a learning rate of 0.3 and a momentum of 0.2. The model was created using 10-fold cross validation to avoid overfitting on the data set. Classifier accuracy on the set was 86.486%. 32 of the 37 instances were correctly classified. Broken down by class, the "Positive" classification had a precision of 0.88 and a recall of 0.917, and the "Negative" classification had a precision of 0.833 and a recall of 0.769.

The second classifier used was the WEKA J48 decision tree. The model was created using 10-fold cross validation to avoid overfitting. The first classification tree produced by the model is represented in Figure 35. First the instances are divided by the /d/ locus equation slope. Speakers with a /d/ slope of less than 0.396 are classified as having speech dysarthria. The remaining speakers are then separated by the value of the /g/ locus equation y-intercept. Speakers with an intercept above 418 Hz are classified as "Negative," and the rest are classified as positive. This tree has an accuracy of 81%, correctly classifying 30 of 37 instances. Precision for the "Negative" class is 0.815, and recall is 0.917. The "Positive" class had a precision of 0.800 and a recall of 0.615. The accuracy of this tree is promising, but recall for the "Positive" class is low, and the model is likely being thrown off by the unusually high /g/ locus equation y-intercept values for the dialect speakers. To more fully investigate the weight that each consonant held for the classification of speakers with dysarthria, the following three decision trees were generated using subsets of the initial feature set.

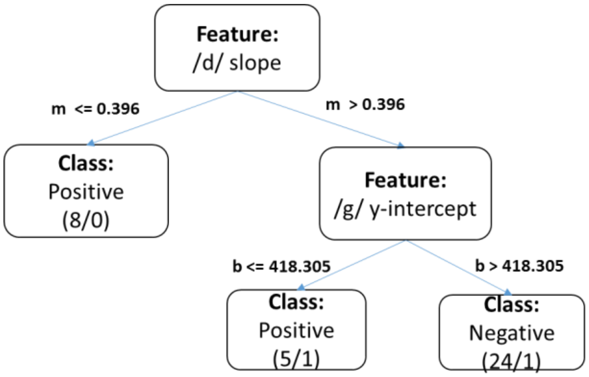


Figure 35: Dysarthria Classification Tree

The first reduced feature set included only the locus equation coefficients for /d/. The decision tree generated using this set had the same classification accuracy as the previous tree: 30 of 37 instances correctly classified. The "Negative" class precision dropped to 0.793, and recall rose to 0.958. The "Positive" class precision rose to 0.875, but the recall dropped to 0.538. The decision tree is visualized in Figure 36. The tree is interesting as it mirrors the patterns observed in section 5.1. Speakers with a /d/ locus equation slope between 0.396 and 0.698 are classified as "Negative," and speakers with slopes falling outside that range are classified as "Positive." This follows the pattern of speakers with dysarthria having locus equation slopes falling either above or below the expected median range.

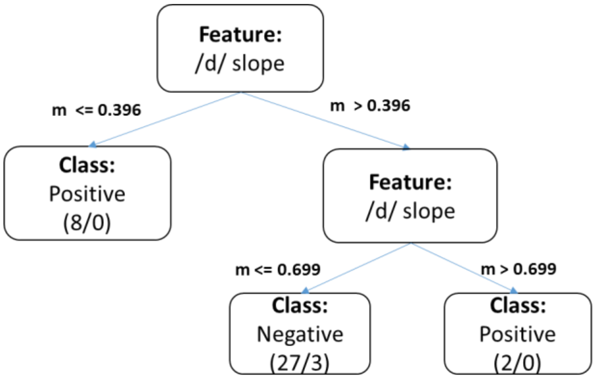


Figure 36: /d/ Classification Tree

The second reduced feature set had the locus equation coefficients for /b/. The classification accuracy drops for this tree. Only 26 speakers are classified correctly, for an accuracy of 70.27%. The "Negative" class has a precision of 0.724 and a recall of 0.875, and the "Positive" class had a precision of 0.625 and a recall of 0.385, a significant decline. The decreased accuracy shows that the /d/ locus equation coefficients were more significantly different for speakers with dysarthria than for speakers without it. The decision tree generated on this set is seen in Figure 37 below. Note that even though accuracy dropped, the tree did divide the set by slope rather than y-intercept once again.

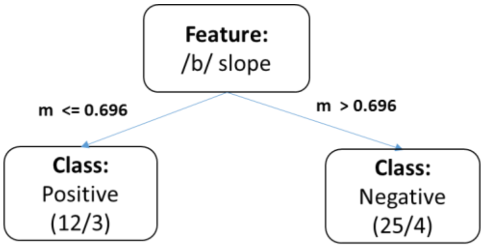


Figure 37: /b/ Classification Tree

The final reduced feature set focused on the /g/ locus equation coefficients. This was the only decision tree to divide the speaker set based on the y-intercept values rather than the slope. The classification accuracy for this tree went up to 83.78%, the highest seen for a decision tree.

The "Negative" class had a precision of 0.875 and a recall of 0.875, and the "Positive" class had a precision and recall of 0.769 each. The results of this tree must be treated with caution, as the /g/ coefficients for speakers from the dialect corpus ("Negative" speakers) were unusually shallow and likely thrown off by unbalanced vowel sets. That being said, the /g/ locus equation slopes for speakers with dysarthria were unusually high, which in turn pushed the /g/ locus equation y-intercept values down. The decision tree classifies the speakers with lower /g/ y-intercept values as "Positive," so the tree does reflect this aspect of the feature set. This decision tree is seen in Figure 38.

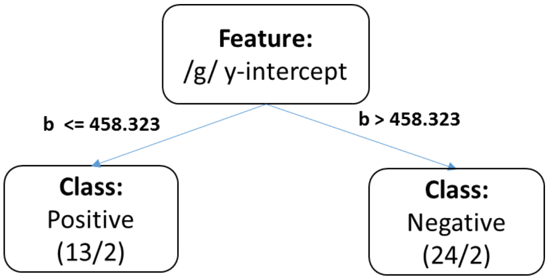


Figure 38: /g/ Classification Tree

In summary, locus equations provide useful information as features for classification of speech with dysarthria and speech without. The multilayer perceptron classifier had the highest classification accuracy at 86%, most likely because the perceptron is capable of capturing softer boundaries in the data than a decision tree, which relies on hard cut-off points for divisions. However, the multilayer perceptron is a black box algorithm, meaning the decisions made within the model are opaque to the user. The decision tree classifier had a competitive classification accuracy at 81%, and the rules built by the model contained interesting information regarding the nature of the locus equation coefficients in relation to the speech dysarthria. The /d/ locus equation held the most weight for classification of the speakers. This may be because the /d/

consonant is easy to move around, leading to over or under coarticulation, and so the slope values can deviate from the norm in either direction. The /g/ locus equation values were the next most important. The /g/ locus equation slope was significantly higher than the norm for most speakers with dysarthria, which also pushed the /g/ locus equation y-intercept values to be uniquely low. This heavy coarticulation of the /g/ sound can be attributed to the difficulty of producing "g". Infants produce "b" and "d" consonants before "g" consonants because they are easier (Ferguson and Farwell, 1975). A speaker with dysarthria is likely to also have a hard time moving the tongue fully back into position for a true /g/, and so the sound is heavily coarticulated. The /b/ locus equation seemed to hold the least amount of weight, although many of the speakers with dysarthria did have /b/ locus equations with noticeable smaller slopes than expected. While the classification results for speakers with and without dysarthria were not perfect, they were promising. Locus equations alone are not sufficient for classification of dysarthria, and it is doubtful they would be sufficient for other speech disorders either. They do, however, contain useful and easily visible information about the speech. Coarticulation is an important aspect of speech, and its continuous nature makes it hard to capture. Locus equations are useful features to be included along with other characteristics of speech such as speed and vowel shape.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

At the beginning of this work, six research questions were posed as topics that would be investigated over the course of the study. The six were (1) Can locus equations be automatically generated from more conversational speech? Does this affect the integrity of the locus equation? (2) Are the generated locus equations accurate enough for recovery of place of articulation? Can machine learning methods be applied to recover place of articulation across more than one manner class? (3) Are the locus equation coefficients valuable features for classification of dialect? (4) Do locus equations still form for speakers with dysarthria? (5) Are the locus equation coefficients valuable features for recognition of speech dysarthria? (6) Do locus equations contain enough information to serve as valuable features in speech recognition and classification? The last question can be seen as the overarching theme of study for this thesis—the usefulness of locus equations in the domain of speech recognition and classification.

The first question was an expansion on previous work done with locus equations. The studies done previously all used very carefully crafted token recordings as the basis for locus equations. If locus equations were to be used as a feature in most speech recognition and classification problems, however, they would need to be present in more than just careful laboratory speech. The speech in the NSP, although still collected in the laboratory, was not produced with locus equations in mind. Using tokens from multisyllabic words and sentences along with the basic CVC tokens provided insight into the generality of locus equations. The answer to question one is both yes and no. At this point in time, locus equations cannot be fully automatically generated from conversational speech. The issue here is with speech alignment—although locus equations could be generated from the sentences, formant values had to be drawn from exactly the right place for the equations to be valid. For this to be fully automated, a speech alignment system would need to be near perfect. The second issue with automatic alignment for this study was outliers caused by errors in Praat measurement of formants. For locus equations with a larger number of tokens, outliers were easily found and discarded. For equations with smaller numbers of tokens, they were a problem. The solution to this problem seems to be one of three things. The first solution is to only generate equations once a large number of transitions have been taken, at least upwards of 30 or 40. From here outliers can be detected and discarded. The second solution is to wait for an increase in accuracy of speech tools like Praat. The third solution is to develop a system which automatically detects potential outliers and re-measures them with updated settings.

The second question was a continuation of the first, and also an expansion on previous work done with locus equations. Using the semi-automatically generated locus equations coefficients to identify the place of articulation was a good way to double check the validity of the equations. Previous studies have proven that locus equations coefficients should indicate place of articulation, so if the automatic equations did not it would count as proof against their validity (Fowler, 1994; Krull 1988; Sussman et al., 1991). The expansion is a continuation of the debate that took place between Sussman and Fowler in the 1990s (Fowler, 1994; Sussman et al., 1991; Sussman and Shore, 1996). Fowler created locus equations for consonants other than voiced stops, and claimed that the coefficients began to overlap with one another, making locus equations insufficient for recovery of place of articulation. The experiment described in 4.2 does not address this argument, as manner of articulation is included as a feature in the feature set.

The experiment does show that consonants across multiple manner classes can be classified by place of articulation with fairly high accuracy if manner of articulation is given as a feature.

The third question was the first exploration into locus equations as features indicative of something else—in this case, speaker dialect. The results in this experiment were resoundingly negative. Classifier accuracy stayed below 10% across multiple different machine learning methods, and an investigation into the consonant space of the equations showed that there was no significant difference between the dialects. This result still provided valuable information, namely that locus equations are robust across a variety of speech, meaning that if they do serve as valuable features elsewhere the user can be confident that the locus equations are not being affected by small individual speaker influences. Although the locus equations could not classify the dialects included in the NSP corpus, they did vary in response to a few of the speakers with heavier accents. A study using locus equations for classification of more significantly different speech, like foreign accents, would be an interesting continuation. It would also be useful to conduct a study of locus equations across dialects with more caution in regard to the vowels. Speakers from different dialects produce different vowels when speaking—for instance, some dialects tend to front back vowels. This could have had an effect on the locus equations. Conducting an experiment where all of the vowels were carefully held steady across dialects could either improve classification using locus equations, or erase the small differences that were present.

With the fourth question, the research moves away from dialects and into speech disorders. The UA corpus of speakers with dysarthria contained a lot of variety, with speaker intelligibility ranging from 4% to 93%. Only voiced stops were examined using this corpus. The locus equations were still generated using the automated Praat script, but the automatic outlier detection was not used. This was both because there were a smaller number of tokens, and because speakers with dysarthria were expected to create locus equations with a higher number of outliers in them. Instead, outliers were hand checked and fixed. Although the regression lines tended to have weaker correlations for speakers with dysarthria, the $F2_{vowel}$, $F2_{onset}$ mappings still created valid locus equations. The validity of these equations was checked in the same way the dialect locus equations were checked; by comparison with previous experiment results, and by application of a classification algorithm for recovery of place of articulation.

Question five was addressed in section 5.3, where locus equation coefficients were used as features for a classifier that labelled speakers either “positive” or “negative” for speech dysarthria. Unlike the classification results with the dialect data, the classifier systems developed here were up to 86% accurate. Additionally, the way the decision tree classifier used the locus equations provided information about the speech itself. Although the classifier was not perfect, this experiment provided support for locus equations as valuable features for speech classification and identification systems.

Overall, locus equations seem like a promising feature for inclusion in speech recognition and classification systems. The issue of obtaining the alignment points at which to draw F2 values for the locus equations poses the biggest threat to locus equations as a useful feature. This is not an insurmountable issue—it is likely speech alignment programs will continue to improve rapidly. In the meantime, hand-alignment of data, though work intensive, is a valid alternative for preparing the data. Otherwise, locus equations have proven to be present even in more continuous, less-controlled speech. Classifier systems have been proven to handle semi-automatically generated locus equations well. Locus equations are useful for recovery of place of articulation of many consonants, especially if manner is included as a feature. Finally, locus equations are robust to small variations in speech, but they do reflect coarticulation and the consonant space of a person.

There are many projects that could be pursued as future work into locus equations as features. One such project would be the development of a program to automatically recognize and resample outliers in the locus equation plots to improve accuracy and automation. Another project would be to create locus equations for non-native English speakers with heavy foreign accents. Past work has been done on the use of locus equations to describe characteristics of a language (Everett, 2008). It would be interesting to see how locus equations help to represent foreign accents. A third project would be to expand the study of locus equations as feature sets for classification of speech disorders. There are a few speech disorders, such as fronting, that deal almost entirely with positioning of articulators for consonants in the mouth. Locus equations might be especially relevant for disorders such as these. Finally, it would be interesting to see if locus equations can be worked into speech models, to help as a predictor of what speech should sound like in speech recognition technology.

REFERENCES

Blumstein, Sheila E., and Kenneth N. Stevens. "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 66.4 (1979): 1001. Web.

Boersma, P and Weenink, D. 2016. Praat: doing phonetics by computer, http://www.praat.org.

Brancazio, Lawrence, and Carol A. Fowler. "On the Relevance of Locus Equations for Production and Perception of Stop Consonants." *Perception & Psychophysics* 60.1 (1998): 24-50. Web.

Chennoukh, Samir. "Locus Equations in the Light of Articulatory Modeling." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 102.4 (1997): 2380. Web.

Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. Speech Communication, 48, 633-644.

"The CMU Pronouncing Dictionary." *The CMU Pronouncing Dictionary*. Carnegie Melon University, n.d. Web. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.

Daniloff, R. and R. Hammarberg. “On Defining Coarticulation” *Journal of Phonetics* 2 (1973): 239-248.

Delattre, Pierre C. "Acoustic Loci and Transitional Cues for Consonants." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 27.4 (1955): 769. Web.

"Dysarthria." *Dysarthria*. American Speech-Language-Hearing Association, n.d. Web. <http://www.asha.org/public/speech/disorders/dysarthria/#comm_better>.

Everett, Caleb. "Locus Equation Analysis as a Tool for Linguistic Fieldwork." *Language Documentation and Conservation* 2.2 (2008): 185-211. Print.

Ferguson, Charles A., and Carol B. Farwell. "Words and Sounds in Early Language Acquisition." *Language* 51.2 (1975): 419-39. Web.

Fowler, Carol A. "Invariants, Specifiers, Cues: An Investigation of Locus Equations as Information for Place of Articulation." *Perception & Psychophysics* 55.6 (1994): 597-610. Web.

Gibson, Terrie, and Ralph N. Ohde. "F2 Locus Equations: Phonetic Descriptors of Coarticulation in 17- to 22-Month-Old Children." *J Speech Lang Hear Res Journal of Speech Language and Hearing Research* 50.1 (2007): 97. Web.

"IPA Chart." *IPA Home*. International Phonetic Association, n.d. Web. <https://www.internationalphoneticassociation.org>.

Iskarous, Khalil, Carol A. Fowler, and D. H. Whalen. "Locus Equations Are an Acoustic Expression of Articulator Synergy." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.*128.4 (2010): 2021-032. Web.

Kim, Heejin, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame. "Dysarthric Speech Database for Universal Access Research."*INTERSPEECH Conference* (2008): n. pag. Web.

Krull, Diana, and B. Lindblom. “Comparing Vowel Formant Data Cross linguistically.” *PERILUS* 15 (1992): Web.

Krull, Diana, and B. Lindblom. "Sorting Stops by Place in Acoustic Space." (n.d.): n. pag. Web.

Krull, Diana. "Relating Acoustic Properties to Perceptual Responses: A Study of Swedish Voiced Stops." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 88.6 (1990): 2557. Web.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Ladefoged, Peter. *A Course in Phonetics*. Boston: Thomson Wadsworth, 2006. Print.

Potter, Ralph K., Harriet C. Green, and George A. Kopp. *Visible Speech*. New York: Van Nostrand, 1947. Print.

LINDBLOM, B. (1963). *On vowel reduction* (Report No. 29). Stockholm: Royal Institute of Technology, Speech Transmission Laboratory.

Montgomery, Allen, Paul E. Reed, Kimberlee A. Crass, H. Isabel Hubbard, and Joanna Stith. "The Effects of Measurement Error and Vowel Selection on the Locus Equation Measure of Coarticulation)." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 136.5 (2014): 2747-750. Web.

Nagoya Institute of Technology. 2010. Open-source large vocabulary csr engine julius, rev. 4.1.5.

Nearey, T. M., and S. E. Shammass. "Formant Transitions as Partial Invariants in the Identification of Voiced Stops." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 79.S1 (1986): n. pag. Web.

Sussman, Harvey M., and Jadine Shore. "Locus Equations as Phonetic Descriptors of Consonantal Place of Articulation." *Perception & Psychophysics* 58.6 (1996): 936-46. Print.

Potter, Ralph Kimball, George Adams Kopp, and Harriet Green Kopp. *Visible Speech*. New York: Dover Publications, 1966. Print.

Rabiner, L.r. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE Proc. IEEE* 77.2 (1989): 257-86. Web.

Stevens, Kenneth N. *Acoustic Phonetics*. Cambridge, MA: MIT, 1998. Print.

Stevens, Kenneth N., and Arthur S. House. "Perturbation of Vowel Articulations By Consonantal Context: An Acoustical Study." *Journal of Speech Language and Hearing Research J Speech Hear Res* 6.2 (1963): 111-28. Print.

Sussman, Harvey M., Courtney T. Byrd, and Barry Guitar. "The Integrity of Anticipatory Coarticulation in Fluent and Non-fluent Tokens of Adults Who Stutter." *Clinical Linguistics & Phonetics*25.3 (2010): 169-86. Web.

Sussman, Harvey M., Helen A. McCaffrey, and Sandra A. Matthews. "An Investigation of Locus Equations as a Source of Relational Invariance for Stop Place Categorization." (1991): n. pag.*ResearchGate*. Web. 11 Apr. 2016.

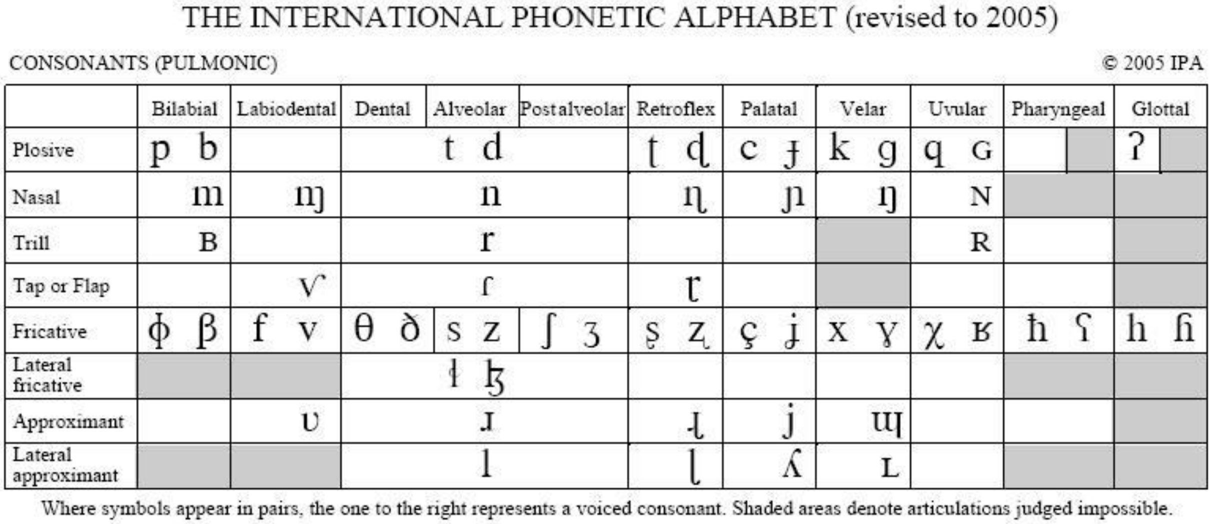
Sussman, Harvey M., Kathryn A. Hoemeke, and Farhan S. Ahmed. "A Cross-linguistic Investigation of Locus Equations as a Phonetic Descriptor for Place of Articulation." *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.* 94.3 (1993): 1256. Web.

VoxForge. 2006-2011. http://www.voxforge.org

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book*. Cambridge: Entropic Cambridge Research Laboratory, 1999. Print.

APPENDIX A

IPA CHART



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap		ⱱ		ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

APPENDIX B

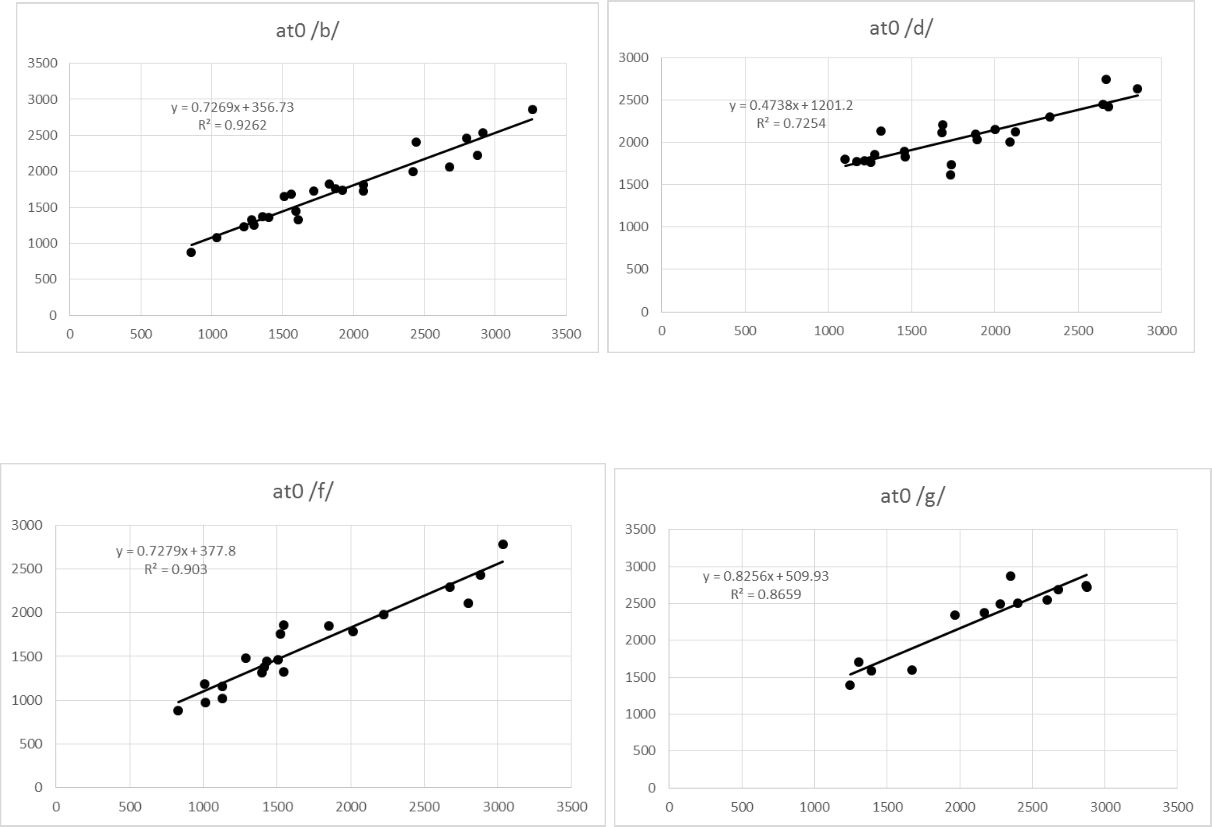
NSP TOKENS

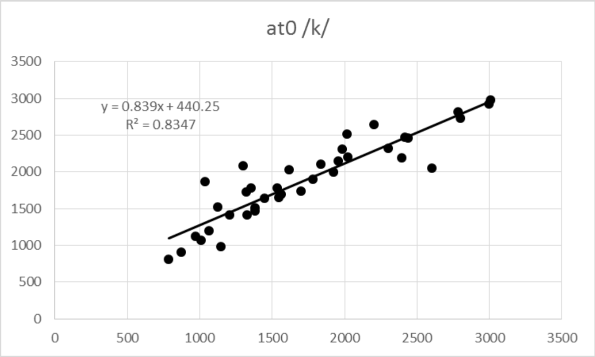
CVC Tokens				**HP Sentences**		**Multisyllable Tokens**	
Number	**Word**	**Number**	**Word**	**Number**	**Sentence**	**Number**	**Sentence**
0	bean	**43**	luck	**0**	A bicycle has two wheels.	**1**	absentee
1	bite	**44**	lull	**1**	A round hole won't take a square peg.	**7**	alligator
2	boat	**45**	lung	**2**	A spoiled child is a brat.	**8**	amphibian
3	boil	**46**	main	**3**	Ann works in the bank as a clerk.	**9**	anonymous
4	bull	**47**	math	**4**	Banks keep their money in a vault.	**15**	bazooka
5	cab	**48**	meal	**5**	Bob was cut by the jacknife's blade.	**16**	bikini
6	calm	**49**	mile	**7**	Cut the meat into small chunks.	**23**	clarinet
7	can	**50**	mill	**8**	Eve was made from Adam's rib.	**26**	coronation
8	caught	**51**	mob	**9**	Follow this road around the bend.	**29**	deactivate
9	coal	**52**	pal	**10**	For dessert he had apple pie.	**33**	disapproval
10	code	**53**	pen	**11**	Get the bread and cut me a slice.	**37**	evaporate
11	coin	**54**	pin	**13**	He rode off in a cloud of dust.	**40**	feminism
12	con	**55**	poke	**16**	Her hair was tied with a blue bow.	**41**	forecast
13	cool	**56**	pool	**18**	I ate a piece of chocolate fudge.	**42**	functionary
14	cot	**57**	pull	**21**	I've got a cold and a sore throat.	**43**	gallop
15	cough	**58**	rice	**22**	Keep your broken arm in a sling.	**46**	guitar
16	death	**59**	rip	**23**	Kill the bugs with this spray.	**51**	influenza
17	dig	**60**	sail	**24**	Maple syrup is made from sap.	**55**	kazoo
18	dime	**61**	sell	**25**	My jaw aches when I chew gum.	**58**	macaroni
19	dock	**62**	sour	**30**	Paul took a bath in the tub.	**64**	mispronounce
20	doll	**63**	south	**31**	Paul was arrested by the cops.	**65**	museum
21	doubt	**64**	tape	**32**	Peter dropped in for a brief chat.	**66**	nectarine
22	dull	**65**	tool	**33**	Playing checkers can be fun.	**77**	peninsula
23	fade	**66**	towel	**34**	Please wipe your feet on the mat.	**111**	victorious
24	fail	**67**	town	**36**	Ruth had a necklace of glass beads.		
25	feed	**68**	tube	**37**	Ruth poured herself a cup of tea.		
26	fell	**69**	voice	**42**	The bird of peace is the dove.		
27	fire	**70**	void	**45**	The bride wore a white gown.		
28	fool	**71**	walk	**51**	The chicken pecked corn with its beak.		
29	foul	**72**	wall	**54**	The cow gave birth to a calf.		
30	full	**73**	wet	**56**	The dealer shuffled the cards.		
31	gap	**74**	wool	**62**	The gambler lost the bet.		
32	good	**75**	wrong	**69**	The lion gave an angry roar.		
33	guide			**71**	The nurse gave him first aid.		
34	head			**73**	The pond was full of croaking frogs.		
35	heal			**74**	The poor man was deeply in debt.		
36	hill			**79**	The stale bread was covered with mold.		
37	home			**80**	The story had a clever plot.		
38	keep			**81**	The super highway has six lanes.		
39	lit			**82**	The swimmer dove into the pool.		
40	loud			**84**	The thread was wound on a spool.		
41	love			**88**	They tracked the lion to his den.		
42	loyal			**91**	Tighten the belt by a notch.		

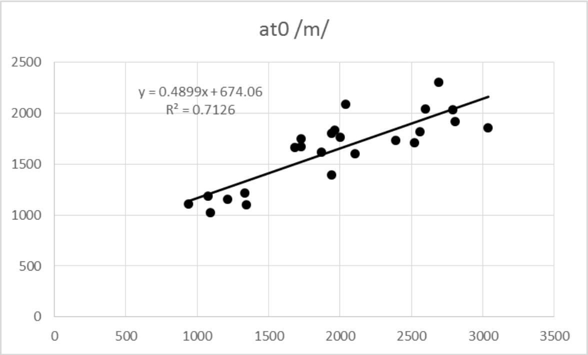
APPENDIX C

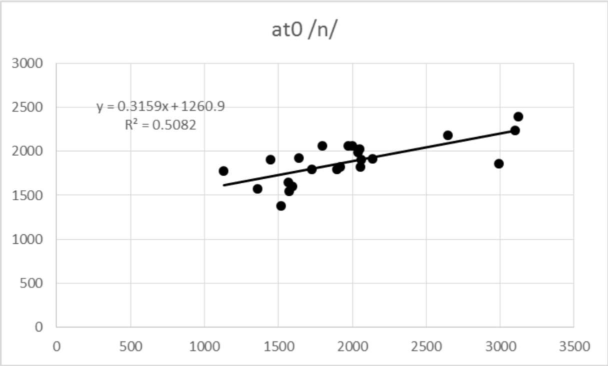
DIALECT LOCUS EQUATIONS

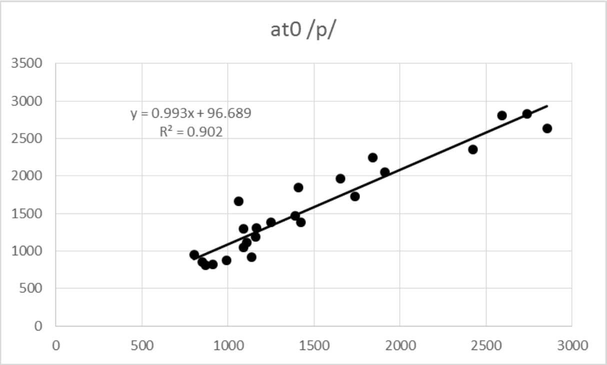
Locus equations for speaker at0. F2 Vowel (Hz) is graphed along the x-axis, and F2 Onset (Hz) is graphed along the y-axis.

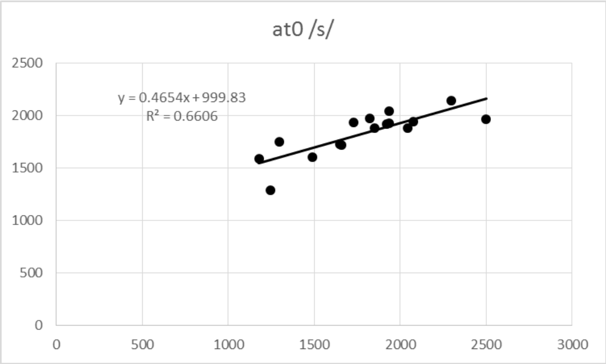


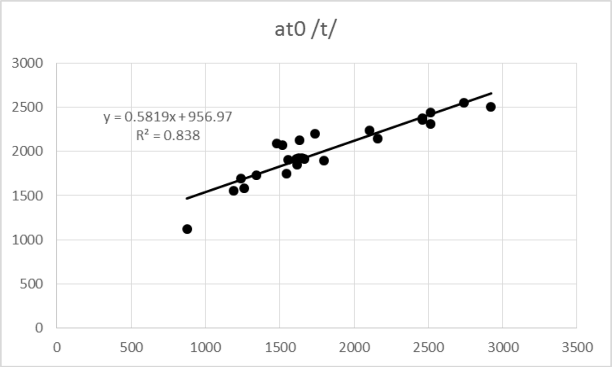

at0 /k/
y = 0.839x + 440.25
R² = 0.8347
3500
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500


at0 /m/
y = 0.4899x + 674.06
R² = 0.7126
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500


at0 /n/
y = 0.3159x + 1260.9
R² = 0.5082
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500

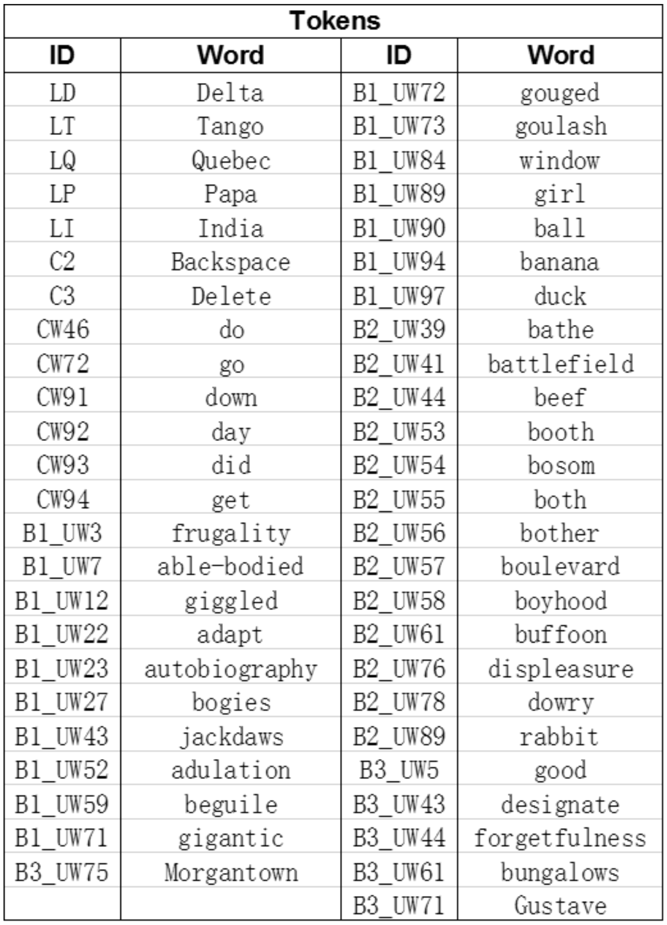

at0 /p/
y = 0.993x + 96.689
R² = 0.902
3500
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000


at0 /s/
y = 0.4654x + 999.83
R² = 0.6606
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000


at0 /t/
y = 0.5819x + 956.97
R² = 0.838
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500

APPENDIX D

UA TOKENS

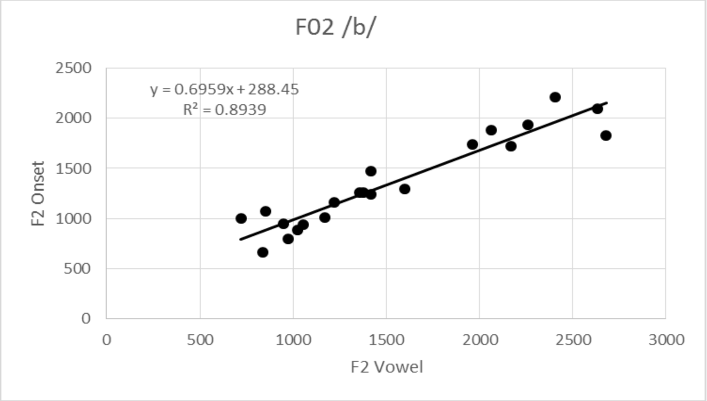


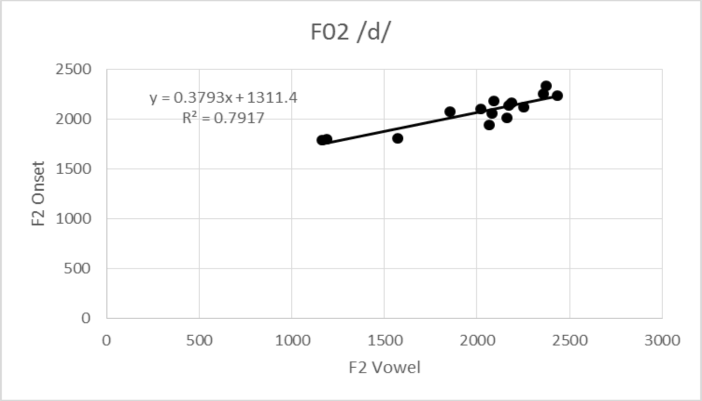
Tokens			
ID	**Word**	**ID**	**Word**
LD	Delta	B1_UW72	gouged
LT	Tango	B1_UW73	goulash
LQ	Quebec	B1_UW84	window
LP	Papa	B1_UW89	girl
LI	India	B1_UW90	ball
C2	Backspace	B1_UW94	banana
C3	Delete	B1_UW97	duck
CW46	do	B2_UW39	bathe
CW72	go	B2_UW41	battlefield
CW91	down	B2_UW44	beef
CW92	day	B2_UW53	booth
CW93	did	B2_UW54	bosom
CW94	get	B2_UW55	both
B1_UW3	frugality	B2_UW56	bother
B1_UW7	able-bodied	B2_UW57	boulevard
B1_UW12	giggled	B2_UW58	boyhood
B1_UW22	adapt	B2_UW61	buffoon
B1_UW23	autobiography	B2_UW76	displeasure
B1_UW27	bogies	B2_UW78	dowry
B1_UW43	jackdaws	B2_UW89	rabbit
B1_UW52	adulation	B3_UW5	good
B1_UW59	beguile	B3_UW43	designate
B1_UW71	gigantic	B3_UW44	forgetfulness
B3_UW75	Morgantown	B3_UW61	bungalows
		B3_UW71	Gustave

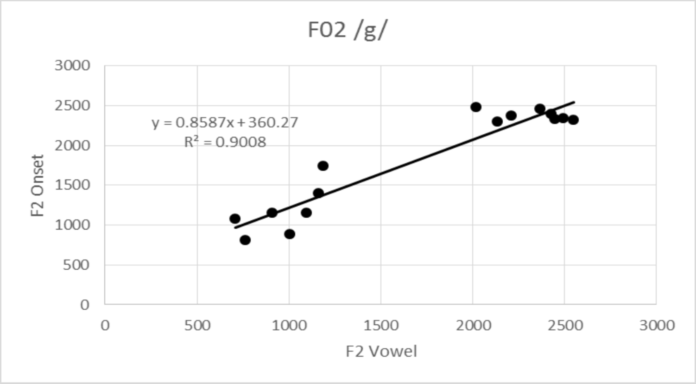
APPENDIX E

DYSARTHRIA LOCUS EQUATIONS

Locus equations for speaker F02. F2 Vowel (Hz) is graphed along the x-axis, and F2 Onset (Hz) is graphed along the y-axis.


F02 /b/
y = 0.6959x + 288.45
R² = 0.8939
F2 Onset
F2 Vowel


F02 /d/
y = 0.3793x + 1311.4
R² = 0.7917
F2 Onset
F2 Vowel


F02 /g/
y = 0.8587x + 360.27
R² = 0.9008
F2 Onset
F2 Vowel