Modeling Coarticulation in Continuous Speech

Brian O. Bush
M. S. Industrial and Systems Engineering, Ohio University, 1996
B. A. Computer Science and Applied Mathematics, Hiram College, 1994

Presented to the
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree
Doctor of Philosophy
in
Computer Science & Engineering

October 2015

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph. D. dissertation of

Brian O. Bush

has been approved.

Dr. Alexander Kain, Thesis Advisor
Associate Professor

Dr. Kyle Gorman
Software Engineer, Google, Inc.

Dr. Xubo Song
Professor

Dr. Kris Tjaden
Professor

Dr. Alison Hill
Assistant Professor

Acknowledgments

It is not an easy or fair task to acknowledge all the people who made this thesis possible with only a few words. However, I will try to do my best to extend my great appreciation to everyone who helped me scientifically, emotionally and financially through out this study.

I would like to express my deepest gratitude to my advisor, Dr. Alexander Kain, for his excellent guidance, instruction, patience, and providing me with an excellent atmosphere for doing research. I also would like to thank my initial advisor, Dr. John Paul Hosom, whom opened my mind to the world of speech. I can't imagine having better role models for both science and teaching.

I would like to thank my thesis committee members, your comments and direction have made this body of work possible. I am grateful to our department's Academic Program Coordinator, Patricia Dickerson who kept the department running smoothly and was always ready to help. I am also grateful to Steven Bedrick for managing the computing resources in addition to his professorial tasks. I would like to thank my colleagues at the Center for Spoken Language Understanding (CSLU) at OHSU, RuleSpace and Sensory, for their friendship, support and enlightening discussions. Thanks to Leo Bush for his assistance in correcting vowel formants in spectrograms of dysarthric speakers. Special thanks to Ken Williams, Akiko Amano-Kusumoto, Brian Snider, Pieter Vermeulen, Jan van Santen, Hideo Kitajima, Luis Rabelo and Bob Ehlen.

I gratefully acknowledge the funding sources that made my studies possible. I was funded by the National Science Foundation (NSF Grant: IIS-0915754) for the first three years and by Oregon ETIC for the remaining two years of my studies.

I dedicate this thesis to the next generation, my children: Hana, Joe and Leo. You are the pride and joy of my life. I love you more than anything and I appreciate all your patience and support during my studies.

At last, I don't know how to begin with saying thank you to my dearest wife and best friend, Ryoko. I love you for everything, for being so understanding and for putting up with me through yet another chapter in our wild adventure. You are the best thing that has ever happened to me.

Brian O. Bush
Oregon Health & Science University
October 2015

Contents

List of Tables

List of Figures

Abstract

Modeling Coarticulation in Continuous Speech

Brian O. Bush

Doctor of Philosophy
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine

October 2015
Thesis Advisor: Dr. Alexander Kain

Coarticulation is the articulatory influence of one phonetic element (e. g., vowel or consonant) on another. Our primary motivation to modeling coarticulation is to explain how a speech signal is influenced by phonetic context. Other straight forward applications of modeling coarticulation might include compact parametric speech synthesis techniques. Thus far, modeling coarticulation in speech has been largely limited to short sequences and/or limited phonetic context.

This dissertation describes a coarticulation model for continuous speech. The document first introduces a methodology for modeling formant frequency in continuous speech, allowing examination of syllable-level coarticulation in natural contexts. The model represents continuous trajectories as a combination of overlapping local trajectories, which are represented by a weighted-addition of acoustic event targets by sigmoidal coarticulation functions characterized by slope and position. An explicit coarticulation model will better explain and possibly predict the variability in speech that currently is only handled sufficiently with large amounts of data.

We present a straight-forward approach to parameter estimation that is achieved through a search of acoustic event global targets along with local slope and position parameters. We then validate our estimation procedure on a synthetic corpus that is generated using known targets. Using our target estimation procedure, we demonstrate that we can recover most of the targets.

Those targets not exactly recovered are only slightly differing with the resulting error negligible. Thus, we are confident that we can estimate actual unknown formant frequency targets from speech. Then using a male clear speech corpus, we estimate model parameters using the described algorithm. The resulting model formant frequency targets are comparable with well-known published targets. Coarticulation parameters are investigated to show the model is capturing known acoustic-phonetic relationships.

Next, we present results from a perceptual evaluation that was conducted by means of a speech intelligibility test to examine whether resynthesis from model parameters produces speech that is as intelligible as vocoded speech using observed formant frequency, thus validating the model, by demonstrating that the model is capturing the necessary components of speech necessary for intelligibility. To show the applicability of our coarticulation model we present results to three experiments, namely (1) investigation of clear and conversation formant frequency targets, (2) formant tracking, and (3) measuring coarticulation in dysarthric speech detection

We first investigate acoustic feature differences in clear (CLR) and conversational (CNV) speech. We begin this investigation by examining observed formant frequency centers of vowels between clear and conversational speech for two distinct corpora derived from a single speaker: CVC words, and vowels in unrestricted phoneme contexts. Using the continuous speech corpus, we then apply our modeling approach to estimate formant frequency targets for both CLR & CNV styles. Using clear speech formant frequency targets we model conversational speech, and conversely, conversational formant frequency targets with clear speech. We demonstrate that CLR and CNV speech targets are quite distinct in the CVC corpus, however, quite similar in the continuous speech corpus. We conclude that, for this speaker, CLR and CNV speech can be represented by formant frequency targets derived from either style, however this holds only in general speech.

In our formant tracking experiment, we derive 24 sets of speaker dependent formant frequency targets using data from TIMIT, where we focus on vowel, nasal and approximant target estimation. We compare model formant trajectories with hand-corrected published trajectories. We present these results combined and broken-down into phoneme-class based results for both male and female speakers. Overall, we find our model exhibits lower overall root-mean-square error relative to WaveSurfer across all broad phoneme classes.

Finally, we investigate the characterization of dysarthric speech by demonstrating formant frequency target vowel-space differences between disordered and non-disordered speakers. Applying our model to both disordered and non-disordered speech, we show several attributes of model parameters could be employed as a robust indicator of intelligibility. These specific attributes are F1/F2 targets to measure vowel space, goodness of fit between modeled and observed formant

trajectories. Both measures had a linear relationship with assessed intelligibility.

This research represents definitive progress towards the modeling coarticulation for continuous speech and the applications detail success in modeling of speech. However, there is much more research that remains in efficient parameter estimation to make presented applications tractable. Finally, we conclude with some thoughts as to which paths we might now take to further this research.

Chapter 1

Introduction

Conversational speech produces several syllables per second. During speech production the brain coordinates the movement of the lips, tongue and other articulators from one position to the next. The articulators transition smoothly during production of adjacent vowels and consonants. This phenomenon is known as *coarticulation*. To date there is no comprehensive data-driven model of coarticulation for arbitrary sequences of acoustic events.

1.1 Summary of Previous Approaches

Historically, formant frequency targets of a few selected phonemes were estimated using isolated phoneme sequences, limited at most consonant-vowels-consonants and often with some manual intervention [23, 68, 82, 22, 13, 99, 15, 70]. Most notable is the work of Broad and Clermont [13] in which the authors produced several models of formant transitions in consonant-vowel (CV) and CV/d/ contexts, where the prevocalic consonant was limited to voiced stops. Their more elaborate model used a linear combination of coarticulation functions (modeled with exponential functions) and target values. Niu *et al.* [79] expanded Broad and Clermont's model to a broader set of consonants (/b,d,g,p,t,k,s,sh,t,f/ and four vowels), but limited the modeling to the vowel centers in their application to model coarticulation in dysarthric speech. Amano *et al.* [4] further improved upon that approach by modeling the entire formant trajectory of the vowel region and moving to a sigmoid function. A sigmoid was chosen because it restricts the coarticulatory effects to be smoothly and monotonically increasing or decreasing, specifically in cases where the consonant being modeled is an approximant (/y,l,w,r/). Vowels were limited to monophthong vowels.

Our previous work [6, 18] investigated fully-automatic and comprehensive estimation of formant frequency targets for a single male speaker, but was limited to modeling CVC triphones. The brute-force parameter estimation approach highlighted the fact that the formant target space can have solutions that are nearly equal in error, which would leave gradient-descent methods with

sub-optimal solutions.

1.2 Thesis and Proposed Approach

In this dissertation, we will advance the state of the art in the area of modeling coarticulation for arbitrary sequences of acoustic events. Our primary objective of this thesis is to show that formant frequency of continuous speech can be effectively modeled using a linear combination of global acoustic event-specific targets and local coarticulation parameters. This data-driven model uses formant frequency targets that are speaker-specific, but independent of speaking style and phonemic context. We demonstrate that these targets, including classes of sounds where formants are not directly observable, can be automatically derived from continuous speech. In this work, we hypothesize that formant frequency targets are global, yet speaker specific. Furthermore, we handle diphthongs and affricates by decomposing them into two separate events. Secondary objectives are (1) validation of modeling approach using a synthetic corpus, (2) perceptual study to demonstrate model is capturing important aspects of speech required for intelligibility, (3) investigating clear and conversational speech targets, (4) automatic error correction in formant tracking, and (5) dysarthria diagnosis wherein we might be able to measure coarticulation differences between disordered and non-disordered speakers.

1.3 Outline

The remainder of this dissertation is organized as follows:

Chapter 2 briefly explains the speech production process and the coarticulation phenomenon, and the provides a literature review of coarticulation modeling approaches that have been proposed and discuss their limitations and applicability to general speech.

Chapter 3 describes the developed coarticulation model. First, we formulate a basic triphone model for modeling local coarticulation of three consecutive acoustic events. Then we describe a global model of coarticulation that is capable of modeling continuous speech. This model represents continuous trajectories as a combination of overlapping local trajectories. We will discuss the approach developed for parameter estimation. For this research, we use traditional formant frequency. As a validation procedure for parameter estimation, we use a synthetic corpus to demonstrate the robustness of our estimation procedure. We then estimate both acoustic event targets and coarticulatory parameters on a continuous speech corpus in two distinct experiments: (1) all acoustic event targets are estimated, and (2) only sonorants event targets are estimated while

other consonants are fixed. We investigate coarticulatory parameters and compare estimated targets with previously published targets. Finally, we present results from two perceptual evaluations that were conducted by means of a speech intelligibility test to examine whether resynthesis from model parameters produces speech that is as intelligible as vocoded speech using observed formants, thus validating the model, its estimation procedure and demonstrating that the model is capturing the necessary components of speech necessary for intelligibility.

Chapter 4 reports on investigating the application of our model to clear and conversational speech for a single male speaker. First, we examine vowel centers for CVC words and vowel centers in unrestricted phoneme contexts to highlight important style-specific differences in two separate corpora. Next, we estimate formant frequency targets for both clear (CLR) and conversational (CNV) speech for both corpora. Next we optimize coarticulation parameters of CLR speech using CNV targets, and vice-versa with CNV speech. Examining RMS error from the resulting trajectories, we find that CLR and CNV targets almost yield the same error for the continuous corpus. We present target error surfaces of the continuous corpus to demonstrate the feasibility of our target estimation technique, and the similar error surfaces between speech styles. Finally, we also perform an analysis of coarticulation parameters to highlight differences between CLR and CNV speech in both corpora.

Chapter 5 reports on our technique and consequent results in applying our continuous coarticulation model to the problem of formant tracking of the first and second formants. In this application we analyze and present results for 24 speakers (twelve male and twelve female) and compare with other formant tracking techniques and their results. We will measure success on a standard published corpus of hand corrected formant tracks where we show phoneme class-based results.

Chapter 6 reports on applying the model to Dysarthria diagnosis and analyzing coarticulatory effects. In this application, we will demonstrate a centralization in vowel space for disordered speakers as compared with non-disordered both as observed vowel centers in F1/F2 space along with formant frequency targets. Using six disordered speakers and two non-disordered speakers, we show differences between the two groups in vowel-space measurements, goodness of fit and coarticulation parameters.

Chapter 7 summarizes the proposed methodology and its contributions and merits along with results presented in this dissertation, evaluates a functional interpretation of the findings, and provides an overall discussion about coarticulation modeling along with possible extensions to the model.

Finally, Appendix A describes four numerical attributes for each acoustic event (used in Chapter 3 in perceptual validation for generating decoy words) and includes a mapping from WorldBet

to the alphabet used throughout this dissertation. Appendix B details the contrived F2 formant frequency targets used in the validation of the estimation validation procedure in Chapter 3. Appendix C details the formant frequency targets used in estimation techniques in Chapters 3 and 5. Appendix D details the sentences used in the perceptual experiment in Chapter 3.

Chapter 2

Background

In order to develop a comprehensive model of coarticulation, it is important to understand theories of human speech production and perception. This section also provides background on some of the more important models of coarticulation that have been developed and in doing so introduces concepts and terminology in speech processing.

2.1 Speech Production

Speech production starts with the flow of air pushed by the respiratory system. The respiratory system can be divided into the upper and lower systems. The upper respiratory system is composed of mouth, nose and nasal cavity, pharynx and larynx. The lower respiratory system consists of the trachea, bronchi, lungs and diaphragm. See Figure 2.1 for a pictorial view of the upper respiratory and vocal apparatus. The respiratory system provides energy for sound production in the form of a relatively constant air pressure. The larynx has two horizontal folds of tissue, these are the vocal folds. The gap between the folds is known as the glottis. The vibrating vocal folds convert this air pressure to a series of rapid air pulses. These rapid air pulses are created by repeated vibratory cycles, which produce a voiced sound (a "buzzy sound"). This voiced sound is then filtered and modified by the vocal tract resonators, producing voice as we know it. In this system, loudness is defined by the volume of air passing, which increases the amplitude of the sound pressure wave. In contrast, pitch is an increase or decrease in vocal fold vibration.

One of the most common models of speech production is called the "source filter" model which was proposed in 1848 by Johannes Müller and detailed by Fant [34]. This model is composed of three distinct parts: (1) a sound source, (2) the tube through which this sound is passed and (3) and radiation of the sound from the mouth. This model can be used to describe the speech signal in the spectral domain. In this model of speech production, the vocal-tract is considered as an acoustic tube with varying cross-sectional area. The shape of the vocal tract determines how the

source is to be filtered. Varied filtering leads to distinct resonance patterns associated with vowels and consonants. This model is quite effective at describing the overall characteristics of speech, such as the overall spectral shape of sounds and formant shapes along with their locations based on the shape of the tract. Another example is that the model explains the overall increase in formant frequencies in female speakers, which is due to the vocal tract length differences in male and female speakers.

2.2 Articulation

The process of vocal tract modification is called articulation and is performed through the dynamic control of the articulators: jaw, tongue, lips, velum and soft palate. Articulators not only change the configuration of the vocal tract, but can also generate sounds (noise) that is independent of the vocal fold vibration. These two noise sources are: frication and plosion. Frication is produced by forcing air through a narrow constriction in the mouth. Frication is the sound source for the phonemes /f/ and /s/. Plosion, however, is the buildup of of air pressure behind an obstruction (such as the lips) and then quickly removing that obstruction which results in a sudden burst of air being released from the mouth.

There are three primary states the glottis can be in: closed, vibrating or open. When the glottis is **closed**, no air can pass and no sounds are created. In the **vibrating** state, the vocal folds vibrate and create "voiced sounds". Finally, the glottis can be wide **open** (as in normal breathing) where "voiceless sounds" can be produced. During production of an utterance (a sequence of vowels and consonants), articulators change the configuration of the vocal tract to generate the resonances that are associated with each sound. Therefore, it can be viewed that production is simply a continuous stream of configuration specifications without any specific phoneme or time boundaries. These specifications are commonly called articulatory gestures, or simply gestures. Gestures are basically an instruction to the motor system for an articulator to achieve a particular constriction in the vocal tract. An example of an articulatory gesture is the closing of the glottis or a raising of the tongue body. However, these gestures do not occur in linear order and overlap one another greatly.

2.3 Coarticulation Phenomena

Perceptually, we consider speech as a sequence of discrete units called phonemes. Each of these phonemes have a specific and different set of articulatory requirements. These discrete units are

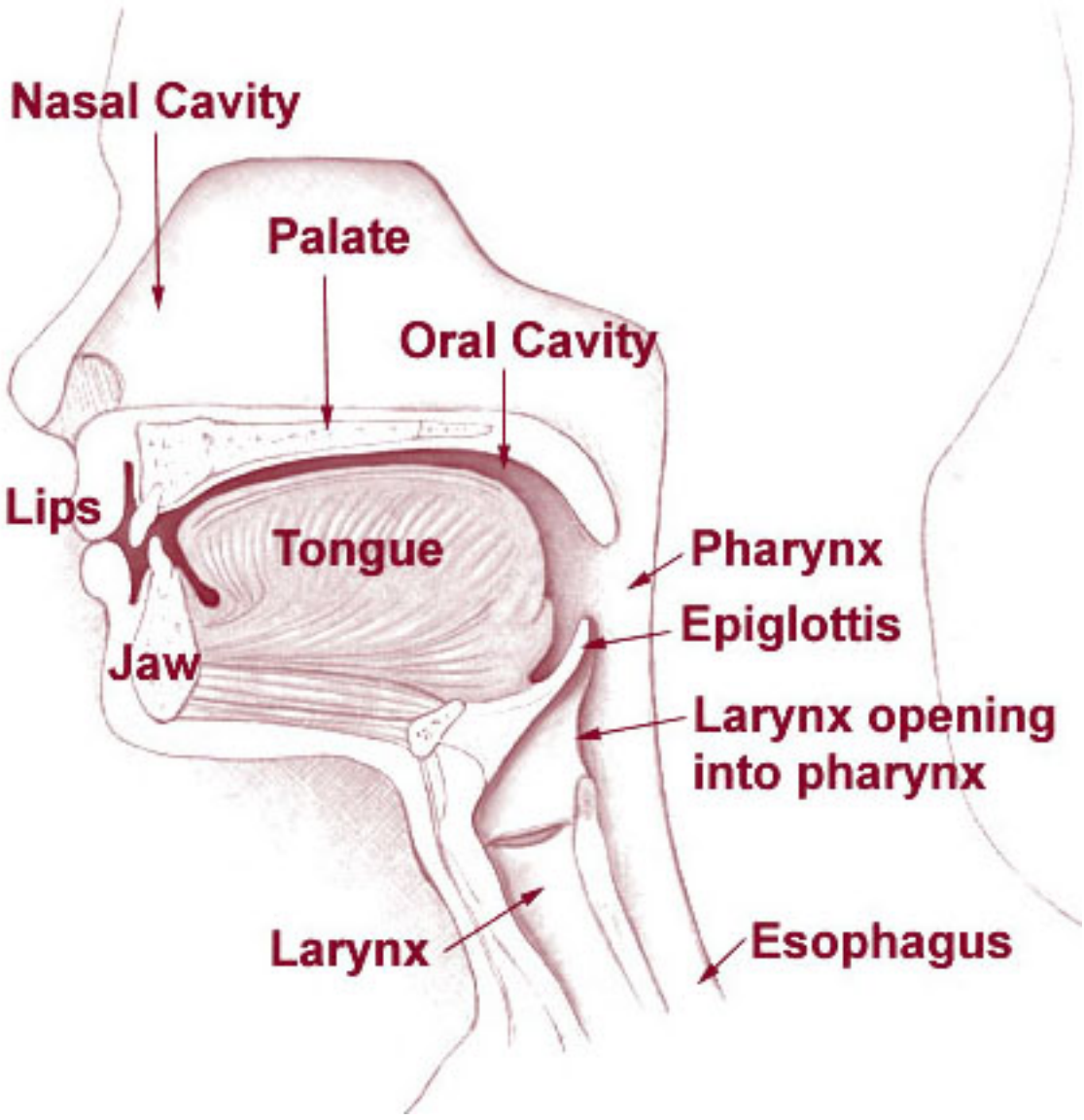


Figure 2.1: Human vocal apparatus used to produce speech. Illustration by Arcadian, 2007

combined into a continuous acoustic stream. As the speech system is unable to change states or location instantly, the overlap of articulatory gestures of neighboring phonemes is inevitable. Coarticulation is the mechanism underlying this overlapping of articulatory gestures from neighboring phonemes. Despite this complexity, however, there is much known about coarticulation [57, 68, 81, 83, 100, 111]:

- Dependency is largely confined to immediate neighbors, but can extend to more remote neighbors.
- It is plausible, that the "degree of articulation" of a phone heightens its coarticulatory impact on neighbors. The phrase "degree of articulation" refers to the precision and speed of the articulators with varying degrees of effort, from "conversational" (relaxed) to "clear" (precise). Generally, but not necessarily, faster speech has a lower degree of articulation. [111].
- The overall coarticulatory pattern depends on the duration of phonetic segments, with shorter segments generally being more influenced by their neighbors.
- Within a phonetic segment, the influence of a neighbor generally increases as one moves away from the phone center to the neighboring phone boundary.
- Some phonemes exert more coarticulatory influence than others. For example, /w/ has a stronger effect on the following vowel than /z/.
- The acoustic manifestation of a phoneme is also influenced by non-phonemic factors, such as prosodic and paralinguistic factors (speech rate, pitch, loudness, fluency, etc.).
- Coarticulation involves multiple articulators. Acoustic manifestations of coarticulation depend on the degree to which these articulators can move independently as well as on their target positions.

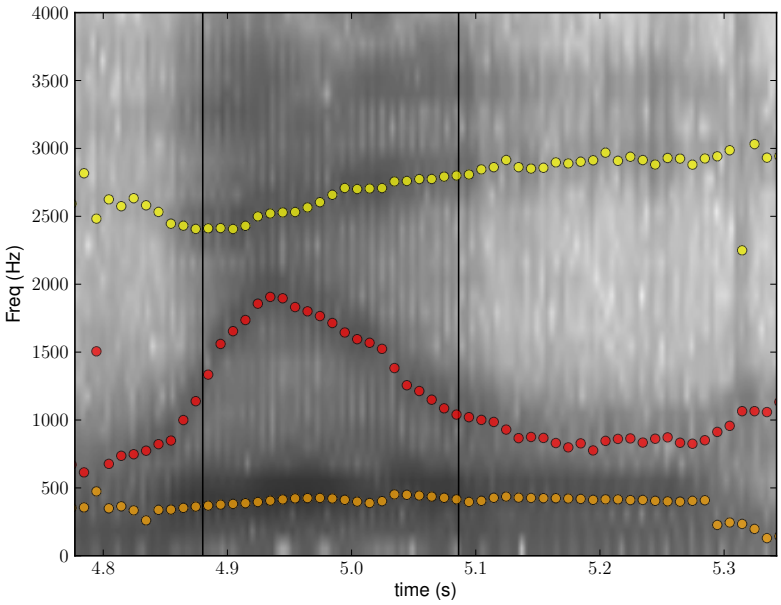
Formants are the spectral energy peaks in speech that show the resonant properties of the vocal tract. Formants provide reliable information of speech, as they reflect the dynamic changing filter properties. However, the mapping between formant structure and perception is complicated and lacks a direct one-to-one mapping from acoustic signal to perceived phoneme. In fact, there is a general agreement in the literature that the variability and unsegmentability of speech is due primarily to the universal phenomenon of coarticulation [35]. The primary attraction of using formants in modeling is that the first two formants can be directly related to the articulatory

movement of the tongue. The first formant is proportional to how "low" the tongue, whereas the second formant is proportional to the "frontness" of the tongue. For example, the vowel /iy/ (as in b**ee**t) is both high and front. These characteristics of the first two formants allows acoustic data to generate articulatory models, which we will use in this work.

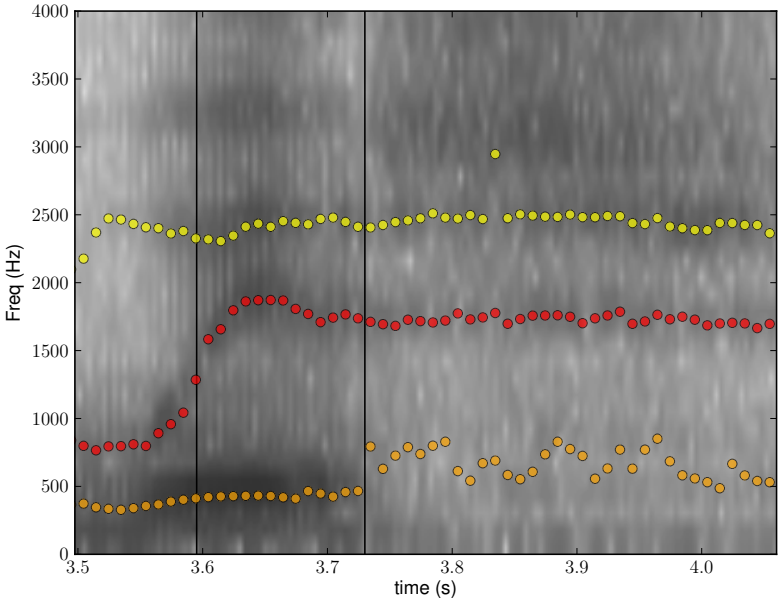
When the words 'will' and 'wish' are spoken, a listener will perceive and categorize the initial consonant as an invariant /w/ phoneme. However, if examining the spectrogram it will be difficult to precisely segment the /w/ from the following vowel. A series of phoneme are not simply concatenated together as we perceive them. See Figure 2.2a and Figure 2.2b for spectrograms of "will" and "wish". If we were to examine a spectrogram for each of the aforementioned words, we would note that features of /w/ would vary between contexts. Basically, there are no acoustic invariants that consistently characterize the /w/ and similarly for other consonants. As much as the vowel affects the consonant, the consonant affects the vowel. Potter and Gordon [90] showed early spectrographic evidence of this and noted that "when vowel sounds are preceded or followed by the consonants, the consonant usually exerts an influence that appears as a blending of the vowel bars (formants)." This acoustic variation is due to coarticulation. Sussman *et al.* introduced the "theoretical puzzle known as the non-invariance problem" [101] that defined the "litmus test" for invariance as the place of articulation in a stop consonant followed by a vowel since these utterances are typical of coarticulated speech. Since the work by Sussman *et al.*, investigations into phonetic invariance has looked primarily on coarticulated CV segments with the point of focus being place of articulation (formants).

Lindblom [69] explained phonetic variation with his H&H theory (hyperspeech and hypospeech) wherein varying degrees of articulatory effort are used in different circumstances. Lindblom basically postulates that a speaker assesses the needs of a listener and balances the effort required in producing speech against the needs for producing speech which is sufficiently discriminated. Thus the speaker modifies the articulation in response to situational and environmental demands along a continuum of hyperspeech and hypospeech. Hyperspeech is defined as carefully articulated speech, which we will refer to as CLR. Individual phonemes are more easily identified in hyperspeech, since the variance in production of speech sound is less and the effect of reduction and coarticulation is minimal. In contrast, hypospeech is basically speech with limited effort and more variance in speech sounds with greater phonetic reduction and coarticulation.

One view, is that coarticulation is a type of reduction. This is manifested as a reduced displacement (basically an undershooting of the abstract acoustic target) and a shift in movement towards the surrounding context, which is done to ease articulation [76, 69]. Basically, coarticulation is the result of a lack of articulatory effort, much in the same way that a reduced vowel might be thought



(a) Formants for "w-ih-l" (will) spoken in the CLR style



(b) Formants for "w-ih-sh" (wish) spoken in the CLR style

Figure 2.2: Examples of coarticulation

of in contrast to a carefully produced vowel. This minimization of effort is only reverted when the environment (or listener) pushes the speaker to produce clearer speech in order to maintain intelligibility.

However, this view does not derive naturally from a coproduction theory of coarticulation [38, 39, 8] or the closely related model of gestural phonology [16]. Coarticulation is not explicit in both theories, but rather a temporal and spatial overall of gestures. Reduction can be represented as a decrease in the amplitude or size of a gesture and occurs in connected speech through an adjustment in the overall duration of a series of gestures. However, hyperarcticulation can be represented as an increase in movement of a gesture. If the amplitude of a gesture is increased without an increase in the overall time of articulation, there will be a greater overlap of adjacent gestures, which in turn results in greater coarticulation.

Target undershoot occurs when there is insufficient time for an articulator to reach its target position target position. If an articulator has too much inertia to move between two targets in the available time it doesnt travel as far as it would if there was more time. Target undershoot can occur in both vowels and consonants. As we discussed earlier, articulatory gestures overlap, which means that the movement of an articulator occurs at the same time as the movements of another articulator.

2.4 Coarticulation Theories

Coarticulation models can be divided into two primary groups: (1) "look-ahead model" (also called anticipatory coarticulation) by Henke [46], and (2) the "time locked coarticulation" model (also called **coproduction**) by Bell-Berti and Harris [8]. The look-ahead model addresses the issues from a purely phonological standpoint, where each phoneme is an bundle of articulatory features. When a phoneme has a specific gesture specified, i. e., [+ROUNDED], the speech system as a whole moves toward the goal of lip rounding in advance of the phoneme that is marked [+ROUNDED]. This model suggests that the central motor control of speech is constantly seeking goals that are changing. Thus, the system is scanning and analyzing utterances in advance and initiates the required gestures toward an articulatory target as soon as there are no conflicting requirements on the specified articulators. In contrast, the "time-locked coarticulation model" suggest that articulation is context-independent and the articulatory gestures overlap. This means there are no scanning mechanisms in the speech motor control, but a simpler sequencing of articulator feature requests (or configurations). Coarticulation in this model is the simple result of temporal overlap between independent articulatory gestures associated by neighboring segments. The overlap in

between configurations is not a movement towards some goal in a priori, but an unavoidable consequence of limitations on articulators in modifying their configurations. In between these two groups is a hybrid model of Perkell and Chiang [86] which decomposed anticipation into two phases: the first depending on phonological constraints and the second one depending on the dynamics of the articulators.

2.5 Coarticulation Modeling

Somewhat parallel to these studies about anticipation are computational models have been developed to describe the whole coarticulation phenomenon. We will examine coarticulation modeling in three distinct, but overlapping areas: phonetic sciences, speech synthesis and automatic speech recognition.

2.5.1 Phonetic Sciences

Early work by Öhman [81] focused on tongue movements and suggested that fast consonant gestures are superimposed on slow movements of the tongue dorsum required to realize vowels. Locus equations [68] attempted to solve the non-invariance problem. Locus equations are a linear regression fit to data points formed by plotting onsets of F2 transitions along the y-axis and their corresponding mid-vowel nuclei along the x-axis. Locus equations capture the F2 transition acoustic variability and perceptual invariance at the categorical level [101]. These linear equations are:

$$F2_{onset} = k \cdot F2_{vowel} + c \tag{2.1}$$

where the coefficients k and c are slope and intercept, respectively. The slope parameter is generally interpreted as indicating place of articulation in addition to the extent of CV coarticulation [64].

In the past few decades, experiments using locus equation metrics have provided some consistency and order in the production and perception of stop consonants across various vowel contexts. Sussman and Shore write "The vowel-context induced variability of the F2 transition, so evident when individual tokens are compare, gives way, when displayed as locus equation plots, to lawful acoustic representation of the entire stop place category" [102, p.101]. However, this relational invariance is limited to stop consonants and fricatives [51], but has not been generalized.

Outside of speech synthesis, much of the work has focused on understanding coarticulation through a single acoustic feature, namely the second formant (F2) and have not addressed the behavior of multi-dimensional features. The concept of the virtual formant became prevalent in the form of locus equations [68, 20, 14, 101, 12], which attempted to measure the extent of

coarticulation and quantify coarticulation overlap. Locus equations define a linear relationship between the F2 values of different vowels measured at the start of the consonant-vowel transition and the vowel center. This linear relationship predicts F2 values of a consonant from an F2 value of a vowel, whereby the consonant is fixed and the vowel varies. The slopes and intercepts of these lines vary systematically as place of articulation of the consonant varies [65]. F2 frequency differences are larger at the vowel center and move proportionally to the same location at the transition start, suggesting that they may converge to a single value (or range) inside the consonant. This value is the virtual F2 value of the consonant. These linear relationships have also been shown to hold for the first formant (F1) [14].

Besides locus equations, several models of coarticulation have been proposed. Öhman [82] developed a model of coarticulation in vowel-consonant-vowel (VCV) utterances. This model was expressed in terms of the vocal-tract shape by the formulation:

$$s(x,t) = v(x) + k(t) \cdot [c(x) - v(x)] \cdot w_c(x) \tag{2.2}$$

where $s(x,t)$ is the shape of the vocal tract at position x and time t, $v(x)$ is the vocal tract shape corresponding to a given vowel, $c(x)$ is the vocal-tract shape of the consonant, $k(t) \in [0,1]$ and is an interpolation term and finally, $w_c(x)$ is a term that defines the amount by which $c(x)$ resists coarticulation. This model was quite successful in describing context-dependent variations of vocal-tract shapes using context independent descriptions of vowels and consonants. However, this model is only capable of modeling VCV utterances and Öhman stated the difficulty in modeling coarticulation between consonants, such as consonant-vowel-consonants (CVC) utterances.

Lindblom's [68] first model found a direct relationship between vowel duration and the amount of undershoot in the first three formants. The experiment described in this study focused on the speech of a single male speaker of Swedish and involved CVC words involving the eight short (lax) vowels in three consonantal contexts (/b/, /d/ and /g/ – that is bxb, dxd, gxg, where x is a short vowel). This produced 24 CVC words which were produced in four carrier phrases constructed to vary the stress on the target word. The resulting vowel duration had a large variation between 80 and 300 ms. In this range of durations, undershoot increased considerably from long to short durations. That is, as duration decreases, undershoot increases. This study showed that the process of vowel reduction is continuous – and most importantly, that this process is not towards vowel centralization, but the effect of consonant-vowel coarticulation. This was demonstrated in that the formant frequencies (in the vowel) vary as a function of vowel duration and consonant context. In other words, as duration decreases the formants tend to undershoot the acoustic target. Lindblom [68] introduced the concept of an acoustic target, an ideal context-free configuration which, in the

case of vowels, is represented by the asymptotic values towards which formant frequencies aim. Limblom's study showed that targets are quite often not realized.

Lindblom was able to predict the values of the lowest two formants (F1 and F2), which were a function of vowel target formant frequency, consonantal context and vowel duration. The core idea of this model is that the shorter the vowel, the more influence the surrounding consonant. Thus, the shorter it is, the more the observed formant frequencies differ from the target formant frequencies. This combination of consonantal articulatory influence with a decreased vowel duration is vowel undershoot, as there will be insufficient time to achieve the vowel target. The general formula for F2 takes the following form [68]:

$$F_o = k \cdot (F_i - F_t) \cdot e^{-a \cdot DUR} + F_t \tag{2.3}$$

where F_o is the steady-state frequency of the second formant at vowel mid-point of a CVC, F_i is the initial value of F_n at the start of a vowel, F_t is the ideal vowel target (one per vowel and independent of consonantal context) for F2, DUR is the vowel duration, $DUR > ln(k)/a$, k and a are fixed per symmetric consonant environment. Note that this is formulated for symmetrical consonantal contexts. With this model, half of the variance in the data could be explained. This method of vowel production is referred to as the target-undershoot model because it assumes that the articulators (and thus the formants) generally fail to reach their ideal target at the vowel midpoint.

Lindblom [68] also showed that the degree of formant undershoot is dependent on speaking style, word stress, vowel duration and neighboring consonants. More recent work by Moon and Lindblom [76] presented a revised model of vowel undershoot. Vowel duration was the primary factor; attributes related to speech style impact the amount of formant undershoot. Results demonstrated that vowel duration and F2 formant frequencies of clear speech vowels tend to be longer and less reduced than in conversational style. Furthermore, the authors report that clear speech was shown to be characterized by higher formant velocity than conversational speech.

Broad and Clermont [13], in an effort to describe formant dynamics, developed several models of formant transition in CV and CV/d/ contexts, where the prevocalic consonant was limited to voiced stops. Their more elaborate model used a linear combination of of coarticulation functions (modeled with exponential functions) and target values:

$$F = \alpha(t) \cdot (T_{C_1} - T_V) + T_V + \beta(t) \cdot (T_{C_2} - T_V) \tag{2.4}$$

where $F(t)$ is the estimated formant vector as a function of of time t, T_V is the target formant

vector of the vowel, T_{C_1} and T_{C_2} are the target formant vectors of the initial and final consonants, respectively. The first term represented the formant transition from consonant C_1 to vowel V. This coarticulatory effect is proportional to the target difference and scaled by a coarticulation function $\alpha(t)$. The last term models a similar effect of the consonant C_2 on the vowel V, with $\beta(t)$ the coarticulation function. If we let $\delta(t) = 1 - \alpha(t) - \beta(t)$, then Eq.2.4 becomes:

$$F = \alpha(t) \cdot T_{C_1} + \delta(t) \cdot T_V + \beta(t) \cdot T_{C_2} \tag{2.5}$$

which shows that the estimated formant vector of the vowel at any point in time is simply a linear combination of the target formant vectors of the phonemes C_1, V and C_2. This formulation models target functions as a vector, combining F1, F2 and F3, while the coarticulation functions, α and β are scalars. Therefore this models formant movement synchronously.

In an application of Broad and Clermont's CVC model, Niu and van Santen [79] applied the model to a broader set of consonants with the goal to measure coarticulation in dysarthric speech. Amano-Kusumoto and Hosom [4] further improved the upon Niu and van Santen's approach by modeling the entire formant trajectory of the vowel. Furthermore, in cases where the prevocalic or postvocalic consonant was an approximant, the modeled region was extended into the consonant nucleus. The coarticulation function was modified from an exponential to a sigmoid function to restrict the coarticulatory effects to be smoothly and monotonically increasing or decreasing, specifically for cases where the consonant being modeled is an approximant (/y,l,w,r/). Vowels were limited to monophthong vowels.

The aforementioned methods and models highlight the complex nature of coarticulation and the difficulties in constructing models. In addition, those models were limited to at most CVC triphones (in several models, the consonants are limited to a specific class) and thus have limited applicability.

2.5.2 Speech Synthesis

Analysis of formant trajectories dates back to the 1940s, with the first availability of spectrograms. Synthesis and perception experiments documented the direction and extent of the second formant (F2) in signaling stop place [67]. Later work in phonetics using an articulatory-acoustic model within the acoustic theory of production [100] also found acoustic relations to place but could not explain the variability shown in empirical studies. A result in deriving rules for synthesis was the concept of "locus" (also referred to a formant target), which was defined as *a place on the frequency scale at which a transition begins or to which it may be assumed to "point"* [23], which differed

from the locus equation paradigm first set forth by Lindblom [68] and popularized by Sussman *et al.* [101]. This research was strongly connected to early speech synthesis which generated speech using equations for formant trajectories [62]. Formant analysis was dominant until the late 1980's when concatenative synthesis became feasible.

Atal [5] introduced a low bit-rate coding technique for analyzing speech, called "temporal decomposition". Within temporal decomposition, coarticulation of speech can be represented by a linear combination of (articulatory) targets. Here we consider a target as the planned location of an articulatory gesture, which is the movements of the lips, tongue and other articulators. Assume an utterance of speech was produced by K movements, which move to and from K articulatory targets. The linear combination that approximates the observed speech, $y(n)$, can be represented as follows:

$$\hat{y}(n) = \sum_{i=1}^{K} w(k)\phi_k(n) \tag{2.6}$$

where $\phi_k(n)$ is a target vector for a "phone-like entity" k and $w(k)$ is a target vector and movement defined by a target function for the k^{th} target.

As Atal's intent was to develop an economical coding method, no attempt was made to understand the relationships between discovered target functions and the underlying phoneme. This approach resulted in several targets per phoneme. Kappers and Marcus [30] extended Atal's work with the primary intent to discover a one-to-one mapping from labeled phoneme to articulatory target directly from speech. This approach resulted in not all phonemes having targets with vowels being more stable than consonants and bursts of plosives problematic. However, this approach implies the correspondence of target functions and underlying acoustic events.

In Text-to-speech (TTS), the Asynchronous Interpolation Model (AIM) was developed to represent units for TTS with an explicit model of coarticulation [55, 54, 56]. The components of AIM are basis vectors, which represent phoneme target values, and weight trajectories, which describe how the basis vectors (or phonemes) change over time. Coarticulation is modeled by overlapping weight trajectories. The different rates at which articulation changes (e. g., fast changes in manner of articulation with slow changes in tongue position) are modeled by different sets of weight trajectories. However, the weight trajectory is a linear mapping between zero and one, which would assume that coarticulation is a linear event.

2.5.3 Automatic Speech Recognition

Most current ASR systems model coarticulation by way of Wickelgren's theory [110] wherein speech units are coded as context-dependent units. For example, the phoneme /ae/ in the word "bat" is

represented using a unit that is entirely different from the phoneme /ae/ in the word "tab". In Hidden Markov Model (HMM) based ASR systems, the states are associated with these context-dependent units. According to Wickelgren, "By assuming (context-sensitive) allophones to be the basic unit of articulation, ... it is trivial to account for how the "same phoneme" in different phonemic some respects at all levels of the speech process" [110]. Though conceptually simple, this model has several disadvantages. First, the number of units required to represent all phonemes in all contexts would require substantial amounts of training data. Second, the Wickelgren model assumes that coarticulation is limited to adjacent phonemes. Kent and Minifie demonstrated that coarticulation *may* be present at a distance of up to six phonemes from the primary phoneme [57]. If Wickelgren's model is expanded to account for contexts over longer phonetic durations, an exponential number of units and data would be required.

A number of approaches to the problem of better modeling of coarticulation were proposed by Deng [29, 25, 26] and others [94, 61]. Deng presented the "trended HMM" [25] which uses the property that speech parameters change over time in a predictable manner is captured by polynomial equations that describe how spectral features will change over time within a phoneme. The trended HMM may be viewed as an extension of a semi-Markov Model [93], in which the duration of each phoneme is explicitly modeled. This improved ability to model both duration and speech dynamics has not yielded significantly better performance than the use of standard context-dependent phonemic units. Possible reasons for lack of improved performance despite a more structured model may include:

- The polynomials may be modeling speaker-dependent attributes, influences of speech rate, or background or channel noise, in addition to coarticulatory effects.
- Several polynomials are used in succession to describe feature dynamics within a phoneme. As there is no boundary constraints for these polynomials, it allows for large and possibly unrealistic discontinuities in the features within a phoneme.

In other work by Deng, coarticulation in ASR is modeled not at the level of phonemes, but at the level of distinct phonetic features (such as nasality or voicing) from which phonemes are composed [29]. In this distinctive-feature approach, the features associated with the phoneme are not required to change synchronously at the phoneme boundary, but the change in features may overlap, or be asynchronous. This approach embodies a model of coarticulation that allows the phonetic features of one phoneme to influence the phonetic features of a neighboring phoneme. Such distinctive-feature overlap is an important component of coarticulation, but Deng's model only addresses coarticulation at this high level. In particular, because Deng's model operates at

the level of distinctive features, there is no specification of how coarticulation affects the speech signal at the level of acoustic features. Such modeling at the acoustic level is evidently critical, as the overlapping feature approach did not perform as well as standard context-dependent HMMs on the task of phoneme classification on the TIMIT [42] corpus.

Deng has also proposed the use of locus equations within a speech-recognition framework [26]. However, despite the inclusion of an explicit model of formant trajectories, the proposed model assumes that all vowels do attain their target values. Such an assumption is not implied in the locus models used in TTS, and in fact it is well known that many vowels do not attain their target values, especially at moderate to fast speaking rates. In addition, Deng used a single locus equation to model only the second formant, thereby constraining the coarticulation model to a single phonetic feature, namely the place of articulation.

The proposed model was trained and evaluated only on consonant-vowel-consonant (CVC) regions containing stop consonants, which is a highly constrained environment. Results of evaluation showed that the "locus HMM" consistently outperformed the baseline HMM system, but not state-of-the-art HMMs. It is quite possible that the weaknesses of assuming that each vowel attains its target values and limiting the locus equation to only the second formant were the primary factors preventing performance superior to state-of-the-art systems. In addition, Deng acknowledged that robust formant-tracking would be a "significant challenge for our future research"[26] in the development of a complete ASR system.

Other work has led to the development of a "Hidden Dynamic Model" [94]. In this framework, the concept of target values that are modified by coarticulation is explicitly modeled. However, as in the trended HMM, the filter used to smooth the target values over time is not constrained to follow known properties of speech. In addition, coarticulation is assumed to occur synchronously for all speech parameters within a frame of speech, when in fact, the properties of manner, place and height of articulation all are quasi-independent and vary asynchronously.

2.6 Problems with Previous Modeling Approaches

In summary, most prior work in modeling coarticulation has resulted in models that explain how the speech signal is influenced by phonetic context. However, modeling has been largely limited to short sequences and/or limited phonetic context. There is an expectation that data-driven algorithms can extract relevant properties from speech data. However, this lack of explanatory power results in the need for large corpora for training in applications such as speech synthesis and automatic speech recognition. This limitation restricts how well the models fit observed data, since

the models lack predictive power, they are unnecessarily complex, rely on manual intervention and focus on coarticulation in limited contexts at the expense of a more general effective solution.

To address these limitations, we introduce a data-driven coarticulation model that models the formant frequency of continuous speech. This model uses a linear combination of global acoustic event-specific formant targets and local coarticulation parameters. These global targets are speaker-specific, but independent of speaking style and phonemic context. We show that these targets can be automatically derived from continuous speech, including classes of sounds where formants are not directly observable. We demonstrate that this model captures necessary information that is necessary for intelligible speech. We also explore further applications of this model in (1) clear and conversational speech analysis, (2) formant tracking, and (3) dysarthria diagnosis.

Chapter 3

Coarticulation Model for Continuous Speech

In this chapter, we present a methodology that models formant trajectories of continuous speech.[1] The model represents continuous trajectories as a combination of overlapping local trajectories, which are represented by a weighted-addition of acoustic event targets by sigmoidal coarticulation functions (or activation functions in some studies) characterized by slope and position. Estimation is achieved by a brute-force search over global acoustic event targets and local context-dependent and utterance-specific coarticulation parameters.

As a first step of validation (Section 3.5), we evaluate the parameter estimation method by creating a synthetic corpus with F2 speech trajectories created with known targets. Using these contrived trajectories, we estimate targets and coarticulation parameters using our estimation method.

Next, we estimate both acoustic event targets and coarticulatory parameters on a continuous speech corpus (Section 3.6) in two distinct experiments: (1) all acoustic event targets are estimated, and (2) only sonorants event targets are estimated while other consonants are fixed. We investigate coarticulatory parameters and compare estimated targets with previously published targets.

Finally, as a second step of validation, in Section 3.7, we present results from two perceptual evaluations that were conducted by means of a speech intelligibility test. The primary goal was to test whether resynthesis from model parameters produces speech as intelligible as vocoded speech using observed formants, along with natural speech as a control. Using the both sets of targets estimated in Section 3.6 and coarticulation parameters, we re-create speech trajectories to produce synthesized speech. We show that our modeled speech was nearly as intelligible as speech using observed formant trajectories, thus validating the model, its estimation procedure

[1] Portions of this chapter was published in Bush and Kain [19]

and demonstrating that the model is capturing the necessary components of speech necessary for intelligibility.

3.1 Introduction

Previous work [6, 18] investigated fully-automatic and comprehensive estimation of formant frequency targets for a single male speaker, but was limited to modeling CVC triphones. This approach has been expanded to modeling coarticulation in continuous speech. In this work, we focus on modeling formant frequencies, but this modeling approach could be applied in modeling formant bandwidths and energy. In the model herein we use the term "acoustic target" or simply "target" as the acoustic invariant of a phonological feature. It is a theoretical construct that may not be precisely realized, especially if there are nearby (or conflicting) targets.

While in principle the proposed model could be applied to any interpolatable feature representing the speech signal, such as line spectral frequencies (LSF), we use traditional formant frequency (F1–F4) trajectories as speech features. Since formants constitute a concise acoustic representation that is closely aligned with a specific vocal tract configuration, we will be focusing our investigation on formant trajectories. In addition, we assume a phoneme segmentation is provided, which could be derived through forced alignment [49]. Because some phonemes contain more than one acoustic event, we map the phoneme segmentation to an acoustic event segmentation $e_1, e_2, \ldots, e_N$ with event boundary beginning and ending times $b_1, b_2, \ldots, b_{N+1}$. Most phonemes map to one acoustic event, but diphthongs, affricates, and released stops map to two events (e. g., closure and release for released stops). This flexibility in mapping could be expanded for phonemes that have distinct regions, such as unvoiced plosives into three events: the closure, burst, and aspiration, e. g., /t/ = /t$_c$/, /t$_b$/ and /t$_h$/. For convenience, we also define event center times $c_i = (b_i + b_{i+1})/2$, $i = 1, 2, \ldots, N$.

In our modeling, the acoustic event targets and coarticulation parameters are unknown parameters and estimated from speech data. In our approach, we hypothesize that these targets are global, whereas the coarticulation parameters are local and context-dependent. The global target hypothesis assumes that regardless of speech style, rate, etc. we are using the same scalar targets. This implicitly states that required flexibility in speech resides in the coarticulation parameters and are local to a specific triphone.

Beyond the model structure, nothing is assumed about these parameters aside from constraints. Thus, we can assume that their estimated values will provide an unbiased view into coarticulation. The advantages of a parametrized formant trajectory model are: (1) allows analysis of formant

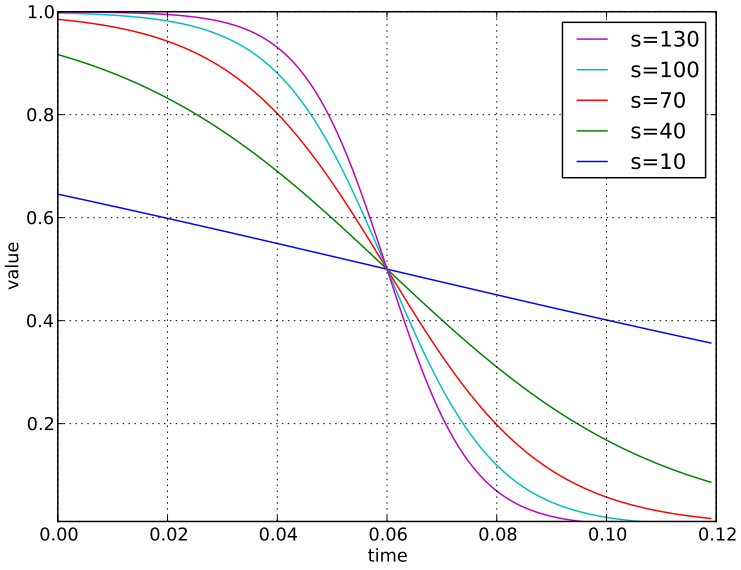


Figure 3.1: Example of σ coarticulation function for discrete values of s, spanning 0 to 120 ms, with p value at 60 ms.

targets separately from the coarticulation effects, (2) allows parametrization of the coarticulation effects, and (3) allows synthesis of formant trajectories with provided phoneme duration.

3.2 Local Coarticulation

During analysis, an individual continuous feature trajectory $X(t)$ is divided into consecutively overlapping local regions comprising three acoustic events such that the local feature trajectories $x_i(t)$ are defined for the interval $t \in [c_{i-1}, c_{i+1}], i = 2, \ldots, N-1$; for example, $x_2(t)$ spans from the center of event e_1 to the center of event e_3. This local trajectory is modeled as

$$\begin{aligned} \hat{x}_i(t) \quad = \quad & f_{i,L}(t; s_{i,L}, p_{i,L}) \cdot T_L + \\ & f_{i,C}(t) \cdot T_C + \\ & f_{i,R}(t; s_{i,R}, p_{i,R}) \cdot T_R \end{aligned} \tag{3.1}$$

with $L = e_{i-1}$, $C = e_i$, and $R = e_{i+1}$. Eq. 3.1 is a convex linear combination of T_L, T_C, and T_R, representing global feature *target* values for the left, center, and right acoustic event type, respectively, weighted by $f_{i,L}(t)$, $f_{i,C}(t)$, and $f_{i,R}(t)$, representing local coarticulation functions for the left, center, and right acoustic events, respectively, in this local event context. For all contexts,

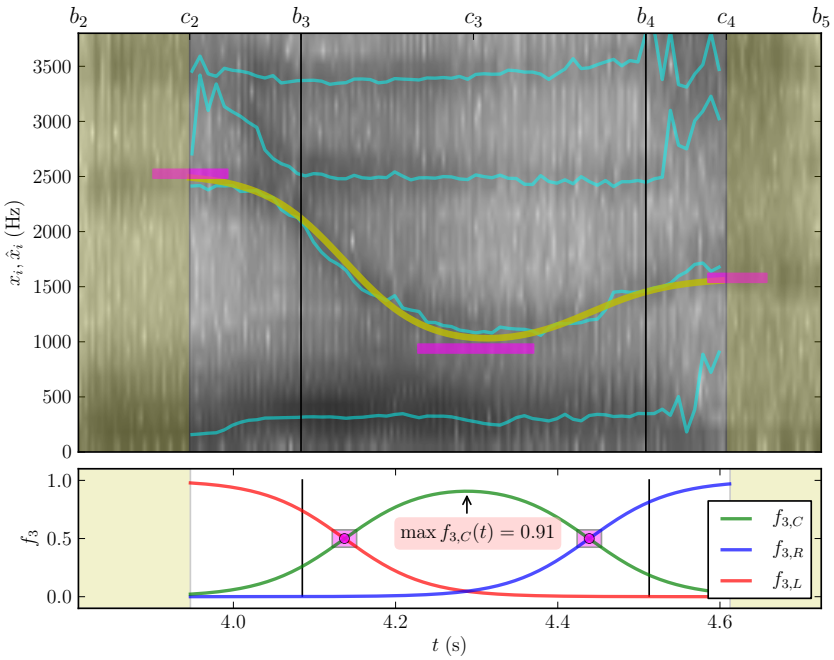


Figure 3.2: Local context example using the word "use" (/y-uw-z/). Upper panel shows observed local formant frequency trajectories $x_3(t)$ (cyan), model F2 formant frequency trajectory $\hat{x}_3(t)$ (yellow), and model formant frequency targets T (magenta bars). Lower panel shows coarticulation functions with sigmoid centers p (magenta circles), and sigmoid slopes s (diagonals of magenta boxes).

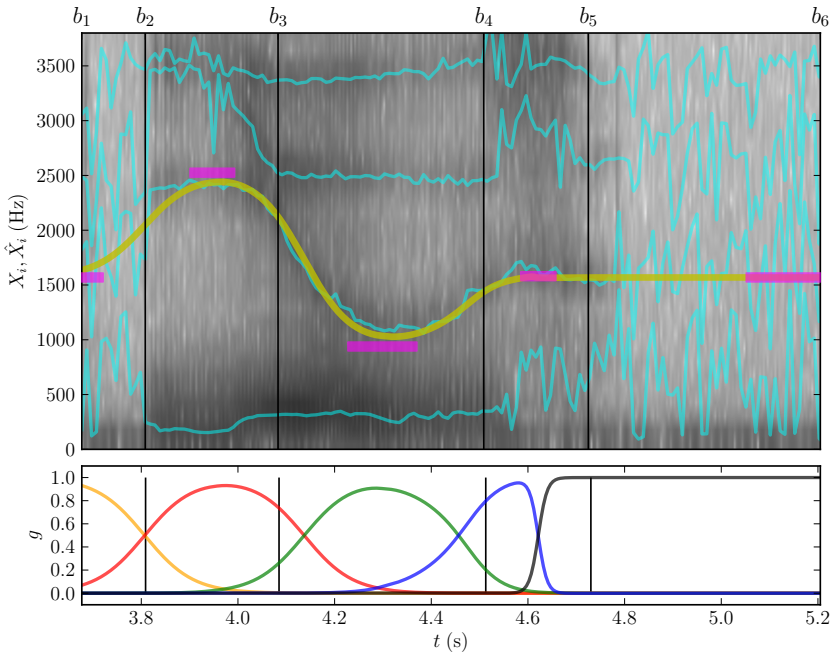


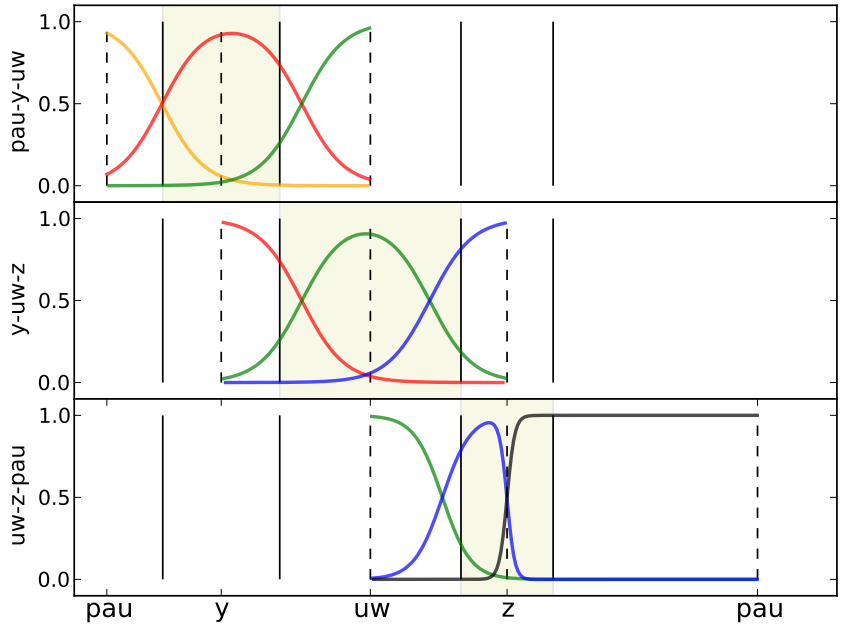
Figure 3.3: Continuous context example using the word "use" (y-uw-z) that shows formant frequency trajectories $X(t)$, as well as $\hat{X}(t)$ for F2.

coarticulation functions, f, are based on the sigmoid $\sigma(t; s, p) = (1 + e^{s \cdot (t-p)})^{-1}$, which can be seen in Figure 3.1 with varying values of s parameter, where

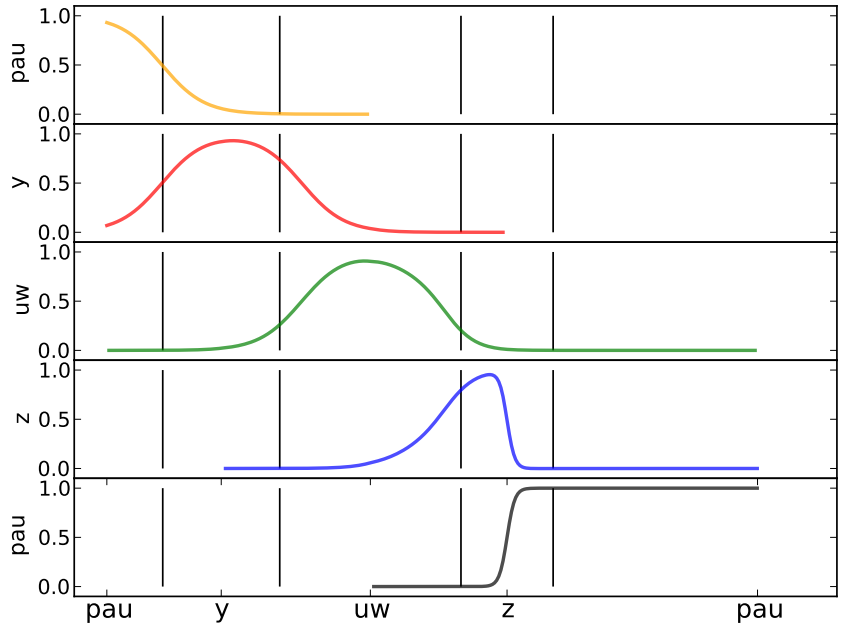
$$\begin{aligned} f_{i,L}(t; s_L, p_L) &= \sigma(t; s_L, p_L), \quad s_L > 0 \\ f_{i,R}(t; s_R, p_R) &= \sigma(t; -s_R, p_R), \quad s_R > 0 \\ f_{i,C}(t) &= 1 - f_{i,L}(t; s_L, p_L) - f_{i,R}(t; s_R, p_R) \end{aligned} \tag{3.2}$$

where $\{s_L, s_R\}$ represent sigmoid *slope* (slow versus fast transition, from the left to center acoustic event and from the center to right acoustic event, respectively), and $\{p_L, p_R\}$ represent sigmoid midpoint *position* (and point of maximum slope), measured relative to their respective phoneme boundaries (b_i and b_{i+1} for the i^{th} context). Figure 3.2 illustrates the concept.

We assume the existence of a single target value T for each acoustic event type; thus they are defined globally (their number depends on the number of distinct acoustic event types). Furthermore, we assume that slope parameters s_L and s_R are shared among contexts of the same type (i. e., different tokens with identical acoustic event labels). Finally, position parameters p_L and p_R are modeled separately for each token.



(a) local coarticulation functions in all contexts



(b) global coarticulation functions

Figure 3.4: Example illustration of local and global coarticulation functions for F2 formant frequency of the word "use" of the male speaker. The dotted lines indicate the centers of the acoustic events and solid lines their boundaries. Coarticulation functions are uniquely colored per acoustic event.

3.3 Continuous Coarticulation

In a second step, local coarticulation functions modeling identical acoustic events are cross-faded to form global, continuous coarticulation functions

$$g_{e_i}(t) = \begin{cases} t < c_{i-2} & 0 \\ c_{i-2} \leq t < c_{i-1} & \alpha(t) \cdot f_{i-1,e_i}(t) \\ c_{i-1} \leq t < c_i & \alpha(t) \cdot f_{i,e_i}(t) + \beta(t) \cdot f_{i-1,e_i}(t) \\ c_i \leq t < c_{i+1} & \alpha(t) \cdot f_{i+1,e_i}(t) + \beta(t) \cdot f_{i,e_i}(t) \\ c_{i+1} \leq t < c_{i+2} & \beta(t) \cdot f_{i+1,e_i}(t) \\ t \geq c_{i+2} & 0 \end{cases} \tag{3.3}$$

where $\beta(t) = 1 - \alpha(t)$, for $i = 3, 4, \ldots, N-2$, where for each case, $\alpha(t)$ is a distinct cross-fading function defined to be zero at the left side of the interval, and monotonically increasing to one at the right side of the interval. This operation can be thought of as a smooth transfer of knowledge about a particular acoustic event from one local model to the next. For the second case in Eq. 3.3, the $i - 2^{\text{nd}}$ local model contains no information about e_i, and thus we are cross-fading with zero, equivalent to a fade-in; analogously for the fifth case and the $i + 2^{\text{nd}}$ local model, where we are performing a fade-out. For $i = 1, 2, N-1, N$ the formula is modified to handle these edge cases correctly. Note that cross-fading preserves convexity, i. e.

$$\sum_{i=1}^{N} g_{e_i}(t) = 1,\ t \in [c_1, c_N] \tag{3.4}$$

For an example illustration refer to Figure 3.4. The final model representation of the continuous feature trajectory is calculated via

$$\hat{X}(t) = \sum_{i=1}^{N} g_{e_i}(t) \cdot T_{e_i} \tag{3.5}$$

implementable by matrix multiplication for discrete t (such as consecutive synthesis frame times). An example is shown in Figure 3.3. For a sequence of N acoustic events we require $N - 2$ local triphone models.

3.4 Estimating Formant Frequency Parameters

In this section, we detail the error function used along with parameter constraints and the model parameter estimation method.

3.4.1 Error Function

In order to estimate optimal coarticulation and target parameters for a specific trajectory, we define a per-utterance root-mean-square error (RMSE)

$$G = \sqrt{\frac{1}{\sum_{t=0}^{N} w(t)} \sum_{t=0}^{N} w(t) \cdot \left(X(t) - \hat{X}(t)\right)^2} \tag{3.6}$$

where X and $\hat{X}$ are the observed and estimated trajectories, N is the number of frames. The weighting factor $0 \leq w(t) \leq 1$ indicates our confidence in the formant observation, which is derived from acoustic event class-based weights detailed in Table 3.1. These values were determined through trial-and-error and knowledge of formant accuracy in specific classes. In previous work [18], we used formant bandwidth alone to derive this weight, where a wide bandwidth provided little confidence in formant measurement. We use RMSE as our error measure since it is measured in the same units as the data (Hz), rather than in squared units, and is representative of the size of a "typical" error.

Acoustic Event	Weight
Vowel	1.0
Approximant	1.0
Nasal	0.6
Remaining acoustic events	0.01

Table 3.1: Acoustic event class-based weights, which represent our confidence in the accuracy of formant observations.

3.4.2 Constraints

For valid and acoustically correct estimation, we apply the following constraints:

- For each triphone $f_L + f_R \leq 1$
- $10 \leq s_L \leq 140$ and $10 \leq s_R \leq 140$
- The range of p_L is from the center of event e_{i-1} to the center of event e_i for the $i - 1^{\text{th}}$ triphone

- The range of p_R is from the center of event e_i to the center of event e_{i+1} for the $i-1^{\text{th}}$ triphone
- $200 < F1 < 1000, 400 < F2 < 2800, 900 < F3 < 4000, 3000 < F4 < 4000$
- $F2 - F1 > 200, F3 - F2 > 200, F4 - F3 > 200$; minimum distance between neighboring formants is 200 Hz

3.4.3 Parameter Search

Model parameters are estimated for a single formant using a nested brute-force search with the outer search being over a valid target range and the inner search on the utterance level searching a subset of coarticulation parameters. The outer target search can be more efficiently search using a steepest ascent hill-climbing technique with three restarts and a maximum of 200 evaluations. Using multiple restarts simulates a parallel search, similar in principle to the way population based algorithms function, which helps avoid local minima in multi-modal solution spaces [95, 48].

For simplicity, we concatenate a series of utterances to form a single utterance. Using this utterance and a specific set of initial targets, we will iterate through all phonemes several times ($maxiter$). In each of these passes, we randomize the phoneme order. For a specific phoneme, we sweep the target space over the prescribed interval for the lowest error. At each target value, we determine the coarticulation parameters for the entire utterance that leads to the lowest overall error (in the Optimize call). This error trajectory over the target space are referred to as E_{sweep}, with an example of /w/ for F2 in Figure 3.5. This target for the specific phoneme is now used in further estimations. This algorithm is detailed in Algorithm 1. The algorithm for Optimize, which uses a specific set of targets and finds the set of coarticulation parameters with lowest overall error is described in Algorithm 2.

Targets can be initialized using (1) observed or (2) published formant values such as those used in the Klatt synthesizer by Allen *et al.* [2]. In experiments conducted within this chapter, targets are initialized to mean observed formant measured at center of the acoustic event. Coarticulation parameters are initialized with s set to 120 and p set to the acoustic event/phoneme boundary (0.0). The bounds F_{min}, F_{max} are formant specific bounds that are defined in Sec 3.4. The minimum for formants higher than F1 is defined as $\max(F_{min}, F_{i-1} + 200)$ for any formant i.

We perform this search for formant frequencies F1, F2, F3 and F4 separately and in that order. The ordering ensures the minimum formant bound (for formants higher than F1) are at least 200 Hz larger than the previous formant maximum. We will validate this search algorithm in Sec 3.5 by constructing artificial trajectories from known targets and estimating the known targets.

Algorithm 1 Brute-force target estimation

for $it = 1$ **to** $maxiter$ **do**
  Randomize phoneme order
  for p in phonemes **do**
    for $t = F_{min}$ **to** F_{max} **do**
      Optimize(t)
      $t \leftarrow t + jump$
    end for
    Store best target for p
  end for
  Note error
end for

Algorithm 2 Optimize coarticulation parameters for an utterance

Require: global targets

for $i = 0 \ldots N-1, N, N-1, \ldots 1, 0, 1, \ldots N$ **do**
  if $i = 0$ **then**
    Search $s_{i,L}$ and $p_{i,L}$
  else if $i = N-1$ **then**
    Search $s_{i,R}$ and $p_{i,R}$
  else
    Search $s_{i,R}$, $p_{i,R} = p_{i+1,L}$, and $s_{i+1,L}$ {General case}
  end if
end for

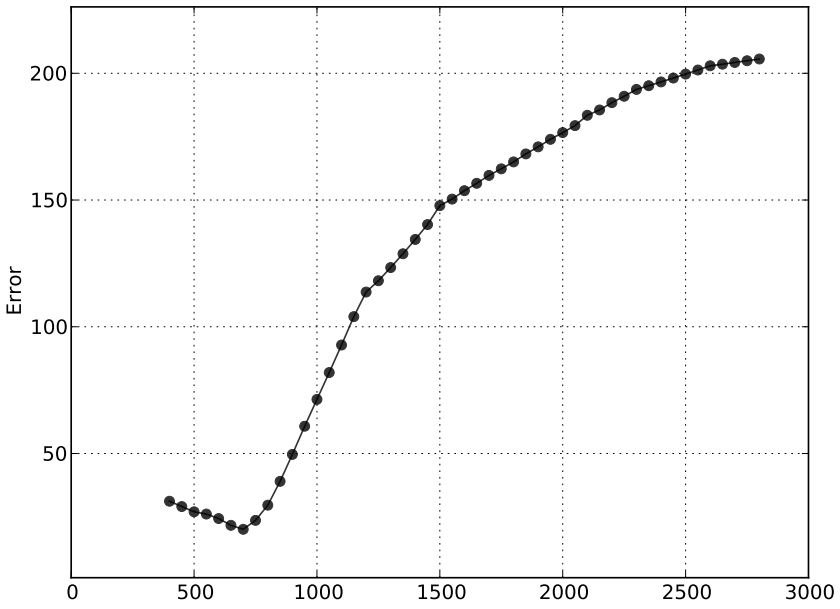


Figure 3.5: Example E_{sweep} of /w/ for F2

3.5 Estimation Validation

This investigation is to validate the our estimation procedure can recover formant frequency targets from a speech trajectory. We create synthetic trajectories from known targets with the goal to rediscover these targets. We consider the validation successful if we recover the original targets or if there are minor differences in discovered targets with small overall error.

3.5.1 Corpus

The corpus was derived from actual speech of 70 sentences from a male speaker recorded in the CLR style. When recording, he was instructed to speak as clearly as possible, as if communicating with a hearing impaired or cognitively impaired listener and asked to "enunciate consonants more carefully and with greater effort than in conversational speech and avoid slurring words together" [45]. For this experiment we are concerned only with the formant F2. Using realistic formant targets based on acoustic-phonetic knowledge, we automatically estimated local coarticulation parameters to fit the extracted F2 formant trajectory. We added noise to the trajectories with zero mean and standard deviation of $\pm$ 15 Hz. The formant targets used are are listed in Appendix B.

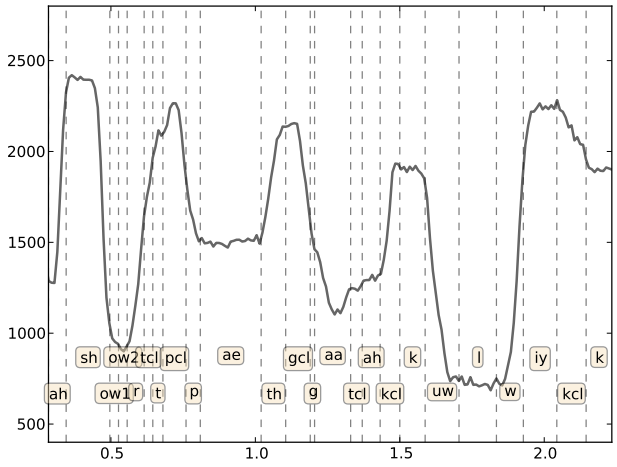


Figure 3.6: Example synthetic trajectory

3.5.2 Results

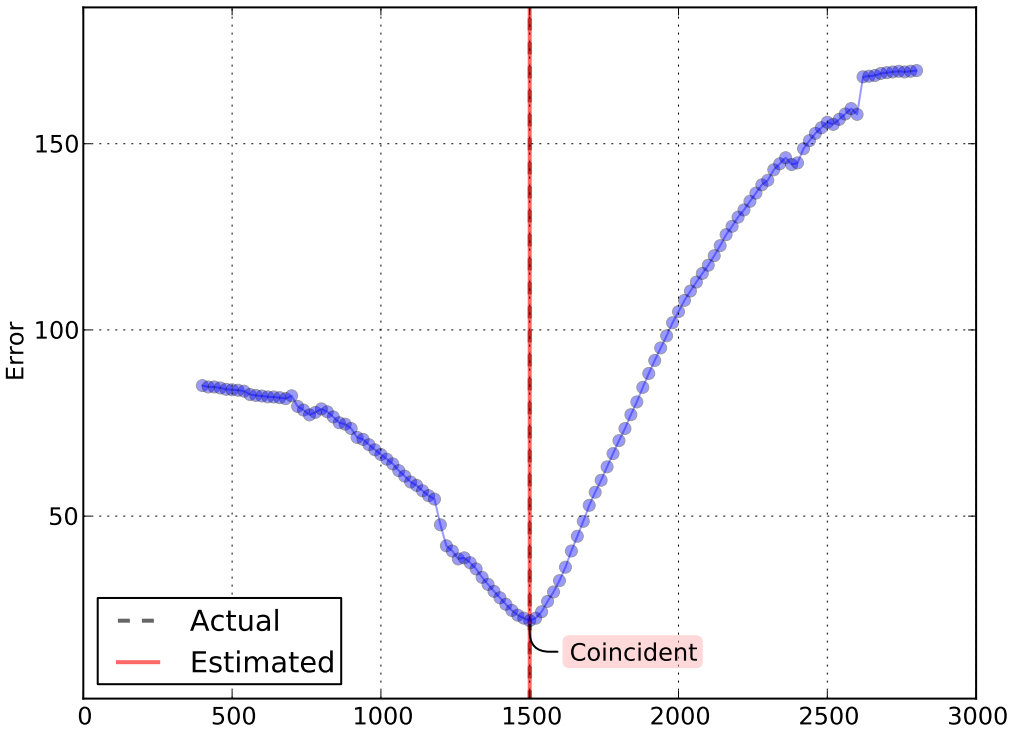
We examine the recovery of coarticulation and target parameters separately. In both cases, we don't achieve exact targets and parameters, due to both near-optimal errors and with the addition of noise to the original trajectories.

Coarticulation Parameter Recovery

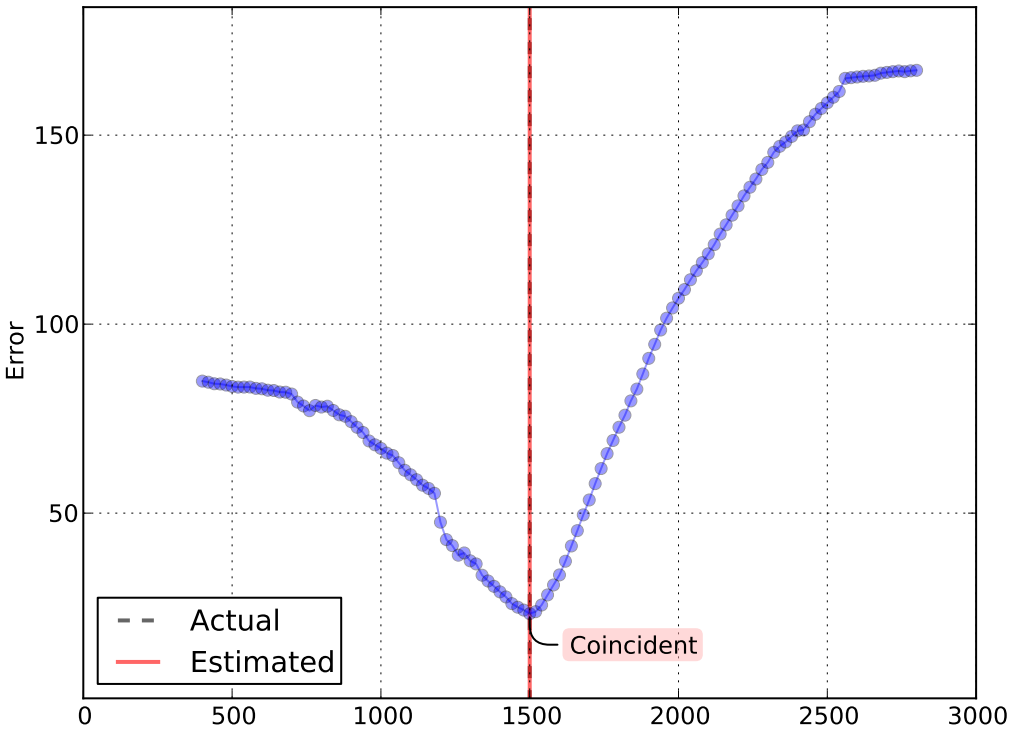
We compared the coarticulation parameters that were estimated with those that were used in creating the speech trajectories. The seventy sentences required 2643 local models which requires 5286 s parameters. For a local model, we break down the three acoustic events into two separate diphones, e. g., e_1, e_2, e_3 become $e_1 - e_2$ and $e_2 - e_3$. In our analysis, only diphones that had target differential of at least 300 Hz and did not contain a pause were considered, which resulted in 2810 parameters. Of these parameters, 2040 (or 72.6%) were exactly matching, with the remaining 770 (or 27.4%) non-matching. The non-matching s-values are broken down by 499 of $s = 30$, 159 of $s = 60$, 99 of $s = 90$ and 13 of $s = 120$. The largest class of non-matching was $s = 30$ and the 499 components are mostly composed of closures where formants are largely non-existent. Table 3.2 details the errors in coarticulation parameters and provides examples of the most frequent diphones per group.

Target Recovery

Using observed targets, we have 23 acoustic event targets that match with the original targets. The acoustic events that were missed In Table 3.3 we detail the RMS errors during various iterations

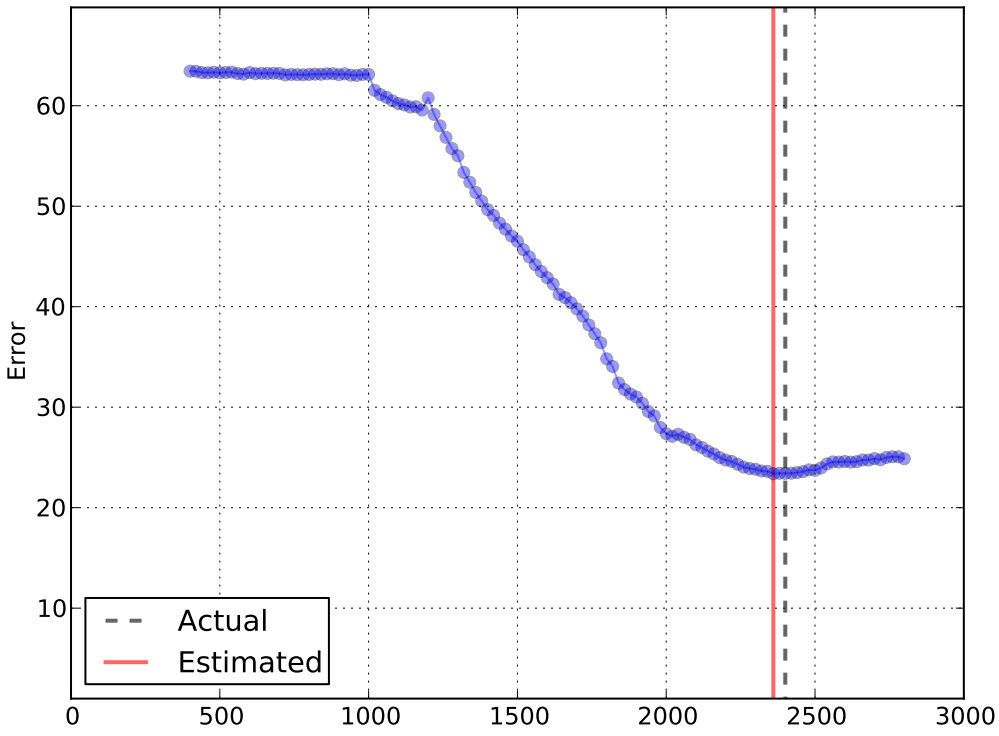


(a) E_{sweep} for /ae/ for observed targets

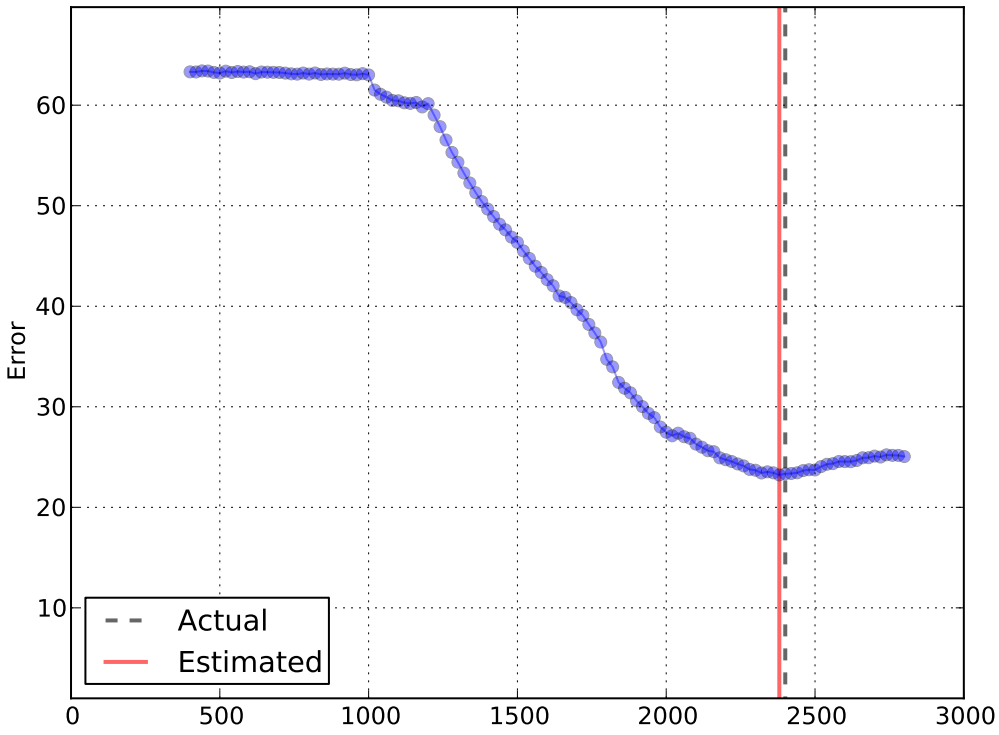


(b) E_{sweep} for /ae/ with targets after first iteration

Figure 3.7: Example E_{sweep} of /ae/ for F2 after one iteration



(a) E_{sweep} for /sh/ for observed targets



(b) E_{sweep} for /sh/ with targets after first iteration

Figure 3.8: Example E_{sweep} of /sh/ for F2 after one iteration

Difference in s value	Total Occurrence	Examples
30	499	pcl-p (25), bcl-b (22), dh-ah (20), s-tcl (18)
60	159	dh-ah (11), pcl-p (7), bcl-b (7), s-tcl (6)
90	99	s-tcl (6), pcl-p (5), d-er (4), er-dh (4)
120	13	n-dcl (2), ay2-n (1), ow2-dcl (1), t-s (1)

Table 3.2: Example diphones that occur frequently with differing coarticulation (s) parameters from original. Examples with frequency in parenthesis.

Target Set	RMS Error	Targets Matched
Actual	21.8	-
Observed	33.2	23/50
Iteration 1	22.3	34/50
Iteration 2	21.3	37/50
Iteration 3	21.1	39/50

Table 3.3: RMS errors of different target sets with the synthetic corpus along with the count of targets where estimated equals the actual.

parameter estimation. In addition, the error using the actual targets that was used to construct the corpus. The error using observed targets is also noted, which is the starting point for the estimation process. Examining the results from the first iteration, we note a significant reduction in both RMS error and target RMS error. Further iterations lead to reductions in RMS error measure along with the number of targets matched, however, the target RMS error increases in iteration two. We hypothesize that the parameter estimation process is over-fitting and fitting the noise that has been applied. This example shows that we can achieve proper targets in a single iteration, that are significantly better than using observed targets alone. It should be noted that our RMS error greater than zero because the trajectories had noise applied.

In Table 3.4 we detail the acoustic event targets that do not match precisely with their actual targets. We can see that the majority are closures and plosives, which are difficult to estimate especially in the presence of noise. We will investigate specific acoustic event error surfaces of /ae/ and /sh/ using both observed targets and targets from the first iteration, in Figure 3.7 and

Acoustic event	Actual Target	Estimated Target	Difference
b	1100	1050	50
ch	2500	2450	50
chcl	1950	1850	100
d	1900	2000	100
dcl	1800	1950	150
er	1350	1300	50
f	2000	1950	50
jhcl	2000	2100	100
p	1700	1750	50
pcl	2300	2350	50
s	2600	2550	50
sh	2400	2350	50
t	2100	2000	100
tcl	2150	2200	50
v	1800	1900	100
z	2100	2150	50

Table 3.4: Sixteen acoustic events that were not exactly matching in estimated target and actual after the first iteration.

Figure 3.8, respectively.

3.5.3 Discussion and Conclusions

In both the coarticulation and acoustic event targets, the majority of parameters were recovered. In particular, targets were matched in most cases with the cases where there were slight discrepancies, the overall error was nearly the same. In coarticulation parameters, almost 73% of the parameters were recovered. The remaining coarticulation parameters are for diphones that have low weighting. The reasoning for not exactly recovering the parameters overall is that the original trajectories had noise applied. Thus, we can assume that our parameter search method is sufficient.

3.6 Parameter Analysis

In this section, we analyze coarticulation and target parameters estimated from continuous speech in two separate experiments. The first experiment will estimate all acoustic events targets, while the second will only estimate vowel, nasal and approximant acoustic events targets while other consonants are fixed. First, we will discuss the corpus used in this study. Then we will examine results of both targets and coarticulation parameters. In examining target values, we compare both resulting estimated target sets from our experiments with known published values. Finally, we will examine coarticulation parameters with notable class-specific patterns in s parameters, context-specific relationships discovered and an investigation of asynchronicity in model highlighted by p parameters.

The first experiment (Experiment 1) estimates all formant frequency targets for each acoustic event. The initial values for these acoustic events uses the median observed value in the center of the event. The second experiment (Experiment 2) uses the same approach except several conditions: (1) All consonants not estimated are set to their respective values based on Allen's values [2], (2) closures are set to their respective consonants, e. g., /pcl/ is mapped to the value of /p/. (3) only sonorants and diphthong target values are estimated, (4) diphthongs were initialized to the observed mean values as these are not available in Allen *et al.* targets. Specifically, sonorants include monophthong vowels, approximants and nasals. Both experiments were estimated in the same manner, where formant frequency targets were estimated with a single iteration in target search as described in Section 3.4.

3.6.1 Corpus

The corpus was composed of 70 syntactically correct, but semantically anomalous sentences, e. g., *They slide far across the tiny clock*, created by randomizing and exchanging words and grammar structures from material described in [108]. All speech materials were scripted (read speech). One male, native speaker of American English with no professional training in public speaking recorded the same speech material in two different speaking styles: CLR and CNV.

The recording was carried out in a sound-treated booth (Whisperroom MDL4260) located inside a control room. Recordings were made using a head-mounted close-talking microphone (AKG HSC200), positioned approximately 5 cm and off axis from the speakers mouth. The speaker recorded the materials at his own pace by operating a computer program. A technician listened to each sentence and the speaker was asked to record a sentence again when pronunciation or style were not satisfactory. The speech signals were captured and stored digitally at a sampling

rate of 22.05 kHz with 16-bit resolution.

Initial estimates of phoneme identities and boundaries in each waveform were obtained using an existing forced-alignment system [49]. Then, a trained labeler checked and adjusted phoneme identities and boundaries manually. Formants were automatically estimated using a standard formant tracker [96, 103]. Formants were not manually corrected.

3.6.2 Results

To visually validate the accuracy of estimated targets we will examine observed formant and model trajectories in a per acoustic event manner. We will focus on vowels, approximants and nasals. In each plot, we present the raw observed formants (blue) and model (red) centered on the labeled acoustic event nucleus. The final estimated target for each acoustic event (from Experiment 1) is labeled with a magenta circle at time zero. Starting with front (Figure 3.9) and back (Figure 3.10) vowels we see a high amount of variation in the observed formants, even though most are consistent. Importantly, we see a high amount of variation in our model, which allows the model to fit the varied observations.

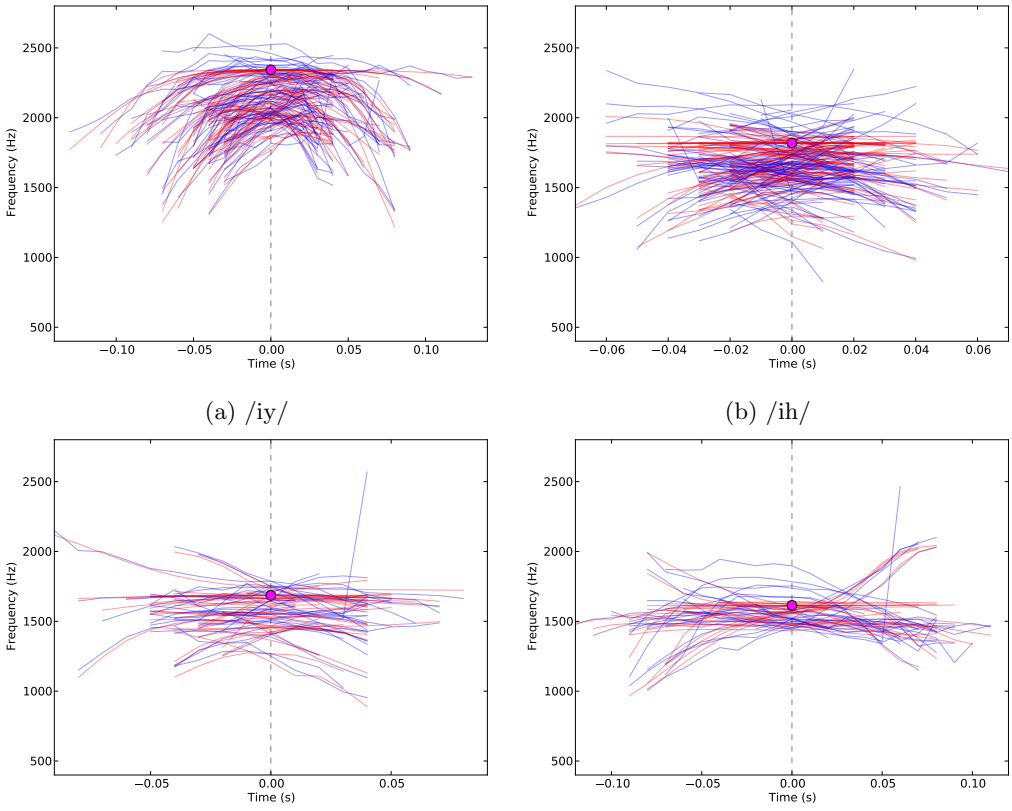
As we move from vowels to approximants, in Figure 3.11, we see much more variation in observed formant trajectories. Note that some phonemes are not as prevalent as others, e. g., /y/. In Figure 3.12, we see nasals have much more variation in the observed formant trajectory as expected.

The overall RMS error for all sentences in the corpus broken down by experiment in Table 3.5, where the second experiment with limited flexibility in target selection yielded overall less error. We hypothesize that the lower error is due to using more appropriate consonant targets that lead to a better fit by removing the tendency of sudden jumps in vowel-consonant transitions where consonant targets could be at upper (or lower) extreme values.

Experiment	**F1**	**F2**	**F3**
all	64	166	254
vowels, nasals and approximants	63	161	247

Table 3.5: RMS error of each formant overall dependent on scope of acoustic event estimation.

In Experiment 1, where we estimate all acoustic event targets, we compared our estimated targets with two sets of published formant targets: vowels and consonants. In Figure 3.13a, we compare our estimated vowel targets results with the steady-state vowel targets derived by



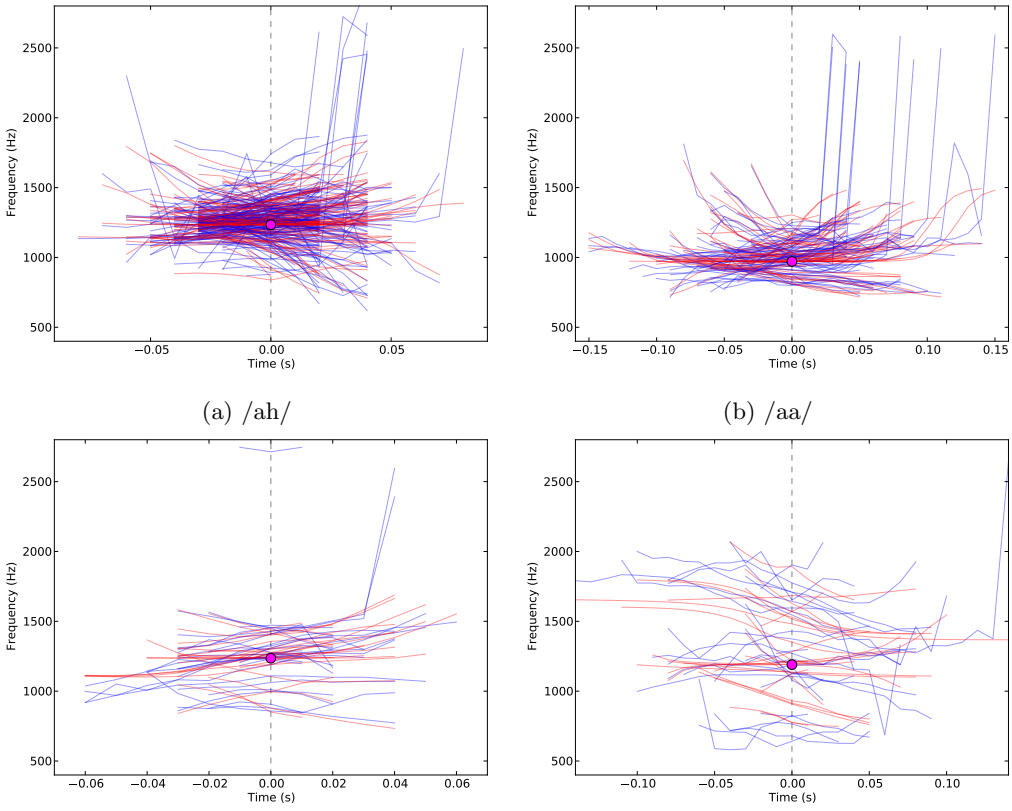
(a) /iy/ (b) /ih/

(c) /eh/ (d) /ae/

Figure 3.9: Observed and model trajectories for front vowels. Blue trajectories are observed formants and red are model trajectories.

Hillenbrand *et al.* [47] targets, which is a more recent study of the seminal work by Peterson *et al.* [87] study. In Figure 3.13b we compare our consonants with Allen *et al.* [2] targets that were estimated manually for speech synthesis. Note that diphthongs and affricates are not included here since we model those phonemes as two acoustic events and these values aren't available in either published dataset.

We can see the estimated vowels track quite well with the common front to back vowel pattern visible in F1 and F2. We note the highest variance in F2. The estimated consonants showed much more variation in comparison with Allen targets, with some acoustic events taking on target values in extreme positions, e. g., /h/, /th/ and /z/ in F3. We hypothesize that the reasoning underlying these high target values, is to simply reduce the error function since formant measurements in these regions are noisy. Furthermore, we see nasals /m/ and /ng/ F2 target values differing more than expected. However, approximants are as expected, except for F2 of /l/. It should be noted,



(a) /ah/ (b) /aa/

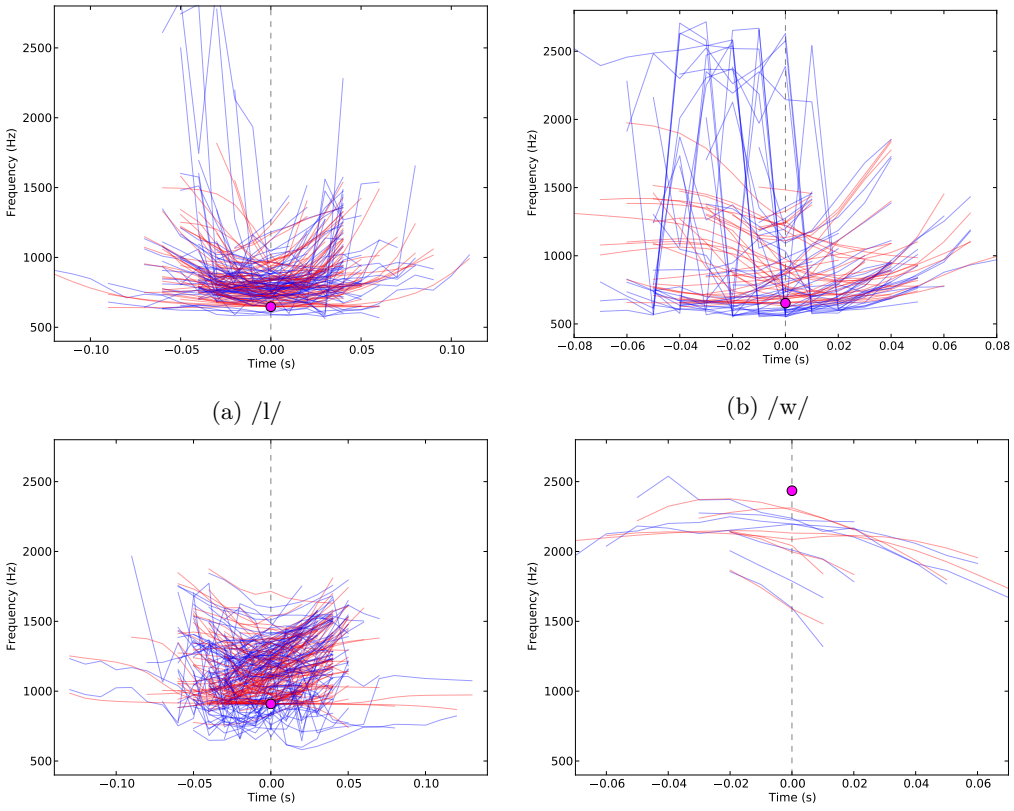
(c) /uh/ (d) /uw/

Figure 3.10: Observed and model trajectories for back vowels. Blue trajectories are observed formants and red are model trajectories.

that we expect some variation due to speaker-specific physical traits.

In Experiment 2, where only vowel, nasal and approximant acoustic events were estimated, we see more consistency in estimated targets. In Figure 3.14 we see less variation from Hillenbrand vowels, than in the first experiment. More importantly, we see that F2 the nasals and approximants are more similar to those defined by Allen with F1 remaining more or less unchanged. Specifically, one can see evidence of a velar "pinch" in /r/, where we see the second and third formants of vowels immediately before and after velar stops (e. g., /g/).

Evaluating the coarticulation parameters, we will first look at asynchronous movement in p and then trends in s parameters based on three common transitions. As a measure of asynchronicity, we calculated the median of the per-model standard deviations of p_L and p_R combined for F1-F3 to be 14.1 ms. In Figure 3.20 we see a histogram of the standard deviations. In Figure 3.19, we see the distribution of p values to be mostly normal and centered around zero. This variation



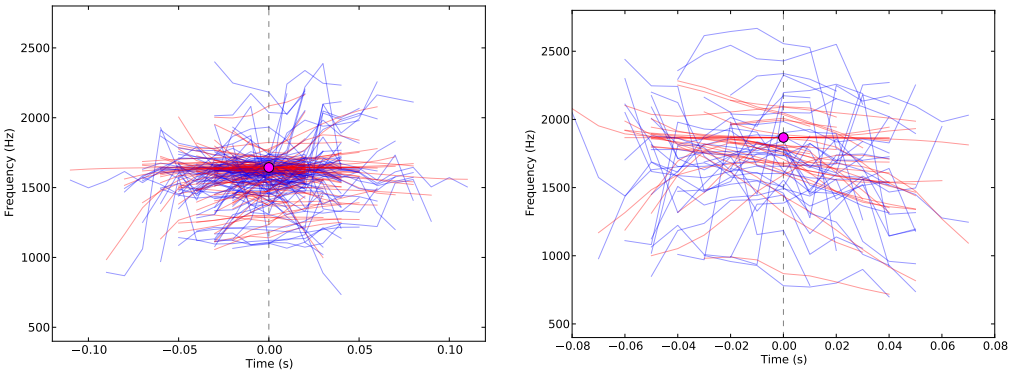
(a) /l/ (b) /w/

(c) /r/ (d) /y/

Figure 3.11: Observed and model trajectories for approximants. Blue trajectories are observed formants and red are model trajectories.

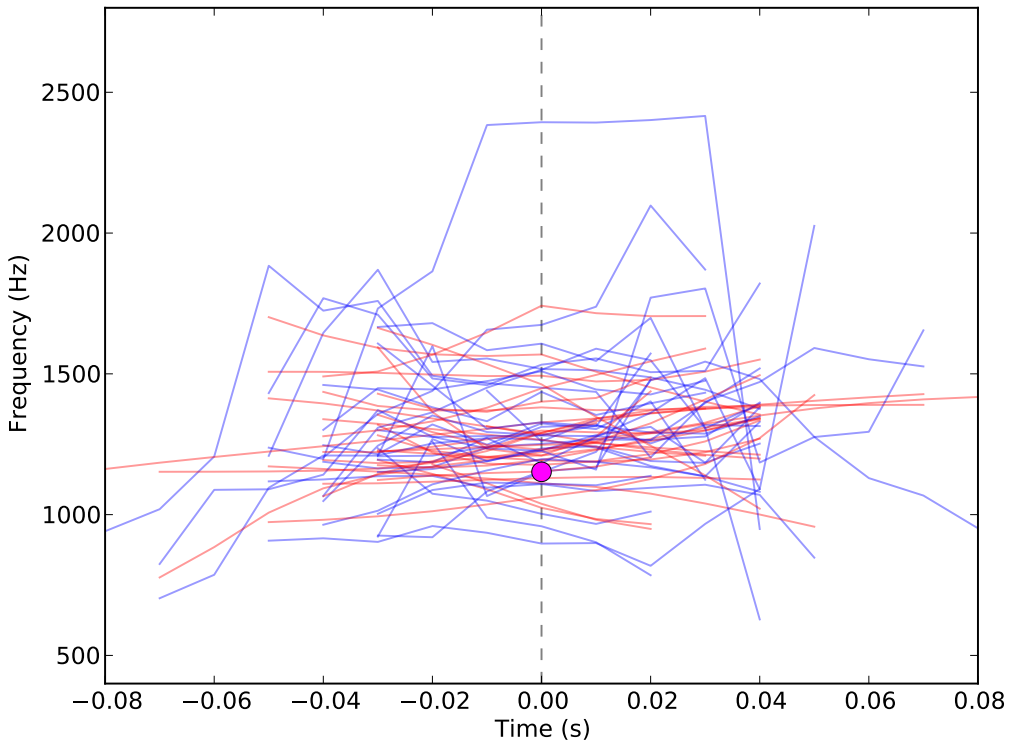
strengthens the case for asynchronous modeling that we present here.

We now examine a series of notable class-specific boundaries for F2 and the coarticulation values at that boundary. We can see in Figure 3.15 that we see two peaks at 40 and 130 where we would expect vowel to nasal transitions to be primarily high. There is a minor trend noticeable with as p values increase we see s values cluster around lower values (20 and 40) of s. Examining the underlying reasons, the target for both /m/ and /ng/ are surprisingly low. In Experiment 2, the target for /m/ and /ng/ were higher and more in-line with expected values. This minor change effects the s histogram to be primarily high, which can be seen in Figure 3.16. This falls in-line with our expectations as F2 formants in vowels typically have a sudden ramp into nasals. Similarly, fast vowel-nasal transitions, we see primarily fast transitions in fricative-closure transitions in Figure 3.18. In contrast to the former relationships, we can see in Figure 3.17 we



(a) /n/

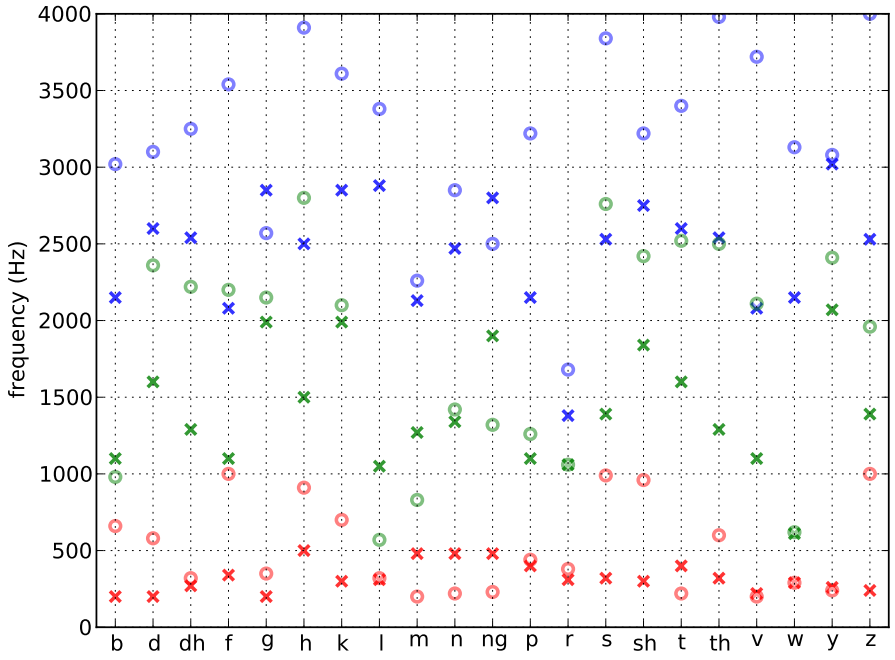
(b) /ng/



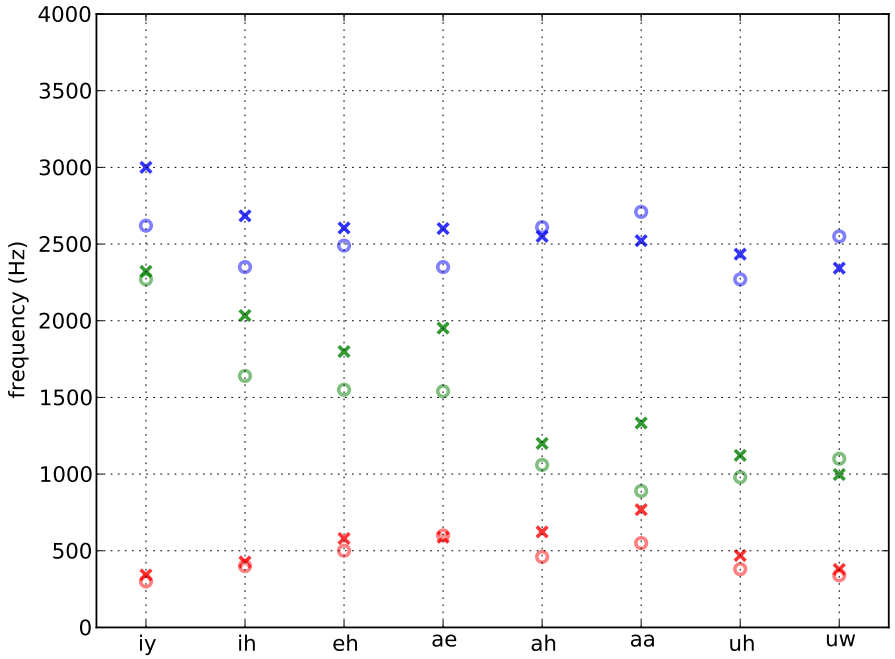
(c) /m/

Figure 3.12: Observed and model trajectories for nasals /m/, /n/ and /ng/. Blue trajectories are observed formants and red are model trajectories.

have an opposite relationship, with a majority of the s parameters low. This follows our acoustic-phonetic knowledge, where vowel to nasal transitions are usually rapid, and vowel to approximant are slow.

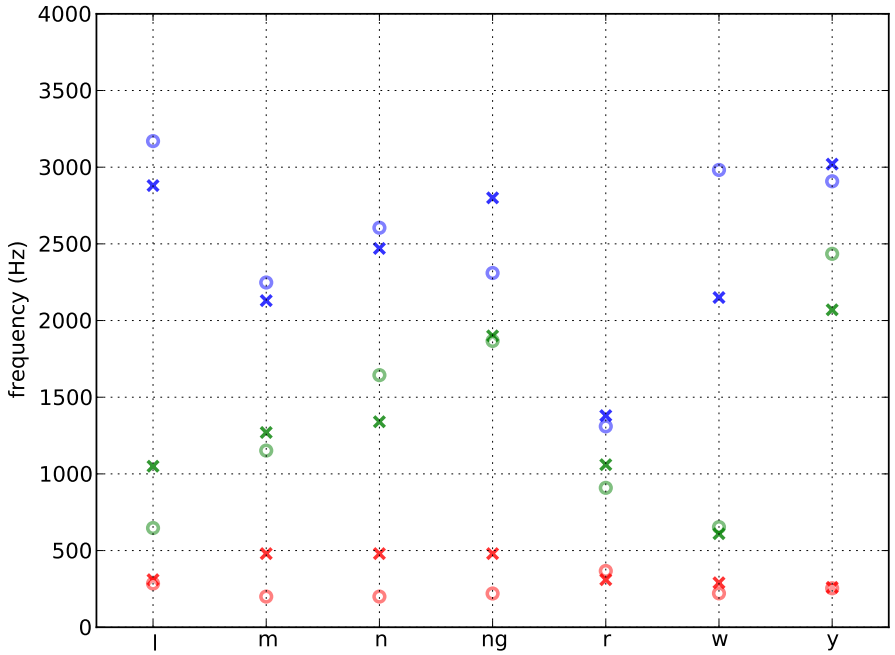


(a) Estimated consonant and semi-vowel targets (circles) and Allen (x). F1 (red), F2 (green) and F3 (blue).

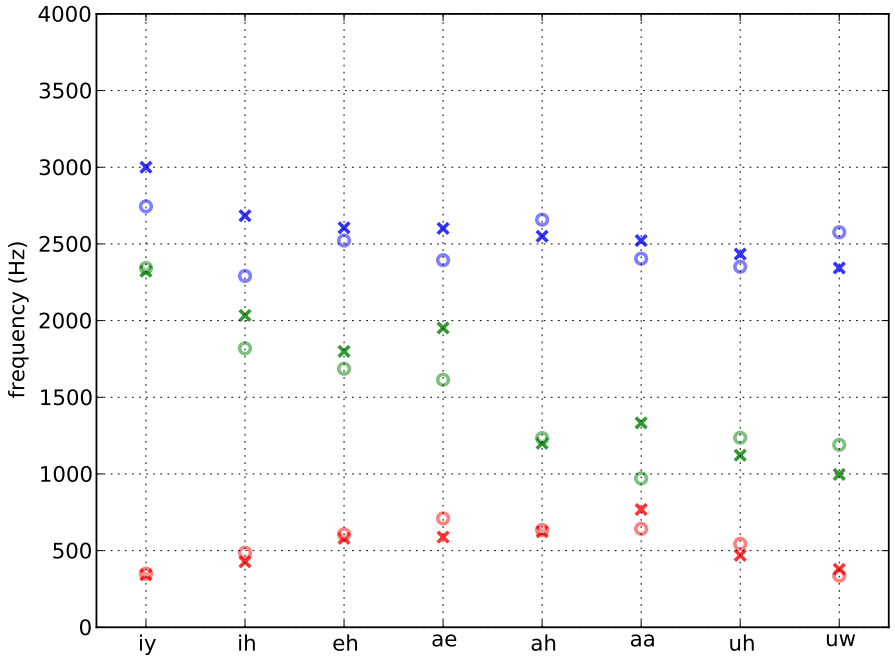


(b) Estimated vowel targets (circles) compared with Hillenbrand (x). F1 (red), F2 (green) and F3 (blue).

Figure 3.13: Global vowel and consonant target comparisons from Experiment 1.



(a) Estimated targets for nasals and approximants (circles) and Allen (x). F1 (red), F2 (green) and F3 (blue).



(b) Estimated vowel targets (circles) compared with Hillenbrand (x). F1 (red), F2 (green) and F3 (blue).

Figure 3.14: Global sonorant targets comparisons with remaining consonants fixed to Allen's values in Experiment 2.

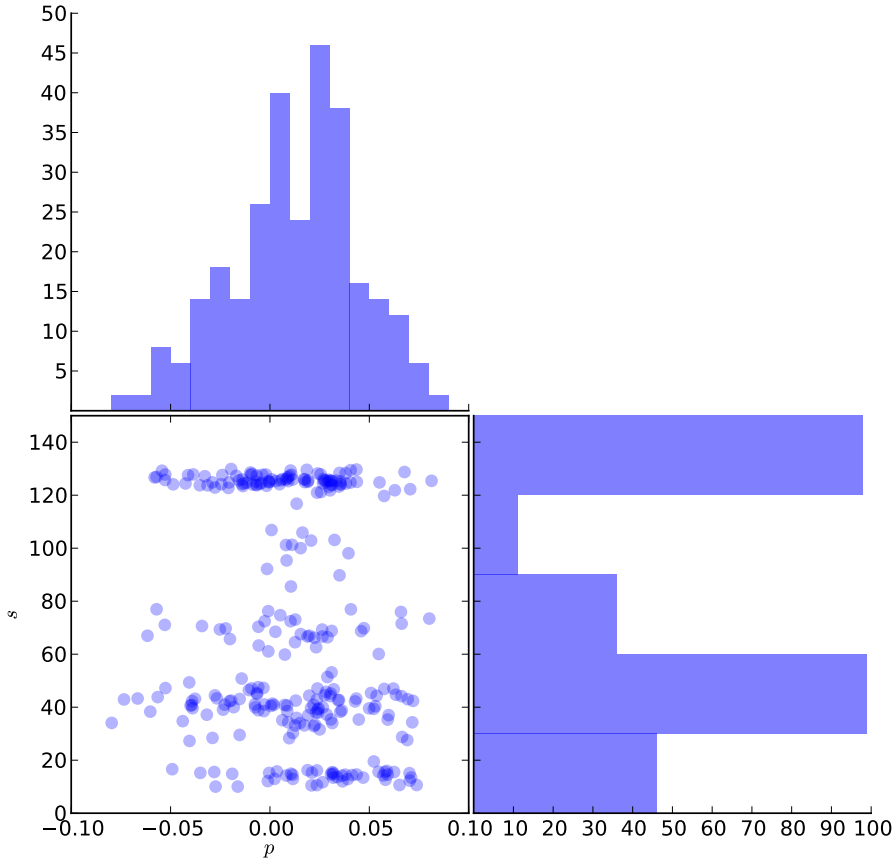


Figure 3.15: Coarticulation parameter relationship in vowel-nasal context in Experiment 1 for F2

3.6.3 Discussion and Conclusions

In this section we analyzed coarticulation parameters from a single-speaker corpus. Overall, we find the formant targets for vowels, approximants and nasals are largely in their expected locations. In examining the observed formant trajectories in vowels, approximants and nasals we see that our coarticulation model trajectories are able to approximate the observed trajectories, which highlights the flexibility in our model and reiterates the wealth of variance in normal speech.

We find that the coarticulation parameter s is loosely associated with our knowledge of formant movement in boundary conditions. Notably, we find many s values high in vowel to nasal transitions, conversely, we find many s values low in vowel to approximant transitions. Both of these follow our acoustic-phonetic knowledge that is easily visible in spectrograms of these transitions. The parameter p plays an important role in handling both minor boundary labeling errors, along with formant transitions that are seemingly asynchronous from our findings.

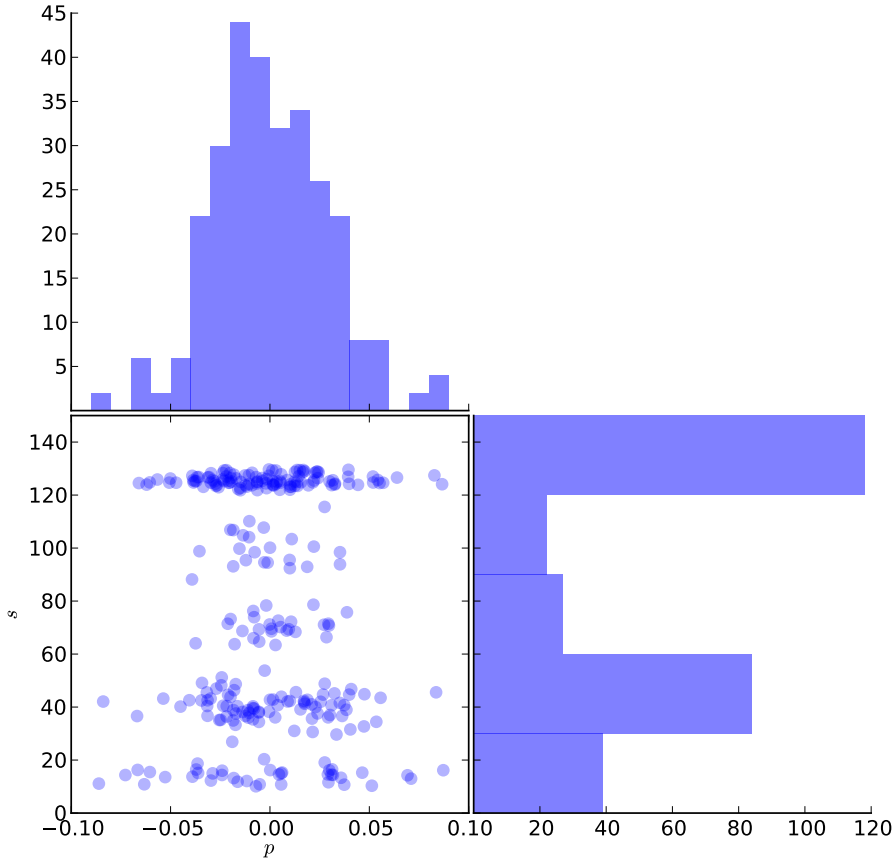


Figure 3.16: Coarticulation parameter relationship in vowel-nasal context from Experiment 2 for F2, with improvements in /m/ and /ng/ targets

3.7 Perceptual Study

In this section, we conduct a perceptual tests using model derived synthetic speech to ensure the model is capturing the important aspects of speech that is required for intelligibility. Specifically, our experiment aims to demonstrate that our resynthesized speech using model parameters is equivalent to observed formants.

3.7.1 Stimulus Creation

Using the corpus presenting in Section 3.6.1, we conducted a perceptual evaluation using a speech intelligibility test to compare vocoded speech from model trajectories with vocoded speech from observed trajectories, as well as with natural speech. Thus, three stimulus conditions were created for each sentence: natural, vocoded-observed, vocoded-model, for a total of 70 sentences $\times$ 3

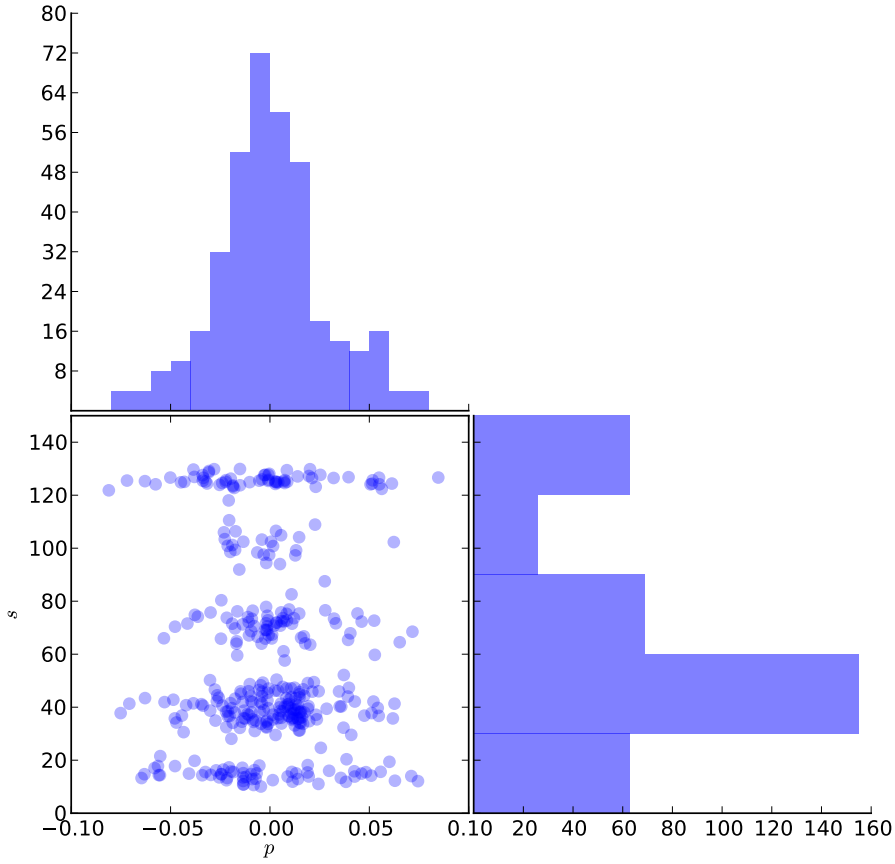


Figure 3.17: Coarticulation parameter relationship in vowel-approximant context for F2

conditions = 210 stimuli. These 210 stimuli were divided into 10 partitions consisting of 21 stimuli in a Latin squares design. In this experiment, we first estimated the formant frequency targets for all acoustic events using a single iteration in target search. Then using those global targets, we estimated coarticulation parameters using the approach detailed in Section 3.4.

The natural condition is unmodified speech and used as a control. The vocoded-model condition was created by replacing the formant frequencies with those from the model trajectories for formant frequency F1-F3. The fourth formant was set to a constant formant frequency of 3800 Hz with the bandwidth set to 500 Hz. In contrast, the vocoded-observed condition used observed formants and bandwidths derived from the freely-available WaveSurfer formant tracker [96]. The fourth formant was also set to the same fixed frequency and bandwidth to match the vocoded-model condition. The higher formants frequencies (F5-F8) and bandwidths (B5-B8) were set to their mean values in both conditions. Both vocoded conditions used energy and pitch from the natural condition.

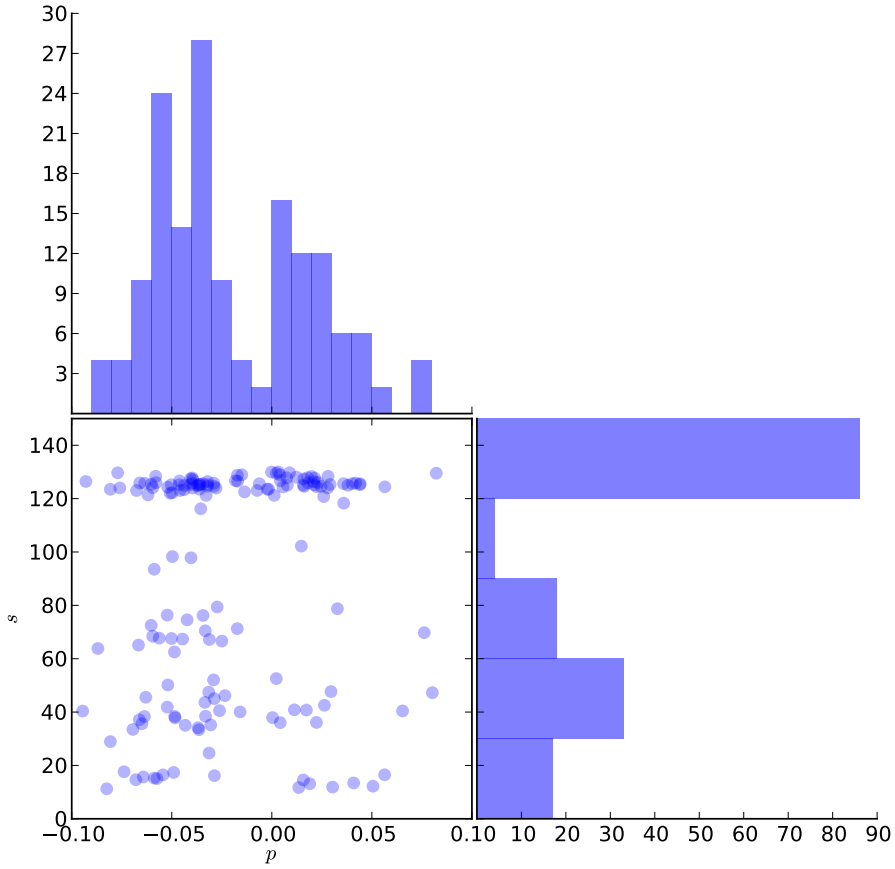
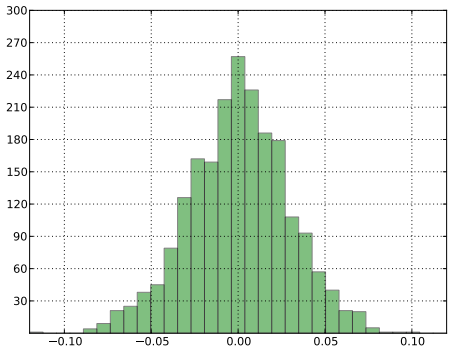


Figure 3.18: Coarticulation parameter relationship in fricative-closure context for F2

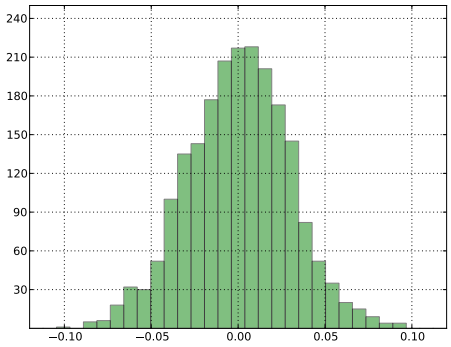
All stimuli were loudness normalized using an A-weighted RMS measure [77] and 12-talker babble noise was added to prevent saturation effects. Our babble noise source consisted of 20 adults simultaneously reading different magazine articles. The noise was stable over time and had a relatively flat spectrum from 200 to 1000 Hz, decreasing thereafter about 6 dB/octave [41]. The energy of the noise was set to a signal-to-noise ratio of +3 dB. The implementation of our synthesizer is similar to a Klatt formant synthesizer [62]. Based on the first three formant frequency and bandwidth values we constructed the all-pole formant filter. The fourth formant frequency was set to a constant 2800 Hz and energy and pitch trajectories were preserved from natural condition.

3.7.2 Administration

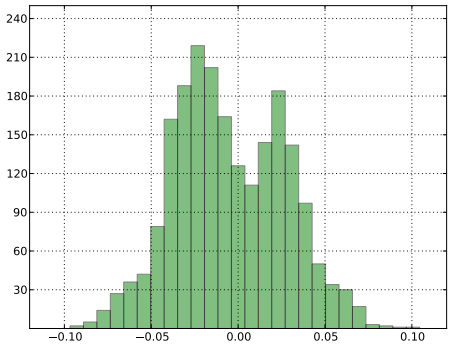
Amazon's Mechanical Turk crowd-sourcing service was used to provide human subjects. Workers were paid $0.50 USD per Human Intelligence Task (HIT) completed. Each HIT was quickly



(a) F1



(b) F2



(c) F3

Figure 3.19: Histograms of p values per formant

performed with an average hourly rate of approximately nine dollars USD. Each HIT was completed by unique individuals located in the United States who had a previous approval rating of 95% or

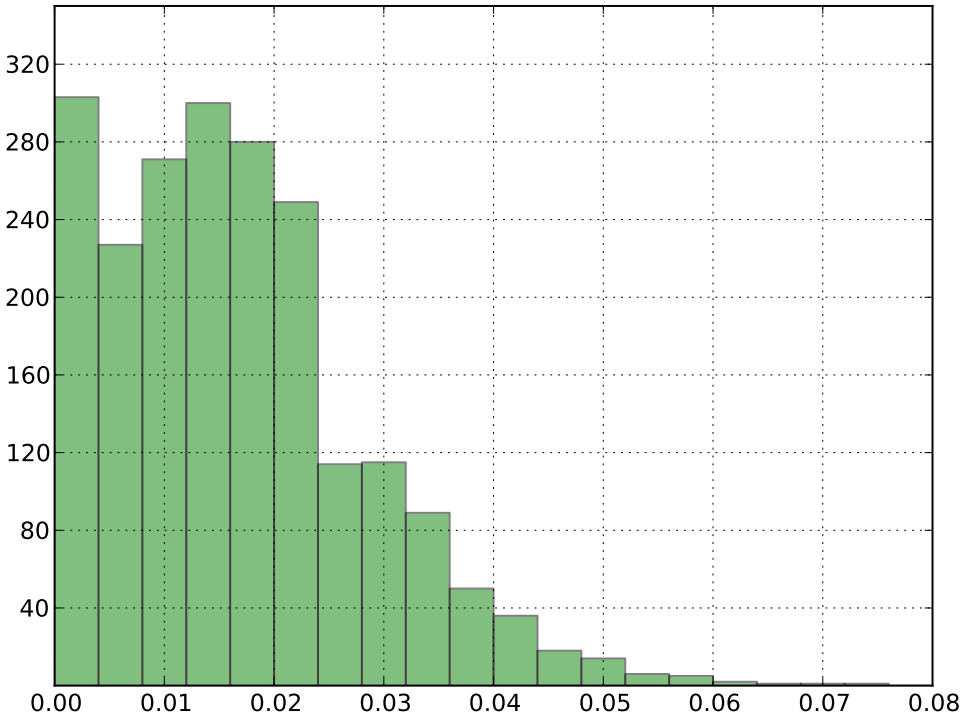


Figure 3.20: Standard deviations of p over F1, F2 and F3

higher and had completed at least 500 HITs. Workers with a high rating were found to produce results as measured by high accuracy on the natural condition. Each HIT contained 21 sentences composed of seven sentences per condition, presented in random order. The listener was presented the audio in conjunction with the sentence. The listener was asked to choose a word from a list of four possible answers to the statement: "Choose the missing word that was heard in the sentence, e. g., "place the wagon while _____ a ship at home". Where the answer, "springs", was displayed among three decoy terms (e. g., "strings", "slings"' and "strains"). Decoy terms were selected based on the closest phonetic similarity to the target term, using a list of commonly used words with a further requirement of containing the same number of phonetic elements. Finally we manually removed words that didn't match the target part of speech. The similarity measurement measured the average phonetic distance from the target term, where distance between any two phonemes was defined as the Euclidean distance of a four-dimensional manually derived description (sonority, manner, place and height) of each phoneme, which is defined in Appendix A. The values

for sonority, manner, place and height have been mapped to a [0, 10] range and are derived from two sources: (1) values for vowels that reconstruct the front-back/open-close dimensions of the International Phonetic Association (IPA) chart, and (2) related work by Kain *et al.* introduced in [52].

3.7.3 Results

Word	Natural	Model	Observed	Often Confused with
tool	8/10	0/10	3/10	pool
chick	8/10	0/10	5/10	tick
store	7/10	2/10	8/10	storm
sea	9/10	2/10	9/10	sea
sun	10/10	3/10	10/10	fun(5)
frosty	10/10	3/10	9/10	costly(6)
fence	9/10	4/10	8/10	vents(4)
cushion	10/10	4/10	10/10	mission(3)
state	10/10	5/10	9/10	spate(4)
punch	10/10	5/10	7/10	bunch(3)
west	10/10	6/10	10/10	rest(3)
spring	10/10	6/10	9/10	string(3)
saw	9/10	6/10	10/10	jaw(3)
hogs	9/10	6/10	5/10	hugs(2)
corns	9/10	6/10	7/10	horns(2)

Table 3.6: Specific words and their accuracy from 10 blocks with consistently poor performance in the vocoded-model condition. Vocoded-observed condition results for those words shown for context.

HITs that had high error rates ($> 50\%$ incorrect) on the seven stimuli in the natural condition were discarded, which was empirically found in previous experiments [18]; discarded HITs were resubmitted until completed. A total of 100 adults (10 blocks) participated, with self-reported normal hearing and unfamiliar with the goals of the study. The average proportion of words

Word	Model Count	Observed Count
week	10/10	10/10
wagon	10/10	10/10
tank	10/10	9/10
stone	10/10	9/10
ship	10/10	10/10
rider	10/10	10/10
rice	10/10	10/10
pup	10/10	10/10
planks	10/10	9/10
period	10/10	9/10
lazy	10/10	8/10
lake	10/10	10/10
home	10/10	10/10
grass	10/10	10/10
girl	10/10	4/10

Table 3.7: Words from 10 blocks that were consistently perceived correctly for vocoded-model condition in the intelligibility test. The observed-model condition was detailed as well.

heard correctly is summarized in Table 3.8. The top fifteen errors for the vocoded-model condition are detailed in Table 3.6 along with the errors for the vocoded-observed condition. The highest error word for vocoded-model condition is "tool", however, it also has a high error rate in vocoded-observed. In contrast, "store" has a high error rate in vocoded-model, but not in vocoded-observed (80% versus 20%, respectively). In examining the specific errors more closely, they can be divided into three primary groups: (1) acoustic event target too low, yielding a poor fit, (2) noisy formant observations yielded a poor fit, and (3) possible ill-fit due to context from sentence.

In order to investigate the impact of improved targets on the perceptual test, we performed a second experiment where we reduce the targets being estimated to just vowels, nasals and approximants. Results are summarized in Table 3.9.

Stimuli	Results	Confidence Interval
natural	95% (9%)	(93.6-97.3)
vocoded-observed	86% (14%)	(83.6-89.0)
vocoded-model	78% (17%)	(74.1-81.0)

Table 3.8: First experiment: Average intelligibility rate for each condition averaged across all listeners with +3 dB SNR. Standard deviation in parenthesis.

Stimuli	Results	Confidence Interval
natural	95% (9%)	(93.5-97.3)
vocoded-observed	88% (14%)	(85.6-91.3)
vocoded-model	74% (17%)	(70.5-77.5)

Table 3.9: Second experiment: Average intelligibility rate for each condition averaged across all listeners with +3 dB SNR. Standard deviation in parenthesis.

3.8 Discussion and Conclusions

We formulated and applied a continuous coarticulation model to formant frequency trajectories of a single speaker. The technique presented is data-driven and estimates global vowel and consonant formant targets for a particular speaker using automatic methods. In validating our model, we separated the target estimation validation from the actual coarticulatory model validation.

In target estimation, we demonstrated using known targets that we can recover those targets almost completely, even in the presence of noise. Next, in validating our coarticulation model, we used resynthesis as the mechanism to determine if our model was capturing the required details of speech in order to maintain intelligibility. In this experiment, we used two sets of targets to reconstruct the speech: (1) fully estimated targets, and (2) estimated vowels, nasals and approximants with remaining consonants fixed. The fully estimated targets performed best, however notably falling below vocoded-observed. While the vocoded-model condition did not perform as well as the vocoded-observed condition, the difference of 8% can be attributed to a small subset of keywords and issues. Especially in light of the compactness of the model parameter set versus the raw formant trajectories, with a compression ratio of approximately 1:12 for our corpus. It is important to highlight that the vocoder used is primitive and the lack of naturalness may have negatively

impacted the intelligibility of both vocoded-observed and vocoded-model conditions. Sommers and Barcroft [97] and Sommers *et al.* [98] hypothesized that variation in acoustic features that are relevant to speech perception could degrade speech intelligibility. In their experiments, degraded intelligibility could have been due to synthesized speech utilizing features that are different from what humans would normally produce.

We also examined coarticulation parameters to understand the parameter relationships with our acoustic-phonetic knowledge. First, we compared the estimated formant frequency targets with Hillenbrand *et al.* and Allen *et al.* to show that our targets are realistic and comparable to our acoustic-phonetic knowledge. Next, we investigated s parameters and found that s values tend to follow known acoustic relationships. Finally, we show that p parameters are varied enough to handle labeling variation and justify an asynchronous model.

Future work might include allowing some phoneme targets to have ranges or multiple points instead of single value targets; e. g., it is possible that targets are different in the onset versus coda of a stressed syllable. Relaxing targets to ranges is in contrast to our hypothesis that targets are global. In addition, simply seeking the lowest error leads to extremely high (or low) and unrealistic formant frequency targets for many consonants. A possible solution would be to utilize known targets for those phonemes and allow minor speaker-specific adjustments. In terms of formant trajectory modeling, one of the deficiencies in a smooth fit is in the translation of local to global coarticulation functions can exhibit subtle discontinuities. However, these discontinuities could be removed by low-pass filter.

Due to the difficulty of reliably estimating formants in a speech signal, another option would be Line Spectral Frequencies (LSFs), which can be computed directly from LPC coefficients and can be interpolated over time. The LSF values may also be grouped into separate streams, e. g., lower LSF values and higher LSF values to represent low-frequency and high-frequency regions of the signal. A key property of these streams is that they are interpolatable; this is needed because linearly weighted sums are used. We could also perform parameter search through synthesis of F1-F4 and compare model spectrum to actual in contrast to the current approach of error computation, which would remove the difficulties in formant tracking and mapping issues with LSFs. The primary computational constraint in this approach would be the synthesis portion.

Chapter 4

Application: Clear and Conversational Speech

Do CLR and CNV speech share phonetic formant targets? In this chapter, we will investigate the impact of speech style on speech features.[1] Our hypothesis is that CNV speech is a more coarticulated version of CLR speech where the differences between style are not in the target formant frequency space, but the coarticulation parameters.

In this chapter we will first introduce the concept of CLR and CNV speech along with a brief discussion of previous research in the acoustic properties of CLR speech. Next, we will detail the two parallel-style corpora used in this study, one CVC with the other unrestricted in context. Next, we will examine differences in CLR and CNV speech of observed phoneme centers for eight monophthong vowels in both CVC and unrestricted contexts. Using both corpora, we will estimate formant frequency targets for both CLR and CNV, along with coarticulation parameters. We will examine the resulting estimated targets from CLR and CNV speech separately, along with the coarticulation parameters, focusing primarily on the s parameters. We will compare these targets with previously published targets to show similarities in vowels and other resonant phonemes. Finally, we will use CLR formant frequency targets to represent CNV speech and CNV formant frequency targets to represent CLR speech. Here we will investigate the ability of a style-specific targets to model speech by measuring trajectory error.

4.1 Introduction

Conversational speech is speech that is elicited when speakers are instructed to "speak in the same manner as you would in ordinary conversation" [88]. On the other hand, clear speech is obtained by instructing speakers to "speak clearly, as you would when talking with hearing-impaired listeners"

[1] Portions of this chapter was published in Bush and Kain [18]

[88]. These two specific types of speech, referred herein as CNV and CLR speech, respectively are commonly called "conversational" and "clear" [89] speech. These labels have been imprecise since "clear" speech might imply intelligibility of the perceived speech, whereas "conversational" might imply speech produced in part of a dialog. It may also be possible that "conversational" speech might, in some cases, be equal in intelligibility to that of "clear" speech. Thus, we will refer to the style of speech that is generated in response to the aforementioned instructions.

4.2 Characteristics of Clear Speech

There are many acoustic and phonetic differences between CLR and CNV speech. Examining the studies that highlight these acoustic differences, we will group these acoustic differences into: prosodic (specifically, fundamental frequency, energy and phoneme duration), spectral (formant), and phonological features.

For prosodic features in CLR speech, we see an increase in mean and variability (or range) of the fundamental frequency (F0) [89, 10, 63]. Picheny *et al.* [89] reported greater root-mean square (RMS) intensities for unvoiced stop consonants in CLR speech than in CNV speech. Phoneme durations were lengthened, especially in the tense vowels [89, 36, 37]. In addition, pause durations were longer and their occurrence was more frequent. As a result of prolonged phoneme durations and increased pause durations, the speaking rate was significantly decreased from 160-200 words per minute in CNV speech to 90-100 words per minute in CLR speech [89, 63]. The consonant-vowel energy ratio (CVR) was increased in CLR speech, particularly for stops and fricatives (i. e., energies of the consonants had greater relative energy in CLR speech) [10]. In another study, this increased CVR was only found in affricates in CLR speech [63]. Increased amplitude modulation for low modulation frequencies (up to 3-4 Hz) of CLR speech was also found on a limited number of speakers [63].

For spectral features of CLR speech, vowel formant frequencies showed an expanded vowel space [89, 36, 10]. Long-term average spectra had higher energies at higher frequencies [63], which can be interpreted as decreased spectral tilt. Even though formant undershoot has been observed for both CLR and CNV speech in /w/-/x/-/l/ contexts (where x is a monophthong vowel), the amount of F2 displacement from target was significantly less in CLR speech [3]. Similarly, Picheny *et al.* [89] showed that formant displacement was dependent on vowel duration more for the lax vowels. Additionally, Moon *et al.* detailed that the formant frequencies showed more variation in lax vowels [76]. In terms of intelligibility, there have been several studies which demonstrated that speakers with larger vowel spaces are more intelligible than speakers with reduced spaces

[11, 44]. Interestingly, van Son and Pols [107] found that speaking style does not impact vowel formant trajectories. Specifically, an experiment with a single speaker reading Dutch text at normal and then fast speaking rate, after normalization, showed that vowel formant trajectories had no significant differences. However, it should be noted that normal speech in van Son *et al.*'s work is closer to our CNV speech definition.

For phonological features, studies have shown a number of differences between CNV and CLR speech. Vowel reduction (i. e., vowels becoming schwa-like), degemination (i. e., two similar phonemes merged into one sound), and alveolar flaps occurred more often in CNV speech. In contrast, bursts of the stop consonants in word final position tended to be released more often in CLR speech. Phoneme insertion of a neutral vowel /ax/ after a voiced consonant occurred more often in CLR speech [89, 63].

In summary, we can identify six major facts that are largely consistent in studies about CLR speech, namely (1) fundamental frequency (F0) mean and range for CLR speech are increased relative to CNV speech, (2) phoneme durations are longer in CLR speech, (3) amplitude modulation is increased for CLR speech, (4) vowel spaces are increased in CLR speech, (5) there are higher energies at higher frequency regions in CLR speech, and (6) phoneme insertions (e. g., schwa) occur with more frequency in CLR speech.

4.3 Methods

4.3.1 Parallel Style Corpora

To investigate differences in CLR and CNV speech, we will use two separate corpora that were derived from a single speaker: (1) a CVC corpus is introduced to show extreme differences between these two styles that are commonly highlighted, and (2) a generalized continuous speech corpus where vowels are unrestricted in their context.

A male native speaker of American English, with no professional training in public speaking recorded the same speech material in both CLR and CNV speaking styles. In conversational speech (CNV), the speaker was asked to speak as if talking with a colleague at a natural pace [88]. In clear speech (CLR), the speaker was asked to "enunciate consonants more carefully and with greater effort than in CNV speech and avoid slurring words together" [45].

All recordings were carried out in a sound-treated booth (Whisperroom MDL4260) located inside a control room. Recordings were made using a head-mounted close-talking microphone (AKG HSC200), positioned approximately 5 cm and off axis from the speakers mouth. The speaker recorded the materials at his own pace by operating a computer program. A technician

listened to each sentence and the speaker was asked to record a sentence again when pronunciation or style were not satisfactory. The speech signals were captured and stored digitally at a sampling rate of 22.05 kHz with 16-bit resolution.

In both corpora, initial estimates of phoneme identities and boundaries in each waveform were obtained using an existing forced-alignment system [49]. Then, a trained labeler checked and adjusted phoneme identities and boundaries manually. Formants were automatically estimated using a standard formant tracker [96, 103]. Formants were not manually corrected.

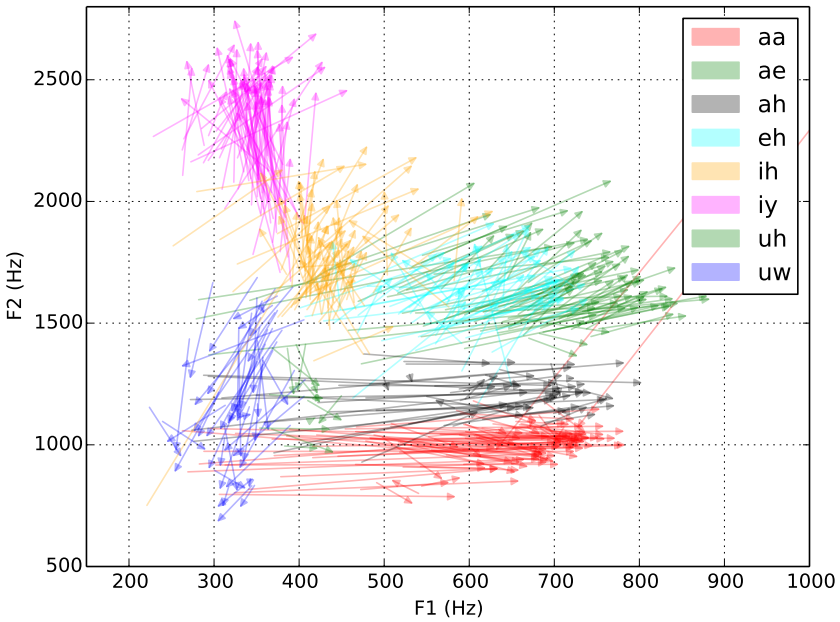
The parallel Style CVC corpus is composed of 242 meaningful CVC words (e. g. "cat", "well") constructed from a combination of 23 initial and final consonants, and eight monophthong vowels, spoken in a carrier sentence. Diphthongs are not represented in this corpus. Since /ao/ is often pronounced as /aa/ in West-Coast American English we merge these two phonemes into /aa/, an example of this merging of sounds is the words *caught* and *cot*. Each token was rendered *twice* in *both* styles, with a total of $242 \times 2 \times 2 = 968$ CVC tokens. The carrier sentences provided neutral meaning and had the same phoneme /d/ just before the target word in a sentence final context (e. g. "I know the meaning of the word *will*"). The parallel style continuous corpus used was introduced in Chapter 3, Section 3.6.1.

4.3.2 Analysis

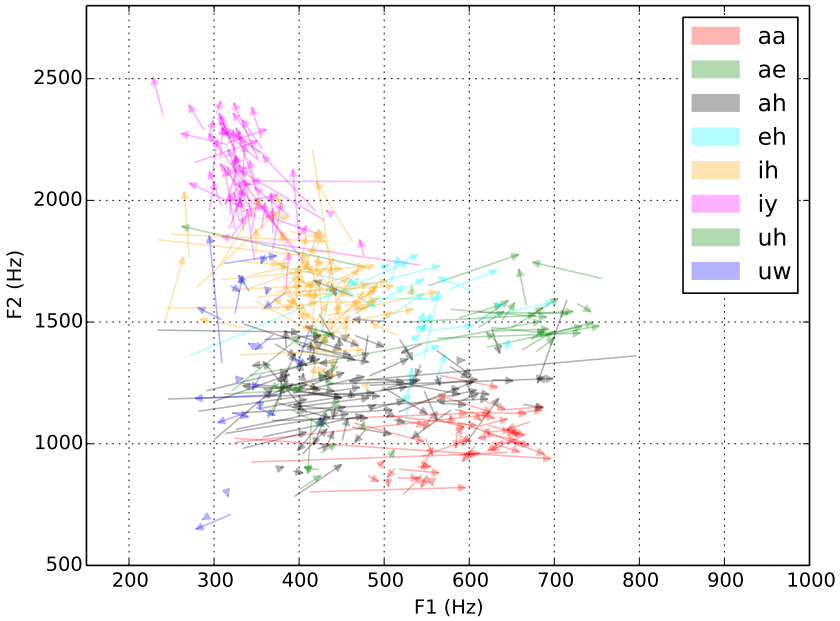
For both corpora, we aligned vowels of matching CLR and CNV sentences, wherein we extracted the center observed value for each vowel for formants F1 and F2. Figure 4.1a and Figure 4.1b show the F1/F2 formant frequency relationship between the two styles at vowel centers, for vowels in CVC and unrestricted contexts, respectively. Averaging the CNV to CLR vectors arrives at a more concise view for both corpora in Figures 4.2. Note the expanded vowel space of CLR speech, as compared to CNV speech [37], which is more pronounced in the CVC corpus indicated by longer arrows for all vowels (Figure 4.2a), but present nonetheless in the unrestricted context corpus. This finding indicates much less overall change from CNV to CLR in general speech.

4.4 Results

First, we estimate formant frequency targets for all phonemes for F1, F2 and F3, in order for both corpora. We can see the resulting formant frequency targets for eight monophthong vowels in Figures 4.3a and 4.3b. Examining F2 targets from the CVC corpus in Figure 4.1, we can see that CLR and CNV targets are quite distinct from each other, save /aa/, /ah/ and /uw/. However, in the unrestricted corpus, F2 targets in Figure 4.2 are noticeably little difference between CLR and

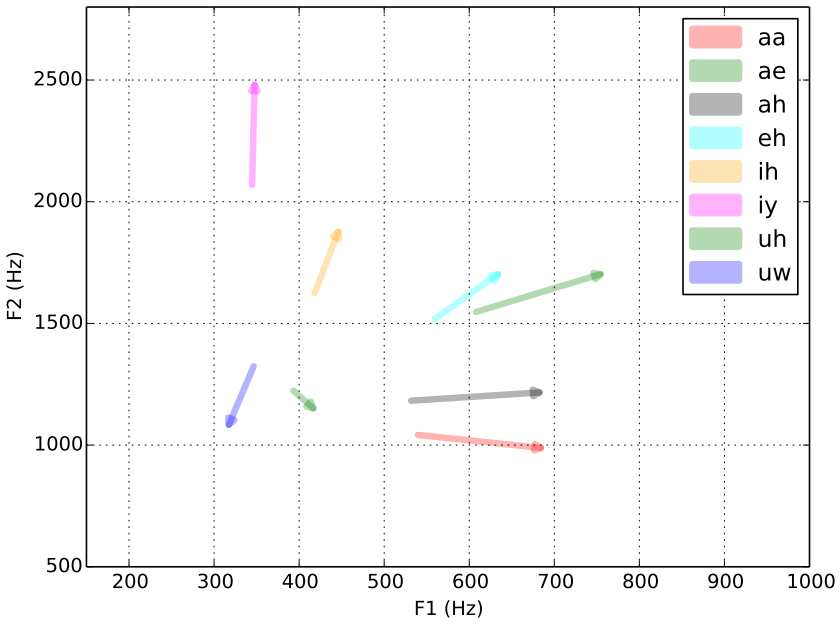


(a) Vowel context in CVC words

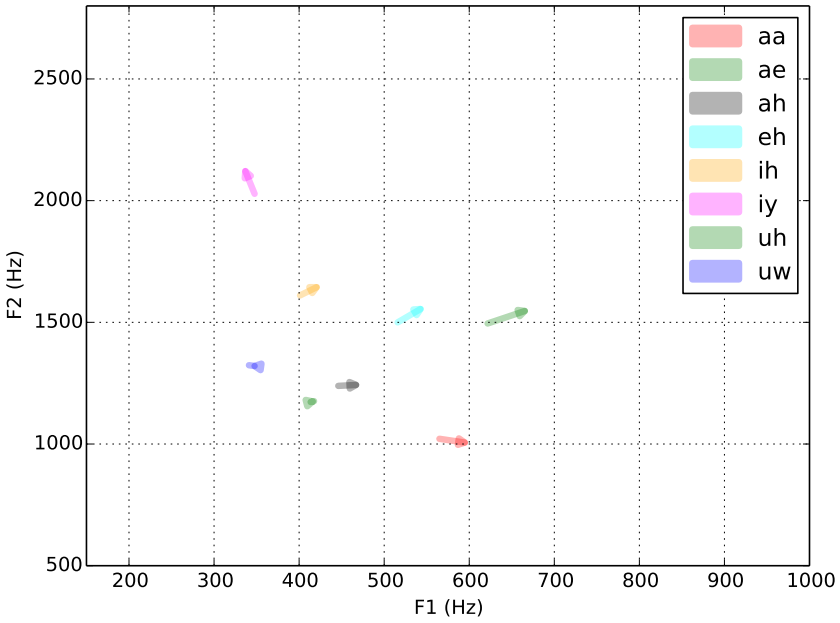


(b) Vowel context unrestricted

Figure 4.1: F1/F2 formant frequency relationship between CNV and CLR at vowel centers, arrows pointing from the former to the latter for vowels in CVC and unrestricted vowel contexts.

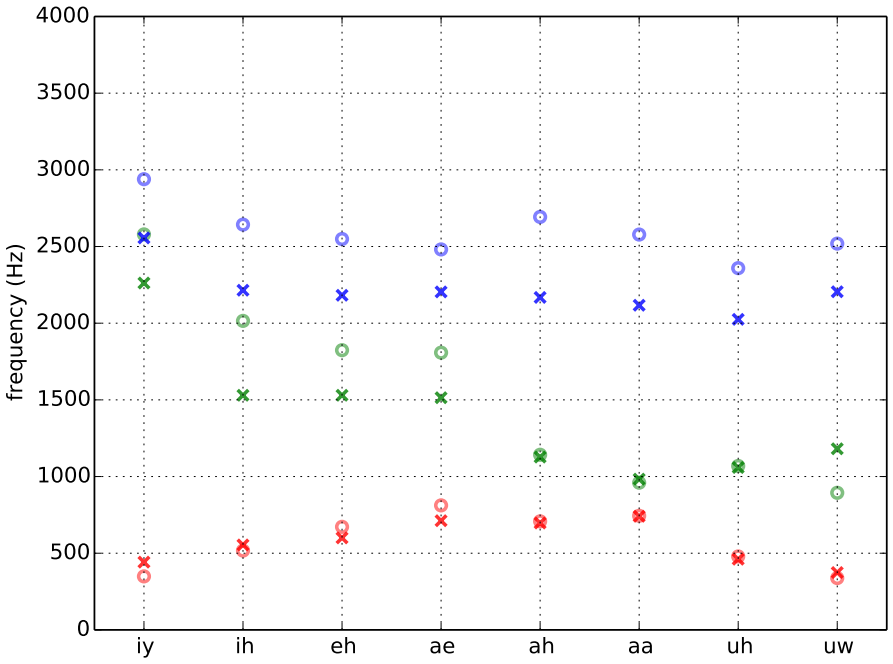


(a) CVC vowel context averaged

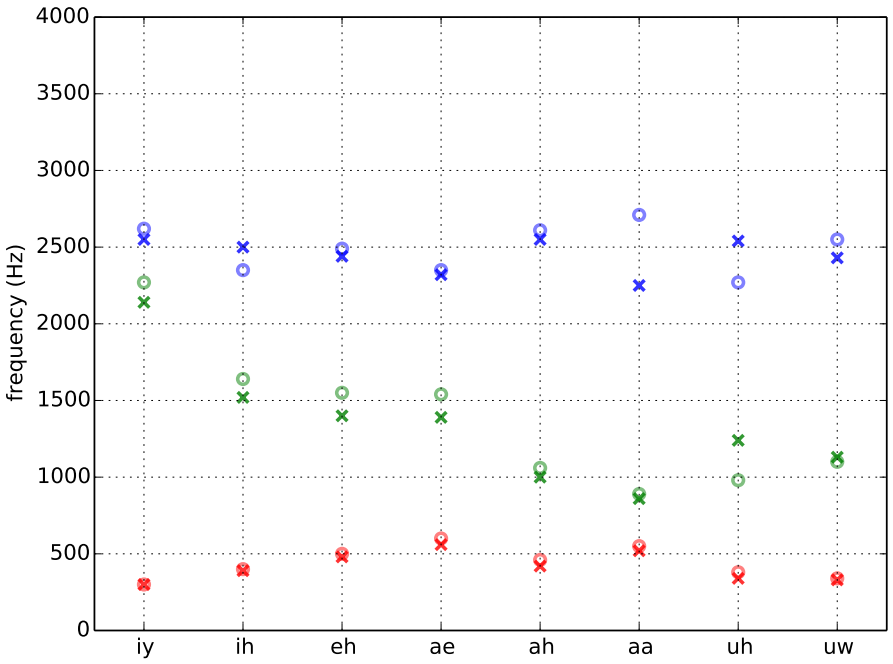


(b) Unrestricted vowel context averaged

Figure 4.2: Averaging F1/F2 formant frequency relationship between CNV and CLR at vowel centers, arrows pointing from the former to the latter for both CVC vowel contexts and all-context continuous speech.

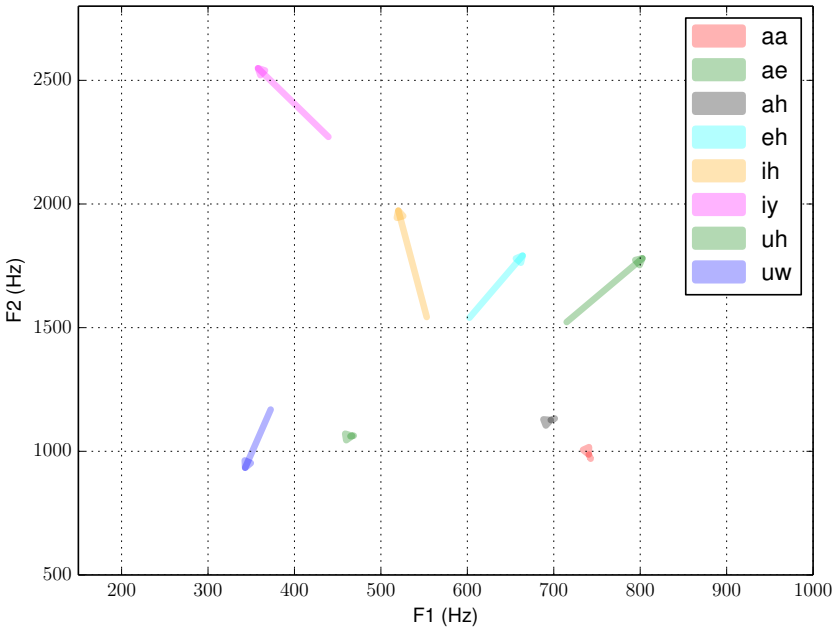


(a) Vowel targets in CVC contexts.

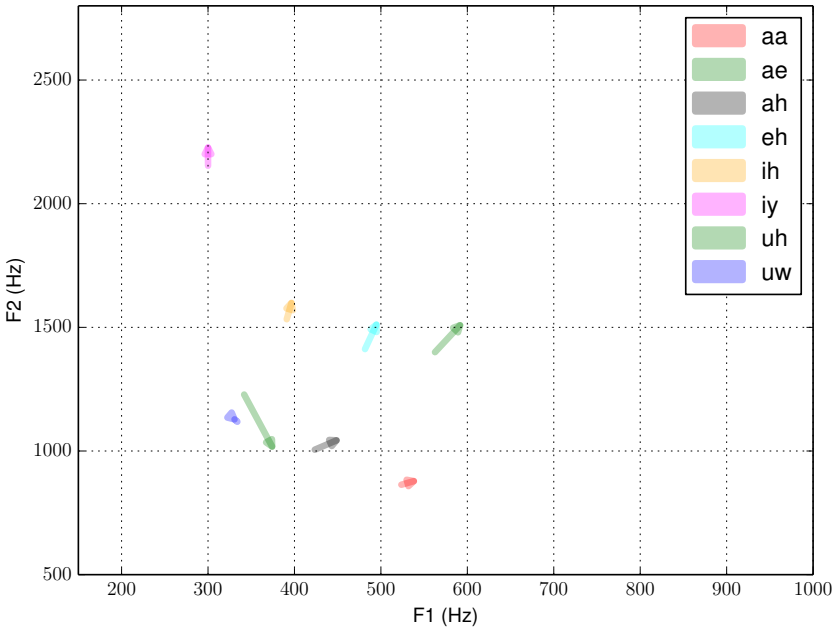


(b) Vowel targets in unrestricted vowel contexts.

Figure 4.3: Estimated vowel targets comparisons for CLR (circles) and CNV (x) speech from CVC and unrestricted vowel contexts. Showing formants F1 (red), F2 (green) and F3 (blue).



(a) CVC corpus



(b) Continuous corpus

Figure 4.4: F1/F2 formant frequency targets relationship between CNV and CLR, arrows pointing from the former to the latter for both CVC vowel contexts and all-context continuous speech.

CNV than in the CVC corpus.

4.4.1 Targets

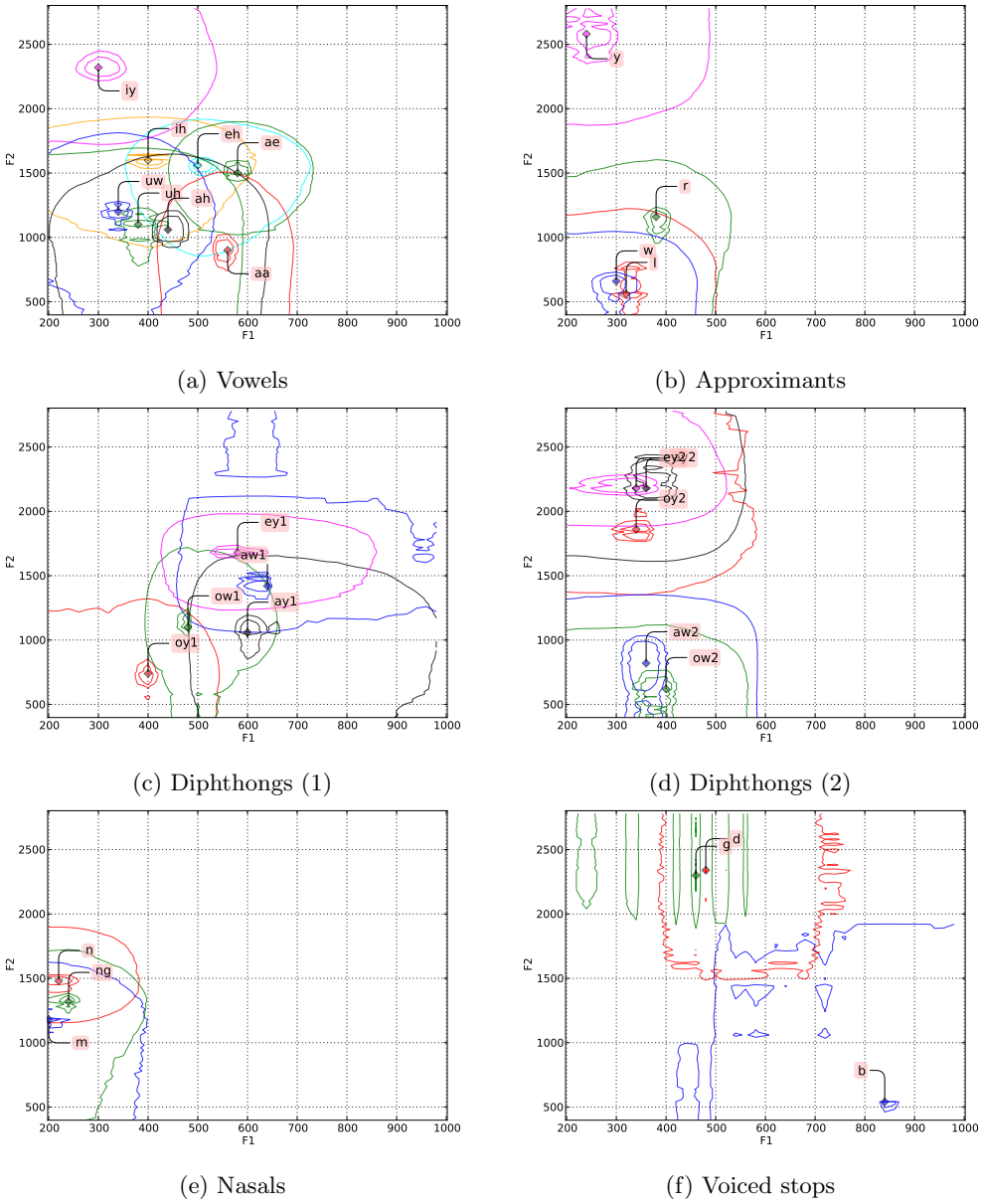
The formant target contours in F1/F2 space are shown for the continuous corpus in Figure 4.5 for CLR speech and Figure 4.6 for CNV speech. The contours are generated by the normalized error for each point in F1/F2 space for the global phoneme-specific error sweeps, resulting in a hill-like contour with bands at 0.5, 1 and 2%. We note that some vowel contours appear to have target ranges instead of points, e. g., /y/. The formant targets for vowels and consonants are largely in their expected locations based on acoustic-phonetic knowledge [2, 47]. We also note the there are little differences in matching CLR and CNV error surfaces.

4.4.2 Coarticulation Parameters

For a meaningful analysis of coarticulation parameters s_L, p_L, s_R, p_R, we only included tokens with neighboring targets that differ by at least 300 Hz to obtain meaningful values. For this study, using only F2, we estimated coarticulation parameters using the style-specific targets that we discovered in Section 4.4.1, with CLR speech using CLR targets. Examining the maximum value of the center coarticulation function, $\max F_C(t)$ of F2 for all local triphone models that were vowels for both corpora.

In the CVC corpus, we found the mean values CLR = 0.92 and CNV = 0.59. Figure 4.7a shows a histogram for these values. As these two distributions failed normality tests, we used the Mann-Whitney test [72], a non-parametric test to measure significance. The Mann-Whitney $U = 5569, p < 0.05r = 5.73$, thus a statistically significant discriminator between the two styles. In the continuous corpus, we found the mean values CLR = 0.83 and CNV = 0.80. Figure 4.7b shows a histogram for these values. Similar to the CVC corpus, these two distributions failed normality tests, we used the Mann-Whitney test [72], a non-parametric test to measure significance. The Mann-Whitney $U = 35430, p = 0.16$, which fails significance.

In the CVC corpus, we observe that the CLR style histograms have many more occurrences of higher values than the CNV style. This was expected since their trajectories are more likely to reach their target in CLR style ($\max F_C(t) \approx 1$), while undershoot is more prevalent in CNV style ($\max F_C(t) \ll 1$). However, in the continuous corpus, we find that CLR and CNV speech look surprisingly similar, which is confirmed with the similar mean value and failing significance between both styles.



(a) Vowels

(b) Approximants

(c) Diphthongs (1)

(d) Diphthongs (2)

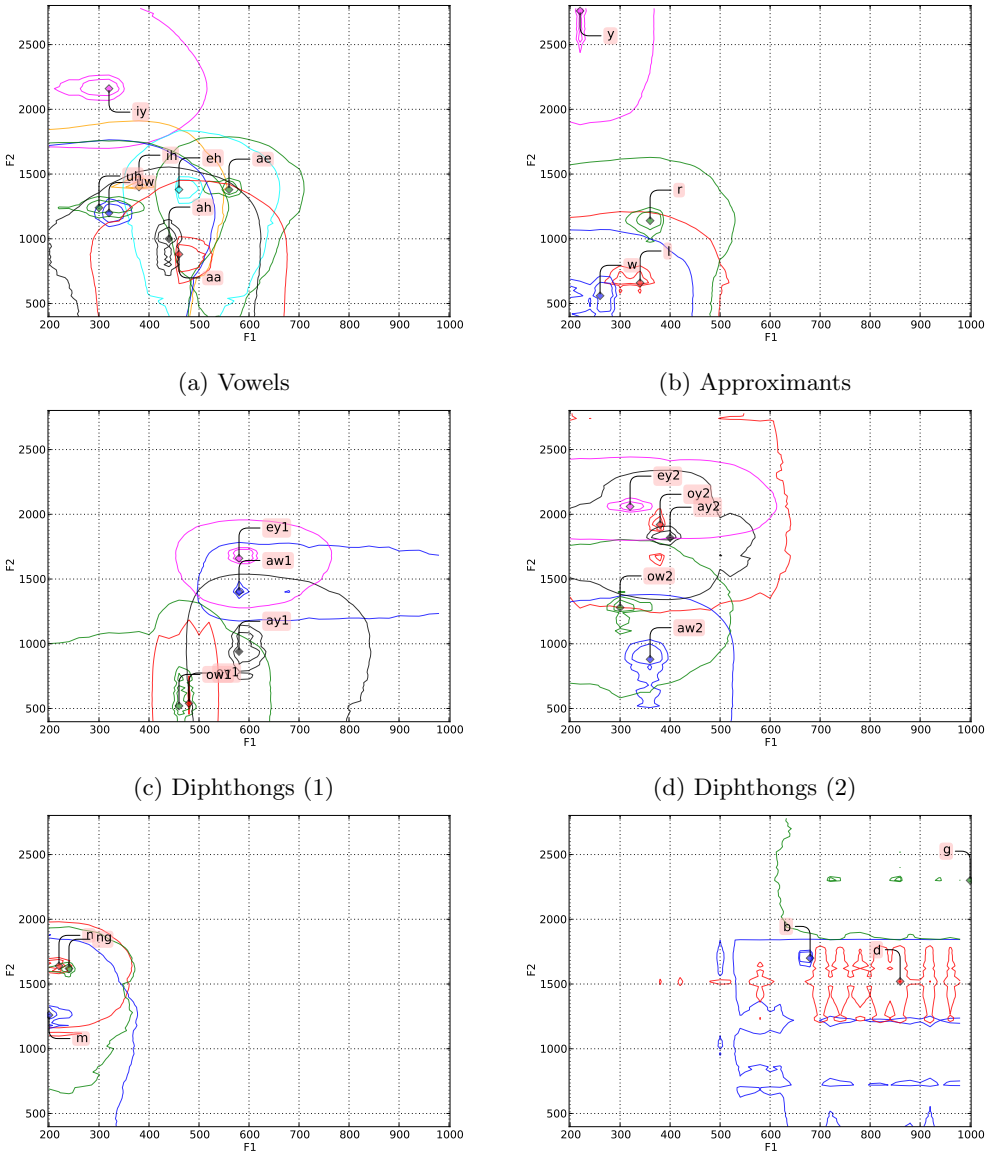
(e) Nasals

(f) Voiced stops

Figure 4.5: Formant targets for CLR speech and iso-contours based on the global minimum of E_{sweep}

4.4.3 Goodness of Fit

We now examining RMS error from the resulting trajectories where we model both speech styles with matched (i. e., CLR speech and CLR targets) and mismatched targets (i. e., CLR speech and CNV targets). Examining results from the CVC corpus, in Table 4.1, we observe that the diagonal is the lowest error since speech style and targets are matched in this case. We also note that CLR



(a) Vowels

(b) Approximants

(c) Diphthongs (1)

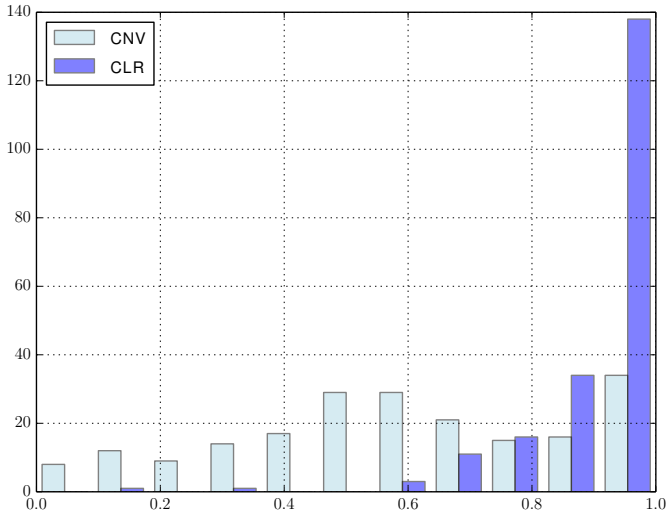
(d) Diphthongs (2)

(e) Nasals

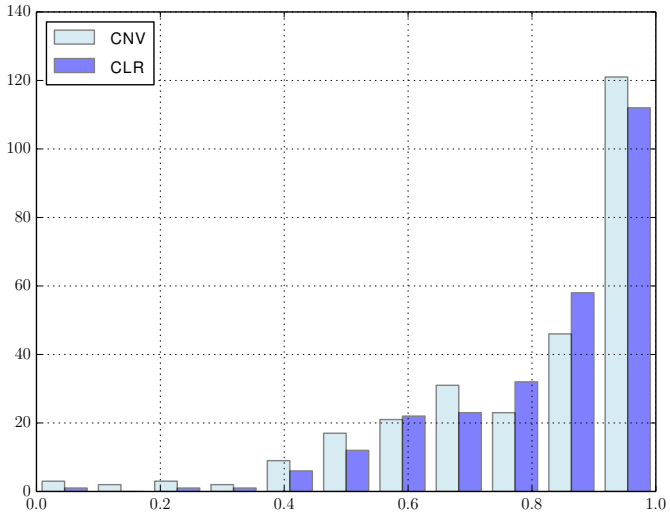
(f) Voiced stops

Figure 4.6: Formant targets for CNV speech (with observed data where available) and iso-contours based on the global minimum of E_{sweep}

speech does quite poorly when modeled with CNV targets, which is expected given the differences in speech styles from observations of vowel centers in Figure 4.2a. Similar to the CVC corpus, the continuous corpus in Table 4.2 has the property with the diagonal having the lowest error. However, we see that regardless of target style, we can model CLR speech equally well. We believe this is an artifact from CLR and CNV formant frequency targets being quite similar, as evidenced



(a) CVC corpus



(b) Continuous corpus

Figure 4.7: Histogram of maximum center coarticulation function values, $\max F_C(t)$, for both CLR and CNV style for both corpora. Analysis is limited to local triphones with the center event a monophthong vowel.

by Figure 4.4.

Style / Targets	CLR	CNV
CLR	170 (101)	218 (117)
CNV	151 (83)	131 (74)

Table 4.1: Average model errors for CVC corpus and standard deviations (in Hz), using targets estimated from one style (column headers) and evaluated on another style (row headers).

Style / Targets	CLR	CNV
CLR	166 (57)	166 (60)
CNV	160 (55)	147 (52)

Table 4.2: Average model errors for continuous corpus and standard deviations (in Hz), using targets estimated from one style (column headers) and evaluated on another style (row headers).

4.5 Discussion and Conclusions

We analyzed coarticulation in both CLR and CNV speech in two separate corpora. Our goal is discover if CLR and CNV speech share targets. In our initial analysis, we examined the observed formant frequency centers of vowels in both corpora. The CVC specific corpora showed the expected differences between both speech styles. Interestingly, the more general continuous corpora showed a decreased difference between speech styles. This finding is similar to van Son and Pols [107] where they found that speaking style does not impact vowel formant trajectories. In their experiment, a single speaker reading Dutch text at normal and then fast speaking rate, after normalization, showed that vowel formant trajectories had no significant differences.

Estimated formant frequency targets mirrored our findings in observed vowel centers with differences between speech styles primarily in the CVC corpus. The continuous corpus shows little difference between speech styles when examining F1/F2 targets. In examining coarticulation parameters, again, we find little no significant difference between speech styles in the continuous corpus. This is contrasted by a significant difference in speech styles in the CVC corpus. Finally, in measuring the goodness of formant trajectory fit we find little evidence to convince us that CLR and CNV speech targets are significantly different in our continuous speech corpus.

In summary, for this speaker, we find little differences between CLR and CNV speech styles across a large corpus of speech where context is unrestricted. This is evidenced by observations and estimated targets that fit the observed formant trajectories. The sharp contrast between speech styles in CVC contexts is possibly explained by the limited context and the target CVC word being sentence final.

Chapter 5

Application: Formant Tracking

Studies concerned with the fundamentals of speech production often use formant (vocal track resonances, or VTR) features; however, traditional formant tracking is known to make many errors, despite recent advances. If the phoneme segmentation is given, our coarticulation model can improve initial estimates of formant tracks.

In this chapter, we describe a new automatic formant tracking algorithm that combines traditional formant tracking and the model introduced in Chapter 3 to derive formant tracks. Our hypothesis is that a coarticulation model fitted to automatic formant estimates will allow a reduction in overall error with the true formant trajectories. We will evaluate the first two formants of our model using a subset of the TIMIT speech corpus [42] that have been manually corrected after an initial automated pass [27].

5.1 Introduction

Fant [34] defined formant frequencies of a vowel to be the eigenfrequencies of the vocal tract, as evidenced by energy in the acoustic power spectrum. Acoustic-Phonetic literature has equates the formant frequencies with peaks in the spectrum. Physically, they correspond to resonances of the vocal tract.

Many research studies in speech analysis and synthesis rely upon accurate formant frequency measurements. There are many examples of formant tracking in research studies, with a few examples in such diverse domains such as dysarthric speech [78, 60], stuttering analysis [91], early diagnosis of obstructive sleep apnea [113], enhancement of vowel intelligibility in dysarthric speech by formant modification [53], and diagnosis of learning disabilities [9] by examining phoneme transition boundaries of formants. All of the aforementioned areas relied upon correct formant measurements for both analysis and synthesis. However, formant trajectories are known to be inaccurate without human annotators correcting by visually inspecting a spectrogram and using

phonetic knowledge. This manual correction of trajectories is a labor intensive task and requires specific training and knowledge. This is especially true in cases where the spectrogram is ambiguous. Therefore, extensive research utilizing formants is limited due to the cost of formant correction in corpus development.

In general, formant tracking is decomposed into two stages: (1) estimating the frequency and bandwidth of each from short segments of the speech signal, (2) linking together these estimates to arrive at an appropriate track. In the first stage, these estimates are estimated from peaks in the spectral representations or from the roots of the polynomial underlying the linear predictive model. This estimation is often difficult because of spurious peaks in the LPC spectrum and the weakening or splitting of formants. Using roots of predictive models is plagued by difficulty in setting rules for identifying roots as a formant. In the second stage, tracks are formed by linking together the estimates from stage one and is usually achieved by continuity constraints, for example, transition cost of tracks estimated using dynamic programming [103, 112].

Using dynamic programming has the difficulties of selecting proper strength of continuity constraints at phoneme boundaries with large transitions,e. g., consonant-vowel boundaries. This was addressed to include phonemic context in the estimation and tracking, which relied upon forced alignment with phonetic information and segmentation [66]. Another study used distributions of formant and bandwidth combinations in a Bayesian approach [33]. Recently, using Kalman filtering techniques [85, 74]. Each of these methods showed an improvement over formant trajectories produced by freely-available formant tracking algorithms [96]. Even with these approaches, tracking errors still persist and solutions to problems like formant track smoothing (e. g., correcting for large frequency jumps) are performed in an *ad hoc* manner.

5.2 Methods

The algorithm is tested using natural speech utterances and the performance is compared against formant tracks obtained by the conventional method using continuity constraints only. First we will detail the corpus, then the algorithm using our coarticulation model. Finally we will go over the results that presented in a summarized fashion along with gender-specific results. Finally we will close with a brief discussion.

5.2.1 Corpus

The VTR database is a proper subset of the The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (or TIMIT for short, Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)) [42] is detailed by Zue [114], contains recordings of phonetically-balanced prompted English speech. It was recorded using a Sennheiser close-talking microphone at 16 kHz sample-rate with 16 bit sample resolution. TIMIT contains a total of 6300 sentences (5.4 hours), consisting of 10 sentences spoken by each of 630 speakers from eight major dialect regions of the United States (seven distinct regions with the eighth region being the "army brats"). All sentences were manually transcribed at the word level, force aligned to phonemes by a speech recognizer and then hand-validated. The prompts for the 6300 utterances consist of 2 "dialect sentences" (SA), 450 phonetically compact sentences (SX) and 1890 phonetically-diverse sentences (SI).

The TIMIT Corpus documentation suggests training (approximately 70%) and test sets, as described in Table 5.1. The training set contains 4620 utterances, but usually only SI and SX sentences are used, resulting in 3696 sentences from 462 speakers. The test set contains 1344 utterances from 168 speakers. The core test set, which is the abridged version of the complete testing set, consists of 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). With the exception of SA sentences which are usually excluded from tests, the training and test sets are disjoint. The experiments herein will use the core test set.

This speech corpus has been a standard database for the speech recognition community for several decades and is still widely used today, for both speech and speaker recognition experiments. This is not only because each utterance is phonetically hand labeled along with gender and dialect region, but also because it is considered small enough to analyze completely and large enough to be realistic.

Set	**Speaker Count**	**Sentence Count**	**Speech Hours**
Training	462	3696	3.14
Core Test	24	192	0.16
Complete Test	168	1344	0.81

Table 5.1: TIMIT corpus training and test details.

The TIMIT phone transcriptions provide a speech sound label for each waveform sample. A frame is considered non-speech if labeled as a pause (pau, epi, h#). Speech-labeled frames are classified into one of six phonetic categories to reveal any phone-dependent error patterns. The six

TIMIT phonetic categories are vowel, semivowel/glide, nasal, fricative, affricate, and stop. Each speech frame is assigned the phonetic category. Diphthongs are excluded from this analysis.

A baseline set of automatic formant measurements was extracted using WaveSurfer, which is built upon the Snack Sound Toolkit [96], which is a freely available signal processing tool that performs formant tracking. The first two formants were extracted using settings: LPC order was set to 18, Hamming window at 128 with a pre-emphasis factor of 0.97. As formant measurements are only every 10ms, our comparisons will be at that scale. The VTR database contains center frequency and bandwidth for the first three formants.

Corrections were performed using a knowledge-based approach based on waveform, word and phoneme transcriptions, phonemic boundaries and wideband spectrum. In corrections by Deng *et al.* [27], specific attention was given to regions of VTR values during consonant-to-vowel (CV) and vowel-to-consonant (VC) transitions.

5.2.2 Algorithm

The formant tracking algorithm consists of three phases: (1) formant tracking using WaveSurfer, (2) formant frequency target estimation, (3) coarticulation function estimation using estimated targets. Phases two and three are performed for each formant separately, starting with F1. In the formant tracking phases, we use a traditional formant tracker to generate candidate formant trajectories.

In the target estimation phases, we estimate targets specifically for each speaker using all available speech examples in the core test subset. The algorithm for target estimation is described in Chapter 3. However, we only estimate targets for nasals, approximants and monophthong vowels since we are in a resource-constrained situation where there are not enough examples of all consonants in various contexts to reliably estimate their target value. The remaining targets were set to values defined in Appendix C and if not defined were set to mean observed values. Furthermore, in this study, we set closures of their respective phoneme target to have the same target value, e. g., /p/ = /pcl/. Finally, using targets derived from previous method, we estimate s and p parameters for each underlying triphone model to arrive at a final trajectory. This resulting trajectory is our new formant track.

5.2.3 Error Analysis

Formant-tracking error rate is calculated for all speech sounds. This set will include monophthong vowels, approximants and nasals. Quantitative performance will be detailed by phonemic class-based error as RMS error (in Hz):

$$Error(X) = \sqrt{\frac{\sum_{t=0}^{N} \cdot \left(X(t) - \hat{X}(t)\right)^2}{N}} \tag{5.1}$$

where $X(t)$ and $\hat{X}(t)$ are the reference and estimated formant trajectories, N is the number of frames. This is computed separately for each formant. In computing phoneme class errors, we use the TIMIT-specified phoneme begin and end times. This error is reported over all speech-labeled frames and further categorized by gender (male, female).

5.3 Results

In Table 5.2 we find the results between WaveSurfer and our model. Over both formants (F1 and F2) we find WaveSurfer's average error of 246 Hz versus our model with an error of 174 Hz. Examining results by gender we find the male overall results were 244 Hz for WaveSurfer, with 164 Hz for our model. In contrast, we find female overall results were 250 Hz for WaveSurfer, with 191 Hz for our model. We presume that the higher error rates in female results can be largely attributed to using male-specific formants for consonants defined by Allen *et al.*[2].

Examining the tracking approach closely, we can see in Figure 5.1a the model has a smooth transition between vowel /ih/ and the consonant /s/. Continuity in consonants is exhibited in Figure 5.1b with /pcl/. In the first female example, Figure 5.1c, we can see examples of formant track jump F2 to F3 in the phoneme /m/. Another example is in Figure 5.1d, where /zh/ of F2 jumps up to F3. In both cases, the model more gracefully handles such transitions.

5.4 Discussion and Conclusions

We have considered the problem of estimating vocal tract resonances from an observed acoustic waveform, in the context of a coarticulatory model. This formant-tracking approach uses traditional formant tracking and utilizes knowing the phoneme identity a priori. The algorithm consists of two phases: (1) traditional formant tracking, (2) utilizing the presented coarticulation model to correct errors from the formant tracker. The algorithm is tested using natural speech utterances and the performance is compared against formant tracks obtained by conventional method, which

	WaveSurfer		Model	
Phonetic class	**F1**	**F2**	**F1**	**F2**
vowels	111	261	90	204
approximants	118	283	98	231
nasals	159	413	126	292
fricatives	235	356	159	194
affricates	288	416	177	217
stops	204	323	150	218
overall per formant	**169**	**323**	**126**	**221**

Table 5.2: VTR tracking errors (in Hz) measured by the RMS error between the reference and estimated values over all frames for a particular broad phonetic class. This table shows our coarticulation model used in formant tracking. Results are listed for F1 and F2, and for each of the six phonetic classes separately and combined per formant.

	WaveSurfer		Model	
Phonetic class	**F1**	**F2**	**F1**	**F2**
vowels	111	224	91	171
approximants	128	292	104	237
nasals	155	374	126	269
fricatives	240	361	155	173
affricates	308	415	190	192
stops	219	326	153	196
overall per formant	**176**	**312**	**127**	**200**

Table 5.3: Male specific speaker VTR tracking errors (in Hz) measured by the RMS error between the reference and estimated values over all frames for a particular broad phonetic class separately for F1 and F2, and combined per formant across all 192 test utterances.

uses only continuity constraints. Our new algorithm significantly reduces the formant-tracking error rate for male speakers by 31% for F1 and 44% for F2, while reducing the error rate for female speakers by 24% for F1 and 28% for F2. We suspect the difference between male and female error

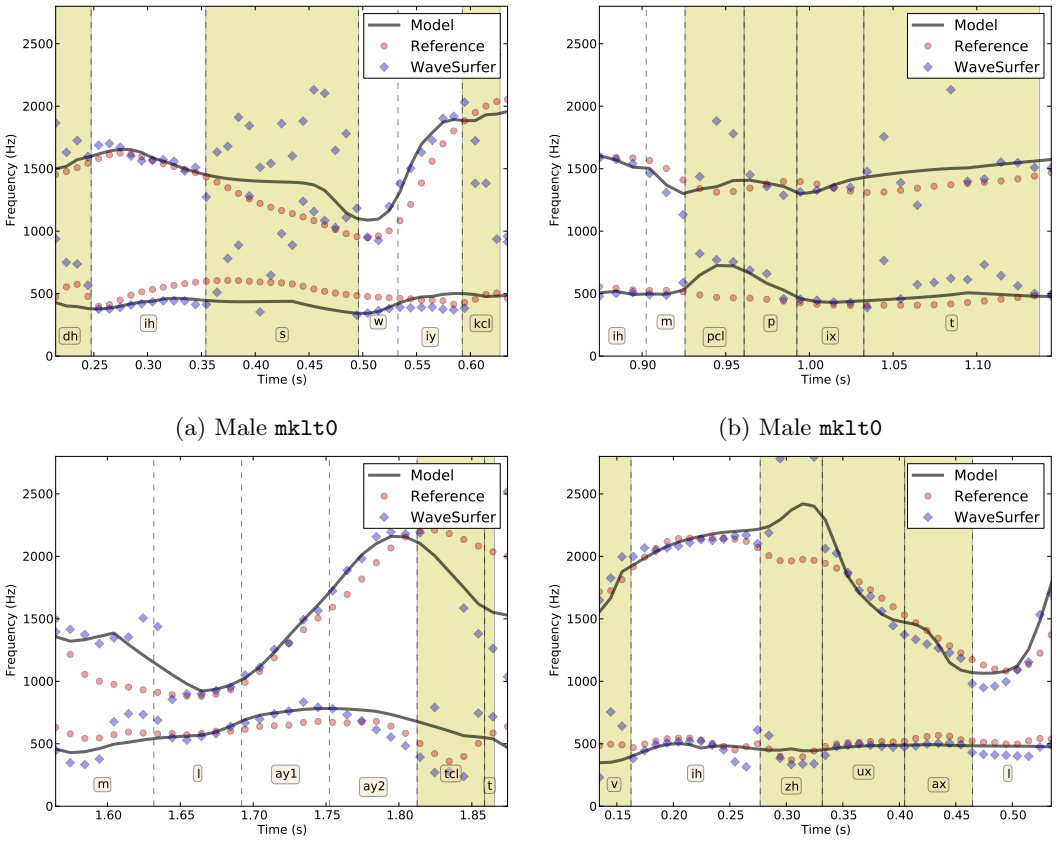
	WaveSurfer		Model	
Phonetic class	**F1**	**F2**	**F1**	**F2**
vowels	111	330	88	266
approximants	98	266	85	219
nasals	165	476	126	330
fricatives	225	345	166	240
affricates	242	418	147	274
stops	177	317	146	256
overall per formant	**157**	**343**	**123**	**260**

Table 5.4: Female specific speaker VTR tracking errors (in Hz) measured by the RMS error between the reference and estimated values over all frames for a particular broad phonetic class separately for F1 and F2, and combined per formant.

rates can be largely attributed to using male-specific formants for consonants defined by Allen *et al.*[2] as indicated by the larger RMS error in female model error rate.

The VTR database used herein provides an initial benchmark of ground truth for the first three formant frequency values. The values in the VTR database, however, should be interpreted with caution because baseline tracks were initially obtained via a first-pass automatic algorithm [28]. It is not known how much manual intervention was required and what types of errors were corrected. In particular, VTR tracks are observed to not always track high-energy spectral regions, e. g., F2 values of /iy/ aren't precise. Despite the presence of various labeling errors in the VTR database, it is still useful to obtain initial performance metrics of formant tracking algorithms on real speech. In summary, we find that using the presented algorithm significantly reduces the formant-tracking error rate for male speakers by 31% for F1 and 44% for F2, while reducing the error rate for female speakers by 24% for F1 and 28% for F2. It should be noted that F2 is likely to have more errors than F1 [27] and our approach significantly reduces the F2 error rate. However, the algorithm occasionally gives choppy or introduces discontinuous formant frequency estimates due to cross-fading of coarticulation functions. Despite this limitation, the overall performance of the algorithm is better than those of traditional formant estimation techniques.

Our results compare similarly to Mehta *et al.* [74] for F1 on the same dataset, however F2 outperforms by a distinct margin across all classes and overall. In comparing our method with Lee



(a) Male `mklt0`

(b) Male `mklt0`

(c) Female `fpkt0`

(d) Female `fdhc0`

Figure 5.1: Example estimated formant track using WaveSurfer and our coarticulation model along with the reference VTR trajectories for both male and female speakers. Regions where targets not estimated highlighted in yellow.

et al. [66], an advantage is that the model can represent undershoot/overshoot as evidenced by maximum center phoneme contribution max F_C can be less one, and often is in CNV speech. This allows us to model coarticulation functions that have undershoot (trajectories that do not reach the target). In contrast, Lee's algorithm assumes a constant level of coarticulation and is unable to model undershoot incorrectly. This would result in less than ideal fit, especially if using targets across varying styles of speech.

This work integrates acoustic-phonetic knowledge seamlessly into the formant tracking problem and is a post-processing step to traditional formant tracking. It should be re-emphasized that manually corrected phoneme boundaries along with phoneme identities were used in the described corpus. Still more work would be required to automate boundary adjustment, but forced alignment [49] would a straight-forward initial solution.

Chapter 6

Application: Dysarthria Detection

Dysarthria is a motor speech disorder, resulting in impaired movement of the muscles used for speech production, including the lips, tongue, vocal folds, and/or diaphragm. The assessment of dysarthric speech intelligibility involves an expert judge and a subjective rating system. This intelligibility assessment is usually part of a larger battery of assessments to gauge the severity of the disease or progress of treatment. Can deficiencies in formant frequency targets be measured using our coarticulation model automatically? To investigate this question, we estimate targets of dysarthric speakers to generate vowel space measurements and compare with non-disordered speaker whose targets are considered typical. We will also examine coarticulation parameters between disordered and non-disordered, focusing primarily on the max F_C measure in local triphone models.

6.1 Introduction

Dysarthria is a diverse group of motor speech disorders that typically impair intelligibility. There are several factors of dysarthric speech that might be captured in the proposed coarticulation model: (1) *target shift* (in physical ability) in dysarthric speakers might differ from non-disorder speakers (spectral targets differ between disordered and non-disordered) and (2) *increased coarticulation* [104] as the contextual influence on articulation might be greater in dysarthric speech. Both of these model attributes may aid in diagnosis of the disorder, e. g., for assessing the presence or severity of dysarthria.

Formant frequencies have been studied extensively in prior research with individuals with a variety of communication disorders. Kim *et al.* [59] analyzed acoustic variables associated with speech intelligibility in speakers with dysarthria. They found that acoustic vowel space (computed with F1/F2) along with three other variables were significantly correlated with speech intelligibility.

The primary aim of this application is to automatically assess intelligibility of dysarthric speech.

Perceptual studies and analysis of dysarthric speech has been the standard methodology for intelligibility assessment [50]. Automatic assessment of intelligibility of dysarthric speech is difficult and has only been approached lightly in current literature. We will limit the scope of our study and focus on dysarthrias associated with either Cerebral Palsy or head traumas.

There have been several experiments that have used varying methods to measure intelligibility. Fowler and Housum [40] extracted words of interest from context and played them to subjects at the rate of one every five seconds. The subjects then wrote down what they believe they heard along with a confidence score (1 to 5). Bradlow *et al.* [11] asked subjects to transcribe entire sentences and provide keywords from each sentence. The sentence was considered correct if all five keywords were transcribed correctly. Bard *et al.* [7] also extracted words in the same manner as Fowler and Housum but added noise to the recording to increase difficulty in recognition. In all three methods, the accuracy of transcription over all subjects was considered a measure of intelligibility.

Dysarthria is a broad term for a related set of disorders that affect the movement of speech articulators, which impair intelligibility. There are approximately six major sub-types of dysarthria [21], with five of the major types being:

1. **Spastic dysarthria** is due to upper motor neuron lesions, pyramidal tract damage and especially lesions due to the facial and hypoglossal cranial nerves for jaw and tongue movement. Phonation in spastic dysarthria is harsh, strained with a low sustained pitch [31]. Hypernasality often accompanies /b/, /k/, /p/ and /s/. Bursts of loudness, slow rate of speech and reduced onset time distinction between voiced and unvoiced stops are also associated with spastic dysarthria [43].

2. **Hyperkinetic dysarthria** is due to lesions in the basal ganglia, and often accompanies other involuntary movement. Harsh phonation is comparable to that in spastic dysarthria, although hypernasality is more common and involuntary movements tend to superimpose on voluntary articulations. Commonly slowness in speech is another trait.

3. **Hypokinetic dysarthria** is associated with Parkinsons, and due to lesions in the basal ganglia, or either anti-psychotic medication or blows to the head. Hypokinetic dysarthria results in mono-pitch hoarse phonation with very low monotonous volume. Compulsive syllabic repetition can also occur. It can result in difficulty initiating voluntary speech, or sudden interruption of movement during speech [31].

4. **Ataxic dysarthria** is caused by damage to cerebellar control of respiration, phonation, and

articulation. However, it is primarily characterized by pronounced bursts of loudness. Equal and excessive stress on each spoken syllable is also common. Dis-coordination results in slurred and slow speech, where patients sounds inebriated [31].

5. **Flaccid dysarthria** is caused by damage to lower motor neurons. This may result in complete paralysis of one or more vocal folds, causing breathiness, low volume, increased nasality and monotonous pitch. In unilateral paralysis the jaw may deviate to the weakened side while the tongue moves to the stronger side, sometimes resulting in drooling [31]. Even with many overlapping behaviors between these subgroups, there may also exist some clear delineations. For example, slow speech in spastic dysarthria is more often caused by lengthened syllables, relative to ataxic dysarthria, which is categorized by longer pauses [84].

The final major type is the the mixed condition, a combination of two or more aforementioned types are present. These sub-types were largely based on the primary neurological conditions that cause the dysarthric symptoms.

Common characteristics of dysarthric sub-types have been examined the perceptual characteristics [21]. Within this particular study, 38 perceptual dimensions of speech were analyzed by a panel of three judges. The judges assessed each perceptual characteristics of speech on a seven-point severity scale, with one being normal and seven being severe. The inter-judge reliability was 85% with more recent work showed similar reliability [17]. Seven of the 38 perceptual characteristics were deemed useful in describing the disordered speech [21] across five of the seven neurological groups, which were: monopitch, monoloudness, harsh voice, imprecise consonants, short phrases, reduced stress and distorted vowels.

Many dysarthrias affect the ability to position the tongue properly. Therefore vowels distortions are a common characteristic of dysarthria. The F1 frequency is proportional to how "low" the tongue is, and the F2 frequency is proportional to the "frontness" of the tongue [24]. The vowel space consisting of all possible F1/F2 combinations of vowels that lie on the extrema of the F1/F2 space. The area enclosed by the triangle (or quadrilateral) formed by connecting three (or four) of the corner vowels decreases in proportion to intelligibility [71], which is known as "vowel centralization".

6.2 Methods

There are several approaches that we will employ to measure disordered speech: (1) using estimated targets examine vowel space in F1/F2 space defined by the four-vowel quadrilateral of vowels /iy/,

/ae/, /aa/ and /uw/, (2) goodness of fit, which will examine how well the modeled trajectory fits the observed in terms of RMS error between both trajectories, and (3) coarticulation by examining model parameters, specifically max F_C. We will examine the vowel space in the context of observed values to highlight the limited, yet varied, range of possible observations.

An expanded vowel space, increases the area of possible observations. This expansion implies a potential increase in spectral separability of these four vowels. This larger space may be critical for intelligibility [109, 105, 73, 106]. In fact, this has been measured by observing that less vowel reduction occurs in intelligible speech [76], which could be a result of (1) increased articulatory effort, (2) longer segmental durations, or (3) both. Finally, Bradlow [11] reported that the vowel space of intelligible speech is more spread out than in less intelligible speech, thus we expect to find the vowel-space is smaller in disordered speech. We will describe the disorder speech corpora that is used in this study. In addition to disordered speech, we will use a CLR speech corpora consisting of two speakers that are considered non-disordered speech.

In estimating formant frequency targets we added a regularization term to the error function. This was due to the small number of exemplars for each speaker, which prevents unrealistic values of targets being estimated. We define a regularization function that is computed over all T acoustic events:

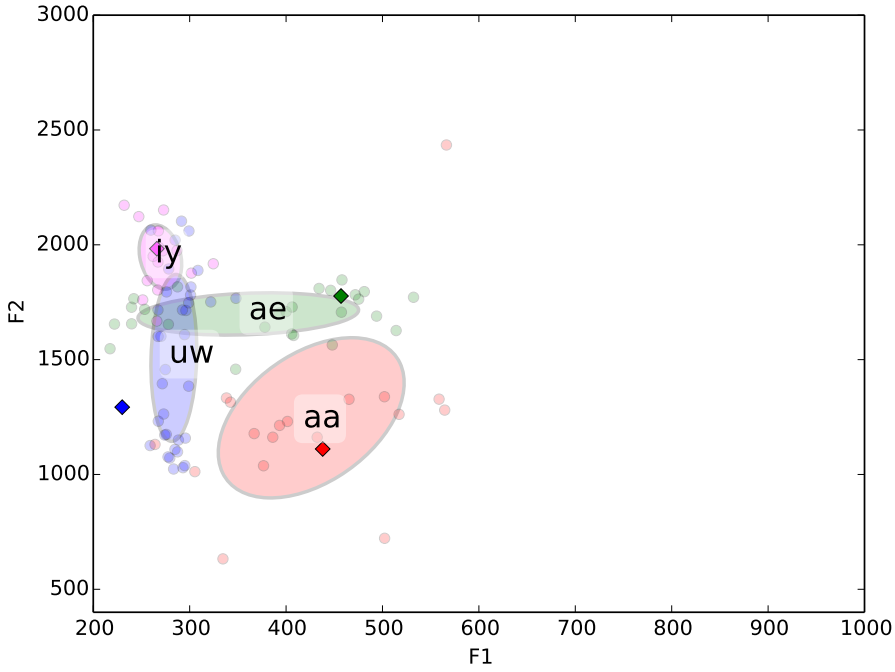
$$R = \alpha \cdot \sum^{T} (obs_t - est_t) \tag{6.1}$$

where *obs* is the median observed formant frequency value, and *est* is the estimated formant frequency target. In our experiments, we used a small value of α of 0.01, which was found through experimentation. Thus, we extended the error formula using in estimated formant frequency targets defined in Section 3.4.1 to include this regularization function:

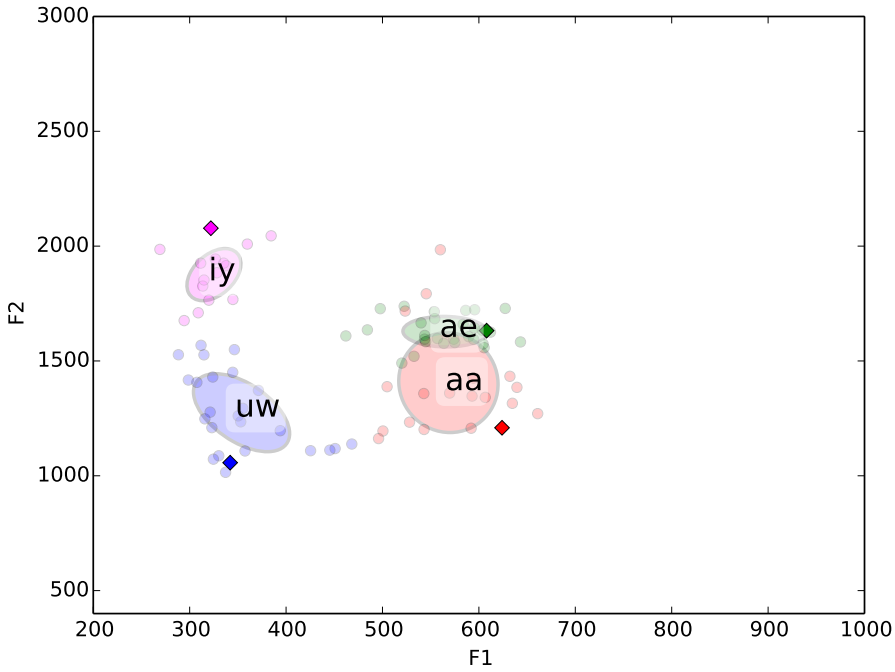
$$G = R + \sqrt{\frac{1}{\sum_{t=0}^{N} w(t)} \sum_{t=0}^{N} w(t) \cdot \left(X(t) - \hat{X}(t)\right)^2} \tag{6.2}$$

6.2.1 Corpora

For our disordered speech corpora, we used the Nemours database of dysarthric speech [75] which is composed of eleven disordered speakers. Of these eleven, we selected six speakers `LL`, `BB`, `MH`, `FB` (mild), `SC` and `JF`, primarily due to accuracy in formant tracking. In Table 6.1 the average percent (Overall) correct identifications for each talker are detailed along with the Frenchay speech assessment scores [32]. There was no speech assessment done for speaker `FB` whose dysarthria was

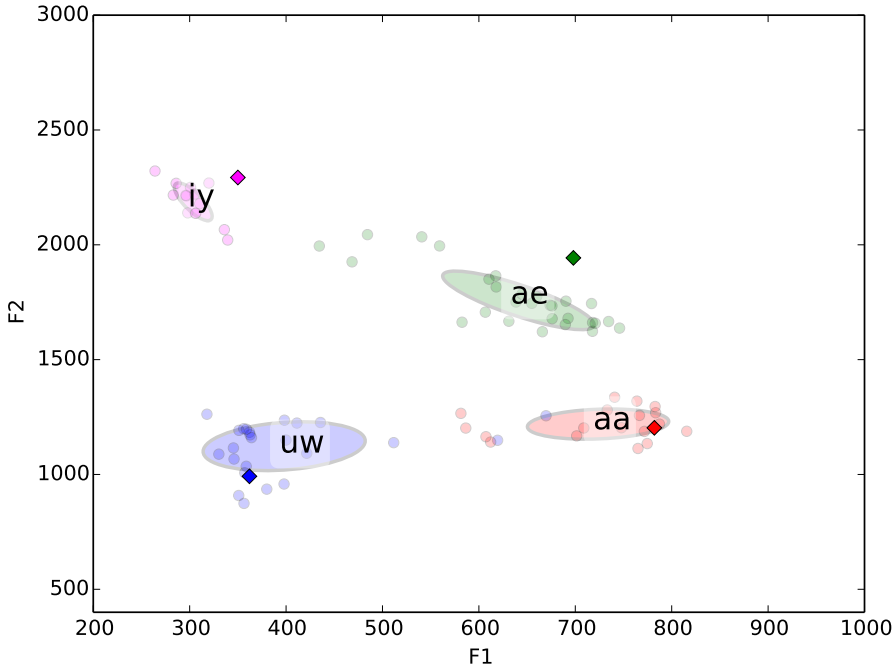


(a) Disordered speaker LL

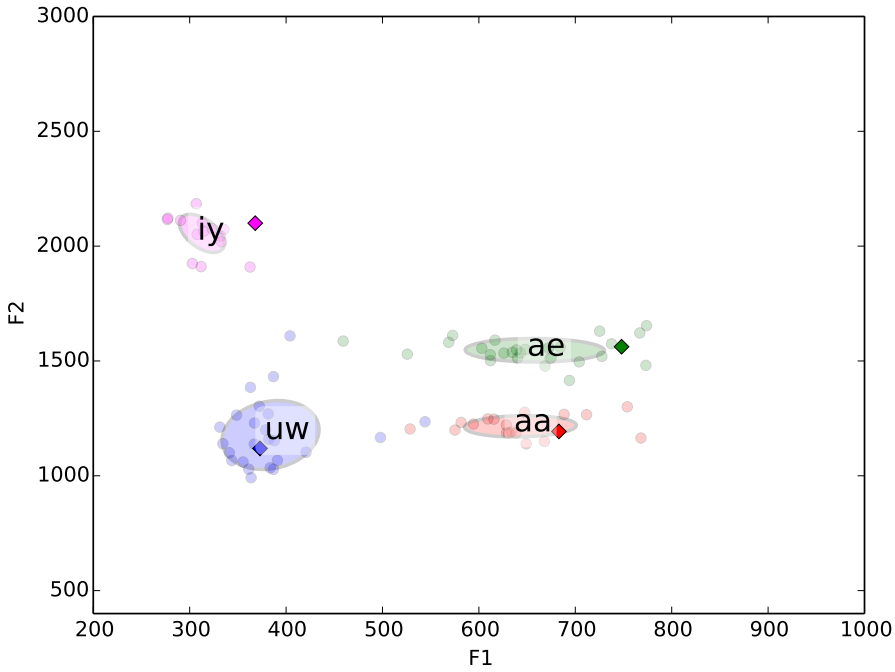


(b) Disordered speaker BB

Figure 6.1: Observed vowel centers with one standard deviation per vowel for disordered speakers. Diamonds indicate estimated targets.

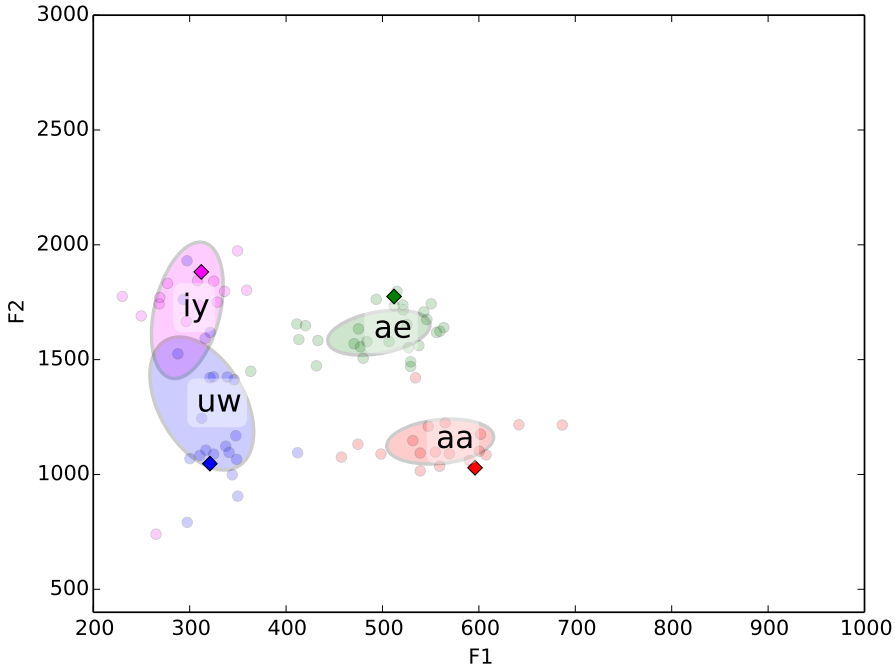


(a) Disordered speaker MH

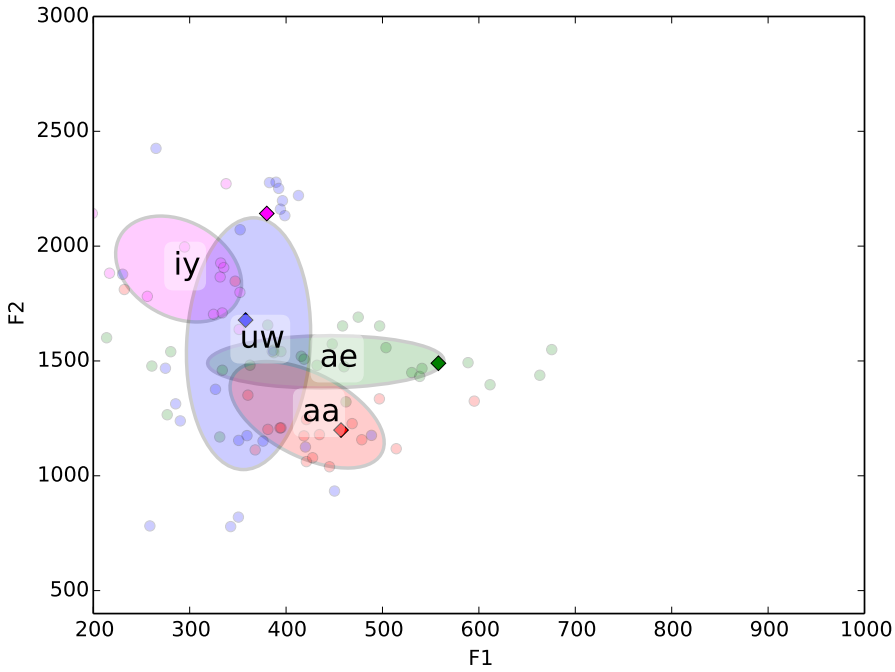


(b) Disordered speaker FB

Figure 6.2: Observed vowel centers with one standard deviation per vowel for disordered speakers. Diamonds indicate estimated targets (continued).

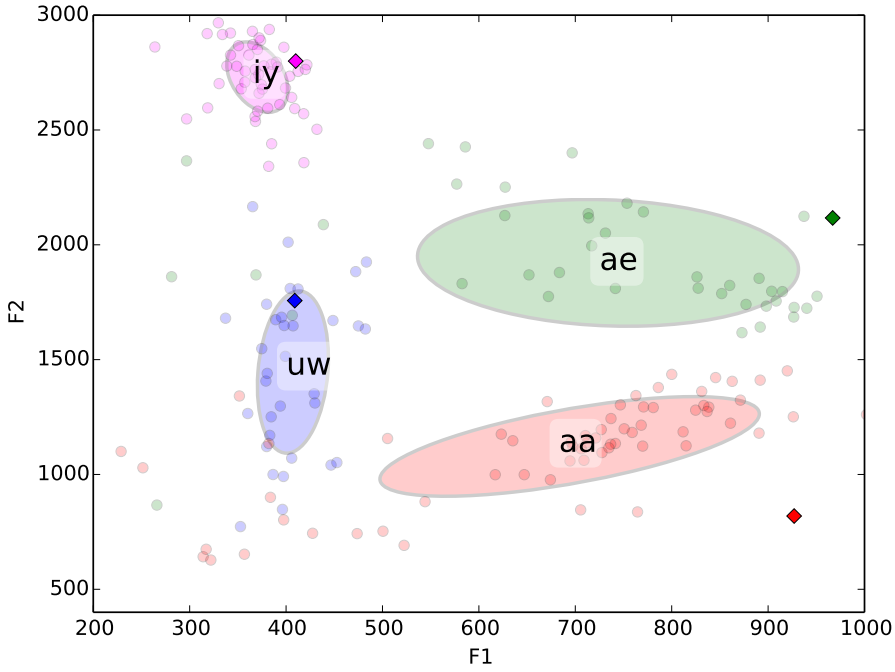


(a) Disordered speaker SC

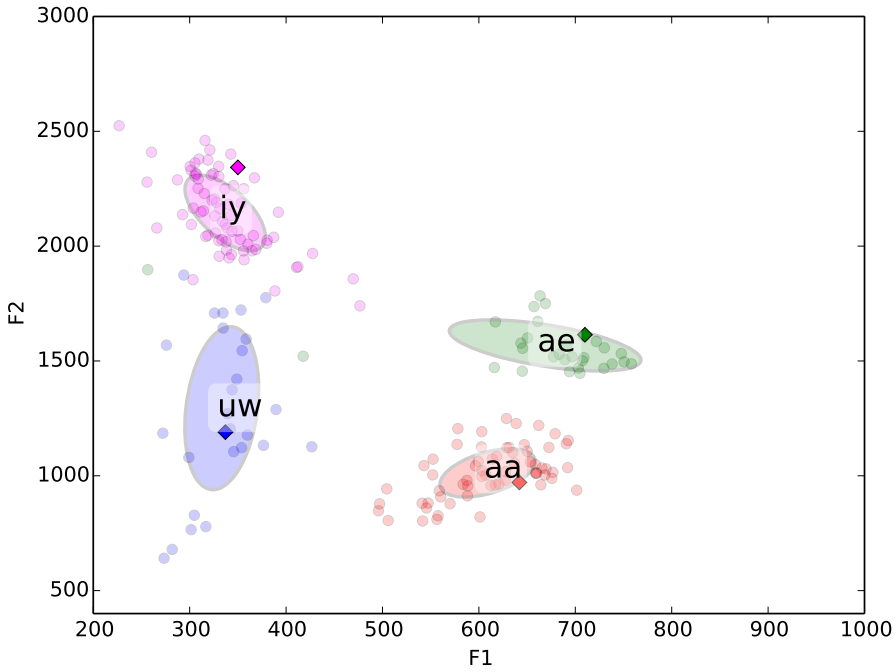


(b) Disordered speaker JF

Figure 6.3: Observed vowel centers with one standard deviation per vowel for disordered speakers. Diamonds indicate estimated targets (continued).



(a) Non-disordered female speaker



(b) Non-disordered male speaker

Figure 6.4: Observed vowel centers with one standard deviation per vowel for non-disordered speakers. Diamonds indicate estimated targets (continued).

extremely mild. The sub-disorders of dysarthria these speakers have not been identified, however, the assessment scores might be somewhat indicative to type.

Speaker	Intelligibility	Tongue	Overall
FB	-	-	92.9
MH	6	6	92.1
BB	8	7	89.7
LL	6	7	84.4
JF	4	4	78.5
SC	3	2	51.5

Table 6.1: Speakers ranked according to average percent correct identifications with the Frenchay Dysarthria assessment scores.

Each nonsense sentence in the database is of the form "The X is Ying the Z." Specific sentences were generated by randomly selecting X and Z ($X \neq Z$) without replacement from a set of 74 monosyllabic nouns and selecting Y without replacement from a set of 37 disyllabic verbs. This process produced 37 sentences from which another 37 sentences were generated by swapping the X and Z tokens in the original set. Thus, over the complete set of 74 sentences, each noun and verb was produced twice by each talker. The target words (X, Y and Z) were chosen based on constraints similar to those used by, for example, Kent *et al.* [58] to provide closed-set phonetic contrasts (e. g., place, manner, and voicing contrasts) within an associated set of four to six words. Thus, all of the target words within a set differ in a single phoneme so that they may be used as alternatives in a closed-response perceptual test for intelligibility.

Our non-disordered speech corpus consisted of two speakers, one male and one female. Similar to the disordered speech corpus, this corpus was composed of nonsense sentences. Each speaker produced 70 syntactically correct, but semantically anomalous sentences, e. g., *They slide far across the tiny clock*, created by randomizing and exchanging words and grammar structures from material described in [108]. All corpora discussed here were composed of scripted speech materials (read speech).

6.3 Results

We estimate targets and coarticulation parameters for each speaker in the Nemours corpus and the non-disordered speakers. Herein we discuss the resulting findings in the context of non-disordered speakers for comparison. These results offer only evidence of a possible trend or patterns. The primary reasoning is that we a small population of disordered speakers with differing underlying disorder types that manifest speech problems in different manners. Furthermore, our non-disordered population size is only two speakers, thus we cannot generalize what attributes a healthy population would take on.

6.3.1 Vowel Space

The Shoelace formula [92] states that given a polygon P made up of points $\{(x_1, y_1), (x_2, y_2)...(x_n, y_n)\}$, the area can be calculated as:

$$Area = \left| \frac{(x_1y_2 - x_2y_1) + (x_2y_3 - x_3y_2) + ... + (x_ny_1 - x_1y_n)}{2} \right| \tag{6.3}$$

Using the Shoelace formula, we computed the area defined by the quadrilateral in F1/F2 space by the vowels /iy/, /ae/, /aa/ and /uw/. Overall, we see a general overall trend of a constricted vowel space. This is in contrast to the non-disordered speakers in Figure 6.4a and Figure 6.4b. Looking at individual observed vowel centers and formant frequency targets each disordered speaker, we find several interesting features. Starting with speakers `LL` (Figure 6.1a) and `SC` (Figure 6.3a), we see all vowels shifted low in F1 space. In F2, `LL` has a considerably low value for /iy/. Speakers `BB` (Figure 6.1b) and `JF` (Figure 6.3b) both display a very centralized F1/F2 space for all vowels. Speakers `MH` (Figure 6.2a) and `FB` (Figure 6.2b) both have a somewhat restricted vowel space, but is not as constricted as others (e. g., `LL`, `BB`), which is consistent with these particular speaker's intelligibility scores (Table 6.1) and for `FB` is consistent with the diagnosis of mild dysarthria.

In Figure 6.5, we see a somewhat linear relationship between the vowel-space created by the estimated formant frequency targets of the four vowels. We find an adjusted r-squared value of 0.614 with six samples and one independent variable. Speaker `SC` has a smaller vowel space, but the intelligibility drops off significantly. In computing the linear regression line, we have removed speaker `SC` due to this speaker's severe dysarthria and low tongue ratings in the Frenchay assessment. The right-most speaker with 100 percent intelligibility is our non-dysarthric male speaker.

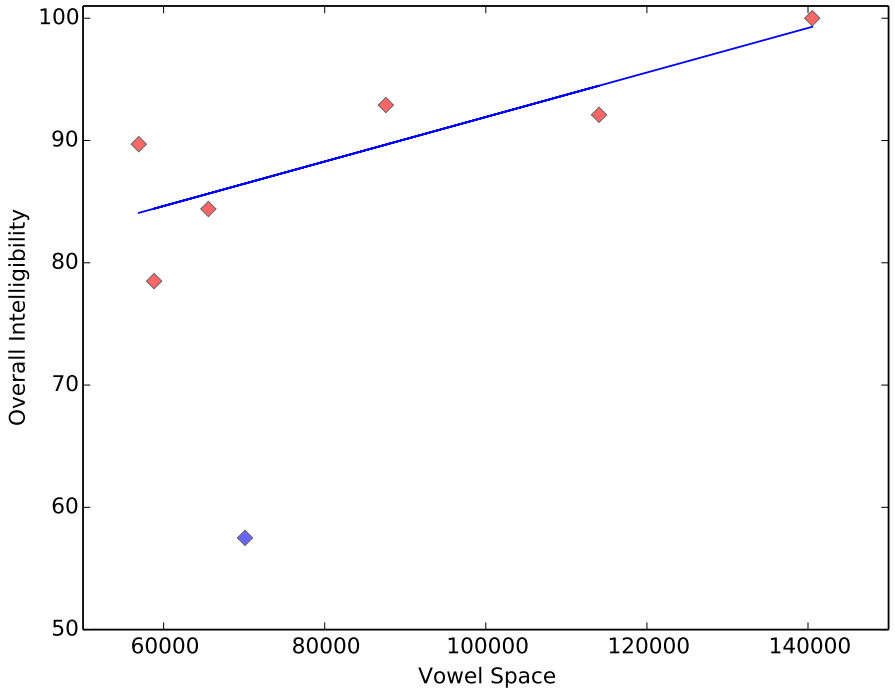


Figure 6.5: Relationship of formant frequency target vowel-space to intelligibility. Linear regression line excludes speaker `SC (blue)`.

6.3.2 Goodness of Fit

We measured the average deviation from model for both non-disordered and disordered speaker, which can be seen in Table 6.2, which shows the average RMS error. As we can see by the error rates, the non-disordered male and female speakers have less errors than the disordered speakers indicating a poor fit to model trajectories. One notable finding is the `FB` (with mild dysarthria) has a good fit that is consistent with their diagnosis. In Figure 6.6 we can see an almost linear relationship between increased model fit error (of F2) and a decrease in intelligibility. We find an adjusted r-squared value of 0.452 with six samples and one independent variable. Once again, speaker `SC` has been omitted in computing the linear regression line as this speaker has severe dysarthria.

6.3.3 Coarticulation Parameters

Examining coarticulation parameters is another form of measurement to see the space of possible observations in formant trajectory movement. Specifically, histograms of max F_C show a compact measure of the variation in articulation. Figures 6.8a and 6.8b show coarticulation values with

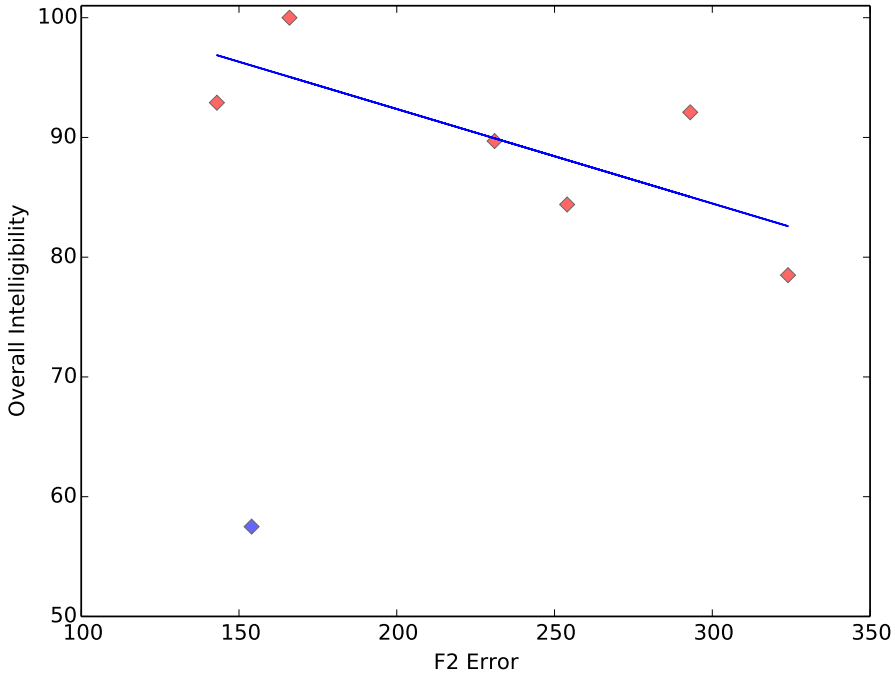
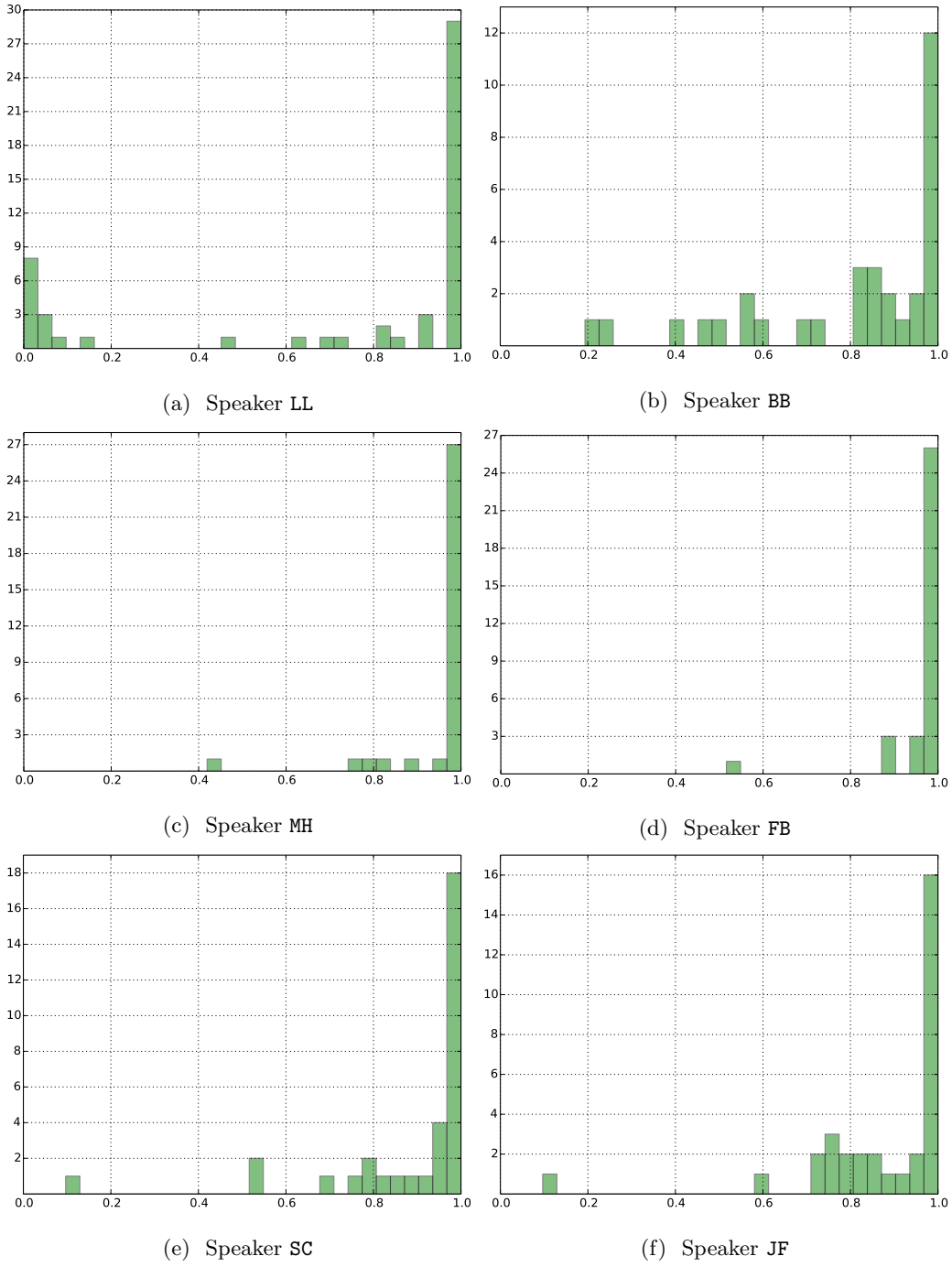


Figure 6.6: Relationship of model fit (of F2) to intelligibility. Linear regression line excludes speaker `SC (blue)`.

Experiment	**F1**	**F2**
male	64	166
female	82	186
`MH`	110	293
`BB`	100	231
`LL`	111	254
`FB`	59	143
`JF`	112	324
`SC`	110	154

Table 6.2: RMS error compared between non-disordered and disordered speakers.

most close to one. Similarly, in disordered speakers, we see most of the values approaching one. Notable is the lack of values between zero and one in disordered speech. However, there is little to show a clear problem in disordered histograms.



(a) Speaker LL

(b) Speaker BB

(c) Speaker MH

(d) Speaker FB

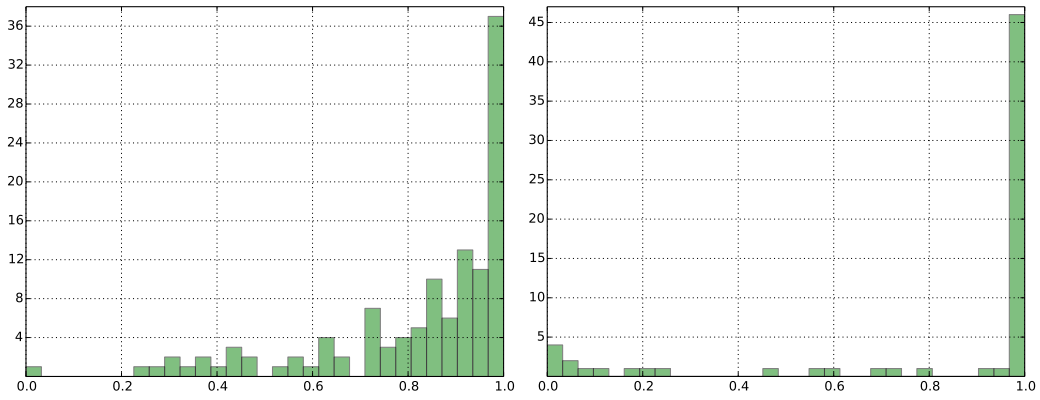
(e) Speaker SC

(f) Speaker JF

Figure 6.7: Histogram of max F_C values representing coarticulation for disordered speakers.

6.4 Discussion and Conclusions

Traditionally, subjective methods were used in the assessment of dysarthric speech intelligibility. These methods are manually administered. We have presented a concept towards the automatic



(a) Speaker `male` (b) Speaker `female`

Figure 6.8: Histogram of max F_C values representing coarticulation for non-disordered speakers.

assessment of intelligibility in dysarthrias. We showed visually in F1/F2 space that observed vowel centers of vowels /iy/, /ae/, /aa/ and /uw/ are greatly centralized, yet varying for each speaker. We successfully applied our coarticulatory model to each disordered speaker, with estimated targets of the same four vowels exhibiting the same centralized behavior. We then measured how we observed trajectories fit the modeled trajectories to find that disordered speakers were not being fit well, even thought the target is appears to be in a valid location. All of these findings were in contrast to the non-disordered speakers. However, in terms of coarticulatory attributes, specifically max F_C, we find similarities between disordered and non-disordered speakers.

In summary, using estimated target deviations and goodness of fit between model and observed formant trajectories might form a simplistic measure for intelligibility estimation. We demonstrated linear relationships in both vowel-space and model to observed error to intelligibility ratings. The variation seen in coarticulation parameters seem insignificant and a larger population of speakers would be necessary to uncover any patterns.

Our primary limitations which are primarily that we have small subset of disordered speakers, which prevent us from deriving a generalized model that predicts intelligibility from vowel space measurements. Furthermore, speakers have varying disorders that lead to various issues in dysarthria characteristics. A more homogeneous population of disordered speakers might uncover interesting patterns in coarticulation that are not visible with the current corpus. Future work might utilized a coarticulatory model of dysarthric speech may provide the ability to "decoarticulate" disordered speech thus increasing intelligibility.

Chapter 7

Conclusion

7.1 Summary

We have investigated coarticulation modeling, specifically, how individual phonemes are influenced by surrounding phonemes. Within this study, we have proposed a data-driven technique for deriving this model from speech alone.

In Chapter 3, we formulated and applied a coarticulation model to continuous speech. In validating the proposed model, we first validated the parameter estimation technique by recovering targets used in a synthetic corpus for F2. We demonstrated that we can recover those targets used to create the trajectories almost completely, even in the presence of noise. Extending the validation to coarticulation model itself, we demonstrated that the model is capturing the required details of speech in order to maintain intelligibility. This validation step entailed resynthesizing speech by replacing observed formants with model trajectories. Finally, we examine coarticulation parameters to find that s parameters are capturing relationships that are known to exist from our acoustic-phonetic knowledge, and p parameters exhibit much variation which provide evidence that justifies an asynchronous modeling approach.

In Chapter 4, we investigated the effect of speech style using our coarticulation model. We first investigated observed vowel centers for eight monophthong vowels in two separate corpora with the same speaker: CVC words, and vowels in unrestricted contexts. First and foremost, vowels in CVC contexts had a larger vowel space. Second, the movement of vowels in CVC contexts from CNV to CLR was also larger than vowels in unrestricted contexts. This was further confirmed in examining estimated formant frequency targets and coarticulation parameters in both corpora. For the speaker that we studied, we find little differences between CLR and CNV speech styles across a large corpus of speech where context is unrestricted.

In Chapter 5, we applied our coarticulation model to the problem of formant tracking. Speaker specific formant frequency targets were estimated the first two formants for 24 speakers, composed

of two males and one female from each of the eight TIMIT dialect regions. Once targets were estimated, the coarticulation parameters of the 192 utterance trajectories with the appropriate set of targets. This estimated trajectory was used as the resultant formant trajectory and compared with hand-corrected published trajectories. Overall, the formant-tracking error rate for male speakers by 31% for F1 and 44% for F2, while reducing the error rate for female speakers by 24% for F1 and 28% for F2. We suspect the difference between male and female error rates can be largely attributed to using male-specific formants for consonants as indicated by the larger RMS error in female model errors. This application seamlessly integrates acoustic-phonetic knowledge into the formant tracking problem, and since the approach is a post-processing step to traditional formant tracking, the combined approach can easily utilize improved formant tracking algorithms.

Finally, in Chapter 6, we analyzed vowel-space defined by formant frequency targets and coarticulation in dysarthric and non-dysarthric speakers. Using both the vowel space defined by estimated formant frequency targets and goodness of fit between model and observed formant trajectories, we showed the foundations of a possible automated intelligibility estimation technique. However, in examining the coarticulation measure $\max F_C$ in dysarthric speakers showed little differences compared to non-disordered speakers. Thus, we find differences primarily in the formant frequency target space between disordered and non-disordered speakers.

7.2 Contributions of this thesis

The following contributions were made in this thesis:

- In this study, we presented a new generic data-driven methodology to estimate vowel and consonant targets for a single speaker. The model described herein is not bound to specific phonemic context, e. g., CVC. Our method assumes the existence of global targets frequencies per acoustic event. We speculate that some phoneme targets have ranges or multiple targets; e. g. it is possible that targets are different in the onset vs coda of a stressed syllable.

- Presented a method for estimating model parameters. In previous studies, we demonstrated the existence of local optima applying our model to simple CVC tokens, which highlighted the difficulties in finding consistent formant targets.

- Demonstrated that our parameter estimation approach can recovered formant frequency targets from a synthetically derived corpus.

- An intelligibility test validated that resynthesized sentences using vocoded-modeled formant trajectories were nearly as intelligible as those using vocoded-observed formant trajectories.

- Applied our coarticulation model to several problem domains: (1) formant tracking where we corrected formant trajectories, (2) discovered model parameters can demonstrate statistically significant differences between CLR and CNV speech, and (3) measure centralization of vowel spaces in dysarthric speech using estimated formant frequency targets and goodness of fit between model trajectories and observed.

7.3 Constraints and Limitations

It should be noted that the corpora used in this research were depended upon manually labeled phoneme identity and boundaries. In some applications, the phoneme identity could be known and we could use force alignment to determine boundaries. Therefore, more work would be needed to fully automate the approaches that we investigated.

The parameter estimation method detailed in Chapter 3 is a brute-force approach. Our primary goal was to explore the parameter space and characterize the error surfaces in parameter estimation. To enable this method to be more readily available a more efficient method would need to be developed. One idea might be to train on a subset of the data as opposed to using a large speaker database. Another approach to speed estimation would be to use acoustic-phonetic knowledge to set initial values of targets and s parameters. The results from chapters were specific to several speakers. Therefore, these results may not be easily generalized to other speakers and applications without further investigation.

Specifically in Chapter 6 there are slight inconsistencies in formant tracking that were manually corrected when extreme errors were detected. In producing a generalized solution in dysarthria speech processing, formants frequency analysis should be avoided and perform modeling by synthesis wherein spectral error is used. However, this would incur a high computational cost, thus an improved optimization method for model parameters should be investigated first.

7.4 Future Work

There are several directions that the research presented in this dissertation could be pursued in the future. Many of the areas discussed would require an expanded set of speakers in order to properly generalize the patterns discovered.

- **Speaker Recognition**, coarticulatory model parameters could be employed as the basis of a model that learns speaker specific attributes, such as structural differences (vocal tract length,

gender, etc.) between speakers, and stylistic differences (due to speaking style, prosody, speaker idiosyncrasy). This model and parameters could be used in a speaker ID task.

- **Speech enhancement**, where conversational speech is modeled, model parameters are transformed, and finally speech resynthesis. As we have discovered that s parameters can be used to determine CNV or CLR speech, however, the exact transformation would need study.
- **Second language acquisition**, often specific sounds in English are difficult for non-native speakers to pronounce (e. g. liquids /r/ and /l/ are often difficult to properly pronounce for Japanese studying English). Subjects speaking a non-native language may fail to properly reach specific targets, where this is trivial in a population of native speakers.
- **Text-to-speech**, as a continuation of work done in Chapter 3, we could use a coarticulation model with specific acoustic event targets with context-dependent coarticulation parameters to generate synthetic speech for a specific speaker. This could be used to change the style from CLR to CNV speech to make the speech more natural.
- **Text-to-formants**, similar to work done by Abari *et al.* [1], using an approach similar to using our coarticulation model in TTS to generate formant trajectories using parameters derived from a specific speaker.
- **Parkison's Disease and Multiple Sclerosis diagnosis**, through a longitudinal study or diagnosis, a coarticulation model based intelligibility measure could serve as a bio-marker of disease prior to any significant reduction in intelligibility.

Appendix A

Acoustic Event Properties

In this dissertation, we use OGIbet throughout. The sonority, manner, place and height have been mapped to a [0, 10] range and are derived from two sources: (1) values for vowels that reconstruct the front-back/open-close dimensions of the International Phonetic Association (IPA) chart, and (2) related work by Kain *et al.* introduced in [52].

OGIbet	**Worldbet**	**IPA**	**Example**	**Sonority**	**Manner**	**Place**	**Height**
aa	A	ɑ	father	4.0	2.0	6.0	1.0
ae	@	æ	bat	4.0	2.0	4.0	1.5
ah	^	ʌ	but	4.0	2.0	6.0	2.0
ao	>	ɔ	caught	4.0	2.0	6.0	2.0
aw	aU	aʊ	about	4.0	1.0	6.0	2.25
ax	&	ə	above	4.0	2.0	5.0	2.5
ay	aI	aɪ	bye	4.0	0.0	5.25	2.25
b	b	b	ban	1.5	7.0	1.0	7.0
ch	tS	tʃ	church	0.75	6.5	3.5	6.0
d	d	d	do	1.5	7.0	3.0	7.0
dZ	dZ	dʒ	judge	2.0	6.5	3.5	6.0
dh	D	ð	thy	2.0	6.0	2.0	6.0
eh	E	ɛ	bet	4.0	2.0	4.0	2.0
el	l=	l	bottle	4.0	2.5	3.0	6.0
em	m=	m	bottom	3.5	3.5	1.0	7.0
en	n=	n	button	3.5	3.5	3.0	7.0
f	f	f	fine	0.5	6.0	1.0	6.0
g	g	ɡ	go	1.5	7.0	8.0	7.0
ih	I	ɪ	bit	4.0	2.0	4.5	3.5
iy	i	iː	beet	4.0	2.0	4.0	4.0

OGIbet	Worldbet	IPA	Example	Sonority	Manner	Place	Height
k	kh	k^h	can	0.75	7.0	8.0	7.0
l	l	l	lent	4.0	3.0	3.0	6.0
m	m	m	me	3.5	4.0	1.0	7.0
n	n	n	no	3.5	4.0	3.0	7.0
ng	N	ŋ	sing	3.5	4.0	8.0	7.0
o	o	o	perro	4.0	2.0	6.0	2.0
ow	oU	oʊ	boat	4.0	2.0	6.0	3.5
oy	>i	ɔi	boy	4.0	0.0	5.0	3.0
p	ph	p^h	pan	0.75	7.0	1.0	7.0
r	r	ɻ	amor	4.0	3.0	5.5	2.0
s	s	s	sign	0.5	6.0	3.0	6.0
sh	S	ʃ	assure	0.5	6.0	4.0	6.0
t	th	t^h	tan	0.75	7.0	3.0	7.0
th	T	θ	thigh	0.5	6.0	2.0	6.0
uh	U	ʊ	book	4.0	2.0	5.5	3.5
uw	u	u	boot	4.0	2.0	6.0	4.0
v	v	v	vine	2.0	6.0	1.0	6.0
w	w	w	went	4.0	3.0	6.0	2.0
y	j	j	yes	4.0	3.0	4.0	4.0
z	z	z	resign	2.0	6.0	3.0	6.0
zh	Z	ʒ	beige	2.0	6.0	4.0	6.0

Appendix B

Synthetic F2 Formant Targets

Synthetic F2 formant targets used in experiments in Chapter 3.

Phoneme	**Target**	**Phoneme**	**Target**
aa	950	k	1900
ae	1500	kcl	2050
ah	1200	l	750
aw_1	1350	m	1200
aw_2	1000	n	1500
ay_1	1150	ng	1400
ay_2	1900	ow_1	1000
b	1100	ow_2	900
bcl	1850	oy_1	700
ch	2500	oy_2	1750
chcl	1950	p	1700
d	1900	pau	1500
dcl	1800	pcl	2300
dh	1800	r	1200
eh	1550	s	2600
er	1350	sh	2400
ey_1	1850	t	2100
ey_2	2050	tcl	2150
f	2000	th	2300
g	1950	uh	1200
gcl	2150	uw	1200
h	2400	v	1800
ih	1700	w	700
iy	2250	y	2300
jh	2350	z	2100
jhcl	2000		

Appendix C

Allen Formant Frequency Targets

Formant frequencies used in synthesis by Allen *et al.* [2].

Phoneme	**F1**	**F2**	**F3**
aa	700	1220	2600
ae	635	1575	2450
ah	620	1220	2550
b	200	1100	2150
ch	350	1800	2820
d	200	1600	2600
dcl	200	1600	2600
dh	270	1290	2540
eh	575	1605	2515
f	340	1100	2080
g	200	1990	2850
h	500	1500	2500
ih	435	1700	2585
iy	300	2045	2960
jh	260	1800	2820
k	300	1990	2850
l	310	1050	2880
m	480	1270	2130
n	480	1340	2470
ng	480	1900	2800
p	400	1100	2150
pau	500	1500	2500
r	310	1060	1380
s	320	1390	2530
sh	300	1840	2750

Phoneme	**F1**	**F2**	**F3**
t	400	1600	2600
th	320	1290	2540
uh	475	1140	2370
uw	335	1075	2200
v	220	1100	2080
w	290	610	2150
y	260	2070	3020
z	240	1390	2530

Appendix D

IEEE Harvard Sentences

Seventy sentences from IEEE-Harvard sentences [80] were used in experiments in Chapter 3 and Chapter 4.

No.	Sentence
1.	Sell the evening after the pine fence.
2.	Wipe the stone closely after its booth.
3.	Place the wagon while losing a ship at home.
4.	A box in a pine creek is a strong rider.
5.	A quick season swam strongly for the west.
6.	Homes and springs lay under sweet sun.
7.	Run a ship below sixteen quick soldiers.
8.	The grand evening lost a hot ship.
9.	The watchful woman plunged before the rare chick.
10.	A straight wagon served the perfect girl.
11.	Use a wheel to cook a hard sheet.
12.	The new juice strikes a hot period.
13.	Kick water far above the tool.
14.	Strike the fast chicken to get the sun.
15.	The wise ball slid after the tiny trash.
16.	Slide a chick or face the hill before the lake.
17.	The small rope broke below the dull lake.
18.	Keep the saw under a hot state.
19.	The hole under the stone leaves rusty planks.
20.	A short path got a cool week.
21.	Feed the grass and mend a rod for the rosebush.
22.	The sweet cat swam to the strong saw.
23.	Strike the girl to lift the great dune.
24.	A river of salmon ruled to cure a path.
25.	A bright canoe slid on the coat.

No.	Sentence
26.	A nice state of evening is perfect to face.
27.	Slide the cars through a stray blue lake.
28.	A crooked ball was using the small floor.
29.	Always pass corns across the mouse.
30.	The sea rained and left over the same fires.
31.	Hit the face but give the box to a draft.
32.	The soft swan reached a great breeze.
33.	Simple dishes help cows and ships.
34.	Seize the morning above the winding bright hogs.
35.	She helped, read, and served a lazy evening.
36.	The wide button stalled after a shiny sea.
37.	A cool fall reached under the raging mouse.
38.	The great state hung a frosty rosebush.
39.	Bind your feet with a wide circus.
40.	Slide a bus under the quick young pirate.
41.	A huge sickness hit the same sun.
42.	The next show failed a frosty spring.
43.	Run the seven ships above the tank.
44.	The tall books move the frosty planks.
45.	A crooked box reads to a smooth coat.
46.	Reach the pencil or pass the saw through a rope.
47.	The frosty load tore across the pink rider.
48.	Face the heat while breaking a coat at the source.
49.	A perfect hike rains on the fine snow.
50.	The silver sea took the dirty gang.
51.	The tank was cutting round balls and rice.
52.	Seven clear mice kicked in the cars.
53.	Turn the wagon over the firm costly grass.
54.	A hard shoulder pressed the early storm.
55.	We often come to the pink snowdrifts.
56.	Break a coat but hit the edge before the store.
57.	Seven trucks were hoisted below a well-oiled mouse.
58.	Lead the kittens and run the chicks to the porch.
59.	The dying maze got lost in the tall field.
60.	Turn a lemon when it writes badly.

No.	Sentence
61.	Hop a box before the winding button.
62.	The steady hours were fed over a droopy saw.
63.	The seven sums rule this large cushion.
64.	The ship against the stone bound a strong morning.
65.	Move the maze near the sleeping pup.
66.	Run the wagon before wise response.
67.	Women and stockings run in fast juice.
68.	The weak rosebush broke under a dirty punch.
69.	The blue gang stalled a truck above the cow.
70.	A boy ran and placed fine garbage.

Bibliography

[1] K. Abari, T. G. Csapó, B. P. Tóth, and G. Olaszy. From text to formantsindirect model for trajectory prediction based on a multi-speaker parallel speech database. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] J. Allen, M. S. Hunnicutt, and D. Klatt, editors. *From text to speech: The MITalk system.* Cambridge University Press, 1987.

[3] A. Amano-Kusumoto and J.-P. Hosom. The effect of formant trajectories and phoneme durations on vowel intelligibility. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4677–4680. IEEE, 2009.

[4] A. Amano-Kusumoto and J.-P. Hosom. Effect of speaking style and speaking rate on formant contours. In *ICASSP*, pages 4202–4205, 2010.

[5] B. Atal. Efficient coding for LPC parameters by temporal decomposition. In *Proc. Int. Conf. Acoust., Speech and Signal Processing*, pages 81–84, 1983.

[6] B. O. Bush, J.-P. Hosom, A. Kain, A. Amano-Kusumoto. Using a genetic algorithm to estimate parameters of a coarticulation model. In *Interspeech*, pages 2677–2680, 2011.

[7] E. Bard, C. Sotillo, A. Anderson, G. Doherty-Sneddon, and A. Newlands. The control of intelligibility in running speech. In *Proceedings of the XIIIth International congress of Phonetic Sciences*, volume 4, pages 188–191, 1995.

[8] F. Bell-Berti and K. S. Harris. A temporal model of speech production. *Phonetica*, 38(1-3):9–20, 1981.

[9] A. Bradlow, N. Kraus, T. Nicol, T. McGee, J. Cunningham, S. Zecker, and T. Carrell. Effects of lengthened formant transition duration on discrimination and neural representation of synthetic cv syllables by normal and learning- disabled children. *Journal of Acoustical Society of America*, 106:2086–2096, 1999.

[10] A. R. Bradlow, N. Krause, and E. Hayes. Speaking clearly for children with learning disabilities: sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, 46:80–97, 2003.

[11] A. R. Bradlow, B. M. Torretta, and D. B. Pisoni. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272, 1996.

[12] L. Brancazio and C. A. Fowler. On the relevance of locus equations for production and perception of stop consonants. *Perception & psychophysics*, 60(1):24–50, 1998.

[13] D. J. Broad and F. Clermont. A methodology for modeling vowel formant contours in CVC context. *Journal of the Acoustical Society of America*, 81(1):155–165, 1987.

[14] D. J. Broad and F. Clermont. Linear scaling of vowel-formant ensembles (VFEs) in consonantal contexts. *Speech Communication*, 37(3-4):175–195, 2002.

[15] D. J. Broad and F. Clermont. Target-locus scaling methods for modeling families of formant transitions. *Journal of Phonetics*, 38(3):337–359, 2010.

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

[17] K. Bunton, R. D. Kent, J. R. Duffy, J. C. Rosenbek, and J. F. Kent. Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of speech, language, and hearing research*, 50(6):1481–1495, 2007.

[18] B. O. Bush and A. Kain. Estimating phoneme formant targets and coarticulation parameters of conversational and clear speech. In *ICASSP*, pages 8017–8021, 2013.

[19] B. O. Bush and A. Kain. Modeling coarticulation in continuous speech. In *Interspeech*, pages 193–197, 2014.

[20] S. Chennoukh, R. Carre, and B. Lindblom. Locus equations in the light of articulatory modeling. *The Journal of the Acoustical Society of America*, 102(4):2380–2389, Oct. 1997.

[21] F. L. Darley, A. E. Aronson, and J. R. Brown. Differential diagnostic patterns of dysarthria. *Journal of Speech, Language, and Hearing Research*, 12(2):246–269, 1969.

[22] P. C. Delattre. From acoustic cues to distinctive features. *Phonetica*, 18:198–230, 1968.

[23] P. C. Delattre, A. M. Liberman, and F. S. Cooper. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27(4):769–773, July 1955.

[24] J. R. Deller, J. G. Proakis, and J. H. Hansen. *Discrete-time processing of speech signals*. IEEE New York, NY, USA:, 2000.

[25] L. Deng, M. Aksmanovic, D. Sun, and J. Wu. Speech recognition using hidden markov models with polynomial regression functions as nonstationary states. In *IEEE Transactions on Speech and Audio Processing*, pages 507–520, 1994.

[26] L. Deng and D. Braam. Context-dependent markov model structured by locus equations: Applications to phonetic classification. *The Journal of the Acoustical Society of America*, 96(4):2008–2025, 1994.

[27] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan. A database of vocal tract resonance trajectories for research in speech processing. In *ICASSP*, pages 369–372, 2006.

[28] L. Deng, L. J. Lee, H. Attias, and A. Acero. A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–557. IEEE, 2004.

[29] L. Deng and D. X. Sun. A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features. *The Journal of the Acoustical Society of America*, 95(5):2702–2719, 1994.

[30] A. M. V. Dijk-Kappers and S. M. Marcus. Temporal decomposition of speech. *Speech Communication*, 8(2):125 – 135, 1989.

[31] J. R. Duffy. *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[32] P. Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980.

[33] K. Evanini, S. Isard, and M. Liberman. Automatic formant extraction for sociolinguistic analysis of large corpora. In *Proc. of Interspeech*, volume 165, pages 1655–1658, 2009.

[34] G. Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

[35] E. Farnetani and D. Recasens. Coarticulation models in recent speech production theories. *Coarticulation: Theory, data and techniques*, pages 31–65, 1999.

[36] S. H. Ferguson and D. Kewley-Port. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 112(1):259–271, 2002.

[37] S. H. Ferguson and D. Kewley-Port. Talker Differences in Clear and Conversational Speech: Acoustic Characteristics of Vowels. *J Speech Lang Hear Res*, 50(5):1241–1255, 2007.

[38] C. A. Fowler. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8(1):113–133, 1980.

[39] C. A. Fowler. Current perspective on language and speech production: A critical overview. 1985.

[40] C. A. Fowler and J. Housum. Talkers' signaling of new and old words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5):489–504, 1987.

[41] T. Frank and C. H. Craig. Comparison of the auditec and rintelmann recordings of the nu-6. *Journal of Speech and Hearing Disorders*, 49(3):267–271, 1984.

[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. FFM TIMIT Acoustic-Phonetic Continuous Speech Corpus Secondary (Far Field) Microphone Recordings CD-ROM, 1993. National Institute of Standards and Technology, NTIS Order NO. PB95-504569.

[43] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang. Audiovisual phonologic-feature-based recognition of dysarthric speech. Citeseer, 2006.

[44] V. Hazan and D. Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the American Academy of Audiology*, 116(5):3108–3118, 2004.

[45] K. S. Helfer. Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal of the American Academy of Audiology*, 9:234–242, 1998.

[46] W. L. Henke. Dynamic articulatory model of speech production using computer simulation. Technical report, DTIC Document, 1966.

[47] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler. Acoustic Characteristics of American English Vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111, May 1995.

[48] J. H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, 1975.

[49] J.-P. Hosom. Automatic phoneme alignment based on acoustic-phonetic modeling. In *Interspeech*. Citeseer, 2002.

[50] K. C. Hustad. Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopaedica*, 58(3):217–228, 2006.

[51] K. Iskarous, C. A. Fowler, and D. Whalen. Locus equations are an acoustic expression of articulator synergy. *The Journal of the Acoustical Society of America*, 128:2021–2032, 2010.

[52] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *The Journal of the Acoustical Society of America*, 124(4):2308–2319, 2008.

[53] A. Kain, J.-P. Hosom, X. Niu, J. van Santen, M. Fried-Oken, and J. Staehely. Improving the Intelligibility of Dysarthric Speech. *Speech Communication*, 49(9):743–759, 2007.

[54] A. Kain and J. Santen. A speech model of acoustic inventories based on asynchronous interpolation. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[55] A. Kain and J. van Santen. Compression of acoustic inventories using asynchronous interpolation. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 83–86. IEEE, 2002.

[56] A. Kain and J. van Santen. Unit-selection text-to-speech synthesis using an asynchronous interpolation model. In *Proceedings of 6th ISCA Workshop on Speech Synthesis*, 2007.

[57] R. D. Kent and F. D. Minifie. Coarticulation in Recent Speech Production Models. *Journal of Phonetics*, 5(2):115–133, 1977.

[58] R. D. Kent, G. Weismer, J. F. Kent, and Rosenbek. J.c.: Toward phonetic intelligibility testing in dysarthria. In *In: Journal of Speech and Hearing Disorders*, pages 482–499, 1989.

[59] Y. Kim, R. D. Kent, and G. Weismer. An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech, Language, and Hearing Research*, 54(2):417–429, 2011.

[60] Y. Kim, G. Weismer, R. D. Kent, and J. R. Duffy. Statistical models of f2 slope in relation to severity of dysarthria. *Folia Phoniatr Logop*, 61:329–335, 2009.

[61] K. Kirchhoff. Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments. In *PROCEEDINGS OF ICSLP*, pages 891–894, 1998.

[62] D. H. Klatt. Review of Text-to-Speech Conversion for English. *The Journal of the Acoustical Society of America*, 82(3):737–793, 1987.

[63] J. C. Krause and L. D. Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America*, 15(1):362–378, 2004.

[64] D. Krull. Second formant locus patterns as a measure of consonant-vowel coarticulation. *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS)*, 5:43–61, 1987.

[65] D. Krull. Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS)*, 10:87–108, 1989.

[66] M. Lee, J. Van Santen, B. Mobius, and J. Olive. Formant tracking using context-dependent phonemic information. *Speech and Audio Processing, IEEE Transactions on*, 13(5):741–750, 2005.

[67] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1, 1954.

[68] B. Lindblom. Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, 35(5):1773–1781, 1963.

[69] B. Lindblom. Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling*, pages 403–439. Springer, 1990.

[70] B. Lindblom and H. M. Sussman. Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, 40(1):1–19, 2012.

[71] H.-M. Liu, F.-M. Tsao, and P. K. Kuhl. The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America*, 117(6):3879–3889, 2005.

[72] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[73] P. A. McRae, K. Tjaden, and B. Schoonings. Acoustic and perceptual consequences of articulatory rate change in parkinson disease. *Journal of Speech, Language, and Hearing Research*, 45(1):35–50, 2002.

[74] D. D. Mehta, D. Rudoy, and P. J. Wolfe. Kalman-based autoregressive moving average modeling and inference for formant and antiformant trackinga). *The Journal of the Acoustical Society of America*, 132(3):1732–1746, 2012.

[75] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell. The nemours database of dysarthric speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1962–1965. IEEE, 1996.

[76] S. J. Moon and B. Lindblom. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 9(1):40–55, 1994.

[77] B. Moore. *An Introduction to the Psychology of Hearing*. Acad. Press, 2003.

[78] M. Mulligan, J. Carpenter, J. Riddel, M. K. Delaney, G. Badger, and P. K. R. Tandan. Intelligibility and the acoustic characteristics of speech in amyotrophic lateral sclerosis (als). *Journal of Speech and Hearing Research*, 37:496–503, 1994.

[79] X. Niu and J. P. H. van Santen. A formant-trajectory model and its usage in comparing coarticulatory effects in dysarthric and normal speech. In *Proc. of the Third International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy, Dec. 2003.

[80] I. of Electrical, E. Engineers, I. Audio, and E. G. S. Committee. *IEEE Recommended Practice for Speech Quality Measurements*. IEEE Std. The Institute of Electrical and Electronics Engineers, 1969.

[81] S. E. G. Öhman. Coarticulation in VCV Utterances: Spectrographic Measurements. *The Journal of the Acoustical Society of America*, 39:151–168, 1966.

[82] S. E. G. Öhman. Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2):310–320, 1967.

[83] J. Olive, A. Greenwood, and J. Coleman. *Acoustics of American English speech: A Dynamic Approach*. Cambridge University Press, Cambridge, 1987.

[84] Y. Ozawa, O. Shiromoto, F. Ishizaki, and T. Watamori. Symptomatic differences in decreased alternating motion rates between individuals with spastic and with ataxic dysarthria: an acoustic analysis. *Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP)*, 53(2):67–72, 2001.

[85] I. Ozbek and M. Demirekler. Vocal tract resonances tracking based on voiced and unvoiced speech classification using dynamic programming and fixed interval kalman smoother. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4217–4220, March 2008.

[86] J. S. Perkell and C. Chiang. Preliminary support for a" hybrid model" of anticipatory coarticulation. In *Proceedings of the 12th International Conference of Acoustics*, pages A3–A6, 1986.

[87] G. E. Peterson and H. L. Barney. Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, 24(2):175–184, March 1952.

[88] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28:96–103, 1985.

[89] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446, 1986.

[90] R. K. Potter and G. E. Peterson. The representation of vowels and their movements. *The Journal of the Acoustical Society of America*, 20(4):528–535, 1948.

[91] B. Prakash. Acoustic measures in the speech of children with stuttering and normal nonfluency - a key to differential diagnosis. In *Proc. of the workshop on Spken Language Processing*, pages 49–57, 2003.

[92] H. Pretzsch. *Forest Dynamics, Growth and Yield: From Measurement to Model*. Springer, 2009.

[93] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*, page 370. Prentice Hall, Englewood Cliffs, NJ, 1993.

[94] H. B. Richards and J. S. Bridle. The HDM: A Segmental Hidden Dynamic Model of Coarticulation. In *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP 1999), volume I*, pages 357–360, 1999.

[95] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, December 2002.

[96] K. Sjölander and J. Beskow. WaveSurfer — an open source speech tool. In *Proc. of ICSLP*, pages 464–467, 2000. http://www.speech.kth.se/snack/.

[97] M. S. Sommers and J. Barcroft. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *Journal of the Acoustical Society of America*, 119(4):2406–2416, 2006.

[98] M. S. Sommers, L. C. Nygaard, and D. B. Pisoni. Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, 96(3):1314–1324, 1994.

[99] R. Sproat, editor. *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*. Kluwer, 1998.

[100] K. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, Mass., 1998.

[101] H. M. Sussman, H. A. McCaffrey, and S. A. Matthews. An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, 90(3):1309–1325, 1991.

[102] H. M. Sussman and J. Shore. Locus equations as phonetic descriptors of consonantal place of articulation. *Perception & psychophysics*, 58(6):936–946, 1996.

[103] D. Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition costs. *The Journal of the Acoustical Society of America*, 82(S1):S55–S55, 1987.

[104] K. Tjaden. An acoustic study of coarticulation in dysarthric speakers with parkinson disease. *Journal of speech, language, and hearing research*, 43(6):1466, 2000.

[105] K. Tjaden and G. E. Wilding. Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 47(4):766–783, 2004.

[106] G. S. Turner, K. Tjaden, and G. Weismer. The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 38(5):1001–1013, 1995.

[107] R. J. Van Son and L. C. Pols. Formant movements of dutch vowels in a text, read at normal and fast rate. *The Journal of the Acoustical society of America*, 92:121, 1992.

[108] N. Vaughan, D. Storzbach, and I. Furukawa. Sequencing versus nonsequencing working memory in understanding of rapid speech by older listeners. *Journal of the American Academy of Audiology*, 17(7):506–518, 2006.

[109] G. Weismer, J.-Y. Jeng, J. S. Laures, R. D. Kent, and J. F. Kent. Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica*, 53(1):1–18, 2001.

[110] W. A. Wickelgren. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76:1–15, 1969.

[111] J. Wouters and M. Macon. Control of spectral dynamics in concatenative speech synthesis. *IEEE Trans. Speech and Audio Proc.*, 9(1):30–38, January 2001.

[112] K. Xia and C. Espy-Wilson. A new strategy of format tracking using dynamic programming. In *Proc. Int. Conf. Spoken Lang. Process (ICSLP)*, pages 1–4, 2000.

[113] A. Yadollahi and Z. Moussavi. Formant analysis of breath and snore sounds. In *Proc. of IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 2563–2566, 2009.

[114] V. Zue, S. Seneff, and J. Glass. Speech database development at mit: Timit and beyond. *Speech Communication*, 9(4):351–356, 1990.

Biographical Note

Brian O. Bush was born on September 15, 1972 in Chillicothe, Ohio. He received his bachelor degrees in both Applied Mathematics and Computer Science from Hiram College in 1994. He attended Ohio University for graduate school, where he received a Masters of Science in Industrial and Systems Engineering in 1996, before fleeing to industry. He co-founded RuleSpace, LLC in 2005 that built NLP solutions for network filtering products that powered some of the world's largest search engines and cell-phone carriers. He served as the chief architect until RuleSpace was acquired by Symantec Corp in 2010. Brian is an author on six peer-reviewed conference proceedings, two peer-reviewed workshop papers, two journal articles and one U.S. Patent (no. 8,271,263).