Optimization of Acoustic Feature Extraction from Dysarthric Speech

by

Thomas M. DiCicco, Jr.

B.S., Boston University (2001)
M.S., University of California, San Diego (2003)

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES & TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOMEDICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2010

Signature of Author: ____________________________
Harvard-MIT Division of Health Sciences & Technology
August, 2009

Certified by: ____________________________
Rupal Patel, PhD, CCC-SLP
Associate Professor of Speech Language Pathology, Northeastern University
HST Speech & Hearing Biosciences & Technology Affiliated Faculty Member
Thesis Supervisor

Accepted by: ____________________________
Ram Sasisekharan, PhD
Director, Harvard-MIT Division of Health Sciences and Technology
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

Optimization of Acoustic Feature Extraction from Dysarthric Speech

by

Thomas M. DiCicco, Jr.

Submitted to the Harvard-MIT Division of Health Sciences & Technology in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biomedical Engineering

Abstract

Dysarthria is a motor speech disorder characterized by weak or uncoordinated movements of the speech musculature. While unfamiliar listeners struggle to understand speakers with severe dysarthria, familiar listeners are often able to comprehend with high accuracy. This observation implies that although the speech produced by an individual with dysarthria may appear distorted and unintelligible to the untrained listener, there must be a set of consistent acoustic cues that the familiar communication partner is able to interpret. While dysarthric speech has been characterized both acoustically and perceptually, most accounts tend to compare dysarthric productions to those of healthy controls rather than identify the set of reliable and consistently controlled segmental cues. This work aimed to elucidate possible recognition strategies used by familiar listeners by optimizing a model of human speech recognition, Stevens' Lexical Access from Features (LAFF) framework, for ten individual speakers with dysarthria (SWDs). The LAFF model is rooted in distinctive feature theory, with acoustic landmarks indicating changes in the manner of articulation. The acoustic correlates manifested around landmarks provide the identity to articulator-free (manner) and articulator-bound (place) features. SWDs created weaker consonantal landmarks, likely due to an inability to form complete closures in the vocal tract and to fully release consonantal constrictions. Identification of speaker-optimized acoustic correlate sets improved discrimination of each speaker's productions, evidenced by increased sensitivity and specificity. While there was overlap between the types of correlates identified for healthy and dysarthric speakers, using the optimal sets of correlates identified for SWDs adversely impaired discrimination of healthy speech. These results suggest that the combinations of correlates suggested for SWDs were specific to the individual and different from the segmental cues used by healthy individuals. Application of the LAFF model to dysarthric speech has potential clinical utility as a diagnostic tool, highlighting the fine-grain components of speech production that require intervention and quantifying the degree of impairment.

Thesis Supervisor: Rupal Patel
Title: Associate Professor of Speech Language Pathology, Northeastern University
HST Speech & Hearing Biosciences & Technology Affiliated Faculty Member

Acknowledgements

First and foremost, I would like to sincerely thank my advisor, Dr. Rupal Patel, for her time, guidance, and patience. Her work ethic, intellect, and leaderships skills motivated me during my time as a member of the Communication Analysis and Design Laboratory (CadLab). I am grateful that she allowed me a great deal of freedom but was always available with advice and support. As I move on to new endeavors I can only hope that I have absorbed and am able to apply at least some of the qualities that have enabled her to excel.

Thank you to my committee members Dr. Robert Hillman, Dr. Deb Roy, and Dr. Janet Slifka for selflessly investing their time and intellect. I was very lucky to have a committee that came from different specialties and possessed such a broad range of knowledge. They challenged me and provided feedback that guided and intimately shaped this work. This thesis would not have been possible without their help and talents.

I would like to acknowledge Dr. Tannin Schmidt and Dr. Cara Stepp for their editing skills.

To my fellow members of the CadLab, both past and present, you made lab an enjoyable place to be everyday, and sometimes, especially recently, late into the night.

To my classmates (SHBT, EECS, and Sloan) I enjoyed having the opportunity to work with and learn from such bright individuals. You taught me how to question and think in so many ways. Cara and Manny, you guys were my best friends and sanity at MIT.

Thank you to my friends for standing with me in all that I do and for the many great memories you have helped create in my life.

To my parents and sister, thank you for your unconditional love and for encouraging me to make the most of my opportunities.

Lastly, to my new wife Andrea you make my already charmed life that much better. I apologize that I have been essentially absent for the two months of our marriage and thank you for understanding and support. I am excited for our new life together.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Dysarthria is a motor speech disorder characterized by weak or uncoordinated movements of the speech musculature. Level of impairment, measured using speech intelligibility, can range from mild to severe. Interestingly, although unfamiliar listeners struggle to understand speakers with severe dysarthria, familiar listeners are often able to comprehend with high accuracy (Beukelman & Yorkston, 1980; D'Innocenzo, Tjaden, & Greenman, 2006; Deller, Hsu, & Ferrier, 1991; DePaul & Kent, 2000; Hustad & Cahill, 2003; King & Gallegos-Santillan, 1999; Liss, Spitzer, Caviness, & Adler, 2002; Neilson & O'Dwyer, 1984; Spitzer, Liss, Caviness, & Adler, 2000; K. K. Tjaden & Liss, 1995). This observation implies that while dysarthric speech may appear distorted and unintelligible to the untrained listener, there must be a set of consistent acoustic cues that the familiar communication partner is able to interpret. This notion of consistent yet distorted speech is expected given that dysarthria is a motor execution disorder rather than a programming deficit[1]. While dysarthria has been characterized acoustically and perceptually, most previous work has focused on cataloging differences between dysarthric and typical speech (Ansel & Kent, 1992; Duffy, 2005; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young, & Quinn, 1980) and less so on identifying the segmental cues that are preserved in dysarthria and potentially used by familiar listeners.

[1] For comparison, consistent production would not be expected for individuals with apraxia because apraxia is a programming disorder.

We hypothesized that speakers with dysarthria use a different set of segmental cues than healthy speakers given their physiological constraints. This phenomenon known as cue trading (i.e. substitution, deletion, or addition of expected or unexpected cues), in which speakers with dysarthria (SWDs) rely on alternative cue combinations to compensate for their impairments, has been observed with prosodic cues (e.g. the use of duration instead of pitch to signal questions; Le Dorze, Ouellet, & Ryalls, 1994; Patel, 1999, 2002a, 2002b, 2004; Patel & Campellone, 2009; Vance, 1994) but had not yet been explored at the segmental level.

The cue trading hypothesis was tested by assuming Stevens' Lexical Access from Features (LAFF) paradigm as a model for human speech segmentation and lexical recognition (Liu, 1996; Park, 2008; Stevens, 2002). Given that dysarthria is heterogeneous in its presentation, affecting respiratory, phonatory, articulatory, and resonatory subsystems of speech production, acquiring large amounts of speech data is difficult. Therefore, using an established model of human speech recognition limited the exploration space.

The base LAFF model (Park, 2008) was assumed as a benchmark for recognition of healthy speech and could be construed as an approximation of the strategies used by unfamiliar communication partners when attempting to decode dysarthric speech (Figure 1.1). The LAFF paradigm incorporated spectral and temporal components derived from empirical studies of healthy speech production and perception. Given that previous studies of dysarthric speech have documented that SWDs do not manifest phonetic features in a manner consistent with healthy speech (Ansel & Kent, 1992; Duffy, 2005; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980), the LAFF model in its base form was inadequate for dysarthric speech. Therefore, per-speaker optimization of the model for ten SWDs was performed to emulate potential recognition strategies employed by familiar listeners.

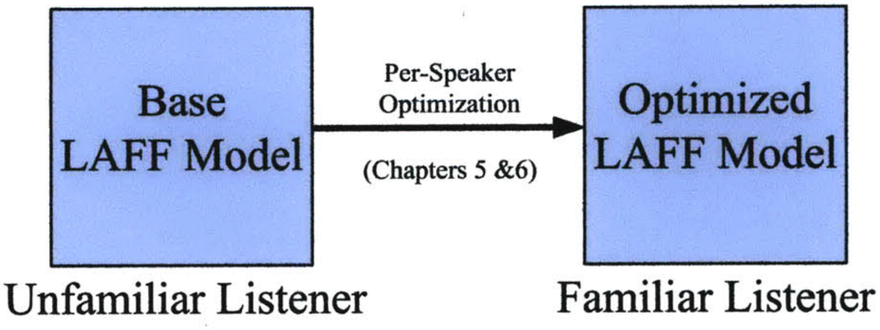

Figure 1.1: Schematic describing the relationship between the base and optimized LAFF models and the analogy of the corresponding level of listener familiarity.

1.1 Distinctive Feature-Based Speech Recognition

The LAFF model is an example of a knowledge-based approach to speech recognition, relying heavily on an understanding of speech production and perception. Rooted in distinctive feature theory, the model assumes that the speech waveform can be represented as a sequence of context-independent binary feature bundles describing the state of the vocal source, vocal tract, and articulators (Jakobson et al., 1952; Chomsky & Halle, 1968). Acoustic cues for distinctive features are present in the speech waveform and it has been suggested that these cues are used for speech segmentation and recognition (Stevens, 2002).

The assumptions of feature-based speech recognition are summarized in Figure 1.2. A speaker's phonetic intent is mapped to a sequence of binary-valued features that are used to drive production and which are manifest in the acoustic waveform. Features are then inferred from the acoustic waveform by a listener (human or machine). Finally, listeners rely on durational, linguistic, and contextual constraints to aid in predicting the speaker's intent. The present work focused on the relationship between speaker intent and measured acoustic features for SWDs.

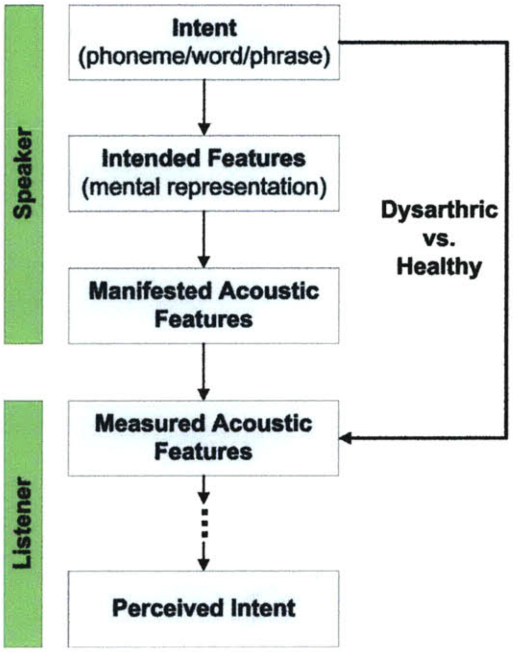

Figure 1.2: Assumptions of phonetic feature-based speech recognition.

1.2 Thesis Overview

It was hypothesized that SWDs would use different sets of acoustic cues to signal segmental contrasts than healthy speakers. To test this hypothesis, the LAFF framework was optimized on a per-speaker basis with the goal of identifying segmental cues that enabled accurate speech discrimination. The LAFF model assumes that the speech stream carries information non-uniformly, with regions of abrupt spectral change being salient points of high information content. These regions, known as landmarks, mark changes in manner of articulation. Also, there is evidence that spectral-temporal information around landmarks and between adjacent landmark pairs provides cues to further elucidate articulator-free and articulator-bound features. Initial analysis of a database of SWDs revealed that landmarks manifested in dysarthric speech did not resemble models of acoustic landmarks expected in healthy speech (DiCicco & Patel, 2008).

The dissertation is comprised of 8 chapters which detail the motivation, background, methods, and results of optimizing the LAFF framework for individual SWDs. In Chapter 1 the motivations and underlying hypothesis of the work are presented.

Chapters 2 and 3 present the relevant background for this work. Chapter 2 reviews distinctive feature theory and Stevens' landmark-based system of lexical speech recognition. For the LAFF model, segmental cues are conveyed by the combination of acoustic landmarks and the correlates extracted from around landmarks. Chapter 3 describes dysarthria and its acoustic and perceptual manifestation in further detail.

Chapter 4 presents the analysis of a dysarthric speech database using the base LAFF model, which was derived from studies of healthy speech production and perception (Park, 2008). Only connected speech productions were used for analysis because cue combinations employed for connected speech may differ from those used for single-word productions given greater motoric demands (Kent, Kent, Rosenbek, Vorperian, & Weismer, 1997). Results revealed that SWDs had difficulty producing the acoustic targets typically associated with changes in manner of articulation (DiCicco & Patel, 2008). Based on these results, shortcomings of the base LAFF model and possible means of improvement for dysarthria are discussed.

Chapters 5 and 6 outline the methods used to determine the acoustic representation of consistently controlled segmental cues, in the form of acoustic landmarks, in dysarthric speech. Chapter 5 deals with extraction of potential landmark candidates and attempts to improve upon existing peak detection and insertion rates. Chapter 6 is centered on identifying speaker-specific acoustic correlates that relate landmark type and validity. In each chapter the corresponding results from both healthy and dysarthric speech are presented.

Chapter 7 describes construction of landmark sequences using higher-level temporal and language information. Landmark sequences produced by the speaker-optimized and base LAFF models are compared to highlight the impact of per-speaker optimization.

Chapter 8 is a summary of the key findings, limitations, and significance of the thesis. Future advancements of the LAFF framework for SWDs as well as possible extensions of the model, including its potential use in clinical settings, are suggested.

The results provide evidence that SWDs produce discriminable segmental cues. These cues, however, are manifested differently than those produced by healthy speakers. While there was overlap in the types of acoustic correlates identified for healthy and dysarthric speakers, the combinations of correlates used to convey landmark type and validity were unique to each SWD. Individuals with severe dysarthria typically required larger sets of correlates for landmark models suggesting that familiar listeners may require highly complex models of lexical recognition to cope with the inherent variability of dysarthric speech.

This work presents a methodology to discover residual segmental cues produced by impaired speakers and specifically illustrates the potential of using the LAFF framework as a tool for understanding dysarthric speech production and perception, with findings having implications for the assessment and treatment of dysarthria. Phonetic features relate articulation and acoustics; therefore, features have the capacity to identify causes for intelligibility deficits resulting from disruptions to a multitude of the speech production subsystems and to track progress during clinical intervention (Ansel & Kent, 1992; J. D. Kent, G. Weismer, J. F. Kent, & J. C. Rosenbek, 1989; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980). Severity of dysarthria is currently assessed using coarse perceptual (i.e. clinician-dependent) measures that qualitatively note articulatory deficiencies but do not quantify their extent.

Landmark analysis provides a quantitative set of fine-grain objective measures that are capable of localizing intelligibility deficits to production of specific manner class(es) and are correlated with severity of impairment.

Chapter 2

Distinctive Feature-Based Models of Speech Recognition

With the goal of identifying sets of discriminable segmental cues reliably produced by speakers with dysarthria (SWDs), Stevens' Lexical Access from Features (LAFF, 2002) paradigm was optimized for ten individuals with dysarthria. The LAFF framework is both a model of human speech segmentation and recognition and also a framework for feature-based automatic speech recognition (ASR). Given the latter, results from landmark analysis of dysarthric speech may have implications for the design and use of customized voice-driven communication aids dependent upon ASR (discussed further in Section 8.1.6).

The LAFF model is rooted in distinctive feature theory, which suggests that the speech waveform can be represented as a sequence of binary feature bundles with context-independent features that describe the state of the vocal folds, vocal tract, and articulators (Jakobson, Fant, & Halle, 1952). Chomsky & Halle (1968) defined a minimal, yet sufficient set of linguistically-distinct binary features to describe all speech sounds (see Table 2.1 for an example representation of the word 'vote' in terms of distinctive features). Distinctive features are the simplest unit of phonological structure; phonemes can be constructed using hierarchically organized bundles of distinctive features with a single feature possibly distinguishing different phonemes.

Table 2.1: Representation of the word 'vote' using bundles of distinctive features. Articulator-free features are in italics and articulator-bound features are located below them. The distinctive features shown come from a set suggested by Stevens (1999, 2002).

	Vote		
Feature	*/v/*	*/o/*	*/t/*
consonantal	+	-	+
sonorant	-		-
continuant	+		-
strident	+		
lips	+		
tongue blade			+
round	-	+	
anterior			+
distributed			-
spread glottis	-		+
constricted glottis	-		-
stiff vocal folds	-		+
slack vocal folds	+		-
high		-	
low		-	
back		+	
advanced tongue root		+	
constricted tongue root		-	

2.1 Distinctive Feature Systems

Various systems of distinctive features that distinguish the phonemes of a language or multiple languages have been suggested (see Baltaxe, 1978 for further discussion). This work relied on the distinctive feature hierarchy originally proposed by Ladefoged and Halle (1988) and adopted by Stevens (2002) in the LAFF framework. In this hierarchy, high-level features known as articulator-free (or manner) features describe the state of the vocal tract irrespective of the position of speech articulators. Moving down the feature hierarchy, articulator-bound (or place) features describe the state of the articulators (lips, tongue blade, tongue body, soft palate, pharynx, glottis, and vocal folds) within the articulator-free class. The relevant set of articulator-bound features is dependent upon the articulator-free class. Top-level articulator-free features were the exclusive focus of this work because automatic feature extraction had not been performed previously for dysarthric speech and accurate extraction of articulator-free features is a prerequisite for detection of articulator-bound features.

2.1.1 Articulator-Free Features

Articulator-free features describe the degree of constriction in the vocal tract, thus indicating manner of articulation. Articulator-free features categorize a segment of speech into one of three broad phonetic classes: vowel, glide, or consonant (Juneja & Espy-Wilson, 2003). Vowels are produced with a relatively open vocal tract. For glides there is a constriction in the vocal tract but the constriction is insufficient to produce an acoustic discontinuity. Consonants are produced via a constriction in the vocal tract that results in an abrupt spectral change.

Vowel and glide segments are each represented via a single articulator-free feature ([+*vowel*] and [+*glide*], respectively) while consonantal segments ([+*consonantal*]) are described further using the features *sonorant*, *continuant*, and *strident* (Liu, 1996; Park, 2008; Stevens, 2002). The feature *sonorant* is relevant for all *consonantal* segments. [+*sonorant*] indicates a segment produced with no pressure build-up in the vocal tract, unattenuated vibration of the vocal folds, and a lack of turbulent noise production (Stevens, 2007)[2]. [-*sonorant*] sounds are produced with a constriction in the oral cavity resulting in a pressure buildup and turbulent noise production and reduced or ceased vocal fold vibration. Nasal consonants and liquids are [+*sonorant*] while obstruent consonants are [–*sonorant*]. Obstruent consonants are divided further into stops and fricatives using the feature *continuant.* [+*continuant*] indicates an incomplete closure in the vocal tract and continuous airflow. [-*continuant*] signals a complete closure and cessation of airflow through the vocal tract. Lastly, the feature [*strident*]

[2] All vowels and glides are [+*sonorant*]. However, because the value of the feature is implied it is not necessary to define the feature *sonorant* for vowel and glide contexts.

distinguishes fricatives based upon the amount of high frequency energy. A summary of the described articulator-free features and a proposed hierarchical organization are shown in Table 2.2 and Figure 2.1, respectively.

Articulator-free features segment speech sounds according to broach phonetic class. By doing so articulator-free features greatly limit the number of potential word candidates. Previous analysis of a 20,000 word dictionary revealed that given a sequence of broad classes, the expected number of word candidates was about 25 (~ 0.1%) with approximately 1/3 of words being specified uniquely (Huttenlocher & Zue, 1984; Shipman & Zue, 1982).

Table 2.2: Summary of the articulatory-free/manner features incorporated into Stevens' Lexical Access From Features model. For each feature, the articulatory correlate and broad phonetic categories, when relevant, are provided (Liu, 1996; Stevens, 2002).

Phonetic feature	Articulatory correlate	Vowels, Glides	Sonorant consonants	Fricatives	Stops
consonantal	Constriction resulting in acoustic discontinuity	-	+	+	+
sonorant	No pressure build up in vocal tract		+	-	-
continuant	Incomplete constriction			+	-
strident	High amplitude energy at high frequencies			+: z, ʃ -: ð, f	

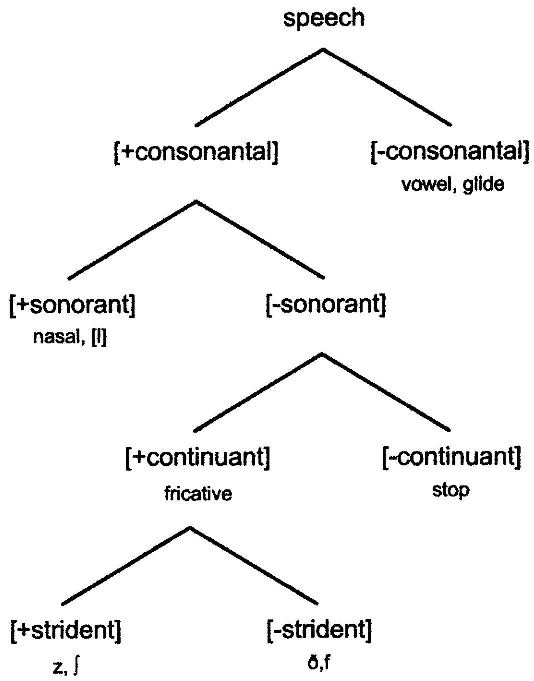

Figure 2.1: Hierarchical organization of articulator-free features in the LAFF paradigm (Liu, 1996).

2.2 Lexical Access from Features

Expanding upon distinctive feature theory, Stevens' LAFF paradigm (2002) assumes that the speech stream carries information non-uniformly with high concentration of information located within and around regions of abrupt spectral change (Figure 2.2). These regions referred to as ***landmarks*** segment the speech stream into three broad phonetic classes: consonant, vowel, and glide. Landmarks are articulatory production targets and serve as perceptual foci for listeners (Liu, 1996). There is evidence that the acoustic representation of articulator-free features is most salient near landmarks, with spectral-temporal information around landmarks and between adjacent landmark pairs providing cues to the identity of articulator-free and articulator-bound features (Jenkins, Strange, & Edman, 1983; Jongman, 1989; Ohde & Stevens, 1983; Stevens, 1985, 2002; Tartter, Kat, Samuel, & Repp, 1983).

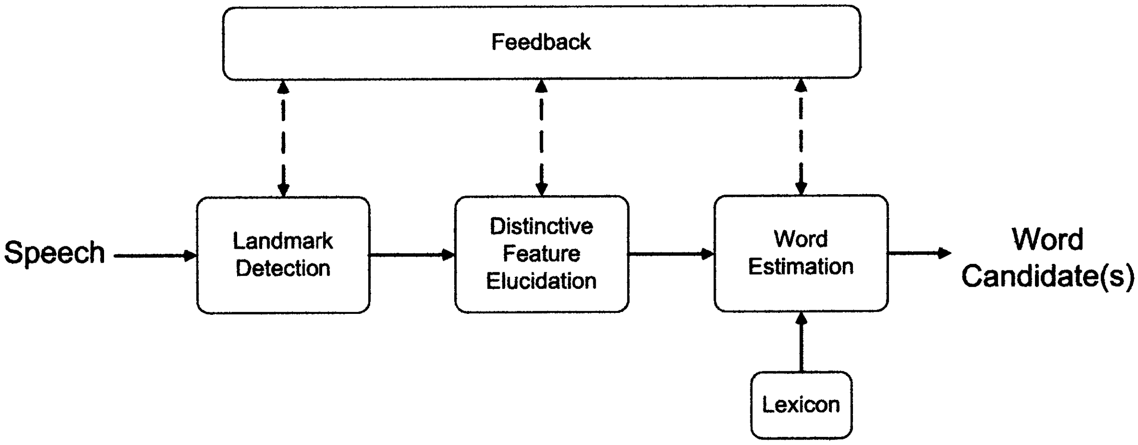

Figure 2.2: Stevens' LAFF model of lexical speech recognition (Liu, 1996; Park, 2008; Stevens, 1992, 2002).

Corresponding to each of the broad classes, there are three categories of acoustic landmarks. Consonant landmarks occur at consonantal closures and releases. Vowels landmarks are localized at syllabic peaks and glide landmarks are associated with syllabic dips.

For the consonantal class there are three landmark types: glottis ('*g*'), sonorant ('*s*'), and burst ('*b*'); and associated with each type is a sign indicating closure (-) or release (+). Table 2.3 summarizes the relationship between various phonetic categories and consonantal landmarks. Glottis landmarks indicate a transition to ('+*g*') or cessation ('–*g*') of free vocal fold vibration. Sonorant landmarks occur during a voiced region in which there is a closure ('-*s*') or release ('+*s*') of a nasal or a liquid. Burst landmarks denote a constriction resulting in an acoustic discontinuity, with a stop or affricate burst release represented by '+*b*' landmarks and cessation of frication or aspiration noise denoted by '–*b*' landmarks.

Table 2.3: Relationship between phonetic category and consonantal landmark type (Liu, 1996). Abrupt-consonantal landmarks are produced by a tight constriction involving one of the primary articulators (lips, tongue blade, tongue body). Abrupt landmarks are acoustic representations of non-primary articulator movement. Outer landmarks are found at consonant/vowel or vowel/consonant borders. Intraconsonantal landmarks occur within a pair of outer AC landmarks. Lastly, intervocalic landmarks appear outside an outer AC pair.

	Phonetic Category	Landmark Type
Outer abrupt-consonantal (AC)	Fricative closure or release	g(lottis)
	Flap closure or release	...
	Stop closure	...
	Unaspirated stop release	...
	Aspirated stop release	b(urst)
	Nasal closure or release	s(onorant)
	Lateral closure or release	...
Intraconsonantal abrupt-consonantal (AC)	Stop closure or release	b(urst)
	Fricative closure or release	...
	Affricate release	...
	Nasal → Fricative	g(lottis)
	Fricative → Nasal	...
Intraconsonantal Abrupt (A)	Velopharyngeal closure or release	g(lottis)
Intervocalic Abrupt (A)	Glottal stop closure or release	g(lottis)
	Voiceless /h/ onset or offset	...

Consonantal landmarks indicate a change in broad class and in most cases a change in articulator-free feature(s) (Park, 2008). Based upon the landmark type, changes in feature bundle values can be inferred (Table 2.4). Glottis landmarks ('*g*') indicate a change in voicing and thus correspond to changes in value of the feature *sonorant*[3]. Burst landmarks ('*b*') do not correspond to a change in articulator-free feature because silence is expected on one side of the landmark.

[3] A glottis landmark carries additional ambiguity. If the landmark is due to the closure/release of an unaspirated stop, then the landmark conveys the features [+*consonantal*], [-*sonorant*], and [-*continuant*]. If a '*g*' landmark is associated with a fricative closure or release it implies the feature [+*consonantal*], [-*sonorant*], and [+*continuant*]. Lastly, if associated with an /h/ or glottal stop then a glottis landmark conveys the feature [-*consonantal*]. While not implemented in this work, formant movements around the landmark could theoretically be used to resolve the identity of glottis landmarks.

They do, however, provide the articulator-free feature values of [+*consonant*, -*sonorant*] on the side of the burst opposite the silence region. Lastly, sonorant landmarks imply a change in the feature *consonant* since they are associated with closures or releases in the oral cavity during continuous voicing. Further disambiguation of feature bundle values can be achieved by looking at pairs of adjacent consonantal landmarks[4].

Table 2.4: Changes in articulator-free features implied by landmark type (Park, 2008).

+*g* – landmark		**-*g* – landmark**	
[+*consonant, -sonorant*] OR Silence	[+*sonorant*]	[+*sonorant*]	[+*consonant, -sonorant*] OR Silence
+*b* – landmark		**-*b* – landmark**	
Silence	[+*consonant, -sonorant*]	[+*consonant, -sonorant*]	Silence
+*s* – landmark		**-*s* – landmark**	
[+*consonant, +sonorant*]	[-*consonant*]	[-*consonant*]	[+*consonant, +sonorant*]

Whereas consonantal landmarks typically indicate change in broad class, vowel and glide landmarks signal syllabic peaks and dips, respectively. The vowel and glide categories are each described by a single landmark. Vowel landmarks ('V') are characterized by local maxima in the first formant (F1) and waveform amplitude. Glide landmarks ('G') are characterized by decreases in F1 and waveform amplitude (Sun, 1996).

This work focused solely on consonantal landmarks because previous studies of dysarthric speech have identified production of consonants as a more likely source of error than vowels and glides due to the complex articulator movements required (Ansel & Kent, 1992;

[4] Example: the landmark pair <-*g*, -*b*> implies that the bounded segment has the features [+*consonant*, -*sonorant*, +*continuant*] (Park, 2008).

Deller et al., 1991; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980). Basic vowel landmark analysis using the vowel landmark detector suggested by Howitt (2000a, 2000b) was previously performed by DiCicco & Patel (2008). In agreement with previous production studies, vowel landmarks were identified in dysarthric speech at rates much similar to those for healthy speakers when compared to detection of consonantal landmarks (DiCicco & Patel, 2008). Glide landmarks have not been examined for SWDs because a robust glide landmark detector, for typical or dysarthric speech, has not yet been developed.

2.3 Dysarthric Speech and the Base LAFF Model

It was hypothesized that the base LAFF model would be inadequate for dysarthric speech since the model incorporates spectral and temporal components derived from studies of healthy speech production and perception. Given the extent and degree of speech motor impairment, SWDs do not reliably produce the same sets of acoustic correlates as healthy speakers for many phonetic features (Ansel & Kent, 1992; Duffy, 2005; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980). For example, reduced range of articulator movement and velocity may result in prolonged segment duration, shallow formant trajectories, weak consonantal release, incomplete closure of the velopharyngeal port, and/or difficulty forming complete closures in the glottis and oral cavity (Duffy, 2005; Logemann & Fisher, 1981; Platt, Andrews, & Howie, 1980). These deficiencies would be expected to alter the manifestation of acoustic landmarks. For instance, an inability to buildup pressure behind a constriction in the oral cavity results in less prominent obstruent bursts. A lack of abruptness would potentially degrade detection of burst landmarks within the base LAFF model. Also, velopharyngeal incompetence introduces nasal zeros into the spectrum, potentially leading to extraction of superfluous sonorant landmarks. Lastly, reduced articulator movement and velocity alter segment duration. In the base LAFF model, temporal parameters related to segmental transitions are used for measuring acoustic correlates around landmarks. Differences in segmental transitions and duration may preclude temporal values determined from healthy speech from being relevant for dysarthric speech.

Despite these differences in production from healthy speakers, SWDs are often able to produce speech that is recognized with high accuracy by familiar communication partners (Beukelman & Yorkston, 1980; D'Innocenzo et al., 2006; Deller et al., 1991; DePaul & Kent, 2000; Hustad & Cahill, 2003; King & Gallegos-Santillan, 1999; Liss et al., 2002; Neilson & O'Dwyer, 1984; Spitzer et al., 2000; K. K. Tjaden & Liss, 1995). This observation suggests that SWDs may be relying on alternative combinations of acoustic cues to signal segmental information and that as long as these cues are consistent and reliable, listeners can learn to decode the speaker's intent. This work sought to apply the LAFF framework to identify the spectral and temporal manifestation of discriminable segmental cues in dysarthria.

Chapter 3

Dysarthria

Dysarthria is a motor speech disorder characterized by weak or imprecise movements of the speech musculature (Duffy, 2005; Yorkston, Beukelman, Strand, & Bell, 1999). Motor speech disorders account for a significant portion of acquired communication disorders (37%), with dysarthria being the diagnosis in almost half (46%) of all cases (Duffy, 2005). Individuals with dysarthria have the ability to comprehend and plan speech but suffer in their ability to produce the coordinated movements of the vocal tract and articulators. Dysarthria is secondary to damage or abnormal development of the brain or the nerves that control the speech muscles. Level of impairment, typically measured in terms of intelligibility, can vary from mild to severe. Common causes for dysarthria include stroke and various other neurological traumas, Parkinson's disease (PD), multiple sclerosis (MS), and cerebral palsy (CP). Acoustic-perceptual hallmarks of dysarthria include imprecise consonants, vowel centralization, slow rate, monopitch, monoloudness, and hypernasality (Boutsen, Bakker, & Duffy, 1997; Darley, Aronson, & Brown, 1969; Gentil, 1990; Kent et al., 1997; Portnoy & Aronson, 1982; K. Tjaden, Rivera, Wilding, & Turner, 2005; K. Tjaden & Wilding, 2004; Ziegler & von Cramon, 1983a, 1983b, 1986; Ziegler & Wessel, 1996). Given the heterogeneity of impairment, dysarthria serves as a challenging disorder from which to identify consistently controlled segmental cues.

3.1 Distinctive Features and Dysarthric Speech

Dysarthric speech has been studied acoustically and perceptually. Several studies have related listener confusion of distinctive features to reduced intelligibility in dysarthric speech, with frequency of error across features strongly correlated with single-word intelligibility scores. In single real-word consonant-vowel-consonant (CVC) productions from 32 spastic and 18 athetoid males, Platt, Andrews, et al. (Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980) noted imprecise production of fricative and affricate consonants, reduction of the vowel quadrangle, and difficulty with anterior tongue articulation. Errors of voicing and place of articulation were six times more common than errors in manner of articulation. There were fewer vowel errors than consonant errors. Vowels lying along the extremes of the vowel triangle posed the most difficulty. The speech deficiencies exhibited by spastic and athetoid speakers were similar regardless of their intelligibility scores and frequency of errors increased with reduced intelligibility.

Similarly, Ansel and Kent (1992) analyzed the CVC single-word productions from 16 individuals with mixed cerebral palsy. They focused on seven phonetic (voicing, manner, or place) contrasts: syllable-initial voicing, syllable-final voicing, stop-nasal, fricative-affricate, front-back vowel, high-low vowel, and tense-lax vowel. Associated with each of these phonetic contrasts were acoustic correlates that were measured and analyzed to look for differences between healthy and dysarthric contrast pairs. The correlates were derived from studies of healthy speech production and perception. Acoustic analysis of the correlates revealed that speakers with dysarthria (SWDs) were able to produce all but one contrast, the tense-lax contrast. However, when recordings were used for a listener transcription task, average discrimination accuracy across all contrasts was only 56%. This mismatch implies that SWDs were able to acoustically differentiate most contrast pairs but did so in a different manner than healthy speakers.

Chapter 4

Landmark Extraction using the Base LAFF Model

In the LAFF model, segmental cues are inferred by extracting candidate landmarks (Section 4.1.1) and then using the acoustic correlates measured from around candidate peaks to determine landmark type and validity (Section 4.1.2). This chapter presents the methods of the base LAFF model (Park, 2008) and the results from analysis on a database of speakers with dysarthria (SWDs). These baseline results guided the development of the methods used to perform model optimization for individuals with dysarthria (Chapters 5 and 6).

4.1 Acoustic Feature Extraction

4.1.1 Extraction of Candidate Landmark Peaks

Extraction of acoustic landmarks was performed by first taking a broadband (6 ms Hanning window) spectrogram every 1 ms from a speech waveform (Figure 4.1; Liu, 1996; Park, 2008; Stevens, 2002). The short window provided good temporal resolution and the high frame rate allowed for tracking of rapid acoustic changes. Next, the spectrogram was divided into 6 coarse frequency bands (Band 1: 0 – 0.4 kHz, Band 2: 0.8 – 1.5 kHz, Band 3: 1.2 – 2 kHz, Band 4: 2 – 3.5 kHz, Band 5: 3.5 – 5 kHz, and Band 6: 5 – 8 kHz)[5] and a two-pass approach was used to estimate peaks in the band energy waveforms (see Table 4.1 for a description of the frequency bands). The first pass discovered regions of large spectral change, while the second-pass

[5] Bands were inspired by the perceptual studies of Shannon et al. (1995) and supplemented by Liu (1996).

precisely localized peaks in the energy waveforms. Peaks remaining after the two-passes were deemed candidate landmarks. On the first pass, the energy waveform in each band was computed using a 16 ms smoothing window and the energy in each band was quantified using the average squared spectral magnitude across the band. For each energy waveform, an *n*-point rate-of-rise (ROR) was calculated. Similar to a first difference, rate-of-rise estimated temporal rate of change. However, instead of relying on samples directly adjacent to the target point, an *n*-point ROR uses the values located *n*/2 samples before and after the target sample. For the first pass, a 40 point (or 40 ms since the frame rate is 1 ms) ROR was calculated for each frequency band. Peaks in the ROR waveforms were extracted using Mermelstein's recursive peak-picking algorithm (Mermelstein, 1975). Candidate '+*g*' landmarks were associated with abrupt, 5 dB or more, increases in band 1 energy while potential '–*g*' landmarks were placed at abrupt decreases in Band 1 energy (i.e. frequencies below 400 Hz). Potential burst and sonorant landmarks were associated with increases or decreases, on the order of ±7 dB or more, in Bands 2 – 6.

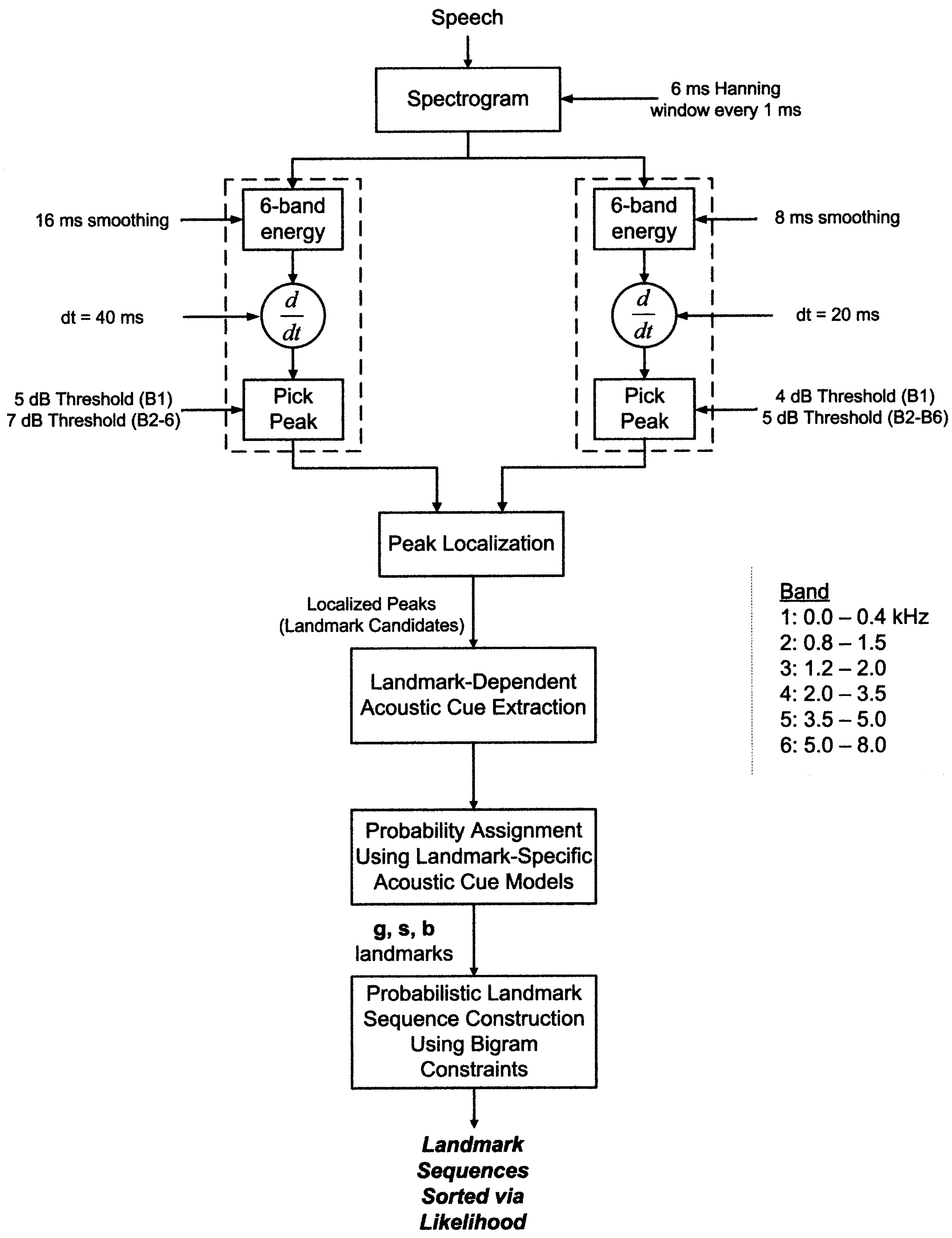

Figure 4.1: LAFF consonantal landmark detection scheme (Liu, 1996; Park, 2008).

Table 4.1: Description of the six energy bands used for landmark detection in the base LAFF model (Fell, MacAuslan, Chenausky, & Ferrier, 1999). Band ranges were inspired by Shannon et al. (1995) and Liu (1996).

Band	*Bound*	*Purpose*	
1	0 – 400 Hz	To capture F_0	
2	800 – 1500 Hz	For intervocalic consonantal segments a zero is introduced in this range.	At a sonorant consonant closure, spectral prominences above F_1 show an abrupt decrease in energy. Onsets & offsets of aspiration and frication noise will lie in at least one of these four bands.
3	1200 – 2000 Hz (F2 ~ 1500 Hz)		
4	2000 – 3500 Hz (F3 ~ 2500 Hz)		
5	3500 – 5000 Hz		
6	5000 – 8000 Hz	Used for silence detection for stops	

Candidate landmarks from the first pass of processing were reevaluated on the second pass. The general procedure was the same but parameter values were decreased to precisely localize landmarks in time. An 8 ms smoothing window was used to calculate energy waveforms, a 20 point rate-of-rise was used to estimate rates of change, and thresholds of ±4 dB for Band 1 peaks and ±5 dB for Bands 2 – 6 were used.

After the two-passes, coarse- and fine-pass peaks were combined to localize landmark candidates. Coarse- and fine-pass peaks were paired if they were in the same frequency band, were sufficiently close (within 15 ms) to one another, and were of the same sign (plus or minus). At this point in the processing, localized peaks from Band 1 were designated as glottis landmarks. Burst and sonorant landmark candidates required an additional step, temporal clustering of peaks from Bands 2 – 6. Grouping of related peaks was an example of the max-cut problem, which calls for partitioning a weighted graph into subsets that maximize the weights on the arcs connecting the groups (Goemans & Williamson, 1995; Marti, Duarte, & Laguna, 2009). In this example, the nodes of the graph were the acoustic peaks and the time differences between peaks were the arc weights. Clustering was performed with the additional constraints that peaks were sufficiently close to one another (within 50 ms), clusters did not overlap in time, and clusters did not contain multiple peaks from the same frequency band. Clusters that contained at least 3 peaks were deemed landmark candidates.

Peak extraction thresholds suggested by Park (2008) were developed for the TIMIT Database. The Nemours Database of Dysarthric Speech (Section 4.2.1; Menéndez-Pidal, Polikoff, Peters, Leonzio, & Bunnell, 1996) required reduced thresholds for peak extraction than those suggested for TIMIT. The TIMIT Database was recorded using a head-mounted close-talking microphone while the Nemours Database was made using a free-field microphone located on a desk. Thus, it would be expected that the Nemours recordings would have a lower speech-to-noise ratio, requiring lower thresholds. For the Nemours control speaker, thresholds were set by performing a grid-search of coarse- and fine-pass values over the training set. Thresholds were started high[6], resulting in low peak detection and insertion rates, and then systematically decreased. Decreasing the threshold led to sharp increases in detection rate. This pattern continued up to a point at which changes in the detection rate began to level-off. The threshold just prior to the breakpoint was set as the peak threshold (glottis thresholds: 4.5 and 3 dB for coarse and fine peaks, respectively; sonorant/burst thresholds: 6 and 4 dB). The results shown for dysarthric speakers in this chapter were achieved using the thresholds from the Nemours control speaker. A further discussion of threshold adjustment is discussed in Section 5.1.1.

4.1.2 Landmark Probabilities

Following peak localization, landmark candidates were assigned a probability using multivariate Gaussian mixture models for each landmark type. Inputs to each model included correlates specific to the landmark type (Table 4.2). Three correlates, all pertaining to low

[6] 10 dB, which was larger than Liu's coarse threshold (1996)

frequency (sub-400 Hz) energy, were used to describe glottis landmark candidates. Abruptness was the height of the ROR peak from the fine pass of processing. The left and right sonorant levels were estimates of Band 1 energy from before and after the candidate peak, respectively. The left sonorant level was defined as the lowest energy level for +*g* landmarks or the highest energy level for –*g* landmarks in the 0-400 Hz band that spanned at least 20 ms between the current and preceding glottis landmarks. The right sonorant level was the highest energy level for +*g* landmarks or the lowest energy level for –*g* landmarks between the current and proceeding glottis landmarks. This spanning criterion attempted to minimize the influence of extremely short perturbations in energy (Park, 2008).

Table 4.2: Spectral cues used to describe each landmark type (Park, 2008).

Landmark Type	Cues
(g)lottis	Abruptness at peak, Sonorant level on left, Sonorant level on right
(b)urst	Abruptness at peak, Silence on left/right, Frication on right/left
(s)onorant	Abruptness at peak, Energy on left & right, Change in spectral tilt

Potential burst and sonorant landmark candidates came from the set of peaks in Bands 2 – 6. For both landmark types the abruptness measure was derived from the rate-of-rise for the frequency band spanning 1200 – 8000 Hz. For burst landmarks, measures of the adjacent silence and frication spanning at least 10 ms over a 50 ms window were calculated using this same broadband frequency range. Like their burst counterparts, models for sonorant landmarks also relied on abruptness and broadband energy on each side of the candidate peak. An additional correlate, temporal change in spectral tilt, was also used to describe sonorant landmarks. Spectral tilt was the ratio of low-to-high frequency (0-360:0-5000 Hz) energy and the difference of spectral tilt across the landmark candidate served as the correlate value.

For each landmark instance[7], two probability models were created to assign the likelihoods that a set of correlates was associated with an expected (i.e. detected) landmark (P(cues|l_i)) and with a superfluous (i.e. inserted) landmark (P(cues|l_i')). Models were created using correctly detected and falsely inserted landmarks from training data. The composition of the training and test sets for both control speakers and individuals with dysarthria is described in Section 4.2. Gaussian mixture models were created for each landmark instance using expectation maximization (EM; Duda, Hart, & Stork, 2001; Hastie, Tibshirani, & Friedman, 2001). The prevalence of detected and inserted peaks from the training data served as prior probabilities (P(l_i) and P(l_i')). From Bayes Theorem, the posterior probability that a given set of cues was the associated landmark instance was given by:

$$P(l_i \mid cues) = \frac{P(cues \mid l_i) * P(l_i)}{P(cues \mid l_i) * P(l_i) + P(cues \mid l_i') * P(l_i')}$$

$$\equiv P(True) = 1 - P(False)$$

where $P(cues \mid l_i)$ and $P(cues \mid l_i')$ were the conditional probabilities estimated via EM.

4.1.3 Landmark Sequence Construction

After landmark extraction and probability assignment, a language model in the form of bigram constraints can be used along with the landmark probabilities to construct possible landmark sequences sorted by likelihood. In Park's probabilistic implementation of the LAFF model, a graph pruning scheme that eliminated low probability sequence pathways was suggested. Landmark sequence construction which utilizes higher-level information was not a large component of this work and will not be discussed further until Chapter 7. Instead, this work focused on understanding the manifestation of acoustic landmarks in dysarthric speech,

[7] The term 'landmark instance' refers to a landmark sign and type pair, e.g. +*g* or –*s*.

which required the extraction of expected landmark candidates and accurate assignment of landmark probabilities. These two steps were directly applicable for testing the hypothesis that SWDs convey landmarks differently than healthy speakers. The base LAFF model proved unsuccessful for these two tasks when applied to dysarthric speech.

4.2 Databases of Dysarthric and Controlled Speech

A major challenge in dysarthric speech research is collecting a suitable number of quality recordings on which to perform analysis. SWDs fatigue easily, there is large variability in production, and recording often has to occur outside of a sound booth or without a head-mounted microphone.

Only connected speech productions were used for this work because it was hypothesized that cue combinations employed for connected speech may differ from those used for single-word productions given greater demands for controlling airflow, vocal fold tension, vocal tract configuration, and articulator placement (Kent et al., 1997). To date, phonetic feature analysis of dysarthric speech had primarily been limited to single-word productions. The use of automated approaches and databases containing sentence-productions, however, enabled examination of dysarthric speech at the connected level.

Time-aligned phonetic transcriptions were necessary to compare landmarks extracted by the model to reference landmarks expected by the annotations. Also, while it is hypothesized that phonetic features and landmarks are context-independent and robust across speakers and languages (Chomsky & Halle, 1968; Stevens, 2002), the LAFF framework contains a probabilistic component that would be expected to benefit from larger amounts of training data. Therefore, larger sets of recordings from SWDs were desirable. Lastly, phonetically-rich sentences were preferred to ensure robustness. Using these criteria, the only suitable and publicly available database was the Nemours Database of Dysarthric Speech[8].

4.2.1 Nemours Database of Dysarthric Speech

The Nemours Database of Dysarthric Speech contains recordings from 10 young males (in their 20's and 30's; exact ages not documented) with dysarthria of varying severities secondary to either cerebral palsy ($N_{CP} = 7$) or head trauma ($N_{HT} = 3$), and a single control speaker (Menéndez-Pidal et al., 1996). Prior to data collection, individuals with dysarthria were examined by a speech-language pathologist using the Frenchay Dysarthria Assessment (FDA; Enderby, 1983). While diagnostic classification of motor control was noted, the dysarthria subtype was not documented in the original database[9]. For each speaker, as part of the FDA the clinician systematically assigned a perceived intelligibility rating on a scale of 0-8 (where '8' corresponded to highest intelligibility).

Table 4.3: Intelligibilities and primary cause of dysarthria for the ten individuals from the Nemours Database of Dysarthric Speech. Mean speaker sentence intelligibility was reported to be 3.4 (SD = 2.7).

Subject	Intelligibility	Etiology
DYS1	0	Cerebral Palsy
DYS2	1	Cerebral Palsy
DYS3	1	Cerebral Palsy
DYS4	2	Cerebral Palsy
DYS5	3	Head Trauma
DYS6	3	Cerebral Palsy
DYS7	4	Head Trauma
DYS8	4	Head Trauma
DYS9	8	Cerebral Palsy
DYS10	8	Cerebral Palsy

[8] The Waisman Center at the University of Wisconsin is in the process of constructing a large database of dysarthric speech. However, the project's IRB precluded recordings from being shared outside of the center (communication with R. D. Kent).

[9] For this reason the effect of dysarthria subtype was not investigated in this work. There is precedence for ignoring dysarthria subtype. The collective findings of Platt and colleagues on speakers with spastic and athetoid cerebral palsy (Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980) and Ansel and Kent (1992) in speakers with mixed cerebral palsy suggest a general homogeneity in the characteristics of speech produced by speakers with cerebral palsy, regardless of subtype, despite the fact that cerebral palsy is often heterogeneous in its presentation.

Each speaker produced 74 nonsense sentences of the form "The *noun*$_i$ is *verb*-ing the *noun*$_j$" (Figure 4.2). Recordings were made using a free-field desktop microphone and sampled at 16 kHz. The lexicon consisted of 71 monosyllabic nouns and 37 disyllabic verbs (counting the 'ing'). Nouns were used in pairs (e.g. both "The *noun*$_i$ is *verb*-ing the *noun*$_j$" and "The *noun*$_j$ is *verb*-ing the *noun*$_i$" were produced). The lexicon can be found in Appendix A. The combinations of nouns and verbs were unique to each individual with dysarthria. A single control speaker produced the corresponding set of utterances for each SWD. The database included manually-verified time-aligned phonetic labels of the recordings produced by each SWD. Phonetic labeling of the utterances produced by the control speaker was performed using a modified version of the hybrid automatic phonetic labeling scheme suggested by Kominek, Bennett, and Black (2003, Appendix B). Training and test sets were created for each speaker by dividing the recordings into two equal-sized collections. Sentences were split along noun-pairs so that only one instance of each noun was found in each set.

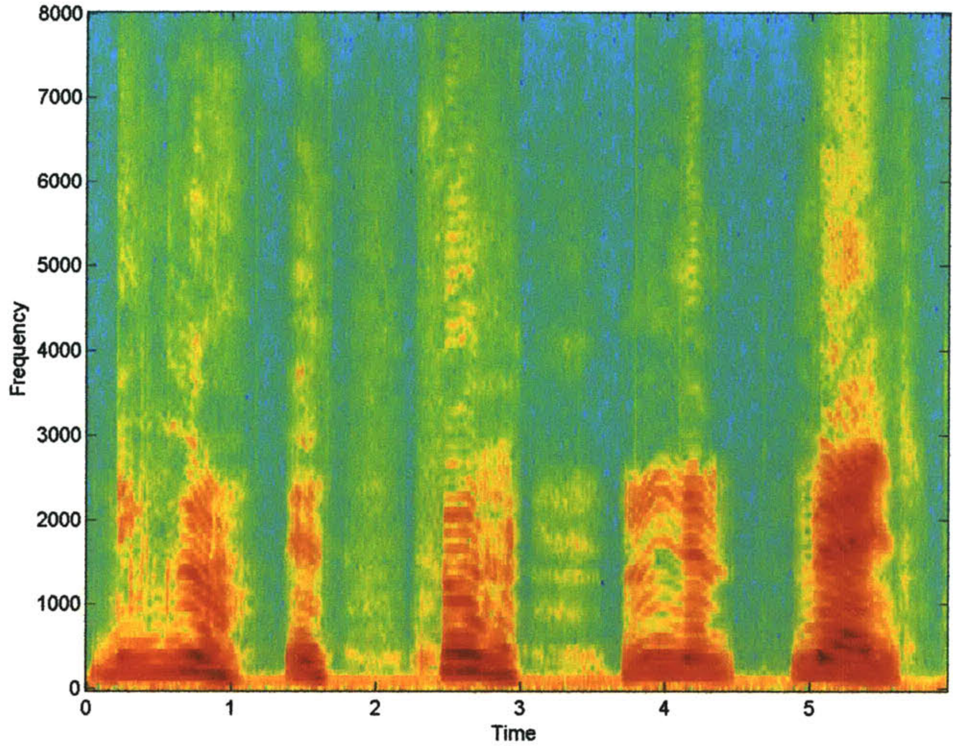

Figure 4.2: Example utterance from the Nemours Database of Dysarthric Speech. The sentence is "The goo is surfing the batch" and was produced by a speaker with severe dysarthria.

4.2.2 TIMIT Database

The TIMIT Database was used as an additional reference for validation and baseline comparison (Fisher, Doddington, & Goudie-Marshall, 1986; Seneff & Zue, 1988; Zue, Seneff, & Glass, 1990). The TIMIT database contains recordings from 630 English speakers, 438 males and 192 females with 10 utterances produced by each subject. For consistency with the Nemours Database, female speakers from the TIMIT Database were excluded. Of the 10 sentences produced by each speaker, two were recorded by all subjects with the intent of exploring dialectal variation, five came from a list of 450 phonetically-compact utterances with more frequent occurrence of phonetic contexts thought to be difficult, and the remaining three were selected from a list of 1,890 phonetically-diverse sentences. Consistent with Park (2008), the two sentences related to dialect were not included in the present analysis so as not to bias landmark models towards a limited set of phonetic content.

Recordings in the TIMIT Database were made in a quiet room using a head-mounted noise-cancelling microphone and sampled at 16 kHz. Time-aligned phonetic transcription was manually performed for all recordings. The TIMIT database was divided into the recommended training and test sets with the training set containing 321 speakers (73%) and the test set 117 speakers, with no overlap across sets.

4.2.3 Hypothetical Landmark Sequences

A phoneme-to-landmark mapping algorithm was used to automatically define expected sequences of landmarks (Park, 2008). For each utterance, the mapping algorithm utilized the time-aligned broad class transcriptions to hypothesize the corresponding sequence and timing of acoustic landmarks (Table 4.4). The hypothetical landmark sequences allowed for calculation of detection, deletion, substitution, and insertion rates. If an observed landmark was the same sign and type and located within 30 ms of a hypothesized landmark, then the hypothesized landmark was classified a detection. Previous work has shown that a 30 ms analysis window was sufficiently long to allow for a majority of automatically extracted landmarks to be associated with corresponding hand-labeled landmarks (Liu, 1995). At the same time, this window was short enough that observed landmarks were not frequently paired with hand-labeled landmarks from neighboring acoustic phonetic events. A hypothesized landmark for which no landmark of the same sign and type was extracted within the acceptable analysis window was considered a deletion. If an observed landmark was the same sign but different type and within 30 ms of a hypothesized landmark, then the hypothesized landmark was judged to be a substitution. Lastly, if a landmark was extracted from the waveform but did not correspond to a hypothesized landmark then it was classified as an insertion. Landmark rates (detection, deletion, substitution, and insertion) were calculated by dividing the counts of each event by the number of landmarks hypothesized according to the phonetic sequences.

Table 4.4: Mapping of adjacent broad class pairs to consonantal landmark instances (Park, 2008).

	Proceeding Broad Class						
Preceding Broad Class	Vocalic	Sonorant	Flap	Glottal Stop	Voiced /h/	Fricative	Silence
Vocalic		-s	-s			-g	-g
Sonorant	+s		-s	+s		-g	-g
Flap	+s	+s		+s		-g	-g
Glottal Stop		-s	-s			-g	-g
Voiced /h/						-g	-g
Fricative	+g	+g	+g	+g	+g		-b
Silence	+g	+g	+g	+g	+g	+b	

4.3 Analysis of Healthy and Dysarthric Speech Using the Base LAFF Model

Landmark analysis of the SWDs from the Nemours Database and the control speakers from the Nemours and TIMIT Databases was performed using an implementation of the base LAFF model (Park, 2008). The results are separated according to the two main components of the model:

1) Extracting candidate landmarks (Section 4.3.1)

2) Using the acoustic correlates measured from around candidate peaks to determine landmark type and validity (Section 4.3.2).

4.3.1 Peak Extraction

Candidate peak detection and insertion rates for the healthy speakers and SWDs are shown in Figure 4.3. For all landmark instances, expected peaks were extracted at higher rates for controls than SWDs. Also, the dysarthric speaker group inserted more unexpected landmarks than the control speakers. These inserted peaks have lexical significance and thus may mislead or confuse listeners.

Given the large variability in impairment severity exhibited by the Nemours' SWDs, it was important to look at the results for each speaker. Table 4.5 displays the peak detection and insertion rates for each individual with dysarthria and the Nemours healthy control. For all SWDs, landmark-specific and average detection rates were always less than for the control; landmark-specific and average insertion rates were always higher. Average detection rates ranged from 56% to 86% for the SWDs, compared to an average detection rate of 91% for the healthy control. Average insertion rates showed even greater variability among the SWDs ranging from 282% to 1107%, compared to 174% for the healthy speaker. Figure 4.4 is a plot of average peak detection and insertion rates as functions of speaker intelligibility. Average detection rate was positively correlated with intelligibility while average insertion rate was negatively correlated with intelligibility (DiCicco & Patel, 2008). Both correlations were significant ($\alpha < 0.05$). Recalling that intelligibility is inversely related to severity of impairment, mildly impaired individuals produced expected landmarks at a higher rate while producing fewer insertions. Severely impaired speakers produced expected landmarks at reduced rates but more frequently inserted superfluous landmarks.

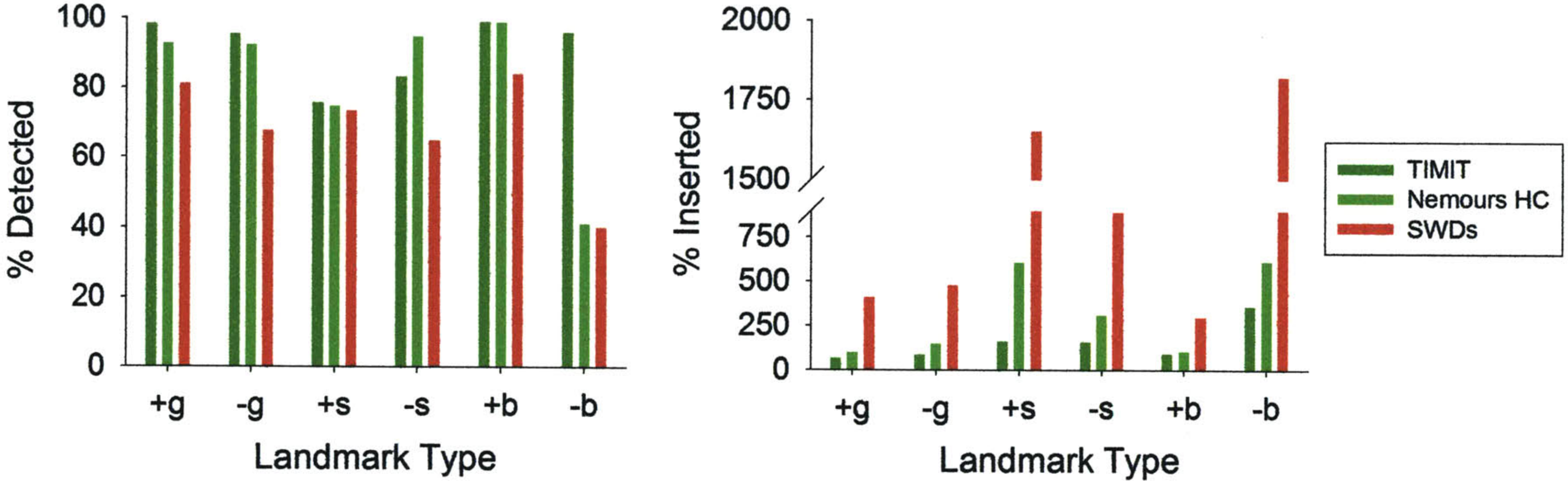

Figure 4.3: Landmark-specific candidate peak detection (left) and insertion (right) rates for the TIMIT Database (dark green), the healthy control speaker from the Nemours Database (light green), and the ten speakers with dysarthria (SWDs) from the Nemours Database. In the Nemours Database burst closures (*-b*) and sonorant releases (*+s*) were more sparsely represented than the other landmark instances.

Table 4.5: Per-speaker candidate peak detection and insertion rates for the individuals with dysarthria (DYS) and the healthy control speaker (HC) from the Nemours Database. In the Nemours Database burst closures (-*b*) and sonorant releases (+*s*) were more sparsely represented than the other landmark instances.

	Intelligibility	g-Detection	g-Insertion	s-Detection	s-Insertion	b-Detection	b-Insertion	Average Detection	Average Insertion
DYS1	0	55%	890%	60%	2208%	58%	1121%	56%	1107%
DYS2	1	74%	778%	73%	1715%	85%	870%	76%	915%
DYS3	1	79%	285%	69%	865%	75%	522%	77%	418%
DYS4	2	70%	219%	57%	765%	77%	406%	70%	332%
DYS5	3	61%	840%	50%	905%	62%	459%	60%	756%
DYS6	3	66%	283%	69%	1072%	71%	544%	67%	443%
DYS7	4	84%	315%	74%	1104%	88%	590%	84%	480%
DYS8	4	91%	309%	73%	945%	81%	480%	86%	428%
DYS9	8	83%	203%	84%	709%	84%	360%	83%	303%
DYS10	8	84%	189%	73%	660%	84%	334%	83%	282%
HC		92%	123%	89%	389%	89%	197%	91%	174%

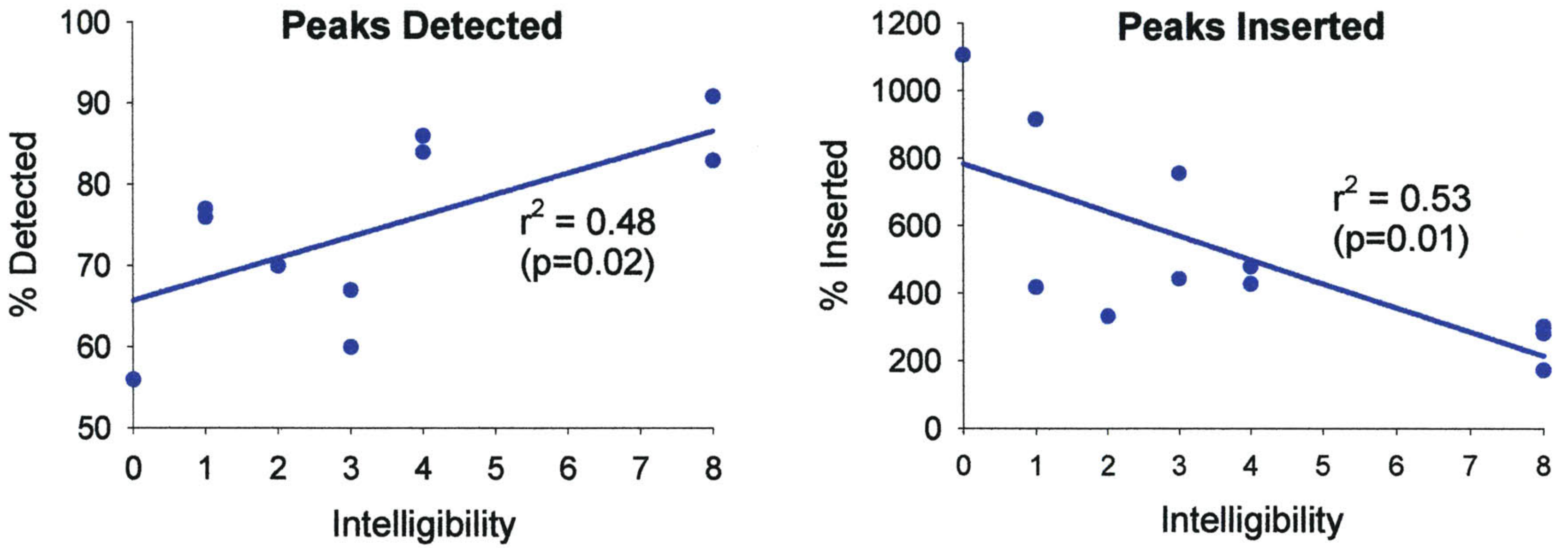

Figure 4.4: Candidate peak detection and insertion rates as functions of sentence-level intelligibility. Linear correlations and correlation coefficients are also shown. The intelligibility for the control speaker was assumed to be maximal (8).

4.3.2 Landmark Probabilities

In the LAFF model, following extraction of candidate landmarks, peaks are assigned a probability using landmark-specific acoustic correlates. Glottis and burst landmarks were described by three correlates. Sonorant landmarks were described by four correlates. Statistical models, in the form of Gaussian mixture models for each landmark instance, were built using correlates extracted from around candidate peaks from training data. Candidate peaks extracted from the withheld test sets were assigned probabilities by these models. For analysis of the TIMIT test set, probability models were created from the TIMIT training set. For the Nemours Database, the healthy subject's training set was used to create landmark models for himself as well as each SWD. Probability models built from the TIMIT training set were also used for the speakers from the Nemours Database but resulted in reduced discrimination rates compared to when using those from the Nemours healthy control.

Figures 4.5 through 4.7 are histograms of the assigned probabilities for detected and inserted landmarks from the TIMIT Database, the Nemours healthy control, and the ten individuals with dysarthria, respectively. An ideal probability model would assign probability of 1 to expected landmarks while assigning probability of 0 to superfluous inserted peaks. Visual inspection of the histograms suggests that the models were sufficient for glottis landmarks (both + and -). For all speaker groups there were large peaks in probability near 1 for detections and 0 for insertions. Across all speakers the models were less than ideal for sonorant and burst landmarks. For SWDs many detected sonorant releases (+*s*) and closures (-*s*) and detected burst closures (-*b*) were assigned low probabilities.

To provide a more quantitative-based comparison, the type I and type II errors for each landmark instance were calculated (Figure 4.8). Type I error was the percentage of correctly extracted peaks (i.e. detected peaks) assigned a probability less than 0.5[10]. Type II error was the percentage of falsely inserted landmark candidates assigned a probability greater than 0.5[11]. The two error metrics were useful indicators of performance because in landmark sequence construction the probability that an extracted landmark candidate is not a landmark is 'P(*false*) = 1-P(*true*)'. The type I error rates for all landmark instances were consistently higher for SWDs, suggesting that the acoustic correlates manifested around expected candidate peaks were different than those expected by the model. For type II errors, there was not a consistent discrepancy across healthy and dysarthric speaker groups.

[10] Type I error = 1 - sensitivity

[11] Type II error = 1 - specificity

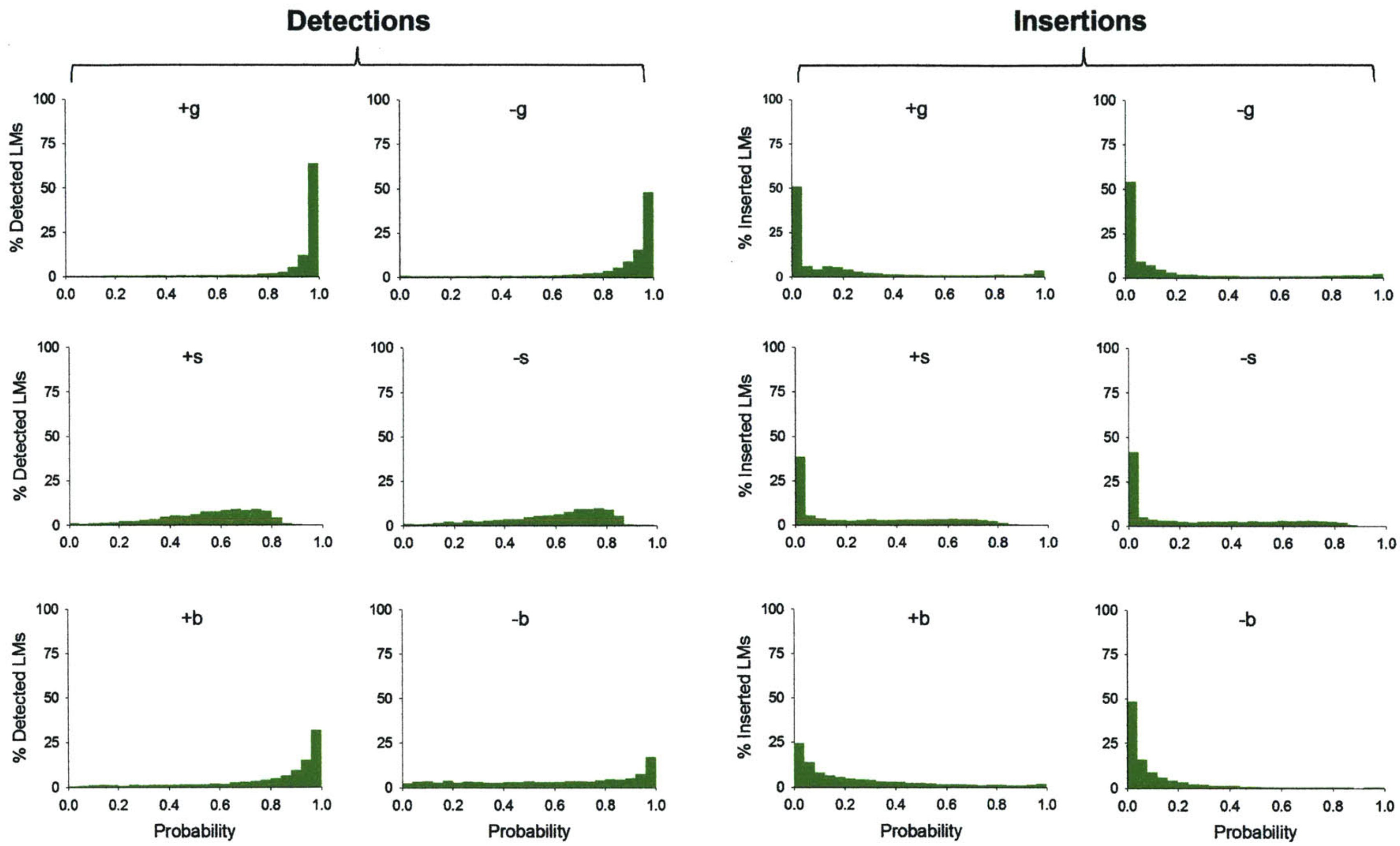

Figure 4.5: Histograms of the probabilities assigned to detected and inserted candidate landmarks peaks extracted from the TIMIT test set recordings.

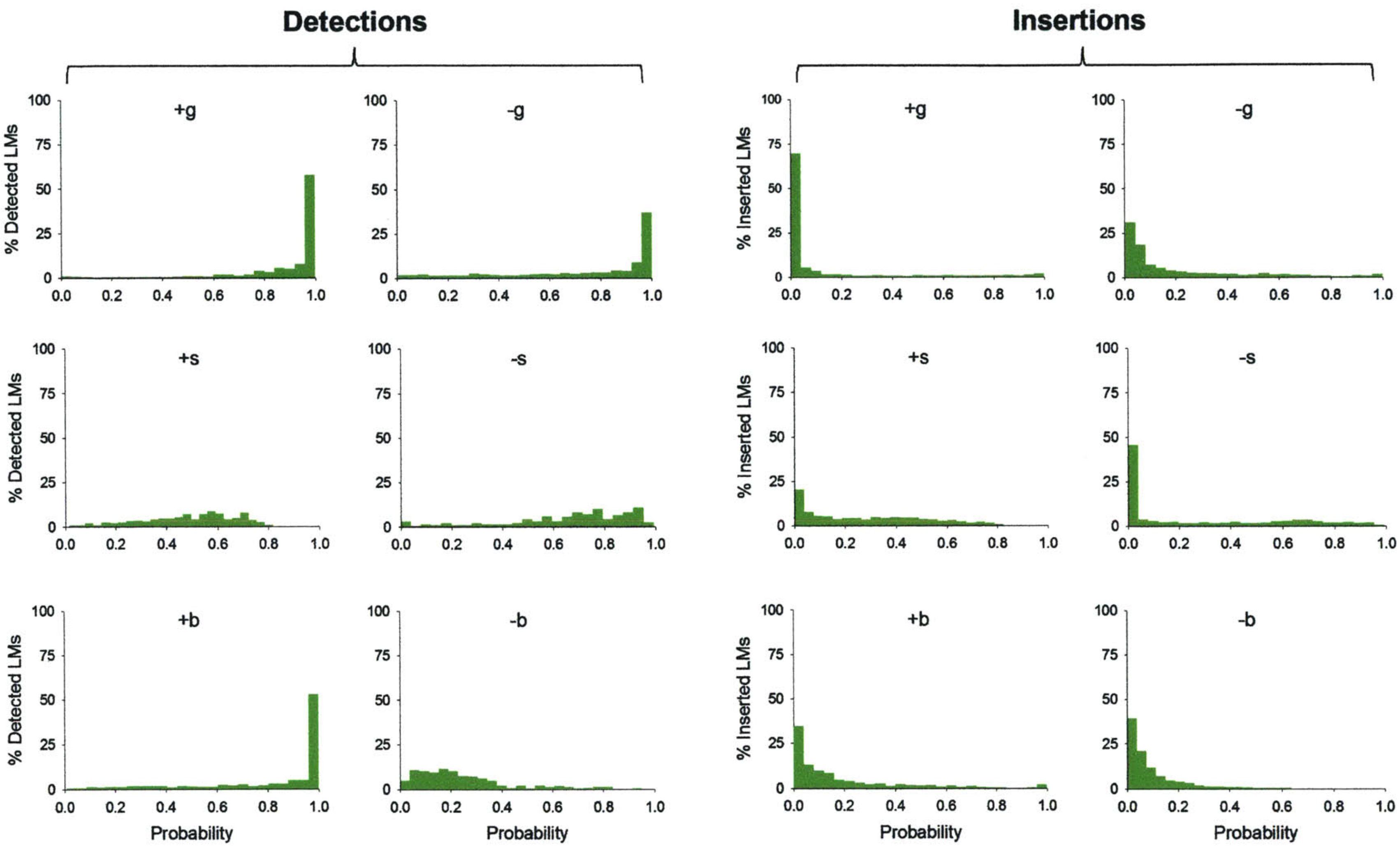

Figure 4.6: Histograms of the probabilities assigned to detected and inserted candidate landmarks peaks extracted from the test set recordings for the healthy control speaker from the Nemours Database.

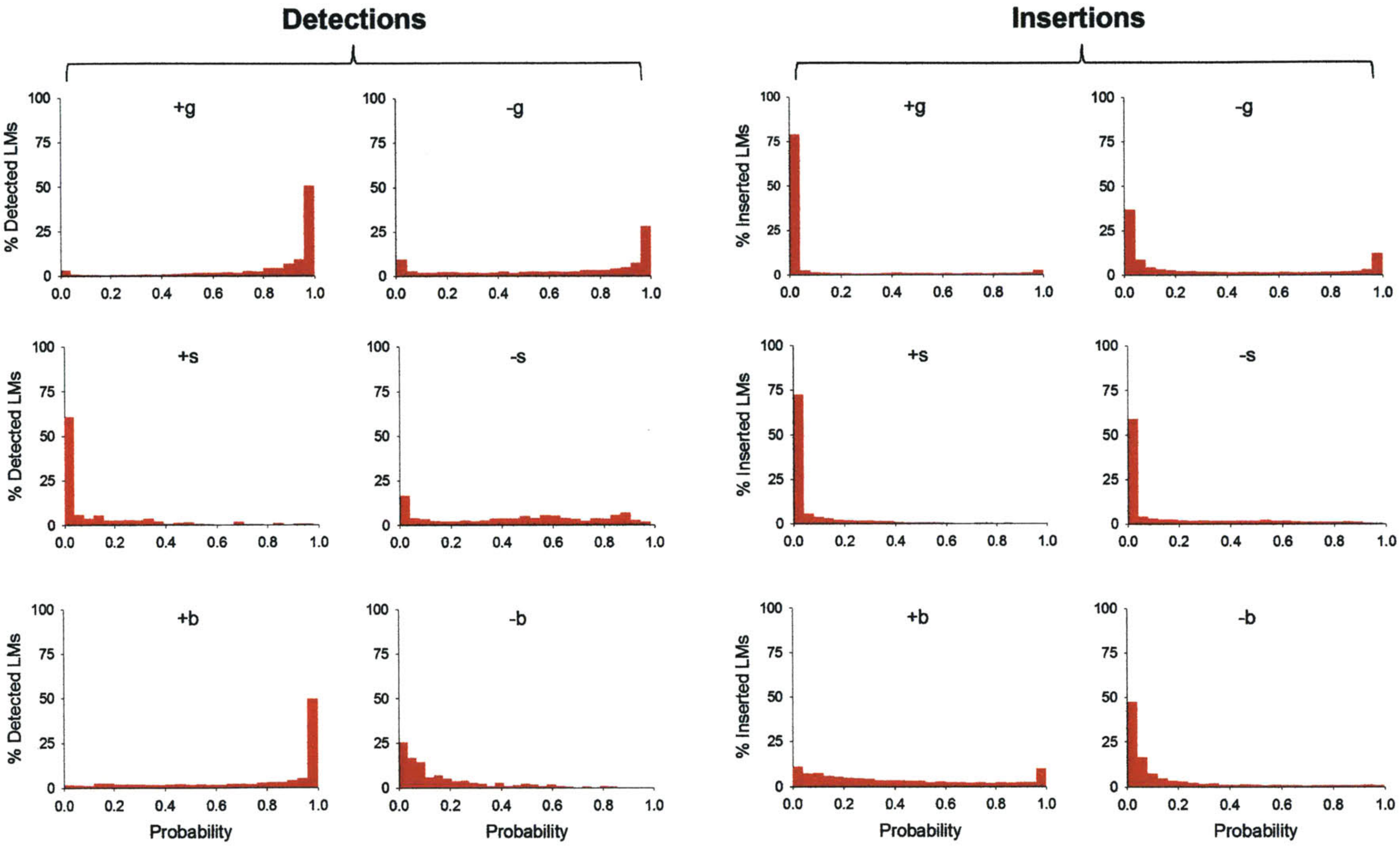

Figure 4.7: Histograms of the probabilities assigned to detected and inserted candidate landmarks peaks extracted from the test set recordings for the ten individuals with dysarthria from the Nemours Database.

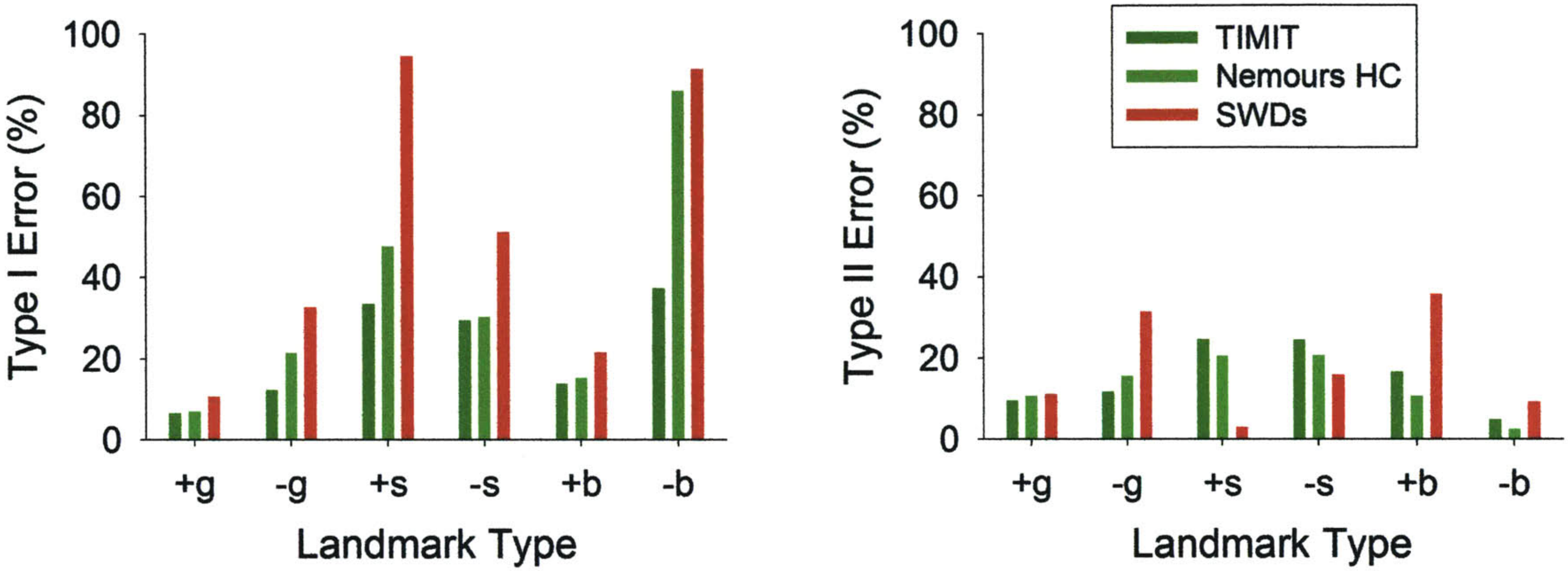

Figure 4.8: Type I (left) and type II (right) error rates associated with the TIMIT Database (dark green), the healthy control speaker from the Nemours Database (light green), and the ten SWDs from the Nemours Database.

Per-speaker analyses were also important for probability assignment given the wide range of speaker impairment. Figures 4.9 and 4.10 are histograms of the assigned probabilities for an individual with mild dysarthria (DYS10) and a speaker with severe dysarthria (DYS1), respectively. For the mildly impaired speaker (DYS10) a majority of detected glottis landmarks and detected burst release (*+b*) landmarks were assigned high probabilities (Figure 4.6). For the severe speaker (DYS1), the probabilities of detected landmarks were more widely distributed (Figure 4.7). For the speaker with mild dysarthria, the histograms for all inserted landmark instances, except burst releases (*+b*), roughly resembled ideal distributions. For the severe speaker, inserted burst releases (*+b*) and voicing offsets (*-g*) were poorly discriminated.

Tables 4.6 and 4.7 display the type I and type II error rates, respectively, for each SWD and the healthy control from the Nemours Database. While candidate peak detection rates for SWDs were always lower and peak insertion rates were always higher than those of the control, there were multiple instances where the type I or type II error rate associated with probability assignment for an individual with dysarthria was less than for the control. Plotting average type I and II error as functions of intelligibility revealed a significant ($\alpha < 0.01$) correlation between type I error and intelligibility (Figure 4.11). There was little correlation, however, between type II error and intelligibility. These findings suggest that more intelligible, less impaired individuals consistently manifested the expected acoustic correlates when producing an expected landmark. Unintelligible, highly impaired speakers, however, failed to manifest these correlates even when successfully producing an expected acoustic peak. The lack of correlation between type II error and intelligibility implies that while severely impaired speakers inserted more unexpected landmarks, the measured correlates for most of these insertions still differed from correlates associated with valid expected landmarks. This pattern was also observed for healthy and mild speakers.

Table 4.6: Per-speaker type I error rates for the individuals with dysarthria (DYS) and the healthy control speaker (HC) from the Nemours Database.

	Intelligibility	+g	-g	+s	-s	+b	-b	Average
DYS1	0	30%	42%	100%	67%	29%	91%	41%
DYS2	1	16%	34%	97%	91%	21%	100%	32%
DYS3	1	13%	36%	94%	41%	38%	80%	32%
DYS4	2	14%	35%	94%	47%	32%	97%	33%
DYS5	3	5%	63%	100%	60%	22%	92%	33%
DYS6	3	13%	27%	100%	51%	29%	85%	27%
DYS7	4	9%	18%	88%	31%	15%	81%	19%
DYS8	4	4%	36%	98%	23%	21%	86%	23%
DYS9	8	6%	25%	86%	59%	4%	100%	20%
DYS10	8	6%	29%	89%	43%	18%	94%	24%
HC		7%	21%	48%	30%	15%	86%	18%

Table 4.7: Per-speaker type II error rates for the individuals with dysarthria (DYS) and the healthy control speaker (HC) from the Nemours Database.

	Intelligibility	+g	-g	+s	-s	+b	-b	Average
DYS1	0	9%	49%	2%	11%	45%	23%	24%
DYS2	1	7%	33%	1%	6%	43%	12%	18%
DYS3	1	23%	28%	6%	24%	20%	3%	18%
DYS4	2	24%	36%	5%	25%	26%	0%	20%
DYS5	3	4%	12%	1%	15%	34%	4%	9%
DYS6	3	14%	40%	4%	17%	33%	4%	20%
DYS7	4	26%	21%	4%	26%	34%	3%	19%
DYS8	4	20%	29%	3%	12%	34%	13%	19%
DYS9	8	17%	46%	4%	19%	33%	4%	22%
DYS10	8	13%	29%	2%	20%	27%	6%	17%
HC		11%	16%	20%	21%	12%	3%	13%

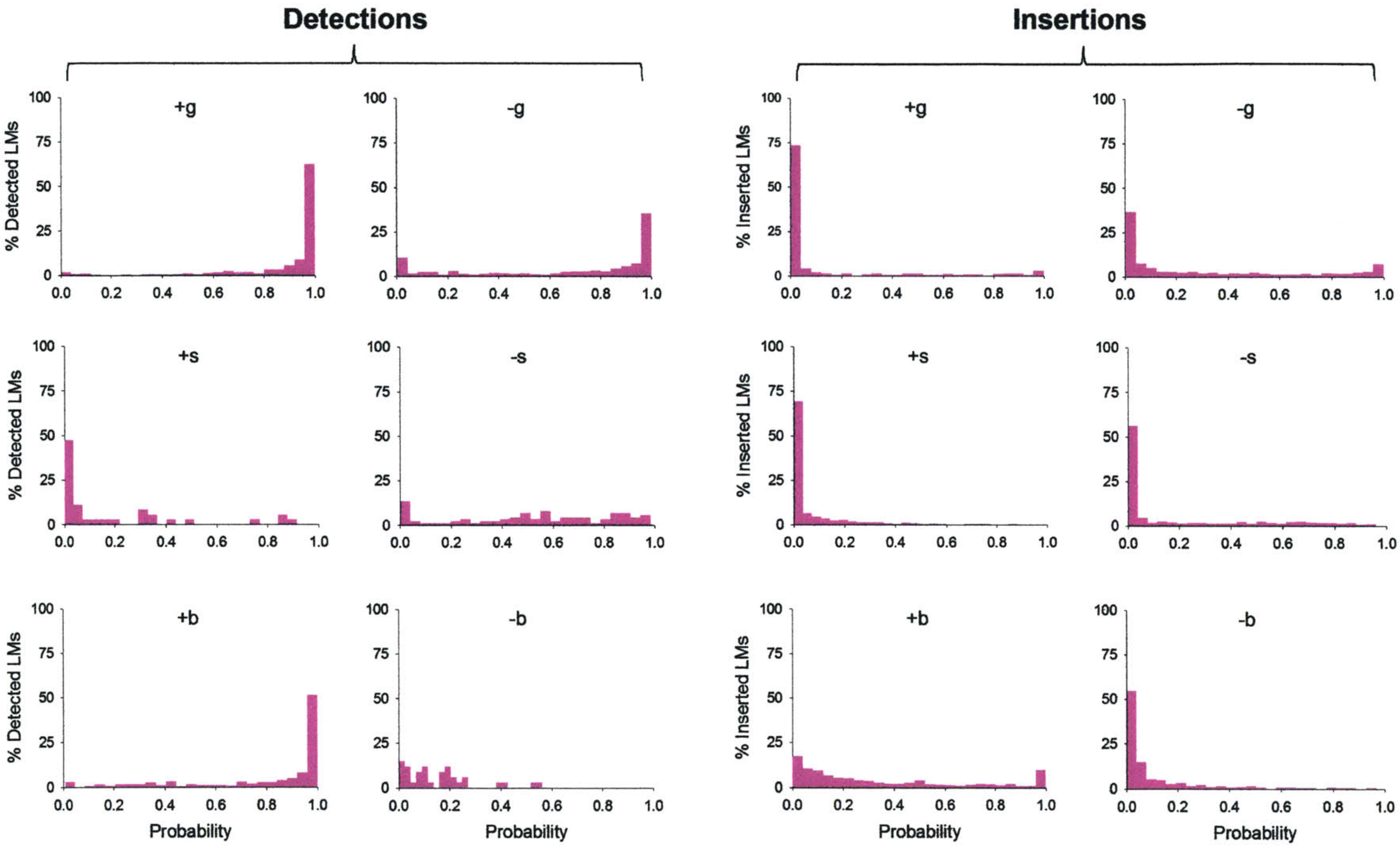

Figure 4.9: Histograms of the probabilities assigned to detected and inserted candidate landmarks peaks extracted from the test set recordings for an individual with mild dysarthria (DYS10).

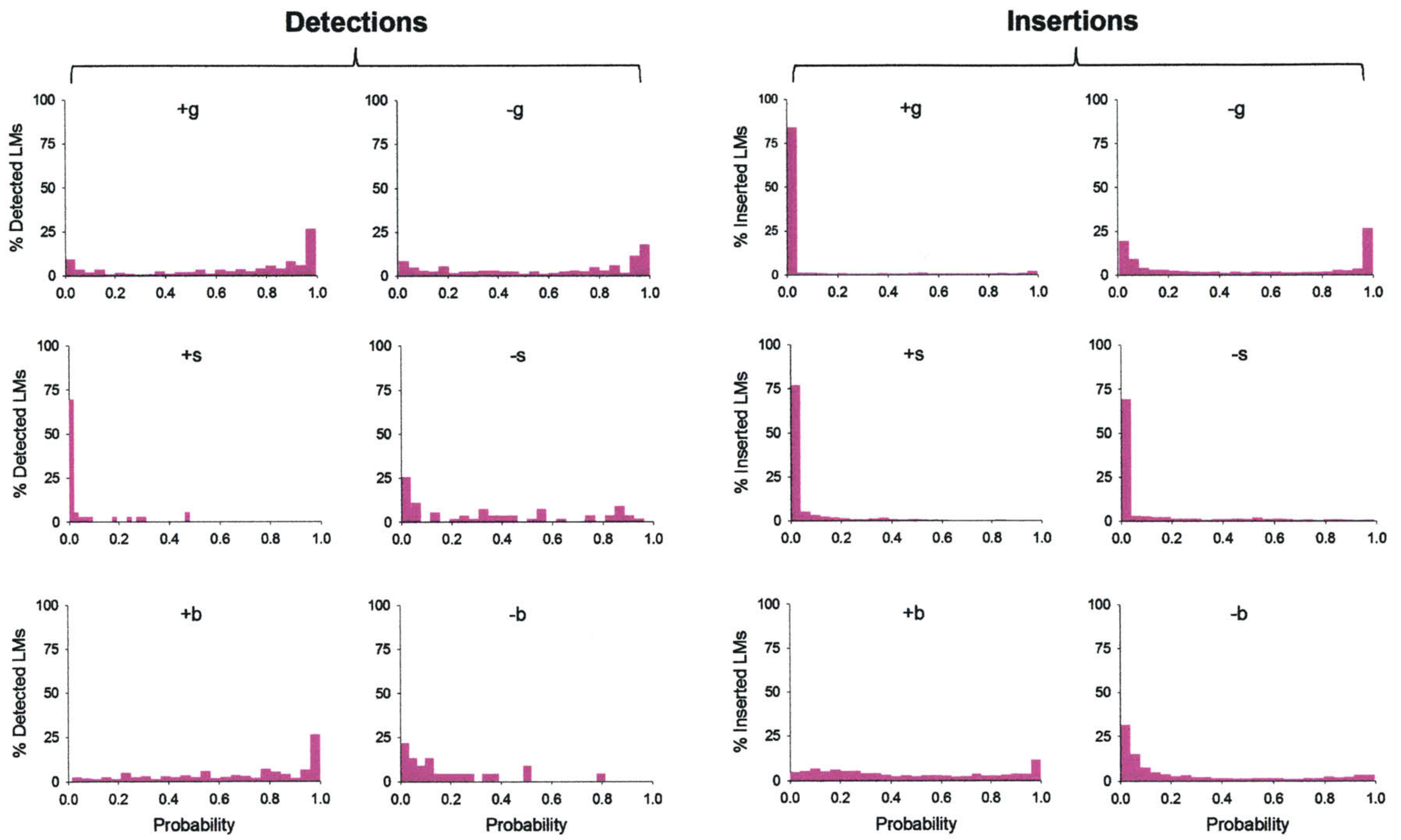

Figure 4.10: Histograms of the probabilities assigned to detected and inserted candidate landmarks peaks extracted from the test set recordings for an individual with severe dysarthria (DYS1).

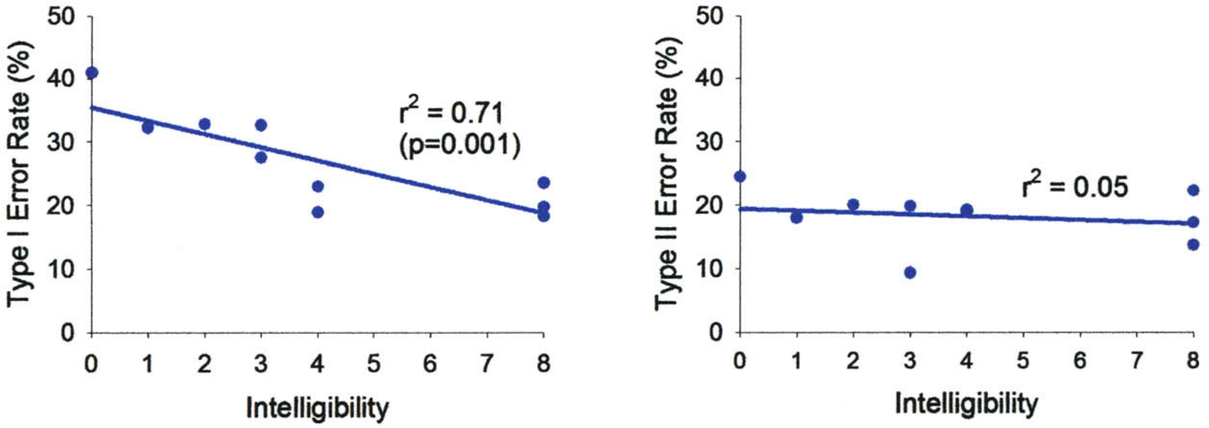

Figure 4.11: Type I (left) and type II (right) error rates as functions of speaker intelligibility. Linear correlations and the correlation coefficients are also shown. The intelligibility for the control speaker was assumed to be maximal (8).

4.3.3 Discussion of Baseline LAFF Analysis

Landmark analysis provided a means to relate acoustic phonetic events to underlying articulatory behavior. Detected landmarks that were assigned a high probability corresponded to production of expected acoustic targets. Assuming the base LAFF framework as a reference for typical human speech recognition, results suggest that SWDs from the Nemours Database often failed to achieve the extreme articulatory positions necessary to signal acoustic landmarks, as evidenced by diminished peak detection rates (Figure 4.3, Table 4.5). This is consistent with previous production studies that noted inadequate narrowing at the point of articulation for fricatives and sonorant consonants and incomplete contact for stops (Logemann & Fisher, 1981; Platt, Andrews, & Howie, 1980). When SWDs were able to signal candidate peaks, they manifested acoustic correlates differently than healthy speakers, as evidenced by reduced type I error rates. Elevated peak insertion rates for individuals with dysarthria suggest that SWDs produced many acoustic events that falsely indicated a point of lexical salience. In general, frequency of errors, in all forms, increased as degree of speaker impairment increased.

Agreeing with previous acoustic and perceptual studies of dysarthric speech (Ansel & Kent, 1992; Duffy, 2005; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980), findings from the base LAFF model suggest that imprecise and variable articulation results in distortion or deletion of information in the dysarthric speech stream. The observation that familiar listeners are often able to comprehend seemingly unintelligible speech agrees with the notion that segmental cues are distorted but still present (Deller et al., 1991; Neilson & O'Dwyer, 1984). This observation and the results of the base LAFF model on dysarthric speech led to the hypothesis that individuals with dysarthria convey acoustic landmarks differently than healthy speakers given their physiological constraints. However, the expected manifestation of landmarks and the potential tradeoff between cues could not be predicted from analysis of the base model alone.

Within the LAFF paradigm, acoustic landmarks are described by the frequency bands used to search for candidate peaks and by the acoustic correlates measured from around these peaks. If the hypothesis of segmental cue trading is correct then familiar listeners are either:

- Using different frequency bands, than those prescribed for healthy speakers, to detect candidate landmarks

  and/or

- Relying on unique sets of acoustic correlates to infer phonetic features.

Each component of the LAFF model was optimized for individual speakers to identify the manifestation of residual segmental cues in dysarthric speech. Optimization of candidate landmark extraction is discussed in Chapter 5 while identification of speaker-specific acoustic correlates is detailed in Chapter 6 (Figure 4.12).

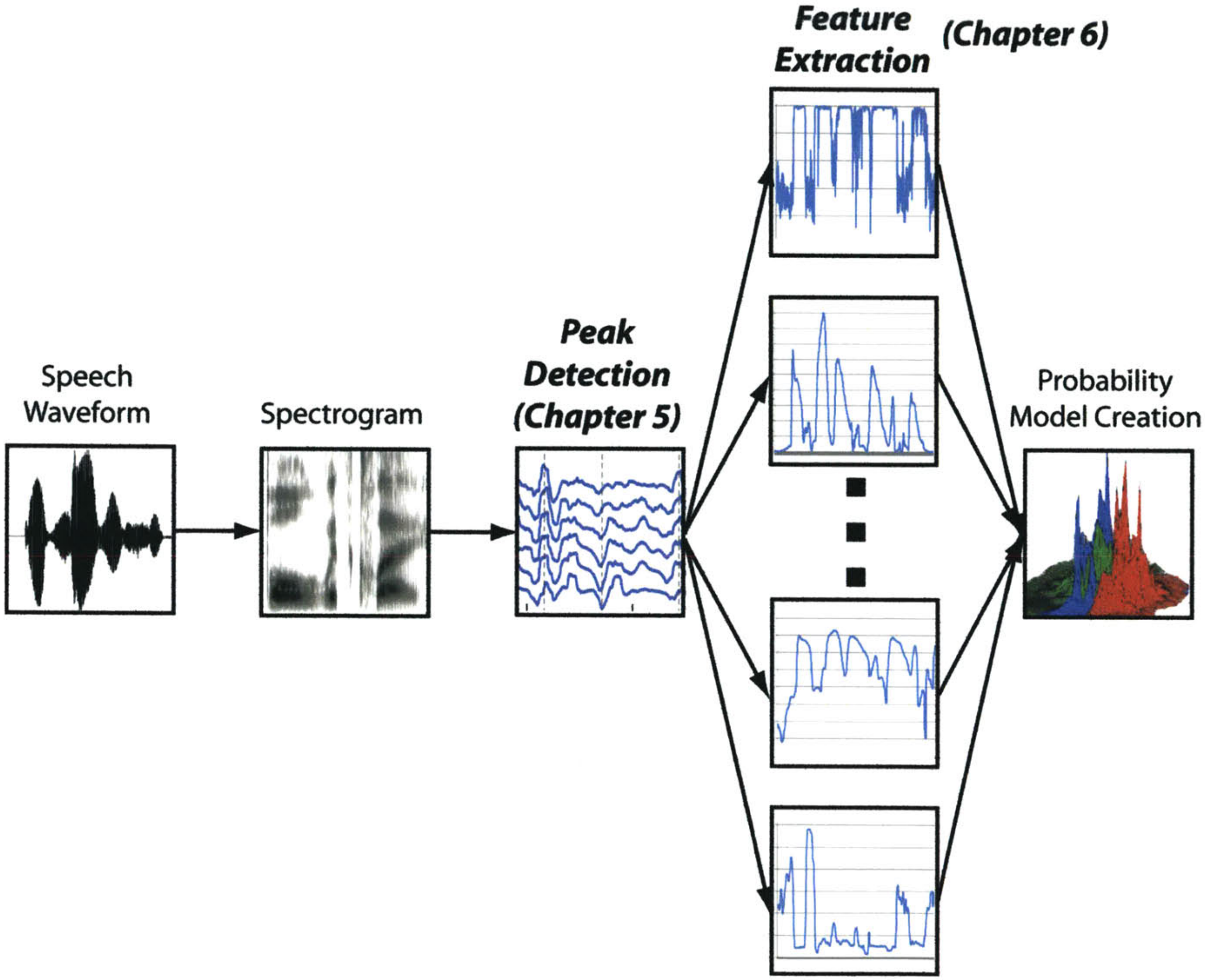

Figure 4.12: Overview of the LAFF paradigm, highlighting the components optimized for dysarthric speech. Chapter 5 presents a method to select optimal thresholds and frequency bands for a single speaker. Chapter 6 discusses identification of speaker-optimal acoustic correlates associated with each landmark instance.

Chapter 5

Landmark Candidate Extraction

In the LAFF model, extraction of candidate landmarks is performed by searching for peaks in six coarse frequency bands. Given the extent and degree of speech motor impairment, the appropriateness of the frequency bands suggested by Liu (1996) for healthy speakers was unclear. Initial expectation was that the same or similar bands would suffice for SWDs because the bands were already coarse and collectively covered the entire sampled frequency range. Also, baseline landmark analysis revealed that the thresholds used for peak extraction were too high (DiCicco & Patel, 2009). Weaker peaks in the energy waveforms were expected due to an inability to fully release constrictions or to form complete closures in the vocal tract.

To detail the manifestation of candidate landmark peaks in dysarthric speech, thresholds were tuned specific to each SWD (Section 5.1.1). Next, an expanded set of frequency bands was explored for candidate peaks (Section 5.1.2). The expanded sets of glottis bands incorporated additional low frequency bands targeted at F0 and F1. For sonorant and burst landmarks, broader bands were used given the increased variability in production of dysarthric speech. In the base LAFF model, sonorant and burst landmarks were extracted using multiple frequency bands while glottis landmarks were extracted using only a single band. In this work multiple glottis bands were also explored to determine if redundancy proved beneficial. Optimal sets of bands were selected by minimizing an error criterion that heavily penalized deletion of expected landmarks. A diagram summarizing the methods used to describe landmark manifestation within the assumptions of the LAFF model is shown in Figure 5.1.

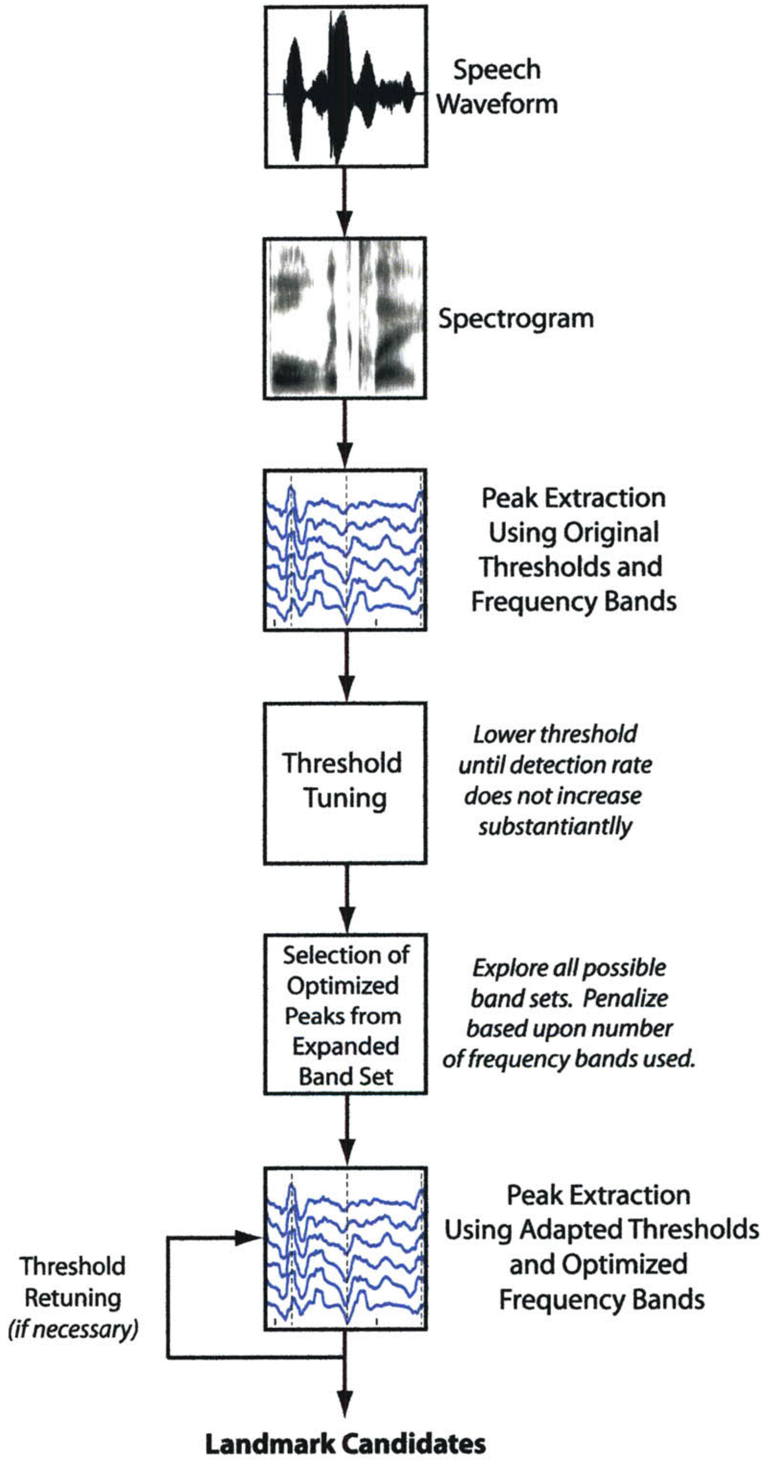

Figure 5.1: Optimization of candidate landmark peak extraction.

5.1.1 Threshold Adjustment

Using the training set for each SWD, optimal coarse- and fine-pass thresholds for each landmark type were found by first initializing the coarse-pass threshold to 10 dB and the fine-pass threshold to a smaller value (¼ of the coarse threshold)[12]. Next, the coarse-pass threshold was systematically decreased. Following each iteration, extracted peaks were compared to expected peaks to calculate detection and insertion rates. As the threshold was decreased, the rate of increase of detection rate decayed until reaching a point at which the detection rate plateaued (with the insertion rate continuing to grow). Plotting detection rate as a function of threshold, the abscissa corresponding to the "elbow" of the curve was selected as the value for the coarse-threshold (Figure 5.2)[13]. After setting the coarse-pass value, the fine-pass threshold was selected by increasing the value until the detection rate began to decay.

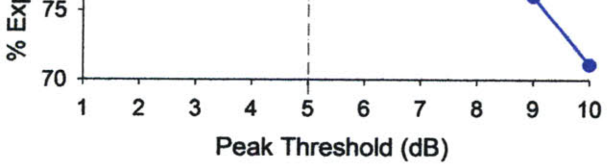

Figure 5.2: Example detection rate-vs-coarse-pass threshold curve. The value selected as the optimal threshold is indicated by the dotted line. Results are from a speaker with mild dysarthria.

[12] 10 dB was greater than the highest values suggested by Liu (1996) or Park (2008).

[13] Elbow estimation was done visually but various automated methods for estimating the optimal operating point have been suggested (Perkins & Schisterman, 2006). For locating the elbow in the curve visually, simple two piece-linear fits were compared.

5.1.2 Band Optimization

For each speaker, after setting the appropriate thresholds an expanded set of potential frequency bands was searched for candidate landmark peaks and optimal bands were selected for each landmark type. Table 5.1 lists the full set of frequency bands explored. The list includes the bands originally used for the LAFF model, additional static bands suggested by Chen, Jung, & Park (2007), and static bands related to the third through sixth formants (several of the originally suggested bands were already targeted to F1 and F2). Dynamic bands reliant on fundamental frequency and formant and bandwidth values were used as well. Dynamically varying bands did not, however, prove beneficial and are therefore not discussed further for optimal band selection[14].

Similar to the base LAFF model, a broadband spectrogram was first computed and the rate-of-rise of energy in each band was calculated using the speaker-optimized threshold values. After pairing up coarse- and fine-pass peaks, clustering of peaks was performed using the max-cut algorithm along with the constraints that peaks were sufficiently close to one another (within 50 ms), clusters did not overlap in time, and clusters did not contain multiple peaks from the same frequency band. Different sets of candidate frequency bands were explored for glottis and non-glottis landmarks because using all bands simultaneously often merged peaks related to closely adjacent landmarks. The most common merge typically involved nearby <+*b*,+*g*> pairs. Also, using only low frequency-related (F1 and below) bands for glottis landmarks eliminated potential confounding of feature identification downstream.

[14] The reliability of fundamental frequency and formant extraction from dysarthric speech using automated methods, specifically those found in Wavesurfer (2005), have not been thoroughly investigated; therefore, one possible explanation why dynamic bands may not have been useful was inadequate accuracy of extracted frequency values.

Table 5.1: Frequency bands explored during optimal band selection. Different sets were used for glottis and non-glottis landmarks. Bands dependent upon dynamic fundamental and formant frequency estimation are denoted using F#. B# refers to the dynamically extracted formant bandwidths. Dynamic frequency values were estimated using Wavesurfer (2005). Static F3 – F6 bands were suggested by Gaudrain et al. (2007). 640-2800, 0-(F3-1000), and F3-8000 were recommended by Bitar (1996). The remaining bands were suggested by Liu (1996) and N. Chen et al. (2007).

Glottis Landmarks		Sonorant & Burst Landmarks	
Frequency (Hz)	Correlate	Frequency (Hz)	Correlate
0 – 400	F0	800 – 1500	Nasal zero
100 – 400	F0	1200 – 2000	F2 for low vowels
300 – 900	F1	2000 – 3500	F2 for high vowels
400 – 800	F1	3500 – 5000	Lower frequency noise
0 – 2000	Lower Frequencies	5000 – 8000	Higher frequency noise
250-650	F1	3000 – 4000	F3
0-900	F0 and F1	0 – 8000	All Frequency Noise
0 – F1	Sub-F1	3000 – 8000	Fricative Noise
F0	F0 (closest spectral peak)	0 – 2000	Sub-mid frequencies
F1	F1 (closest spectral peak)	640 – 2800	F1 and F2 (syllabicity)
		1800 – 3240	F3
		2800 – 3550	F4
		3250 – 4450	F5
		3800 – 5900	F6
		F2	F2 (closest spectral peak)
		F3	F3 (closest spectral peak)
		F4	F4 (closest spectral peak)
		$(\text{F1-}\frac{B1}{2})-(\text{F1+}\frac{B1}{2})$	First formant bandwidth centered around F1
		$(\text{F2-}\frac{B2}{2})-(\text{F2+}\frac{B2}{2})$	Second formant bandwidth centered around F2
		$(\text{F3-}\frac{B3}{2})-(\text{F3+}\frac{B3}{2})$	Third formant bandwidth centered around F3
		$(\text{F4-}\frac{B4}{2})-(\text{F4+}\frac{B4}{2})$	Fourth formant bandwidth centered around F4
		0 – (F3-1000)	Sub-mid frequencies
		F3 - 8000	Mid-frequencies & above

After extracting peaks in the candidate frequency bands, optimal selection of bands was performed for each individual from the Nemours Database. A key component to selecting bands was also setting the appropriate number of peaks required to declare a cluster of peaks a landmark. To determine both the best frequency bands and the corresponding required number of peaks, the error function shown below was calculated for all possible <bands, required number of peaks> pairs for each landmark type:

$$Error(b,p) = \left(\alpha * DeletionRate + InsertionRate\right) * Description\ Length$$

b corresponded to a specific set of frequency bands belonging to the complete set B and p was the number of required peaks with $p \leq ||b||$. The description length penalized larger sets of bands and was given by the expression $1+\|b\|^2 \log(N)/N$, where N was the number of training samples (Rissanen, 1978). α was set to 5 to heavily bias the function towards the deletion rate because deletions cannot be recovered and are thus more severe of an error than insertions[15]. Using the data from the speaker's training set, the error function was calculated for each landmark type. The <bands, numbers of peaks> pair that minimized the error function was selected as the optimal set of parameters.

5.1.3 Results and Discussion

Candidate peak detection and insertion rates associated with using the base model (*Base*), following threshold tuning (*Adjustment*), and after band optimization (*Optimized*) are presented in Table 5.2. The thresholds appropriate for the Nemours healthy speaker were used to calculate the baseline rates. Rates were computed using the test sets withheld from threshold tuning and band selection. Threshold adjustment resulted in increases of detection rate ranging from 5% to 36% but also increases in insertion rate ranging from 24% to 106%. During threshold tuning,

[15] Using a larger weighting factor did little to improve detection rate.

speakers with severe dysarthria typically required the lowest thresholds, possibly due to a greater inability to form complete closures and fully release constrictions. Low thresholds resulted in large increases in both detection and insertion rates. After threshold adjustment, detection rates varied less across speaker but insertion rates ranged considerably, with less intelligible speakers exhibiting very high insertion rates. While elevated insertion rates could ultimately pose a problem in landmark sequence construction, it was encouraging that threshold tuning resulted in noticeable improvements in detection rate because deleted landmarks cannot be recovered in the model.

Table 5.2: Detection and insertion rates for all speakers from the Nemours Database. Included are the overall and relative effects of threshold adjustment and band optimization. The baseline rates were acquired using thresholds appropriate for the healthy speaker from the Nemours Database (HC).

					Relative Increase				Relative Increase	
	Base Settings		Adjusted Threshold		Adjusted:Base		Optimized Bands		Optimized:Base	
Subject	Detected	(Inserted)	Detected	(Inserted)	Detected	(Inserted)	Detected	(Inserted)	Detected	(Inserted)
DYS1	56%	(1141%)	77%	(1830%)	+37%	(+60%)	72%	(1559%)	+29%	(+37%)
DYS2	76%	(973%)	86%	(1387%)	+13%	(+42%)	83%	(1249%)	+9%	(+28%)
DYS3	77%	(480%)	86%	(631%)	+12%	(+31%)	85%	(581%)	+10%	(+21%)
DYS4	70%	(388%)	88%	(650%)	+26%	(+68%)	87%	(571%)	+24%	(+47%)
DYS5	60%	(792%)	76%	(1370%)	+26%	(+73%)	73%	(1184%)	+22%	(+50%)
DYS6	67%	(485%)	78%	(738%)	+17%	(+52%)	75%	(648%)	+12%	(+33%)
DYS7	84%	(571%)	89%	(667%)	+6%	(+17%)	87%	(619%)	+4%	(+8%)
DYS8	86%	(511%)	91%	(640%)	+5%	(+25%)	90%	(579%)	+5%	(+13%)
DYS9	83%	(356%)	89%	(492%)	+7%	(+38%)	88%	(454%)	+6%	(+27%)
DYS10	83%	(330%)	87%	(389%)	+5%	(+18%)	86%	(370%)	+4%	(+12%)
HC	91%	(174%)	-		-		91%	(168%)	0%	(-3%)

While it was unclear whether band optimization on a per-speaker basis would prove effective or yield significant differences between speakers, the results were encouraging in that improvements were noted over using just the base LAFF frequency bands. For all speakers, peaks in certain bands were more prevalent in detections than in insertions. This phenomenon was especially noticeable for glottis landmarks. Intensity plots of the percentage of detected and inserted landmarks that contain peaks in each of the indicated frequency bands for the Nemours control speaker, an individual with mild dysarthria, and a speaker with severe dysarthria are shown in Figures 5.3 through 5.5, respectively. For all three speakers there were bands that were red for detections, indicating that more than 80% of detected landmarks contained peaks in the specified bands, but were blue for insertions, indicating that less than 50% of inserted landmarks contained peaks in these same bands. Also, there were differences in the number of peaks present in the clusters associated with detected and inserted landmarks (Figure 5.6). Detected landmarks typically had more peaks in the associated frequency bands than inserted landmarks. This pattern was observed for all speakers. Utilizing dissimilarity in cluster structure between detected and inserted landmarks enabled some insertions to be eliminated without seriously impacting detection rates (Table 5.2). For all speakers, optimal band selection decreased insertion rates relative to their post-threshold adjustment values, with relative decreases as large as 33%. Detection rates for all speakers also decreased but to a much lesser extent, with relative decreases ranging from 1% to 6%. If desired, the error function used during optimization could be altered as a means of impacting detection rates even less but this would come at the cost of additional insertions.

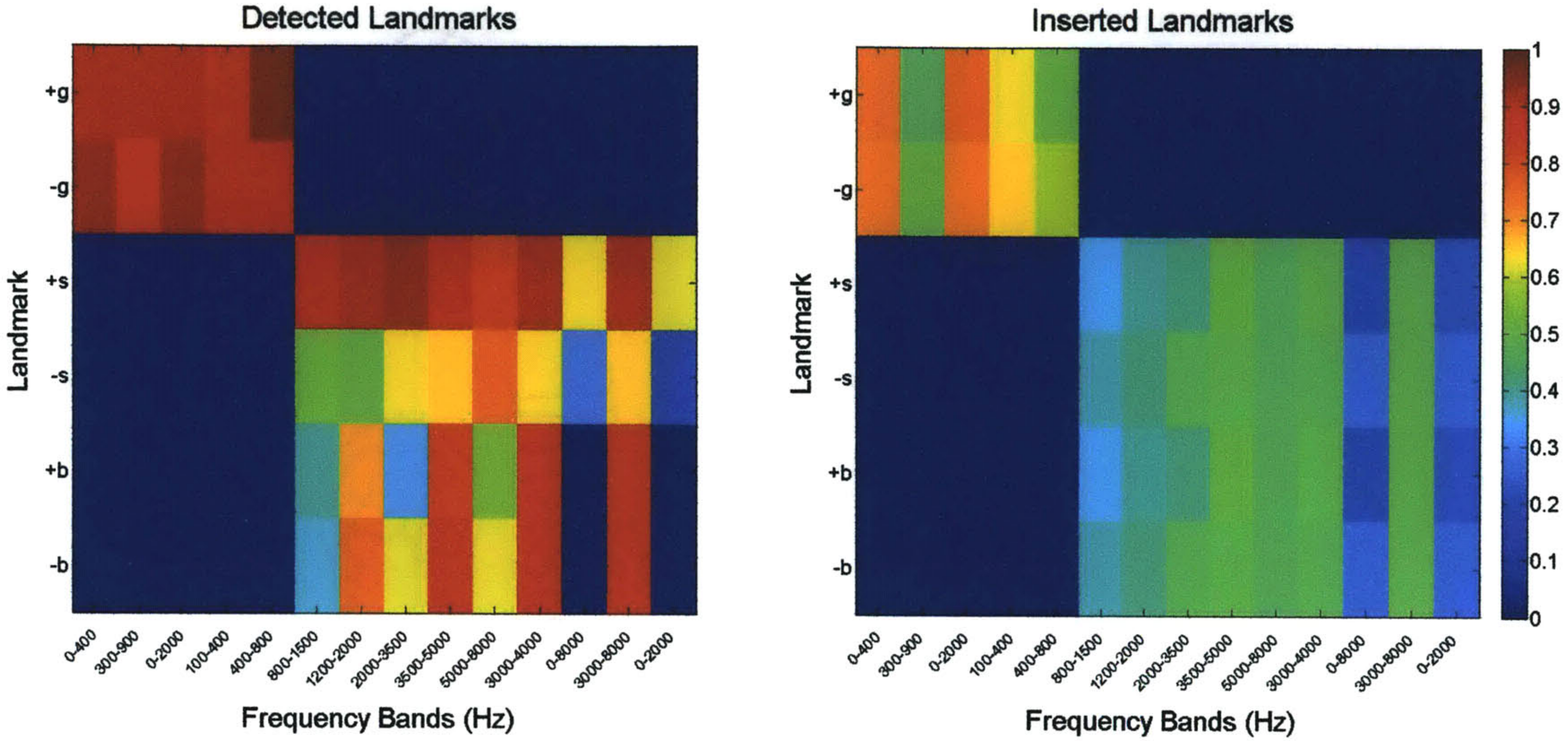

Figure 5.3: Comparison of the prevalence of peaks in specific frequency bands for detected and inserted landmarks from the healthy speaker (HC) of the Nemours Database. For compactness only a limited subset of the bands explored are actually shown.

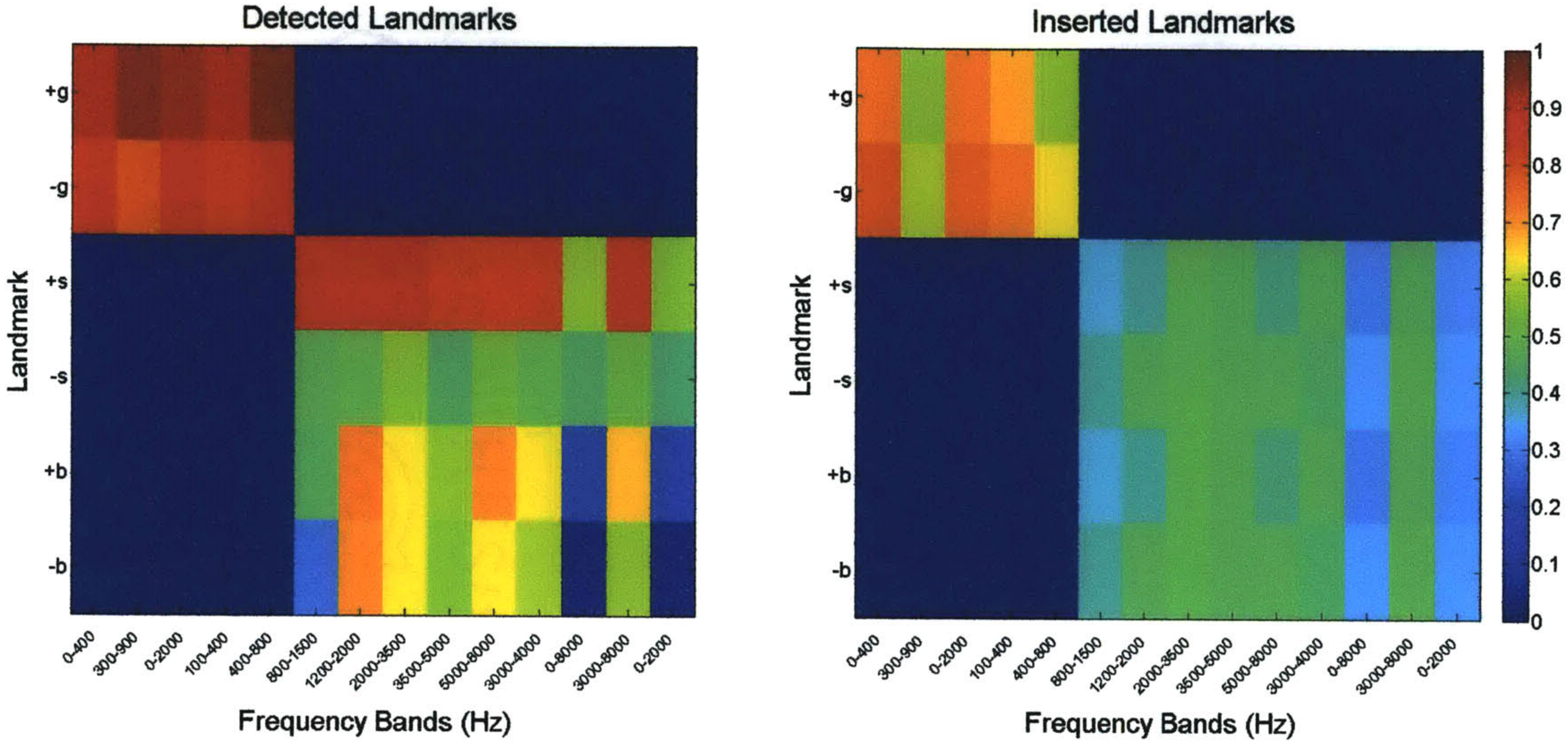

Figure 5.4: Comparison of the prevalence of peaks in specific frequency bands for detected and inserted landmarks from an individual with mild dysarthria (DYS10). For compactness only a limited subset of the bands explored are actually shown.

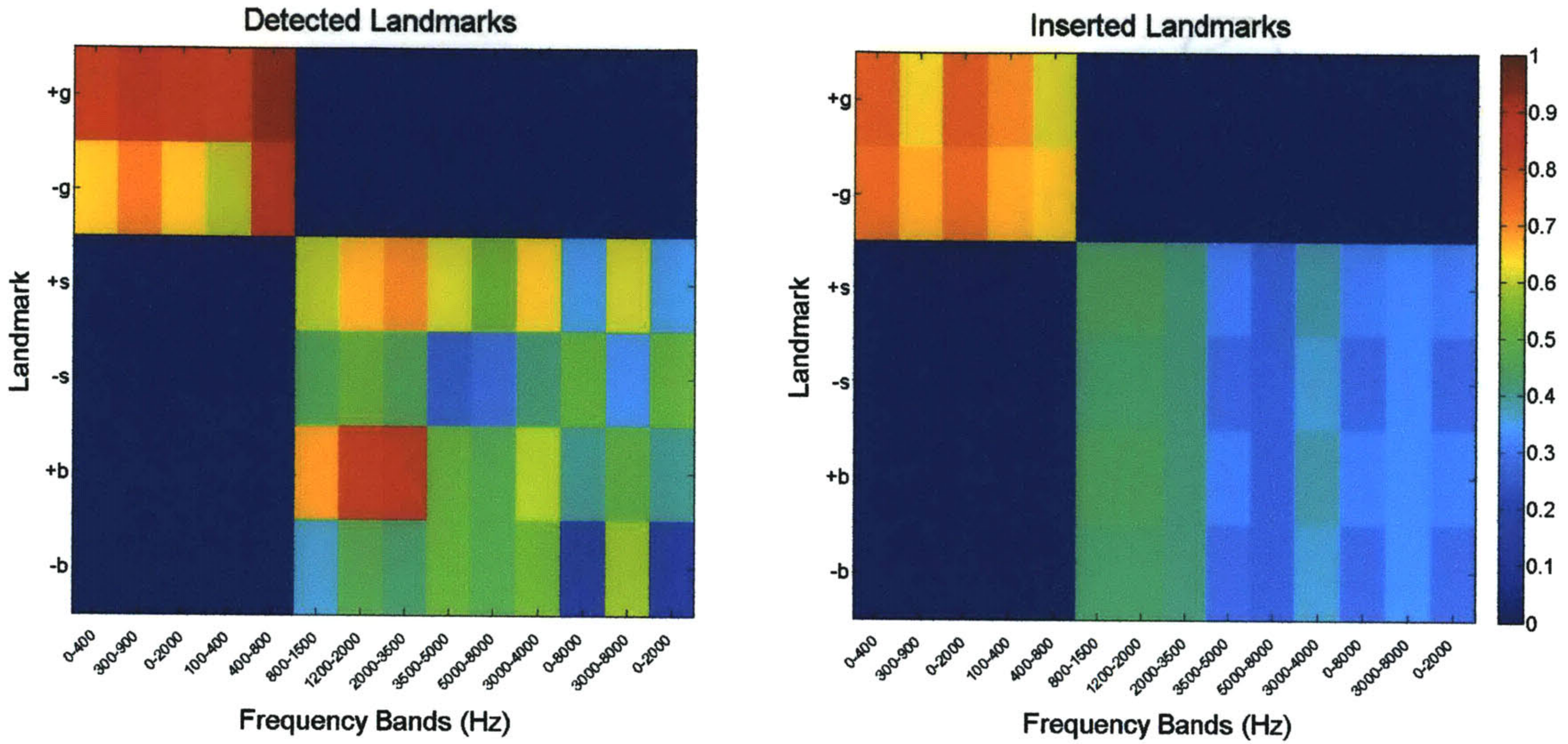

Figure 5.5: Comparison of the prevalence of peaks in specific frequency bands for detected and candidate landmarks from an individual with severe dysarthria (DYS1). For compactness only a limited subset of the bands explored are actually shown.

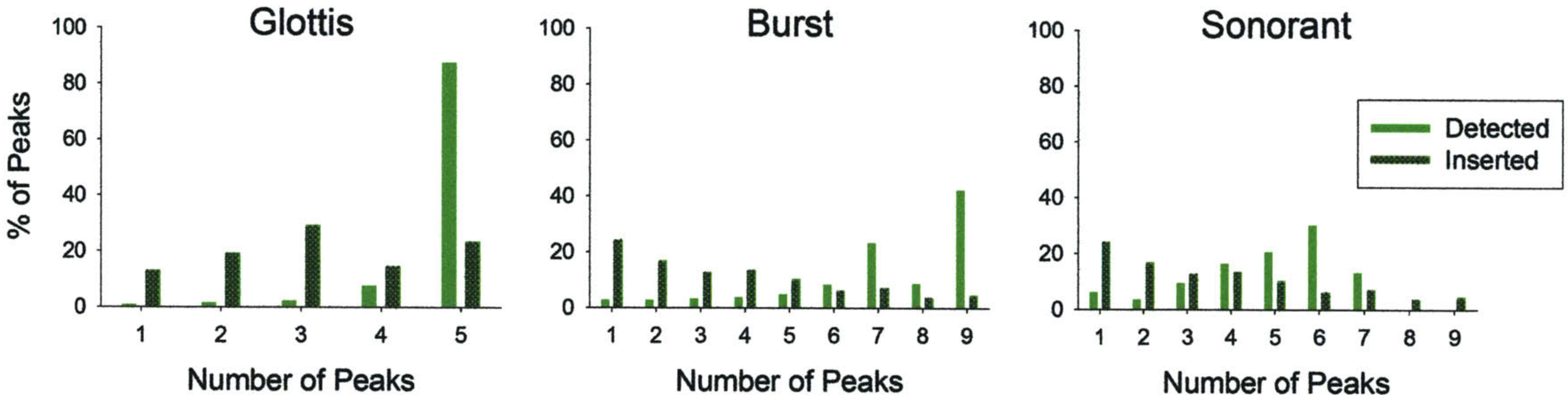

Figure 5.6: Percentage of clusters that contained peaks in a specified number of frequency bands. For compactness, the results shown were acquired using a limited number of bands. For glottis landmarks, the first five glottis bands listed in Table 5.1 were used. For sonorant and burst landmarks the first nine non-glottis bands listed in Table 5.1 were used. Results are from an individual with mild dysarthria (DYS10).

The data in Table 5.2 demonstrate that threshold adjustment and optimal band selection resulted in improved detection of candidate landmarks for SWDs. To determine if individuals with dysarthria were using unique sets of peaks compared to healthy speakers, the peaks deemed optimal were compared. Tables 5.3 through 5.5 contain the per-speaker sets of optimized bands for each landmark type. The sets of bands used for burst and sonorant landmarks closely align with the baseline set prescribed by Liu, with the exception that broadband energy from 0 - 8000Hz was frequently included (especially for burst landmarks). Optimization always suggested using multiple bands for glottis landmark extraction. While there was large overlap between several of the low frequency bands, this redundancy reinforced the likelihood that a cluster of peaks was actually a glottis landmark. Figure 5.6 is an example of this pattern with 95% of detected glottis landmarks containing peaks in four or five of the five low-frequency regions while 32% of inserted glottis landmarks only had peaks in one or two bands. This pattern was noted for all speakers in the Nemours Database. Given that band optimization only suggested novel band sets for glottis landmarks, improvements in insertion rates were primarily due to reduced glottis insertions (Table 5.6). For sonorant and burst landmarks, band selection reduced post-threshold insertion rates by less than 10%. Glottis landmark optimization, however, reduced insertion rates 10% to 24% for SWDs.

Table 5.3: Optimal frequency bands identified for glottis landmarks for each speaker from the Nemours Database. All frequencies are in Hz.

Speaker	[0-400]	[300-900]	[0-2000]	[100-400]	[400-800]	[250-650]	[0-900]
DYS1		X		X	X		
DYS2		X		X	X		
DYS3			X		X	X	
DYS4		X		X			X
DYS5	X	X			X		
DYS6				X	X	X	X
DYS7		X		X	X		
DYS8	X			X		X	X
DYS9		X		X	X		
DYS10		X	X	X	X		
HC		X		X	X		

Table 5.4: Optimal frequency bands identified for sonorant landmarks for each speaker from the Nemours Database. All frequencies are in Hz.

	[800-1500]	[1200-2000]	[2000-3500]	[3500-5000]	[5000-8000]	[3000-4000]	[0-8000]	[3000-8000]	[0-2000]	[640-2800]	[1800-3240]	[2800-3550]	[3250-4450]	[3800-5900]
DYS1		X	X	X	X		X	X						
DYS2	X	X			X	X			X		X			
DYS3		X			X	X			X		X		X	
DYS4	X	X	X				X							X
DYS5		X	X	X	X			X						
DYS6						X	X	X		X		X		
DYS7							X			X		X		X
DYS8		X	X	X	X	X			X					
DYS9		X			X		X				X		X	
DYS10			X		X	X	X			X				
HC		X		X		X	X	X						

Table 5.5: Optimal frequency bands identified for burst landmarks for each speaker from the Nemours Database. All frequencies are in Hz.

	[800-1500]	[1200-2000]	[2000-3500]	[3500-5000]	[5000-8000]	[3000-4000]	[0-8000]	[3000-8000]	[0-2000]	[640-2800]	[1800-3240]	[2800-3550]	[3250-4450]	[3800-5900]
DYS1		X	X	X	X	X			X					
DYS2	X						X				X		X	X
DYS3					X	X			X	X				
DYS4		X	X		X		X						X	
DYS5			X	X				X	X			X		
DYS6	X				X	X	X			X				
DYS7	X		X	X	X				X					
DYS8							X			X		X		X
DYS9	X		X		X	X	X							
DYS10	X					X	X				X			X
HC	X	X	X		X	X	X							

Table 5.6: Change in per-landmark type insertion rate following band optimization. The post-threshold adjustment values were the reference insertion rates.

Speaker	Decrease in Glottis Insertion Rate	Decrease in Burst Insertion Rate	Decrease in Sonorant Insertion Rate
DYS1	22%	6%	8%
DYS2	17%	5%	4%
DYS3	16%	3%	3%
DYS4	21%	7%	6%
DYS5	24%	3%	9%
DYS6	21%	8%	8%
DYS7	20%	2%	7%
DYS8	17%	6%	1%
DYS9	16%	3%	2%
DYS10	10%	1%	2%
HC	6%	1%	2%

The results from this chapter suggest that peak detection for glottis landmarks was not optimal in the base LAFF model and that selection of frequency bands beyond the set recommended by Liu (1996) was of little value for sonorant and burst landmarks. Individuals with dysarthria often were not able to produce the articulatory movements necessary to properly buildup pressure in the vocal tract or to completely release constrictions, resulting in less abrupt segmental contrasts. Threshold adjustment increased detection rate but was met with an increase in insertion rate. While the magnitude of candidate peaks was reduced, the sets of bands selected for SWDs suggest that, like healthy speakers, SWDs signal acoustic landmarks using coarse frequency bands that collectively span a broad range of frequencies. Ultimately there were not large differences between the bands identified for healthy and dysarthric speakers. This implies that if the work's underlying hypothesis was correct, SWDs must be relying on different combinations of acoustic correlates manifested around landmarks than healthy speakers.

Chapter 6

Per-Speaker Acoustic Correlate Identification

Park's implementation of the LAFF framework used sets of three or four acoustic correlates for each landmark type. These correlates did not prove reliable for discriminating between detected and falsely inserted landmarks for speakers with severe dysarthria from the Nemours Database (Section 4.3.2). To test the hypothesis that SWDs use alternative sets of segmental cues, a system aimed at identifying discriminable acoustic correlates for a single speaker was implemented. The identification process involved three steps:

1.) Calculating an expanded set of acoustic correlates without regard to landmark type.

2.) Using a mixture of greedy and global optimization techniques to select speaker-optimized correlates.

3.) Building probability models for each landmark instance using speaker adaptation.

A diagram summarizing the complete correlate optimization scheme is shown in Figure 6.1

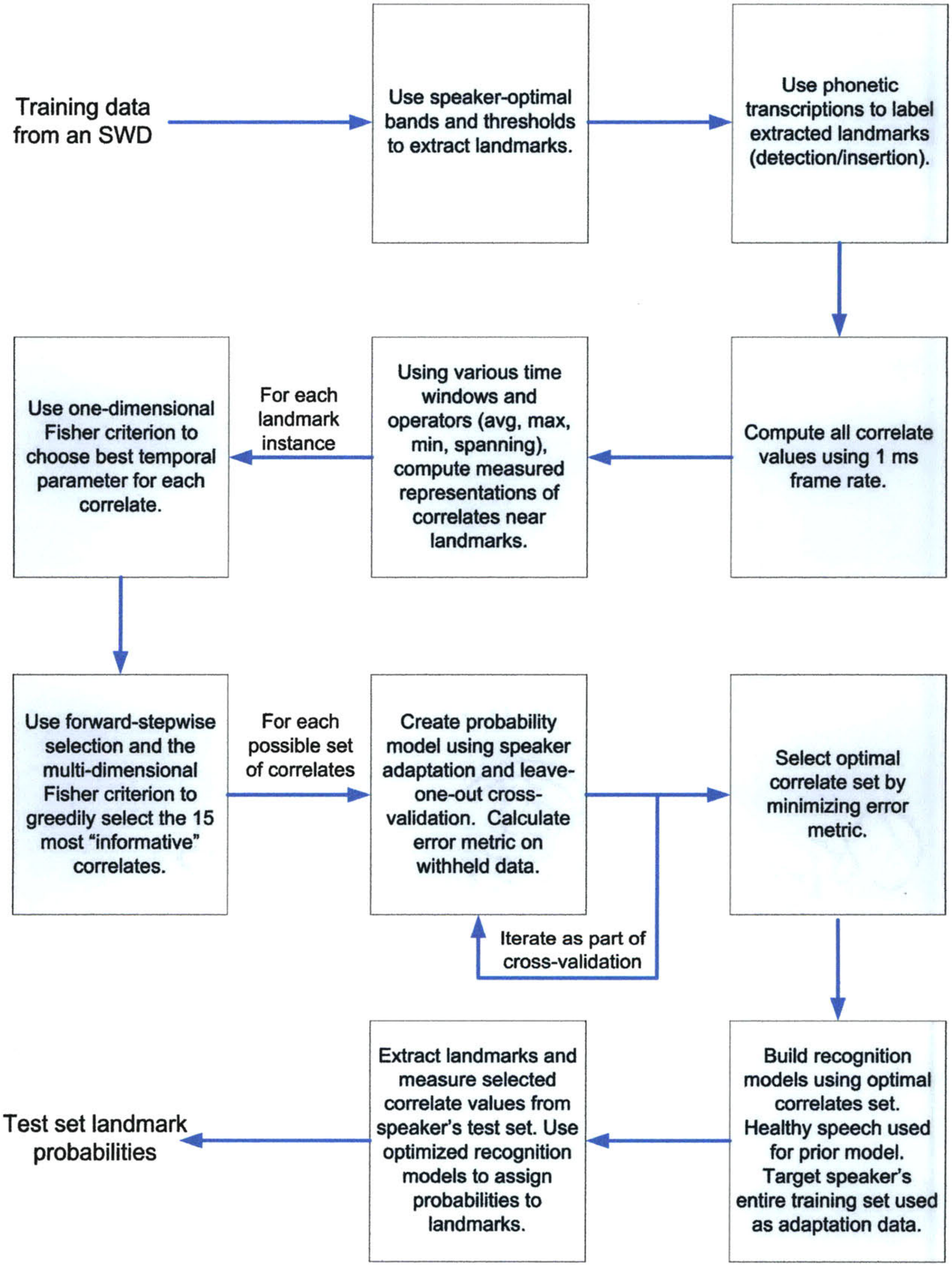
Training data from an SWD
Use speaker-optimal bands and thresholds to extract landmarks.
Use phonetic transcriptions to label extracted landmarks (detection/insertion).
Compute all correlate values using 1 ms frame rate.
Using various time windows and operators (avg, max, min, spanning), compute measured representations of correlates near landmarks.
For each landmark instance
Use one-dimensional Fisher criterion to choose best temporal parameter for each correlate.
Use forward-stepwise selection and the multi-dimensional Fisher criterion to greedily select the 15 most "informative" correlates.
For each possible set of correlates
Create probability model using speaker adaptation and leave-one-out cross-validation. Calculate error metric on withheld data.
Iterate as part of cross-validation
Select optimal correlate set by minimizing error metric.
Build recognition models using optimal correlates set. Healthy speech used for prior model. Target speaker's entire training set used as adaptation data.
Extract landmarks and measure selected correlate values from speaker's test set. Use optimized recognition models to assign probabilities to landmarks.
Test set landmark probabilities

Figure 6.1: Schematic of process for identifying optimal correlates for a single speaker.

6.1 Candidate Acoustic Correlates

The sets of acoustic correlates used in the base LAFF model were based on empirical observations of healthy speech (Park, 2008). The correlates for glottis landmarks relied on low frequency energy measures while the correlates for sonorant and burst landmarks utilized broadband energy above F1. Analysis with the base LAFF model revealed that while these correlates were adequate for discriminating between detected and inserted instances of landmarks for healthy speakers and individuals with mild dysarthria, they did not accurately model expected landmarks for individuals with moderate-to-severe impairment, as evidenced by elevated type I error rates.

To determine if SWDs rely on different sets of acoustic correlates than healthy speakers, an expanded set of correlates, identified in the acoustic phonetic literature, was incorporated (Table 6.1). Similar to the base LAFF correlates, many were static frequency-bounded measures of energy and rate-of-rise. Correlates dependent upon dynamic fundamental and formant frequency estimates were also included. Finally, several other correlates implemented and ultimately selected during optimization included the zero crossing rate (ZCR), multiple measures of periodicity including the autocorrelation coefficient (Rabiner & Juang, 1993) and the normalized cross-correlation function (NCCF, also referred to as voicing probability; Talkin, 1995), and onset/offset energy (Espy-Wilson, 1992; Salomon, Espy-Wilson, & Deshmukh, 2004). Onset/offset energy was similar to rate-of-rise but used non-overlapping frames to prevent obscuring of brief, abrupt changes. While previous studies have noted hallmarks of dysarthric speech, dysarthria is often heterogeneous in its presentation. Therefore, all correlates were calculated for each landmark instance in the event that a SWD used correlates not typically associated with a specific landmark instance.

For most of the correlates, the values of the correlate to the left and right as well as the difference across the landmark were computed. In Park's implementation of the LAFF model, correlates were calculated using a temporal spanning criterion where the search window was bounded by the previous/proceeding landmark. The appropriateness of using this spanning criterion, which was aimed at eliminating very short perturbations in energy, for SWDs was unclear. Therefore, for measuring correlates, average, minimum, and maximum values calculated over a time window were utilized along with the spanning level. In all, over 300 acoustic correlate measures (each computed using various time windows) were calculated for each landmark. Selection and adjustment of the relevant temporal parameters is described in the next section.

Table 6.1: Correlates explored during optimization. For the correlates specified by frequency bounds, several representations of the correlate were computed including energy, spectral peak, and rate-of-rise. Energy values were the sum of the squared spectral values located within the specified bounds, divided by the number of channels. Spectral peak (SP) measures were calculated from the maximum spectrum value located within the bounds. Rate-of-rise (ROR), a form of temporal derivative, was calculated using the energy waveform within the frequency bounds. For all correlates, values were computed to the left and right of candidate peaks and differences across landmarks were also measured. Correlate values were calculated using the minimum, maximum, average, and spanning operators using a specified time window. For minimum, maximum, and average values the time window was set as a free parameter. For calculating the spanning minimum/maximum values, the window between the target peak and the previous/proceeding landmark was searched for the lowest/highest correlate value that continuously occurred for the duration specified by a free temporal parameter. For each correlate representation, an optimal time window was selected. Relevant citations - a: Liu (1996) b: Park (2008); c: Bitar & Espy-Wilson (1996); d: Mercier et al. (1990); e: N. Chen et al. (2007); f: Juneja & Espy-Wilson (2008); g: Bitar & Espy-Wilson (1995); h: Espy-Wilson et al. (2007); i: Bitar (1998); j: Waibel & Lee (1990); k: Espy-Wilson (1992); l: Salomon et al. (2004); m: Aye (2009); n: Talkin (1995); o: Gaudrain et al. (2007).

Correlate	Description	Correlate	Description
0-400Hz	Low frequency (voicing)[a]	3250-4450	F5[o]
100-400Hz	Low frequency (voicing)[c]	3800-5900	F6[o]
250-650 Hz	Low frequency (voicing)[d]	F0	Fundamental frequency (dynamic). Closest spectral peak.
250-850 Hz	Low frequency [d]	F1	First formant (dynamic). Closest spectral peak.
300-900 Hz	Low frequency (F1)[e]	F2	Second formant (dynamic). Closest spectral peak.
400-800 Hz	Low frequency (F1)[e]	F3	Third formant (dynamic). Closest spectral peak.
0-900 Hz	Low frequency (F0 and F1)[f]	F4	Fourth formant (dynamic). Closest spectral peak.
0-2000 Hz	Sub-mid frequency[c]	(F1-B1/2)-(F1+B1/2)	Energy contained within first formant bandwidth (dynamic).
$0-(F3_{avg}-1000)$	Sub-mid frequency[f]	(F2-B2/2)-(F2+B3/2)	Energy contained within second formant bandwidth (dynamic).
640-2800 Hz	Syllabicity, low-to-mid frequency[c]	(F2-B2/2)-(F3+B3/2)	Energy contained within third formant bandwidth (dynamic).
800-1500 Hz	Nasal zero[a]	(F4-B4/2)-(F4+B4/2)	Energy contained within fourth formant bandwidth (dynamic).
1200-2000 Hz	F2 for low vowels[a]	0-F1	Sub-F1, low frequency (dynamic).
1200-3500 Hz	F2[e]	F1-8000 Hz	Supra-F1 (dynamic).
2000-3000 Hz	Syllabicity, mid-frequency[a]	(0-2000)	Ratio of low frequency to broad (mid and

		:(2000-8000)	high) frequency[g]
2000-3500 Hz	Mid-frequency, F2 for high vowels[a]	(0-360):(0-5000)	Spectral tilt (low frequency density)[b]
($F3_{avg}$-1k)- ($F3_{avg}$+1.7k)	Mid-frequency[i]	(0-360):(0-8000)	Spectral tilt (low frequency density)
3500-5000 Hz	Lower frequency noise[a]	(0-F1):(F1-8000)	Sub-F1 density
5000-8000 Hz	High frequency noise[a]	Degree of openness	F1-F0 (using closest spectral peaks, dynamic)[j]
3000-8000 Hz	Fricative noise[e]	M1	Spectral center of gravity[i]
($F3_{avg}$-1000)-8000 Hz	Mid and high frequency[a]	RMS energy	Root mean square energy[j]
2000-8000 Hz	Mid and high frequency[g]	Energy onset/offset	First difference broadband energy calculated using non-overlapping frames[k,l]
0-8000 Hz	Broad frequency[e]	Glottal energy onset/offset	First difference low frequency energy calculated (100-400 Hz) using non-overlapping frames[k,l]
1200-8000 Hz	Broad (supra-F1) frequency[b]	ZCR	Zero crossing rate[m]
0-5000 Hz	Broad frequency[b]	Autocorrelation coefficient	Autocorrelation coefficient, (periodicity measure)[n]
1800-3240	F3[o]	NCCF	Normalized cross-correlation function (voicing probability)[n]
2800-3550	F4[o]	R1	Normalized 1st cross-correlation coefficient (periodicity measure)[g]

6.2 Correlate Optimization

Previous studies of dysarthric speech have noted that rate of speech production and articulator velocity was often reduced, (Hirose, Kiritani, & Sawashima, 1982; Murdoch, 1998; Yunusova, Weismer, Westbury, & Lindstrom, 2008) resulting in temporal differences compared to healthy speech (Ansel & Kent, 1992; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980; Vijayalakshmi & Reddy, 2006; Ziegler, Hartmann, & Hoole, 1993). Since temporal parameters directly impact measured correlate values, for each correlate representation the associated time window was set as a free parameter. Windows ranged over a wide grid, from 5 ms to 120 ms with 5 ms steps. The lower bound was set to track rapidly occurring events, specifically burst releases (Salomon et al., 2004). The upper bound was derived by noting that segmental cues in healthy speech were sometimes located as far as 80 ms away from the relevant landmark (Poeppel, Idsardi, & van Wassenhove, 2008). This value was then increased by 50% as a conservative estimate for dysarthric speech.

The per-speaker optimal thresholds and frequency bands determined in Chapter 5 were used to identify candidate landmark peaks from the Nemours training set. For each extracted landmark, correlate values were computed over the entire range of time windows. Time-aligned phonetic labels were used to identify extracted landmarks as detections, substitutions, and insertions. After landmark identification, the feature space was pared down by selecting an optimal temporal window for each of the correlates. This was done in a greedy fashion via maximization of the Fisher criterion.

6.2.1 Fisher Criterion

The Fisher criterion (FC) represented the effectiveness of a given feature vector for characterizing competing classes (true or false landmarks), using the distance between classes as the selection criterion (Luebke & Weihs, 2005). The FC was defined relative to the ratio of the between-class scatter and the within-class scatter (Bitar, 1998). The within-class scatter was defined as:

$$S_w = \frac{1}{n}\sum_{i=1}^{C} S_i$$

where S_i was proportional to class i's covariance matrix and was equal to:

$$S_i = \sum_{j=1}^{n_i} (x_j^{(i)} - m_i)(x_j^{(i)} - m_i)^T$$

C referred to the number of classes, m_i was the sample mean for class C_i, $x_j^{(i)}$ was the j^{th} sample from class C_i, and n_i was the number of observations of C_i. Elements along the diagonal of S_W were proportional to the sum of the class sample variances along a feature dimension, across classes. The between-class scatter was defined as:

$$S_B = \sum_{i=1}^{C} \frac{n_i}{n}(m - m_i)(m - m_i)^T$$

where m_i represents the sample mean (or class centroid) for C_i and m was the sample mean for all data. Respectively, m_i and m were given by:

$$m_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^{(i)}$$

$$m = \sum_{i=1}^{C} \frac{n_i}{n} m_i$$

Elements along the diagonal of S_B were proportional to the sum of the distances between class sample means and the global mean for a single feature dimension.

The Fisher criterion was defined as the ratio of the trace of the between-class scatter matrix and the trace of the within-class scatter matrix:

$$FC = \frac{trace(S_B)}{trace(S_W)}$$

The higher the FC value the better a feature set maximized the distances among class means and sample mean.

Selection of a temporal window for each acoustic correlate involved one-dimensional feature vectors and two classes; therefore, the Fisher criterion simplified to (Pekalska, Harol, Lai, & Duin, 2005):

$$FC = \frac{\left\| m_{true} - m_{false} \right\|}{\sqrt{\sigma_{true}^2 + \sigma_{false}^2}}$$

where m_{true} and σ_{true}^2 and m_{false} and σ_{false}^2 were the mean and variances for the correlates for detected and inserted landmarks, respectively. For each correlate, the FC was calculated for each time window and the window that maximized the FC was selected.

The goal of parameter selection was ultimately to minimize the error associated with probability assignment. Minimization of an error-based metric would have required creating probability models for each <correlate, time window> pair. This was computationally too demanding for the given search space. The Fisher criterion, however, only required calculating class means and variances, making it a computationally efficient selection heuristic for time window selection.

6.3 Feature Selection

After selecting temporal parameters for each correlate, the feature space of over 300 correlate representations was still too large to explore all possible correlate combinations[16]. Therefore, forward feature selection was performed to further reduce the search space. Starting with the empty set, informative correlates were incrementally added by again using the Fisher criterion as the selection heuristic[17]. Forward selection was ceased when the feature set reached fifteen correlates. Previous acoustic phonetic research has typically recommended between three and six correlates to describe various acoustic features (Hasegawa-Johnson et al., 2005; Juneja & Espy-Wilson, 2008; Park, 2008). Combined with the fact that only a limited amount of training data was available, fifteen features would be expected to be too large to reliably and generally describe acoustic features in dysarthric speech. Working with a set of correlates larger than desired, however, allowed for possible higher-order dependencies that may have been missed during stepwise selection to be explored. Empirically, fifteen correlates was a manageable number of features on which a global search could be performed, corresponding to over 32,000 (2^{15}) possible correlate groupings.

Identification of the best combination of correlates from the remaining fifteen again required a selection criterion. While the Fisher criterion was used for paring down the feature space, it was not utilized for final correlate selection because the Fisher criterion was a measure of class separability and not a direct measure of the error associated with probability assignment. Instead, a heuristic that relied equally on the average squared error associated with detected landmarks and the average squared error associated with inserted landmarks was used:

[16] Examination of all possible sets would have been preferred to account for possible higher-order dependencies between correlates.

[17] The one-dimensional approximation for the Fisher criterion was not valid for forward feature selection.

$$e(s,l_j,c_k)=\left[\bar{e}^2_{detected}(s,l_j,c_k)+\bar{e}^2_{inserted}(s,l_j,c_k)\right]\ *\mathrm{DL}_{Penalty}$$

where

$$\bar{e}_{detected}(s,l_j,c_k)=\frac{1}{\|D\|}\sum_{d\in D}(1-p_d)^2 \ \ \& \ \ \bar{e}_{inserted}(s,l_j,c_k)=\frac{1}{\|I\|}\sum_{i\in I}p_i^2$$

$$\text{and } \mathrm{DL}_{Penalty}=1+\|c_k\|^2*\log(N)/N$$

D and *I* were the sets of detected and inserted landmarks from the training set for speaker *s*, l_j referred to a single landmark instance, c_k was a specific set of correlates of cardinality $\|c_k\|$, and $\mathrm{DL}_{Penalty}$ was the minimum description length penalty (Rissanen, 1978). Equal sized sets of detected and inserted landmarks from the training set were used with *N* being the total number of landmarks.

Given that deletions could not be recovered, for optimization of candidate landmark extraction the associated error metric was weighted heavily towards minimizing deletion rate. For probability assignment, discrimination of inserted landmarks was deemed equally important as for detected landmarks. The square of the error was used as the basis for the detected and inserted error components to more heavily penalize probabilities farther from the correct values (1 for detections and 0 for insertions).

To calculate the error heuristic, probability models were necessary for each landmark instance. However, given that the training set for each SWD contained just 37 utterances, the amount of data available was extremely sparse. Therefore, probability models were constructed via a speaker adaptation technique known as maximum a *posteriori* (MAP) adaptation. Speaker adaptation, discussed in more detail in the next section, entailed training the recognition models using correlate values measured from another speaker(s) and then tuning the recognition models using only the training data from the target speaker.

For each individual with dysarthria, after paring down to 15 correlates, each possible combination of correlates served as input parameters to Gaussian mixture models. Recordings from the Nemours control speaker (HC) were used to build prior probability models and these models were adapted using recordings from the target speaker's training set. Various other combinations of non-target speaker data were used for creating prior models, including all other SWDs along with the control, all other SWDs without the control, and all individuals of greater intelligibility than the target speaker. In most cases, incorporating speech from other SWDs either had little effect or negatively impacted test set error rates. This observation suggests that incorporating speech from non-target SWDs added extraneous noise into the discrimination models and failed to model the inherent variability of the target speaker.

Following model creation, average error rates were calculated. To avoid overlapping the data used for model adaptation and error estimation, leave-one-out cross-validation (LOOCV) was performed (Duda et al., 2001). Data from 36 of a speaker's 37 training recordings were used for model adaptation followed by calculating the error for each landmark instance on the withheld utterance. This procedure was iterated using each recording once as the held-out utterance. For each landmark instance, the mean of the 37 trials served as the estimate of the error metric. The correlate combinations that minimized the instance-specific error metrics were selected as the optimal sets. Following correlate selection, new recognition models were created using healthy speech for the prior models and the full set of recordings from the target speaker's training set as adaptation tokens. Optimized landmark models were then used to assign probabilities to landmarks from the speaker's test set and type I and II error rates were calculated.

6.3.1 Speaker Adaptation

With sufficient training data, speaker-dependent recognition systems outperform their speaker-independent counterparts, with error rates typically on the order of one-half or one-third (C. H. Lee, Lin, & Juang, 1991). Speaker-specific data, however, is usually too sparse to create true speaker-dependent models. To overcome data limitations, speaker-specific data can be used to tune the parameters of recognition models created using non-target speaker data. Often models from a speaker-independent system are used as prior models. Several approaches to speaker adaptation (SA) have been developed, including maximum *a posteori* (MAP)-based estimation (Gauvain & Lee, 1994; C. H. Lee & Gauvain, 1993), maximum linear regression (MLLR; Leggetter & Woodland, 1995), eigenvoices (Kuhn, Junqua, Nguyen, & Niedzielski, 2000), and Bayesian adaptation (Surendran & Lee, 2001). This work relied on MAP-based adaptation because it provides robust parameter estimates for small amounts of adaptation data, especially when compared to MLLR and eigenvoices (Shinoda, 2005).

Detailed derivations of MAP estimators for Gaussian mixture models are found in Gauvain and Lee (1994). The prior (pre-adaptation) recognition models were created via expectation maximization (EM) using non-target speaker data (EM; 2001; Hastie et al., 2001). The adapted parameters were then estimated via another round of EM. The equations used to update the parameters during the maximization step were:

$$\hat{\omega}_k = \frac{(\nu_k - 1) + \sum_{t=1}^{T} \pi_{kt}}{\sum_{k=1}^{K} (\nu_k - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T} \pi_{kt}}$$

$$\hat{m}_k = \frac{\tau_k \mu_k + \sum_{t=1}^{T} \pi_{kt} x_t}{\tau_k + \sum_{t=1}^{T} \pi_{kt}}$$

$$\hat{\Sigma}_k = \frac{\mu_k + \sum_{t=1}^{T} \pi_{kt} (x_t - m_k)(x_t - m_k)^t + \tau_k (\mu_k - \hat{m}_k)(\mu_k - \hat{m}_k)^t}{\alpha_k - p + \sum_{t=1}^{T} \pi_{kt}}$$

where $\hat{\omega}_k$ was the estimator of the k^th mixture weight in the adapted model, ν_k was the mixture weight of the k^th component in the prior model, π_{kt} was the soft assignment of observed sample x_t to cluster k from the expectation step, $\hat{m}_k$ was the estimator of the mean of the k^th mixture component, τ_k was the learning rate, μ_k was the mean of the k^th component in the prior model, $\hat{\Sigma}_k$ was the covariance estimator of the k^th mixture component, T was the number of adaptation tokens, and p was the cardinality of the correlate set. Adapted parameter estimates were a weighted sum of the prior parameters and parameters estimated from the observed speaker-specific data. With limited adaptation tokens (i.e. small T), the resulting model parameters were essentially the prior parameters. As the number of speaker-specific tokens increased, model performance approached that of the speaker-dependent system (Shinoda, 2005).

6.4 Impact of Correlate Optimization

Selection of speaker-specific correlates decreased average type I and type II error rates for each individual with dysarthria. A comparison of the baseline and optimal rates are plotted in Figure 6.2. Per-speaker error rates for each landmark instance are presented in Table 6.2. Optimal type I error rates for SWDs ranged from 14% to 30%, while type II error rates ranged from 6% to 12%. For four of the ten individuals with dysarthria, type I error rate was roughly less than or equal to the baseline type I error for the Nemours healthy control (18%). Optimal type II error rates for all SWDs were less than the baseline type II rates for both control groups (Nemours HC: 13%, TIMIT: 14%). Comparing each speaker's results following correlate optimization to their baseline results (Figure 6.3), reductions in type I and II error rates ranged from 6% to 33% and from 22% to 64%, respectively[18]. For nine of the ten SWDs type I error was reduced by at least 12% and type II error by at least 40%.

In several cases the set of correlates selected for certain landmark instances resulted in higher error rates than the base model correlates (Tables 6.3 and 6.4)[19]. The most prominent jumps were in type II error rates for *+s* and *–b* landmarks. These two landmark instances were sparse in the Nemours corpus and thus overall average type I and II error rates were not significantly affected. There were several additional reasons why optimization may have resulted in an increase in the type I or II error rate. First, selection of temporal parameters and paring down of possible correlates was performed using a measure of class separability (the Fisher criterion) and not an error-based measure. Second, because correlate selection was performed in a stepwise manner, there was again no guarantee that the global best set of available correlates would be selected. Lastly, since the error function gave equal weight to the squared errors associated with detections and insertions, it was possible that one of the error rates (typically the type I error) increased slightly while the other rate decreased significantly. Thus, while one of the rates increased the sum of the type I and II errors still decreased.

[18] The results presented in Chapter 4 acquired using the base model correlates were used as the results for baseline comparison. Since speaker adaptation was not performed as part of the analysis with the base LAFF model, it might seem inappropriate to compare results from the optimal model to the baseline results. However, as Figure 6.4 reveals, speaker adaptation using the base model correlates had little impact on the baseline results, especially when compared to the impact of correlate optimization.

[19] Average error rates were not recomputed when the base model correlates outperformed the selected correlates for a landmark instance. The results, as presented, provide a more objective measure of the impact of the implemented correlate identification scheme.

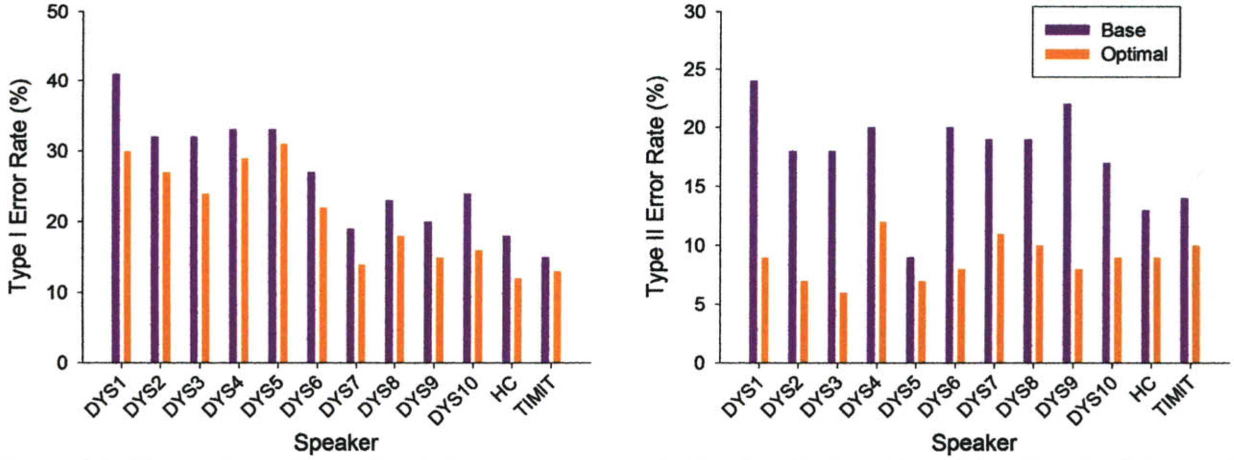

Figure 6.2: Comparison of type I and II error rates acquired using the base (purple) and optimal (orange) correlate sets. For all speakers, including the controls, the optimal error rates were always less than the baseline rates.

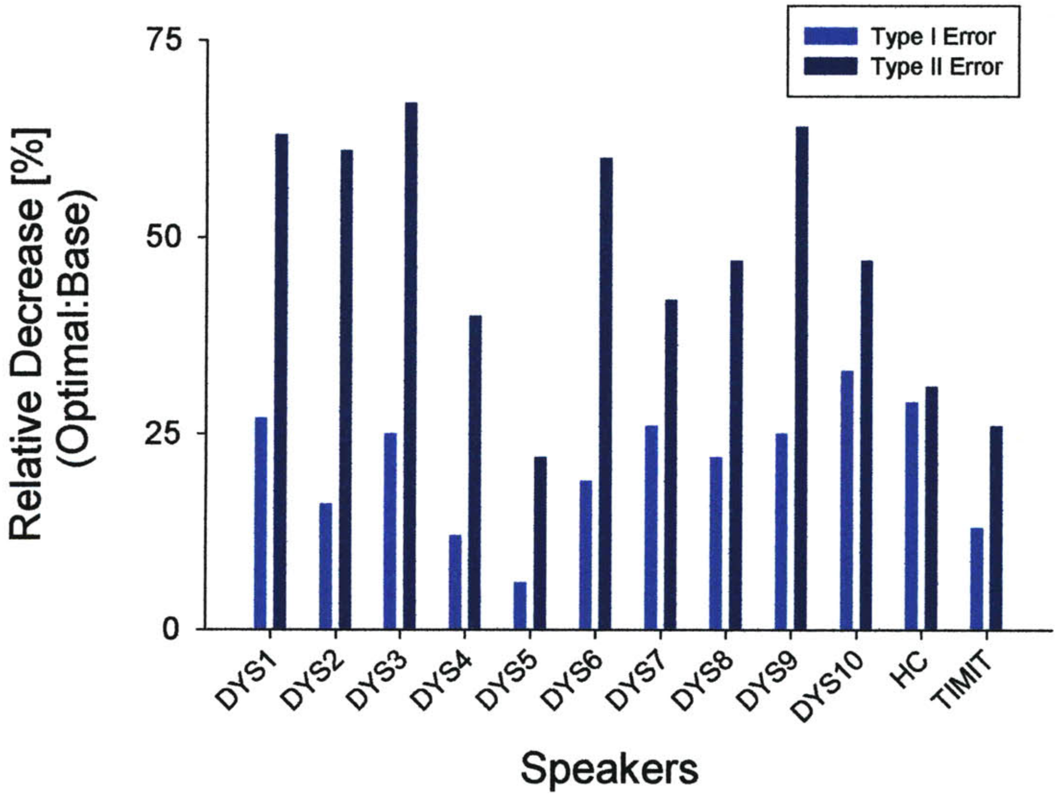

Figure 6.3: Relative decrease (defined as the [Baseline-Optimal]/Baseline*100%) of type I (blue) and type II (dark blue) error rates for each speaker with dysarthria and the healthy control sets (HC and TIMIT).

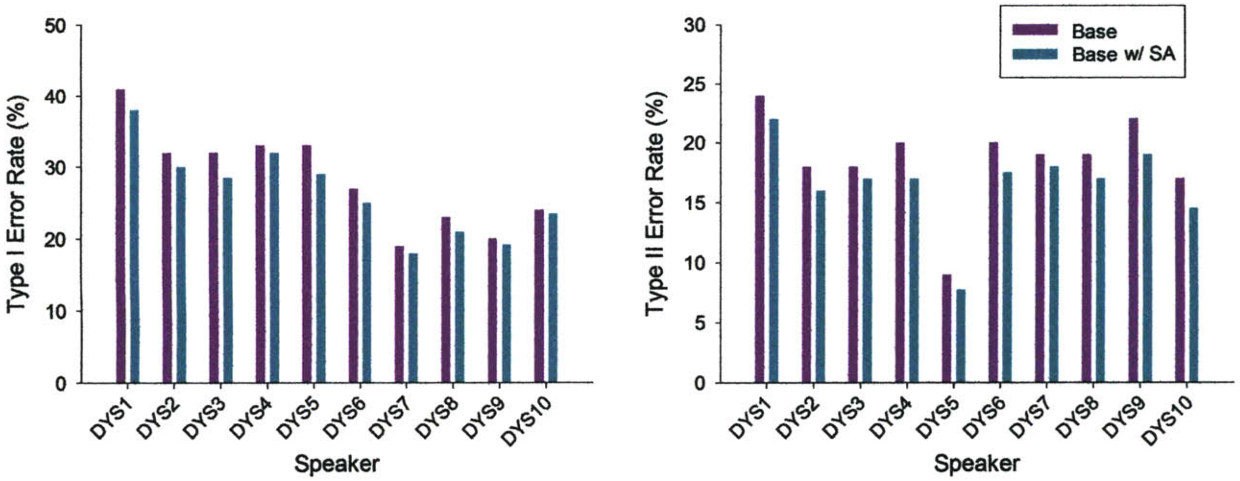

Figure 6.4: Impact of speaker adaptation on type I and II error rates when using the base correlates for discrimination models. For *Base* results, discrimination models were created using the base model correlates measured from the Nemours healthy control. For *Base with Speaker Adaptation (SA)*, prior models were created using base correlates measured from the Nemours healthy speaker. These models were then adapted using target-speaker training data.

Table 6.2: Per-speaker type I and II error rates following correlate optimization and speaker adaptation. Error rates are listed for each landmark instance. The average rates are displayed on the right.

	+g	*-g*	*+s*	*-s*	*+b*	*-b*	Average
Speaker	**Type I (Type II)**	**Type I (Type II)**	**Type I (Type II)**	**Type I (Type II)**	**Type I (Type II)**	**Type I (Type II)**	**Type I (Type II)**
DYS1	19% (9%)	41% (9%)	39% (7%)	31% (12%)	33% (12%)	17% (7%)	30% (9%)
DYS2	15% (6%)	32% (11%)	54% (4%)	43% (9%)	22% (10%)	70% (1%)	27% (7%)
DYS3	12% (11%)	31% (8%)	65% (1%)	32% (4%)	26% (10%)	0% (1%)	24% (6%)
DYS4	17% (12%)	36% (13%)	71% (1%)	33% (27%)	25% (14%)	37% (2%)	29% (12%)
DYS5	7% (5%)	53% (8%)	76% (3%)	43% (13%)	26% (15%)	47% (6%)	31% (7%)
DYS6	10% (10%)	25% (16%)	39% (4%)	25% (7%)	27% (7%)	54% (1%)	22% (8%)
DYS7	11% (14%)	15% (9%)	35% (15%)	30% (4%)	12% (13%)	0% (10%)	14% (11%)
DYS8	6% (9%)	31% (11%)	20% (8%)	21% (3%)	18% (14%)	43% (1%)	18% (10%)
DYS9	6% (7%)	28% (17%)	24% (4%)	19% (5%)	6% (6%)	24% (4%)	15% (8%)
DYS10	8% (10%)	24% (10%)	22% (11%)	24% (11%)	16% (8%)	6% (4%)	16% (9%)
HC	5% (8%)	15% (11%)	22% (6%)	18% (6%)	9% (6%)	18% (4%)	12% (9%)

Table 6.3: Relative decrease ([Baseline-Optimal]/Baseline*100%) in type I error rates for the individuals with dysarthria (DYS) and the healthy control speaker (HC) from the Nemours Database.

Speaker	*+g*	*-g*	*+s*	*-s*	*+b*	*-b*
DYS1	37%	2%	61%	54%	-14%	81%
DYS2	6%	6%	44%	53%	-5%	30%
DYS3	8%	14%	31%	22%	32%	100%
DYS4	-21%	-3%	24%	30%	22%	62%
DYS5	-40%	16%	24%	28%	-18%	49%
DYS6	23%	7%	61%	51%	7%	36%
DYS7	-22%	17%	60%	3%	20%	100%
DYS8	-50%	14%	80%	9%	14%	50%
DYS9	0%	-12%	72%	68%	-50%	76%
DYS10	-33%	17%	75%	44%	11%	94%
HC	25%	27%	54%	37%	35%	78%

Table 6.4: Relative decrease ([Baseline-Optimal]/Baseline*100%) in type II error rates for the individuals with dysarthria (DYS) and the healthy control speaker (HC) from the Nemours Database.

Speaker	*+g*	*-g*	*+s*	*-s*	*+b*	*-b*
DYS1	0%	82%	-250%	-9%	73%	70%
DYS2	14%	67%	-300%	-50%	77%	92%
DYS3	52%	71%	83%	83%	50%	67%
DYS4	50%	64%	80%	-8%	46%	-400%
DYS5	-25%	33%	-200%	13%	56%	-50%
DYS6	29%	60%	0%	59%	79%	75%
DYS7	46%	57%	-275%	85%	62%	-233%
DYS8	55%	62%	-167%	75%	59%	92%
DYS9	59%	63%	0%	74%	82%	0%
DYS10	23%	66%	-450%	45%	70%	33%
HC	22%	29%	69%	70%	48%	-29%

To understand the impact of each component of correlate optimization, type I and II error rates acquired using the base correlates (*Base*) were compared to the error rates achieved with correlate identification but prior to speaker adaptation (*Pre-SA*), and the rates following both correlate selection and speaker adaptation (*Optimal,* Table 6.5). Table 6.6 lists the relative change in error from each step to the next and from the baseline to the optimal results. For the six most severely impaired speakers (DYS1 - DYS6), there was minimal change in type I error rate following correlate selection but prior to speaker adaptation (*Pre-SA:Base*). However, for five of these six individuals (excluding DYS5), there were reductions in type II error rate ranging from 15% to 39%. For speakers DYS7-DYS10, prior to adaptation there were larger reductions in type I error rates but minimal changes in type II rates. Following speaker adaptation, the type I error rate decreased for all speakers except DYS5 for whom the rate remained essentially constant. For speakers DYS1-DYS4 and DYS6 the relative reduction in type I error was greater following speaker adaptation than after correlate selection. For DYS7-DYS10, however, the relative decrease in type I error was similar to or less following adaptation than after correlate identification. One plausible explanation why speakers with more severe dysarthria did not see marked improvements in type I error until after adaptation, whereas individuals with milder forms of impairment saw larger gains following correlate selection, was that the speech from the more intelligible dysarthric speakers more closely resembled typical healthy speech. Further supporting this explanation was that the two largest reductions in type I error following adaptation were observed for speakers with the lowest intelligibilities (DYS1 and DYS3, who had intelligibility scores of 0 and 1, respectively).

Speaker adaptation had a more consistent impact on type II error, with all SWDs showing large reductions ranging from 29% to 62%. Additionally, for nine of the ten individuals with dysarthria the relative change as a result of adaptation was greater than the change following correlate selection. For the remaining individual, DYS2, the change following adaptation, 36%, was similar to the relative decrease following adaptation, 39%. From a production standpoint, these results suggest that the manifestation of superfluous inserted landmarks may be highly speaker-specific in dysarthric speech, even for intelligible SWDs. Currently the salience of inserted landmarks, even in healthy speech, is unclear. Future work focused on the acoustic properties and perceptual impact of insertions would be of value to understanding the significance of these potentially misleading cues on speaker intelligibility.

Table 6.5: Type I and II error rates following each stage of analysis. Baseline rates were calculated using Park's correlate sets (2008). Pre-SA (pre-speaker adaptation) rates were acquired using the correlates identified for the target speaker and with models trained on the Nemours control speaker's training set (for the TIMIT results training was done on the TIMIT training set). Optimal rates were acquired using speaker-specific sets of correlate and adapted models. Speaker adaptation for the TIMIT Database was not performed. For HC the system itself was speaker-dependent because only speech from HC was used.

Speaker	Intelligibility	Type I Error (Baseline)	Type II Error (Baseline)	Type I Error (Pre-SA)	Type II Error (Pre-SA)	Type I Error (Optimal)	Type I Error (Optimal)
DYS1	0	41%	24%	41%	16%	30%	9%
DYS2	1	32%	18%	30%	11%	27%	7%
DYS3	1	32%	18%	31%	13%	24%	6%
DYS4	2	33%	20%	34%	17%	29%	12%
DYS5	3	33%	9%	31%	11%	31%	7%
DYS6	3	27%	20%	27%	16%	22%	8%
DYS7	4	19%	19%	15%	20%	14%	11%
DYS8	4	23%	19%	20%	19%	18%	10%
DYS9	8	20%	22%	17%	21%	15%	8%
DYS10	8	24%	17%	19%	18%	16%	9%
HC		18%	13%	-	-	12%	9%
TIMIT		15%	14%	-	-	13%	10%

Table 6.6: Relative decrease of type I and II error rates following correlate selection but prior to speaker adaptation (Pre-SA:Base) and following speaker adaptation (Optimal:Pre-SA). The two rightmost columns display the overall impact of correlate identification and speaker adaptation (visualized in Figure 6.3). The relative decrease for X:Y was defined as $[Y_{error} - X_{error}]/Y_{error}*100\%$.

Speaker	Type I Error Pre-SA:Base	Type II Error Pre-SA:Base	Type I Error Optimal:Pre-SA	Type II Error Optimal:Pre-SA	Type I Error Optimal:Base	Type II Error Optimal:Base
DYS1	-1%	33%	27%	44%	**27%**	**63%**
DYS2	7%	39%	9%	36%	**16%**	**61%**
DYS3	2%	28%	24%	54%	**25%**	**67%**
DYS4	-2%	15%	14%	29%	**12%**	**40%**
DYS5	7%	-22%	-1%	36%	**6%**	**22%**
DYS6	0%	20%	18%	50%	**19%**	**60%**
DYS7	22%	-5%	5%	45%	**26%**	**42%**
DYS8	11%	0%	12%	47%	**22%**	**47%**
DYS9	15%	5%	12%	62%	**25%**	**64%**
DYS10	21%	-6%	16%	50%	**33%**	**47%**
HC	-	-	29%	31%	**29%**	**31%**
TIMIT	-	-	13%	29%	**13%**	**26%**

6.4.1 Identified Correlates

Optimal sets of correlates were examined to investigate whether the combinations of correlates were unique to each SWD. Tables 6.7 through 6.9 are the sets of correlates for the Nemours healthy control (HC), a mild speaker (DYS10), and a severe speaker (DYS10), respectively (the optimal sets for each SWD in the Nemours Database are found in Appendix C). The associated probability distributions are also shown in Figures 6.6 through 6.9[20].

Starting with the Nemours healthy control (HC), correlates identified for glottis landmarks were measures of periodicity and low frequency energy (sub-400Hz and sub-2000Hz, respectively). The correlate sets for sonorant landmarks included frequency bands typically associated with nasal zeros ([800-2000] Hz) and F2 through F4. Both sets of sonorant correlates also included a measure of periodicity. Given that there was large overlap between the speaker-specific sets of frequency bands used for extraction of sonorant and burst peaks, burst landmarks often resulted in extraction of candidate sonorant peaks and vice-versa. The periodicity measures discriminated between the different landmark types. Both sets of burst correlates for the control speaker contained broad frequency measures, mid-to-high ([3500-5000] Hz) frequency energy, and voicing measures (for the same reason as sonorant landmarks). For each landmark instance for speaker HC, the optimal sets contained between four and six correlates (Table 6.7). The probability distributions post-optimization for HC (Figure 6.6) closely resembled ideal distributions, with most probability mass located around 1 for detections and 0 for insertions.

[20] The histograms for the TIMIT test set are found in Figure 6.5. These distributions, however, will not be discussed because the focus of this work was on the Nemours Database and the individuals with dysarthria

Table 6.7: Speaker-specific correlates identified for the control speaker (HC) from the Nemours Database.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Left mean voicing probability -Right mean voicing -Max ROR [0-400] Hz probability -Right span max [100-400] Hz energy -Difference of mean [300-900] Hz energy	-Left span max voicing probability -Min ROR [0-400] Hz -Max ROR [0-2000] Hz -Difference of mean voicing probability -Difference of mean [300-900] Hz energy	-Difference of mean [800-2000] Hz energy -Left span min [2000-3000] Hz energy -Difference of mean [3500-5000] energy -Difference of mean voicing probability	-Right mean autocorrelation coefficient -Difference of mean [800-2000] Hz energy -Difference of mean [2000-3500] Hz spectral peak (SP) -Min ROR [3500-5000] Hz -Min ROR [0-8000] Hz -Difference of mean [300-900] Hz energy	-Right span max voicing probability -Onset energy [0-8000] Hz -Left mean [0-8000] Hz SP -Difference of span [640-2800] Hz energy -Max ROR [3500-5000] Hz -Left span min [0-400] Hz SP	-Right mean autocorrelation coefficient -Difference of span [100-400] Hz energy -Left mean RMS -Right mean [0-(F3-1000)] Hz energy -Difference of mean [3500-5000] Hz energy -Right span min RMS

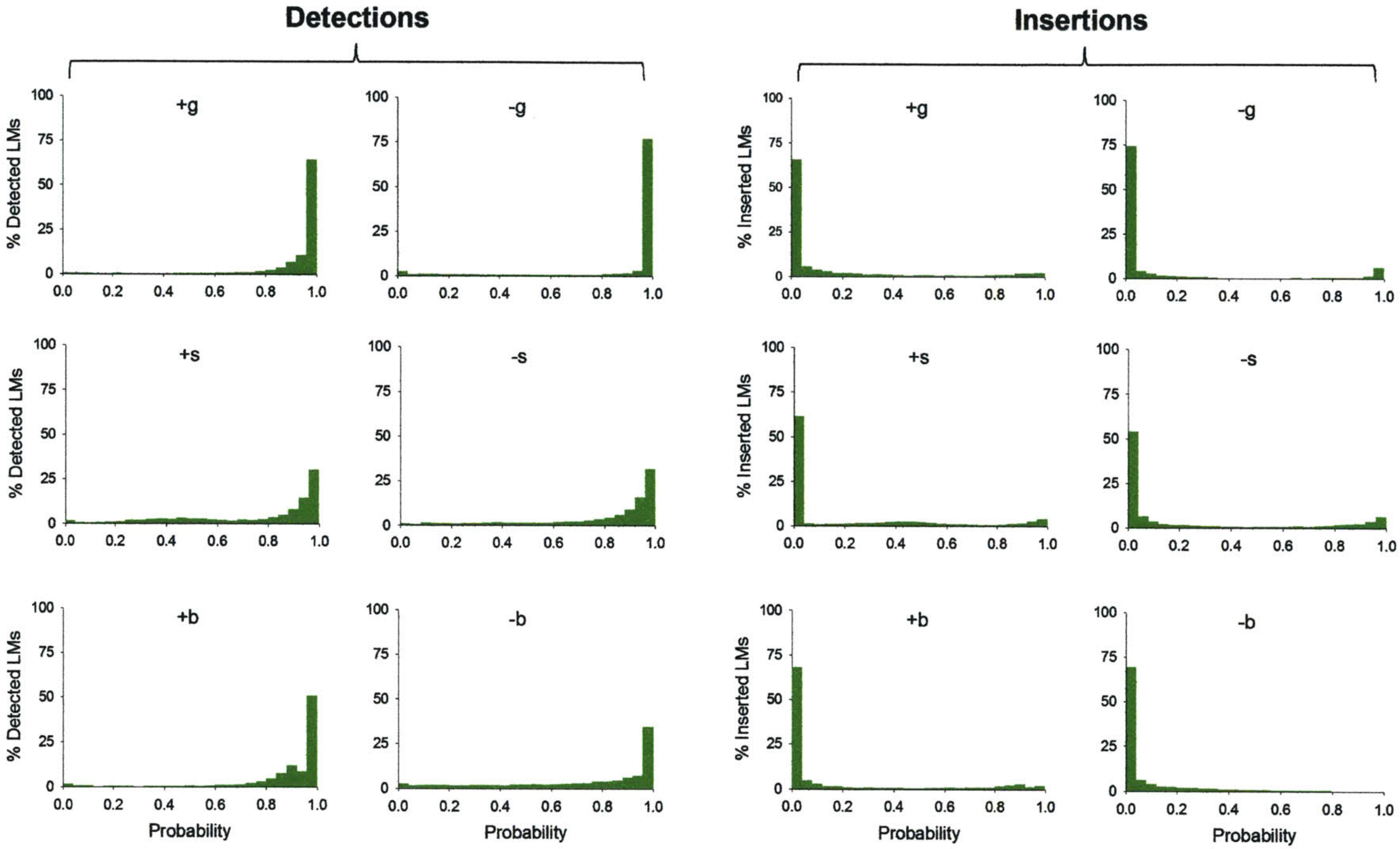

Figure 6.5: Probability distributions for the TIMIT test set acquired using the optimal acoustic correlates. Landmark models were trained using the TIMIT training set.

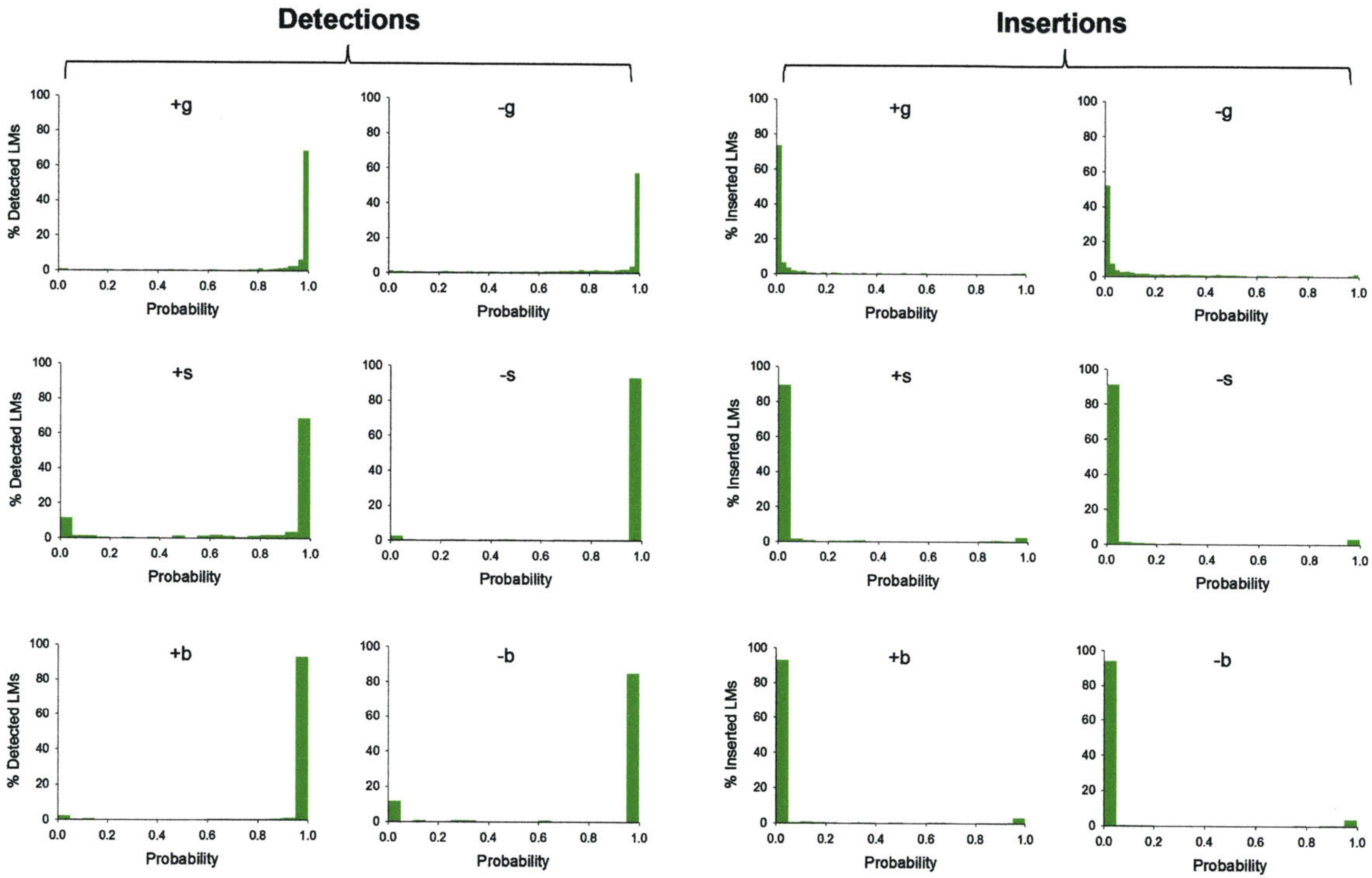

Figure 6.6: Probability distributions for the control speaker (HC) from the Nemours Database acquired using the speaker's optimal acoustic correlates. Landmark models created trained using the training portion of the control speaker's data.

For the correlates identified for the individual with mild impairment (DYS10, Table 6.8), there were many similarities to the healthy control's. For glottis landmarks, correlates were associated with voicing and changes in low frequency energy. Sonorant correlates included voicing measures and energy associated with nasal zeros and F2 through F4. The correlates identified for burst landmarks were measures of changes in broad frequency and mid-to-high frequency energy and voicing probability. On average fewer correlates per landmark instance were identified for DYS10 than for HC. The description length penalty incorporated into the error metric was dependent upon the number of training tokens used, with the penalty for a given number of correlates being higher when fewer exemplars were available. Because of the difference in training set size for each SWD (37 utterances) and the control speaker (370 utterances), larger correlate sets were penalized more heavily for the individuals with dysarthria.

Table 6.8: Speaker-specific correlates identified for an individual with mild dysarthria (DYS10).

+g	*-g*	*+s*	*-s*	*+b*	*-b*
-Max ROR [0-400] Hz -Left span min [100-400] Hz energy -Right mean voicing probability -Glottis onset energy [0-400] Hz -Right mean RMS -Right mean [0-2000] Hz energy	-Min ROR [0-400] Hz -Left span max voicing probability -Difference of span reflection coefficient -Right span min RMS -Left span max [0-400] Hz energy	-Right span max voicing probability -Right span max [2000-3500] Hz energy -Difference of mean energy contained within F3 bandwidth -Onset energy [0-8000] Hz -Difference of span RMS -Max ROR [0-8000] Hz	-Right span min mean [0-400] Hz SP -Difference of mean [640-2800] -Min ROR [5000-8000] Hz -Difference of span [1200-8000] Hz energy	-Max ROR [3500-5000] Hz -Left mean RMS -Left span min voicing probability -Right span max [800-1500] Hz -Onset energy [0-8000] Hz	-Right span min [0-8000] Hz spectral peak (SP) -Left span max [800-2000] Hz -Left span max voicing probability

The probability distributions for DYS10 (Figure 6.7) showed marked improvement compared to the distributions achieved using the base model correlates (Figure 4.9). Except for detected sonorant releases (+*s*), which still were discriminated with 78% sensitivity (compared to 11% originally), each of the distributions was properly skewed to the desired extreme (0 or 1).

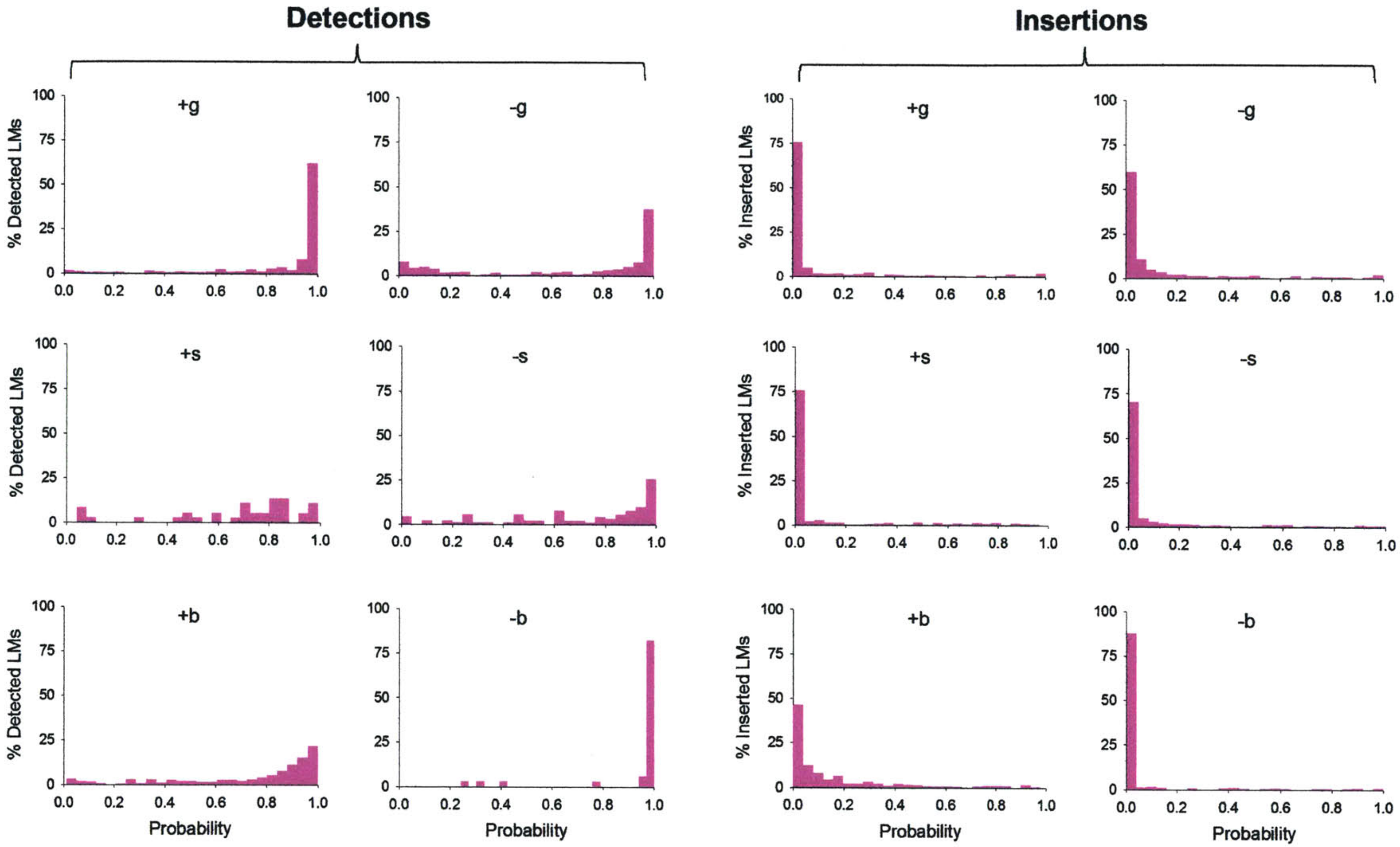

Figure 6.7: Probability distributions for an individual with mild dysarthria (DYS10) acquired using the speaker's optimal acoustic correlates. Landmark models were created using the training portion of the control speaker (HC)'s data and then adapted using DYS10's training set.

There was overlap in the types of correlates selected for the individual with severe dysarthria (DYS1; Table 6.9, Figure 6.9) and the correlates identified for the healthy speaker and the mild speaker. Glottis correlates included voicing and low frequency measures. Selected sonorant correlates pertained to energy in the ranges of F2 through F4. Spectral tilt, one of the base correlates suggested by Park (2008), was also identified for sonorants. Finally, burst landmarks were described by changes in broad frequency energy.

A difference that was observed between the sets of correlates identified for DYS1 and those for DYS10 and HC was the number of correlates selected for each landmark instance. For the severe speaker an average of 6.2 correlates were selected for each instance. 5.3 and 4.8 correlates per landmark instance were selected on average for HC and DYS10, respectively. Generalizing this result, slightly larger sets of correlates were typically identified for speakers with lower intelligibility (Figure 6.8). Larger sets of redundant correlates were discouraged in correlate optimization through the use of the description length penalty. The description length incorporated both the number of correlates and the number of training tokens. Because equal numbers of detections and insertions were used for training, fewer tokens were typically available for severe SWDs due to lower detection rates. For severe speakers the gain in error rate achieved by using more correlates was still sufficient to overcome this penalty. Larger correlate sets imply that more complex models were necessary to discriminate productions from speakers with severe dysarthria, possibly due to increased variability of production. These results also suggest that familiar communication partners may have developed highly intricate models of lexical recognition, with larger sets of overlapping correlates reinforcing recognition of ambiguous landmarks.

Table 6.9:Speaker-specific correlates identified for an individual with severe dysarthria (DYS1).

+g	*-g*	*+s*	*-s*	*+b*	*-b*
-Max ROR [0-400] Hz -Right span max voicing probability -Left span min voicing probability -Difference of mean [300-900] Hz energy -Max ROR [800-1500] Hz -Max ROR [3500-5000] Hz	-Min ROR [0-400] Hz -Left span max [300-900] Hz spectral peak (SP) -Right mean voicing probability -Difference of mean reflection coefficient -Difference of mean energy contained within bandwidth of F1 -Min ROR [0-2000] Hz	-Right mean voicing probability -Right span max [800-2000] Hz energy -Difference of mean [1200-2000] Hz energy -Right mean [2000-3000] Hz energy -Right mean [(F3-1000)-(F3+1700)] Hz energy -Right span max spectral tilt [0-360]:[0-8000] Hz	-Left span max voicing probability -Right mean voicing probability -Difference of span [800-1500] Hz SP -Difference of mean [1200-2000] Hz energy -Difference of mean [2000-3500] Hz energy -Difference span spectral tilt [0-360]:[0-8000] Hz -Offset energy [0-8000] Hz	-Onset energy [0-8000] Hz -Left mean voicing -Difference of span autocorrelation coefficient -Max ROR [800-1500] Hz probability -Right mean [1200-8000] Hz SP -Right of mean ZCR	-Right mean [100-400] Hz energy -Left span max [800-1500] Hz SP -Min ROR [2000-3500] Hz - Difference of mean [3500-5000] Hz SP -Difference of span spectral tilt [0-360]:[0-5000] Hz -Offset energy [0-8000] Hz

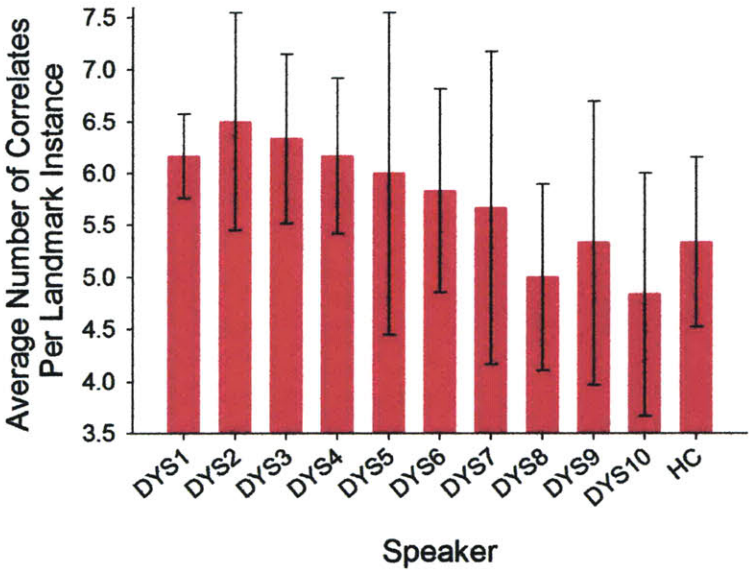

Figure 6.8: Average number of optimal correlates per-landmark instance for each individual from the Nemours Database. Standard deviation is shown.

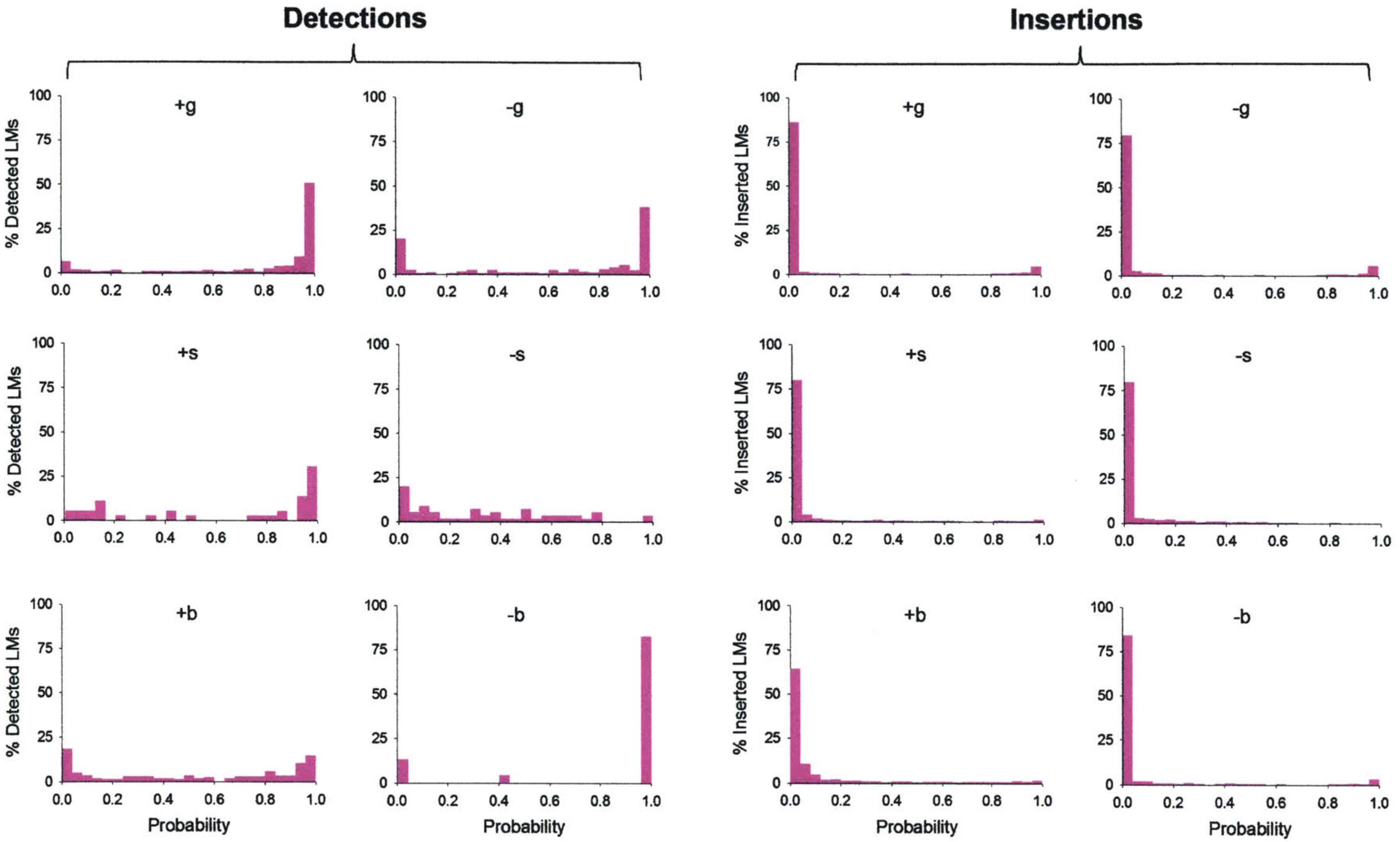

Figure 6.9: Probability distributions for an individual with severe dysarthria (DYS1) acquired using the speaker's optimal acoustic correlates. Landmark models were created using the training portion of the control speaker (HC)'s data and then adapted using DYS1's training set.

6.4.2 Uniqueness of Identified Speaker-Specific Correlates

While there were differences in the number of correlates identified for each SWD, there was also overlap between the types of correlates selected and those suggested in the base model. To determine if the speaker-specific sets of correlates were unique to the speaker or were instead collections of redundant correlates that complemented discrimination (Keyser & Stevens, 2006; Slifka, 2005a; Stevens, Keyser, & Kawasaki, 1986), the optimal correlate combinations for each SWD were applied to the control datasets (TIMIT and HC)[21]. Models were created using the TIMIT and Nemours healthy control (HC) training sets. Applying the trained models to the associated test sets, there were consistent increases in both type I and II error rates over their optimal values (Table 6.10). For HC the relative increase in type I error rate ranged from 12% to 92%. Larger increases in error rates typically occurred when using the correlate combinations from speakers with more severe dysarthria. This effect was more apparent when adding together the increases in type I and II error rate. For the TIMIT test set, larger type I error rates were also observed when using the sets of correlates from less intelligible SWDs.

[21] In terms of the temporal parameters used for measuring correlates, both the time window identified for the associated SWD and the optimal window identified from control speaker data were evaluated. The type I and II error rates reported correspond to the pair with the lowest total error.

Table 6.10: Type I and II error rates for the Nemours control speaker (HC) and the TIMIT test set acquired using the correlates identified as optimal for each individual with dysarthria. In parentheses are the relative increases over the optimal error rates for HC and TIMIT (Table 6.5).

Speaker	HC Type I Error (% Increase)	HC Type II Error (% Increase)	TIMIT Type I Error (% Increase)	TIMIT Type II Error (% Increase)
DYS1	18% (+49%)	15% (+62%)	19% (+47%)	21% (+110%)
DYS2	19% (+57%)	14% (+58%)	18% (+40%)	22% (+117%)
DYS3	16% (+35%)	12% (+29%)	24% (+85%)	14% (+39%)
DYS4	23% (+92%)	13% (+46%)	17% (+32%)	25% (+151%)
DYS5	14% (+18%)	16% (+73%)	14% (+9%)	14% (+40%)
DYS6	18% (+53%)	14% (+52%)	16% (+22%)	21% (+111%)
DYS7	14% (+13%)	15% (+64%)	21% (+59%)	17% (+72%)
DYS8	15% (+26%)	12% (+30%)	20% (+52%)	19% (+91%)
DYS9	20% (+63%)	8% (-9%)	16% (+25%)	23% (+134%)
DYS10	13% (+12%)	13% (+40%)	14% (+8%)	22% (+123%)

6.5 Significance of Correlate Optimization

The results provide evidence that SWDs produce discriminable segmental cues. These cues, however, are manifested differently than those produced by healthy speakers. Correlate optimization was successful in reducing type I and II error rates for all SWDs. Also, using correlates identified as optimal for an individual with dysarthria greatly impaired discrimination of healthy speech. This implies that the correlate pairings suggested for the SWDs were not simply composed of redundant combinations of correlates that complemented discrimination but instead were specific to a speaker.

Larger sets of correlates were deemed as optimal for speakers with more severe forms of impairment. Also, for discrimination of healthy speech, using sets of correlates for speakers with severe dysarthria typically resulted in larger increases in type I and total (type I + type II) error rates compared to when using the correlate combinations selected for individuals with mild dysarthria. These results suggest that discriminable cue patterns in severely dysarthric speech are more complex than in intelligible speech possibly due to increased variability in production.

While results confirm that SWDs produce different discriminable segmental cues than healthy speakers, the magnitude of results must be interpreted conservatively. In addition to identifying differences in production due to impairment, partial reductions in error rates associated with optimization may have also been achieved by leveraging the repetitive sentence structure in the Nemours Database and/or by accounting for natural per-speaker variability. The repetitive sentence structure, while likely leading to lower error rates, was consistent for all speakers. Analysis of a database with broader syntactical structure would remedy this limitation. Speaker adaptation partially models natural per-speaker variability. Therefore, regardless of speaker impairment, lower error rates would be expected following adaptation. Future work performing adaptation for individual healthy speakers would provide a baseline reference for expected between-speaker variation.

Chapter 7

Landmark Sequence Construction

Following extraction of potential peaks and assignment of probabilities, sequences of landmarks can be constructed using higher-level information including language constraints and durational cues. Language information, specifically syntax, would be expected to be the same for both healthy and dysarthric individuals. Segmental duration, however, may vary significantly between SWDs and healthy speakers (Ansel & Kent, 1992; Platt, Andrews, & Howie, 1980; Vijayalakshmi & Reddy, 2006; Ziegler, Hartmann, & Hoole, 1993).

In this chapter, landmark sequences produced by the speaker-optimized and base LAFF models are compared to highlight the combined impact of improved peak detection and optimized correlate selection. A brief introduction to language and duration modeling and to sequence construction is presented prior to the results.

7.1 Language Model

Only a limited number of adjacent landmark pairings are possible according to the articulator movements associated with each landmark. For example, it is not be possible to have a burst release (*+b*) next to a sonorant closure (*–s*) because sonorant landmarks can only occur within a voiced segment. Likewise, two voicing onset landmarks (+*g*) can not be directly adjacent. From the articulatory constraints, bigram models of allowable landmark sequences were constructed. Bigrams were chosen over higher order models such as trigrams because of limited training data and the binary nature of landmarks (onset/offset and release/closure; Park, 2008).

In a bigram model, associated with each allowable landmark pair is the conditional probability that the right-hand landmark would occur given the left-hand landmark. Maximum likelihood estimates (MLE) for bigram probabilities can be calculated by counting the number of occurrences of each landmark instance and each allowable pair (<l_L,l_R>, Jurafsky, Martin, & Kehler, 2000):

$$P(<l_L,l_R>) = P(l_R \mid l_L) = \frac{C(l_L,l_R)}{C(l_L)}$$

A single bigram model was created for all speakers from the Nemours Database by using the TIMIT Database (Table 7.1). The TIMIT corpus spanned broader phonetic and syntactic contexts than the Nemours Database, providing a more robust and general estimate of the bigram probabilities. Incorporating the language model, the probability of transitioning from a base landmark to a target landmark was proportional to the product of the landmark pair's bigram probability and the probabilities that landmarks skipped in the transition were not actually landmarks (Figure 7.1).

Table 7.1: Allowable landmark pairings and associated bigram probabilities derived from the TIMIT Database (Park, 2008).

		RIGHT HAND LANDMARK						
		+g	*-g*	*+b*	*-b*	*+s*	*-s*	*Silence*
LEFT HAND LANDMARK	*+g*		56%			9%	35%	
	-g	34%		45%	15%			6%
	+s		66%			1%	33%	
	-s		44%			56%	1%	
	+b	90%			10%			
	-b	13%		62%				25%
	Silence	40%		60%				

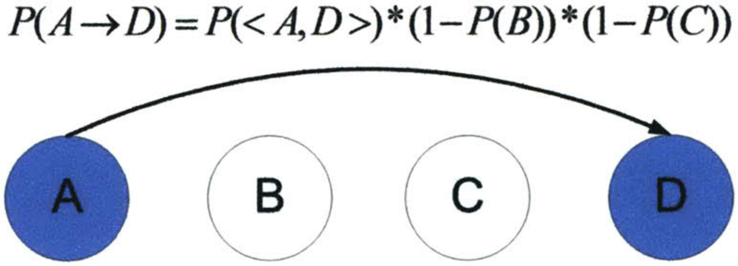

Figure 7.1: The probability of transitioning from a base landmark (A) to a target landmark (D). The probability of skipping a landmark was one minus the probability of the landmark.

7.2 Durational Constraints

Previous studies have noted temporal differences between healthy and dysarthric speech (Ansel & Kent, 1992; Hirose et al., 1982; Kent et al., 1979; Murdoch, 1998; Platt, Andrews, & Howie, 1980; Vijayalakshmi & Reddy, 2006; Yunusova et al., 2008; Ziegler et al., 1993). Figures 7.2 and 7.3 are the distributions of segment durations, calculated using the time-aligned transcriptions, for the sixteen legal landmark pairings from a healthy speaker (HC) and a speaker with severe dysarthria (DYS1), respectively. Maximum segment duration was consistently higher for the speaker with severe dysarthria than the control. The distributions typically showed one or two skewed bell-shaped masses, with a bounded tail on the left. Such properties are consistent with a single or mixture of Rayleigh distributions (Juneja & Espy-Wilson, 2008; Mitchell & Jamieson, 1993). A Rayleigh distribution is defined by a single parameter corresponding to the mode (M) of the function:

$$f(x;M) = \frac{x}{M} \exp(\frac{-x^2}{2M^2})$$

Speaker-specific duration models for each landmark pair were estimated using data from the target speaker's training set. One or two Rayleigh components were used for each duration model, depending upon the shape of the training set distribution. Mixture models were rarely used for SWDs, however, given the limited training data. Also, for several landmark pairings, including <+*s*, -*s*> and <-*b*,+*b*>, the number of examplars was so sparse that a uniform density was assumed for SWDs. Given that segment duration may vary significantly, future work should focus on modeling temporal aspects of dysarthric speech. A first step of analysis could be performing model adaptation similar to that described in Chapter 6 for landmark discrimination.

Duration models did not have a large influence on overall error rates, because transition probabilities already penalized skipping over likely landmarks. Duration models did, however, eliminate very short landmark pairings, specifically inserted burst and sonorant landmarks located within several milliseconds a change in vocal fold vibration. A majority of these burst and sonorant landmarks were due to broad energy changes that occurred as a result of voicing onset/offset and not due to production of an obstruent or sonorant consonant.

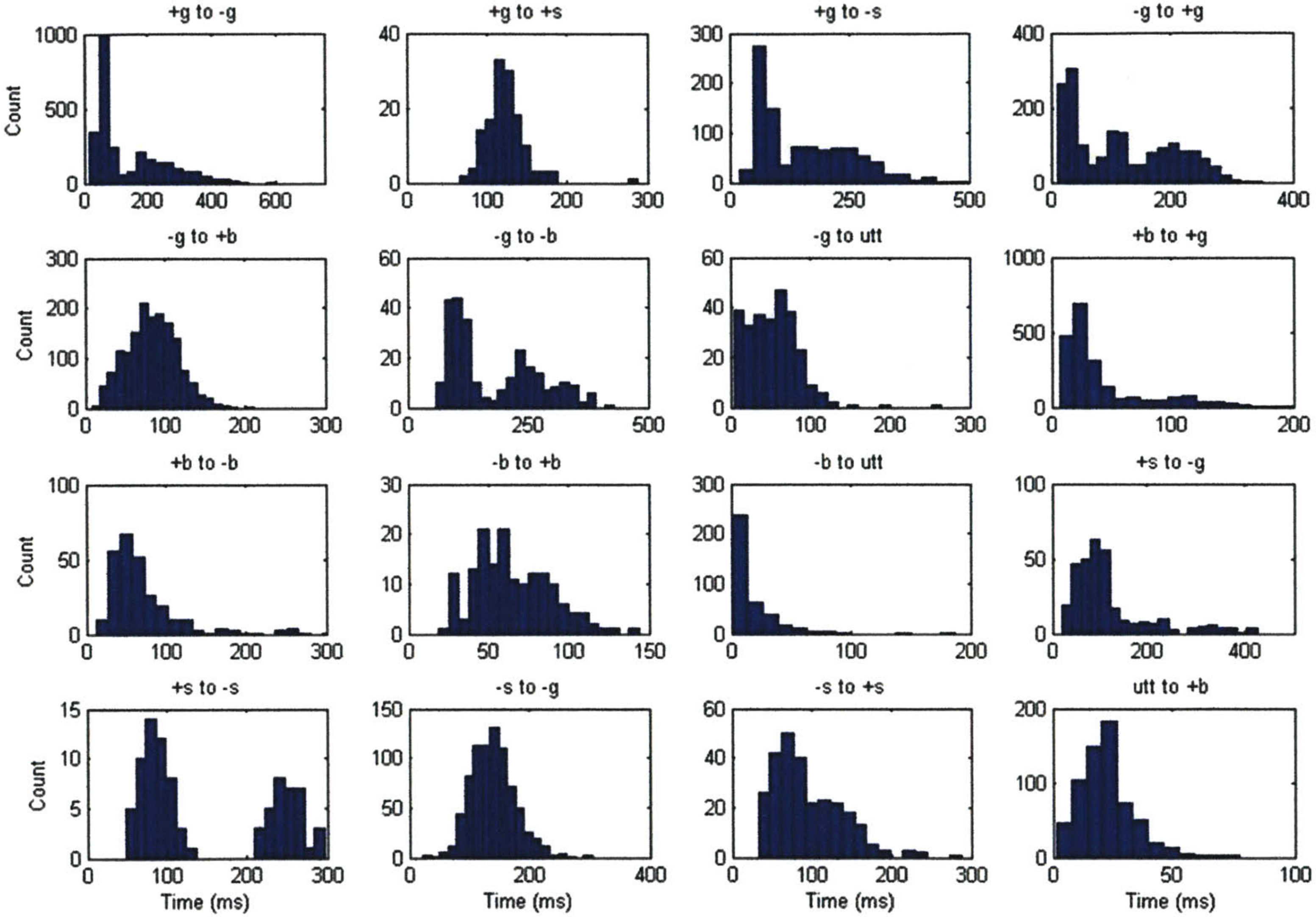

Figure 7.2: Distributions of segment durations, calculated using the time-aligned phonetic transcriptions, for the sixteen legal landmark pairings from the Nemours healthy control (HC).

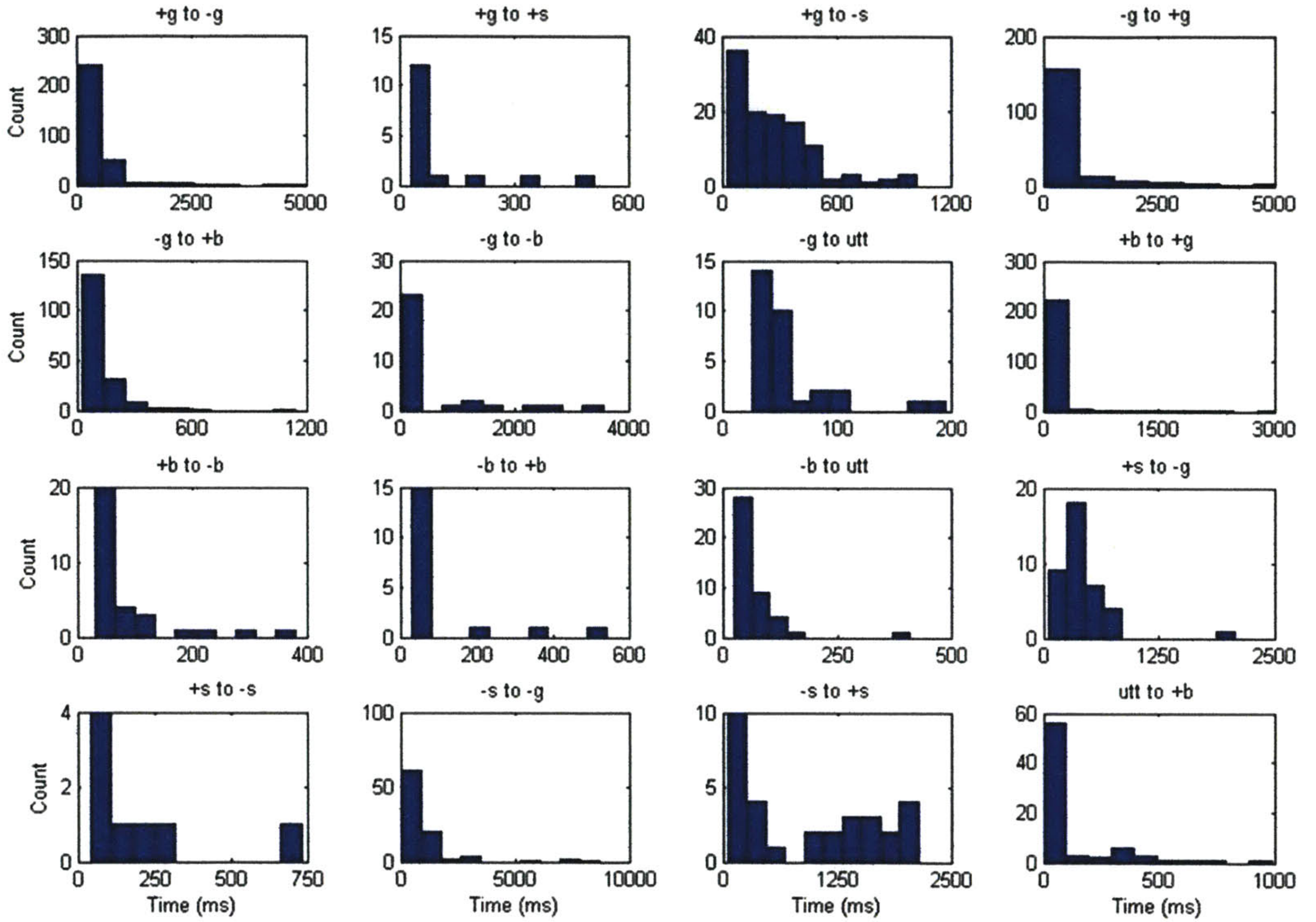

Figure 7.3: Distributions of segment durations, calculated using the time-aligned phonetic transcriptions, for the sixteen legal landmark pairings from an individual with severe dysarthria (DYS1).

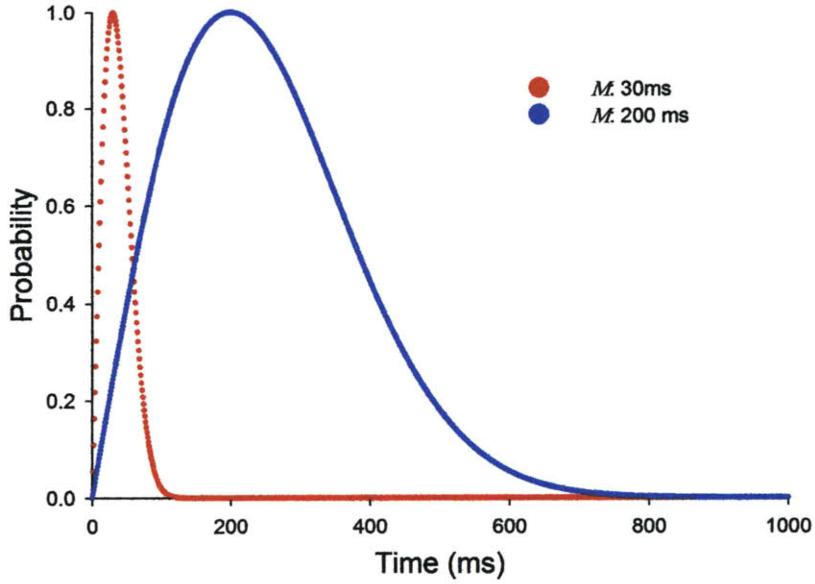

Figure 7.4: Example Rayleigh distributions with different modes.

7.3 Sequence Construction

Sequence construction attempts to find the most probable landmark sequence or multiple likely sequences. In the LAFF model, the likelihood of an individual sequence (*S*) was proportional to product of the bigram transition probabilities (P_B), the segment duration probabilities (P_D), and the individual landmark probabilities (P(*true*) or P(*false*)):

$$P(S) = P_B(S)P_D(S)P_L(S)$$

$$\text{where} \quad \begin{array}{l} P_L(S) = \prod_{s \in S} P(true \mid s) \prod_{\tilde{s} \notin S} P(false \mid \tilde{s}) \\ \text{and } P(false \mid \tilde{s}) = 1 - P(true \mid \tilde{s}) \end{array} .$$

To calculate the most likely sequence(s), various dynamic programming methods have been described including Viterbi alignment (Viterbi, 1967) and A* search (Dechter & Pearl, 1985). An N-best generalization of the Viterbi search algorithm was implemented. Viterbi alignment pinpoints the most likely sequence by finding the best path to each component node. This was generalized to produce multiple sequence hypotheses by keeping a stack of N-best paths to each node (Ström, 1996).

7.4 Results

Landmark detection and error rates using the most probable (N=1) utterance sequence are compared for the base LAFF model and the speaker-optimized models in Figure 7.5. For each SWD, selection of optimal frequency bands and correlates resulted in increased landmark detection rate. Following optimization, detection rates for SWDs ranged from 40% to 77%, compared to 19% to 61% for the base model. In terms of relative improvement, detection rate increased by at least 13% for all SWDs with a maximum relative increase of 118% for the most severe speaker (DYS1, Figure 7.6). Also, the relative change in detection rate for seven of ten SWDs was at least 33% (Figure 7.6). Post-optimization insertion rate was lower for seven of ten SWDs, even though lowering of thresholds substantially increased the number of superfluous candidate peaks. For two of the three remaining SWDs, the relative increase in insertion rate after optimization was less than 20% of the baseline value.

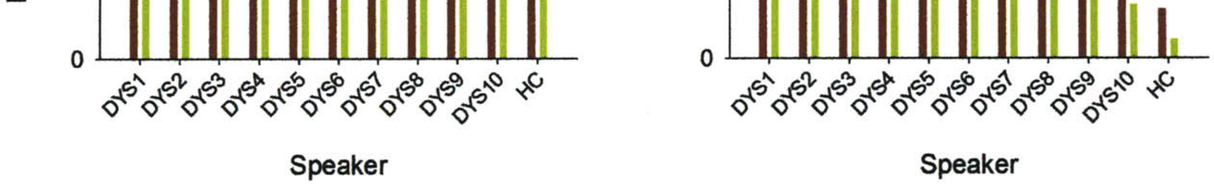

Figure 7.5: Baseline and optimal detection and insertion rates acquired using the most likely (N=1) landmark sequence for each utterance.

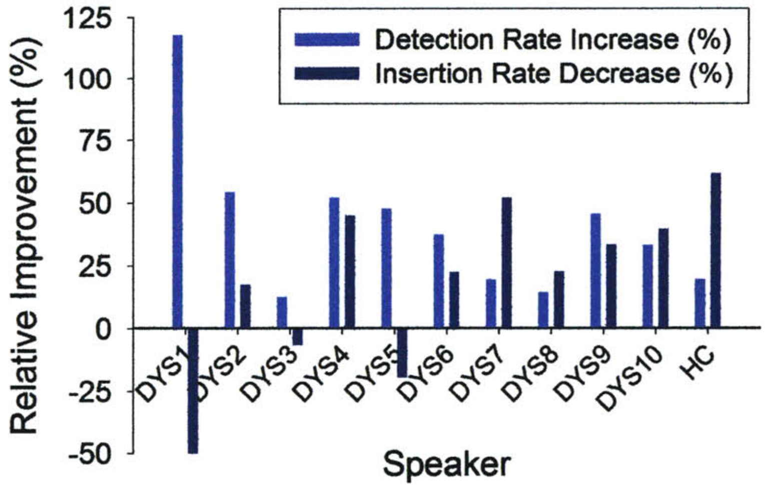

Figure 7.6: Relative increase of detection rate and decrease of insertion rate following per-speaker optimization. Results were acquired using the most likely (N=1) sequence of landmarks for each utterance.

Relying on a single sequence to represent the landmarks in an utterance eliminated many extracted landmarks that were expected according to the phonetic labels. Actual landmarks were not included in the most likely sequence because their individual probabilities were low, expected adjacent landmarks were not present, or nearby superfluous landmarks had high probabilities. An expanded hypothesis of the actual landmarks present was created by merging together multiple landmark sequences. Using the five-best (N=5) search alignments for each utterance resulted in detection rates ranging from 44% to 88% post-optimization, compared to 22% to 64% for the base model (Figure 7.7). The relative increase of detection rate ranged from 13% to 93%, with an increase of at least 25% for seven of ten SWDs (Figure 7.8). Insertion rates for nine of ten SWDs were lower post-optimization compared to baseline rates.

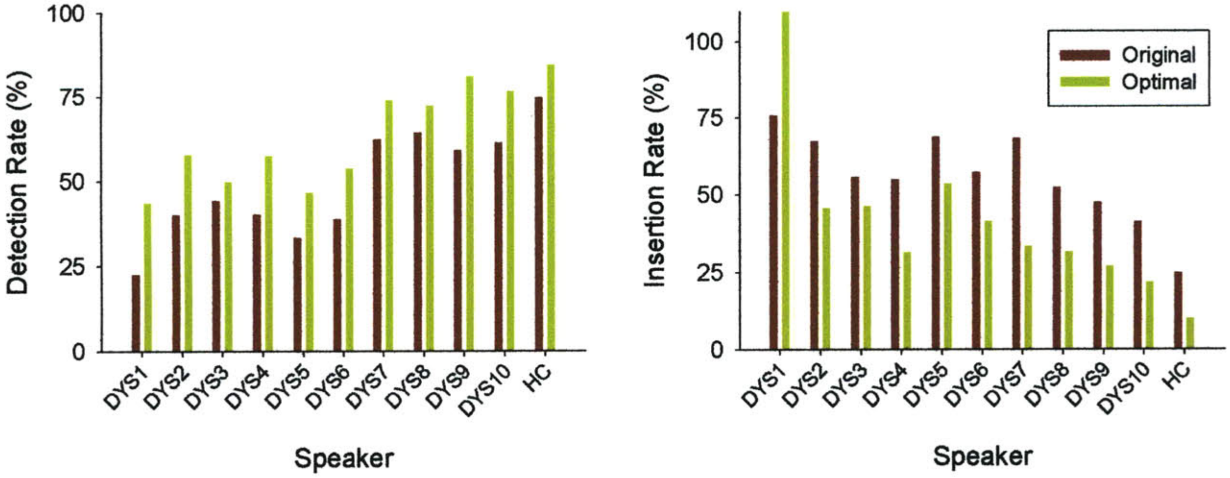

Figure 7.7: Original and optimal detection and insertion rates acquired using the landmarks contained in the five (N=5) most likely sequences for each utterance.

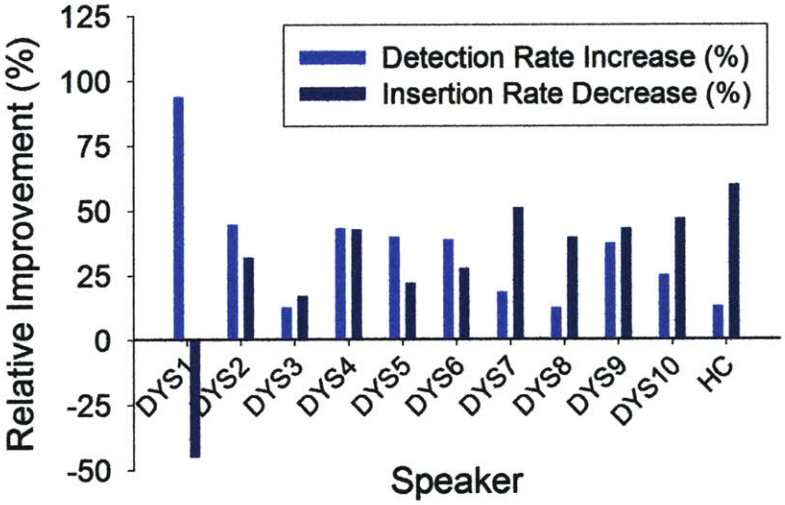

Figure 7.8: Relative increase of detection rate and decrease of insertion rate following per-speaker optimization. Results were acquired by merging the five (N=5) most likely sequences of landmarks for each utterance.

While not a focus of development in this work, construction of landmark sequences emphasizes the overall impact of per-speaker model optimization. Detection thresholds were lowered resulting in increased extraction of expected peaks and also superfluous peaks. Speaker-specific correlates and the resulting recognition models assigned high probabilities to a majority of expected landmarks and low probabilities to a majority of inserted landmarks. As a result, during sequence construction expected landmarks were typically selected while superfluous landmarks were filtered out. For nine of ten SWDs the net effect of optimization was increased detection rate, with insertion rate either decreasing or changing minimally. For the remaining individual with dysarthria (DYS1), while insertion rate increased by almost 50% following optimization, detection rate doubled when using the best or five-best sequences to represent the landmarks present.

Chapter 8

Conclusion

The overall objective of this dissertation was to identify sets of consistently produced acoustic cues by speakers with dysarthria, possibly explaining how familiar listeners are able to decipher seemingly unintelligible utterances. It was hypothesized that the manifestations of reliable acoustic cues in dysarthria would differ from those for healthy speakers, yet the nature and extent of these differences was not understood. The LAFF paradigm (Stevens, 2002), a phonetic feature-based model of human speech recognition, was optimized for ten individual speakers with dysarthria (SWDs). Analysis was performed using only connected speech production because discriminable cue combinations for connected speech may differ from those for individual words (Kent et al., 1997). Optimization of the model involved adjusting thresholds and selecting speaker-dependent frequency bands that improved candidate landmark extraction and identifying speaker-specific acoustic correlates that reliably discriminated between expected and inserted landmarks. While the methodology presented in this work was aimed at discovering residual segmental cues produced by dysarthric speakers within the context of the LAFF model, it is conceivable that the collective process could be generalized for speakers with various impairments as well as for alternative acoustic phonetic frameworks.

The major findings of this work were:

1.) Individuals with dysarthria frequently failed to produce segmental cues resembling those found in healthy speech.

    a. SWDs produced expected landmark candidates less frequently, likely due to an inability to form complete closures in the vocal tract and fully release consonantal constrictions. Also, SWDs inserted superfluous, potentially misleading landmarks at higher rates than healthy speakers. The extent of both behaviors was dependent upon impairment severity (DiCicco & Patel, 2008).

    b. When SWDs produced expected candidate peaks, they often failed to produce correlate manifestations similar to those measured in healthy speech, as evidenced by elevated type I error rates.

2.) SWDs produced candidate landmark peaks less abruptly, requiring lower peak extraction thresholds.

    a. Individuals with severe impairment typically required lower thresholds than mildly impaired speakers, possibly due to further reduction of articulator range of motion and velocity.

    b. Lowering of thresholds resulted in increased detection and insertion rates. Optimal band selection for each speaker partially offset increases in insertion rates following threshold adjustment while having minimal impact on the improved detection rates.

    c. Threshold adjustment and band selection decreased the range over which landmark detection rates varied.

3.) Incorporating redundancy into extraction of candidate glottis landmarks reduced the number of extracted superfluous peaks, while minimally impacting extraction of expected landmarks. This was observed for both SWDs and healthy individuals.

4.) Similar to healthy speakers, individuals with dysarthria signaled abrupt acoustic events using coarse frequency bands spanning a broad range.

   a. For sonorant and burst candidate landmark extraction, there was overlap between the sets of bands selected for each SWD and the base model bands (Liu, 1996).

   b. Identified glottis bands, while different from the single band in the base model, were similar across SWDs and the control speaker.

5.) The manifestations of acoustic correlates that reliably signaled landmark type and validity were unique to individual SWDs.

   a. Discovery of speaker-specific optimal correlates resulted in decreased type I and II error rates (i.e. improved sensitivity and specificity) for target speaker.

   b. Correlates identified for an individual with dysarthria resulted in elevated type I and II error rates (i.e. decreased sensitivity and specificity) when used for discrimination of healthy speech.

6.) Higher-dimensional landmark recognition models were identified for individuals with more severe dysarthria.

   a. The average number of correlates per-landmark instance was dependent upon level of impairment, despite more heavily penalizing larger correlate sets for speakers with more severe dysarthria than for individuals with mild dysarthria or healthy speakers[22].

[22] The penalty was on the number of training tokens. Because individuals with severe dysarthria typically produced fewer expected landmarks there were fewer exemplars available for training.

b. These results suggest that familiar communication partners may have developed highly complex models of lexical recognition where larger sets of correlates are necessary to cope with the inherent variability of dysarthric speech and to possibly reinforce recognition of ambiguous landmarks.

7.) For building prior models for speaker adaptation, using speech from other SWDs rarely provided any benefit and in many cases degraded landmark discrimination. This observation suggests limited overlap in the variability of production between the SWDs in the Nemours Database. It is possible, however, that with a larger set of SWDs adequate similarities between small sets of speakers could be found.

8.) Per-speaker model optimization improved the accuracy of constructed landmark sequences. For nine of ten SWDs, the net effect of model optimization was increased detection rate, with insertion rate either decreasing or changing minimally.

8.1 Limitations and Future Directions

There were several limitations of this work related to the shortcomings of the Nemours Database, a lack of control over variability between speakers, the appropriateness of dysarthria-specific duration models, and extraction of an incomplete set of features within the LAFF framework.

8.1.1 The Nemours Database

The Nemours Database used male speakers with mixed etiologies (cerebral palsy and head trauma) and the dysarthria subtype (spastic, flaccid, unilateral upper motor neuron, ataxic, hyperkinetic, hypokinetic, and mixed; Duffy, 2005) was not documented during assessment. Therefore, it was not possible in this work to investigate the influence of gender or dysarthria subtype on landmark manifestation. The Nemours Database was selected because it was the only publicly available collection of phonetically-broad sentence-level productions from individuals with dysarthria. Analysis of a more controlled database, in terms of subject demographic and dysarthria subtype, would allow for the results of landmark analysis and model optimization to be related to the underlying lesion location and motor disturbance.

An additional limitation of the Nemours Database was that recordings were sometimes noisy because a tabletop free-field microphone was used. Non-speech sounds, often due to speaker movement, were occasionally audible. Use of a head-mounted microphone would improve the speech-to-noise ratio of the recorded waveform.

The Nemours Database contained only ten SWDs and a limited number of recordings from each individual. Results from larger datasets, both in terms of the number of speakers and the amount of speech from each individual, could form the basis upon which to draw stronger, more general conclusions. Therefore, assembling a larger, more tightly-controlled set of recordings from individuals with dysarthria would be of value. Recruiting and recording from SWDs, however, is extremely difficult and time-consuming. The Waisman Center at the University of Wisconsin is in the process of constructing a large database of dysarthric speech but recordings are not currently available outside of the center (per communication with R. D. Kent).

8.1.2 Sources of Per-Speaker Variability

While results confirm that SWDs produce different discriminable segmental cues than healthy speakers, the magnitude of results must be interpreted within the scope of the work. In addition to identifying differences in production due to impairment, partial reductions in error rates associated with optimization may have also been achieved by leveraging the repetitive sentence structure in the Nemours Database and/or by accounting for natural per-speaker variability. The repetitive sentence structure, while likely leading to lower error rates, was consistent for all speakers. Analysis of a database with broader syntactical structure would remedy this limitation. Speaker adaptation partially models natural per-speaker variability. Therefore, regardless of speaker impairment, lower error rates would be expected following adaptation. Future work performing adaptation for individual healthy speakers would provide a baseline reference for expected between-speaker variation.

8.1.3 Appropriateness of Dysarthria-Specific Duration Models

Segment duration is typically longer and more variable in dysarthric speech (Ansel & Kent, 1992; Kent et al., 1979; Platt, Andrews, & Howie, 1980; Vijayalakshmi & Reddy, 2006; Ziegler et al., 1993). Therefore, temporal models currently used for healthy speech may not prove adequate or may require significant alteration. Given sparse training data for each SWD and the closed syntactical structure of the Nemours Database, building reliable per-speaker duration models for all possible landmark pairings proved challenging. Therefore, future work focused on modeling temporal aspects of dysarthric speech over broad phonetic and syntactical contexts should be performed.

8.1.4 Incomplete State of the LAFF model

In its current state, the LAFF model only extracts landmarks and acoustic correlates related to the articulator-free features of *consonantal*, *sonorant*, and *continuant*[23]. Additional advancements of the LAFF model that would benefit all speakers include elucidation of the lower-level articulator-free feature *strident*, incorporating vowel (Howitt, 2000b; Slifka, 2005b) and glide landmarks to further constrain allowable landmark sequences, and development of

[23] The value for the feature continuant can be inferred in most situations using adjacent landmark pairs.

methods for automatic extraction of articulator-free features. Sets of correlates for determining several articulator-free features using the spectral-temporal information around landmarks have been suggested (Bitar, 1998; Bitar & Espy-Wilson, 1996; Hasegawa-Johnson et al., 2005). In the present work, it was demonstrated that extraction of correlates related to manner features required per-speaker optimization. Given that errors of place of articulation are more common than errors in manner of articulation in dysarthric speech (Ansel & Kent, 1992; Juneja & Espy-Wilson, 2008; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980), it would therefore be expected that extraction of articulator-free features would also require per-speaker identification.

8.1.5 Perceptual Analysis

While not related to addressing limitations, several potential perceptual studies are suggested by this work. First, studies aimed at understanding the impact of insertions on intelligibility, both in healthy and dysarthric speech, would benefit our understanding of lexical recognition. Second, this work demonstrated that distinct sets of acoustic correlates can be identified that improve landmark discrimination by a machine listener. Studies investigating the perceptual importance of speaker-specific correlates could potentially extend these findings to human listeners. Given that many of the correlates identified for SWDs were energy-based measures, one possible experimental paradigm would be to distort frequencies related to speaker-specific correlates and to examine the impact on familiar listener recognition. Understanding the mechanism by which familiar communication partners recognize dysarthric speech would contribute both to our understanding of the dysarthric speech stream and also to our general understanding of human speech recognition.

8.1.6 Voice-Driven Augmentative and Alternative Communication (AAC)

Results from per-speaker optimization suggest that additional work investigating the use of feature-based automatic speech recognition (ASR) for SWDs is warranted. Customized ASR could enable individuals with dysarthria to interact more richly and naturally within their surroundings, regardless of listener familiarity. However, previous attempts to use standard data-driven recognition models, specifically hidden Markov models (HMMs), have proven inadequate for individuals with severe impairment, likely due to insufficient training data (Blaney & Wilson, 2000; Bowes, 1999; Deller et al., 1991; Doyle et al., 1997; Ferrier, Shane, Ballard, Carpenter, & Benoit, 1995; Kotler & Thomas-Stonell, 1997; Rosengren, Raghavendra, & Hunnicutt, 1995; Thomas-Stonell, Kotler, Leeper, & Doyle, 1998).

Phonetic feature-based recognition relies on context-independent units, potentially minimizing the need for large amounts of training data. Significant gains in candidate landmark detection and discrimination achieved via optimization of the LAFF model suggest that additional work related to feature-based ASR for SWDs be performed. It should be noted, however, that the LAFF model may not be the most appropriate acoustic phonetic framework upon which to perform automatic recognition of dysarthric speech. The LAFF paradigm was specifically used for this work because it is a theory of human speech recognition and the assumptions and constraints of the model enabled testing of the dissertation's underlying hypothesis.

Espy-Wilson's Event Based System (EBS; Espy-Wilson et al., 2007; Juneja & Espy-Wilson, 2008) is a feature-based ASR platform that takes a different approach to feature-based recognition. Eliminating extraction of candidate landmark peaks as a precursory step to measuring acoustic correlates, correlates are instead measured on a frame-by-frame basis.

Measured correlates serve as input parameters to a hierarchy of classifiers which assign to each frame the likelihoods that it belongs to each of five broad phonetic classes. Per-frame probabilities are used by an alignment algorithm to partition an utterance into likely sequences of broad classes. From the broad classes, landmarks are then estimated.

Automatic extraction of segmental peaks proved difficult for speakers with severe dysarthria. Even following band selection and threshold adjustment, detection rates were reduced compared to healthy speakers and individuals with mild dysarthria. Therefore, EBS's approach to ASR could potentially benefit SWDs, given that it places less emphasis on regions of abrupt spectral change. Applying the finding that SWDs use distinct sets of correlates to convey changes in manner, it may be possible to identify correlates that convey the per-frame broad class, instead of the landmark type. Results from a trial analysis of the Nemours Database using the base implementation of EBS, developed for and trained on healthy speech, are found in Appendix D. It should be noted that for EBS (or any other feature-based model) to be a viable platform for ASR, both manner and place features must be identified with high accuracy. Given the high frequency of errors in place of articulation in dysarthria (Ansel & Kent, 1992; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980), adequate extraction of place features could prove prohibitive given limited training data.

8.2 Implications: Assessment and Monitoring of Treatment Efficacy

This work illustrates the potential of using the LAFF framework as a tool for understanding dysarthric speech production and perception, with findings having implications for the assessment of impairment and monitoring of progress during speech therapy. Given that phonetic features relate articulation and acoustics, features have the capacity to identify causes for intelligibility deficits resulting from disruptions to a multitude of the speech production subsystems (Ansel & Kent, 1992; Kent, Weismer, Kent, & Rosenbek, 1989; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980). Dysarthria is currently assessed using perceptual (i.e. clinician-dependent) measures, such as the Assessment of Intelligibility of Dysarthric Speech (AIDS; Yorkston & Beukelman, 1981), that qualitatively note articulatory deficiencies on a broad level and do not quantify their extent.

Using the base LAFF framework as a model of a typical, unfamiliar listener, landmarks represent acoustic targets for normal speech production and perception (Liu, 1996; Park, 2008; Stevens, 2002). Reduced landmark detection rates indicate a failure to properly produce closures or releases in the glottis, vocal tract, or nasal port. Insertions potentially indicate sporadic or irregular voicing (glottis and sonorant landmarks), velopharyngeal inadequacy (sonorants), and unexpected frication or aspiration (bursts). Type I error rates provide a measure of articulatory precision, indicating how often observable expected landmarks are manifested in a form similar to landmarks in typical speech. Finally, measured segmental durations can be compared to typical healthy values to evaluate coordination of articulator movements and to screen for temporal irregularities possibly impairing speaker intelligibility.

Unlike current perceptual measures, landmark analysis provides a quantitative set of objective measures that are capable of localizing intelligibility deficits to production of specific manner class(es). For example, using a standard articulation test, a clinician (depending upon his/her ability) may be able to note that a speaker has difficulty producing obstruent bursts. This perceptually-based observation fails to identify if the speaker is having difficulty forming a complete closure behind which to build up pressure in the vocal tract or is failing to fully release the closure due to insufficient articulatory displacement. Landmark analysis, on the other hand, can not only identify the impaired component of production (closure/release) but can also quantify the degree of impairment via the associated landmark deletion rate. Compared to current assessment protocols, another unique aspect of landmark-based analysis is that it can identify inserted acoustic events that potentially confound listener perception.

Using landmark analysis as a tool for describing and understanding dysarthric speech requires phonetically-transcribed speech from which reference landmark sequences can be generated. Thus, methods that rapidly label dysarthric speech with minimal user input are required for landmark analysis to be viable in clinical settings. While current ASR systems fail to achieve high accuracies for speakers with severe dysarthria (Blaney & Wilson, 2000; Bowes, 1999; Deller et al., 1991; Doyle et al., 1997; Ferrier et al., 1995; Kotler & Thomas-Stonell, 1997; Rosengren et al., 1995; Thomas-Stonell et al., 1998), using standardized reading passages would limit possible alignments. Constraining the phonetic sequence should improve the quality of suggested alignments, reducing the amount of time necessary for manual inspection and correction.

Tracking of progress via landmark analysis could be performed in a similar manner as assessment, monitoring changes in error rates for different manner classes using the base LAFF model. Improved detection rates would indicate that a speaker was producing more perceivable segmental boundaries. Decreased insertion rates would denote that a speaker was creating fewer extraneous landmarks that may be degrading his/her intelligibility. Finally, reduced type I error rates would suggest that as a result of the prescribed speech therapy, articulatory precision had improved.

By providing a more detailed description of an individual's speech than current assessment protocols, landmark analysis has the potential to standardize and quantify assessment of dysarthria while also providing a measure of treatment efficacy during speech intervention.

Further, by highlighting the fine-grain components of speech production that require intervention, landmark analysis may enable clinicians to personalize therapies, potentially minimizing treatment time and yielding more immediate improvement in intelligibility.

Appendix A. List of Tokens for the Nemours Database

Nouns		Verbs
back	faith	bearing
bad	fake	chewing
badge	fat	daring
bag	fate	going
bait	fay	heaping
bake	fife	knowing
base	fight	leaping
bash	fin	licking
bass	fine	lifting
bat	five	listing
bat	gin	living
batch	goo	mowing
bath	inn	owing
bathe	Jew	pairing
beet	knew	reaping
beige	kong	searching
Bert	lot	serving
bet	mat	shooing
bin	moo	singing
bit	pat	sinning
bite	phase	sipping
boat	pin	sitting
boot	rot	sleeping
butt	shin	stewing
chew	shoe	suing
chin	sin	surfing
con	sue	surging
coo	thin	sweeping
dew	tin	tearing
dial	two	wading
die	vat	waiting
dime	watt	waking
din	who	waning
dive	yacht	waving
face	zoo	wearing
fade		weeping
		weighing

Appendix B.

Time-Aligned Phonetic Labeling

To transcribe recordings from the control speaker in the Nemours Database, an automatic phonetic labeler similar to that described by Kominek, Bennett, & Black (2003) was implemented. This scheme makes use of an HMM-based forced alignment recognizer paired alongside a template (dynamic time warping, DTW)-based recognizer to suggest two separate phonetic alignments. *Ad hoc* statistical measures comparing the two transcriptions highlight alignments likely requiring manual inspection. To label productions from the control speaker of the Nemours Database, the TIMIT Database was used to create individual phoneme models for the HMM-based recognizer while utterances synthesized with CMU's Festvox (Black & Lenzo) served as templates for the DTW-based aligner. Annotation of utterances from the Nemours healthy control were created by time-averaging the phonetic boundaries suggested by the two labelers. Manual inspection and correction of alignments was then performed for all utterances. Figure B.1 is a diagram illustrating the automatic labeling process.

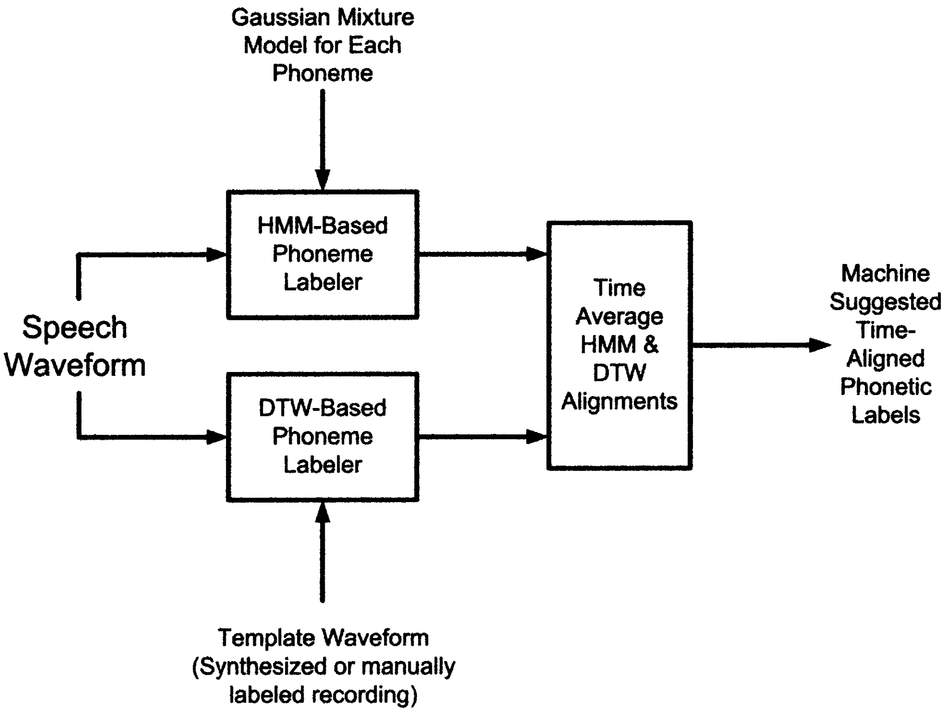
Gaussian Mixture Model for Each Phoneme
HMM-Based Phoneme Labeler
Speech Waveform
DTW-Based Phoneme Labeler
Time Average HMM & DTW Alignments
Machine Suggested Time-Aligned Phonetic Labels
Template Waveform (Synthesized or manually labeled recording)

Figure B.1: Automatic labeling of an utterance using a mixture of HMM and DTW-based recognizers. This procedure was adapted from Kominek et al. (2003).

Appendix C.

Optimal Correlates Identified for Each Speaker from the Nemours Database

Table C.1: Speaker-specific correlates identified for speaker DYS1.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Right span max voicing probability -Left span min voicing probability -Difference of mean [300-900] Hz energy -Max ROR [800-1500] Hz -Max ROR [3500-5000] Hz	-Min ROR [0-400] Hz -Left span max [300-900] Hz spectral peak (SP) -Right mean voicing probability -Difference mean reflection coefficient -Difference of mean energy contained within bandwidth of F1 -Min ROR [0-2000] Hz	-Right mean voicing probability -Right span max [800-2000] Hz energy -Difference of mean [1200-2000] Hz energy -Right mean [2000-3000] Hz energy -Right mean [(F3-1000)-(F3+1700)] Hz energy -Right span max spectral tilt [0-360]:[0-8000] Hz	-Left span max voicing probability -Right mean voicing probability -Difference span [800-1500] Hz SP -Difference mean [1200-2000] Hz energy -Difference of mean [2000-3500] Hz energy -Difference span spectral tilt [0-360]:[0-8000] Hz -Offset energy [0-8000] Hz	-Onset energy [0-8000] Hz -Left mean voicing -Difference span autocorrelation coefficient -Max ROR [800-1500] Hz probability -Right mean [1200-8000] Hz SP -Right of mean ZCR	-Right mean [100-400] Hz energy -Left span max [800-1500] Hz SP -Min ROR [2000-3500] Hz - Difference of mean [3500-5000] Hz SP -Difference span spectral tilt [0-360]:[0-5000] Hz -Offset energy [0-8000] Hz

Table C.2: Speaker-specific correlates identified for speaker DYS2.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Max ROR [800-1500] Hz -Left span min [100-400] Hz energy -Right span max F0 amplitude -Right span max [640-2800] Hz energy -Right span max voicing probability -Left span min voicing probability	-Min ROR [0-400] Hz -Left span max voicing probability -Difference of span reflection coefficient -Right span min RMS -Left span max [0-400] Hz energy -Onset energy [0-8000] Hz	-Max ROR [5000-8000] Hz -Right mean spectral tilt [0-360]:[0-5000] Hz -Right mean [2000-3000] Hz energy -Mean F1 amplitude across peak -Onset energy [0-8000] Hz -Right span max voicing probability	-Right span min mean [0-400] Hz SP -Difference of mean [640-2800] -Min ROR [5000-8000] Hz -Difference of span [1200-8000] Hz energy -Right mean [0-400] Hz SP -Right span min [0-8000] Hz energy -Left mean voicing probability	-Max ROR [1200-8000] Hz -Right span max [0-400] Hz SP -Difference span voicing probability -Left span min F0 -Left span min [0-8000] Hz SP -Right mean ZCR -Left mean [0-8000] Hz energy -Left span min [300-900] Hz energy	-Right span min [0-8000] Hz SP -Left span max [800-2000] Hz -Left span max voicing probability -Right mean [0-400] Hz SP -Offset energy [0-8000] Hz

Table C.3: Speaker-specific correlates identified for speaker DYS3.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Max ROR [800-1500] Hz -Right span max voicing probability -Left mean autocorrelation coefficient -Left mean RMS -Glottal Offset energy -Right span max [640-2800] Hz energy	-Min ROR [0-400] Hz -Left span max voicing probability -Difference of span reflection coefficient -Right span min RMS -Left span max [0-400] Hz energy -Onset energy [0-8000] Hz -Left span max [300-900] Hz energy	-Right span max voicing probability -Right mean spectral tilt [0-360]:[0-8000] Hz -Right span max [1200-8000] Hz SP -Onset energy [0-8000] Hz -Difference span [800-2000] Hz energy	-Right span min [800-2000] Hz energy -Offset energy [0-8000] Hz -Right mean [0-400] Hz SP -Difference of mean [640-2800] -Difference of span [1200-8000] Hz energy -Left mean voicing probability	-Max ROR [1200-2000] Hz -Right span max [0-(F3$_{avg}$ -1000)] Hz energy -Right mean [640-2800] Hz energy -Right mean [2000-3500] Hz SP -Left mean [0-8000] Hz energy -Max ROR [1200-8000] Hz	-Min ROR [0-8000] Hz -Left span max [1200-8000] Hz SP -Difference span [5000-8000] Hz SP -Left mean [5000-8000] Hz SP -Offset energy [0-8000] Hz -Mean F2 amplitude across peak -Difference mean [0-8000] Hz energy

Table C.4: Speaker-specific correlates identified for speaker DYS4.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Difference mean autocorrelation coefficient -Left mean RMS -Right span max voicing probability -Difference mean [300-900] Hz energy -Right span max [640-2800] Hz energy	-Min ROR [0-400] Hz -Difference span [2000-3000] Hz energy -Difference mean [800-1500] Hz SP -Left span max voicing probability -Right span min RMS	-Difference mean RMS -Left mean F0 amplitude -Difference span [100-400] Hz energy -Left mean ZCR -Right span max [2000-3000] Hz energy -Difference span [3500-5000] Hz SP -Left span min [5000-8000] Hz energy	-Min ROR [0-400] Hz -Min ROR [1200-2000] Hz -Right mean RMS -Difference mean [($F3_{avg}$-1000)-($F3_{avg}$+1700)] Hz energy -Left span max [2000-3000] Hz energy -Right span min [1200-8000] Hz SP -Offset energy [0-8000] Hz	-Left mean [0-8000] Hz energy -Max ROR [0-8000] Hz -Max ROR [5000-8000] Hz -Right mean reflection coefficient -Right span max spectral tilt [0-360]:[0-8000] Hz -Right mean [0-8000] Hz energy	-Right span min spectral tilt [0-360]:[0-5000] Hz -Left mean [0-8000] Hz energy -Left span max [1200-8000] Hz SP -Right mean [0-360] Hz energy -Offset energy [0-8000] Hz -Difference span [5000-8000] Hz SP

Table C.5: Speaker-specific correlates identified for speaker DYS5.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Right span max voicing probability -Left span min [100-400] Hz energy -Difference mean [0-F1] Hz energy -Difference mean [300-900] Hz energy -Difference mean autocorrelation coefficient -Right span max [640-2800] Hz energy	-Difference mean voicing probability -Difference mean reflection coefficient -Left span max [100-400] Hz energy -Right span min [300-900] Hz energy -Offset energy [0-8000] Hz -Min ROR [0-400] Hz -Difference mean [800-1500] Hz SP	-Difference span [0-F1] Hz energy -Right mean spectral tilt [0-360]:[0-5000] Hz -Left span min [1200-2000] Hz SP -Left span min [2000-3500] Hz energy -Mean F4 amplitude across peak	-Left mean voicing probability -Right span min [100-400] Hz energy -Min ROR [1200-2000] Hz -Difference mean [1200-8000] Hz energy	-Max ROR [0-8000] Hz -Max ROR [2000-3500] Hz -Left mean [F1-8000] Hz energy -Right span max spectral tilt [0-360]:[0-8000] Hz -Difference mean [3500-5000] Hz SP -Right mean [0-8000] Hz energy -Difference mean F1 amplitude -Right mean voicing probability	-Right mean [100-400] Hz energy -Offset energy [0-8000] Hz -Difference span [5000-8000] Hz SP -Left mean [0-8000] Hz energy -Right span min spectral tilt [0-360]:[0-5000] Hz

Table C.6: Speaker-specific correlates identified for speaker DYS6.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Right span max F0 amplitude -Left mean spectral tilt [0-360]:[0-5000] Hz -Mean F1 amplitude across peak -Glottal onset energy -Left mean voicing probability	-Min ROR [0-400] Hz -Left mean RMS -Left span max voicing probability -Mean F2 amplitude across peak -Glottal offset energy -Right mean [250-650] Hz energy -Difference span voicing probability	-Right span max voicing probability -Left span min RMS -Difference mean F2 amplitude -Difference mean F4 amplitude -Difference span [($F3_{avg}$-1000)-($F3_{avg}$+1700)] Hz energy -Difference span spectral tilt [0-360]:[0-5000] Hz	-Right span min RMS -Difference mean F4 amplitude -Difference span [0-F1] Hz energy -Left span max spectral tilt [0-360]:[0-8000] Hz -Right span min [0-8000] Hz energy -Offset energy [0-8000] Hz	-Max ROR [1200-8000] Hz -Onset energy [0-8000] Hz -Left mean spectral tilt [0-360]:[0-5000] Hz -Left span min voicing probability	-Right span min F0 amplitude -Right span min spectral tilt [0-360]:[0-5000] Hz -Left mean voicing probability -Difference mean [0-360] Hz energy -Offset energy [0-8000] Hz -Min ROR [3500-5000] Hz

Table C.7: Speaker-specific correlates identified for speaker DYS7.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Right span max voicing probability -Left mean RMS -Difference mean F1 amplitude -Right mean [0-400] Hz SP	-Min ROR [0-400] Hz -Left mean [800-1500] Hz SP -Left span max voicing probability -Glottal offset energy -Right span min [0-8000] Hz energy	-Right span max voicing probability -Difference span voicing probability -Left span min spectral tilt [0-360]:[0-8000] Hz -Difference span [1200-2000] Hz energy -Right mean [5000-8000] Hz energy -Right span max [0-360] Hz energy -Max ROR [1200-8000] Hz	-Right mean autocorrelation coefficient -Right span min [3500-5000] Hz SP -Offset energy [0-8000] Hz -Difference mean F3 amplitude	-Max ROR [0-8000] Hz -Max ROR [5000-8000] Hz -Difference span [0-400] Hz SP -Right span max autocorrelation coefficient -Left span min RMS	-Min ROR [2000-3500] Hz -Right mean F0 amplitude -Difference mean spectral tilt [0-360]:[0-5000] Hz -Left span max [1200-2000] Hz SP -Difference mean [3500-5000] Hz SP -Left span max [1200-2000] Hz energy -Left span max [300-900] Hz energy -Difference span [1200-8000] Hz SP

Table C.8: Speaker-specific correlates identified for speaker DYS8.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Right span max [0-400] Hz SP -Right span max [0-F1] Hz energy -Max ROR [0-400] Hz -Difference span voicing probability	-Left span max V.P. -Left span max [0-F1] Hz energy -Min ROR [0-400] Hz -Right span min autocorrelation coefficient -Onset energy [0-8000] Hz	-Difference mean F2 amplitude -Right mean [5000-8000] Hz SP -Right mean [1200-2000] Hz energy -Max ROR [3500-5000] Hz	-Right span min V.P. -Offset energy [0-8000] Hz -Difference mean [2000-3500] Hz energy -Right span min [5000-8000] Hz SP -Right span min [3500-5000] Hz energy -Right span min [0-8000] Hz energy	-Max ROR [5000-8000] Hz -Left span min V.P. -Left mean [0-($F3_{avg}$ - 1000)] Hz energy -Right mean [2000-3000] Hz energy -Max ROR [0-400] Hz -Onset energy [0-8000] Hz	-Left span max RMS -Difference mean [3500-5000] Hz SP -Difference span [5000-8000] Hz SP -Difference mean [0-8000] Hz energy -Right mean [800-1500] Hz energy

Table C.9: Speaker-specific correlates identified for speaker DYS9.

+*g*	-*g*	+*s*	-*s*	+*b*	-*b*
-Max ROR [0-400] Hz -Max ROR [0-2000] Hz -Difference mean autocorrelation coefficient -Right span max V.P. -Difference mean [0-360] Hz energy -Glottal onset energy -Right span max voicing probability -Difference mean [640-2800] Hz energy	-Left mean [0-400] Hz SP -Left mean V.P. -Min ROR [0-400] Hz -Right span min RMS -Difference span [300-900] Hz energy	-Max ROR [1200-2000] Hz -Max ROR [3500-5000] Hz -Right span max [0-400] Hz SP -Right mean RMS -Mean F1 amplitude across peak	-Right span min [0-400] Hz SP -Left mean RMS -Difference span [($F3_{avg}$-1000)-($F3_{avg}$+1700)] Hz energy -Right mean [2000-3500] Hz energy	-Max ROR [2000-3500] Hz -Left mean voicing probability -Difference span ZCR -Onset energy [0-8000] Hz -Difference span [3500-5000] Hz energy	-Right mean autocorrelation coefficient -Left span max F0 -Right mean spectral tilt [0-360]:[0-5000] Hz -Left mean [0-8000] Hz energy -Offset energy [0-8000] Hz

Table C.10: Speaker-specific correlates identified for speaker DYS10.

+g	*-g*	*+s*	*-s*	*+b*	*-b*
-Max ROR [0-400] Hz -Left span min [100-400] Hz energy -Right mean voicing probability -Glottis onset energy [0-400] Hz -Right mean RMS -Right mean [0-2000] Hz energy	-Min ROR [0-400] Hz -Left span max voicing probability -Difference of span reflection coefficient -Right span min RMS -Left span max [0-400] Hz energy	-Right span max voicing probability -Right span max [2000-3500] Hz energy -Difference of mean energy contained within F3 bandwidth -Onset energy [0-8000] Hz -Difference of span RMS -Max ROR [0-8000] Hz	-Right span min mean [0-400] Hz SP -Difference of mean [640-2800] -Min ROR [5000-8000] Hz -Difference of span [1200-8000] Hz energy	-Max ROR [3500-5000] Hz -Left mean RMS -Left span min voicing probability -Right span max [800-1500] Hz -Onset energy [0-8000] Hz	-Right span min [0-8000] Hz SP -Left span max [800-2000] Hz -Left span max voicing probability

Table C.11: Speaker-specific correlates identified for the control speaker (HC) from the Nemours Database.

+g	*-g*	*+s*	*-s*	*+b*	*-b*
-Left mean voicing probability -Right mean voicing -Max ROR [0-400] Hz probability -Right span max [100-400] Hz energy -Difference of mean [300-900] Hz energy	-Left span max voicing probability -Min ROR [0-400] Hz -Max ROR [0-2000] Hz -Difference of mean voicing probability -Difference of mean [300-900] Hz energy	-Difference mean [800-2000] Hz energy -Left span min [2000-3000] Hz energy -Difference of mean [3500-5000] energy -Difference of mean voicing probability	-Right mean autocorrelation coefficient -Difference of mean [800-2000] Hz energy -Difference of mean [2000-3500] Hz SP -Min ROR [3500-5000] Hz -Min ROR [0-8000] Hz -Difference of mean [300-900] Hz energy	-Right span max voicing probability -Onset energy [0-8000] Hz -Left mean [0-8000] Hz SP -Difference of span [640-2800] Hz energy -Max ROR [3500-5000] Hz -Left span min [0-400] Hz SP	-Right mean autocorrelation coefficient -Difference span [100-400] Hz energy -Left mean RMS -Right mean [0-(F3-1000)] Hz energy -Difference of mean [3500-5000] Hz energy -Right span min RMS

Appendix D.

EBS - An Alternative Acoustic Phonetic Framework

Espy-Wilson's Event Based System (EBS; Espy-Wilson et al., 2007; Juneja & Espy-Wilson, 2008) takes a different approach to feature-based recognition than the LAFF model, eliminating extraction of landmarks as a first step to measuring acoustic correlates. Instead, acoustic correlates are computed on a frame-by-frame basis. These correlates serve as input parameters to a hierarchy of classifiers which assign to each frame the likelihoods that it belongs to each of five broad phonetic classes (vowel, stop, sonorant consonant, fricative, and silence). These likelihoods are used by a probabilistic segmentation algorithm to partition an utterance into the most likely sequence(s) of broad classes. Landmarks are then suggested by the model using the broad class sequence.

Extraction of landmarks proved difficult for speakers with severe dysarthria. Even following band selection and threshold adjustment, detection rates were decreased and insertion rates increased as speaker intelligibility degraded. The LAFF model was appropriate for testing the hypothesis that individuals with dysarthria convey acoustic landmarks differently than healthy individuals given that it is a model of human speech recognition. However, EBS deserves attention in future work as a potential alternative for dysarthric ASR because it removes the need to detect candidate landmark peaks as a precursory step.

D.1 EBS Overview

EBS relies on a slightly different collection and organization of manner features than LAFF. These features are *sonorant*, *syllabic*, and *continuant.* The articulatory correlates and broad articulation classes for each of these manner features are summarized in Table D.1. The hierarchical organization of the features is shown in Figure D.1. [+*sonorant*] indicates a lack of constriction or a constriction insufficient to produce turbulence. Vowels and sonorants have the distinctive feature of [+*sonorant*] while obstruent consonants are [–*sonorant*]. [+*syllabic*] signals an open vocal tract and distinguishes vowels from sonorant consonants. Lastly, fricatives are characterized by the feature [+*continuant*], while stop consonants are [–*continuant*].

Table D.1: Summary of the manner (articulatory-free) features incorporated into the EBS model (Juneja & Espy-Wilson, 2008). For each feature, the articulatory correlate and broad phonetic class, when relevant, are provided.

Phonetic feature	Articulatory correlate	Vowels	Sonorant consonants	Fricatives	Stops
sonorant	No constriction, insufficient constriction to result in turbulence	+	+	-	-
syllabic	No pressure build up in vocal tract	+	-		
continuant	Incomplete constriction			+	-

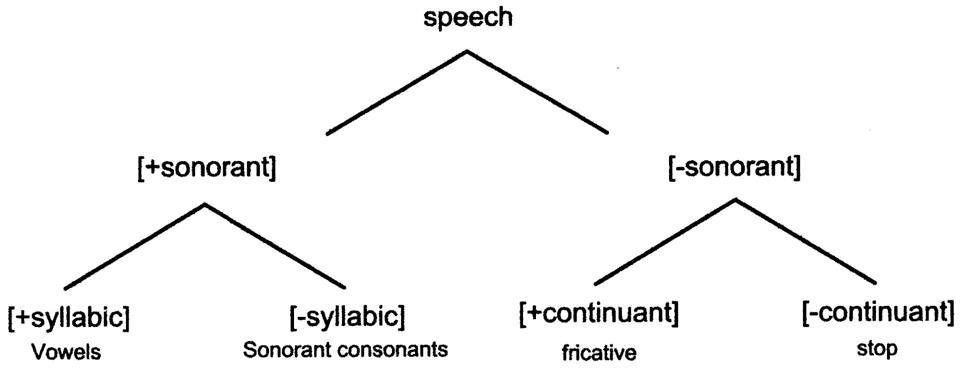

Figure D.1: Organization of manner (articulator-free) features in EBS (Juneja & Espy-Wilson, 2003, 2008).

Unlike LAFF which places emphasis on detecting acoustic landmarks, EBS tracks a variety of acoustic correlates to perform broad class segmentation. Tracking is performed on a frame-by-frame basis at a 5 ms frame rate. Subsets of acoustic correlates are used as inputs to a chain of classifiers corresponding to the manner hierarchy shown in Figure D.2. The subset of acoustic correlates for each classifier is manner-dependent. For a single frame, the outputs of the classifiers correspond to the probabilities that the frame belongs to each of the five broad classes (vowel, stop, sonorant consonant, fricative, or silence).

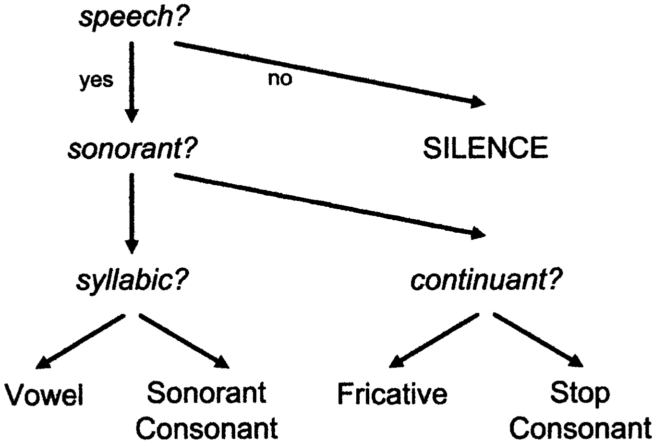

Figure D.2: Hierarchy of manner classifiers. The outputs of the classifiers are combined to yield the posterior probabilities of a speech segment being in each of the broad phonetic classes (Juneja & Espy-Wilson, 2008)

Stating the broad class segmentation task formally, let $O = \{o_1, o_2, ..., o_T\}$ represent the sequence of acoustic correlates, where o_T is the vector of parameters at time t. The most probable sequence of broad classes, $\hat{B}$, and their durations, $\hat{D}$, are given by the relationship:

$$\hat{B}\hat{D} = \arg\max_{BD} P(BD \mid O)$$

The posterior probability of a frame being in one of the five broad class at time t is calculated by traversing the manner hierarchy to the appropriate broad class node. For example, the posterior of frame t being part of a sonorant consonant (SC) is:

$$P_t(SC \mid O) = P_t(speech, sonorant, syllabic' \mid O)$$

$$= P_t(speech \mid O) P_t(sonorant \mid speech, O) P_t(syllabic' \mid sonorant, O)$$

As only subsets of acoustic correlates are needed for each feature, the above equation can be rewritten as:

$$P_t(SC \mid O) = P_t(speech \mid x_t^{speech}) P_t(sonorant \mid speech, x_t^{sonorant}) P_t(syllabic' \mid sonorant, x_t^{syllabic})$$

Representing broad class B_i with the set of relevant features $\{f_1^i, f_2^i, ... f_{N_{B_i}}^i\}$, the broad class at time t as b_t, and the sequence of broad classes $\{b_1, b_2, ... b_{t-1}\}$ as b^{t-1}, allows for the probability $P(BD \mid O)$ to be expanded in terms of the manner features of each broad class:

$$P(BD \mid O) = \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i + \sum_{j=1}^{i-1} D_j} P_t(B_i \mid O, b^{t-1})$$

$$= \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i + \sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i \mid x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i, b^{t-1})$$

where $\left\{1+\sum_{j=1}^{i-1} D_j,...,D_i+\sum_{j=1}^{i-1} D_j,\right\}$ are the indices of the frames that occupy B_i.

Assuming $x_t^{f_k^i}$ is independent of b^{t-1} and given $\left\{f_1^i,...,f_k^i\right\}$, the expression for $P(BD\,|\,O)$ can be rewritten as:

$$P(BD\,|\,O)=P(B)P(D\,|\,B)\prod_{i=1}^{M}\prod_{t=1+\sum_{j=1}^{i-1}D_j}^{D_i+\sum_{j=1}^{i-1}D_j}\prod_{k=1}^{N_{B_i}}\frac{\mathrm{P}_t(f_k^i\,|\,x_t^{f_k^i},f_1^i,...,f_{k-1}^i)}{\mathrm{P}_t(f_k^i\,|\,f_1^i,...,f_{k-1}^i)}$$

where $P(B_i)$ is the bigram prior probability of broad class B_i, and $P(D_i\,|\,B_i)$ is the probability of duration D_i given broad class B_i. $P(D_i\,|\,B_i)$ can be calculated using Rayleigh distributions based on mode duration for the broad classes. The posterior probabilities, $\mathrm{P}_t(f_k^i\,|\,x_t^{f_k^i},f_1^i,...,f_{k-1}^i)$, come from the outputs of the support vector machine-based classifiers using a technique known as binning (Drish, 2001; Zadrozny & Elkan, 2001). $\mathrm{P}_t(f_k^i\,|\,f_1^i,...,f_{k-1}^i)$ normalizes imbalances between training set sizes.

To simplify computation, it is assumed that[24]:

$$P(B\,|\,O)\approx\max_{D}P(BD\,|\,O)$$

Working with this assumption a quasi-Viterbi alignment algorithm is used to calculate the N-most probable segmentations using the manner feature probabilities. The segmentation algorithm essentially behaves as a smoothing filter, eliminating short sporadic changes in broad class.

[24] The invariance assumption is similar to that used for Viterbi decoding in HMM-based and segmentation-based speech recognition systems to simplify computation (Glass, Chang, & McCandless, 1996; Juneja & Espy-Wilson, 2008; S. Lee & Glass, 1998).

A summary of the acoustic correlates is found in Table D.2 (Bitar & Espy-Wilson, 1996; Espy-Wilson & Bitar, 1997; Juneja & Espy-Wilson, 2008). Four classifiers, one for each of the phonetic features (*silence*, *sonorant*, *syllabic*, and *continuant*), are trained using positive and negative examples of each feature. For the *silence* classifier, frames from speech serve as the negative class while frames extracted from periods of silence serve as positive training tokens. The relevant set of acoustic correlates from the three previous and two following frames are included with the target frame to form the input vector. For the *sonorant* classifier, frames from sonorant consonants and vowels are trained against frames from fricatives and stops. Four frames prior and one after the target frame are also included in the input vector. For the *syllabic* classifier, frames from vowels (including the 16 frames before and the 24 frames following) are trained against comparable sets of frames from sonorant consonants. Lastly, for the *continuant* classifier, fricative frames (including 4 prior and 4 after) are trained against frames from stop bursts.

Table D.2: Summary of the acoustic correlates used for each manner feature classifier (Bitar & Espy-Wilson, 1996; Espy-Wilson & Bitar, 1997; Juneja & Espy-Wilson, 2008). Per-frame classification accuracies are from analysis of the TIMIT Database (Juneja & Espy-Wilson, 2008).

Phonetic Feature	Acoustic Description	Acoustic Correlates	Per-frame Classification Accuracy (%)
silence		(1) E[0,F3-1000], (2) E[F3,f_s/2], (3) ratio of spectral peak in [0,400Hz] to spectral peak in [400,f_s/2]	93.5
sonorant	Periodic; Strong low frequency energy	(1) Voicing probability (Talkin, 1995) (2) Ratio of E[0,F3-1000] to E[F3-1000,f_s/2], (3) E[100,400], (4) ZCR, (5) Zero-crossing rate of high-pass filtered speech, (6) ratio of spectral peak in [0,400Hz] to spectral peak in [400,f_s/2]	94.4
syllabic	Strong mid frequency energy	(1) E[640,2800], (2) E[2000,3000] (both energies normalized by nearest syllabic dips and peaks)	81.7
continuant	Closure followed by abrupt change in spectrum	(1) Sum of first-difference values across STFT channels, (2) First autocorrelation coefficient normalized by the zeroth coefficient, (3) E[0,F3-1000], (4) E[F3-1000,f_s/2]	95.6

Following frame classification, the alignment algorithm is used to perform broad class segmentation. In previous experiments with the TIMIT Database, 86.7% of broad classes were detected with a 7.2% insertion rate (Juneja & Espy-Wilson, 2008). From the broad class segmentations landmarks can inferred using the mapping from broad classes to landmarks found in Table 2.3.

D.2 Results

Analysis of recordings from the Nemours Database was performed using the base EBS model. Manner classes were described by the correlates suggested by Juneja and Espy-Wilson

(2008) and classifiers were trained on the TIMIT training set. Results resembled those acquired using the base LAFF model trained on healthy speech. Landmark deletion and insertion rates were higher for individuals with more severe forms of impairment (Figure D.3). The correlations between intelligibility and detection, deletion, and insertion rates were all significant ($\alpha<0.05$).

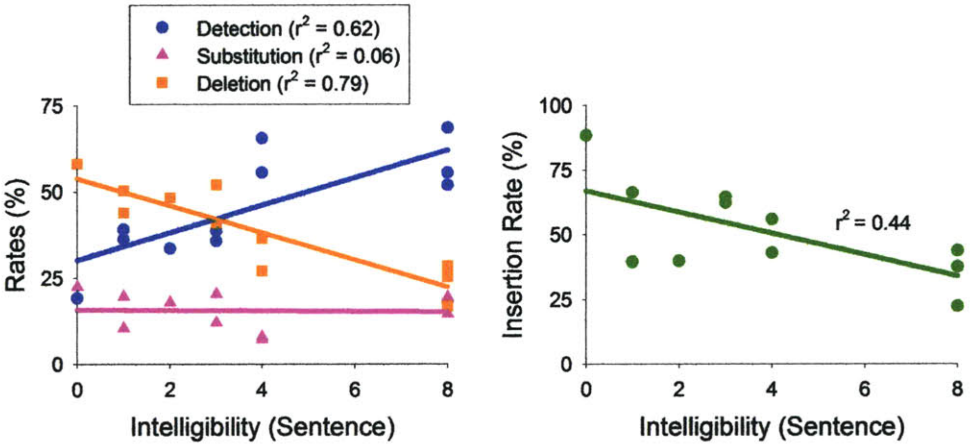

Figure D.3: Detection (blue), substitution (purple), deletion (orange), and insertion (green) rates for landmarks estimated from the broad class sequences suggested by EBS.

The base EBS model incorporates duration models trained on healthy speech. Individuals with severe dysarthria, however, often have very different duration profiles for their segmental units (Ansel & Kent, 1992; Platt, Andrews, & Howie, 1980; Vijayalakshmi & Reddy, 2006; Ziegler, Hartmann, & Hoole, 1993). Thus it would be expected that landmark detection rates would be reduced for speakers with lower intelligibility. To eliminate the influence of the duration models, the per-frame broad class probabilities were compared to the identities expected from the phonetic labeling. Comparing each frame's most-likely broad class suggested by EBS to its expected broad class, classification accuracy showed a strong correlation to intelligibility (Figure D.4, left). The per-frame accuracies, even for the healthy control speaker, were less than those acquired on the TIMIT test set. This was also observed when the base LAFF model was trained on TIMIT data and tested on the Nemours healthy control. One of the reasons for this discrepancy was the difference in recording conditions between the databases. Another reason was that non-sonorant frames were frequently classified as sonorant consonants. EBS is extremely sensitive to sonorants and incorporates a tuning parameter related to calculating sonorant consonant probability. This parameter was set at its default value for TIMIT. For speakers with dysarthria (SWDs) it would be expected that sonorant sensitivity would need to be reduced because excessive nasalization is a common deficiency. To deal with the sonorant sensitivity issue in this initial analysis, frame classification was re-performed where a frame was labeled correct if the mapped identity was found among the first or second most-likely frames (Figure D.4, right). This led to large increases in the detection rates and the correlation between classification accuracy and intelligibility was again significant (α<0.0001). Broad class frame accuracies for the speakers in the Nemours Database are listed in Table D.3. For most speakers sonorant consonants and silence frames were labeled with higher accuracies than the other broad classes.

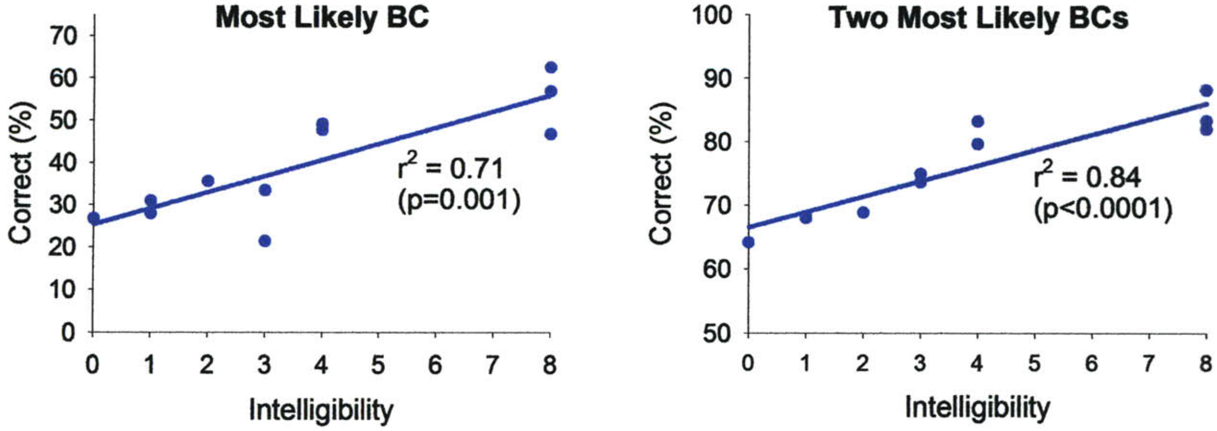

Figure D.4: Per-frame broad class classification accuracy as a function of speaker intelligibility. The left figure displays the percentage of frames correctly classified by the most likely broad class. The right figure is the percentage of frames where the identity of the correct broad class was among the two most likely broad classes.

Table D.3: Per-frame detection rates for each of the broad classes, for each speaker in the Nemours Database. Unenclosed numbers are the detection rates acquired using the most likely frame. Numbers in parentheses are the rates acquired using the two most likely frames.

Speaker	Vowel	Fricative	Stop	Sonorant Consonant	Silence
DYS1	26% (53%)	3% (74%)	7% (49%)	50% (69%)	66% (89%)
DYS2	22% (46%)	6% (84%)	22% (80%)	31% (50%)	91% (97%)
DYS3	14% (64%)	18% (59%)	38% (64%)	73% (86%)	68% (84%)
DYS4	27% (67%)	9% (46%)	12% (44%)	73% (89%)	69% (92%)
DYS5	8% (65%)	3% (87%)	8% (53%)	42% (64%)	89% (95%)
DYS6	25% (72%)	18% (70%)	8% (81%)	75% (87%)	56% (77%)
DYS7	24% (81%)	43% (75%)	27% (59%)	83% (95%)	92% (96%)
DYS8	51% (69%)	35% (87%)	34% (81%)	37% (89%)	89% (94%)
DYS9	62% (85%)	12% (64%)	20% (68%)	71% (98%)	90% (96%)
DYS10	49% (81%)	27% (82%)	31% (64%)	61% (84%)	72% (92%)
HC	57% (88%)	60% (82%)	56% (80%)	79% (97%)	74% (94%)

The potential advantage of EBS over LAFF for dysarthric speech recognition is that it does not rely on first extracting landmarks in order to perform feature extraction. For this strategy to be effective, however, requires that frames away from segmental boundaries be recognized with greater accuracy than frames adjacent to boundaries. Given higher motoric demands it has been suggested (Deller et al., 1995) that SWDs have the greatest difficulty and exhibit the most variability in production of segmental transitions. Assuming this hypothesis, correlates measured directly around landmarks would be expected to be noisier than during less transitory regions. Juneja and Espy-Wilson (2008) showed with the base EBS framework that for healthy speech higher per-frame recognition accuracies were achieved on frames located in the middle 50% of broad class segments compared to those on the periphery. Base analysis of the Nemours Database produced similar results for all SWDs and the healthy control (Figure D.5). Whether the extent of this behavior is sufficient for EBS to serve as a viable platform for ASR is unclear and should be addressed in future work.

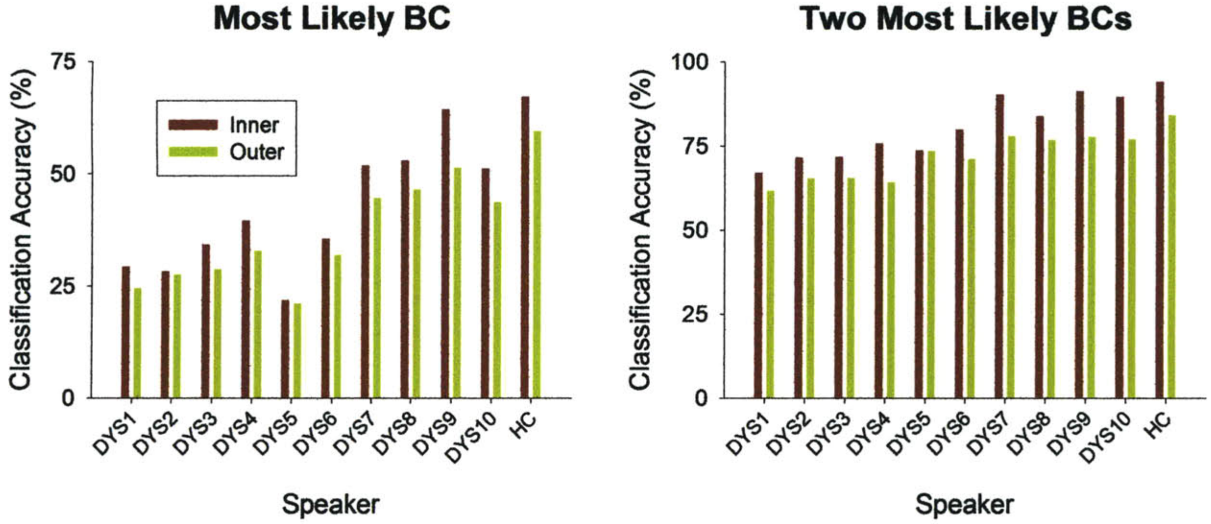

Figure D.5: Comparison of per-frame accuracies of frames located within the middle 50% (Inner) of broad class segments and of frames located within the two outer 25% peripheries (Outer). For all speakers from the Nemours Database, internal frames were recognized at a higher rate than peripheral frames.

The results using the standard EBS model resembled closely those produced by the base LAFF model. Landmark detection and insertion rates were correlated with speaker intelligibility. The promising result from base EBS analysis of the Nemours Database was that frames located away from changes in broad phonetic class were recognized at higher rates than those located near segmental boundaries. Applying the finding from the LAFF model that SWDs use distinct sets of acoustic correlates to convey changes in manner, it may be possible to identify correlates that convey the per-frame broad class. A frame-by-frame approach to speech recognition that does not rely heavily on regions of abrupt change could potentially prove beneficial for SWDs. Therefore, in future work correlate optimization should be performed for EBS to explore its potential as a feature-based ASR platform. It should be noted that for EBS (or any other feature-based model) to be a viable platform for ASR, both manner and place features must be identified with high accuracy. Given the high frequency of errors in place of articulation in dysarthria (Ansel & Kent, 1992; Platt, Andrews, & Howie, 1980; Platt, Andrews, Young et al., 1980), adequate extraction of place features could prove prohibitive given limited training data.

References

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech and Hearing Research, 35*(2), 296-308.

Aye, Y. Y. (2009). Speech Recognition Using Zero-Crossing Features. in *Proceedings of the 2009 International Conference on Electronic Computer Technology*, 689-692.

Baltaxe, C. A. M. (1978). Foundations of Distinctive Feature Theory.

Beukelman, D. R., & Yorkston, K. M. (1980). Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of communication disorders, 13*(1), 33-41.

Bitar, N. (1998). *Acoustic Analysis and Modeling of Speech Based on Phonetic Features.* Ph.D. thesis, Boston University, MA.

Bitar, N., & Espy-Wilson, C. (1995). A signal representation of speech based on phonetic features. in *Proceedings of the 5th Annual Dual-Use Technology and Applications Conference*, 310-315.

Bitar, N., & Espy-Wilson, C. (1996). A Knowledge-Based Signal Representation for Speech Recognition. in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing (ICASSP 1996)*, 29-32, Atlanta, GA, USA.

Black, A., & Lenzo, K. Building Voices in the Festival Speech Synthesis System. Retrieved April 28, 2008, from http://www.festvox.org/bsv

Blaney, B., & Wilson, J. (2000). Acoustic Variability in Dysarthria and Computer Speech Recognition. *Clinical Linguistics and Phonetics, 14*(4), 307-327.

Boutsen, F. R., Bakker, K., & Duffy, J. R. (1997). Subgroups in Ataxic Dysarthria. *Journal of Medical Speech-Language Pathology, 5*, 27-36.

Bowes, D. (1999). Getting it right and making it work! Selecting the right speech input and writing software for users with special needs. in *Proceedings of Technology and Persons with Disabilities*, California State University, Northridge.

Chen, N., Youngsook, J., & Park, C. (2007, May 18, 2008). Automatic Detection of Phonetic Boundaries. Retrieved April 28, 2008, from http://goodie7.tistory.com/attachment/do447.pdf

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English.* New York: Harper & Row.

D'Innocenzo, J., Tjaden, K., & Greenman, G. (2006). Intelligibility in dysarthria: Effects of listener familiarity and speaking condition. *Clinical linguistics & phonetics, 20*(9), 659-675.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12*(3), 462-496.

Dechter, R., & Pearl, J. (1985). Generalized best-first search strategies and the optimality of A*. *Journal of the Association for Computing Machinery (JACM), 32*(3), 505-536.

Deller, J. R., Hsu, D., & Ferrier, L. J. (1991). On the use of hidden Markov modeling for recognition of dysarthric speech. *Computer Methods and Programs Biomedicine, 35*(2), 125-139.

DePaul, R., & Kent, R. D. (2000). A longitudinal case study of ALS: Effects of listener familiarity and proficiency on intelligibility judgments. *American Journal of Speech-Language Pathology, 9*(3), 230-240.

DiCicco, T. M., & Patel, R. (2008). Automatic Landmark Analysis of Dysarthric Speech. *Journal of Medical Speech-Language Pathology, 16*(4), 213-221.

Doyle, P., Leeper, H., Kotler, A. L., Thomas-Stonell, N., O'Neill, C., Dylke, M. C., et al. (1997). Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development, 34*(3), 309-316.

Drish, J. (2001). Obtaining Calibrated Probability Estimates from Support Vector Machines. Retrieved April 28, 2008, from http://www.svms.org/classification/Dris.pdf

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). New York: Wiley.

Duffy, J. R. (2005). *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. St. Louis: Mosby.

Enderby, P. M. (1983). *Frenchay Dysarthria Assessment*. San Diego: College Hill Press.

Espy-Wilson, C. (1992). Acoustic measures for linguistic features distinguishing the semivowels/wjrl/in American English. *Journal of the Acoustical Society of America, 92*, 736–757.

Espy-Wilson, C., & Bitar, N. (1997). The Design of Acoustic Parameters for Speaker-Independent Speech Recognition. in *Proceedings of the European Conference on Speech Communication and Technology (EuroSpeech 1997)*,1239-1242, Patras, Greece.

Espy-Wilson, C., Pruthi, T., Juneja, A., & Deshmukh, O. (2007). Landmark-based approach to speech recognition: An alternative to HMMs. in *Proceedings of Interspeech 2007*, 886-889.

Fell, H., MacAuslan, J., Chenausky, K., & Ferrier, L. (1999). Automatic babble recognition for early detection of speech related disorders. *Communication Abstracts, 22*(5).

Ferrier, L. J., Shane, H. C., Ballard, H. F., Carpenter, T., & Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication, 11*(3), 165-175.

Fisher, W. M., Doddington, G. R., & Goudie-Marshall, K. M. (1986). The DARPA speech recognition research database: specifications and status. in *Proceedings of the DARPA speech recognition workshop*, 93-99.

Gaudrain, E., Grimault, N., Healy, E. W., & Béra, J. C. (2007). Effect of spectral smearing on the perceptual segregation of vowel sequences. *Hearing Research, 231*(1-2), 32-41.

Gauvain, J. L., & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixtureobservations of Markov chains. *IEEE Transactions on Speech and Audio Processing, 2*(2), 291-298.

Gentil, M. (1990). Acoustic Characteristics of Speech in Friedreich's Disease. *Folia Phoniatrica et Logopaedica, 42*, 125-134.

Glass, J. R., Chang, J., & McCandless, M. (1996). A Probabilistic Framework for Feature-Based Speech Recognition. in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, 2277-2280.

Goemans, M. X., & Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the Association for Computing Machinery (JACM), 42*(6), 1115-1145.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., et al. (2005). Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop. *in Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2005)*, 213 - 216. Philadelphia, PA, USA, March 2005.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. New York: Springer.

Hirose, H., Kiritani, S., & Sawashima, M. (1982). Velocity of articulatory movements in normal and dysarthric subjects. *Folia Phoniatrica et Logopaedica, 34*(4), 210-215.

Howitt, A. W. (2000a). *Automatic Syllable Detection for Vowel Landmarks.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Howitt, A. W. (2000b). Vowel Landmark Detection in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, 628-631.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12*(2), 198-208.

Huttenlocher, D. P., & Zue, V. W. (1984). A Model of Lexical Access from Partial Phonetic Information. in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1984)*, 391-394.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to Speech Analysis. The distinctive features and their correlates.* (No. 13): Acoustics Laboratory, Massachusetts Institute of Technology, Technical Report.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Perception & Psychophysics, 34*(5), 441-450.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America, 85*, 1718-1725.

Juneja, A., & Espy-Wilson, C. (2003). Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. in *International Joint Conference on Neural Networks*, Portland, Oregon.

Juneja, A., & Espy-Wilson, C. (2008). A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *Journal of the Acoustical Society of America, 123*(2), 1154-1168.

Jurafsky, D., Martin, J. H., & Kehler, A. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*: MIT Press.

Kent, R. D., Kent, J. F., Rosenbek, J. C., Vorperian, H. K., & Weismer, G. (1997). A speaking task analysis of the dysarthria in cerebellar disease. *Folia Phoniatrica et Logopaedica, 49*(2), 63-82.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *American Journal of Speech-Language Pathology, 54*(4), 482-499.

Keyser, S. J., & Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language, 82*(1), 33-63.

King, J. M., & Gallegos-Santillan, P. (1999). Strategy use by speakers with dysarthria and both familiar and unfamiliar conversational partners. *Journal of Medical Speech-Language Pathology, 7*, 113-116.

Kominek, J. K., Bennett, C., & Black, A. W. (2003). Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. in *Proceedings of the European Conference on Speech Communication and Technology (EuroSpeech 2003)*, Geneva, Switzerland, 313-316.

Kotler, A. L., & Thomas-Stonell, N. (1997). Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment. *Journal of Augmentative and Alternative, 12*, 71-80.

Kuhn, R., Junqua, J. C., Nguyen, P., & Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing, 8*(6), 695-707.

Ladefoged, P., & Halle, M. (1988). Some major features of the International Phonetic Alphabet. *Language*, 577-582.

Le Dorze, G., Ouellet, L., & Ryalls, J. (1994). Intonation and speech rate in dysarthric speech. *Journal of Communication Disorders, 27*(1), 1-18.

Lee, C. H., & Gauvain, J. L. (1993). Speaker adaptation based on MAP estimation of HMM parameters. in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing (ICASSP 1993)*, Minneapolis, MN, 27-30.

Lee, C. H., Lin, C. H., & Juang, B. H. (1991). A study on speaker adaptation of the parameters of continuousdensity hidden Markov models. *IEEE Transactions on Signal Processing, 39*(4), 806-814.

Lee, S., & Glass, J. R. (1998, November, 1998). Real-Time Probabilistic Segmentation for Segment-Based Speech Recognition. in *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language, 9*(2), 171.

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America, 112*, 3022-3030.

Liu, S. A. (1995). *Landmark Detection for Distinctive Feature-based Speech Recognition.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Liu, S. A. (1996). Landmark detection for distinctive feature-based speech recognition. *Journal of the Acoustical Society of America, 100*(5), 3417-3430.

Logemann, J., & Fisher, H. B. (1981). Vocal Tract Control in Parkinson's Disease. *Journal of Speech and Hearing Disorders, 46*, 348-352.

Luebke, K., & Weihs, C. (2005). Improving feature extraction by replacing the Fisher criterion by an upper error bound. *Pattern Recognition, 38*(11), 2220-2223.

Marti, R., Duarte, A., & Laguna, M. (2009). Advanced Scatter Search for the Max-Cut Problem. *INFORMS Journal on Computing, 21*(1), 26-38.

Menéndez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996). The Nemours Database of Dysarthric Speech. in *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA, USA.,1962-1965. Philadelphia, PA, USA.

Mercier, G., Bigorgne, D., Miclet, L., Guennec, L. L., & Querre, M. (1990). Recognition of speaker-dependent continuous speech with KEAL. In *Readings in Speech Recognition* (pp. 225-234): Morgan Kaufmann Publishers Inc.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America, 58*(4), 880-883.

Mitchell, C. D., & Jamieson, L. H. (1993). Modeling duration in a hidden Markov model with the exponential family. in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP1993)*, 331-334.

Murdoch, B. E. (1998). *Dysarthria: A Physiological Approach to Assessment and Treatment*: Nelson Thornes.

Neilson, P., & O'Dwyer, N. J. (1984). Reproducibility and Variability of Speech Muscle Activity in Athetoid Dysarthria of Cerebral Palsy. *Journal of Speech and Hearing Research, 27*, 502-517.

Ohde, R. N., & Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America, 74*, 706-714.

Park, C. (2008). *Consonant Landmark Detection for Speech Recognition.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Patel, R. (1999). Prosody conveys information in severely impaired speech. in *Proceedings of the European Speech Community Association Workshop of Dialogue and Prosody*, 129-135, Veldhoven, The Netherlands.

Patel, R. (2002a). Phonatory control in adults with cerebral palsy and severe dysarthria. *Augmentative and Alternative Communication, 18*(1), 2 - 10.

Patel, R. (2002b). Prosodic control in severe dysarthria: preserved ability to mark the question-statement contrast. *Journal of Speech and Hearing Research, 45*(5), 858-870.

Patel, R. (2004). The Acoustics of Contrastive Prosody in Adults With Cerebral Palsy. *Journal of Medical Speech-Language Pathology, 12*(4), 189-194.

Patel, R., & Campellone, P. (2009). Acoustic and Perceptual Cues to Contrastive Stress in Dysarthria. *Journal of Speech, Language, and Hearing Research, 52*, 206-222.

Pekalska, E., Harol, A., Lai, C., & Duin, R. P. W. (2005). Pairwise selection of features and prototypes. in *Proceedings of the 4-th International Conference on Computer Recognition Systems (CORES)*, 271–278.

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology 163*(7), 670-675.

Platt, L. J., Andrews, G., & Howie, P. M. (1980). Dysarthria of adult cerebral palsy: II. Phonemic analysis of articulation errors. *Journal of Speech and Hearing Research, 23*(1), 41-55.

Platt, L. J., Andrews, G., Young, M., & Quinn, P. T. (1980). Dysarthria of adult cerebral palsy: I. Intelligibility and articulatory impairment. *Journal of Speech and Hearing Research, 23*(1), 28-40.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1071-1086.

Portnoy, R. A., & Aronson, A. E. (1982). Diadochokinetic Syllable Rate and Regularity in Normal and in Spastic and Ataxic Dysarthric Subjects. *Journal of Speech and Hearing Disorders, 47*, 324-328.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*: Prentice Hall.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465–471.

Rosengren, E., Raghavendra, P., & Hunnicutt, S. (1995). How does automatic speech recognition handle severely dysarthric speech? in *Proceedings of the Second TIDE Congress on The European Context for Assistive Technology*, Paris, 336-339.

Salomon, A., Espy-Wilson, C., & Deshmukh, O. (2004). Detection of speech landmarks: Use of temporal information. *The Journal of the Acoustical Society of America, 115*, 1296-1305.

Seneff, S., & Zue, V. (1988, November). Transcription and alignment of the TIMIT database. in *Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, Hawaii.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Shinoda, K. (2005). Speaker adaptation techniques for speech recognition using probabilistic models. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science), 88*(12), 371-386.

Shipman, D. W., & Zue, V. W. (1982). Properties of large lexicons: Implications for advanced isolated word recognition systems. in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 546-549.

Slifka, J. (2005a). Acoustic cues, landmarks, and distinctive features: a model of human speech processing. in *Proceedings of the 6th International Symposium on Natural Language Processing*, Chang Rai, Thailand, 91-96.

Slifka, J. (2005b). Acoustic cues to vowel schwa sequences for high front vowels. *The Journal of the Acoustical Society of America, 118(3)*, 2037.

Spitzer, S. M., Liss, J. M., Caviness, J. N., & Adler, C. (2000). An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech. *Journal of Medical Speech Language Pathology, 8*(4), 285-294.

Stevens, K. N. (1985). Evidence for the role o acoustic boundaries in the perception of speech sounds. In V. A. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*: Academic Press, Inc.

Stevens, K. N. (1992). Lexical access from features. *Speech Communication Group Working Papers, Volume VIII*, Research Laboratory of Electronics, Massachusetts Institute of Technology, 119-144.

Stevens, K. N. (1999). *Acoustic Phonetics*. Cambridge, MA: The M.I.T. Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111*(4), 1872-1891.

Stevens, K. N. (2007). *6.543J The Lexicon and Its Features*, A course offered at Massachusetts Institute of Technology.

Stevens, K. N., Keyser, S. J., & Kawasaki, H. (1986). Toward a Phonetic and Phonological Theory of Redundant Features. in J.S. Perkell and D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes:* Lawrence Erlbaum, Hillsdale. 426-449.

Ström, N. (1996). Continuous speech recognition in the WAXHOLM dialogue system. *TMH-QPSR, 37*(4), 67-96.

Sun, W. (1996). *Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition.* M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Surendran, A. C., & Lee, C. H. (2001). Transformation-based Bayesian prediction for adaptation of HMMs. *Speech Communication, 34*(1-2), 159-174.

Talkin, D. (1995). A Robust Algorithm for Pitch Tracking (RAPT). in W. B. Kleijn & K. K. Paliwal (Eds.), *Speech Coding and Synthesis*: Elsevier. 495-518.

Tartter, V. C., Kat, D., Samuel, A. G., & Repp, B. H. (1983). Perception of intervocalic stop consonants: the contributions of closure duration and formant transitions. *Journal of the Acoustical Society of America, 74*, 715-725.

Thomas-Stonell, N., Kotler, A. L., Leeper, H., & Doyle, P. (1998). Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative & Alternative Communication, 14*(1), 6.

Tjaden, K., Rivera, D., Wilding, G., & Turner, G. S. (2005). Characteristics of the Lax Vowel Space in Dysarthria. *Journal of Speech and Hearing Research, 48*(3), 554-566.

Tjaden, K., & Wilding, G. E. (2004). Rate and Loudness Manipulations in Dysarthria: Acoustic and Perceptual Findings. *Journal of Speech, Language, and Hearing Research, 47*(4), 766-783.

Tjaden, K. K., & Liss, J. M. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical linguistics & phonetics, 9*(2), 139-154.

Vance, J. E. (1994). Prosodic deviation in dysarthria: a case study. *European Journal of Disorders of Communication, 29*(1), 61-76.

Vijayalakshmi, P., & Reddy, M. R. (2006). Assessment of dysarthric speech and an analysis on velopharyngeal incompetence. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3759-3762.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*(2), 260-269.

Waibel, A., & Lee, K. F. (1990). *Readings in Speech Recognition*: Morgan Kaufmann.

Wavesurfer. (2005). Centre for Speech Technology (CTT) at KTH in Stockholm, Sweden. http://www.speech.kth.se/wavesurfer

Yorkston, K. M., & Beukelman, D. R. (1981). *The Assessment of Intelligibility of Dysarthric Speakers*. Austin, TX: PRO-ED.

Yorkston, K. M., Beukelman, D. R., Strand, E. A., & Bell, K. R. (1999). *Management of Motor Speech Disorders in Children and Adults*. Austin, TX: Pro-Ed.

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory Movements During Vowels in Speakers With Dysarthria and Healthy Controls. *Journal of Speech, Language, and Hearing Research, 51*(3), 596-611.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. in *Proceedings of the Eighteenth International Conference on Machine Learning.*

Ziegler, W., Hartmann, E., & Hoole, P. (1993). Syllabic timing in dysarthria. *Journal of Speech and Hearing Research, 36*(4), 683-693.

Ziegler, W., & von Cramon, D. (1983a). Vowel Distortion in Traumatic Dysarthria: A Formant Study. *Phonetica, 40*, 63-78.

Ziegler, W., & von Cramon, D. (1983b). Vowel Distortion in Traumatic Dysarthria: Lip Rounding versus Tongue Advancement. *Phonetica, 40,* 312-322.

Ziegler, W., & von Cramon, D. (1986). Spastic dysarthria after acquired brain injury: An acoustic study. *British Journal of Disorders of Communication, 21*, 173-187.

Ziegler, W., & Wessel, K. (1996). Speech Timing in Ataxic Disorders. *Neurology, 47*, 208-214.

Zue, V. W., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication, 9*(4), 351-365.