PRODUCTION KNOWLEDGE IN THE RECOGNITION OF DYSARTHRIC SPEECH

by

Frank Rudzicz

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Department of Computer Science
University of Toronto

Abstract

Production knowledge in the recognition of dysarthric speech

Frank Rudzicz

Doctor of Philosophy

Graduate Department of Department of Computer Science

University of Toronto

2011

Millions of individuals have acquired or have been born with neuro-motor conditions that limit the control of their muscles, including those that manipulate the articulators of the vocal tract. These conditions, collectively called dysarthria, result in speech that is very difficult to understand, despite being generally syntactically and semantically correct. This difficulty is not limited to human listeners, but also adversely affects the performance of traditional automatic speech recognition (ASR) systems, which in some cases can be completely unusable by the affected individual.

This dissertation describes research into improving ASR for speakers with dysarthria by means of incorporated knowledge of their speech production. The document first introduces theoretical aspects of dysarthria and of speech production and outlines related work in these combined areas within ASR. It then describes the acquisition and analysis of the TORGO database of dysarthric articulatory motion and demonstrates several consistent behaviours among speakers in this database, including predictable pronunciation errors, for example. Articulatory data are then used to train augmented ASR systems that model the statistical relationships between vocal tract configurations and their acoustic consequences. I show that dynamic Bayesian networks augmented with instantaneous theoretical or empirical articulatory variables outperform even discriminative alternatives. This leads to work that incorporates a more rigid theory of speech production, i.e., task-dynamics, that models the high-level and long-term aspects of speech production. For this task, I devised an algorithm for estimating articulatory

positions given only acoustics that significantly outperforms the state-of-the-art. Finally, I present ongoing work into the transformation and re-synthesis of dysarthric speech in order to make it more intelligible to human listeners.

This research represents definitive progress towards the accommodation of dysarthric speech within modern speech recognition systems. However, there is much more research that remains to be undertaken and I conclude with some thoughts as to which paths we might now take.

Acknowledgements

This work was funded at various times by Bell University Labs, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto. Parts of chapter 4 were written in collaboration with Aravind Namasivayam and Talya Wolff. Section 6.5.1 was written in collaboration with Michael Reimer.

Alas, any number of years is far too short a time to live among such excellent and admirable hobbits. The computational linguistics group at the University of Toronto constitutes research of unbound diversity and profundity among which I was delighted and fortunate to dwell.

I thank the students and research associates who helped raise this project to maturity. These include but are not limited to Talya Wolff, Beverly Ho, Heidi Diepstra, and the researchers at the Oral Dynamics Laboratory. Special thanks go to Aravind Namasivayam, Siavash Kazemian, Kinfe Mengistu, and Michael Reimer.

I acknowledge the inspiring researchers in whose footsteps I walked, especially Mark Hasegawa-Johnson, Karen Livescu, and Sam Roweis.

The members of my doctoral committee are diligent examiners of detail yet far-seeing in their vision. I thank Pascal van Lieshout for providing the use of his laboratory, and for the theoretical and biological underpinnings of this work. I thank Fraser Shein for an applied focus, a constant consideration for the end user, and for his guidance during my time at Quillsoft. I thank Gerald Penn for his amazing capacity to find and explain the important theoretical and mathematical aspects of this work, and for many enjoyable lunches. I thank Graeme Hirst for his guidance in all things, for setting expectations higher than necessary, for helping me meet those expectations, and for making me a better person through his example.

Most importantly, I thank my wife Melissa without whom none of this would be possible and for whom any words here would be inadequate. Melissa is my entire life. I dedicate this thesis to her and to our biological collaboration now stored in her abdomen.

- - -

This thesis has passed on. This thesis is no more. It has ceased to be. It's expired and gone to meet its maker. It's a stiff. Bereft of new data, it rests in peace. If I hadn't bound its pages together, it would be pushing up the daisies. It's run down the curtain and joined the bookshelf invisible. This is an ex-thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There are several simplifying assumptions in automatic speech recognition (ASR) that have become particularly ingrained. One such assumption is that the acoustics of speech can be adequately described despite being agnostic to non-surface phenomena. Although ASR takes a few important cues from the biological perception of speech, such as the Mel scale (O'Shaughnessy, 2000), it rarely models physical production explicitly. Secondly, modern ASR is often built assuming that models trained on a sufficiently large set of speakers will adequately capture enough inter-speaker variability to be usable by a typical user. The further one's voice deviates from this aggregate, however, the less likely an ASR system is to function as intended.

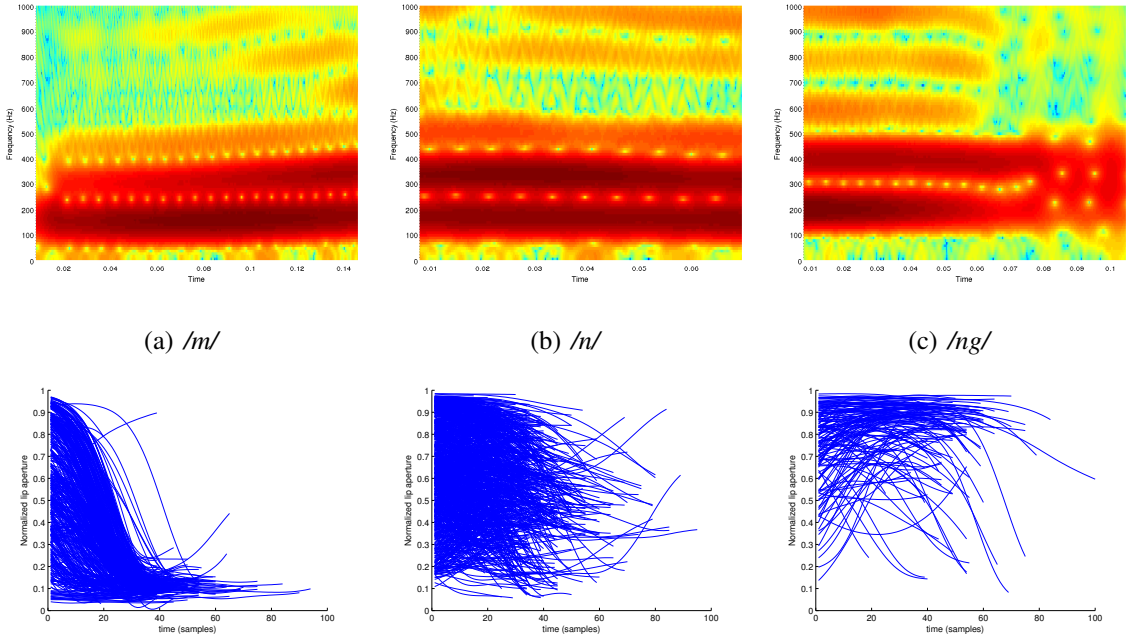
Each of these simplifications is useful in certain contexts but their utility in the presence of more atypical patterns of production can be disputed, especially in cases of speech disorder. One group of such disorders, called dysarthria, is primarily an endogenous phenomenon distinguished by its aberrant mechanics of articulation resulting in highly unintelligible speech that is not accommodating to the traditional assumptions of speech recognition. The intuition behind this work is that more informed models are also more accurate and that by studying the phenomena of dysarthria empirically and encoding the results explicitly we can build more suitable software for those with speech disabilities. Surprisingly, there has so far been relatively little research that incorporates production knowledge into ASR, especially for this population.

This document describes research that has significantly improved automatic speech recognition accuracy for individuals with dysarthria by augmenting acoustic models with articulatory information. This information is obtained through the collection of an extensive new database of dysarthric speech called TORGO. The relationships between acoustics and articulation are especially relevant for these speakers, for whom normal speech production is compromised. The TORGO database is also applied to new models of speech recognition that incorporate high-level abstractions of speech production.

1.1 Central thesis

The thesis presented in this dissertation is that *classification systems built using empirical and theoretical models of speech production can significantly improve recognition accuracy for speakers with dysarthria*. Current automatic speech recognition technology has produced mature and widespread tools for the general public, but a preponderance of error in recognizing and adapting to dysarthric speech has kept such software effectively inaccessible to individuals with severe speech disorders. The driving motivation is therefore to be able to augment the expressive abilities of those with communicative challenges. This motivation is manifested in a secondary component of this dissertation, which provides an inaugural step into the applied areas of augmentative and alternative communication.

The central intuition is that an understanding of the source of a phenomenon will instruct us on how to manage its consequences. This is a useful heuristic in any scientific domain. Figure 1.1 provides a motivating example of the utility of studying the articulatory source of speech. The spectrographic representations of the acoustics of three nasals (the top row of figure 1.1) are superficially indistinguishable and, indeed, classification of these phonemes is difficult. However, by simply observing the degree of lip aperture over time (the bottom row of figure 1.1), we can fairly easily determine which of the three involves bilabial closure and hence classification becomes significantly simpler.



(a) */m/* (b) */n/* (c) */ng/*

Figure 1.1: Example spectrograms for nasals (a) */m/*, (b) */n/*, and (c) */ng/*, and all normalized lip aperture traces over the same phonemes for non-dysarthric speech in the TORGO database. The distinction between nasals can more easily be made, visually at least, by observing the relevant articulators. The */m/* shows the clearest directed movement, namely to close the lips. The lip traces for */n/* and */ng/* are more disordered and their patterns may suggest a dependence on their phonemic contexts. Except where noted, this thesis uses the ARPAbet phone set.

1.2 Three perspectives of the thesis

The work described in this dissertation takes place at the intersection of three disciplines – computer science, speech-language pathology, and rehabilitation science. The following subsections outline this document according to the particular perspectives of these disciplines.

1.2.1 The perspective from computer science and pattern recognition

At one level, dysarthric speech represents a highly entropic data set that is especially challenging for the less discriminative of statistical models. At another level, however, it is apparent that the long-term and speaker-dependent nature of dysarthric speech requires a fundamentally new approach to processing speech data. In either case, additional data is required to construct more complex models, and this data must include supplementary information, such as the articulatory causes of the acoustic consequences.

Chapter 2 provides some background on anatomical aspects of speech production and clinical aspects of dysarthria. Chapter 3 outlines a number of techniques that have been applied to the scenario of atypical speech, including discriminative methods and modifications to traditional hidden Markov models. These methods, however, are limited in their lack of articulatory knowledge. Chapter 4 describes the collection of our own database of dysarthric articulation, called TORGO. Since this database was always intended to be used to train ASR systems, a number of engineering decisions were factored into its design including the selection of stimuli to elicit a broad range of ecologically valid observations and categories. Chapter 5 incorporates this database and the additional statistical knowledge it presents into advanced discriminative classification methods augmented with articulatory knowledge. This leads to completely new conceptualizations of the very task of ASR in terms of higher-level dynamical models of speech production, as described in chapter 6. Finally, Chapter 7 is more applied than its predecessors and sketches a system that transforms dysarthric input audio in order to make it more intelligible to human listeners.

1.2.2 The perspective from speech-language pathology

Pattern recognition algorithms depend pervasively on a perspective that speech is a sequence of very brief and non-overlapping patterns of sound. The true nature of speech, however, involves overlapping long-term dynamics and hidden behaviours that routinely evade the scrutiny of such a perspective. This dissertation highlights areas in which traditional ASR lacks the theoretical underpinnings of speech science and empirically analyzes approaches that unify those theoretical aspects with applied systems. The particular influences of speech science used in this dissertation may suggest other novel ways in which speech theory can be applied to methods that classify speech sounds into words or phonemes, for example.

Chapter 2 provides some background on algorithms and models used in traditional automatic speech recognition. Chapter 3, though mostly referring to research in engineering, demonstrates which aspects of dysarthric speech have traditionally been engaged within ASR. The creation and nature of a new database of dysarthric articulation, called TORGO, is described in chapter 4. In particular, this chapter demonstrates the use of specialized techniques in the collection of articulatory data and provides insight into a number of relationships between articulation and acoustics of dysarthric speech. Further insight may be obtained through the third-party use of this database by speech-language pathologists, clinicians, and speech scientists.

Chapter 5 describes the use of state-of-the-art discriminative classification methods in the recognition of dysarthric speech, given articulatory data. Although these methods are relatively successful, they also indicate a boundary for what is possible using short-term representations of speech. Chapter 6 challenges this boundary by incorporating a number of long-term representations of speech into ASR, including principal differential analysis and Task Dynamics. This chapter may inspire the use of other behavioural representations of speech such as DIVA (Guenther and Perkell, 2004) in similar contexts. Finally, chapter 7 explores various acoustic transformations of dysarthric speech in order to make it more intelligible to human listeners. The aspects of dysarthric speech selected in this study are of particular relevance, and the

demonstration that certain problems that may interfere with intelligibility can be identified and corrected highlights the importance of discovering and systematically identifying those aspects of disordered speech.

1.2.3 The perspective from rehabilitation science

Rehabilitation science often involves the use of modern and breakthrough technologies in order to manage physical and cognitive disabilities, to empirically evaluate those solutions, and to improve the quality of life for users of these technologies. The use of language as a means of personal expression is routinely supported by writing aids both in hardware and software and by systems that either produce speech output or accept spoken phrases and commands. It is therefore crucial that such systems correctly identify the words spoken by individuals with speech disabilities. This ability of speech recognition must necessarily precede the development of applications whose interfaces are tuned to these populations.

Chapter 2 provides some background on algorithms and models used in traditional automatic speech recognition and on anatomical aspects of speech production and clinical aspects of dysarthria. Chapter 3 surveys the prior state-of-the-art in speech recognition, often in semi-clinical contexts, in which user-driven models of speech are used to handle specific disorders. The TORGO database of dysarthric speech is introduced in chapter 4. Since this database includes both detailed acoustic and articulatory data, it can provide clinicians with valuable insight as to how specific motor deficits affect resulting speech intelligibility and may therefore be useful in the development of treatment protocols. Chapter 5 describes experiments with state-of-the-art speech recognition systems with dysarthric speech and indicates the boundaries of what is possible or expected in real-world scenarios. Chapter 6 expands on this work with the presentation of new higher-level models.

Chapter 7 experimentally analyzes modifications to the acoustics of dysarthric speech in order to make that speech more intelligible. This chapter shows that the effects of dysarthria on intelligibility can be mitigated by specific adjustments to observed behaviours of dysarthria,

such as adjusting the phonemes in mispronounced words. This represents the first step towards building fully automatic augmentative systems capable of assisting individuals to communicate more effectively with the general public.

Chapter 2

Background

This chapter summarizes the fundamental conceptual building blocks that are used to construct the thesis throughout this dissertation. Section 2.1 describes the relevant features of the anatomy of speech production, section 2.2 describes some fundamental clinical aspects of dysarthria, section 2.3 overviews some fundamental knowledge in automatic speech recognition, and section 2.4 enumerates a few conceptual frameworks by which speech anatomy can be represented for use in speech recognition.

2.1 Anatomy of speech production

The organs used in speech have all evolved for purposes other than speaking (e.g., breathing, eating) and have only comparatively recently been adapted to speech, making them in some sense a suboptimal communication mechanism (O'Shaughnessy, 2000). The speech organs can be subdivided into three groups: the lungs, the larynx, and the upper vocal tract consisting of the jaw, lips, tongue, and mouth walls. The lungs provide all of the airflow that is transformed by the rest of the vocal tract into the time-varying air pressure waves that constitute speech. During speech, the diaphragm muscle compresses the lungs, producing a pressure of 10–20 cm H_2O, compared with 1–2 cm H_2O required for normal breathing (O'Shaughnessy, 2000). Normal breathing is generally almost inaudible since the air pressure expelled by the

lungs is unobstructed by the vocal tract. Many animals, however, can create vocal noise by sinusoidally obstructing this air flow by means of the larynx, which is supported by the thyroid, cricoid, arytenoid, and epiglottal cartilages (Sundberg, 1977). These cartilages are shown in figure 2.1, with the cricoid cartilage below the vocal fold, and the arytenoid cartilages in the posterior section which can move to abduct or adduct the vocal folds. The vocal folds do not follow muscle contractions directly, but certain muscles are involved in changing the characteristics of the quasi-periodic airflow. Changes to the fundamental frequency (F0) of speech[1] are primarily caused by two laryngeal muscles – the vocalis muscle in the vocal folds and the cricothyroid which can increase F0 by tensing and lengthening the vocal folds by up to 4 mm (Löfqvist, McGarr, and Honda, 1984). The fundamental frequency can also be lowered by active contractions of the thyroarytenoid and sternohyoid muscles (Titze, 1994). If these muscles cannot be controlled (i.e., contracted or relaxed), the vocal folds are tightly adducted and cannot vibrate normally, resulting in harsh, irregular F0 (Schneiderman and Potter, 2002).

The muscular and bony tissue structures above the larynx contribute to speech by either warping the spectral distribution acoustic waves or by generating certain obstruent sounds such as plosives and fricatives. The jaw is an important articulator controlled by the masseter and pterygoid muscles, although its function is largely indirect in that its placement is used to assist the positioning of the tongue and lips. The lips themselves are controlled by a number of muscles. The orbicularis oris surrounds the mouth, protrudes the lips outward, and rounds the lips when contracted. The buccinator is a thin muscle below the cheekbones that stretches toward the mandible and controls retraction and spreading of the lips. The depressor anguli oris and depressor labii inferioris muscles lie below the lips and away from the midsagittal plane and pull the lip corners downwards. These muscles have counterparts levator anguli oris and levator labii superioris above the lips which pull the lips upwards. All of these muscles are controlled by the facial nerve, described below.

[1]The fundamental frequency is the rate of vocal fold vibration and generally corresponds to the aspect of perceived pitch (Stevens, 1998). Harmonics in the speech spectrum occur at multiples of the fundamental frequency, which can be determined using methods described in chapter 7.

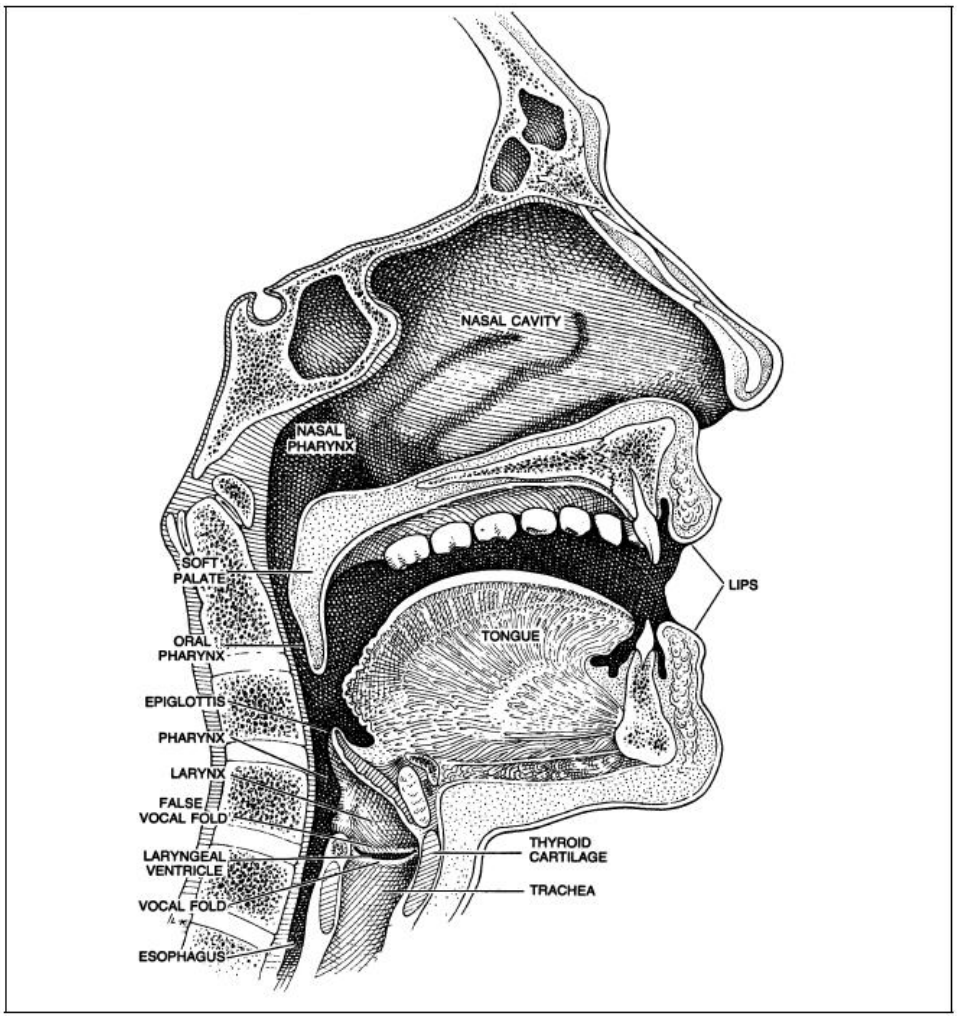


Figure 2.1: The vocal organs, as shown in the midsagittal plane. Illustration by Laszlo Kubinyi in Sundberg (1977).

The tongue is perhaps the most complex and most important articulator in speech, consisting of 12 muscle pairs and tissues (O'Shaughnessy, 2000). The tongue provides almost all movement within the mouth, with the exception of the velum, which lowers and raises the rear of the oral cavity to allow air to pass into the nasal cavity. In normal conditions there is almost no significant lateral tongue movement, though the tongue is highly agile and can be reconfigured between relevant positions in less than 50 ms (Stevens, 1998). The tip and dorsum of the tongue are two important areas that allow quick constrictions to occur at various positions along the vocal tract.

It is useful to conceptualize of the speech production system as a conjunction of at least two parts: a source which generates sound waves, and a filter which shapes those waves. The source is represented in this source-filter model by the glottis, whose rate of vibration provides harmonics at higher multiples of that frequency. The locations of these harmonics are determined by the interaction of the sound waves with the oral cavity walls, but especially by sudden changes in the width of that cavity. These sudden changes are due almost exclusively to the configuration of the tongue, which is the primary causative agent of the filter. Figure 2.2 illustrates this relationship between the physical contour of the tongue and the resulting effect on the distribution of the formants of vowels. The uniform-tube model which artificially describes this phenomenon is described further in section 7.2.

All of the speech musculature is controlled by the brain where voluntary movement is initiated by the motor cortex. However, messages produced by the higher structures are transmitted through highly specialized cranial nerves (CN) that emerge through fissures in the lower brain around the cerebellum and basal ganglia. Figure 2.3 shows the cranial nerves. These nerves carry the impulses that constrict the musculature but also communicate sensory data back to the brain. All of the facial musculature is innervated by the primary facial nerve (CN VII), although submandibular and sublingual motion is also controlled by the intermediate facial nerve and the muscles of mastication are controlled by the trigeminal nerve (CN V). Perhaps most important is the hypoglossal nerve (CN XII) which controls almost all intrinsic and ex-

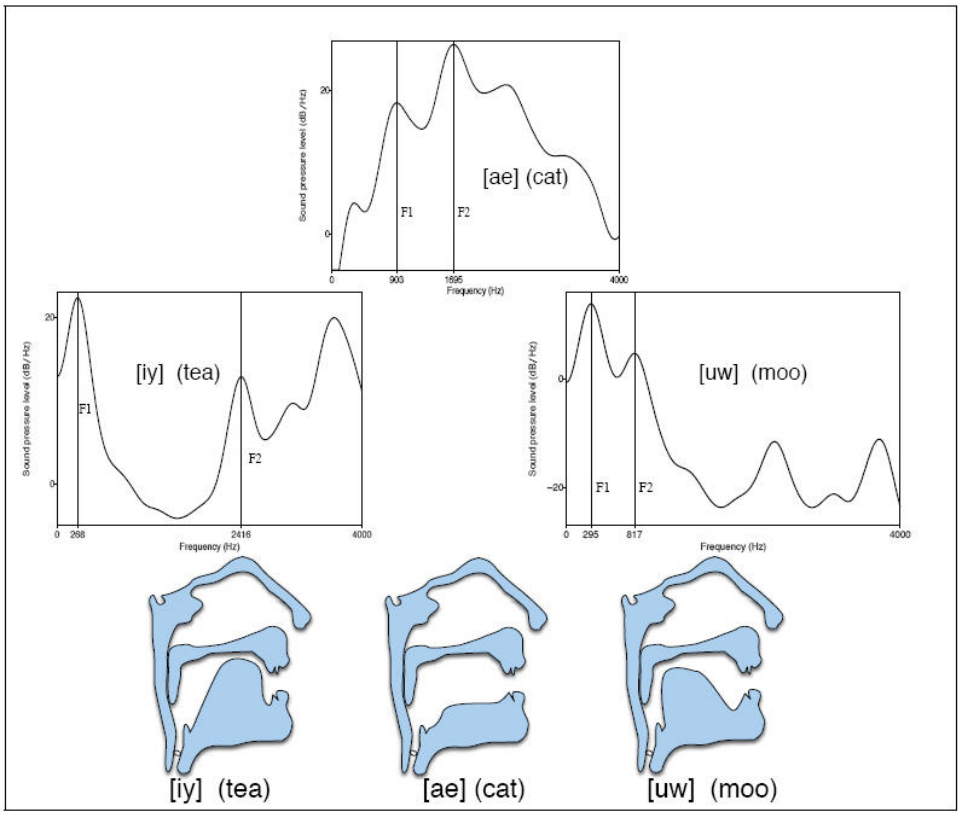


Figure 2.2: Exemplar configurations of the tongue for three English vowels and the resulting spectral envelopes, from Jurafsky and Martin (2009). These examples demonstrate the effect that the dorsoventral (front-back) position of the tongue has on the distribution of F2 and that the superior-inferior (top-down) position has on F1.

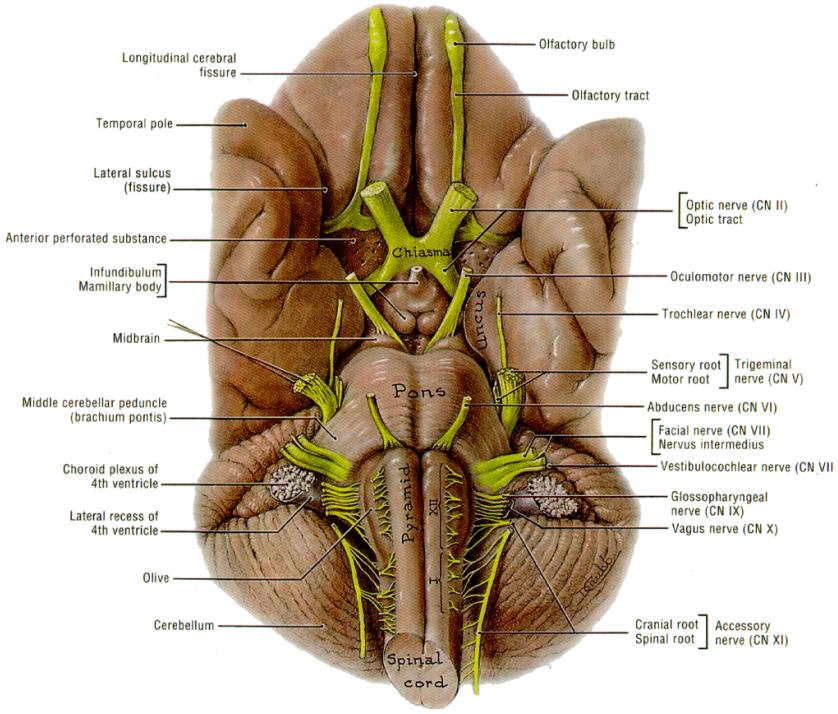


Figure 2.3: Inferior ventral view of the brain highlighting the cranial nerves, from Moore and Dalley (2005).

trinsic muscles of the tongue. When these cranial nerves are disrupted or rendered inoperative, partial paralysis of the respective musculature occurs as signal information is not transmitted. By contrast, if these nerves are activated involuntarily, the muscles can react in relatively unpredictable ways. Specific effects of damage to the cranial nerves are discussed in section 2.2 and modelled loosely in section 6.3.

2.2 Dysarthria

Dysarthria is a set of congenital and traumatic neuromotor disorders that impair the physical production of speech. These impairments reduce or remove normal control of the primary vocal articulators but do not affect the regular comprehension or production of meaningful, syntactically correct language. Congenital causes of dysarthric speech are often manifested by some sort of asphyxiation of the brain, inhibiting normal development in the speech-motor ar-

eas. Of these causes, cerebral palsy is among the most common[2], affecting approximately 0.5% of children in North America (Hasegawa-Johnson et al., 2006b), 88% of whom are dysarthric throughout adulthood (Augmentative Communication Incorporated (ACI), 2007). Later-onset causes are more typically traumatic, including cerebro-vascular stroke affecting approximately 1% of the population aged 45 to 64, and 5% of those aged 65+, with the severity of impairment varying with the amount of cerebral damage (Augmentative Communication Incorporated (ACI), 2007). Other sources of dysarthria include multiple sclerosis, Parkinson's disease, myasthenia gravis (i.e., blocked acetylcholine receptors), and amyotrophic lateral sclerosis (Kent and Rosen, 2004).

Neurological causes of dysarthria involve damage to the cranial nerves that control the articulatory musculature of speech (Moore and Dalley, 2005). For example, damage to the recurrent laryngeal nerve typically reduces control over vocal fold vibration (i.e., phonation), resulting in either guttural or grating raspiness. Inadequate control of soft palate movement caused by disruption of the vagus cranial nerve may lead to a disproportionate amount of air being released through the nose during speech (i.e., hypernasality). More commonly, a lack of tongue and lip dexterity often produces heavily slurred speech and a more diffuse and less differentiable vowel target space (Kent and Rosen, 2004). The lack of articulatory control often leads to various involuntary sounds caused by velopharyngeal or glottal noise, or noisy swallowing problems (Rosen and Yampolsky, 2000). Dysarthria is differentiated from *apraxia*, in which damage to Broca's area in the left frontal lobe reduces the ability to plan rather than to execute speech articulation.

The following subsections describe common phenomena in dysarthric speech, including abnormal speaking rates, fatigue, disfluency, and reduced control of volume, articulation, and pitch.

[2]The earliest record of a scientific understanding of cerebral palsy dates from 1861 when Dr. William John Little described a systematic condition in children characterized by "spastic rigidity of the limbs of new-born children, [and] spastic rigidity from asphyxia neo-natorum" (Little, 1861). This condition gradually came to be known as Little's disease, later generalized to incorporate speech and swallowing difficulties (Posey, 1923), and later still redefined as spastic displegia — a type of cerebral palsy.

2.2.1 Abnormal speaking rates

Dysarthric speech is often between 10 and 17 times slower than regular speech, at about 15 words per minute in the most severe cases (Patel, 1998). Apart from being more laborious for the speaker and listener, slow speech has several acoustic consequences. For example, monosyllabic words that are prolonged by lengthened voiced phonemes (e.g., vowels) are frequently misinterpreted as multisyllabic by human listeners (Kent and Rosen, 2004). Also, if lengthy occlusions precede voiceless plosives such as $/k/$, $/p/$, or $/t/$, listeners often mispartition a single word into two (Raghavendra, Rosengren, and Hunnicutt, 2001). Despite a great amount of inter-speaker variability, dysarthric individuals who can maintain a regular speaking rate are able to repeat individual speech units with fairly normal consistency (Kent and Rosen, 2004).

Abnormally slow speaking rates have been shown to expand the acoustic vowel space, leading to increased intraspeaker variability for those speakers, and more difficult differentiation between phonemes (Kent and Rosen, 2004). Tsao et al. contest the significance of this vowel space expansion in general, but agree that the acoustics of speech is far more variable among slow speakers, including higher interspeaker variability within that group (Tsao, Weismer, and Iqbal, 2006). Simple alterations of speaking rate alone, however, do not account for all unintelligibility of dysarthric speech (Hammen, Yorkston, and Minifie, 1994).

2.2.2 Muscle fatigue and weakness

Low endurance of the facial muscles is often associated with dysarthria, and may be caused by deficiencies at different points of the neuromotor process. Reduced lip and tongue strength and tongue endurance have been associated with Parkinsonism (Solomon, Robin, and Luschei, 2000), stroke (Thompson, Murdoch, and Stokes, 1995), myasthenia gravis (Weijnen et al., 2000), and traumatic brain injury (Goozee, Murdoch, and Theodoros, 2001). Muscle weakness may also limit the amount of air these speakers can release, therefore reducing acoustic energy.

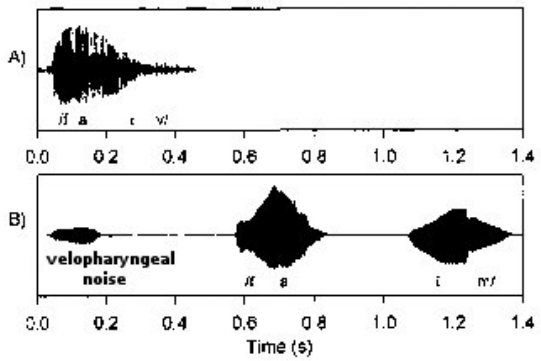


Figure 2.4: Example waveforms of typical (A) and cerebral palsied (B) pronunciation of the word *five* (adapted from Chen and Kostov (1997)).

Despite a clear correlation between dysarthria and facial muscle weakness (Umapathi et al., 2000), the acoustic consequences of that weakness may be less important than other features of disordered speech (Goozee, Murdoch, and Theodoros, 2001). McHenry and Liss (2006) suggest that temporal and spatial inconsistencies of hypokinetic and ataxic dysarthria influence acoustic perception more than increased hypernasality caused by velopharyngeal weakness.

2.2.3 Intense acoustic disfluency

The lack of articulatory control in dysarthria often leads to various involuntary sounds caused by velopharyngeal or glottal noise, or noisy swallowing problems (Rosen and Yampolsky, 2000). Figure 2.4 shows examples of both involuntary noise and involuntary pausing in dysarthric versus normal pronunciations of the word *five*, resulting in two insertion errors and a substitution error in the former [3].

Other types of disfluency commonly associated with dysarthria include hesitation (e.g., false-starts), stuttering, and other involuntary repetition, although these may sometimes result from higher-level linguistic causes (Kent, 2000). These sorts of disfluencies produce severely atypical phrasing which is difficult to understand at the utterance level.

[3]The problem is compounded by a mispronunciation of the labiodental $/v/$ as the labial $/m/$.

2.2.4 Reduced control of articulation and pitch

The most common dysarthric mispronunciations tend to occur with more complex requirements on articulatory movement, namely consonants or consonant clusters. Thubthong et al. (2005) report that among 18 children with CP, word-initial consonants were the most difficult to pronounce, with only a 62.2% rate of accuracy. Of these, alveolar consonants were the most troublesome, with $/r/$ and $/t/$ being correctly articulated 0% and 27.8% of the time, respectively. Vowels and word-final consonants were the most accurately articulated phoneme classes, at 93.7% and 77.1% accuracy respectively. Groups of clustered consonants such as $/tr/$ or $/kw/$ were produced correctly only 11.1% of the time.

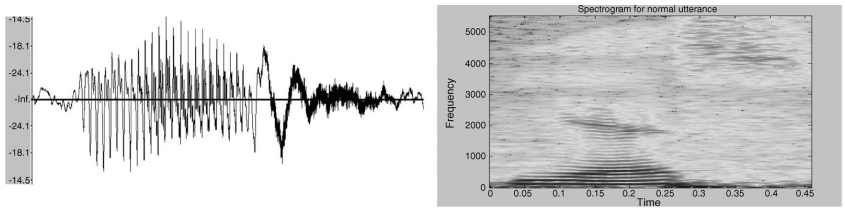
Figure 2.5 shows pronunciation of the word *yes* by both a control and a cerebral palsied individual. The reduced precision of the fricative $/s/$ is likely caused by insufficient jaw movement, and the prolonged duration is almost exclusively due to an extended vowel. Interestingly, the distribution of formants in figure 2.5(b) suggests a pronunciation closer to $/i/$ (O'Shaughnessy, 2000) than to ε as in figure 2.5(a).

Kim et al. (2010) found that speakers with spastic cerebral palsy have drastically reduced displacement of the tongue tip, in position and in velocity. They also found that directed motion of the tongue occurred later than voicing onset relative to the general population, which loosely supports other work that suggest a general difficulty in co-ordinating glottal and supraglottal systems in dysarthria (Chen and Stevens, 2001).

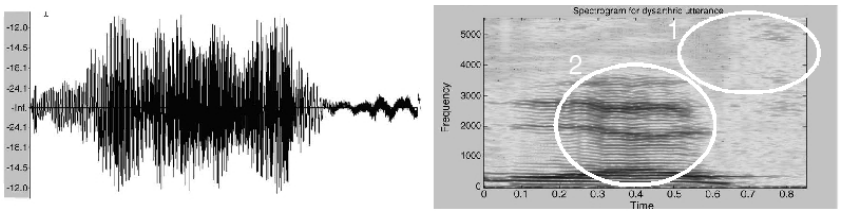
Pitch prosody

Prosody includes changes in fundamental frequency F_0 caused by voluntarily tensing the vocal folds to stress or decline certain syllables for syntactic effect. Proper control of the pitch aspect of prosody has several positive effects on intelligibility, and is also an important conveyor of semantic and emotional content (O'Shaughnessy, 2000).

Dysarthria often reduces voluntary control of the larynx, reducing or diminishing prosody and resulting in machine-like speech (Mori et al., 2005). Despite this reduction, however,



(a)



(b)

Figure 2.5: Example waveforms (dB vs time) and spectrograms (frequency vs time) of (a) typical and (b) cerebral palsied pronunciation of the word *yes* (from Polur and Miller (2006)). Note in the latter the suppression of the /*s*/ fricative in region 1, and the misplaced formants in region 2.

dysarthric speakers retain at least the ability to reliably form differentiable questions and statements using binary vocal pitch contours (Patel, 2002b; Patel, 2002a). Kim, Hasegawa-Johnson, and Perlman (2010) suggest that dysarthric speakers use pitch and intensity cues of lexical stress to a greater degree than non-dysarthric speakers, especially in words that emphasize the second syllable. Inappropriate pitch prosody may affect up to 50% of suspected dysarthric children (Ziegler and Maassen, 2004), but causes can either be physical or learned (e.g., through compensatory behaviour).

2.2.5 Classifying dysarthria

Some of the more clearly defined subgroups of dysarthria include the following:

Spastic Due to upper motor neuron lesions, pyramidal tract damage, and especially lesions to the facial and hypoglossal cranial nerves for jaw and tongue movement, respectively. Phonation is harsh, and strained with low sustained pitch (Duffy, 2005). Hypernasality often accompanies phonemes /p/, /b/, /s/, and /k/. Bursts of loudness, slow rate of speech, and reduced onset time distinction between voiced and unvoiced stops are also associated with spastic dysarthria (Hasegawa-Johnson et al., 2006b).

Hyperkinetic Due to lesions in basal ganglia, and often accompanies other involuntary movement. Harsh phonation is comparable to spastic dysarthria, although hypernasality is more common and involuntary movements tend to superimpose on voluntary articulations. Slowness is also common.

Hypokinetic Associated with Parkinsonism, and due to lesions in the basal ganglia, or to either anti-psychotic medication or blows to the head. Hypokinetic dysarthria results in mono-pitch hoarse phonation with very low monotonous volume. Compulsive syllabic repetition (pallilalia) can also occur. It can result in difficulty initiating voluntary speech, or sudden interruption of movement during speech (Duffy, 2005).

Ataxic Caused by damage to cerebellar control of respiration, phonation, and articulation, but is chiefly characterized by pronounced bursts of loudness. Equal and excessive stress on each spoken syllable is also common. Discoordination results in slurred and slow speech, where patients sound as if inebriated (Duffy, 2005).

Flaccid Caused by damage to the lower motor neurons. May result in complete paralysis of one or more vocal folds, causing breathiness, low volume, increased nasality and monotonous pitch. In unilateral paralysis the jaw may deviate to the weakened side while the tongue moves towards the stronger side, sometimes resulting in drooling (Duffy, 2005).

Despite many overlapping behaviours between these categories, there may also exist some clear delineations. For example, Ozawa et al. (2001) have shown that slow speech in spastic dysarthria is more often caused by lengthened syllables, relative to ataxic dysarthria, which is categorized by longer pauses. Nishio and Niimi (2001) reach a similar conclusion, although they focus on flaccid and hypokinetic dysarthria as predictors for longer pauses. Even if invariant distinctions exist between types of dysarthria for certain features, it is not clear that prior knowledge of differing neuromotor deficiencies can be exploited in ASR.

2.2.6 Evaluating and treating dysarthria

Intelligibility quantifies the degree to which an individual's speech is discernible to human listeners, typically by measuring the average accuracy of word-level transcriptions of utterances across groups of naïve listeners (Kent et al., 1989) (Menendez-Pidal et al., 1996; Hasegawa-Johnson et al., 2006a). If speech samples are phonetically balanced, one can automatically classify the most prevalent errors according to discrete phonetic features of how the vocal tract restricts airflow (manner), where along this tract the narrowest constriction occurs (place), and whether the vocal folds vibrate during production (voicing)[4]. Other procedures that measure intelligibility include the Children's Speech Intelligibility Measure which includes developmental statistics, and the Yorkston-Beukelman-Traynor assessment (Hammen, Yorkston, and Minifie, 1994) which has been computerized and includes factors such as speaking rate and rate of intelligibility. Intelligibility scores are also sometimes accompanied by results of the Frenchay Dysarthria Assessment (Enderby, 1983) that individually scales the strength of the various articulators, respiration, reflex, and rate (Menendez-Pidal et al., 1996).

Since dysarthrias cannot yet be cured with surgery or medication, behavioural interventions are often used to strengthen the articulatory muscles or develop alternate pronunciation strategies to improve intelligibility (Kent, 2000). This behavioural intervention often involves computer-based treatment that can improve intelligibility by exercises and feedback automati-

[4]These measures may be overly simplistic, but are useful in classification (O'Shaughnessy, 2000).

cally generated using speech recognition (Thomas-Stonell et al., 1998) that is just as effective as traditional treatment (Palmer, Enderby, and Hawley, 2007).

The precise relationship between speech repeatability and neurological damage is still an open question. For instance, although dysarthric utterances are highly variable, those dysarthric speakers who have maintained a regular speaking rate appear to be able to repeat individual speech units in isolation with fairly normal reproducibility (Kent and Rosen, 2004; Chen and Kostov, 1997). Furthermore, although intelligibility strongly correlates with recognition accuracy in ASR (Ferrier et al., 1995), consistency does not (Thomas-Stonell et al., 1998).

2.3 Automatic speech recognition (ASR)

The goal of ASR is to decide on the optimal word sequence $W = w_1\ w_2\ ...\ w_n$ to describe an acoustic input speech signal X:

$$W_c = \arg\max_W \frac{P(W)P(X|W)}{P(X)} \tag{2.1}$$

where $P(W)$ and $P(X|W)$ are the optimal language and acoustic models, respectively. The input speech signal, X is typically measured at a constant sampling rate, where the i^{th} measurement, $x[i]$, is quantized according to a constant bit rate (e.g., 8 bit values range from -128 to 127, 16 bit values from -32768 to 32767). This discretized signal is first converted to an alternative form more amenable to machine learning via feature extraction and it is upon the space defined by these features that statistical models are trained and used in classifying sounds, as summarized below.

2.3.1 Feature extraction

Although certain aspects of speech can be identified directly from this superpositional waveform representation (e.g., energy, pitch, broad phoneme classes), most of the information that distinguishes phonemes from one another is found in the relative intensities of the compo-

nent waveforms at their respective oscillating frequencies (Stevens, 1998). As the oscillating frequencies of these waveforms increase, their respective amplitudes become weaker, due to glottal pulse becoming attenuated by its interaction with the vocal tract walls. In order that the information contained at these frequencies are more accurately encoded by the acoustic model, the first stage of feature extraction is typically pre-emphasis, where the signal $x[n]$ is transformed to signal $\hat{x}[n] = x[n] - \alpha x[n-1]$ for some empirically determined parameter $0.9 \leq \alpha \leq 1.0$.

In order to extract spectral information, windows of several consecutive samples must be collectively analyzed. The width of these windows must include enough consecutive samples so that they can encapsulate two complete oscillations of the lowest-frequency waveform to be considered. This requirement of the sampling rate relative to the length of the component waveforms in time is variously referenced with regards to the Nyquist rate (Jurafsky and Martin, 2009). Since human speech mostly occupies the frequencies between 100 Hz and 10 kHz (Stevens, 1998), the minimum window length must be $\approx 2/100\text{Hz} = 0.02\text{s}$. Rarely are analysis windows wider than this, due to the non-stationary nature of the speech signal. In order for spectral features to be extracted, however, stationarity within these windows is assumed to be inviolate.

Consecutive windows do not cover mutually exclusive segments of the audio. Indeed, throughout this thesis consecutive windows overlap in time over half of their lengths (e.g., if analysis windows are 0.16 s wide, each window begins 0.08 s after its predecessor). This offset accounts for rapid changes in the speech signal. Moreover, since simple segmentation of the audio signal results in abrupt cuts to the signal, which can negatively influence proper feature extraction, each window is modified so that the signal tends to 0 at the boundaries of each window according to the popular Hamming window method (Quatieri, 2002), where

$$w[i] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi i}{N}\right) & 0 \leq i \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where N is the number of discrete samples in the analysis window. The discrete Fourier trans-

form then converts this amplitude/time representation to its associated amplitude/frequency representation over the pre-emphasized windows obtained in this manner, $w[n]$,

$$X[k] = \sum_{n=0}^{N-1} w[n] e^{\frac{-j2\pi nk}{N}}, 0 \leq k \leq N. \tag{2.3}$$

This transform depends on Euler's formula $e^{j\theta} = \cos\theta + j\sin\theta$ for imaginary unit j. In this work, the discrete Fourier transform is computed by the fast Fourier transform (FFT) algorithm using a recursive decimation-in-time[5] algorithm that forcibly assumes that N is a power of 2. The FFT computes the frequencies $X[k]$ of a signal in $O(N \lg N)$ time complexity, improving on $O(N^2)$ complexity of linear sequential computation. These frequencies are then analyzed within perceptually-motivated models that imitate the behaviour of the human cochlea (and the neuronal membrane therein) by means of non-linear scaling functions that warps signal frequencies f to be more amenable to feature extraction (Huang, Acero, and Hon, 2001). In this work, occasionally the Bark scale is used

$$Bark(f) = 13\arctan(7.6 \times 10^{-4} f) + 3.5\arctan\left(\left(\frac{f}{7.5E^3}\right)^2\right), \tag{2.4}$$

but in general the spectra resultant from FFT are scaled according to the mel scale. A mel is a unit of pitch that describes the distance between sounds adjacent in their perceptual pitch (Stevens, Volkman, and Newman, 1937). Frequencies obtained by FFT below around 1000 Hz are mapped linearly to the mel scale, and those above 1000 Hz are mapped logarithmically according to

$$mel(f) = 1127\ln\left(1 + \frac{f}{700}\right). \tag{2.5}$$

This scale is analogous to the human loss of sensitivity to pitch differences at higher frequencies. The mel-scaled spectrum contains several aspects of the glottal source of speech that are not particularly useful in distinguishing between phonemes. For example, the spectrum includes fundamental frequency and energy information, which is not as important to speech recognition as details of the filter, i.e., the vocal tract. In order to deconvolve the source from

[5]Decimation in time refers to splitting into sums over even and odd time indices for the purposes of recursion.

the filter, the first step is to take the logarithm of the magnitude spectrum obtained in equation 2.5. The final step visualizes this log spectrum as if it were itself a waveform and takes into consideration that the shape of this pseudo-waveform is characterized by high-frequency oscillations caused by the fundamental frequency, and otherwise by broad peaks and valleys. It is these high-frequency oscillations that correspond to the glottal source, and these broad peaks and valleys that correspond to the shape of the vocal tract (and formant frequencies, generally). Taking the spectrum of the log spectrum separates these two components and has the added benefit that the resulting coefficients are uncorrelated (unlike the spectrum), so that acoustic models used in classification do not have to encode covariances between all features, which reduces the number of parameters necessary in machine learning (Jurafsky and Martin, 2009). This spectrum of the log spectrum, or 'cepstrum', is converted from the windowed speech $w[n]$ to Mel-scaled cepstral coefficients $c[k]$ by

$$c[k] = \sum_{n=0}^{N-1} \log \left(H_k(n) \left| \sum_{l=0}^{L-1} w[l] e^{-j\frac{2\pi}{L}nl} \right| \right) e^{j\frac{2\pi}{N}kn} \tag{2.6}$$

where $H_k(n)$ is the magnitude of the m^{th} filterbank evaluated at the n^{th} linear frequency.

Linear predictive coding (LPC)

Linear Predictive Coding (LPC, also known as autoregressive modelling) estimates the main features of speech using filters $H(z)$ where

$$H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}, \tag{2.7}$$

and $A(z)$ is the inverse filter and $X(z)$ is the z-transform,

$$X(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \frac{1}{1 - az^{-1}}. \tag{2.8}$$

LPC p^{th}-order analysis then predicts the current sample as a linear combination of its past p samples:

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k]. \tag{2.9}$$

Formant candidates in the spectrum can be obtained in each frame by computing the roots of the p^{th}-order LPC polynomial $A(z)$ (Ahadi-Sarkani, 1996) using such methods as Laguerre's method. The i^{th} root can be represented by

$$z_i = \exp(-\pi b_i + j2\pi f_i) \tag{2.10}$$

giving the formant frequency (f_i) and bandwidth (b_i). In order to track formants over time and to decide between many formant candidates, a dynamic programming algorithm is combined with an *a priori* state-based formant transition model (Huang, Acero, and Hon, 2001).

Comparing feature extraction on dysarthric speech

Jayaram and Abdelhamied (1995) compared the effects of two feature sets on the classification of cerebrally palsied dysarthric speech. Their experiments used a single speaker with a limited 22-word vocabulary to compare 128-point FFT coefficients (reduced to 16 by the Turning-point algorithm) against 43 LPC coefficients describing the frequency, amplitude, and bandwidth of formants in each frame. Each feature set provided the input to fully connected feed-forward neural networks, with inputs representing the features for 15 consecutive frames. The network trained on FFT features achieved 76.3% accuracy while the one trained on LPC formant features only achieved 42.5%.

A possible explanation for the discrepancy between FFT and LPC formants that Jayaram and Abdelhamied did not mention is that by not modelling non-formant aspects of consonants, such as the high energy of fricatives or plosives, the LPC network effectively ignores those phonemes which are most troublesome in dysarthric speech, namely the consonants[6]. How-

[6]Formant tracking generally gives very large-bandwidth formant estimates to areas without voiced speech (Huang, Acero, and Hon, 2001).

ever, the direction of the F_2 slope following a plosive has been shown to accurately identify the place of articulation of that plosive (O'Shaughnessy, 2000).

Jamieson et al. compared the results of various coders on the intelligibility of a range of dysarthric speech. Of all lossy coders, GSM 06.10[7] resulted in the most intelligible speech, reducing relative error by as much as 57% in the most severe cases over simple LPC-reconstructed speech (Jamieson et al., 1996).

2.3.2 Classification

Dynamic Time Warping (DTW) can be used to compute the distortion between two speech samples, and hence can be used to identify the phoneme whose template is the least distorted from some unlabelled input. Given vectors of frames representing the reference $R = \langle x_1\ x_2\ ...\ x_n \rangle$ and target $T = \langle y_1\ y_2\ ...\ y_m \rangle$ templates, DTW computes a warping function $m = w(n)$ mapping the time axis n of R onto m of T, as shown in Algorithm 1. This warping function depends on two cost functions. The first is a Euclidean distance function $d(i,j)$ between $i \in R$ and $j \in T$. The second is a penalty $\rho(a,b)$ applied to alignment jumps of a steps in R and of b steps in T (i.e., from alignment (x_i, y_j) to (x_{i+a}, y_{j+b})). Except where noted, in this dissertation $\rho(1,1) = \rho(0,1) = \rho(1,0) = 0$, and $\rho(a,b) = \infty$ for other combinations. Here, the total distance between two sequences is $D(n,m)$ and the warping function is described by the state sequence $(s_1, s_2, ..., s_l)$ where $s_l = (n,m)$ and B is the backtrace matrix where $B(i,j) = (c,d)$ indicates that if pair (x_i, y_j) are aligned, the best previous alignment is (x_c, x_d).

Although this approach treats durational variations as noise to be smoothed out, the extent to which it can overcome the temporal variability in dysarthric speech is unknown. Frequency warping can also be used for classification if the distance function $d(i,j)$ is somehow aware of spectral formant positions, (O'Shaughnessy, 2000).

[7] The Global System for Mobile communications (GSM) with the European Telecommunications Standards Institute (ETSI) fixed-rate 06.10 speech codec.

Algorithm 1: DYNAMIC TIME WARPING

Data: Reference $R = \langle x_1\ x_2\ ...\ x_n \rangle$ and target $T = \langle y_1\ y_2\ ...\ y_m \rangle$ templates.

begin

$D(1,1) = d(1,1), B(1,1) = 1$, and $\forall 2 \leq j \leq M$, $D(1,j) = \infty$

for $i = 2,...,N$ **do**

for $j = 1,..,M$ **do**

$$D(i,j) = \min_{1 \leq q \leq i} \min_{1 \leq p \leq M} [D(q,p) + \rho(i-q, j-p) + d(i,j)]$$

$$B(i,j) = \arg\max_{1 \leq q \leq i} \arg\max_{1 \leq p \leq M} [D(q,p) + \rho(i-q, j-p) + d(i,j)]$$

end

Hidden Markov models (HMMs)

A more popular alternative classification mechanism, hidden Markov models (HMMs), categorizes observable temporal data sequences according to 'hidden' statistical parameterizations and an underlying connected-state structure. Unless otherwise noted, this document refers to continuous HMMs that are defined by a multi-dimensional continuous observation space O with $\mathbf{o}$ being a sequence of length T of observation vectors $o_t \in O$ for $t = 1 \ldots T$ [8], a state space Q (where q_t is the state at time t), an initial state distribution $\pi_i = P(q_0 = i)$, a state transition matrix $A(q_i, q_j)$ describing the *a priori* probability of transitioning from state q_i to q_j, and a distribution $B_i(o)$ defining the probability of observing vector $\mathbf{o}$ in state i. Typically in ASR, the distribution $B_i(o)$ will be a mixture of Gaussians, i.e.,

$$B_i(o) = \sum_{m=1}^{M} \omega_{i,m} \frac{1}{(2\pi)^{d/2} |\Sigma_{i,m}|^{1/2}} \exp\left[-\frac{1}{2} (o - \mu_{i,m})^\top \Sigma_{i,m}^{-1} (o - \mu_{i,m}) \right] \tag{2.11}$$

where d is the number of dimensions in each observation, $|\Sigma|$ is the determinant of Σ, and there are M component Gaussians in each state, $\omega_{i,m}$ is the weight of the m^{th} Gaussian in state i, $\mu_{i,m}$ is its mean, and $\Sigma_{i,m}$ is its covariance.

[8] This is analogous to the observation alphabet in discrete HMMs.

Weights in a Gaussian mixture are subject to the condition

$$\sum_{m=1}^{M} \omega_{jm} = 1. \tag{2.12}$$

The complete parameter set of an HMM, Φ, constitutes all parameters of $a_{ij} = A(q_i, q_j)$, $B_i(o)$, and π_i. In some cases, we are interested in computing the likelihood of a particular observation o given the parameters Φ. This is performed by the Forward algorithm (Huang, Acero, and Hon, 2001),

$$P(\mathbf{o};\Phi) = \sum_{\forall \mathbf{q}} P(\mathbf{q};\Phi) P(\mathbf{o} \,|\, \mathbf{q};\Phi), \tag{2.13}$$

which sums over all possible sequences of hidden states $\mathbf{q}$. Here, the probability of a particular state sequence and the probability of an observation given that state sequence are

$$\begin{aligned} P(\mathbf{q};\Phi) &= P(q_1;\Phi) \prod_{t=2}^{T} P(q_t \,|\, q_{t-1};\Phi) \\ P(\mathbf{o} \,|\, \mathbf{q};\Phi) &= \prod_{t=1}^{T} P(o_t \,|\, q_t;\Phi). \end{aligned} \tag{2.14}$$

More typically, we are interested in finding the state sequence that gives the highest probability given an observation sequence. This is referred to as 'decoding' and is useful since the state sequence reveals the most likely phoneme or word sequence, assuming that individual smaller HMMs, each representing a word or phoneme, are concatenated together with appropriate bigram state transition probabilities. The Viterbi algorithm is used in decoding and determines the most likely state sequence to represent an observation, given an HMM's parameters. It is summarized in algorithm 2. This algorithm can be generalized to produce a ranking of the n state sequences that give the highest probability given the observation (i.e., an 'n-best list'). For the purposes of computational parsimony, this list is often approximated by generalizing the Viterbi algorithm to withhold only a limited number of state sequences up to a particular time. This is often referred to as 'beam search'.

Algorithm 2: VITERBI

Data: Observation sequence **o** (length T) and an HMM with $\Phi = \{A, B, \pi\}$ and Q states.

begin

for $i = [1 \dots Q]$ **do**

$V_1(i) = \pi_i B_i(o_1)$ for $i = [1 \dots Q]$; // Initialization

$\beta_1(i) = 0$

for $t = [2 \dots T]$ **do**

for $j = [1 \dots Q]$ **do**

$V_t(j) = \max_{1 \le i \le Q} \left[v_{t-1} a_{ij} \right] B_j(o_t)$; // Induction

$\beta_t(j) = \arg\max_{1 \le i \le Q} \left[v_{t-1}(i) a_{ij} \right]$

$bestScore = \max_{1 \le i \le Q} [V_T(i)]$; // Termination

$q_T^* = \arg\max_{1 \le i \le Q} [B_T(i)]$

for $t = [T-1, T-2, \dots, 1]$ **do**

$q_t^* = B_{t+1}(s_{t+1}^*)$; // Backtracking

return $Q^* = (q_1^*, q_2^*, \dots, q_T^*)$; // This is the best sequence

end

In order for HMMs to model phonetic rather than lexical acoustics, three-state left-to-right structures are usually used, where the probabilities of $B(o,s)$ are modelled by multivariate GMMs[9].

2.3.3 Computer-assisted interaction

The neurological damage that causes dysarthria usually affects other physical activity, such as slowing keyboard input 150 to 300 fold in severe cases compared with regular users (Hosom et al., 2003; Hux et al., 2000). Since dysarthric speech is often only 10 to 17 times slower than normal (section 2.2.1), ASR is seen as a viable alternative to improve communicativity in computer-assisted interaction.

Alternative augmentative communication (AAC) devices include speech-synthesis machines in which touch-panels or keyboards are used to access and concatenate recorded speech utterances to output sentences, such as the Dynavox 3100 ($7,300 CAD) and DeltaTalker ($12,000 CAD) (Messina and Messina, 2007). Some of these devices include a 'scanning' interface where a button is pressed to iteratively cycle through a list of alternative commands, words, or phrases (Hawley et al., 2007).

Assistive technology for the disabled requires a careful combination of clinical study and engineering in order to accommodate particular disabilities. The complexity and specificity of these requirements may have limited the widespread adoption of such technology (Noyes and Frankish, 1992) which necessitates sophisticated adaptability beyond the training of the acoustic model $P(X|W)$. For example, automatic adjustments to the dialogue flow of an ASR system may be necessary in cases where neurological impairment may make instructions difficult to follow (Hux et al., 2000). In cases where disfluency or slow speech would result in faulty word detection, pressing a physical switch can be used to invoke ASR over the duration of an utterance (Rosen and Yampolsky, 2000), but only when the speaker has enough physical endurance.

[9] Other probabilistic alternatives are discussed later.

Hawley et al. (2007) described an experiment in which 8 disabled users (with either cerebral palsy or multiple sclerosis) controlled non-critical devices in their home (e.g., TV, hi-fi) with ASR. Arrays of microphones were placed between 0.5 and 3.0 m from the usual sitting positions of the participants, and command vocabularies consisted of very simple phrases (e.g., *TV channel up*, *Radio volume down*)[10]. Feedback was provided either through visual displays or by auditory cues. This ASR-based environmental control was compared with a 'scanning' interface. While the ASR interface made more errors (77.3 to 96.9% pre-training, 90.8 to 100% post-training accuracy) than the scanning interface (100% accuracy), the former was significantly faster (7.7s vs 16.9s, on average)[11]. Participants commented that speech control was significantly less tiring than the scanning interface, although subjective preference between the two interfaces was evenly split (Hawley et al., 2007). Similar results were obtained by studying the STARDUST command-and-control system (Green et al., 2003).

One alternative to both speech and keyboard AAC that avoids issues of accuracy in the former and fatigue in the latter is to use controlled durations of the vowel /a/, perhaps with some binary pitch control, to communicate Morse code text to a speech synthesis device (Patel, 2002a). This places some burden on the dysarthric speaker to learn or consult tables of rules, and would not be necessary if ASR accuracy were improved.

Word prediction

Word- or phrase-based prediction can assist cognitively or physically disabled users type text by allowing them to select among suggested completions as they write. This assistive technology has reduced the number of keystrokes required of an individual by as much as $\sim 69\%$ in adaptive-lexicon systems (Swiffin et al., 1987; Matiasek, Baroni, and Trost, 2002), thereby increasing communication speed and allowing improved individual expression (Alm, Arnott,

[10]The types of phrases used would be neatly described by a simple grammar, but the authors appear not to use one.

[11]This comparison may be doubly misleading, however, since the scanning interface requires scanning over irrelevant commands, and higher accuracy can *sometimes* be improved by adding ASR to a scanning interface (Havstam, Buchholz, and Hartelius, 2003).

and Newell, 1992). Prediction is especially valuable to those for whom fatigue or frustration often accompany attempts at communication (Garay-Vitoria and Abascal, 2006).

The current word w_i can be anticipated given an n-gram context augmented by part-of-speech tags t_j. For example, Fazly and Hirst (2003) describe an algorithm that ranks possible completions based on the estimate

$$\begin{aligned} P(w_i|w_{i-1},t_{i-1},t_{i-2}) &\approx \sum_{t_i \in T(w_i)} P(w_i|w_{i-1},t_i)P(t_i|t_{i-1},t_{i-2}) \\ &\approx \sum_{t_i \in T(w_i)} \frac{P(w_i|w_{i-1})P(t_i|w_i)}{P(t_i)} P(t_i|t_{i-1},t_{i-2}) \\ &= P(w_i|w_{i-1}) \sum_{t_i \in T(w_i)} \frac{P(t_i|t_{i-1},t_{i-2})P(t_i|w_i)}{P(t_i)} \end{aligned} \tag{2.15}$$

where $T(w_i)$ is the set of all possible PoS tags associated with word w_i. Combining PoS with lexical context in this way reduces the percentage of keystrokes needed to produce text by $\sim 6\%$ over purely *a priori* statistical methods (Fazly and Hirst, 2003). Other extensions to text prediction to further refine the list of hypothesized completions include the use of grammatical syntax and semantics (Li and Hirst, 2005; Erdogan et al., 2005), as well as trained neural networks (Garay-Vitoria and Abascal, 2006).

Empirically observed improvements in the rate of typed communication with prediction might not overcome improvements gained through the use of speech (see above), but applying the same approach to predicting spoken communication may reduce the amount of effort required for both the dysarthric speaker and their audience. If speech input is coupled with a visual display for output, for example, that display could be updated 'on-the-fly' with the results of predicted queries before those queries are completed.

2.4 Representations for speech production

Articulatory features (AFs) are quantized abstractions of speech production according to distinctive configurations of the vocal tract [12]. They provide an inventory of the types of sounds humans can produce (O'Shaughnessy, 2000; Huang, Acero, and Hon, 2001). The study of AFs in recent phonetics dates back at least to Chomsky and Halle (Chomsky and Halle, 1968), who represented sounds of speech as vectors of binary features (e.g., nasal/non-nasal, voiced/voiceless). That work showed that some context-sensitive phonetic variation could be specified by transformational rules based on phoneme sequences and syntactic trees (e.g., $/p/$ is aspirated if it begins a syllable onset consonant cluster, as in *prim*, but not aspirated if it ends that onset, as in *spin*).

Here, articulatory features are collected into seven categories, each with a number of possible values. For example, a segment of speech can be concurrently voiced, nasal, and static, which represent values for three distinct features. Parallelizing streams of information in this manner allows asynchronous modulation of speech acts across phoneme boundaries, which can partially account for co-articulation effects and speaker variability (Livescu et al., 2007), which are particularly exacerbated in dysarthric speech. Other useful properties reported of AFs include language-independence and reliable recovery from acoustics among regular speakers (Frankel, Wester, and King, 2007). The features used here are based on those of Wester (Wester, 2003; Scharenborg, Wan, and Moore, 2007) and are listed in table 2.1.

In the absense of AF annotations, AF values can be derived directly from phoneme annotations. In this study, we assign to each MFCC frame of data a 7-dimensional vector of AF values based exclusively on the phoneme annotation at that frame. This assignment is derived directly from the phoneme-to-AF transformation table in Frankel et al. (Frankel, Wester, and

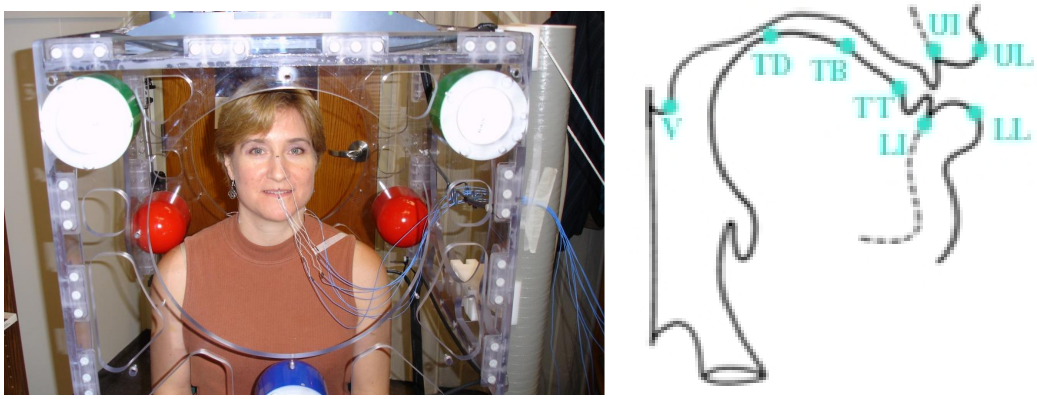
[12] Articulatory features are sometimes called *phonological features* in the literature (e.g., by Clements (Clements, 1985) and by King and Taylor (King and Taylor, 2000)). However, the latter term has largely been superseded by the former in the literature (e.g., by Kirchhoff (Kirchhoff, 1999) and by Metze (Metze, 2007)). In this document, the term *articulatory feature* must be differentiated from articulatory *measurements*, which refer to direct recordings of the vocal tract.

Feature	Description (*and values*)
Manner (**M**)	high-level categorization of speech sound
	approximant, fricative, nasal, retroflex, silence, stop, vowel
Place (**Pl**)	location of primary constriction
	alveolar, bilabial, dental, labiodental, velar, silence, nil
High/Low (**HL**)	anterior position of the tongue
	high, mid, low, silence, nil
Front/Back (**FB**)	ventral position of the tongue
	front, central, back, nil
Voice (**V**)	presence/absence of glottal vibration
	voiced, unvoiced
Round (**R**)	circularity of the lips
	round, non-round, nil
Static (**S**)	movement of articulators (e.g., diphthong)
	static, dynamic

Table 2.1: Articulatory features, a description of their characteristics, and their possible values.

King, 2007). This incorporates recommendations by Wester et al. (Wester, Frankel, and King, 2004) in which the Front/Back feature includes the normally excluded *central* value, and diphthongs are split in half into their component vowels, which are mapped to their corresponding AFs. Unlike Frankel et al. (Frankel, Wester, and King, 2007), we label the Place feature of phonemes /*b*/ and /*m*/ as bilabial rather than labiodental to distinguish these from fricatives /*f*/ and /*v*/.

A more empirical approach to production knowledge is derived from direct measurement of the vocal tract during speech with semi-invasive procedures such as electromagnetic articulography (EMA), magnetic resonance imaging (MRI), X-ray microbeam analysis (Westbury, 1994), or electropalatograph. These procedures capture motions of external (e.g., lips) and internal (e.g, tongue, velum) actuators with sufficient temporal and spatial resolution to accurately reconstruct physical activity (van Lieshout et al., 2007). Electromagnetic articulography uses alternating electromagnetic fields generated by a cube that surrounds the speaker's head to infer the articulators at a rate of 200 Hz to 500 Hz, usually to within an error of 0.5mm (Yunusova, Green, and Mefferd, 2009). These systems produce no audible noise, and the coils do not interfere with regular speech. Figure 2.6 shows typical configurations of the EMA cube and the placement of the receiver coils.



(a) (b)

Figure 2.6: Example configuration of electromagnetic articulography. Subfigure (a) shows a subject connected within the recording environment, and subfigure (b) shows the typical locations of receiver coils on the midsagittal plane (i.e., **V** velum, **TD** tongue dorsum, **TB** tongue body, **TT** tongue tip, **UI** upper incisor, **LI** lower incisor, **UL** upper lip, and **LL** lower lip).

Chapter 3

Related work

Explicit use of articulatory knowledge is still rare in automatic speech recognition (ASR) despite evidence that it is far more speaker-invariant and less ambiguous than the resulting acoustics (King et al., 2007). For example, the nasal sonorants */m/*, */n/*, and */ng/* are acoustically similar but uniquely and consistently involve either bilabial closure, tongue-tip elevation, or tongue-dorsum elevation, respectively. The identification of linguistic intention would, in some cases, become almost trivial given access to the articulatory goals of the speaker.

There have been a number of attempts at improving speech recognition for dysarthric speakers, and other attempts at integrating articulatory knowledge into ASR, but these two efforts have so far not converged. Section 3.1 describes the state-of-the-art in applying ASR to dysarthric speech. Section 3.2 describes mechanisms in the literature that apply articulatory knowledge to speech recognition generally.

3.1 Recognition with dysarthric speech

Since dysarthria is specifically a motor-control disorder, intuition may suggest that the language model $P(W)$ would be effectively regular, but temporal irregularities such as disfluency and vocal variability make estimation of $P(X|W)$ difficult. However, applications of speech recognition for dysarthric speakers have been pervasive, including automatic dictation of spon-

taneous text (e.g., natural communication) (Havstam, Buchholz, and Hartelius, 2003), telephonic access to services (e.g., ticket reservation), and the local control of machines (e.g., wheelchair, domestic appliances) (Noyes and Frankish, 1992; Hawley et al., 2007).

Early work in applying ASR to individuals with dysarthria almost exclusively involved the use of hidden Markov models (HMMs) whose parameters were trained to the general population. Usually, these involved small-vocabulary recognition tasks with word-recognition rates significantly lower for dysarthric speakers, often at least 26.2% lower than the general population (Coleman and Meyers, 1991). For example, given a vocabulary of 40 words, Rodman, Moody, and Price (1985) report mean word-recognition rates of 58.6% for dysarthric speakers compared with 95% for the general population. Deller and Snider (1990) showed that highly-connected HMMs could be evaluated efficiently in linear time (relative to the number of states) for use with speech from an individual with cerebral palsy.

Hux et al. (2000) report similar results with a control subject and an ataxic survivor of traumatic brain injury in three commercial ASR dictation systems, namely Microsoft Dictation, Dragon NaturallySpeaking (DNS), and Kurzweil Education Systems' VoicePad Platinum. None of these systems modeled grammatical syntax, but instead relied on similar architectures that augmented HMM phonetic modelling with n-gram language models[1]. Microsoft Dictation and DNS both accepted whole sentences, whereas VoicePad Platinum only recognized isolated words, perhaps providing the dysarthric speaker with more time to plan out her articulatory movements in the intervening gaps between words.

Figure 3.1 shows average recognition accuracies of these systems across five trials each. Every trial consisted of 10 constant and pre-selected sentences, followed by 10 novel spontaneous sentences. All systems performed significantly better with regular speech, averaging between 83.4% (Microsoft) and 89.9% (Dragon) word-recognition, compared with between 51.9% (VoicePad) and 64.7% (Dragon) for dysarthric speakers. In all cases, novel sentences

[1]Little technical documentation exists for VoicePad itself, although this product was sold in 1997 to Lernout & Hauspie (Felber, 2001), and rebranded as Voice Xpress, which uses this model (Salleh et al., 2000).

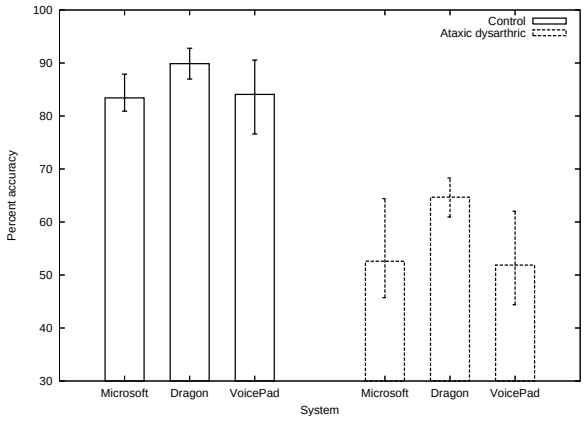


Figure 3.1: Comparison of recognition rates for control and ataxic speakers across Microsoft Dictation, Dragon NaturallySpeaking, and KES VoicePad Platinum, from Hux et al. (2000). Boxes represent average accuracy rates, with errorbars representing minimum and maximum accuracy over 5 trials.

were recognized correctly less frequently than the predetermined ones, but not by a significant margin. Despite periods of 'training' preceding each of the five trials (or during, in the case of VoicePad), these did not result in significant improvement in accuracy. Despite their relatively poor results, however, such commercial ASR systems have been shown to improve accuracy and speed in simple text-entry for physically disabled individuals relative to other modes of input (e.g., scan-and-switch) (Havstam, Buchholz, and Hartelius, 2003; Hawley et al., 2007).

These results show that ASR recognition rates have not improved significantly over a decade prior when Coleman and Meyers found that dysarthric speakers typically have recognition rates between 9.5 and 26.2% lower than normal speakers for isolated-word dictation tasks, on average (Coleman and Meyers, 1991). Several projects have attempted to adapt discriminative models to dysarthric speech without considering the causes or features of dysarthria. These are discussed in sections 3.1.2, and 3.1.3. More recently, attempts have been made to improve ASR rates by focusing on the types of errors made with dysarthric speech. Polur and Miller (Polur and Miller, 2006), for example, produced ergodic HMMs that allow for 'backwards' state transitions. This ergodic structure is meant to capture aspects of dysarthric speech

such as involuntary repetition and disruptions during sonorants (e.g., pauses) and reveals small but definite improvements over the traditional baseline. Morales and Cox (Morales and Cox, 2009) improved word-error rates by approximately 5% on severely dysarthric speech and approximately 3% on moderately dysarthric speech by building weighted transducers into an ASR system according to observed phonetic confusion matrices. The metamodels used in this work are very similar to those used by Matsumasa et al. (2009), described below, except it also involved a language model, albeit one based on the highly restricted Nemours database described in section 4.1. A commonality among all this work is that the actual articulatory behaviour of the dysarthric speech has not been taken into account.

Adapting HMM acoustic models trained to the general population given dysarthric data has also shown to improve accuracy, but not as clearly as training those models exclusively with dysarthric acoustics, especially in the more severe cases (Raghavendra, Rosengren, and Hunnicutt, 2001; Sanders et al., 2002a). In clinical settings, automated methods are increasingly being used to quantify the level of dysarthric severity (Hill et al., 2006; Constantinescu et al., 2010) as automated systems are less costly and less subject to potential bias than human clinicians (Bodt, Huici, and Heyning, 2002; Nuffelen et al., 2009). The accuracy of ASR systems are often used for this purpose (Doyle et al., 1997; Ferrier et al., 1995; Maier et al., 2009).

3.1.1 Adapting acoustic models to dysarthric speech

Improving ASR accuracy usually involves adapting the underlying models to better represent observed data, which normally involves acquiring and transcribing indicative data from the target population, and applying one of many parameter estimation algorithms. Speaker-dependent (SD) models are trained to an individual speaker, and are therefore more accustomed to the peculiarities of any dysarthric individual. In practice, SD models can improve accuracy relatively by 20% to 30% (Huang, Acero, and Hon, 2001), but can restrict use to speakers who have previously trained with the system. Speaker-adaptive (SA) models are initialized by models trained on some non-atomic subset of a target population, but are later trained with a single user. SA

models tend to not be as accurate as SD models given the same amount of user-specific training, but are initially more accurate as they are initialized by real speech data (Huang and Lee, 1993).

Raghavendra, Rosengren, and Hunnicutt (2001) compared what they described as an SA phoneme recognizer and an SD word recognizer on dysarthric speech. They concluded that SA modelling is appropriate for mild or moderate dysarthria, with an empirical relative error reduction (RER) of 22%, but that severely dysarthric speakers are better served by speaker dependence, with 47% RER. Sharma and Hasegawa-Johnson (2010a) contradict these findings, somewhat, in that they find no evidence that the severity of dysarthria is predictive of the relative performance of SD and SA systems. They also conclude that left-right HMMs are better suited to dysarthric speech than transition-interpolated HMMs and that adapting parameters other than transition probabilities give ideal results in this domain. Noyes and Frankish (Noyes and Frankish, 1992) reported SD models attaining between 75% and 99% word-level accuracy for impaired speakers on a small vocabulary, where human listeners could only correctly identify between 7% and 61%. Sawhney and Wheeler (1999) found pronounced gains from SD models, with an RER of $\sim 22\%$ over independent models using an unspecified segmental phoneme recognizer. These experiments, however, used no more than 5 test subjects each, with limited training data. Sanders et al. (2002b) also found that speaker-dependent HMM modelling improved word-error rates by 50% to 100% relative to speaker-independent baselines trained to the Dutch speech corpus, although these experiments involved only two dysarthric subjects, small vocabularies (i.e., between 39 and 336 words) and very small amounts of data (i.e., 8.5 or 12.8 minutes of speech recorded in total from each dysarthric subject). It is difficult to draw reliable conclusions from such small amounts of data.

3.1.2 Support vector machines and dysarthric speech

Hasegawa-Johnson et al. (2006a) compare HMMs and support vector machines (SVMs) within isolated word recognition in a multi-microphone environment for 4 dysarthric speakers with

Vocabulary	45 Words		10 Words (Digits)				
Algorithm	H	HV	H	HV	Word-SVM	WF-WVM	WFV-SVM
Speaker 1	44	55	71	80	97	86	90
Speaker 2	42	49	86	95	70	69	70
Speaker 3	87	89	99	100	90	90	90
Speaker 4	77	80	99	100	97	100	100

Table 3.1: Comparison of word recognition accuracy across 4 speakers with dysarthria of varying intelligibility, and system types (HMMs (H and HV) and SVMs), from Hasegawa-Johnson et al. (2006a).

intelligibility ranging from 19.2% (Speakers 1 and 2) and 29.2% (Speaker 4) to 92.5% (Speaker 3). Across all speakers, word-initial and word-final consonant errors accounted for 27.6% and 30.9% of the errors respectively, with word-medial consonant (18.9%) and vowel errors (22.6%) rounding out the rest [2].

The four speakers each produced 541 phonetically-balanced words which were then processed by 5 systems: The first and second systems use standard triphone HMMs with (< 10) Gaussian observation densities per state and the HTK toolkit (Cambridge, 2007), where the first system (**H**) uses a single microphone and the second (**HV**) trains separate HMMs for each of 7 microphones, and combines their hypotheses using majority voting during recognition. Two types of SVM were trained: the first (**Word-SVM**) uses 10 SVMs that identify individual words given a single microphone, and the second using a bank of 170 binary SVMs trained to identify important features on 17 binary target functions such as sonorance or nasality, each taking one of 10 different types of superframe as input. One word-feature SVM used a single microphone (**WF-SVM**), and the other used majority voting among the 7 microphones (**WFV-SVM**). These results are summarized in table 3.1.

While digit recognition using HMMs alone failed for the speaker most likely to reduce

[2]The authors do not list the number of correctly enunciated phonemes in each category, so it is impossible to compare these statistics against those of Thubthong et al. (2005).

or drop consonants, the SVM, which here relies on fixed-length words, failed for the severely stuttering speaker. The authors conclude that DTW-like features of the HMM give it robustness against large-scale word-length fluctuations, but that the SVM is more robust against dropped or deleted consonants. The authors also conclude that the SVM cannot be meaningfully applied if its inputs differ in temporal length, which ignores *sequence kernels* as a direct means for such a task, as discussed below.

Polynomial DTW kernel

Kernel functions that discriminate among vectors of unequal length often do so by applying a change of variables on the data. For instance, Wan and Carmichael (2005) convert the global DTW measure of distance to a dot product for use in an SVM kernel through spherical normalization. This process translates components of vectors $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ and $\mathbf{Y} = \{y_1, y_2, ..., y_m\}$ to a spherical space defined by their mutual origin (producing $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$). The geodesic distance between two points on this sphere, $d_S(\hat{x_i}, \hat{y_j})$, is defined by the angle between them, so

$$d_S(\hat{x_i}, \hat{y_j}) = \arccos(\hat{x_i}, \hat{y_j}) \tag{3.1}$$

is computed for each local distance. The global distance $D_s(\hat{X}, \hat{Y})$ is the mean angle between all pairs of vectors in $(\hat{X}, \hat{Y})$ plus DTW transition costs, giving the linear SVM kernel

$$K_{linearDTW}(\hat{X}, \hat{Y}) = \cos D_s(\hat{X}, \hat{Y}). \tag{3.2}$$

Wan and Carmichael raise Equation 3.2 to the third power to obtain higher-order, non-linear solution spaces, and compare its classification error against two other word classifiers – a manually endpointed DTW system, and an 11-state HMM with 3 Gaussians per state. These were compared on a set of 7 dysarthric and 7 non-dysarthric speakers uttering between 30 and 40 samples of each of a small set of 11 command words. All data was parameterized with 13 MFCCs on 32ms windows and their first order derivatives, and silences and non-speech bursts were removed using a Gaussian detector.

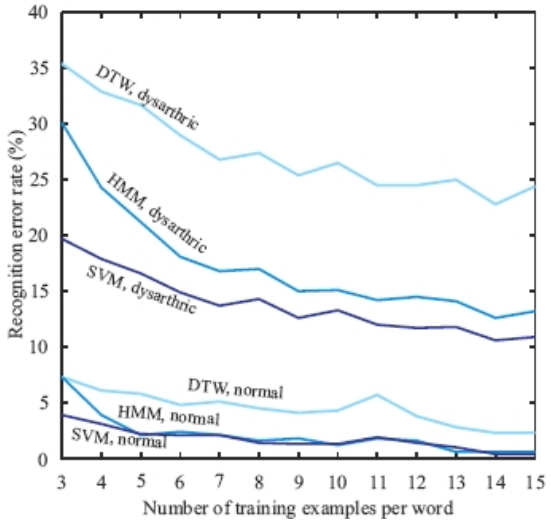


Figure 3.2: Comparison of recognition error rates for 3 word classification techniques (SVM with 3^{rd} order DTW kernel, 11-state HMM, and standard DTW template matching) for dysarthric and non-dysarthric speakers, across training set sizes (from (Wan and Carmichael, 2005)).

Results of this comparison are summarized in Figure 3.2. Standard DTW performs uniformly worse than the other methods within each speaker class, with the SVM reducing error by up to 46% relative to the HMM when training data is most sparse. The success of the DTW-kernel SVM here is significant, since it clearly improves over the contemporary HMM method, and because it demonstrates relative competence amid sparse data. Although these results cannot be directly compared to those of Hux et al. 5 years prior (section 3.1) due to differing application types, both surveys show a clear disparity between dysarthric and non-dysarthric speech recognition.

3.1.3 Neural networks and dysarthric speech

Jayaram and Abdelhamied (1995) compared two artificial neural networks (ANNs) taking fast Fourier transform (FFT) coefficients and formant frequencies as inputs, respectively, given dysarthric speech. Each network was feed-forward, fully connected, and had two hidden layers

where all nodes had sigmoid transfer functions. Both networks outperformed human performance (42.4%), and the Introvoice HMM-based system ($\leq 37.5\%$) on dysarthric speech, with the FFT-trained network reaching $\sim 75\%$ accuracy after 7000 iterations of backpropagation on 440 samples of speech of the cerebrally palsied speaker.

Polur and Miller (2006) obtain ergodic-HMM state sequences given acoustics from a dysarthric speaker and pass these along with the original acoustics to a neural network for word-identification. This approach has reduced error by up to 55.6% relative to HMM-only methods on dysarthric speech (Polur and Miller, 2006). The authors of that study claim that augmenting HMMs with neural networks in this way provides better discrimination by relaxing assumption constraints on the form of the statistical distribution to be modelled.

The only context provided to classification and predictive ANNs for time-varying signals is the physical input itself, rather than the previous states of the machine. This forces the size of these networks to expand at least multiplicatively with the size of the input (Krose and van der Smagt, 1996) in standard back-propagation learning. Recurrent networks can compress contextual information through additional units in the input frame that are activated solely by non-input units, usually with a fixed $+1$ weight. With context units activated solely by output units (a Jordan network), Robinson, Hochberg, and Renals (1996) learn emission probabilities for a phone-classification HMM and achieve 87% word accuracy on a $20,000$ word vocabulary. The Elman network, which receives contextual state from hidden units alone, can often achieve better accuracy than the Jordan network (Wilson, 1995).

3.1.4 Pathological effects on ASR

The unintelligibility of dysarthria is not due to any single behaviour, but to the combination of many articulatory and phonemic phenomena. These phenomena, however, can have unique consequences for automatically recognizing speech.

Increased variability among slow speakers described in section 2.2.1 suggests that even vowel models trained on groups of spastic or ataxic dysarthric or other abnormally slow speak-

ers may not necessarily be indicative of an individual's acoustic behaviour within that group. Muscle fatigue, particularly of the tongue (section 2.2.2), coupled with overall longer speech events, may also lead to nonlinear alterations of acoustic models over the length of an utterance.

Acoustic disfluency (see section 2.2.3) often leads to phonemic insertion errors in or around words containing voiceless plosives (Rosen and Yampolsky, 2000) or voiceless fricatives (/*f*/, /*θ*/, /*s*/, /*sh*/, /*h*/) (Raghavendra, Rosengren, and Hunnicutt, 2001). If dysarthric speech contains hesitations or repetitions, these are likely to be recognized with insertion errors (repeated words) or pairs of adjacent substitution errors (Huang, Acero, and Hon, 2001). If involuntary sounds are consistent, then a recognizer could be trained to ignore them or to classify them as an extraneous syntactic part of speech, at the cost of extra training effort (Lease, Johnson, and Charniak, 2006), though it is not clear that these involuntary sounds *are* consistent within dysarthric speech (Chen and Kostov, 1997).

Pauses of abnormal lengths may also lead to erroneous end-of-speech estimation (Rosen and Yampolsky, 2000). For example, the two final high-energy segments in Figure 2.4A will normally be misinterpreted as two separate words in standard ASR. Chen and Kostov attempt to deal with this problem by using manually-adjustable thresholds on energy levels, and an intra-word search function using trained speaker-dependent 8-state HMMs, and were able to achieve between 81 and 92% recognition on isolated digits (Chen and Kostov, 1997).

3.1.5 Miscellaneous adjustments to traditional processing

There have been many attempts to accommodate dysarthric speech in ASR that have involved more fastidious adjustments to traditional methods. Matsumasa et al. (2009) propose a system for recognition of dysarthric speech that incorporates robust feature extraction and an HMM-based 'metamodel'. In this work, feature extraction involves taking the logarithm of the mel-scaled FFT features, as usual, but instead of taking the discrete cosine transform at this point, principal component analysis is applied to combine 'stable' acoustic features with

their corresponding fluctuations in dysarthric speech[3]. Specifically, for analysis windows of width n, frequencies ω, an 'unstable' dysarthric utterance $X_n(\omega)$ is represented as the sum of a component of 'stable' speech $S_n(\omega)$ and a 'fluctuation element' $H(\omega)$ in the log space

$$\log X_n(\omega) = \log S_n(\omega) + \log H(\omega). \tag{3.3}$$

The stable component of the dysarthric speech is then extracted through PCA on the non-dysarthric speech by

$$\hat{S} = V^\top \log X_n(\omega) \tag{3.4}$$

where V is derived from the eigenvalue decomposition of the centered covariance matrix of the 'stable' speech set and is composed of the M-dimensional eigenvectors v_i corresponding to the L dominant eigenvalues,

$$V = [v_1, \cdots, v_L]. \tag{3.5}$$

By applying PCA instead of the discrete cosine transform to the mel-scaled filter bank output in this way, dimensionality is reduced relative to features expected to correspond to features in non-dysarthric acoustic space. The authors claim that this method alone results in an absolute improvement of 6.1% in word recognition accuracy over the traditional MFCC baseline of 77.1%, although experiments were performed with only one dysarthric speaker (Matsumasa et al., 2007). In addition to this approach to feature extraction, the same group learned a phoneme-classification model $P(p \mid X)$ for phoneme sequences p and approximated equation 2.1 by Bayes's rule and

$$P(W \mid X) \approx P(W \mid p^*) P(p^* \mid X) \tag{3.6}$$

for a phoneme sequence p^* chosen by

$$p^* = \arg\max_{p^* \in \mathscr{P}} P(p \mid X) \tag{3.7}$$

[3]This work is unclear on the definition of stability. Apparently, each single word uttered by a speaker with athetoid cerebral palsy was repeated several times by one or more non-dysarthric speakers, although the authors only explicitly claim to have 'stable' versions of the dysarthric speech. The authors also do not describe whether or how these 'stable' utterances are aligned with their 'unstable' counterparts.

for all possible phoneme sequences $\mathscr{P}$. This phoneme-classification model $P(p|X)$ is built using traditional HMMs, but the *a priori* presence of each phoneme is represented by a tristate HMM that encodes the probability of a phoneme being substituted, deleted, or preceded or followed by an erroneous phoneme insertion. This 'metamodel' resulted in an additional 3.8% improvement in word accuracy over the method of PCA-filtered features described above, although these experiments were performed with only two dysarthric speakers, and their baseline accuracies were abnormally high at 79.1% (Matsumasa et al., 2009).

3.2 Speech recognition with articulatory information

Although adaptation at the acoustic level alone has led to some increase in accuracy for atypical speech, there remains much room for improvement. In order to learn discriminating factors of dysarthric speech, new approaches must combine advanced machine learning with physiological models in order to directly inform causal parameters that would otherwise be hidden.

Articulatory knowledge has had relatively little historical presence in ASR despite evidence that articulatory control is often far more speaker-invariant than the resulting acoustics (Fujimura, 1986). Typically, such knowledge is manifested as decision trees that support state-tying in semi-continuous ASR systems (Young et al., 2006). Here, knowledge of common articulatory features (e.g., nasality in */m/* and */n/*) allows states in HMM models for different phones to be trained on shared data. There have, however, been a few attempts to build more explicit production knowledge into phoneme- and word-recognition systems. For example, appending articulatory measurements to acoustic observations has been shown to reduce phone-error relatively by up to 17% on a non-dysarthric speaker in a standard HMM system (however, if those articulatory measurements are inferred from acoustics, this improvement disappeared) (Wrench and Richmond, 2000). Similar work on incorporating AFs learned discriminatively with maximum mutual information into HMM systems has reduced word-error rates from 25% to 19.8% on English spontaneous scheduling tasks (Metze, 2007).

A number of approaches involved representing words as state-transition graphs embedded within HMMs whose states were conjunctions of discrete phonological or articulatory features. Erler and Deng (1993), Deng and Sun (1994), and Sun and Deng (2002a), for example, annotated words with parallel asynchronous variables representing the lips (which can be closed or rounded), tongue blade, tongue dorsum, velum, and larynx (which represents, e.g., vowelization or aspiration) (Sun, Jing, and Deng, 2000; Deng, 2000; Sun and Deng, 2002b). The manners in which these words could be constructed given these annotations were encoded within HMM transition networks with high-level linguistic constraints such as phrase boundaries, morphemes, syllables, and word stress. Augmenting HMMs in this way can explicitly model coarticulation and phonetic reduction while using a relatively small number of parameters compared with other HMM approaches (Lee, Fieguth, and Deng, 2001). Variations between pronunciations include anticipatory or inertial feature spreading, for example (McDermott and Nakamura, 2006), and are depicted in figure 3.3. Results have been somewhat humble, however, with this 'overlapping-feature' model improving the baseline triphone accuracy of 70.86% to 72.95% on the TIMIT database. However, this feature-rich approach has the advantage of requiring as little as 10% of the training data as the baseline. Richardson, Bilmes, and Diorio (2000) take a similar approach and reduce the size of the state-transition network by placing constraints on articulator velocities and continuity. Their approach reduced word-error rates relative to the state-of-the-art at the time by between 28–35%.

Along these lines, systems incorporating discrete articulatory features derived by neural networks from acoustics into HMMs have shown some improvement over acoustic-only baselines (Fukuda, Yamamoto, and Nitta, 2003; Kirchhoff, 1999). However, these results were often statistically insignificant except in the presence of extreme environmental noise (King et al., 2007). Neural network discriminative classifiers have been shown to correctly identify approximately 53% of simultaneous multivalued articulatory features, on average, for non-dysarthric speech (King and Taylor, 2000; Kirchhoff, 1999; Scharenborg, Wan, and Moore, 2007).

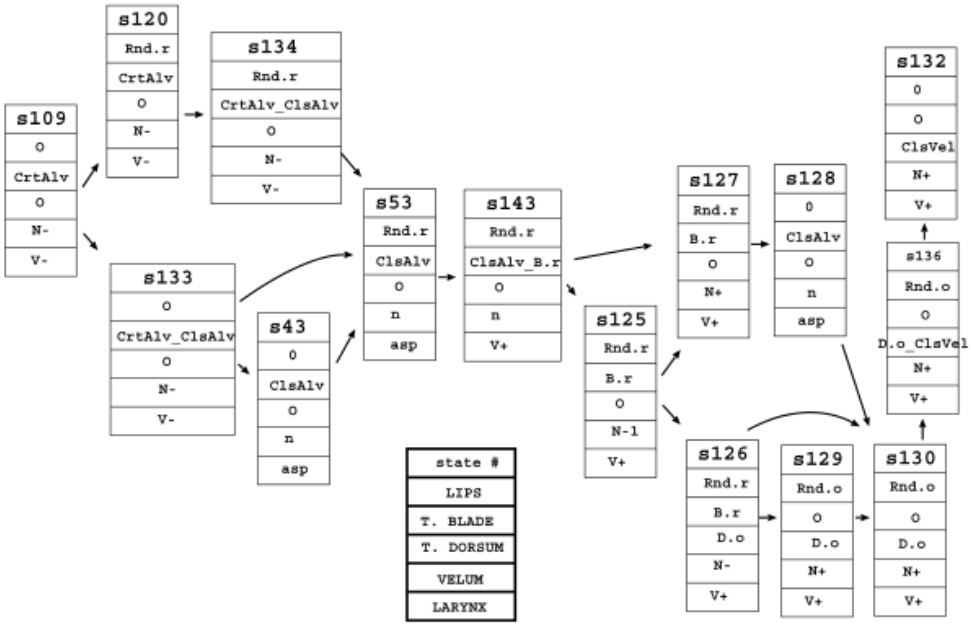


Figure 3.3: State-transition graph of feature-based HMM for the word *strong* as adapted by McDermott and Nakamura (2006) from Sun and Deng (2002a).

A commonality among all of this work is its reliance on non-dysarthric data where articulatory and acoustic patterns are less disordered than in speakers with cerebral palsy and other neuromotor disabilities (Livescu et al., 2007).

3.2.1 Audio-visual speech recognition

Speech recognition by humans often involves visual information to supplement the audio, especially in noisy environments where audio input is distorted or obfuscated (Summerfield, 1992). Physical gesticulation is also a common manner of conveying information between humans, allowing semantic information to be transmitted concurrently among multiple modalities (Rudzicz, 2006). The compulsion of humans to use visual information in deciphering speech is so strong that the visual signal will almost invariably supersede the audio signal if the two conflict in certain circumstances. Specifically, the McGurk effect is an audio-visual illusion in which audio of an individual saying */ba/* will be superimposed on video sequences of that

Figure 3.4: Fictionalized lip reading in profile by machine, from Kubrick (1968).

person uttering either */fa/* or */da/*. Human listeners, naïve or not, will consistently perceive */fa/* in the first combination and */ga/* in the second (McGurk and MacDonald, 1976).

Lip reading by machine is a field with a long tradition (Stork and Hennecke, 1996) that is becoming increasingly tractable for widespread use given a preponderance in low-cost video-capturing systems embedded in today's personal computers (Neti et al., 2000). A fictional depiction of this activity is shown in figure 3.4. Unsurprisingly, the incorporation of additional relevant information into the speech recognition process results in higher rates of correct recognition than acoustic-only ASR (Adjoudani and Benoit, 1995; Potamianos and Graf, 1998; Dupont and Luettin, 2000; Papandreou et al., 2009). However, most research in this area continues to concentrate on small-vocabulary tasks such as digit recognition (Neti et al., 2000).

Coupled HMMs, shown in figure 3.5(b), represent parallel streams of acoustics and visual observations and are commonly used in audio-visual speech recognition (Chu and Huang, 2000; Nefian et al., 2002b). Coupled HMMs are essentially collections of standard HMMs, each representing a unique data source, in which the discrete state nodes are conditioned on the state nodes of the HMMs representing other streams. This work classified pixels in the visual data into 'face' and 'mouth' classes by linear discriminant analysis performed offline on sequences that were manually segmented using chromatic values. The contour of the lips in this binarized space was obtained via a binary chain encoding method (Castleman, 1996), the result of which is shown in figure 3.5(a). The efficacy of this approach is especially pronounced in noisy environments, e.g., at 16 dB SNR. This combined approach gave 69.9% word accu-

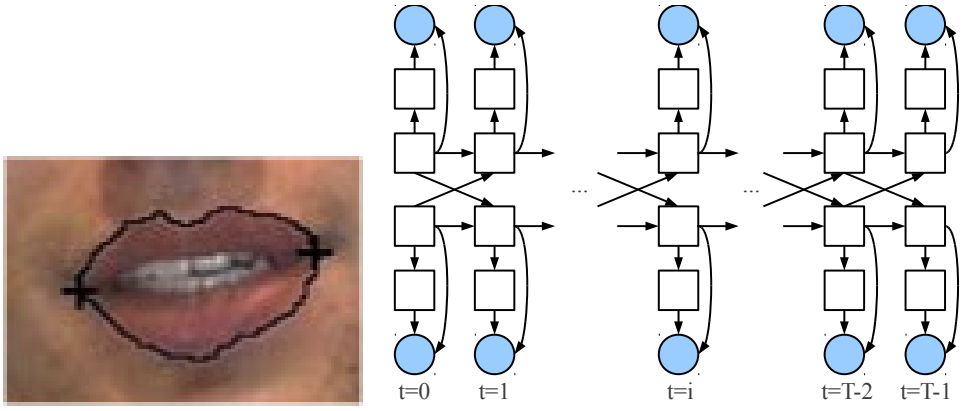


Figure 3.5: Lip contours (a) and Coupled HMM (b) with aligned acoustic and visual observations, each represented by circular nodes, from Nefian et al. (2002b). The coupled HMM is essentially two typical HMMs in parallel for each time step t.

racy compared with 28.26% accuracy for the acoustic-only method (this disparity decreases, naturally, at higher SNR levels). This approach was later shown to outperform a nearly identical factorial HMM model which conditioned all output and Gaussian mixture indices on both the acoustic and visual state sequences (Nefian et al., 2002a).

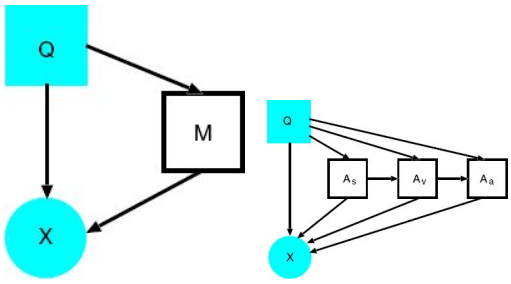
Saenko and Livescu (2006) use a model very similar to that of Nefian et al. (2002b) (although they use more generic topologies, as described in section 3.2.2) for recognizing English digits given audio and visual representations. This model allows the state sequences for the audio and visual streams to be 'de-synchronized' in that state transitions in one do not require state transitions in the other, which is accomplished by an additional variable in the network that measures the degree of asynchrony. On the CUAVE database of audio-visual English digits (Patterson et al., 2002) this work showed a word-error rate as low as 3% in the combined model in a SNR scenario of 12 dB in which the acoustic-only HMM baseline had a word-error rate of 7%. Larger margins were obtained in noisier environments.

3.2.2 Dynamic Bayes networks

Dynamic Bayes networks (DBNs) constitute a statistical framework for representing temporal dynamics of systems of variables (Deng, 2006). The class of DBNs generalize many other popular statistical models such as HMMs and Kalman filters and are represented by directed acyclic graphs of variables. A mathematical description of dynamic Bayes networks is provided in section 5.1.5. This framework is appealing since DBNS allow for both manual configuration of the topologies between relevant defined variables (e.g., articulatory parameters) and powerful statistical machine learning techniques (Zweig, 1998). Here, acoustics can be conditioned into a myriad of parameters including the type of speaker, their pronunciation variation, and their speaking rate, or on higher-level effects, such as prosodic or linguistic structure (Ostendorf, 2000). Recently, dynamic Bayes networks have been applied to the problem of AF classification by Frankel, Wester, and King (2007) under similar conditions as King et al. (2007) above using structures expanded upon in section 5.3 and correctly identified 57.8% of similar multivalued AFs on non-dysarthric speech. These data structures have also been used in modelling inter-dependencies between acoustics and measured articulation in regular speech (Nefian et al., 2002a). Stephenson, Magimai-Doss, and Bourlard (2004) showed that simple Bayes networks relating Mel-frequency cepstral coefficients observations with Wisconsin's X-ray microbeam articulatory data (Westbury, 1994) resulted in a 9% word-error rate reduction when compared with a baseline acoustic-only ASR system.

Markov, Dang, and Nakamura (2006) followed this work with a series of simpler Bayes networks that estimated the likelihood of acoustic observations given discretized articulatory parameters, achieving similar results when combined with an HMM-based ASR system. They point out that modelling speech by non-overlapping disjoint units is not physiologically plausible, and degrades performance with coarticulation phenomena. They suggest that replacing hidden variables in the standard HMM framework with actual articulatory data may result in more realistic and robust acoustic models.

The simple BNs in figure 3.6 determine the output probabilities of three-state phonetic



(a) (b)

Figure 3.6: Simple Bayes networks used to model HMM observation probabilities. Model (a) models only articulatory position data (A) while model (b) includes transitively dependent velocity (A_v) and acceleration (A_a) coefficients, from Markov, Dang, and Nakamura (2006).

HMMs within an ASR system given discrete state variable Q, continuous acoustic variable X, and hidden and discretized articulatory data. The first BN models output probabilities by

$$\begin{aligned} P^{(1)}(X = x_t|Q = q_i) &= \frac{P(X = x_t|Q = q_i)}{P(Q = q_i)} \\ &= \frac{\sum_{j=1}^{K} P(X = x_t, M = m_j, Q = q_i)}{P(Q = q_i)} \\ &= \frac{\sum_{j=1}^{K} P(X = x_t|M = m_j, Q = q_i)P(M = m_j|Q = q_i)P(Q = q_i)}{P(Q = q_i)} \\ &= \sum_{j=1}^{K} P(X = x_t|M = m_j, Q = q_i)P(M = m_j|Q = q_i) \end{aligned} \tag{3.8}$$

and the second by

$$\begin{aligned} P^{(2)}(X = x_t|Q = q_i) = \sum_{j=1}^{K_s}\sum_{n=1}^{K_v}\sum_{m=1}^{K_a} & P(A_s = a_j^s|Q = q_i) \\ & \cdot P(A_v = a_n^v|A_s = a_j^s, Q = q_i) \\ & \cdot P(A_a = a_m^a|A_v = a_n^v, Q = q_i) \\ & \cdot P(X = x_t|A_s = a_j^s, A_v = a_n^v, A_a = a_m^a, Q = q_i). \end{aligned} \tag{3.9}$$

In all graphical depictions of DBNs in this dissertation, filled and empty nodes represent observed and hidden variables, respectively, and square and round nodes are discrete and continuous variables, respectively. During training, when articulatory information is available, the values of the various A variables can be specified, and the BN parameters updated after

alignment using simple maximum likelihood estimation. During recognition, when the articulatory values are hidden, calculation of the output probabilities above is straightforward since all parent nodes represent discrete variables.

Across three normal speakers, using velocity and acceleration parameters did not present much improvement as the model in figure 3.6(b) averaged 84.7% accuracy, against 84.6% for the model in figure 3.6(a). More significantly, the position-only model outperforms standard HMM output models, reducing relative error by as much as $\sim 20\%$. The authors claim that future work involves experimenting with physiologically-inspired models for HMM state transitions (Markov, Dang, and Nakamura, 2006).

Chapter 4

The TORGO database of dysarthric articulation

This chapter describes the acquisition of a new database of dysarthric English speech in terms of aligned acoustics and articulatory data. This database, called TORGO, is the result of a collaboration between the departments of Computer Science and Speech-Language Pathology at the University of Toronto and the Bloorview Kids Rehab hospital in Toronto. The goal of this resource is to provide developers of speech technology with the tools to tailor their systems to atypical speech and to incorporate source information (i.e., knowledge of physical speech production) into these systems. This data is also applicable to linguists, pathologists, and clinicians who are interested in studying atypical speech production.

TORGO currently includes data from seven individuals with speech impediments caused by cerebral palsy or amyotrophic lateral sclerosis and age- and gender-matched control subjects. Each of the individuals with speech impediments are given standardized assessments of speech-motor function by a speech-language pathologist. Similar databases are surveyed in section 4.1 before the data collection procedure used in TORGO is described in section 4.2, including descriptions of subjects, assessments, speech stimuli, and instrumentation. Sections 4.3 and 4.4 describe post-processing techniques and observed aspects of the data, respectively.

4.1 Existing databases

To date, no database of combined acoustic and articulatory dysarthric speech is publicly available. Since dysarthric speakers are in the minority and susceptible to fatigue, collecting data from this population can be particularly challenging. Data collection with dysarthric speakers has usually involved fewer than 5 participants (Hasegawa-Johnson et al., 2006a), frequently producing only about 25 utterances each (Jayaram and Abdelhamied, 1995). Adjustments to regular ASR training in the presence of data sparsity have included using training algorithms that require less data (Wan and Carmichael, 2005) or augmenting 'surface' data (acoustics) with the hidden variables on which they depend (articulation) (Markov, Dang, and Nakamura, 2006).

The A.I. duPont Institute's Nemours database is a popular source of phonemically annotated dysarthric acoustics consisting of 11 dysarthric males with varying degrees of intelligibility and one non-dysarthric male. Each subject utters 74 syntactically invariant and semantically meaningless short sentences and two additional paragraphs (Menendez-Pidal et al., 1996). Each nonsense sentence has the form *The* N_0 *is* V*ing the* N_1, where N_0 and N_1 are unique monosyllabic nouns and V is a monosyllabic verb. The target words, N_0, V, and N_1, were randomly selected without replacement in order to provide closed-set phonetic contrasts (e.g., place, manner, voicing). Here, phonemic annotations are automatically derived by HMM-based forced alignment given known orthography. Each speaker is also associated with a complete Frenchay assessment of motor function. Since no physiological information is included, articulatory features are derived directly from phonemic annotations as described in section 2.4 and provide the bases for production knowledge in section 5.3.

The University of Edinburgh's MOCHA database consists of 460 sentences derived from TIMIT (Zue, Seneff, and Glass, 1989) uttered by a male and a female British speaker without dysarthria (Wrench, 1999). All acoustic data is temporally aligned with electromagnetic articulography (recorded at 500 Hz), laryngography (at 16 kHz), and electropalatography (EPG, at 200 Hz). This study involves eight bivariate articulatory parameters, namely the upper lip

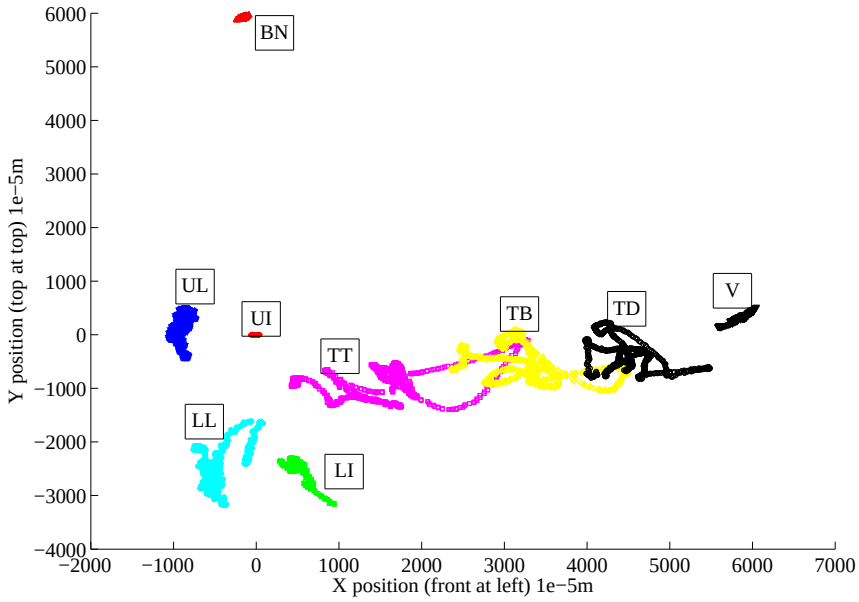


Figure 4.1: The midsagittal motion of the articulators during the phrase "*This was easy for us*".

(UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue blade (TB, 1 cm from the tongue tip), tongue dorsum (TD, 1 cm from the tongue blade), and velum (V). Each parameter is measured in the two dimensions of the midsagittal plane, resulting in a 16-dimensional articulatory configuration, as shown in figure 4.1.

Recently, Yunusova et al. (2008) have collected x-ray microbeam data with 15 individuals with ALS and Parkinson's disease. This data includes point-data in similar positions to the MOCHA database and generally follows the protocol and methodology of the Wisconsin x-ray microbeam database for non-dysarthric speakers (Westbury, 1994). This database only includes 10 stimuli per speaker, however, which is not enough to train ASR systems.

Finally, the Universal Access (UA-Speech) database recorded at the University of Illinois consists of 17 participants diagnosed with cerebral palsy (Kim et al., 2008). Each participant in that database utters 765 isolated words from sources such as digits, the radio alphabet, common dictionary words, and computer commands, which is comparable in scope and in content to the contributions of participants in the TORGO database. However, the UA-Speech database

does not include any connected-word sentence-level utterances nor does it provide access to measurements of the tongue. Furthermore, at the time of this writing, there is little to no annotation that accompanies this data such as phonemic annotation.

The following sections describe our study population, their speech-motor assessment, and the data collection process.

4.2 Data collection

Data collection began in 2008 through collaboration between the departments of Computer Science and Speech-Language Pathology at the University of Toronto, Bloorview Kids Rehab hospital in Toronto, and the Ontario Federation for Cerebral Palsy. The following section describes various aspects of the data collection process.

4.2.1 Subjects

Seven dysarthric subjects (4 male, 3 female) have so far been assessed in this study, covering a wide range of intelligibility. Dysarthric subjects were recruited by a speech-language pathologist at the Bloorview Research Institute in Toronto (Rudzicz et al., 2008). The subjects were between the ages of 16 and 50 years old and have dysarthria resulting from cerebral palsy (e.g., spastic, athetoid, or ataxic). In addition, one subject with dysarthria from a confirmed diagnosis of amyotrophic lateral sclerosis (ALS) was recruited. These individuals were matched according to age and gender with non-dysarthric subjects from the general population. Having an equal number of dysarthric and control speakers is useful for comparing acoustic and articulatory differences, and for analyzing these relationships mathematically and functionally (Hosom et al., 2003; Kain et al., 2007).

Each subject began the data collection process with a short questionnaire that covers general demographic data and health-related questions that can impact speech and language function including various types of motor problems, both gross (e.g., standing, balancing) and fine (e.g.,

writing, swallowing). All participants were required to have a negative history of severe hearing or visual problems and of substance abuse, and to be able to read at a 6^{th} grade elementary level. This was further quantified by requiring that their cognitive function lie above or at level VIII (i.e., Purposeful-Appropriate) on the Rancho scale (Herndon, 1997), which is determined during a pre-visit questionnaire.

Each participant has recorded 3 hours of data (approximately 500 utterances from each dysarthric speaker and 1200 from non-dysarthric speakers).

4.2.2 Assessment

Clinical assessments of motor function and intelligibility in dysarthric speakers are often used by speech therapists for rehabilitation (Kent, 2000) and intelligibility correlates well with ASR accuracy (Ferrier et al., 1995). The motor functions of each experimental subject were assessed according to the standardized Frenchay Dysarthria Assessment (FDA) (Enderby, 1983) by a speech-language pathologist. This assessment is designed to categorize and diagnose individuals with dysarthria while being easily applicable to therapy, sensitive to changes in speech, simple and quick to administer, and easily communicable within professional teams. There exist other assessment measures of oral motor ability, such as the Assessment of Intelligibility of Dysarthric Speech (AIDS) (Yorkston and Beukelman, 1981), which quantifies the intelligibility of single words, sentences, and speaking rates of adults and adolescents with dysarthria. However, these tend to focus only on speech production, whereas the FDA also includes analysis of the movement of the articulators in non-linguistic contexts.

The Frenchay assessment measures 28 perceptually-relevant dimensions of speech grouped into 8 categories, namely reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility as described in Table B.1 in Appendix B. Influencing factors such as speech rate and sensation are also recorded. To measure most of these dimensions, the administering clinician either engages the subject in communication or has the subject perform a simple task (e.g., drinking from a cup of water) while observing their oral movements. The subject's oral

behaviour is rated on a 9-point scale and plotted with a simple bar graph. The assessment provides characterizations of behaviours across this 9-point scale. For example, for the cough reflex dimension, a subject would receive a grade of 'a'(8) for no difficulty, 'b'(6) for occasional choking, 'c'(4) if the patient requires particular care in breathing, 'd'(2) if the patient chokes frequently on food or drink, and 'e'(0) if they are unable to have a cough reflex. The resulting graph provides a high-level overview to the clinician to quickly identify problematic aspects of speech.

The mildly dysarthric speakers were able to participate in all tasks required of them for the assessment. The more severely dysarthric speakers also engaged in all tasks but levels of fatigue and poor breath control inhibited them from completing some of these tasks. Assessment data of this type is useful in analyzing how modifications to ASR software affects achievable accuracy across the spectrum of intelligibility levels. For example, alterations to the process by which vowels are categorized by the machine may have greater impact for those individuals with more atypical tongue movement, as opposed to pronounced velum differences. Table B.1 in Appendix B shows the mean (μ) and standard deviation (σ) of our participants, split by gender, across each of the 28 dimensions of the Frenchay assessment.

4.2.3 Speech stimuli

All subjects read English text from a 19-inch LCD screen placed 60 cm in front of them, or repeated verbal stimuli if the former is not possible. The stimuli were presented to the participants in randomized order from within fixed-sized collections of stimuli. Dividing the stimuli into collections in this manner guaranteed a certain degree of overlap between subjects who speak at vastly different rates, which is the case when dealing with severely dysarthric speakers. There is no dependency relation between the sessions and the presented stimuli. The collected speech data covers a wide range of articulatory contrasts, is phonetically balanced, and simulates simple command vocabularies typical of assistive ASR technology. The following types of stimuli are included:

Non-words These are used to control for the baseline abilities of the dysarthric speakers, especially to gauge their articulatory control in the presence of plosives and prosody. Speakers are asked to perform the following:

- 5 to 10 repetitions of */iy-p-ah/*, */ah-p-iy/*, and */p-ah-t-ah-k-ah/*, respectively. These sequences allow us to observe phonetic contrasts around plosive consonants in the presence of high and low vowels, and have been used in other studies (Bennett, van Lieshout, and Steele, 2007).
- High-pitch and low-pitch vowels maintained over five seconds (e.g., *"Say 'eee' in a high pitch for 5 seconds"*). This allows us to explore the use of prosody in assistive technology, as many dysarthric speakers who have difficulty with articulation can control pitch to some degree (Patel, 2002b).

Short words These are useful for studying speech acoustics without the need for word-boundary detection. These stimuli include formant transitions between consonants and vowels, the formant frequencies of vowels, and acoustic energy during plosive phonemes, as explored by Roy et al. (2001). This category includes the following:

- Repetitions of the English digits 1 to 10, *yes*, *no*, *up*, *down*, *left*, *right*, *forward*, *back*, *select*, *menu*, and the international radio alphabet (i.e., *alpha*, *bravo*, *charlie*, etc.). These words are useful for hypothetical command-and-control software for accessibility.
- 50 words from the the word intelligibility section of the Frenchay Dysarthria Assessment (Enderby, 1983).
- 360 words from the word intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981). These are grouped into phonetically similar words, as was presented in the Nemours database (Menendez-Pidal et al., 1996) (e.g., *hit*, *hat*, and *hut* are a trio of monosyllabic words differing only in their vowel).

- The 10 most common words in the British National Corpus (Clear, 1993).
- All phonetically contrasting pairs of words from Kent et al. (1989). These are grouped into 18 articulation-relevant categories that affect intelligibility, including glottal/null, voiced/voiceless, alveolar/palatal fricatives and stops/nasals; these are shown in table A.1 in Appendix A.

Restricted sentences In order to utilize lexical, syntactic, and semantic processing in ASR, full and syntactically correct sentences are recorded. These include the following:

- Preselected phoneme-rich sentences such as *"The quick brown fox jumps over the lazy dog"*, *"She had your dark suit in greasy wash water all year"*, and *"Don't ask me to carry an oily rag like that"*.
- The Grandfather passage from the Nemours database (Menendez-Pidal et al., 1996).
- 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981). These sentences are designed to highlight perceptual contrasts in speech that are relevant to speaker intelligibility.
- The 460 TIMIT-derived sentences used as prompts in the MOCHA database (Wrench, 1999; Zue, Seneff, and Glass, 1989).

Unrestricted sentences Since a long-term goal is to develop applications capable of accepting unrestricted and novel sentences, we elicited natural descriptive text by asking participants to spontaneously describe 30 images of interesting situations taken randomly from among the cards in the Webber Photo Cards: Story Starters collection (Webber, 2005). These are similar in nature to images used in other standardized tests of linguistic proficiency (Campbell, Bell, and Keith, 2001). This data complements restricted sentences in that they more accurately represent naturally spoken speech, including disfluencies and syntactic variation.

4.2.4 Instrumentation

In each of three sessions, subjects are prepared for either of two instrumental studies. The first involves the use of electromagnetic articulography (EMA) and the other involves video recordings of facial markers using specialized software to extract their positions over time. For EMA, the preparation takes approximately 30 minutes in which sensors are placed on the relevant locations of the speech articulators as described below in section 4.2.4. In the video-based setup, preparation takes about 20 minutes and involves the placement of phosphorescent markers on relevant landmark positions of the face, as described in section 4.2.4. The actual data collection process takes no more than 1 hour thereafter in either the EMA or video configurations. Of the three recording sessions, two are within the EMA environment since we are interested in the motion parameters of the tongue, which are unavailable in the video setup. We perform three sessions for each participant in order to check the reliability and variability of our data over time. Moreover, the literature suggests that EMA can provide a reliable estimate of speaker variability of speech parameters over time (van Lieshout et al., 1997).

Electromagnetic articulograph (EMA) kinematics

The collection of movement data and time-aligned acoustic data is carried out using the three-dimensional AG500 electro-magnetic articulograph (EMA) system (Carstens Medizinelektronik GmbH, Lenglern, Germany) with fully-automated calibration. The 3D-EMA system is considered state-of-the-art technology for studying speech movements and its principles have been elaborated elsewhere (Hoole, Zierdt, and Geng, 2003; van Lieshout, Merrick, and Goldstein, 2008; Yunusova, Green, and Mefferd, 2009; Zierdt et al., 2000). This system allows for 3D recordings of articulatory movements inside and outside the vocal tract, thus providing a detailed window on the nature and direction of speech related activity.

In the AG500 system, six transmitters attached to a clear cube-shaped acrylic plastic structure (dimensions L 58.4 x W 53.3 x H 49.5 centimetres) generate alternating electromagnetic fields as shown in figure 4.2(a). Each transmitter coil has a characteristic oscillating frequency

ranging from 7.5 to 13.75 kHz (Yunusova, Green, and Mefferd, 2009). When sensors (also called *transducers*) are brought into the field, induction generates a weak current oscillating with the same frequencies. The energy in each frequency of the induced complex signal depends on the distance of the sensor from the transmitters and its orientation. The spatial position of the sensor coil in the field is then determined by identifying the strength of the contribution of each transmitter coil via a process of demodulation of the complex signal induced in the sensor (Yunusova, Green, and Mefferd, 2009). The induced voltage values in the sensors are compared to expected values based on a known field model (Zierdt, Hoole, and Tillmann, 1999) and the difference is expressed as root-mean-square (RMS) error. The system translates these voltages into 3D coordinates of sensor positions over time. As will be discussed later, the RMS error is used to position the subject within the recording field and in part to measure the recording accuracy of the system.

As recommended by the manufacturer, the AG500 system is calibrated prior to each session subsequent to a minimum of a 3 hour warm-up time. It is reported that, at or close to the cube's centre, positional errors are significantly smaller (Yunusova, Green, and Mefferd, 2009) compared to the peripheral regions of the recording field within the cube. For our system, the stable volume around the center was roughly $0.008m^3$ (approximately the size of a basketball). Thus, care was taken to ensure that all participants were as close to the cube centre as possible, as shown in figure 4.2(a). The subject positioning within the cube was aided visually by the `Cs5view` real-time position display program (Carstens Medizinelektronik GmbH, Lenglern, Germany). This allowed the experimenter to continuously monitor the subject's position within the cube (repositioning the subject if required) and thereby maintain low RMS error values[1] to ensure good tracking of the sensor coils.

Sensor coils were attached to three points on the surface of the tongue, namely tongue tip (TT – 1 cm behind the anatomical tongue tip), the tongue middle (TM – 3 cm behind the

[1]The `Cs5view` real-time position display flags a coil in red if the RMS error exceeds 30 units; however, the RMS during recording rarely exceeded 8 units across all coils, which is suitable for minimizing position tracking errors (Kroos, 2008; Yunusova, Green, and Mefferd, 2009).

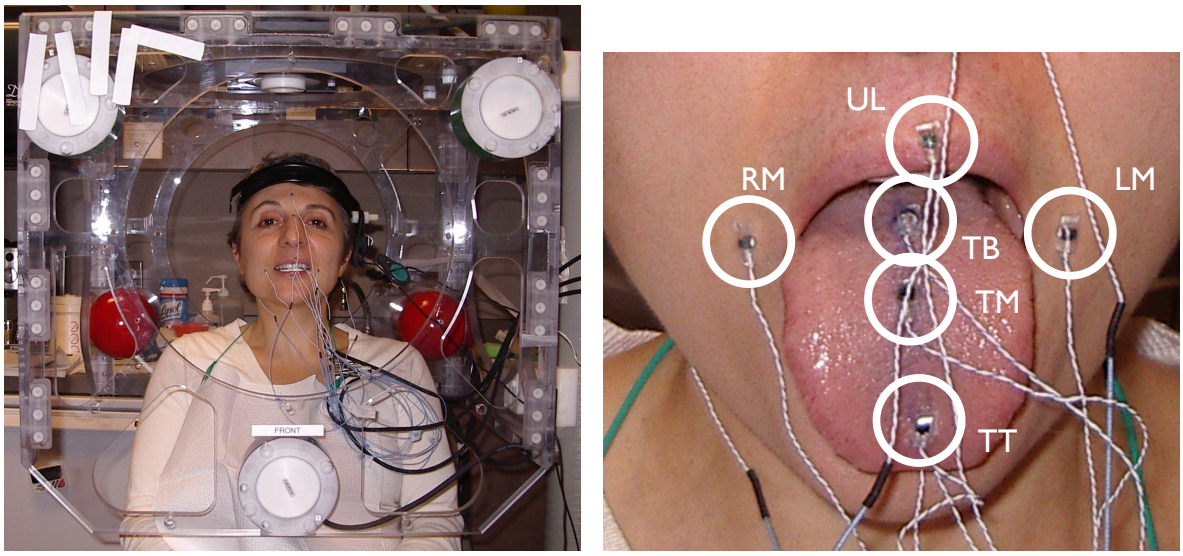
tongue tip coil), and tongue back (approximately 2 cm behind the tongue middle coil). A sensor for tracking jaw movements (JA) is attached to a custom mould made from polymer thermoplastic that fits the surface of the lower incisors and which is necessary for a more accurate and reproducible recording (van Lieshout and Moussa, 2000). Four additional coils are placed on the upper and lower lips (UL and LL) and the left and right corners of the mouth (LM and RM). The placement of some of these coils is shown in figure 4.2(b). Further coils are placed on the subject's forehead, nose bridge, and behind each ear above the mastoid bone for reference purposes and to record head motion. Except for the left and right mouth corners, all sensors that measure the vocal tract lie generally on the midsagittal plane on which most of the relevant motion of speech takes place. Sensors are attached by thin and light-weight cables to recording equipment but do not impede free motion of the head within the EMA cube. Many cerebrally palsied individuals require metal wheelchairs for transportation, but these individuals were easily moved to a wooden chair that does not interfere with the electromagnetic field for the purposes of recording.

Video-based articulatory kinematics

Although EMA provides detailed recordings of the tongue, which is not normally visible, typical use of speech recognition software will not likely involve such measurements. Therefore, we implement a second recording environment whose purpose is to derive more varied surface-level facial information using digital cameras. Here, recorded positions are meant to mimic the type of information that can be extracted from webcam-based face-recognition software.

Here, two digital video cameras are placed equidistant from the subject, at approximately 45 degree angles to their midsagittal plane, to the front-left and front-right of the subject. Video is captured at 60 frames per second and audio at $16,000$ Hz on both cameras. This audio is used for synchronizing the frames from both cameras and for separate acoustic measurements.

Two 250 W black lights are used to illuminate small (2 mm radii) glow-in-the-dark markers placed on the surface of the subject's face at selected points around the lips and over the orbic-



(a) (b)

Figure 4.2: The AG500 electromagnetic articulography system. Figure 4.2(a) shows a participant seated in the center of the EMA cube. Figure 4.2(b) shows the placement coils on the right mouth (RM), left mouth (LM), upper lip (UL), tongue tip (TT), tongue mid (TM), and tongue back (TB).

Figure 4.3: The binocular video recording setup showing the placement of phosphorescent dots on the subject's face.

ularis oris, depressor anguli oris, and depressor labii inferioris muscles as in previous studies on speech production (Craig, van Lieshout, and Wong, 2007) and as shown in figure 4.3.

Facial markers are tracked by specialized vision software based on strong contrasts between the reflection of the markers and the relatively darker background. These positions are converted into 3-dimensional co-ordinates using pairs of aligned video images and an estimated inter-camera calibration (Tsai, 1987). Calibration between cameras is performed by first filming a reference object with a known geometry, namely a cube with 30 cm sides.

Acoustics and microphones

All acoustic data are recorded simultaneously through two microphones. The first is an Acoustic Magic Voice Tracker array microphone with 8 recording elements. The device uses amplitude information at each of these microphones to pinpoint the physical location of the speaker within its 60-degree range and to reduce acoustic noise by spatial filtering and typical amplitude filtering in firmware. This microphone records audio at 44.1 kHz. The second microphone is a head-mounted electret microphone which records audio at 16 kHz. The electromagnetic field produced by this microphone does not demonstrably affect the field of the EMA system, and so it can be worn during all recordings. Using multiple microphones to record speech

can significantly improve the intelligibility of that speech in the presence of acoustic noise (Schwander and Levitt, 1987; Aarabi and Shi, 2004; Shi, Aarabi, and Jiang, 2007).

Signals from the two microphones are temporally aligned using simple cross-correlation, which is a measure of the similarity of one waveform and a second time-lagged waveform. Namely, given the two discrete signals f and g, we compute the complex conjugate[2] of the first, giving signal f^* consisting of real and phase values, and compute the cross-correlation by

$$(f \star g)[n] = \sum_m N^{-1} f^*[m] g[n+m]$$

where N is the length of the longer of the two sequences. The maximum value of this cross-correlation signal is the time delay between the jointly stationary signals, which is the speech signal recorded by both microphones. An example of this alignment is shown in figure 4.4. An alternative mechanism for ensuring alignment between signals involved the construction of an electrical circuit to communicate between the AG500 and the laptop computer connected to the directional microphone. This system was later abandoned because acoustic alignment with cross-correlation would be performed regardless, but its diagram is included in Appendix D.

Finally, acoustic noise reduction is performed using spectral subtraction (Ephraim and Malah, 1985). Here, we assume that the recorded audio $y[n]$ consists of a desired signal $x[n]$ that has been corrupted by additive noise $\upsilon[n]$, i.e.,

$$y[n] = x[n] + \upsilon[n].$$

Here, we assume that the signals $x[n]$ and $\upsilon[n]$ are statistically independent, as are their power spectra $|X(f)|^2$ and $|\Upsilon(f)|^2$ [3], so that the power spectrum of the recorded audio is

$$|Y(f)|^2 \approx |X(f)|^2 + |\Upsilon(f)|^2 .$$

We estimate $|\Upsilon(f)|^2$ by M-frame estimates over periods containing only noise,

$$\left|\hat{\Upsilon}(f)\right|^2 = \frac{1}{M} \sum_{i=0}^{M-1} |Y_i(f)|^2 .$$

[2] If $z = x + iy = Ae^{i\phi}$ is a complex signal with $i^2 = -1$, then its complex conjugate is $z^* = x - iy = Ae^{-i\phi}$.

[3] A power spectrum is defined as $|X(f)|^2 = X(f)X^*(f)$ where $X(f)$ is the Fourier transform at frequency f and $X^*(f)$ is the complex conjugate counterpart.

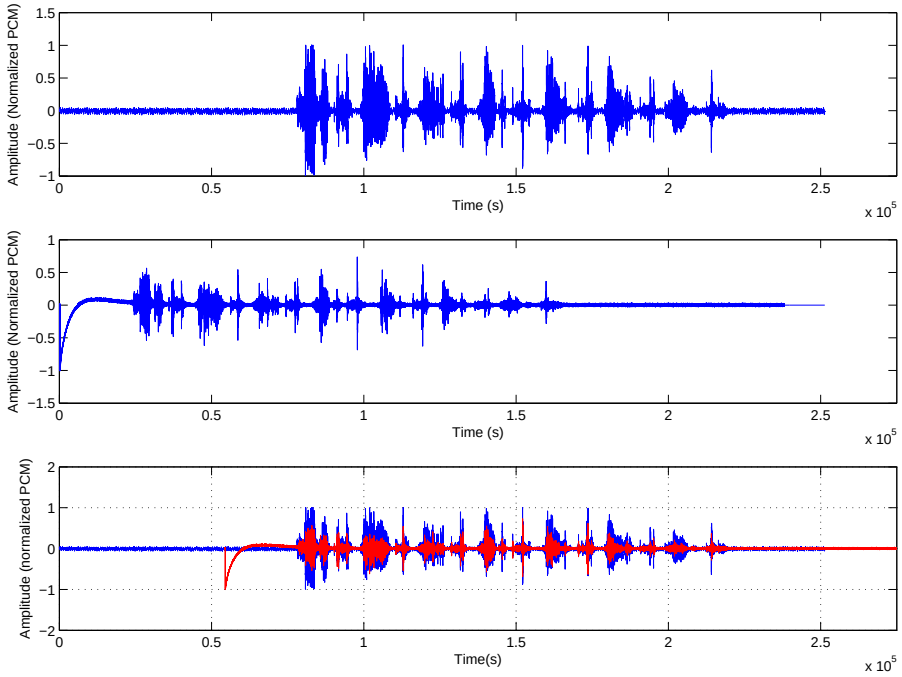


Figure 4.4: Alignment of two acoustic sources with the cross-correlation method. The top waveform is the signal recorded by the head-mounted microphone. The middle waveform is the signal recorded by the directional microphone. The bottom superposition of waveforms is produced by cross-correlation.

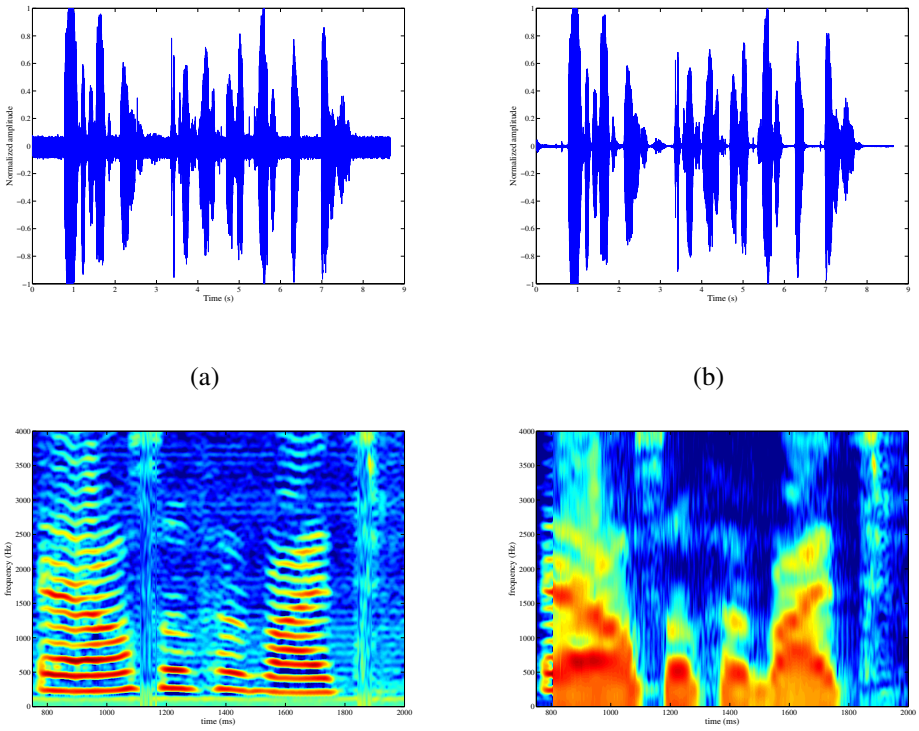


Figure 4.5: Original (a) and enhanced (b) waveforms and spectrograms for audio in TORGO uttered by speaker F03. Enhancement performed by spectral subtraction based on Martin (2001). Note the reduction in noise below 250 Hz and the general smoothing of the spectra in the enhanced spectrogram.

Given a frequency-dependent signal-to-noise ratio

$$SNR(f) = \frac{|Y(f)|^2}{\left|\hat{\Upsilon}(f)\right|^2}$$

we then estimate the source waveform by

$$\left|\hat{X}(f)\right|^2 = \max\left(0, |Y(f)|^2 - \left|\hat{\Upsilon}(f)\right|^2\right) = \max\left(0, |Y(f)|^2\left(1 - \frac{1}{SNR(f)}\right)\right).$$

The implementation used in this thesis is based on work by Martin (2001) in which spectral minima are tracked in each frequency band and are used to smoothe $|\Upsilon(f)|^2$ by minimizing the conditional mean square estimation error at each time step. Figure 4.5 shows the results of applying this technique on audio data in the TORGO database.

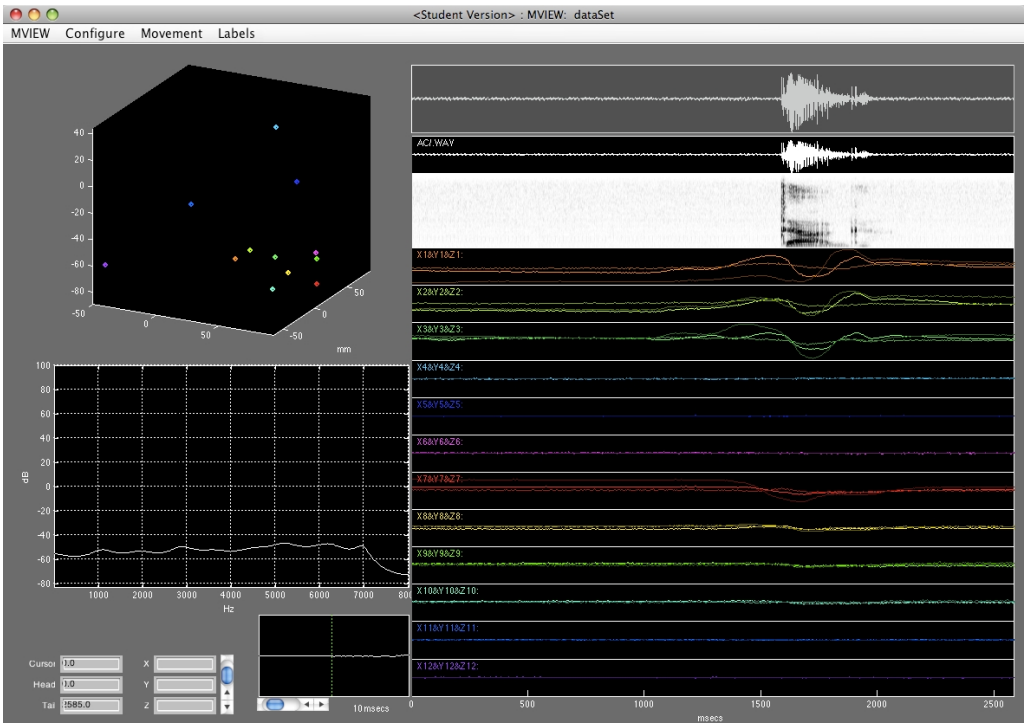


Figure 4.6: The MVIEW visualization environment for TORGO EMA data. The upper-left corner shows the instantaneous 3D locations of 12 articulator points. The upper-right corner shows the acoustic waveform and spectrogram. The lower-right corner shows the motion of each articulator point over time, in mm. The lower-left corner shows controls and the instantaneous spectrum.

4.3 Data post-processing

All data is being phonemically annotated to the TIMIT phone set (Zue, Seneff, and Glass, 1989) by a speech-language pathologist to allow supervised frame-level training of phone-dependent acoustic and kinematic models. These annotations are further checked by two naïve listeners for consistency. EMA recordings are analyzed in MATLAB with the program shown in figure 4.6, which displays instantaneous spectral acoustics and articulatory positions, as well as their respective temporal sequences (Tiede, 2008).

The AG500 EMA system has an expected error specification of up to 0.5mm in each dimension (X, Y, and Z) and an angular error (θ) of less than half of a degree. However, in

reality it is possible that accuracy may vary slightly across different AG500 systems due to set-up and environmental conditions such as ambient room temperature, type of sensor coils used, and existing electromagnetic fields in the room. These conditions may also vary across time (Kaburagi, Wakamiya, and Honda, 2005; Yunusova, Green, and Mefferd, 2009). To estimate more realistic values, we carried out a series of static and dynamic accuracy measurements for the AG500 system. For static measurement, 3-dimensional Euclidean distances between pairs of sensor coils were calculated. The sensors used here were those located on relatively rigid surfaces, namely the forehead, nose bridge, and behind the ear on the skin covering the right and left mastoid bone. Under ideal conditions, the distance between the pairs of sensors should remain constant throughout all trials for a given session. In other words, smaller average standard deviations for the 3D Euclidean distances between pairs of reference coils would imply lower static system noise or relative error. Similar methods have been applied to a camera-based marker tracking system (Craig, van Lieshout, and Wong, 2007) and in other 3D EMA systems (Hoole and Zierdt, 2010; Yunusova, Green, and Mefferd, 2009). This Euclidean RMS method provides a real and accurate measure of intrinsic system noise and relative error for each recording session. The average value was 0.2 mm across all pairs. These numbers may be taken roughly as the lower limit of the system's resolution (Kroos, 2008).

Recent studies have indicated that position errors in dynamic measurements, as opposed to static measurements, may be larger in magnitude and may vary across the three spatial dimensions (Kroos, 2008; Yunusova, Green, and Mefferd, 2009). We therefore ran a set of dynamic accuracy measurements for all coils using a specific tool recommended by the manufacturer. This allows us to estimate dynamic spatial errors as a function of sensor orientation. This accuracy checking tool is a mechanical device that is rigidly fixed in the centre of the cube's recording field and allows user defined manipulations of sets of coils in different orientations and directions. The device is constructed such that sets of coils placed on it can only travel a fixed distance (70 mm) in a particular direction. For the current study, we displaced 3 sets of 4 coils (i.e. $\langle(1..4, 5..8, 9..12\rangle$) across the entire 70 mm distance six times in a row in each

dimension (X, Y, and Z). A custom Matlab algorithm calculated the maximum 3D Euclidean displacement between points in that trial, as well as the average 3D Euclidean displacement. The algorithm automatically finds the coils that are being moved and the dimensions in which they are moving using maximum variance. Ideally, the maximum and average 3D Euclidean displacement values should be as close to 70 mm as possible. The amount of deviation from 70 mm provides an estimate of direction specific spatial accuracy of the system. We calculated the accuracy averaged across all 12 sensor coils per dimension. This was in the range of 0.54 to 0.60 mm in the Z (up/down) dimension, 0.34 to 0.59mm in the X (front/back) dimension, and 0.84 to 1.07 mm in the Y (left/right) dimension.

4.3.1 Data normalization

Position normalizations and corrections for head movements were carried out using custom-made `NormPos` software from the manufacturer of the AG500. The `NormPos` program does a sample-by-sample head normalization by rotating and shifting the coordinate system such that all reference sensors remain in the same 3D location across all samples and trials. Computationally, this is carried out using algorithms similar to 3D pose estimation methods (Kroos, 2008). Such algorithms calculate transformation parameters that can transform head position of a given sample to an experimenter chosen arbitrary reference position (that defines the orientation of the head and the origin of the coordinate system). The transformation parameters are derived by minimizing the sum of the squared distances between the reference sensor coils in the reference position and the actual position in other trials using linear least squares approaches such as (Kroos, 2008). The `NormPos` program stores these transformational parameters as a normalization pattern file. This normalization pattern file is then used to rotate and translate all other (non-reference) sensor coils positions in the remaining trails of the experiment to yield articulation trajectories that are corrected for head movements and with a fixed head-orientation that is identical across trials (and across subjects).

Since the `NormPos` program uses a normalization pattern file that is based on a single trial,

the quality of the head movement correction for the entire experiment depends on the quality of the data from the reference sensor coils in that trial. At times, the quality of data may not be equally good in all reference coils (as in the case of coil detachment and/or position tracking errors). For this reason, researchers have recommended the use of more than two reference sensor coils [4], typically four, to allow for redundancy in the available reference sensor coils (Hoole and Zierdt, 2010). For the present study, the two noise measures that were previously discussed were used to decide which two or three reference sensor coils (of the four available) were suitable to create the normalization pattern file (Hoole and Zierdt, 2010). Generally, the nose bridge and the two sensor coils behind the ears had the least amount of noise and were chosen to create the normalization pattern.

All trials are screened for errors in performance using procedures reported in the literature (Namasivayam and van Lieshout, 2008). Errors in speech production (e.g., coughs, laughs, misarticulations, false starts) are noted. A research associate, carefully reviewed the movement data visually and listened to the acoustic recordings (that were collected simultaneously with the movement data) and then compared these to the error notations that were made during the experiments. Only error-free and fluent portions of trials were used in this study.

4.3.2 Reconstruction of 3D movement from binocular video

There are various techniques that estimate the 3D structure of the face given stereo (i.e., 'binocular') video images. Some of these involve fitting video images to 3D polygonal models that have previously been trained by manual facial landmark identification (e.g., eyebrows, nose, eyes) (Blanz and Vetter, 2003; Park and Jain, 2006), often through the use of generic models of facial structure (Chowdhury and Chellappa, 2003).

In order to estimate the 3D position of a point P viewed simultaneously by multiple cameras, we must first transform its arbitrary location (X_W, Y_W, Z_W) in the coordinate system of the

[4] Two sensors, in principle, are sufficient to characterize the six degrees of freedom related to rigid-body motions (Hoole and Zierdt, 2010).

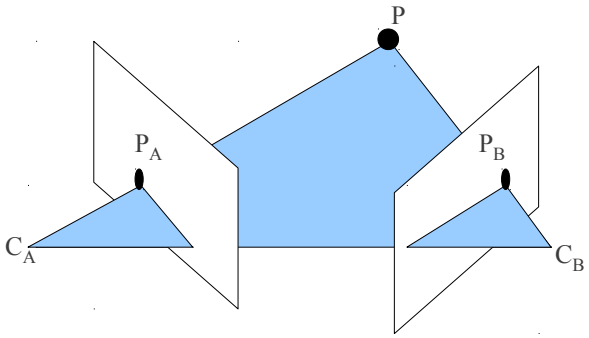


Figure 4.7: Reconstructing 3D coordinates of point P given its projections in the 2-dimensional images of cameras A and B described by their focal points C_A and C_B, respectively.

world with origin $(X_W^{(0)}, Y_W^{(0)}, Z_W^{(0)})$[5] to its location (X_c, Y_c, Z_c) in the coordinate system of the camera with origin $(X_c^{(0)}, Y_c^{(0)}, Z_c^{(0)})$, as visualized in figure 4.7.

This coordinate transformation is performed by a simple translation and rotation

$$\begin{aligned}
\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} &= \mathbf{M} \begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} + \mathbf{T} \\
&= \begin{bmatrix} \cos\varphi\cos\kappa, & \sin\omega\sin\varphi\cos\kappa - \cos\omega\sin\kappa, & \cos\omega\sin\varphi\cos\kappa + \sin\omega\sin\kappa \\ \cos\varphi\sin\kappa, & \sin\omega\sin\varphi\sin\kappa - \cos\omega\cos\kappa, & \cos\omega\sin\varphi\sin\kappa - \sin\omega\cos\kappa \\ -\sin\varphi, & \sin\omega\cos\varphi, & \cos\omega\cos\varphi \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} \\
&\quad + \begin{bmatrix} X_c^{(0)} \\ Y_c^{(0)} \\ Z_c^{(0)} \end{bmatrix}
\end{aligned} \tag{4.1}$$

where ω, φ and κ are the Euler angles describing rotation in the x, y and z axes, respectively, as described by Heikkilä and Silvén (1997). We then estimate the intrinsic camera parameters that describe the focal length f, the scale factor s_u, and the image center (u_0, v_0). The pinhole

[5]The origin in the world's coordinate system is usually determined manually during calibration by identifying some point on a calibration pattern.

model projects a point (x_i, y_i, z_i) to the image plane by

$$\begin{bmatrix} \tilde{u}_i \\ \tilde{v}_i \end{bmatrix} = \frac{f}{z_i} \begin{bmatrix} x_i \\ y_i \end{bmatrix}. \tag{4.2}$$

We can estimate the focal length and the image center by direct linear transformation (Gruen and Huang, 2001). Here, we first solve the linear transformation from object coordinates (X_i, Y_i, Z_i) to image coordinates (u_i, v_i) by introducing a 3×4 homogeneous matrix $\mathbf{A}$,

$$\begin{bmatrix} u_i w_i \\ v_i w_i \\ w_i \end{bmatrix} = \mathbf{A} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}. \tag{4.3}$$

The parameters of $\mathbf{A}$ are then obtained by eliminating w_i, which is a free parameter. This is accomplished by defining matrix $\mathbf{L}$ using a set of N observed calibration points (X_i, Y_i, Z_i) with known position relative to the origin in the world co-ordinate system. Given a set of calibration points, the calibration matrix is defined as

$$\mathbf{L} = \begin{bmatrix} X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0 & -X_1u_1 & -Y_1u_1 & -Z_1u_1 & -u_1 \\ 0 & 0 & 0 & 0 & X_1 & Y_1 & Z_1 & 1 & -X_1v_1 & -Y_1v_1 & -Z_1v_1 & -v_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_i & Y_i & Z_i & 1 & 0 & 0 & 0 & 0 & -X_iu_i & -Y_iu_i & -Z_iu_i & -u_i \\ 0 & 0 & 0 & 0 & X_i & Y_i & Z_i & 1 & -X_iv_i & -Y_iv_i & -Z_iv_i & -v_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & Y_N & Z_N & 1 & 0 & 0 & 0 & 0 & -X_Nu_N & -Y_Nu_N & -Z_Nu_N & -u_N \\ 0 & 0 & 0 & 0 & X_N & Y_N & Z_N & 1 & -X_Nv_N & -Y_Nv_N & -Z_Nv_N & -v_N \end{bmatrix}. \tag{4.4}$$

We then represent $\mathbf{A}$ in one dimension as

$$\mathbf{a} = [a_{11}, a_{12}, a_{13}, a_{14}, a_{21}, a_{22}, a_{23}, a_{24}, a_{31}, a_{32}, a_{33}, a_{34}]^\top,$$

where a_{ij} is the element in the i^{th} row and j^{th} column of $\mathbf{A}$. Then we solve the equation

$$\mathbf{La} = \mathbf{0} \tag{4.5}$$

with least squares (Tsai, 1987; Heikkilä and Silvén, 1997). We can finally obtain our desired parameters by the decomposition

$$\mathbf{A} = \lambda \mathbf{V}^{-1}\mathbf{B}^{-1}\mathbf{F}\mathbf{M}\mathbf{T} \tag{4.6}$$

where **M** and **T** are rotation and translation matrices defined in equation 4.1 and

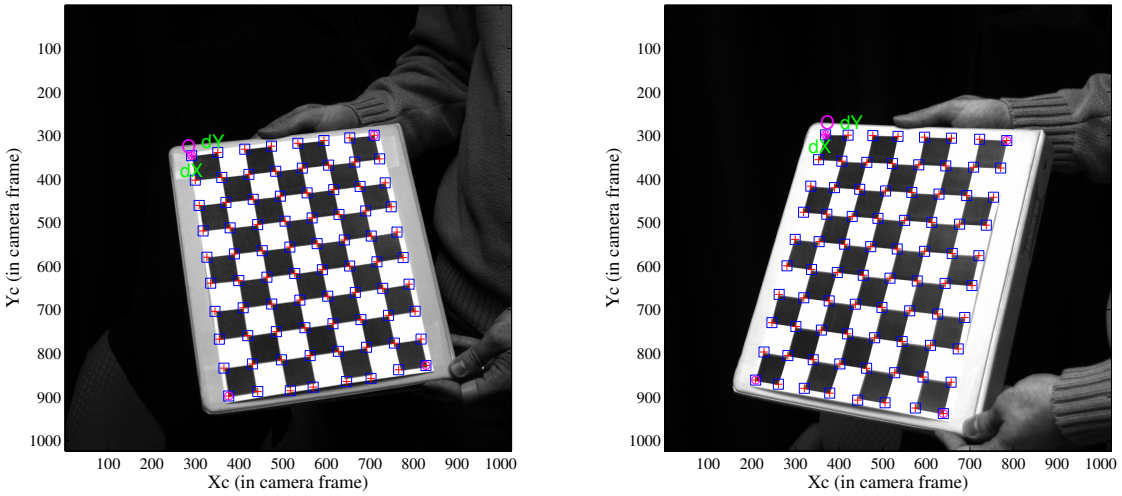
$$\begin{aligned}
\mathbf{V} &= \begin{bmatrix} 1 & 0 & -u_0 \\ 0 & 1 & -v_0 \\ 0 & 0 & 1 \end{bmatrix} \\
\mathbf{B} &= \begin{bmatrix} 1+b_1 & b_2 & 0 \\ b_2 & 1-b_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
\mathbf{F} &= \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}
\end{aligned} \tag{4.7}$$

with (b_1, b_2) providing linear distortion effects, as proposed by Melen (1994). Figure 4.8 shows a calibration pattern used in the recording of the TORGO database, where each square has a known geometry and the intersections of these squares represent our calibration points in equation 4.4.

In practice, lenses also incorporate nonlinear distortions that cannot be estimated by this method. For example, since lenses typically have radial distortion (e.g, the 'fisheye' lens) (Slama, 1980) this distortion is approximated by

$$\begin{bmatrix} \Delta\tilde{u}_i^{(r)} \\ \Delta\tilde{v}_i^{(r)} \end{bmatrix} = \begin{bmatrix} \tilde{u}_i \left(k_1 r_1^2 + k_2 r_2^4 + \ldots\right) \\ \tilde{v}_i \left(k_1 r_1^2 + k_2 r_2^4 + \ldots\right) \end{bmatrix} \tag{4.8}$$

where $k_1, k_2, \ldots$ are the coefficients of radial distortion (typically one or two is sufficient (Heikkilä and Silvén, 1997)) and $r_i = \sqrt{\tilde{u}_i^2 + \tilde{v}_i^2}$. This is performed by a nonlinear estimation and correction process described by Heikkilä and Silvén (1997) and Bouguet (1999) and



(a) (b)

Figure 4.8: Example calibration images of left (a) and right (b) video images in TORGO.

implemented in MATLAB by Bouguet (2010). In TORGO, each phosphorescent marker is manually annotated by a human during post-processing in the first frame of both camera sequences. The centroid of this marker is computed by binarizing the black-and-white video image with an empirically-determined threshold[6] and computing the center-of-mass of the white region within a given window. These windows are allowed to move and stretch in the computed direction of motion given a maximum velocity threshold relative to the video frame rate. This allows the markers to be tracked automatically without human intervention beyond the initial annotation.

Given known rotation and translation matrices ($\mathbf{M}_\alpha$ and $\mathbf{M}_\beta$, and $\mathbf{T}_\alpha$ and $\mathbf{T}_\beta$) for each of our two cameras, α and β relative to a single point of origin in the world co-ordinate system, it is relatively simple to compute the relative translation and rotation between these cameras by generalizing equation 4.1. Extensions to computing this relationship, including estimates of error, are discussed further in Hartley and Zisserman (2004).

[6]This is relatively easy, since the markers are much brighter than their surroundings.

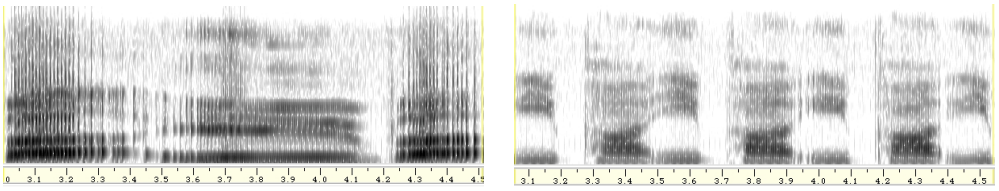
4.4 Aspects of dysarthric speech in TORGO

There are a number of features which differentiate dysarthric and non-dysarthric speech in our recorded data. Table 4.1 shows the proportion of phonemes that were mispronounced according to manner of articulation for dysarthric speech. Plosives are mispronounced most often, with substitution errors exclusively caused by errant voicing (e.g. */d/* for */t/*). By comparison, 5% of corresponding plosives in total are mispronounced in non-dysarthric speech. Furthermore, the prevalence of deleted affricates and plosives in word-final positions, almost all of which are alveolar, does not occur in the corresponding non-dysarthric speech data.

	SUB (%)			DEL (%)		
	i	m	f	i	m	f
plosives	13.8	18.7	7.1	1.9	1.0	12.1
affricates	0.0	8.3	0.0	0.0	0.0	23.2
fricatives	8.5	3.1	5.3	22.0	5.5	13.2
nasals	0.0	0.0	1.5	0.0	0.0	1.5
glides	0.0	0.7	0.4	11.4	2.5	0.9
vowels	0.9	0.9	0.0	0.0	0.2	0.0

Table 4.1: Proportion of phoneme substitution (SUB) and deletion (DEL) errors in word-initial (i), word-medial (m), and word-final (f) positions across categories of manner for dysarthric data.

Figure 4.9 exemplifies some typical acoustic contrasts between dysarthric and non-dysarthric speech in TORGO, namely the divergences in speed. Figures 4.10 and 4.11 show the durations of various steady-state phonemes (i.e., vowels and consonants, respectively) averaged across the dysarthric and control groups of TORGO. All vowels produced by dysarthric speakers are significantly slower than their non-dysarthric counterparts at the 95% confidence interval and can be up to twice as long, on average. This might partially be explained by an increase of brief



(a) (b)

Figure 4.9: Repetitions of */iy pcl p ah/* over 1.5s by (a) a male speaker with athetoid CP, and (b) a female control in the TORGO database. Dysarthric speech is notably slower and more strained than regular speech.

staccato gaps in exhalation during sonorants. We note that the divergence of the nasal consonants are most severe, which may be indicative of poor control of the velum, but the degree of this divergence does not significantly outweigh those among the other consonants.

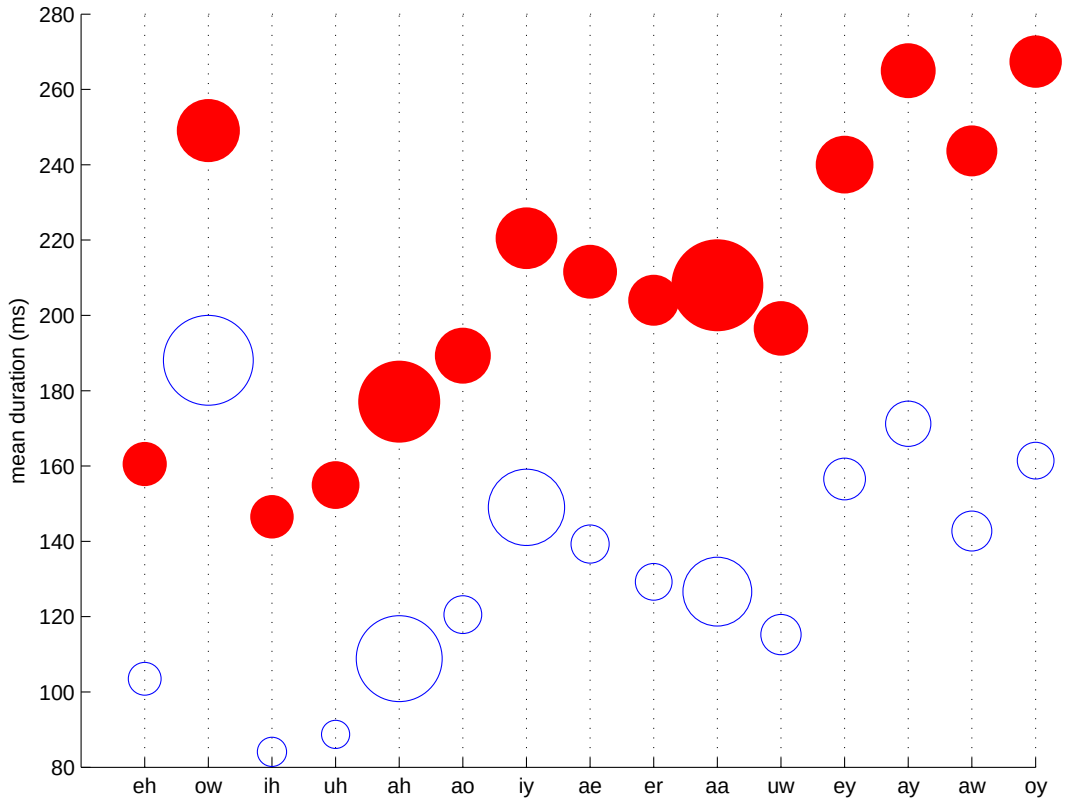


Figure 4.10: Duration of vowels among dysarthric speakers (filled circles) and control speakers (unfilled circles). The heights of the circles correspond to the average duration, in milliseconds, of the associated vowel and the radii of the circles represent one standard deviation of the data. Vowels are sorted from left to right according to increasing divergence between groups, with diphthongs displaying the greatest divergence.

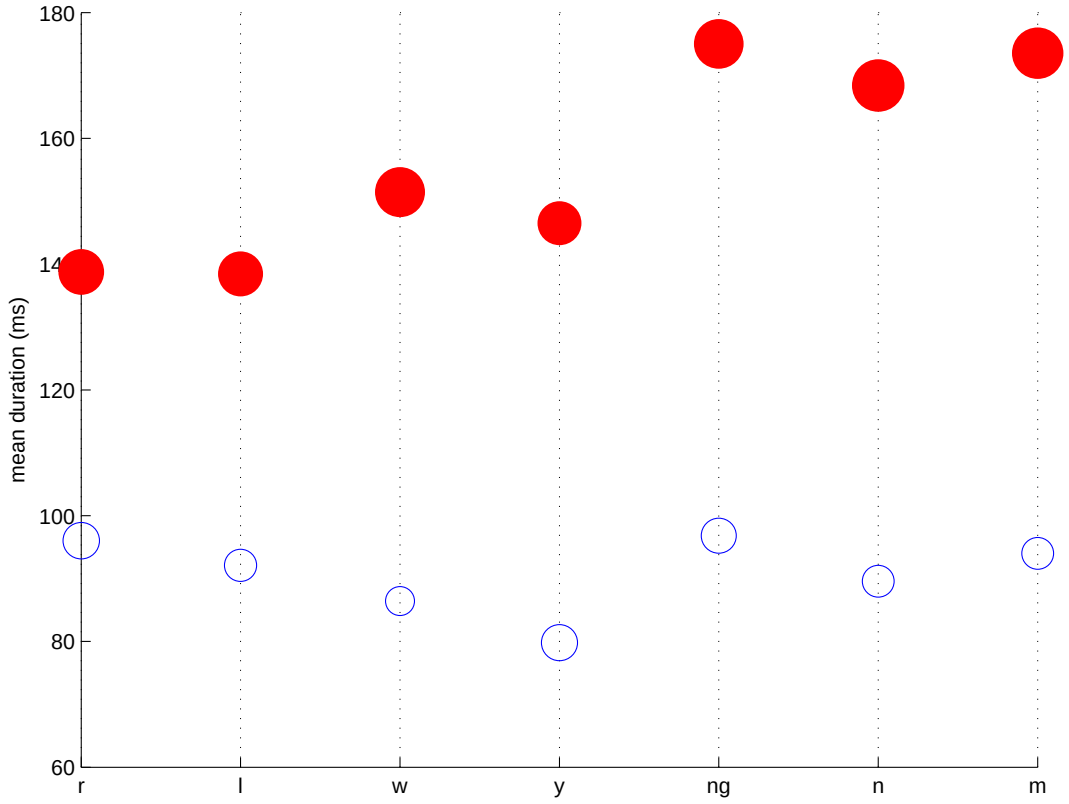


Figure 4.11: Duration of selected consonants among dysarthric speakers (filled circles) and control speakers (unfilled circles). The heights of the circles correspond to the average duration, in milliseconds, of the associated consonant and the radii of the circles represent one standard deviation of the data. Consonants are sorted from left to right according to increasing divergence between groups.

Chapter 5

Discriminative classification with discretized articulation

This chapter describes the use of theoretical and empirical knowledge of the vocal tract for atypical speech in labelling segmented and unsegmented sequences. These combined models are compared against discriminative models such as neural networks, support vector machines, and conditional random fields. These experiments constitute the first divergent step from traditional speech recognition.

Here, we concentrate on EMA recordings in the TORGO and MOCHA databases. Unlike the MOCHA database, TORGO includes points outside the midsaggital plane, namely the two lip corners and one point behind each ear, but not on the velum. In addition to typical issues of speech data collection such as the need to suppress environmental noise, the development of the TORGO database has incurred some additional challenges specific to the population. Decreased control of salivation and an increased risk of a severe gag reflex among cerebrally palsied participants can make placing coils on the tongue very difficult, so approximately 12% of EMA data from dysarthric individuals does not include the rearmost tongue positions. Involuntary movement such as shaking or extension of the neck also presents a problem for video recording, as the points on the face become occluded.

Section 5.1 describes the characteristics of the classification mechanisms used in the following experiments. Sections 5.2 and 5.3 describe experiments using acoustic observations in baseline systems and with discrete theoretical knowledge, respectively. Section 5.4 describes models that are adapted from models including explicit articulatory observations. Finally, section 5.5 summarizes the findings and provides a mechanism to transform between non-dysarthric and dysarthric data spaces.

5.1 Classification methods

Throughout the following experiments we apply five classification methods which are described below.

5.1.1 Hidden Markov models (HMM)

The default baseline is a tristate left-to-right triphone HMM with observation likelihoods at each state computed over mixtures of 16 Gaussians through marginalization amenable to normal expectation-maximization training with Baum-Welch and Viterbi decoding. Details of the HMM used in these experiments are summarized in section 2.3.2. Prior to training each HMM, the Gaussian mixtures for all states are first initialized to a common Gaussian mixture obtained by performing k-means clustering with full covariance over all data for the associated triphone. If fewer than 5 examples of the triphone exist, data for the associated monophonic root are used instead. This approach to dealing with sparse triphone data is taken for all other classification methods as well.

5.1.2 Latent-dynamic conditional random fields (LDCRF)

The discriminative latent-dynamic conditional random field is a sequence classifier differing from the HMM in that its estimation of the distribution over a sequence of labels $\mathbf{l}$ (where the i^{th} label $l_i \in \mathscr{L}$ for some vocabulary of labels $\mathscr{L}$) does not model the observation prior $P(\mathbf{o})$,

as shown in eq. 5.1. This model extends traditional conditional random fields in that it models an intrinsic sequential substructure using hidden states, and differs from 'hidden state' CRFs in that labels are assigned dynamically on a frame-by-frame basis, rather than once to the entire sequence (Morency, Quattoni, and Darrell, 2007).

In CRFs, the parameter set θ defines the weights ($\theta_k \in \theta$) applied to *feature functions* f_k of the graphical model, which are analogous to state and observation variables in HMMs (see Lafferty, McCallum, and Pereira (2001)). In fact, the parameters θ are analogous to logarithms of the conditional probabilities present between variables in HMMs (i.e., transition probabilities and state-specific observation probabilities) and are initialized randomly. In this approach, we wish to measure the likelihood of a particular labelling $\mathbf{l}$ of an observation sequence $\mathbf{o}$ given some parameterization θ. This quantity must be computed over all possible sequences of hidden states (where $\mathbf{q}$ is a particular state sequence) that produce that label sequence, where each state q_i comes from the set $\mathscr{Q}_{l_i}$ of states associable with a particular label l_i at time i. For example, an LDCRF model for phoneme */m/* might have three hidden states (i.e., $|\mathscr{Q}_m| = 3$) which are distinguished from the states in the other phoneme models. In other words,

$$P(\mathbf{l}\,|\,\mathbf{o},\theta) = \sum_{\mathbf{q}:q_i \in \mathbf{Q}_{li}} P(\mathbf{l}\,|\,\mathbf{q},\mathbf{o},\theta)\,P(\mathbf{q}\,|\,\mathbf{o},\theta)\,, \tag{5.1}$$

where $P(\mathbf{q}\,|\,\mathbf{o},\theta)$ is the standard conditional random field formulation that defines state and transition functions (Lafferty, McCallum, and Pereira, 2001; Morency, Quattoni, and Darrell, 2007), namely

$$P(\mathbf{q}\,|\,\mathbf{o},\theta) = \frac{\exp\left(\sum_k \theta_k F_k(\mathbf{q},\mathbf{o})\right)}{\sum_{\mathbf{r}} \exp\left(\sum_k \theta_k F_k(\mathbf{r},\mathbf{o})\right)}, \tag{5.2}$$

where $F_k(\mathbf{q},\mathbf{o})$ is the sum over all state transition feature functions applicable to $\mathbf{q}$ and observation feature functions applicable to $\mathbf{o}$.

Given a training set of labelled sequences $(\mathbf{o}_i,\mathbf{l}_i)$ where $i = 1..N$, we apply conjugate gradient ascent to find the optimal parameter values $\theta^* = \arg\max_\theta L(\theta)$ given the following ob-

jective function:

$$L(\theta) = \sum_{i=1}^{N} \log P(\mathbf{l}_i \,|\, \mathbf{o}_i, \theta) - \frac{1}{2\sigma^2}||\theta||^2, \tag{5.3}$$

which is the log-likelihood of the parametrization given by the conditional log-likelihood of each training sequence $\log P(\mathbf{l}_i \,|\, \mathbf{o}_i, \theta)$ and the Gaussian prior likelihood of θ with variance σ^2. If the parameter space θ is uniformly distributed, as we assume here, σ^2 approaches infinity and we discount the second term. Further details on training LDCRFs can be found in Morency, Quattoni, and Darrell (2007).

The label sequence hypothesis $\mathbf{l}^*$ is obtained by marginalizing over the sets of states $\mathbf{Q}_{l_t}$ given the label l_i at time t,

$$\mathbf{l}^* = \arg\max_{\mathbf{l}} \sum_{\mathbf{q}:\forall q_t \in \mathbf{Q}_{l_t}} P(\mathbf{q} \,|\, \mathbf{o}, \theta^*)\,. \tag{5.4}$$

5.1.3 Neural networks (NN)

Neural networks are parallelizable, multi-layer directed graphs whose arcs and vertices are associated with weights and activation functions, respectively, that can be adjusted during back-propagation to learn complex non-linear solutions (Gluck and Myers, 1999). Neural networks tend to take longer to converge than the Baum-Welch algorithm on HMMs but may be better suited to modelling duration of steady-state patterns (Tebelskis, 1995).

In modelling speech, multiple frames of input are assigned to the input layer, and output is either a single vector identifying the phoneme, or p synthesized frames that predict the next frame of speech assuming each of p phonemes, which are then compared to the actual speech. Tebelskis (1995) claims that despite apparently attractive features, since predictive ANNs use separate networks for each class, the resulting lack of categorical discrimination yields weaker results.

The two types of NN we consider here are the feed-forward multi-layer perceptron (MLP) and the recurrent Elman network (ELM) (Elman, 1990), which are primarily distinguished by the latter's time-delayed replication of the hidden layer as additional contextual input. The

Feature	# hidden units	Feature	# hidden units
Manner	300	Voice	100
Place	200	Round	100
High/Low	100	Static	100
Front/Back	200		

Table 5.1: Number of hidden units per NN, given target feature.

output of each AF NN consists of n nodes, where n is the cardinality of the class being modeled (i.e., either AF or phone), and the i^{th} node is uniquely active when training the i^{th} value of that class. Given the presence of $21,464$ triphones in our data, this approach is not tenable for NNs that recognize triphones. In that case, 15 output neurons are used in which each of the 2^{15} possible binary output combinations are mapped to a unique triphone (or a 'null' triphone not considered in classification). The sizes of hidden layers in AF neural networks are based empirically on similar work on non-dysarthric speech (Scharenborg, Wan, and Moore, 2007; Frankel, Wester, and King, 2007) and shown in table 5.1. All NN triphone classifiers contain 500 hidden units.

Activation functions at each node are tan-sigmoid (i.e., $a(x) = \left[2/\left(1+e^{-2x}\right)\right] - 1$) in the hidden layer, and linear in the output layer, given a weighted sum of all inputs $x = \sum_j \omega_j a_j$, where a_j is the activation of node j and ω_j is the weight of the connection from node j to the current node, as usual. All NN training is performed by resilient back-propagation, which adjusts update values according to sign changes in partial derivatives. Here, the degree of updates is reduced if weights oscillate over several iterations and is increased when weights continually change in the same direction. This approach is faster than standard steepest descent on our data, while only requiring a modest increase in memory.

All networks are fully connected between layers and select the class having the highest posterior probability.

5.1.4 Support Vector Machines (SVM)

General maximum margin classifiers are of increasing interest in ASR due to their robustness against both sparse data (Wan and Carmichael, 2005) and rapid transient changes in acoustic sequences (Niyogi and Burges, 2002). SVMs explicitly minimize a hypothesized upper bound on the expected classification error by orienting a hyperplane between classes such that the norm of its orthogonal vector maximizes the margin between the nearest data. We use a soft-margin SVM here and extend the process to k-class discrimination by training $k(k-1)/2$ binary classifiers, each delineating two class regions (Wu, Lin, and Weng, 2003).

SVMs depend on kernel functions, κ, to describe the distance between two points of data. We consider two of these that differ slightly in the form of their input. The first kernel is a symmetric radial basis function (RBF), that generalizes to non-linear decision boundaries using the following function:

$$\kappa_{RBF}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{y})^2}{2\sigma^2}\right), \tag{5.5}$$

given vectors $\mathbf{x}$ and $\mathbf{y}$, and width parameter σ.

The second kernel, κ_{DTW}, is a sequence kernel that can be generalized to arbitrary sequences $\mathbf{u}$ and $\mathbf{v}$ having non-equal lengths, as proposed recently by Wan and Carmichael (Wan and Carmichael, 2005). This kernel exploits the notion of distance between sequences inherent in dynamic time warping (DTW), and converts it to a form amenable for use in SVMs. The approach is to convert local Euclidean distances between frame vectors to angles by projecting these d-dimensional vectors onto a unit hypersphere H centered α units from their origin in the $(d+1)^{st}$ dimension. Namely, every vector u_i is converted to the unit vector $\hat{u}_i$ sharing an origin with H by

$$\hat{u}_i = \frac{1}{\sqrt{u_i^2+\alpha^2}}\begin{bmatrix} u_i \\ \alpha \end{bmatrix}. \tag{5.6}$$

Given two unit vectors, $\hat{u}_i$ and $\hat{v}_j$ that define points on the surface of H, the angle between them is by definition

$$d_s(\hat{u}_i, \hat{v}_j) = \theta_{\hat{u}_i, \hat{v}_j} = \arccos(\hat{u}_i, \hat{v}_j). \tag{5.7}$$

Now, given these local distances, we apply *symmetric* DTW on whole sequences $\mathbf{u}$ and $\mathbf{v}$ and get the minimum global distance from the non-linear aligned Viterbi path Γ with

$$D_{global}(\mathbf{u}, \mathbf{v}) = \min_{\Gamma} \frac{1}{||\Gamma||} \sum_{p=1}^{||\Gamma||} d_s(\hat{u_p}, \hat{v_p}). \tag{5.8}$$

This distance is then converted to the kernel

$$\kappa_{DTW}(\mathbf{u}, \mathbf{v}) = \cos D_{global}(\mathbf{u}, \mathbf{v}), \tag{5.9}$$

which is symmetric if the symmetric version of DTW is used, which is a requirement for use in SVM classification. In order for the quadratic programming problem to have a definite solution, the kernel must either be a valid dot product (Russell and Norvig, 2003), or satisfy Mercer's condition, which is to say that given a real-valued kernel $\kappa(x, y)$, all square integrable functions $g(x)$ will give $\int \int \kappa(x,y)g(x)g(y)dx\,dy \geq 0$ (Vapnik, 1995). While the cosine over an aggregate of sequences is not strictly a dot-product, it has been shown to be empirically useful in speech classification nonetheless (Wan and Carmichael, 2005). For multi-category classification, directed acyclic graphs can be used to discriminate between pairs of classes until only one remains (Platt, Cristianini, and Shawe-Taylor, 2000).

5.1.5 Dynamic Bayes Networks (DBN)

Popular statistical modelling techniques such as HMMs and CRFs do not permit much flexibility in defining the relationships between variables. Standard HMMs, for example, allow only one hidden variable (in addition, e.g., to the index of a Gaussian mixture observation distribution). In practice, there may be many such variables, or variables whose values are only

intermittently known or recorded over a sequence of observations. Bayes networks provide a popular statistical framework that allows us to determine precise instantaneous conditional relationships. Traditional Bayesian learning is restricted to universal or immutable relationships and does not model dynamic systems or time-varying relationships. Dynamic Bayes networks (DBNs) are directed acyclic graphs connecting random variables that generalize the stochastic mechanisms of Bayesian learning to time sequences[1]. Given an N-variable observation sequence $Y_{1:T}^{(1:N)}$ of arbitrary length T, its likelihood is computed by 'unrolling' a 2-frame DBN to T frames, and multiplying all posteriors,

$$\begin{aligned} P(Y_{1:T}^{(1:N)}) = & \prod_{i=1}^{N} P_{B_1}(Y_1^{(i)}|par(Y_t^{(i)})) \times \\ & \prod_{t=2}^{T}\prod_{i=1}^{N} P_{B_\rightarrow}(Y_t^{(i)}|par(Y_t^{(i)})), \end{aligned} \tag{5.10}$$

where conditional distributions, $B_\rightarrow$ are drawn over adjacent frames in time for the i^{th} state at time t, $Y_t^{(i)}$ by $P(Y_t|Y_{t-1}) = \prod_{i=1}^{N} P(Y_t^{(i)}|par(Y_t^{(i)}))$, given the parents of $Y_t^{(i)}$, $par(Y_t^{(i)})$. This temporal model generalizes both the hidden Markov model, the coupled hidden Markov model, the Kalman filter, and many others (Murphy, 2002). Given a specified topology between variables and a data set $D = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(N)}\}$, where $\mathbf{Y}^{(i)}$ is a sequence of vectors of observed variables, the likelihood of D is

$$P(D; \theta, \mathscr{M}) = \prod_{i=1}^{N} P(\mathbf{Y}^{(i)}; \theta) \tag{5.11}$$

where θ is the set of parameters to the DBN [2]. The maximum likelihood parameterization is obtained by maximizing

$$\mathscr{L}(\theta) = \sum_{i=1}^{N} \log P(\mathbf{Y}^{(i)}; \theta). \tag{5.12}$$

[1] Since DBNs are acyclic, they cannot model sequences in time in which future events influence past ones, so they may not be appropriate for modelling sub-atomic structures in quantum physics; however, that is beyond the scope of this dissertation.

[2] The topology of a DBN, $\mathscr{M}$ is usually implicit in the parameters θ, but there are algorithms that can modify this structure given data, so some definitions explicitly include $\mathscr{M}$.

In general, we have hidden variables so equation 5.12 cannot be decomposed into a sum over log likelihoods of individual nodes given their parents. Instead, $L(\theta) = \log \sum_X P(Y,X\,;\theta)$, where X is the set of all hidden data and $\sum_X$ is the sum over all the possible permutations of X to obtain a marginalization. Given a distribution $\mathscr{H}(\cdot)$ over the hidden variables, we obtain a lower bound

$$\begin{aligned} L(\theta) &= \log \sum_X P(Y,X\,;\theta) \\ &= \log \sum_x \mathscr{H}(X) \frac{P(Y,X\,;\theta)}{\mathscr{H}(X)} \\ &\geq \sum_x \mathscr{H}(X) \log \frac{P(Y,X\,;\theta)}{\mathscr{H}(X)} \\ &= \sum_x \mathscr{H}(X) \log P(Y,X\,;\theta) - \sum_x \mathscr{H}(X) \log \mathscr{H}(X) \\ &= \mathscr{F}(\mathscr{H},\theta), \end{aligned} \tag{5.13}$$

where the inequality results from Jensen's inequality (Jensen, 1906). The expectation-maximization algorithm then maximizes $\mathscr{F}$ with respect to $\mathscr{H}$ and θ, respectively. The maximum at the expectation step occurs when $\mathscr{H}_{K+1}(X) = P(X \,|\, Y\,;\theta_K)$ where K is the iteration of the algorithm (Ghahramani, 1998). In the maximization step, we consider the first term in the penultimate line in equation 5.13, thus

$$\theta_{K+1} \leftarrow \arg\max_\theta \sum_X P(X \,|\, Y\,;\theta_K) \log P(Y,X\,;\theta), \tag{5.14}$$

which gives us the general learning procedure for dynamic Bayes networks.

5.2 Experiment set 1: HMM baselines

Our baseline is designed to test the accuracy of ASR as one adapts continuous speaker-independent HMM systems trained on the general population to dysarthric data, or trained on that dysarthric data alone.

5.2.1 MLLR and MAP adaptation

Adaptation of model parameters is used when the conditions in which those parameters were trained no longer reflects the conditions in which we expect new observations. For example, if a model is trained in a quiet environment but will be used in a noisy one, we wish to adjust the model parameters to reflect this new situation using a small amount of calibration data which typically is much smaller in scope than the original training data. HMM adaptation is performed in the following experiments using a combination of two standard techniques, namely maximum *a posteriori* (MAP) estimation and maximum likelihood linear regression (MLLR). In the first, given a parameter space Φ defined on HMMs as described in section 2.3.2, we assume that we have prior knowledge that can characterize a probability density $p(\Phi)$. Given a set of observation sequences $\mathbf{X}$, the MAP estimate for the ideal parameters is

$$\hat{\Phi} = \arg\max_{\Phi} p(\Phi \mid \mathbf{X}) = \arg\max_{\Phi} \left[p(\mathbf{X} \mid \Phi) p(\Phi) \right]. \tag{5.15}$$

This estimate reduces to the maximum likelihood estimate if $p(\Phi)$ is uniform, i.e., when there is no prior knowledge. Since we use continuous Gaussian mixture HMMs, we assume that the different components are mutually independent, which is standard practice (Huang, Acero, and Hon, 2001) and allows us to split the optimization problem into subcomponents. For example, the prior probability density of a Gaussian mixture b_i is

$$p_{b_i}(\vec{\omega}_i, \vec{\mu}_i, \Sigma_i) = p_{\omega_i}(\vec{\omega}_i) \prod_k p_{b_{ik}}(\vec{\mu_{ik}}, \Sigma_{ik}), \tag{5.16}$$

where $p_{\omega_i}(\vec{\omega}_i)$ is the Dirichlet prior over all M mixture weights in state i. That is,

$$\begin{aligned} p_{\omega_i}(\vec{\omega}_i) &= p_{\omega_i}(\omega_i[1], \omega_i[2], \ldots, \omega_i[M]) \\ &= \frac{\Gamma\left(\sum_{j=1}^{M} \alpha_i[j]\right)}{\prod_{j=1}^{M} \Gamma(\alpha_i[j])} \prod_{j=1}^{M} \omega_i[j]^{\alpha_i[j]-1} \end{aligned} \tag{5.17}$$

where $\alpha_i[j]$ is a hyperparameter associated with the j^{th} argument of the Dirichlet distribution for state i. We can then apply the Lagrange method by

$$\frac{\delta}{\delta \hat{\omega}_i[m]} \left(\log p_{\omega_i}(\vec{\omega}_i) + \sum_{m=1}^{M} \sum_t \xi_t(i,m) \log \hat{\omega}_i[m] \right) + \lambda = 0, \forall m \tag{5.18}$$

with the constraint that $\sum_{m=1}^{M} \hat{\omega}_i[m] = 1$. In equation 6.4, λ is the Lagrange multiplier and $\xi_t(i,m)$ is the probability that the observation at time t was generated by the m^{th} Gaussian of the i^{th} state. The solution is

$$\hat{\omega}_i[m] = \frac{\alpha_i[m] - 1 + \sum_t \xi_t(i,m)}{\sum_{l=1}^{M} (\alpha_i[l] - 1 + \sum_t \xi_t(i,l))}. \tag{5.19}$$

The density function $p_{b_{ik}}(\vec{\mu_{ik}}, \Sigma_{ik})$ in equation 5.16 is the prior probability of the k^{th} Gaussian component in state i. Optimization with respect to the means and covariances of the Gaussians is accomplished in the same manner (Gotoh et al., 1995; Woodland, 2001). Here, the form of the conjugate prior is Gaussian for $\vec{\mu_{ik}}$ multiplied by a Wishart distribution for Σ_{ik} (Gauvain and Lee, 1994). This process is iterative and can be considered as interpolated between speaker-dependent and speaker-independent models (Huang, Acero, and Hon, 2001). Here, this MAP process is embedded within a maximum likelihood regression, as described by Chesta, Siohan, and Lee (1999).

5.2.2 HMM experiments

We categorize each speaker according to his recognition rate on Nemours data using a baseline acoustic model trained on spoken transcripts of the Wall Street Journal (Lamere et al., 2003). The four speakers having word-level recognition rates below 10% with the baseline model are grouped as ‘severe’, the four with rates between 11% and 30% are grouped as ‘moderate’, and the three between 31% and 60% are grouped as ‘mild’. The control speaker had a word-level recognition rate of 84.8%. These initial recognition rates correlate well with subjective sentence-level intelligibility scores among human listeners.

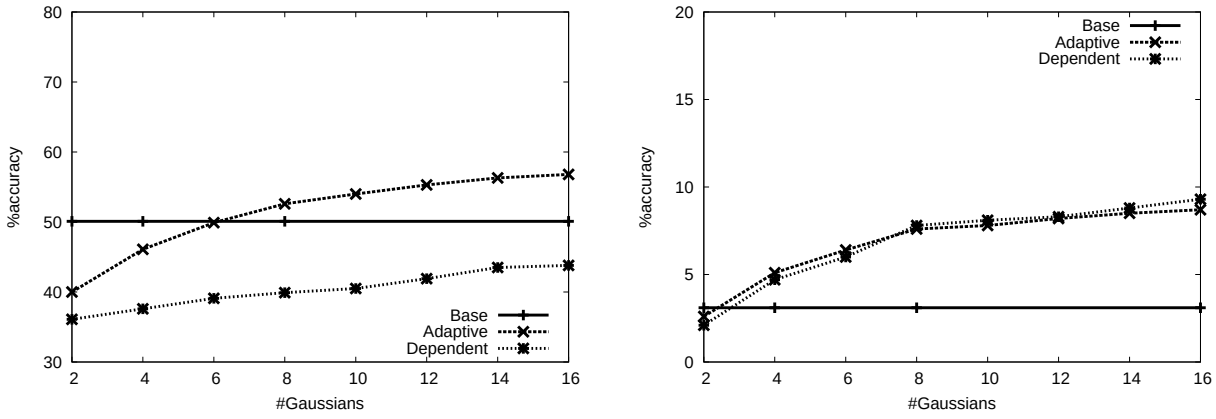
Both the dependent and adaptive models for each speaker are triphone left-right Hidden Markov Models (HMMs) with Gaussian mixture output densities decoded with the Viterbi algorithm on a lexical-tree structure augmented with a context-free grammar. For each speaker, we initialize the HMM acoustic parameters of the dependent model randomly, and initialize the adaptive model with the common WSJ-trained baseline. We independently vary the number

of Gaussians and the amount of training utterances in order to measure how precision and data coverage accommodate the variability of dysarthric speech, and apply the iterative Baum-Welch training algorithm on both models for each speaker.

Increasing the amount of training data from 20 to 132 training sentences per speaker does not show any definite improvement, with accuracy fluctuating around 3% from the mean across trials. The fact that accuracy does not increase suggests that there is not enough data in Nemours to represent intra-speaker variation, and that studies using fewer test subjects may also require more data.

Figure 5.1 shows accuracy increasing monotonically with the number of Gaussians for the mildly and severely dysarthric speakers. In all cases but the most severe, the adaptive models outperform their dependent counterparts and reduce relative error by up to 23.1% in the mild group, by 4.9% in the moderate group, and by 30.7% for the non-dysarthric speaker. This suggests that taking advantage of pre-existing models of the normal population may best suit dysarthric speakers with higher intelligibility. This tends to support the abstract conclusions of Raghavendra et al. (Raghavendra, Rosengren, and Hunnicutt, 2001), except that they also observed a clear superiority of dependency for severely dysarthric speakers. By contrast, we only observe slight SD gains over the baseline as the number of Gaussians increases, possibly due to the distribution of data.

Of the 485 insertion errors Sphinx made among the dysarthric speakers of the Nemours database, /*ih*/ and /*d*/ were the most common with 63 and 51, respectively. The most commonly dropped phonemes by these speakers were /*b*/ (118), /*s*/ (111), /*w*/ (60), /*f*/ (55) and /*l*/ (48), among 649 deletion errors in total. The most common substitutions were /*ng*/ for /*n*/ (125) and surprisingly /*t*/ for /*uw*/ (87), /*ey*/ for /*ih*/ (84) and /*t*/ for /*n*/ (77). These observations suggest that ASR software might be made more accessible to dysarthric speakers by increasing robustness against consonant variations in general. Sawhney and Wheeler perform similar experiments on the Nemours database using an unspecified segmental context-independent phoneme recognizer trained on the TIMIT database (Sawhney and Wheeler, 1999).



Mildly dysarthric speakers

Severely dysarthric speakers

Figure 5.1: ASR accuracy measured against acoustic model precision (i.e., number of Gaussians). Baselines represent models trained on the WSJ corpus.

They report $\sim 42\%$ accuracy on speaker dependent models, and $\sim 25\%$ on independent models averaged across all dysarthric speakers, the former which is significantly higher than observed with Sphinx. Sawhney and Wheeler also show far fewer substitution errors initially, where $/d/$ for $/t/$ (13), $/p/$ for $/t/$ (12), and $/ng/$ for $/t/$, $/ih/$ for $/uw/$ (11) are the most common.

5.3 Experiment set 2: Discrimination with acoustics alone

We begin by considering the effects of dysarthria in systems trained solely from acoustic data, which is a considerably more common scenario than one in which kinematic data are available. However, given phonemic annotations, we can infer articulatory features as representative of articulatory knowledge, as described in section 2.4. We train each classifier both to identify articulatory features from acoustics and to identify phones given both acoustics and their identified AFs. In all cases, acoustic data are sampled at 16kHz and converted to 42-dimensional feature vectors consisting of 0^{th}- to 12^{th}-order Mel-frequency cepstral coefficients, log energy, and δ and $\delta\delta$ coefficients. Neither δ nor $\delta\delta$ are appended to AF components, due to the relative parsimony of tracking changes in step functions. We apply 10-fold cross-validation on random permutations of 90% training and 10% test data for each speaker in the Nemours

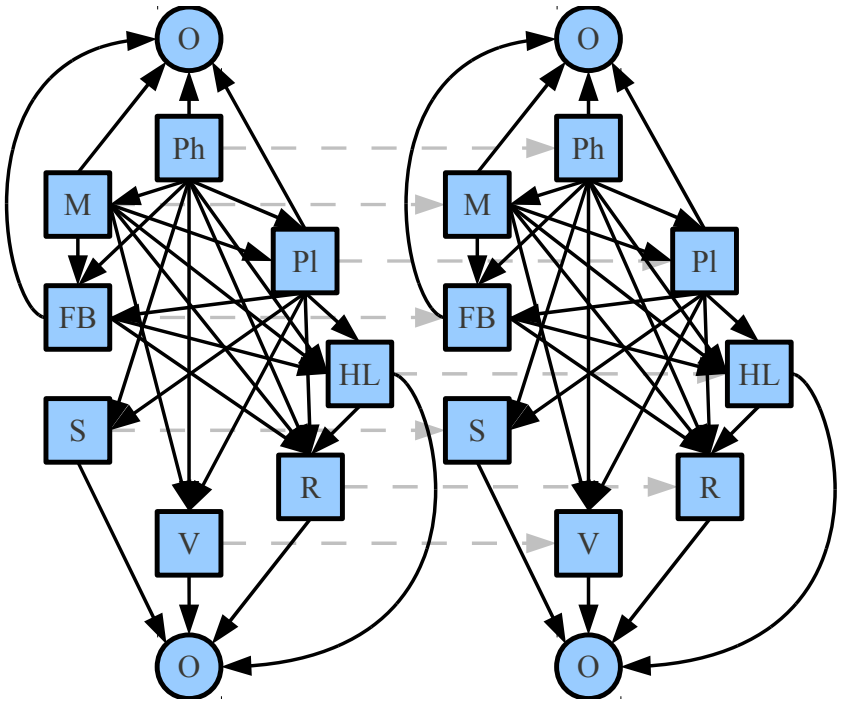
database. Training sets consist of approximately $93,000$ frames per speaker on average.

We test two topologies of AF variables within DBNs. The first is based on similar work by Frankel et al. (Frankel, Wester, and King, 2007), and is shown in figure 5.2a. The second is a sparser version of that DBN with certain conditional dependencies removed in order to reduce the complexity of parameterization, as shown in figure 5.2b. All AFs are observed in the DBN during training but inferred during testing.

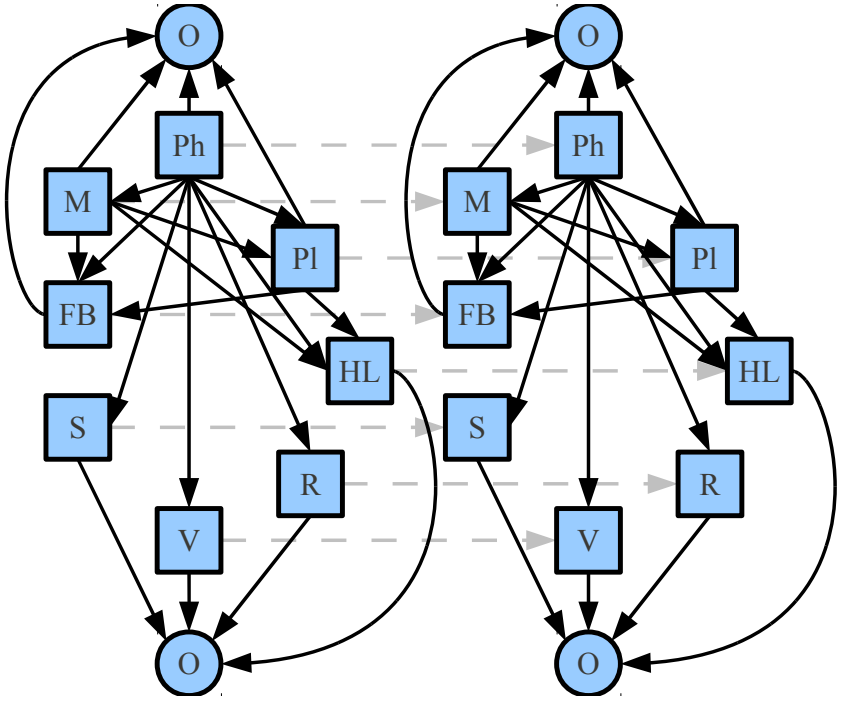
5.3.1 AF classification with acoustics

Frame-level accuracies for each AF averaged over all speakers in the Nemours database are summarized in table 5.2 for each classifier. Both the LDCRF and SVM methods are exceptionally proficient at classifying *Manner* and *Place*, which are highly related, and poor at classifying the *Round* AF despite its low cardinality. This suggests that there is some other aspect of those AFs that affects discriminability, at least for SVMs. The *nil* class is the most poorly recognized in three of the four AFs having it. The most frequently confused pairs for each AF are shown in table 5.3, which is generally consistent with the literature for non-dysarthric speakers (Kirchhoff, 1999).

In general, SVM methods outperform NN on average by 4.9% to 9.3% absolute and provide a 19.8% relative error reduction on dysarthric speech. On the control subject, AF models achieved 74.3% accuracy for MLP, and 77.6% for RBF, on average. Results of the SVM methods with this speaker were comparable though slightly lower than in similar research on non-dysarthric AF recognition by SVM (Chaudhari and Picheny, 2009), although that work included far more training data. Other research on speaker-independent recurrent neural networks for AF recognition on regular speech report frame-level accuracies between 85.9% and 91.8% given $\sim$2.2 million frames (Frankel, Wester, and King, 2007).



(a)



(b)

Figure 5.2: Two-frame dynamic Bayes networks with articulatory features, (a) DBN-F (default), and (b) DBN-F (sparse). Nodes **Ph** and **O** represent phoneme, state, and MFCC observations. All other variables are highlighted in table 2.1. Inter-frame conditional links are dashed for clarity.

	Average accuracy (%)									
		DBN-F			NN		SVM			
Feature	HMM	default	sparse	LDCRF	MLP	ELM	RBF	DTW	μ	σ
Manner	23.8	36.5	32.1	**69.1**	22.1	30.2	66.8	65.4	43.3	19.0
Place	33.9	39.6	34.7	**58.8**	35.5	41.9	**58.3**	56.5	44.9	10.4
Hi/Low	48.6	52.9	49.0	56.2	53.0	**58.7**	55.7	55.9	53.8	3.3
Voice	76.1	77.8	76.3	79.2	78.7	**81.3**	76.8	78.1	78.0	1.6
Front/Back	49.0	48.4	49.4	54.0	48.2	52.1	**55.1**	**55.7**	51.5	2.9
Round	60.4	64.5	60.6	64.8	68.9	**69.7**	55.3	54.0	62.3	5.4
Static	61.3	65.2	63.6	**70.2**	64.2	66.5	67.3	69.2	65.9	2.8
μ	50.4	55.0	52.2	64.6	52.9	57.2	62.2	62.1		

Table 5.2: Classifier accuracies averaged over dysarthric speakers (best of row in bold) for AF recognition.

Feature	1^{st}	2^{nd}
Manner	[vowel]→[approx.] (12%)	[vowel]→[retro.] (8%)
Place	[nil]→[alv.] (10%)	[nil]→[dental] (7%)
Hi/Low	[nil]→[low] (14%)	[mid]→[low] (11%)
Voice	[unvoiced]→[voiced] (68%)	[voiced]→[unvoiced] (32%)
Front/Back	[nil]→[central] (19%)	[nil]→[back] (17%)
Round	[non]→[nil] (26%)	[nil]→[non] (22%)
Static	[stat.]→[dyn] (54%)	[dyn]→[stat] (46%)

Table 5.3: Most frequent errors for each AF ([actual] → [hypothesis] (% total error)).

Effects of dysarthria

Figure 5.3 shows the overall accuracy of each classification technique according to speaker intelligibility as determined by the Frenchay Dysarthria Assessment (see section 4.2.2). These results show a general success of SVM and LDCRF methods across all speakers, especially the less intelligible ones, and a global increase in accuracy with intelligibility. Two speakers perturb this trend, however, with noticeable drops in accuracy as indicated for speakers 'RK' and 'BB' in the figure. These two individuals share exceptionally poor tongue elevation and lateral movement relative to the rest of the group which seems to account for their especially low accuracy with *High/Low* and *Front/Back* AFs, which are predicated on tongue movement and position. According to their Frenchay assessments, 'RK' and 'BB' both had scores of $0/9$ for tongue elevation and scores of $0/9$ and $1/9$ for lateral tongue movement, respectively. Only two other speakers, 'SC' and 'BK', had similarly poor assessments of tongue control, with the latter also having the lowest intelligibility of all speakers.

Table 5.4 shows the recognition rates for the two AFs under consideration against the average of all other AFs given an HMM system. Here, the four speakers identified as having particularly bad tongue movement have recognition rates for *Front/Back* and *High/Low* that are all between 5.3% and 10.2% lower than for other AFs, on average. By contrast, *Front/Back* and *High/Low* AFs are better recognized than other AFs, on average, for all speakers without the identified tongue deficit.

Within these AFs, follow-up analysis revealed linear correlation coefficients up to 0.95 between increased formant deviation and decreased tongue function. While overall intelligibility may be useful in predicting general trends in figure 5.3, it is an aggregate measure of the functions of component articulators, and may be overridden for speakers having more localized disabilities.

	Front/Back	High/Low	avg. other AF
BK	31.2	32.5	37.8
SC	35.3	34.7	41.3
RK	37.1	36.9	47.1
BB	48.6	49.0	55.8
avg. of others	55.3	54.5	54.2

Table 5.4: Recognition rates (% correct) of *Front/Back* and *High/Low* AFs compared with the average recognition rates across all other AFs for 4 speakers and the average of all other speakers given an HMM recognition system.

5.3.2 Phone recognition with acoustics

Finally, we consider whether AFs are useful in identifying phones. For each of our modelling techniques, we construct three triphone classifiers that differ by the nature of their observations. Each of these is trained either with acoustics, with estimated AFs, or with acoustics and estimated AFs concatenated together. Here, AF estimates are derived both from the outputs of models having the same type as the phone classifier, or from the outputs of the LDCRF model which represents the best average AF estimates achievable. No other heterogeneous combination of models is attempted. Given that the LDCRF is the most accurate AF classifier, we find it unlikely that other combinations would yield much greater accuracies.

All models are applied over whole unsegmented utterances as continuous tasks. Specifically, each frame of speech is classified by NN and SVM methods given short windows of input observations, as described earlier. Connected-state models of the same type (i.e., either HMM, LDCRF, and DBN) are connected together so that all phonemes are equally likely to follow all others. This approach is taken to evaluate these models as substitutes to standard acoustic models, as is our intention. The use of language models is explored in section 5.4.4. Accuracy is measured at the frame level by converting estimated triphones to their monophonic roots.

The results in table 5.5 indicate relative error reductions of 8.8% and 11.2% merely by re-

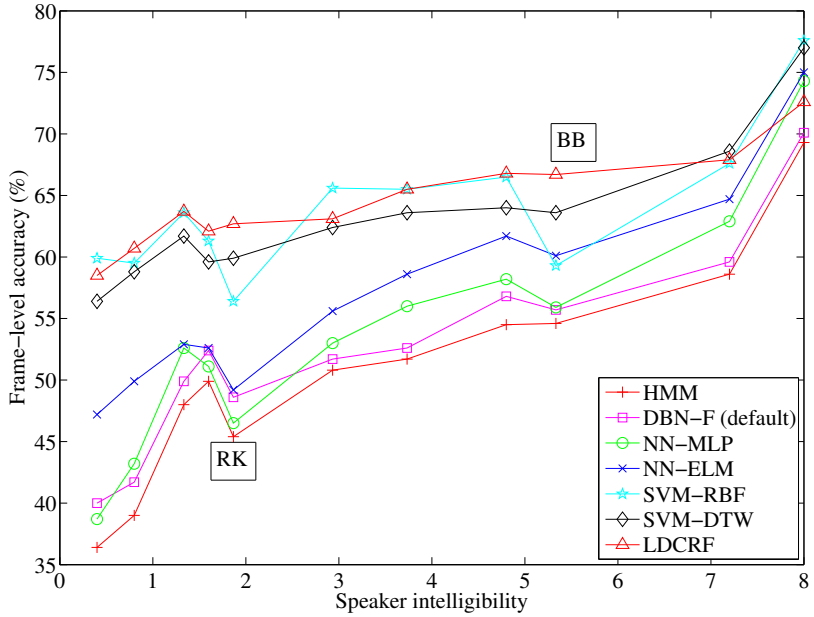


Figure 5.3: Average classifier accuracy against assessed intelligibility level.

placing an HMM model with an SVM-DTW and an LDCRF, respectively, given only dysarthric acoustics, which is significant at the 99% confidence level. Relative error reduction is the absolute difference between the error rates of the two systems under comparison divided by the higher error rate of the two. Extending observation vectors to include AFs reduces error relatively by between 0.5% and 7.1% over associated acoustic-only models, which represent significant improvements at the 99% confidence level for all models except LDCRF. This result shows a clear benefit of incorporating AFs into the input of all but one type of acoustic model. Since the seven AFs are so rarely unanimously correct, they alone cannot be used to infer the respective phone in practice, and further research should investigate whether it is more useful to limit the use of AFs to some subset. No explicit weighting was applied between the MFCC and AF components of heterogeneous vectors, but the relative importance of these parts and their covariances are inferred during training by each of these classifiers implicitly.

	MFCC	AF	MFCC+AF	MFCC+AF$_{LDCRF}$
HMM	33.8	7.4	36.3	37.6
DBN-F (default)	34.1	7.8	37.1	37.9
DBN-F (sparse)	33.4	7.5	37.0	38.1
LDCRF	41.2	16.0	41.5	41.5
NN-MLP	31.9	5.8	34.8	35.3
NN-ELM	36.7	11.7	40.2	40.7
SVM-RBF	38.4	16.2	38.7	40.1
SVM-DTW	39.6	17.9	41.0	41.3

Table 5.5: Phone classification accuracies (%) at the frame level averaged over speakers with dysarthria given various types of observation. Estimated AFs are concatenated with MFCC observations either by using AF estimators of the same type (MFCC+AF) or by using the LDCRF AF estimator (MFCC+AF$_{LDCRF}$).

5.4 Experiment set 3: Initialization from articulation

There is increasing evidence that replacing the Gaussian mixture observation densities of HMMs with limited Bayes nets representing spacial vocal tract kinematics can improve accuracy over acoustic-only models for non-dysarthric speakers (Markov, Dang, and Nakamura, 2006). Although it is impractical to perform articulography on each speaker we wish to model, we can make use of publicly available databases such as MOCHA or TORGO to provide baseline kinematic knowledge that we can adapt to speakers for whom only acoustic data is available. This scenario is explored in this section.

We conflate the instantaneous EMA position data from the MOCHA and TORGO databases by first reducing their dimension to $N_p = 4$ or $N_p = 8$ principal components by singular value decomposition specific to each phone in which $K = 4$, $K = 8$, or $K = 16$ mean vectors are computed according to the sum-of-squares error function. During training, the DBN variable $\mathbf{A}$ is the observed index of the mean vector nearest to the current frame of EMA data at time t.

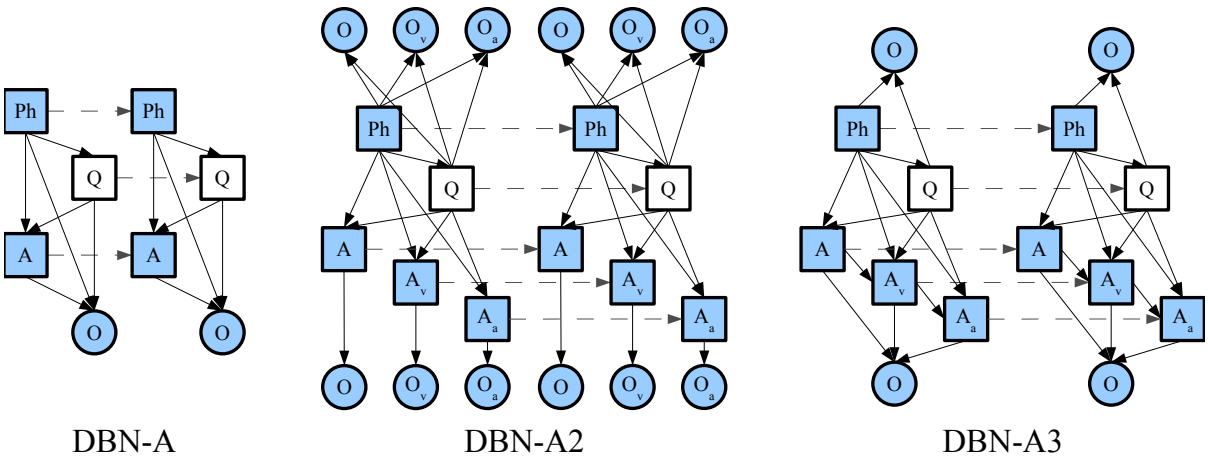


Figure 5.4: Two-frame dynamic Bayes networks with EMA measurements differing by their connectivity. Nodes **Ph**, **Q**, **O**, **A**, $\mathbf{A_v}$, and $\mathbf{A_a}$ represent phoneme, state, MFCC observations, and EMA position, velocity, and acceleration, respectively. Inter-frame conditional links are dashed for clarity.

During inference, this variable is hidden and we marginalize over all its values when computing the likelihood. In this way, DBN-A is essentially a DBN representation of an HMM with the hidden mixture index replaced by observed quantized articulation. Similarly, we follow the same procedure on the velocities and accelerations of the articulators, producing indices $\mathbf{A_v}$ and $\mathbf{A_a}$. These variables are used in alternative DBN topologies DBN-A2 and DBN-A3. In the first, the observation vector is trisected, with each 14-dimensional vector (i.e., MFCC, δ, and $\delta\delta$) being conditioned on **P**, **Q**, and one of **A**, $\mathbf{A_v}$ and $\mathbf{A_a}$. The second alternative structure, DBN-A3, conditions $\mathbf{A_a}$ on $\mathbf{A_v}$, and $\mathbf{A_v}$ on **A** and conditions the 42-dimensional observation vector on all variables. The three kinematic DBN topologies are shown in figure 5.4.

The MOCHA database uniquely includes velum position and the TORGO database uniquely includes left and right lip corners. Both databases include three midsagittal tongue positions, upper and lower lip, and lower incisor positions.

		DBN-A		**DBN-A2**		**DBN-A3**	
N_p	K	MOC.	TOR.	MOC.	TOR.	MOC.	TOR.
4	4	57.6	58.9	56.9	57.4	57.8	57.5
	8	66.8	67.2	66.5	67.2	66.8	67.1
	16	68.9	69.0	69.1	68.8	69.3	69.3
8	4	63.3	62.7	63.4	63.0	63.8	63.6
	8	71.0	70.8	71.1	71.3	71.3	71.6
	16	72.4	72.4	72.2	72.1	72.7	72.7
16	4	64.7	65.0	65.1	65.2	65.2	65.2
	8	72.5	72.6	72.4	72.4	72.7	72.5
	16	73.6	73.8	73.6	73.9	74.0	74.1

Table 5.6: Accuracies of frame-level phone recognition across kinematic DBNs with varying quantities of principal components, N_p, and Gaussians, K, for speaker-dependent, non-dysarthric speech. Data is obtained from the MOCHA and TORGO databases.

5.4.1 Recognition with non-dysarthric speech

The three DBN models are compared on non-dysarthric speech across the number of principal components, N_p, and the number of Gaussians, K, used in quantization. Reducing dimensionality across heterogeneous acoustic/articulatory observations in this way has previously been shown to preserve important features of both articulatory and acoustics (Wrench and Richmond, 2000; Fukuda and Nitta, 2003). Results of frame-level phone recognition are summarized in table 5.6. Across all topologies and data, $N_p = 16$ is significantly more accurate than $N_p = 8$ at the 95% confidence level and $N_p = 4$ at the 99% confidence level. Results across MOCHA and TORGO, and across the three topologies, are statistically indistinguishable. However, both DBN-A2 and DBN-A3 are several times slower than DBN-A to train.

5.4.2 Retraining dysarthric acoustics

We retrain models initialized on non-dysarthric data given new dysarthric acoustics. We retrain each kinematic DBN with dysarthric acoustics by making indices $\mathbf{A}$, $\mathbf{A_v}$, and $\mathbf{A_a}$ hidden after training on non-dysarthric acoustic/articulatory data (MOCHA and TORGO), and retraining on dysarthric acoustics (Nemours and TORGO). All HMM and kinematic DBN models are trained with EM and smoothed junction-tree inference, given their hidden variables. When retraining the HMM, DBN, NN, and LDCRF models to dysarthric speech, we initialize new instantiations with the distributions learned on regular speech and retrain on speaker-specific acoustics until convergence. All training of the fully observed DBN-F is with maximum likelihood, so retraining involves concatenating the non-dysarthric and dysarthric training data and learning once. SVM models from previous sections are not included here, due to the dissimilar manner in which those models are trained. In all cases, training data include all phones observed during testing and are applied to the 46 phones that MOCHA, Nemours, and TORGO have in common. Data are randomly split into 90% training and 10% test data. We split dysarthric TORGO and Nemours data by speaker into three categories according to the level of intelligibility as determined by the Frenchay assessment (Enderby, 1983). Individuals with intelligibility levels between 0 and 25% are 'severe', between 25% and 62.5% are 'moderate', and between 62.5% and 87.5% are 'mild'.

Table 5.7 shows the frame-level accuracy of unsegmented phone labelling on speaker-dependent and speaker-retrained distributions for each model, according to the severity of dysarthria. Here, DBN-A, -A2, and -A3 are trained to mixtures of 16 Gaussian clusters determined by unreduced (16-dimensional) articulatory data. These results show an increasing benefit of retrained over dependent training on dysarthric speech as intelligibility increases, with absolute rates of improvement of 0.86%, 1.96%, and 6.03% on severely, moderately, and mildy dysarthric speech, respectively. Although speaker-dependent kinematic models are more successful than other models, they do not adapt as well as the DBN-F or LDCRF models.

These results are generally consistent with similar work that retrained acoustic-only DBNs

		sev	mod	mild	ctrl
HMM	Depend.	14.1	27.8	51.6	72.8
	Retrain.	16.8	32.1	58.9	-
LDCRF	Depend.	15.2	28.0	51.8	73.5
	Retrain.	16.8	32.4	59.1	-
DBN-F	Depend.	15.0	28.0	51.6	73.3
	Retrain.	16.7	32.3	59.4	-
DBN-A	Depend.	16.4	31.1	54.2	73.6
	Retrain.	16.2	31.7	58.3	-
DBN-A2	Depend.	16.3	31.1	54.3	73.6
	Retrain.	16.3	31.9	58.4	-
DBN-A3	Depend.	16.4	31.3	54.5	73.8
	Retrain.	16.5	32.0	58.7	-
NN-MLP	Depend.	15.5	28.6	51.4	72.6
	Retrain.	16.0	29.0	58.6	-
NN-ELM	Depend.	15.6	30.5	51.2	72.7
	Retrain.	16.1	30.7	57.5	-

Table 5.7: Average accuracy (%) of correctly labelled phones of speaker-dependent and speaker-retrained (EMA-initialized) models, according to the severity of dysarthria.

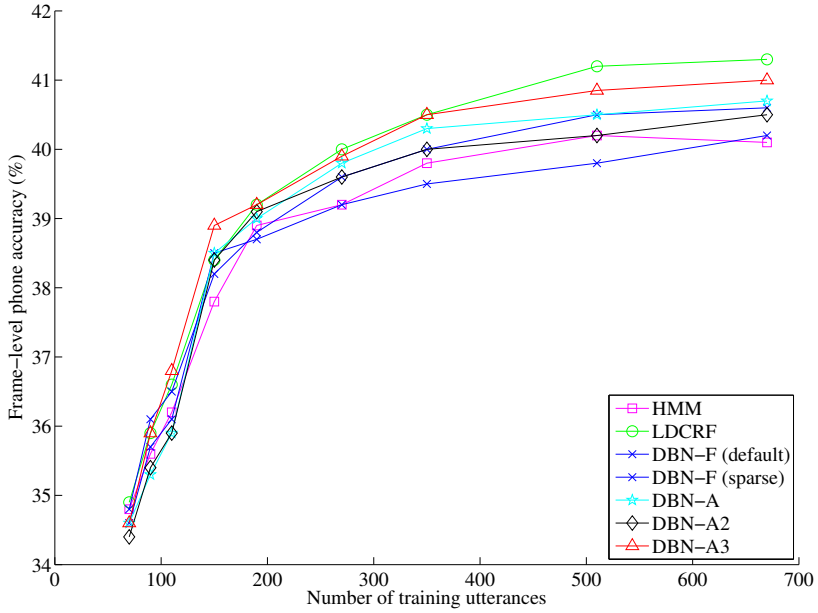


Figure 5.5: Labelling accuracy of four models with increasing amount of dysarthric retraining data.

to Japanese kinematic data (Markov, Dang, and Nakamura, 2006) over 1 or 2 iterations of EM. That work showed error reduction of between 0.7% and 3.8% on phone classification among a selection of alternative speaker-dependent DBNs relative to a baseline DBN. The performances of DBN-F and HMM are also consistent with similar work on non-dysarthric models (Frankel, Wester, and King, 2007).

5.4.3 Effect of sample size

We examine the effect of increased sample size by retraining non-dysarthric models with cross-sections of data selected uniformly at random among all speakers with dysarthria in Nemours and TORGO, and testing on proportionally increasing test sets. Figure 5.5 suggests that as the amount of dysarthric speech is increased, the LDCRF model outperforms all others, with an absolute error reduction of 1.2% over HMM with 670 training utterances for retraining.

5.4.4 The use of language models

Although this work is concentrated on articulatory enhancements to acoustic models, in practice the latter are rarely used alone without some contextual information. Often, bigrams are used in order to weigh the likelihood of transitioning from one phoneme or word to another. Since our data consist of many single-word utterances, we consider phoneme bigrams in which the probability of one phoneme p_t following another p_{t-1} at time t is given by

$$P(p_t \,|\, p_{t-1}) = \frac{N_{(p_{t-1},\, p_t)}}{N_{(p_{t-1})}}, \tag{5.20}$$

where $N_{(p_{t-1})}$ is the total number of occurrences (i.e., whole sequences of frames) of p_{t-1} in the data and $N_{(p_{t-1},\, p_t)}$ is the total number of times p_t immediately follows p_{t-1} in the data. We gather these counts from TIMIT which includes 2472 unique bigrams covering $172,460$ adjacent pairs of phonemes, as determined by the included phonetic annotations. Similarly, the unigram probability of phoneme p_t is determined from the same data by

$$P(p_t) = \frac{N_{(p_t)}}{\sum_\rho N_{(\rho)}}, \tag{5.21}$$

where ρ is iterated over all 61 phonemes in the training data.

In order to implement systems that incorporate either bigram or unigram information, we first train individual HMM and DBN-A models for each phoneme, as before, where training data consist of whole sequences of phonemes. The result is 61 HMMs and 61 DBN-A models, each consisting of 3 states with reflexive and left-to-right transitions. We first connect the HMMs together and the DBN-As together by creating transitions from the last state of each phoneme model to the first state of all other phoneme models of the same type. First, the probabilities associated with these transitions are their bigram probabilities of equation 5.20. Expectation-Maximization is then performed for 2 iterations on each of the large connected HMM and DBN-A models in order to learn reflexive transition probabilities on the last state for each phoneme without over-fitting. This is a common approach producing all-phoneme ergodic models (Miyazawa, 1993). This process is then repeated, but with initial transition probabilities between phoneme models being derived from their unigram probabilities (equation 5.21).

Severity	HMM		DBN-A	
	unigram	bigram	unigram	bigram
sev	17.2	20.8	17.4	21.0
mod	33.4	37.3	34.1	37.9
mild	60.1	63.5	60.5	63.7
ctrl	74.0	74.2	74.2	74.6

Table 5.8: Average frame-level accuracy (%) of unsegmented phoneme labelling given ergodic HMMs and DBN-As with unigram and bigram phoneme transition probabilities.

Given these connected models, the same data as in section 5.4.2 is used to measure the average proportion of correctly labelled phones given phoneme models trained by the speaker-dependent method. Table 5.8 shows the frame-level phoneme recognition accuracies of each model across the same speaker intelligibility levels of table 5.7. While there are clear improvements in accuracy, these are still lower than one would expect if full word-level bigrams were used, given more testing data. Trigram models were not attempted due in part to this relative sparsity of data and to inherent constraints of the implementation.

5.5 Discussion

Preceding sections summarize an extensive series of experiments concerning the recognition of dysarthric speech given knowledge of speech production. Our purpose is to discover which combinations of articulatory knowledge and modelling give improved rates of recognition for individuals with speech disabilities. In situations where no kinematic data is available, incorporating theoretical articulatory knowledge into generative dynamic Bayes networks shows some improvement in phone recognition over traditional HMM models, but far greater improvements are possible through the application of discriminative methods, particularly latent-dynamic conditional random fields. However, generative DBN models that are trained by aligned kinematic electromagnetic articulographic data give the greatest improvement over

standard models, also outperforming acoustic-only discriminative methods.

The following subsections explore possible explanations for some of the behaviour observed in these experiments above.

5.5.1 Synthesizing dysarthric acoustics

We compare the generative abilities of DBN-A and DBN-F on our data. We iteratively set **Ph** to each phone in the available DBN-A and DBN-F models and marginalize over all other variables to get the distribution on **O** from which we sample virtual data for each phone. These generated likelihood functions are fitted with Gaussians and compared with the true MFCC distributions of each phone by means of Kullback-Leibler relative divergence. The likelihood functions generated by DBN-F diverge from true distributions by a factor of 0.22016 on regular speech and by 0.2246 on dysarthric speech. However, while virtual DBN-A data diverge from true data by a factor of 0.1690 for regular speech, speaker-retrained DBN-As for dysarthric speech diverge by 0.3378, on average, from true phone MFCC distributions. This disparity is exemplified in figure 5.6.

5.5.2 Statistical transformation of articulator space

In order to better understand some recognition results, we relate the distributions of the vowels in acoustic and articulatory spaces across dysarthric and non-dysarthric speech. Vowels in acoustic space are characterized by the steady-state positions of the first two formants as determined automatically by applying pre-emphasis and the Burg algorithm (Press et al., 1992). Vowels in articulatory space are characterized by the positions of the articulators when their accelerations are minimum. We fit Gaussians to these data, as exemplified in figure 5.7 for the most frequent vowels in TORGO and compute the entropy of the data within these distributions. Surprisingly, the entropies of these distributions were relatively consistent across dysarthric (34.6 nats) and non-dysarthric (33.3 nats) speech, with some exceptions (e.g., *iy*).

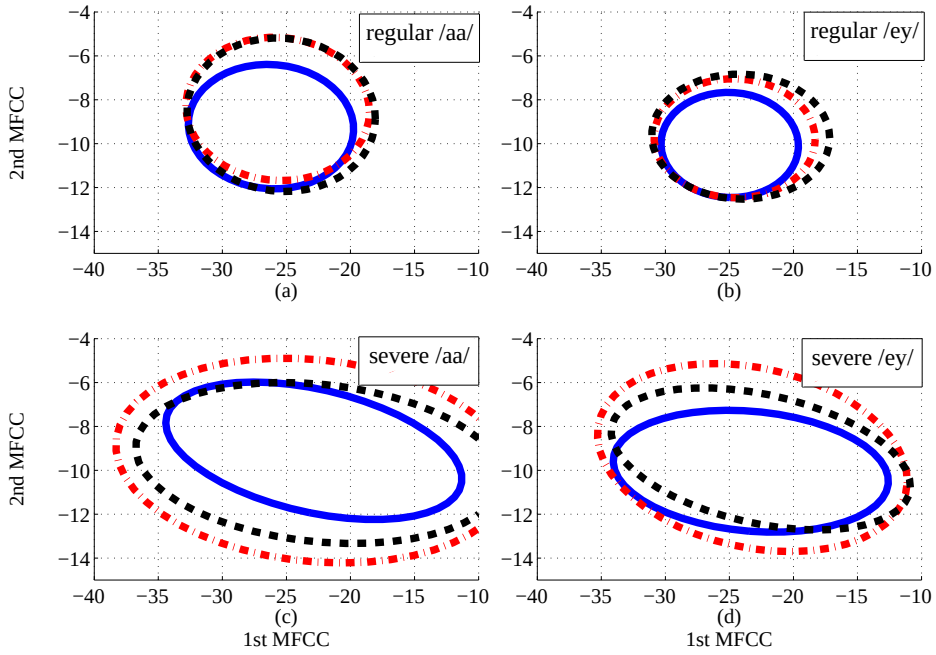


Figure 5.6: Contours representing 2 standard deviations of Gaussians fitted to real data (solid line), samples from DBN-F (dashed line), and samples from DBN-A (dash-dotted line) on the first two mel-frequency cepstral coefficients. Subfigures represent (a) regular speech (/aa/), (b) regular speech (/ey/), (c) severely dysarthric speech (/aa/), and (d) severely dysarthric speech (/ey/).

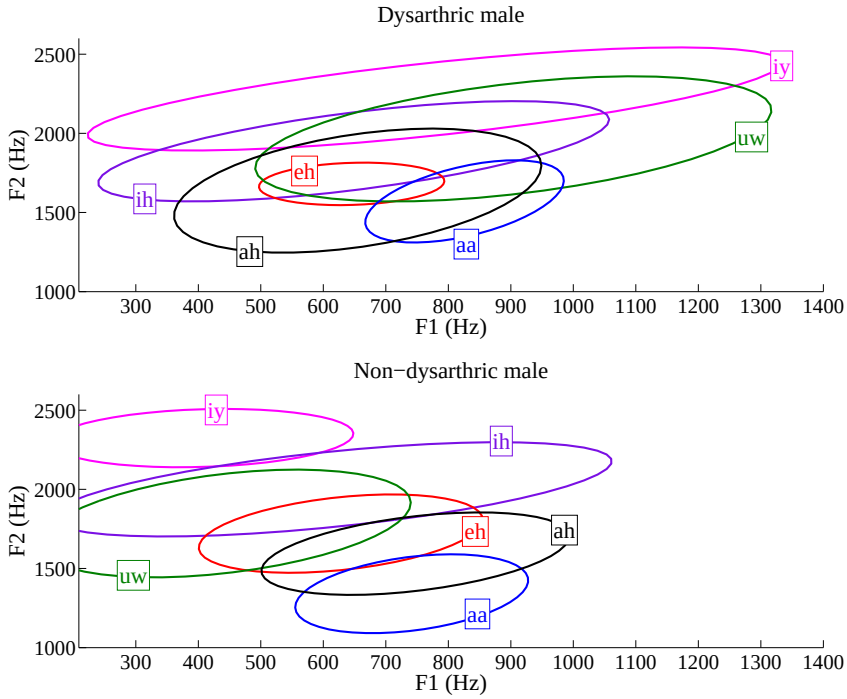


Figure 5.7: Contours showing first standard deviation in F1 vs F2 space for distributions of the six of the most frequent vowels in continuous speech for the dysarthric and non-dysarthric males from the TORGO database.

However, vowel spaces overlap considerably more in the dysarthric case signifying that, while speakers with CP can be nearly as consistent as speakers without dysarthria in the acoustic space, the locations of their targets in that space are not as discernible. Moreover, we note linear correlation coefficients of over 0.95 between F2 standard deviation and the extent of tongue protrusion, as determined by the Frenchay assessment described above. Some research has shown larger variance among dysarthric vowels relative to our findings (Kain et al., 2007). This may partially be due to our use of natural connected speech as data, rather than restrictive consonant-vowel-consonant non-words.

In an attempt to tease apart the acoustic targets in dysarthric speech, and to give them meaningful conditioning articulatory variables within the DBN framework, we learn statistical mappings between dysarthric and non-dysarthric speech. Namely, we learn two functions, f

and g, which produce the expected frames in the acoustic and articulatory spaces of a speaker with dysarthria given corresponding frames for a regular speaker. For each function, we define Gaussian distributions, $\mathbf{N}(\cdot)$, for each phone, p, by the means of the regular and dysarthric speech, respectively $\mu_p^{(x)}$ and $\mu_p^{(y)}$, and the covariances, $\Sigma_p^{(xx)}$, of the regular speech. We can then apply the following statistical transformation function between non-dysarthric acoustic vectors, x, and their dysarthric counterparts, y:

$$\begin{aligned} f(\mathbf{x}) &= E(\mathbf{y}|\mathbf{x}) \\ &= \sum_{i=1}^{P} h_i \left[\mu_i^{(y)} + \Sigma_i^{(yx)} \left(\Sigma_i^{(xx)} \right)^{-1} \left(\mathbf{x} - \mu_i^{(x)} \right) \right], \end{aligned} \tag{5.22}$$

where

$$h_i(\mathbf{x}) = \frac{\alpha_i N\left(\mathbf{x}; \mu_i^{(x)}, \Sigma_i^{(xx)}\right)}{\sum_{j=1}^{P} \alpha_j N\left(\mathbf{x}; \mu_j^{(x)}, \Sigma_j^{(xx)}\right)}, \tag{5.23}$$

α_p is the proportion of the occurrences of phone p in the data, and $\Sigma_p^{(yx)}$ is the cross-covariance matrix in phone p across speakers with and without dysarthria. The function g is identical in articulatory space, but with vectors defined by articulator positions from EMA. We learn cross-covariance matrices on aligned sequences from both sets of speakers. Since each speaker in the TORGO database recites the same set of phrases, we achieve frame-by-frame alignment by applying dynamic time warping on corresponding acoustic segments of pre-annotated speech, and applying the resulting alignment on the raw articulatory data. This is effectively the reverse of the approach suggested by Hosom et al., who propose transforming dysarthric acoustic space to regular acoustic space in order to be made more intelligible (Hosom et al., 2003).

Once we have the transformed acoustic and articulatory spaces of a regular speaker that resemble those of our speaker with dysarthria, we quantize the latter using k-means clustering and train the DBN-A model as described in section 5.4. We then update this model given either dysarthric acoustics only (see section 5.4.2), or aligned dysarthric acoustics and quantized articulation. These three models are then tested with either additional transformed acoustics,

Training Data	Retraining Data	Testing Data	Accuracy (%)
Trans. acous. ∪ Trans. artic.	-	Trans. acous.	72.9
		Dys. acous.	72.6
	Dys. acous.	Trans. acous.	73.7
		Dys. acous.	73.4
	Dys. acous. ∪ Dys. artic.	Trans. acous.	74.3
		Dys. acous.	74.2

Table 5.9: Phoneme accuracy of DBN model trained and retrained across various combinations of transformed regular acoustics and articulation, and dysarthric acoustics and articulation.

or actual dysarthric acoustics. These results are shown in table 5.9. Notably, models tested with the transformed speech show slightly higher accuracies of recognition than models tested on the target dysarthric speech, which may be an artifact of suprasegmental effects of dysarthria on intelligibility. We note that models initialized with transformed regular speech perform better than any dependent or adaptive combination for dysarthric test data in section 5.4.2.

Chapter 6

Task-dynamics in ASR

Although results in previous chapters may be applicable to improving current ASR systems for the dysarthric population, these successes were tempered by the relatively unconstrained nature of the underlying statistical methods and the short-time observation windows. Several fundamental phenomena of dysarthria such as increased disfluency, longer sonorants, and reduced pitch control (Rudzicz et al., 2008) could not be readily represented in any of the methods described before. Representing speech as a sequence of non-overlapping (though restricted) syllabic or phonemic units is the basis for automatic speech recognition, and has been useful in describing certain types of dysarthria where speech is broken into syllables either due to respiratory problems or to improve overall intelligibility (Ziegler and Maassen, 2004). However, such models cannot inherently account for more complex aspects of articulatory organization, for which parallel and self-organizing theories may be more appropriate (Smith and Goffman, 2004). In order to study the long-term dynamics of dysarthria in particular, and speech generally, we require a framework of dynamical systems into which our data can be explored.

The theory of task-dynamics is a combined model of skilled articulator motion and the planning of vocal tract configurations (Saltzman, 1986; Saltzman and Munhall, 1989). This theory introduces the notion that the dynamic patterns of speech are the result of overlapping gestures, which are high-level abstractions of goal-oriented reconfigurations of the vocal tract

such as bilabial closure or velar opening. Indeed, the quantal theory of speech is based on the empirical observation that acoustics depend on a relatively discrete set of distinctive underlying articulatory configurations (Stevens, 1972; Stevens and Keyser, 2010).

This chapter introduces this theory in section 6.1. Section 6.2 describes a method to derive the instantaneous articulatory positions in task-dynamics given only acoustic information. Section 6.3 describes the information content in observed articulatory data and proposes dysarthria as a noisy-channel perturbation of underlying task-dynamics. Section 6.4 describes a mechanism of using inferred task-dynamics to correct errors made by traditional speech recognition, generally. Finally, section 6.5 describes a new method for deriving the parameters of task-dynamics from observed data.

6.1 Tract variables and task dynamics

The interaction between linguistic and motor capabilities is not easily diagnosed, or understood. In psycholinguistic theory the linguistic hierarchy is often decomposed into conceptual, syntactic, morphological and phonological representations independent from the motor system through which these aspects are realized (Levelt, Roelofs, and Meyer, 1999). Articulatory phonology (AP) bridges the gap between phonetics and phonology by encapsulating them as the physical (constraining) and abstract (planning) stages of a single system (Goldstein and Fowler, 2003). Articulatory phonology has also been directly applied to the study of speech disorders such as apraxia (Bahr, 2005), which is believed to affect the exact part of the neurological interface that AP describes (Dogil and Mayer, 1998).

Task-dynamics is a combined model of skilled articulator motion and the planning of abstract vocal tract configurations (Saltzman, 1986). Here, the dynamic patterns of speech are the result of overlapping *gestures*, which are high-level abstractions of reconfigurations of the vocal tract. An instance of a gesture is any articulatory movement towards the completion of some speech-relevant goal, such as bilabial closure, or velar opening. The progenitors of this theory

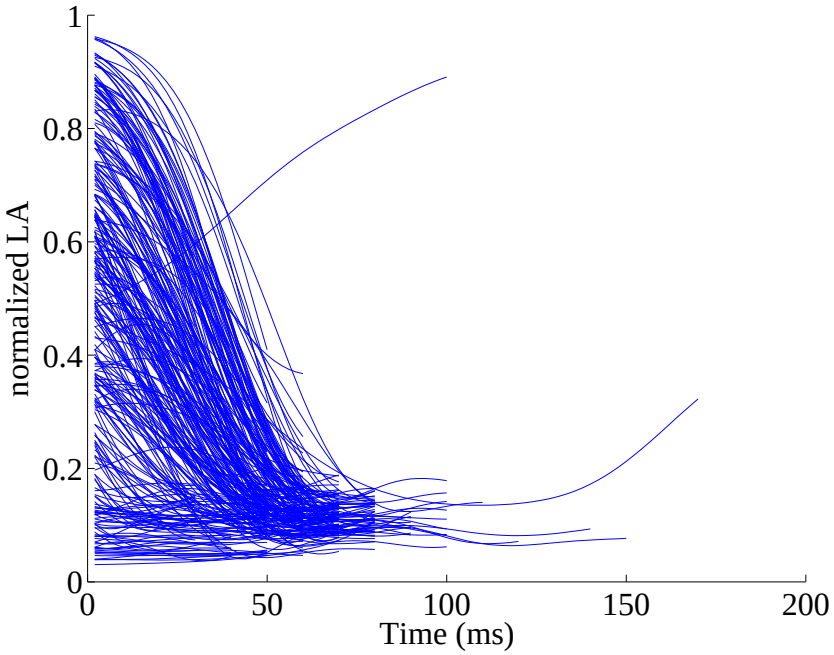


Figure 6.1: Lip aperture (LA) over time for all instances of phoneme */m/* in MOCHA.

claim that all the implicit spatiotemporal behaviour underlying speech is the result of the interaction between the abstract *intergestural* dimension (between tasks) and the geometric *interarticulator* dimension (between physical actuators) (Saltzman and Munhall, 1989). Each gesture in task-dynamic theory occurs within one of the following *tract variables* (TVs): lip aperture (LA), lip protrusion (LP), tongue tip constriction location (TTCL) and degree (TTCD)[1], tongue dorsum constriction location (TDCL) and degree (TDCD), velum (VEL), glottis (GLO), and lower tooth height (LTH). For instance, a gesture to close the lips would occur within the LA variable and would set that variable close to zero, as shown in figure 6.1 where the relevant articulatory goal of lip closure is evident.

For example, the syllable *pub* consists of an onset (*/p/*), a nucleus (*/ah/*), and a coda (*/b/*). Four gestural goals are associated with the onset, namely the shutting of GLO and of VEL, and the closure and release of LA. Similarly, the nucleus of the syllable consists of three goals, namely the relocation of TBCD and TBCL, and the opening of GLO. The presence and extent

[1]Constriction *locations* generally refer to the front-back dimension of the vocal tract and constriction *degrees* generally refer to the top-down dimension.

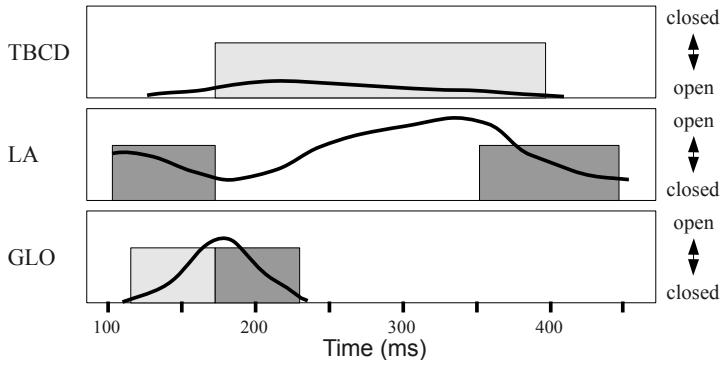


Figure 6.2: Canonical example *pub* from Saltzman and Munhall (1989) representing overlapping goals for tongue blade constriction degree (TBCD), lip aperture (LA), and glottis (GLO). Boxes represent the present of discretized goals, such as lip closure. Black curves represent the output of the TADA system.

of these gestural goals are represented by filled rectangles in figure 6.2. Inter-gestural timings between these goals are specified relative to one another according to human data as described by Nam and Saltzman (2003).

The dynamic influence of each gesture in time on the relevant tract variable is modeled by the following non-homogenous second-order linear differential equation (Saltzman and Munhall, 1989):

$$Mz'' + Bz' + K(z - z^*) = 0, \tag{6.1}$$

where z is a 9-dimensional vector of the instantaneous positions of each tract variable, and z' and z'' are its first and second differentials. Here, M, B, and K are diagonal matrices representing mass, damping, and stiffness coefficients, respectively, and z^* is the 9-dimensional vector of target (equilibrium) positions. This model is built on the assumption that the tract variables are independent and do not interact dynamically, although these matrices could be adjusted to reflect dependencies, if desired (Nam and Saltzman, 2003). If the targets z^* of this equation are known, the identification of linguistic intent becomes possible. For example, given that a bilabial closure occurs simultaneously with a velar opening and glottal vibration, we can identify the intended phone as $/m/$. This represents a dimensionality reduction for classification of an

instantaneous frame of speech from 14 (typical of Mel-frequency cepstral coefficients) to 9, of which only 3 are relevant in the example above.

A primary problem to solve on the path towards task-dynamic speech recognition is how to infer the tract variable values in the first place. Obviously, for general recognition we cannot measure the tract variables directly and will have to estimate them from acoustics. One principled approach may be to first estimate the positions of the physical articulators, and to use these to infer gestural activity. Once the m instantaneous articulator positions, Φ, are known, we can infer the instantaneous impulsion of the tract variables (z) towards their targets with the following direct kinematic equation of active control:

$$\Phi''_A = J^{(P)}(\Phi)\left(M^{-1}\left[-BJ(\Phi)\Phi' - K\left(z - z_0\right)\right]\right) - J^{(P)}J'(\Phi,\Phi')\Phi' \tag{6.2}$$

where $J(\Phi)$ represents the Jacobian matrix of partial derivatives of Φ over time, $J^{(P)}(\Phi) = W^{-1}J(\Phi)^T\left(J(\Phi)W^{-1}J(\Phi)^T\right)^{-1}$ is the Jacobian pseudo-inverse, W is an articulatory weighting matrix specific to each gesture, and M, B, K, and z_0 carry the same meaning as in equation 6.1 (Saltzman and Munhall, 1989)[2]. Obviously, the values of Φ and its derivatives are specific to the geometry of the virtual speaker assumed during simulation.

Articulatory data consists of spoken utterances and their aligned articulator positions as described in chapter 4. In order to convert the articulator space to tract variable space, we transform the midsagittal articulatory data using a combination of principal component analysis and sigmoid activation functions. For example, we describe VEL by calculating the first principal component of velum motion in the midsagittal plane, finding the minimum and maximum deviations from the mean in this transformed space, and applying a sigmoid to that unidimensional space to retrieve a real function on $[0..1]$. Similarly, the first and second principal components of the distance between UL and LL are used for the determination of lip aperture and protrusion, respectively, the first and second principal components of TT are used for the determination of TTCL and TTCD, respectively, and the first and second principal components

[2]Note that an augmented form of this equation includes an orthogonal projection operator that eliminates extraneous motion by including supplementary dissipative forces proportional to articulatory velocity.

of TB are used for the determination of TBCL and TBCD, respectively. Voicing detection on energy below 150 Hz (O'Shaughnessy, 2000) is used to estimate the GLO tract variable.

6.2 Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics

Differences between speakers are the result of purely endogenous phenomena distinguished by their mechanics of articulation. Such distinctions cannot be codified into automatic speech recognition (ASR) systems that are agnostic of speech production, however. Although direct measurements of the vocal tract are not typical during speech recognition, it is nonetheless desirable to find an accurate method of projecting acoustic speech data onto a lower-dimensional space which is more indicative of the linguistic intentions of the speaker, namely, to the space of physical properties of vocal tract motion. Evidence that such inversion takes place during speech perception in humans suggests that the discriminability of speech sounds depends powerfully on their production (D'Ausilio et al., 2009). For example, the motor theory of speech perception is a branch of speech science that assumes that the same processes that are involved in the production of speech are used to decode speech acoustics during perception (Liberman et al., 1957; Liberman et al., 1967; Liberman and Mattingly, 1985). This theory may also be applicable, or at least provide ecological support to our design decisions.

Despite the one-to-many relationship in acoustic-to-articulatory inversion (Roweis, 1999; Ananthakrishnan, Neiberg, and Engwall, 2009), such protestation has not limited research in this area. For example, Richmond et al. (Richmond, King, and Taylor, 2003) estimated the 2-dimensional midsagittal positions of 7 articulators given kinematic data using both a multilayer perceptron and discriminatively trained Gaussian mixture models to within 0.41 mm and 2.73 mm. Toda et al. (Toda, Black, and Tokuda, 2008) achieved almost identical results on the same data by applying expectation-maximization using both minimum mean-squared error and maximum likelihood estimation to a Gaussian mixture mapping function with low-pass

filtering. Simpler approaches achieved similar results (errors less than 2mm, typically around 1mm) using simple vector quantization with an appropriate number of vectors (Hogden et al., 1996; Hogden et al., 2007). This work, called MALCOM (Maximum Likelihood Continuity Map), uses band-limited low-pass filtering of around 8–15 Hz on articulatory trajectories, which is a common practice (Yehia, 2002). These trajectories are low-dimensional and smooth relative to their acoustic counterparts. Here, vector-quantized codes that partition the acoustic space are each associated with a Gaussian probability density in a pseudo-articulatory space defined in the frequency domain based on multidimensional kinematic data. During recognition, acoustics are encoded in VQ space and then a smooth trajectory is computed in the pseudo-articulatory space that maximizes the conditional likelihood given the discretized acoustics (Hogden, 1996; McDermott and Nakamura, 2006).

One commonality in existing work is that the target dimensions consist of the absolute physical positions of points in the vocal tract. Typical points include the upper and lower lips (**UL**, **LL**), the upper and lower incisors (**UI**, **LI**), the tongue tip, body, and dorsum (**TT**, **TB**, **TD**), and the velum (**V**). Despite the popularity of this approach, neither its generalizability among speakers nor its representation of linguistic intent has been justified. Why would the physical position of the upper lip be as explicative of intent or of acoustic consequence as a measure of the distance between the lips, for example?

In this section we estimate features of the vocal tract from acoustics using adaptive kernel canonical correlation analysis (KCCA). We choose features of the vocal tract derived from the theory of task dynamics, as described below.

6.2.1 Adaptive KCCA

Canonical correlation analysis (CCA) is a popular technique in statistical analysis used in a variety of contexts, including communication theory and statistical signal processing, to measure linear relationships between sets of variables. Given vector variables $\mathbf{x} \in \mathbb{R}^{\mathbf{m_x}}$ and $\mathbf{y} \in \mathbb{R}^{\mathbf{m_y}}$, CCA finds a pair of directions $\omega_x \in \mathbb{R}^{m_x}$ and $\omega_y \in \mathbb{R}^{m_y}$ such that the correlation $\rho(\mathbf{x}, \mathbf{y})$ is max-

imized between the two projections $\omega_x^T \mathbf{x}$ and $\omega_y^T \mathbf{y}$. Given joint observations $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 ... \mathbf{x}_N]^T$ and $\mathbf{Y} = [\mathbf{y}_1 \mathbf{y}_2 ... \mathbf{y}_N]^T$, where $\mathbf{x}_i$ co-occurs with $\mathbf{y}_i$, CCA is equivalent to finding projection vectors ω_x and ω_y that maximize

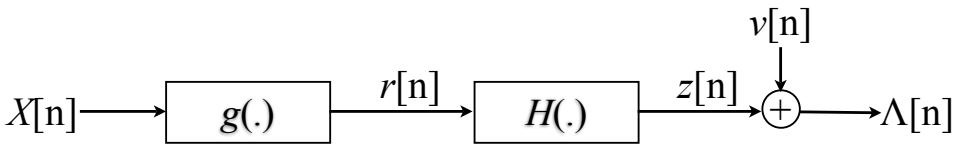
$$\rho(\mathbf{X}, \mathbf{Y}; \omega_x, \omega_y) = \frac{\omega_x^T \mathbf{X}\mathbf{Y}^T \omega_y}{\sqrt{\omega_x^T \mathbf{X}\mathbf{X}^T \omega_x}\sqrt{\omega_y^T \mathbf{Y}\mathbf{Y}^T \omega_y}}. \tag{6.3}$$

Although this method can find good linear relationships between sets of data, it is incapable of capturing nonlinear relationships, which limits its application in many aspects of speech. In order to overcome this limitation, we employ the "kernel trick" in which a nonlinear transformation Φ of the data obtains a higher-dimensional feature space (e.g., $\hat{\mathbf{X}} = \Phi(\mathbf{X})$). In effect, this approach extends the data into a higher dimension with which the categories are more linearly separable. The linear solution of CCA within this higher-dimensional space is equivalent to a non-linear solution in the original data space (Lai and Fyfe, 2000). We can avoid the need to explicitly define Φ, however, since positive definite kernel functions $\kappa(\mathbf{x}, \mathbf{y})$ satisfying Mercer's condition can implicitly map their input to higher-dimensional spaces. We specify a set of such kernels in section 6.2.2.

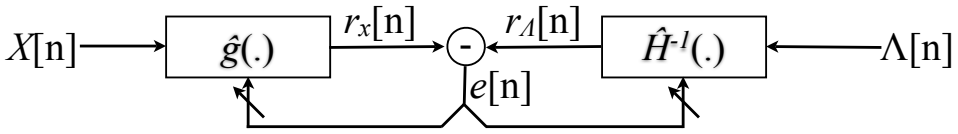
Reformulating eq. 6.3 within a framework of least-squares regression allows us to minimize $\frac{1}{2}\|\mathbf{X}\omega_x - \mathbf{Y}\omega_y\|^2$ such that $\frac{1}{2}(\|\mathbf{X}\omega_x\| + \|\mathbf{Y}\omega_y\|) = 1$. This allows us to solve the following generalized eigenvalue problem on the transformed data $\hat{\mathbf{X}} \in \mathbb{R}^{N \times m'_x}$ and $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times m'_y}$ by the method of Lagrange multipliers:

$$\frac{1}{2}\begin{bmatrix} \hat{\mathbf{X}}^T\hat{\mathbf{X}} & \hat{\mathbf{X}}^T\hat{\mathbf{Y}} \\ \hat{\mathbf{Y}}^T\hat{\mathbf{X}} & \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} \end{bmatrix} \hat{\omega} = \beta \begin{bmatrix} \hat{\mathbf{X}}^T\hat{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} \end{bmatrix} \hat{\omega}, \tag{6.4}$$

where $\hat{\omega} = [\hat{\omega}_x \hat{\omega}_y]^T$ is the concatenation of the transformed direction vectors and β is the Lagrange multiplier. We can now avoid explicit data transformation by applying a kernel function. Since the kernel matrix describing our transformed data, $\mathbf{K}_x = \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \in \mathbb{R}^{N \times N}$, has elements $\mathbf{K}_x[i, j] = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ defined by vectors in our original data space ($\mathbf{K}_y$ is defined similarly for $\hat{\mathbf{Y}}$), we left-multiply eq. 6.4 by $\begin{bmatrix} \hat{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Y}} \end{bmatrix}$, giving



(a) Nonlinear Hammerstein system (feedforward).



(b) System for identifying the parameters of the nonlinear Hammerstein system.

Figure 6.3: The feedforward Hammerstein system and its associated identification system.

$$\frac{1}{2}\begin{bmatrix} \mathbf{K}_x^2 & \mathbf{K}_x\mathbf{K}_y \\ \mathbf{K}_y\mathbf{K}_x & \mathbf{K}_y^2 \end{bmatrix}\alpha = \beta \begin{bmatrix} \mathbf{K}_x^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_y^2 \end{bmatrix}\alpha. \tag{6.5}$$

Here, $\alpha = [\alpha_x \alpha_y]^T \in \mathbb{R}^{2N}$ such that $\hat{\omega}_x = \hat{\mathbf{X}}^T\alpha_x$ and $\hat{\omega}_y = \hat{\mathbf{Y}}^T\alpha_y$ (Vaerenbergh, Via, and Santamaria, 2006b). This gives a generalized eigenvalue problem in the higher-dimensional space where we can minimize $(\mathbf{K}_x\alpha_x + \mathbf{K}_y\alpha_y)/2$ by adjusting α_x and α_y according to our original data space (Vaerenbergh, Via, and Santamaria, 2008).

KCCA and Hammerstein systems

A nonlinear Hammerstein system is a memoryless nonlinear function $g()$ followed by a linear dynamic system $H()$ in series, as shown in Figure 6.3(a). Our goal is to input acoustic observations, $\mathbf{X}$, of Mel-frequency cepstral coefficients (MFCC) to such a system and to infer the associated articulation vectors, Λ. In order to accomplish this accurately, we must learn the parameters of the two components of the Hammerstein system.

A mechanism for identifying these parameters has recently been proposed that takes ad-

vantage of the cascade structure by inverting the linear component, as in Figure 6.3(b), and minimizing the difference, $e[n]$, between $g(\mathbf{X}[n])$ and $H^{-1}(\Lambda[n])$ using KCCA (Aschbacher and Rupp, 2005). Since $H()$ is linear, we can reformulate eq. 6.5 to

$$\frac{1}{2}\begin{bmatrix} \mathbf{K}_x^2 & \mathbf{K}_x\hat{\Lambda} \\ \hat{\Lambda}^T\mathbf{K}_\mathbf{x} & \hat{\Lambda}^T\hat{\Lambda} \end{bmatrix}\begin{bmatrix} \alpha_x \\ \omega_\Lambda \end{bmatrix} = \beta\begin{bmatrix} \mathbf{K}_x(\mathbf{K}_x + c\mathbf{I}) & \mathbf{0} \\ \mathbf{0} & \hat{\Lambda}^T\hat{\Lambda} \end{bmatrix}\begin{bmatrix} \alpha_x \\ \omega_\Lambda \end{bmatrix}, \tag{6.6}$$

where we add a regularizing constant c to prevent overfitting (Aschbacher and Rupp, 2005). Here, ω_Λ provides the parameters of the linear part of the system, $H()^{-1}$, and α_x provides the parameters of the nonlinear part, $g()$. Given a combined average of the output of these two systems, $r = (r_x + r_\Lambda)/2 = (\mathbf{K}_x\alpha_x + \Lambda\omega_\Lambda)/2$, the eigenvalue problem decomposes to two coupled least squares problems:

$$\begin{aligned} \beta\alpha_x &= (\mathbf{K}_x + c\mathbf{I})^{-1}r \\ \beta\omega_\Lambda &= (\Lambda^T\Lambda)^{-1}\Lambda^T r \end{aligned} \tag{6.7}$$

This representation allows us to minimize a Euclidean error measurement $\|r_x - r_\Lambda\|$ by analytically solving for α_x and ω_Λ. In order to estimate articulation at run time, we compute $r_x = \mathbf{K}_x\alpha_x$, since we can construct the kernel matrix from observed acoustics, and then solve for $\Lambda \approx \mathbf{K}_x\alpha_x\omega_\Lambda^{-1}$, since $\Lambda\omega_\Lambda = r_\Lambda \approx r_x = \mathbf{K}_x\alpha_x$.

Adaptive algorithm

Unfortunately, for problems involving large amounts of data, as is typical in speech, the sizes of the kernel matrices described above become prohibitively large. An online algorithm that iteratively adjusts the estimates of α_x and ω_Λ based on subsequent segments of data is therefore desirable. We assume that we have a sliding context window covering L aligned frames from each data source, namely, $\mathbf{x}^{(n)} = [\mathbf{x}_n, \mathbf{x}_{n-1}, ..., \mathbf{x}_{n-L+1}]$ and $\Lambda^{(n)} = [\Lambda_n, \Lambda_{n-1}, ..., \Lambda_{n-L+1}]$. Assuming that we have matrix $\mathbf{K}_{reg}^{(n-1)}$ for the $(n-1)^{th}$ window of speech, and $\hat{\mathbf{K}}_{reg}^{(n-1)}$ is the matrix formed by its last $n-1$ rows and columns, then the regularized matrix for the current

window is

$$\mathbf{K}_{reg}^{(n)} = \begin{bmatrix} \hat{\mathbf{K}}_{reg}^{(n-1)} & \mathbf{k}_{n-1}(\mathbf{x}^{(n)}) \\ \mathbf{k}_{n-1}(\mathbf{x}^{(n)})^T & k_{nn}+c \end{bmatrix}, \tag{6.8}$$

where $\mathbf{k}_{n-1}(\mathbf{x}^{(n)}) = [\kappa(\mathbf{x}^{(n-L+1)}, \mathbf{x}^{(n)}, ..., \kappa(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}]^T$ and $k_{nn} = \kappa(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})$. The inverse of $\mathbf{K}_{reg}^{(n)}$ can also be computed quickly, given the inverse of $\mathbf{K}_{reg}^{(n-1)}$ (Vaerenbergh, Via, and Santamaria, 2006a). We then iteratively update our parameter estimates for ω_Λ and α_x as new data arrives using eq. 6.7. This entire process is summarized in algorithm 3 and is based on work on Wiener systems by Vaerenbergh et al. (Vaerenbergh, Via, and Santamaria, 2006b).

Algorithm 3: The adaptive KCCA algorithm.

begin

Initialize $\mathbf{K}_{reg}^{(0)} = (1+c)\mathbf{I}$

Initialize α_x and ω_Λ with random data

for $n = 1..N$ **do**

Calculate $\mathbf{K}_{reg}^{(n)}$ from $x^{(n)}$ as in eq. 6.8

$\mathbf{r}_x^{(n)} = \kappa(\mathbf{x}^{(n)}, \mathbf{x}^{(n-1)})\alpha_x^{(n-1)}$

$\mathbf{r}_\Lambda^{(n)} = \Lambda^{(n)}\omega_\Lambda^{(n-1)}$

$\mathbf{r}^{(n)} = \frac{\mathbf{r}_x^{(n-1)} + \mathbf{r}_\Lambda^{(n-1)}}{2}$

Calculate $\left(\mathbf{K}_{reg}^{(n)}\right)^{-1}$

Update solutions for α_x and ω_Λ as in eq. 6.7

Normalize solutions with $\beta = \|\omega_\Lambda\|$

end

6.2.2 Experiments

Our experiments evaluate the stability of the error-correction method and the estimation of tract variables from acoustics. We apply four kernel functions, namely the homogenous polynomial ($K_{h\_poly}^{(i)}$), the non-homogenous polynomial ($K_{nh\_poly}^{(i)}$), the radial-basis function ($K_{rbf}^{(\sigma)}$), and the

sigmoid ($K^{(\kappa,c)}_{sigmoid}$) kernels:

$$K^{(i)}_{h\_poly}(x_1,x_2) = (x_1x_2)^i$$
$$K^{(i)}_{nh\_poly}(x_1,x_2) = (x_1x_2+1)^i$$
$$K^{(\sigma)}_{rbf}(x_1,x_2) = \exp\left(-\frac{\|x_1-x_2\|^2}{2\sigma^2}\right)$$
$$K^{(\kappa,c)}_{sigmoid}(x_1,x_2) = \tanh(x_1x_2+c).$$

Training data consists of midsagittal tract variables (10-dimensional vectors) and aligned acoustics (42-dimensional MFCCs) selected from approximately 460 sentences uttered by a male speaker from Edinburgh's MOCHA database (Wrench, 1999). The positions and velocities of the jaw, lips, and tongue, are recorded with electromagnetic articulography as described in section 2.4. These data are then converted to the tract variable space as described in section 6.1. Results reported below are averages of 10-fold cross validation. Until otherwise indicated, the window length $L = 150$.

Stability and convergence during training

The goal of auto-correction is for the Euclidean error ($\mathbf{K}_x\alpha_x - \Lambda\omega_\Lambda$) (i.e., $e[n]$ in Figure 6.3(b)) to approach zero during training. Figure 6.4 shows the best, average, and worst mean squared errors in decibels during training given the homogenous polynomial kernel and 10 random initial parameterizations. This example is indicative of all other kernels whereby a period of fluctuation tends to follow a rapid decrease in error. Table 6.1 shows the total decrease in mean squared error (dB) between the first 20 and last 20 windows of the adaptive KCCA training process. As one increases the order of both the homogenous and non-homogenous kernels, the MSE reduction also increases. In both the tan-sigmoid and radial-basis function kernels, however, our choice of parameters seems to have little discernible effect.

Vaerenbergh et al. apply a nearly identical approach to learning Wiener systems on the comparatively simple problem of estimating a hyperbolic tangent function given univariate input (Vaerenbergh, Via, and Santamaria, 2006b; Vaerenbergh, Via, and Santamaria, 2008),

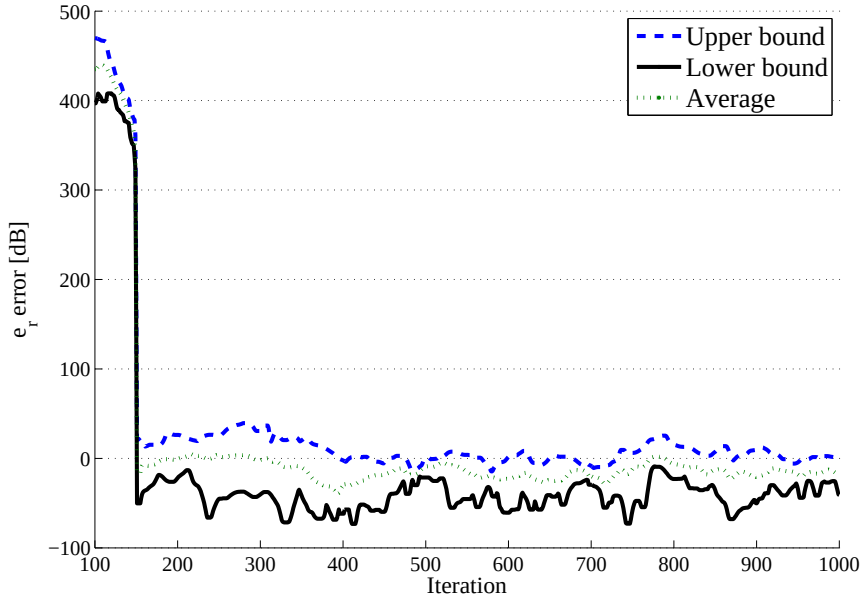


Figure 6.4: Normalization error, $e[n]$, for the first-order homogenous polynomial kernel at window size $L = 150$.

Homogenous polynomial		Nonhomogenous polynomial	
i	MSE reduction	i	MSE reduction
1	421.6	1	441.9
2	403.6	2	413.1
3	394.5	3	382.9
Sigmoid		**Radial-basis function**	
(κ, c)	MSE reduction	σ	MSE reduction
$(0.2, 0.1)$	313.2	0.1	406.5
$(0.2, 0.5)$	321.5	0.5	410.4
$(0.5, 0.1)$	309.7	1.0	406.7
$(0.5, 0.5)$	314.3		

Table 6.1: Total reduction in MSE (dB) between Hammerstein components during training across kernels and parameterizations.

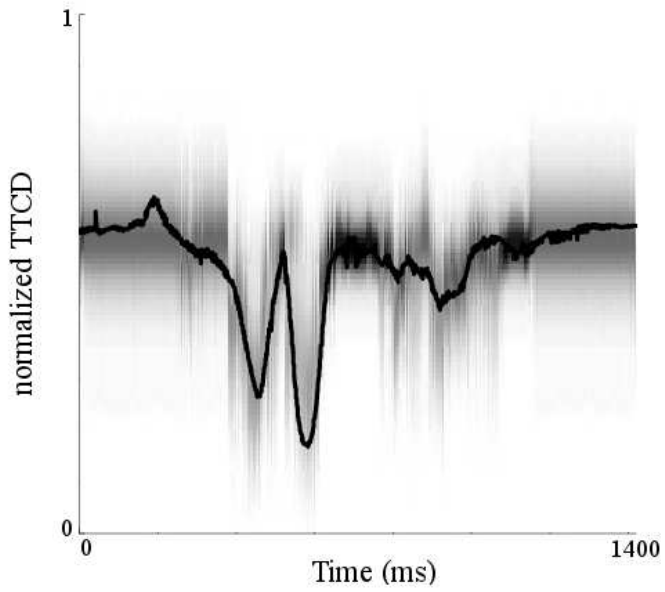


Figure 6.5: Example intensity map of Gaussian mixtures produced by a mixture density network trained to estimate the tongue tip constriction degree. Darker sections represent higher probability. The true trajectory is superimposed as a black curve.

reaching MSE between -30dB and -40dB within 1000 to 1500 iterations. Surprisingly, most of the error in our experiments is dispelled much earlier, within 200 iterations, with MSE fluctuating between -76.9dB and 39.5dB thereafter across all kernels and parameterizations.

KCCA versus mixture density networks

In order to judge the accuracy of the articulatory estimates produced by adaptive KCCA against the state-of-the-art, we consider mixture density neural networks (MDNs) that output parameters of Gaussian mixture probability distributions, as described by Richmond (2003). We train MDNs to estimate the likelihood of tract variable positions given MFCC input and 2 frames of surrounding acoustic context. Figure 6.5 shows an example of the estimated likelihood of tract variable positions over time produced by a trained MDN as an intensity map superimposed with the true trajectory. MDNs are trained on the same data as KCCA. Articulatory estimates for KCCA are smoothed with third-order median filters.

TV	MDN $\mu(\sigma^2)$	KCCA $\mu(\sigma^2)$	TV	MDN $\mu(\sigma^2)$	KCCA $\mu(\sigma^2)$
VEL	−0.28 (0.08)	−0.23 (0.07)	TTCD	−1.60 (0.17)	−1.60 (0.17)
LTH	−0.18 (0.12)	−0.18 (0.14)	TTCL	−1.62 (0.17)	−1.57 (0.16)
LA	−0.32 (0.11)	−0.28 (0.10)	TBCD	−0.79 (0.14)	−0.80 (0.15)
LP	−0.44 (0.12)	−0.41 (0.13)	TDCL	−0.20 (0.11)	−0.18 (0.09)
GLO	−1.30 (0.16)	−1.14 (0.15)			

Table 6.2: Average log likelihoods of true tract variable positions in test data, under distributions produced by mixture density networks (MDNs) and the KCCA method, with variances.

We assess the accuracy of the MDN and KCCA methods by comparing their estimates of the log likelihood of the true articulatory trajectories. A more accurate method will assign a higher probability to the actual trajectory. The likelihood of a frame of articulation is easily computed by MDNs whose output is a probability distribution over tract variable positions. We approximate the likelihood of a frame of articulation in the KCCA approach with the radial-basis kernel by fitting a Gaussian to the estimates of 10 trials having different initial parameterizations. Test data in each trial consists of approximately 60 utterances from our male speaker.

The mean and variance of the log likelihoods of true articulatory positions across all test frames is summarized in Table 6.2 for both methods. According to the t test with $9.6E^4 < n_1 = n_2 < 9.9E^4$ frames and one degree of freedom, KCCA is significantly more accurate than the MDN method at the 95% confidence level for **VEL**, **LA**, **LP**, **TTCL**, and **TDCL** and at the 99% confidence level for **GLO**, and statistically indistinguishable at these levels for the remaining tract variables.

6.2.3 Summary of KCCA approach

Some high-level questions remain. For example, if the eventual aim is to use estimated articulatory trajectories to constrain hypotheses in speech recognition, then it is possible that a quantized representation may be more amenable to training in such systems. A similar (though non-adaptive) kernel-based system has recently been proposed that inverts acoustic to articulatory data according to discrete categories (Zheng et al., 2006). Likewise, a k-means clustering of the tract variable motion estimated by our adaptive KCCA process might be applicable as conditioning variables in dynamic Bayes networks for speech classification (Rudzicz, 2009a).

Our analysis has demonstrated that adaptive KCCA can effectively learn non-linear relationships between co-occurring variables in speech, and perform more accurate acoustic-to-articulatory inversion than the state-of-the-art. This approach combines a semi-analytical (non-statistical) kernel-based approach with an iterative, adaptive learning process and could be used for online trajectory estimation.

6.3 A noisy-channel model of dysarthria

Dysarthria is sometimes characterized as a distortion of parallel biological pathways that corrupt motor signals before execution (Kent and Rosen, 2004; Freund et al., 2005). In this section we cast the speech-motor interface within the mathematical framework of the noisy-channel model. Within this information-theoretic approach, we aim to infer the nature of the motor signal distortions given appropriate measurements of the vocal tract. That is, we ask the following question: Is dysarthric speech a distortion of typical speech, or are they both distortions of some common underlying representation?

First, in section 6.3.1, we ask whether the incorporation of articulatory data is theoretically useful in reducing uncertainty in dysarthric speech. Second, in section 6.3.2, we ask which of the two noisy channel models in figure 6.6 best describe the observed behaviour of dysarthric speech.

Data for this study are collected as described as in chapter 4. Here, we use data from three dysarthric speakers with cerebral palsy (males M01 and M04, and female F03), as well as their age- and gender-matched counterparts from the general population (males MC01 and MC03, and female FC02). For this study we restrict our analysis to 100 phrases uttered in common by all six speakers.

6.3.1 Entropy

We wish to measure the degree of statistical disorder in both acoustic and articulatory data for dysarthric and non-dysarthric speakers, as well as the *a posteriori* disorder of one type of data given the other. This quantification will inform us as to the relative merits of incorporating knowledge of articulatory behaviour into ASR systems for dysarthric speakers. Entropy, $H(X)$, is a measure of the degree of uncertainty in a random variable X. When X is discrete, this value is computed with the familiar

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

where b is the logarithm base, x_i is a value of X, of which there are n possible, and $p(x_i)$ is its probability. When our observations are continuous, as they are in our acoustic and articulatory database, we must use *differential entropy* defined by

$$H(X) = -\int_X f(X) \log f(X) dX,$$

where $f(X)$ is the probability density function of X. For a number of distributions $f(X)$, the differential entropy has known forms (Lazo and Rathie, 1978). For example, if $f(X)$ is a multivariate normal,

$$\begin{aligned} f_X(x_1,...,x_N) &= \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{N/2}\,|\Sigma|^{1/2}} \\ H(X) &= \tfrac{1}{2} \ln\left((2\pi e)^N |\Sigma|\right), \end{aligned} \tag{6.9}$$

where μ and Σ are the mean and covariances of the data. However, since we observe that both acoustic and articulatory data follow non-Gaussian distributions, we choose to represent these

spaces by mixtures of Gaussians. Huber et al. (2008) have developed an accurate algorithm for estimating differential entropy of Gaussian mixtures based on iteratively merging Gaussians and the approximate upper bound of the entropy,

$$\tilde{H}(X) = \sum_{i=1}^{L} \omega_i \left(-\log \omega_i + \tfrac{1}{2} \log((2\pi e)^N |\Sigma_i| \right),$$

where ω_i is the weight of the $i^{th}(1 \leq i \leq L)$ Gaussian and Σ_i is that Gaussian's covariance matrix. This method is used to approximate entropies in the following study, with $L = 32$. Note that while differential entropies *can* be negative and not invariant under change of variables, other properties of entropy are retained (Huber et al., 2008), such as the chain rule for conditional entropy

$$H(Y\,|\,X) = H(Y,X) - H(X),$$

which describes the uncertainty in Y given knowledge of X, and the chain rule for mutual information

$$I(Y;X) = H(X) + H(Y) - H(X,Y),$$

which describes the mutual dependence between X and Y. Here, we quantize entropy with the *nat*, which is the natural logarithmic unit, e (≈ 1.44 bits).

Experiments

We measure the differential entropy of acoustics ($H(Ac)$), of articulation ($H(Ar)$), and of acoustics given knowledge of the vocal tract ($H(Ac\,|\,Ar)$) in order to obtain theoretical estimates as to the utility of articulatory data. Table 6.3 shows these quantities across the six speakers in this study. As expected, the acoustics of dysarthric speakers are much more disordered than for non-dysarthric speakers. One unexpected finding is that there is very little difference between speakers in terms of their entropy of articulation. Although dysarthric speakers clearly lack articulatory dexterity, this implies that they nonetheless articulate with a level of consistency similar to their non-dysarthric counterparts[3]. However, the equivocation $H(Ac\,|\,Ar)$ is an order

[3]This is borne out in the literature (Kent and Rosen, 2004).

	Speaker	$H(Ac)$	$H(Ar)$	$H(Ac\|Ar)$
Dys.	M01	66.37	17.16	50.30
	M04	33.36	11.31	26.25
	F03	42.28	19.33	39.47
	Average	47.34	15.93	38.68
Ctrl.	MC01	24.40	21.49	1.14
	MC03	18.63	18.34	3.93
	FC02	16.12	15.97	3.11
	Average	19.72	18.60	2.73

Table 6.3: Differential entropy, in nats, across dysarthric and control speakers for acoustic *ac* and articulatory *ar* data.

of magnitude lower for non-dysarthric speakers. This implies that there is very little ambiguity left in the acoustics of non-dysarthric speakers if we have simultaneous knowledge of the vocal tract, but that quite a bit of ambiguity remains for our dysarthric speakers, despite significant reductions. Further investigation should confirm the causes of this remnant ambiguity, but potential sources include unmeasured interaction between the glottis and the other articulators (e.g., aberrant voicing) or unmeasured lateral asymmetry in the tongue.

Table 6.4 shows the average mutual information between acoustics and articulation for each type of speaker, given knowledge of the phonological manner of articulation. In table 4.1 we noted a prevalence of pronunciation errors among dysarthric speakers for plosives, but table 6.4 shows no particularly low congruity between acoustics and articulation for this manner of phoneme. Those pronunciation errors tended to be voicing errors, which would involve the glottis, which is not measured in this study.

Table 6.4 appears to imply that there is little mutual information between acoustics and articulation in vowels across all speakers. However, this is almost certainly the result of our

Manner	$I(Ac;Ar)$	
	Dys.	Ctrl.
plosives	10.92	16.47
affricates	8.71	9.23
fricatives	9.30	10.94
nasals	13.29	15.10
glides	11.92	12.68
vowels	6.76	7.15

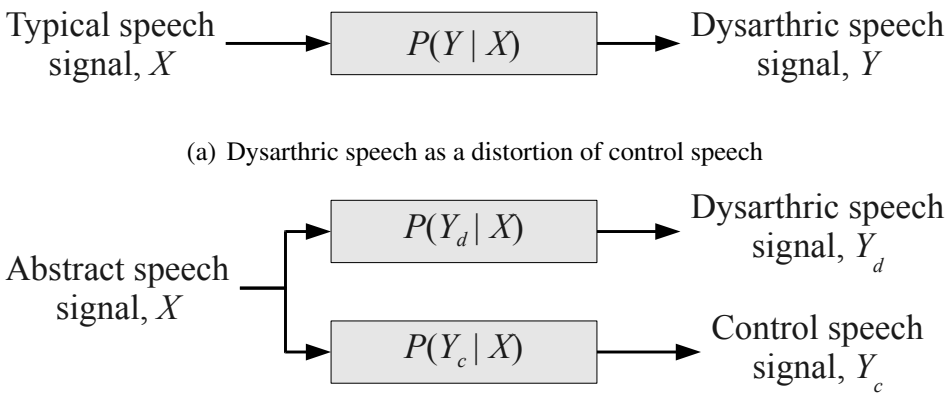
Table 6.4: Mutual information $I(Ac;Ar)$ of acoustics and articulation for dysarthric and control subjects, across phonological manners of articulation.

exclusion of tongue blade and tongue dorsum measurements[4] in order to standardize across speakers who could not manage these sensors. Indeed, the configuration of the entire tongue is known to be useful in discriminating among the vowels (O'Shaughnessy, 2000). An *ad hoc* analysis including all three tongue sensors for speakers F03, MC01, MC03, and FC02 revealed mutual information between acoustics and articulation of 16.81 nats for F03 and 18.73 nats for the control speakers, for vowels. This is compared with mutual information of 11.82 nats for F03 and 13.88 nats for the control speakers across all other manners. The trend is that acoustics are better predicted given more tongue measurements, as expected.

6.3.2 The noisy channel

The noisy-channel theorem states that information passed through a channel with capacity C at a rate $R \leq C$ can be reliably recovered with an arbitrarily low probability of error given an appropriate coding. Here, a message from a finite alphabet is encoded, producing signal $x \in X$. That signal is then distorted by a medium which transmits signal $y \in Y$ according to some distribution $P(Y \mid X)$. Given that there is some probability that the received signal, y, is

[4] We retained the tongue tip, jaw, and four lip measurements.



(a) Dysarthric speech as a distortion of control speech

(b) Dysarthric and control speech as distortions of a common abstraction

Figure 6.6: Sections of alternative noisy channel models for the neuro-motor interface in speakers with dysarthria.

corrupted, the message produced by the decoder may differ from the original (Shannon, 1949).

To what extent can we describe the effects of dysarthria within an information-theoretic noisy channel model? We pursue two competing hypotheses within this general framework. The first hypothesis models the assumption that dysarthric speech is a distorted version of typical speech. Here, signal X and Y represent the vocal characteristics of the general and dysarthric populations, respectively, and $P(Y|X)$ models the distortion between them. The second hypothesis models the assumption that *both* dysarthric and typical speech are distorted versions of some common abstraction. Here, Y_d and Y_c represent the vocal characteristics of dysarthric and control speakers, respectively, and X represents a common, underlying mechanism and that $P(Y_d|X)$ and $P(Y_c|X)$ model distortions from that mechanism. These two hypotheses are visualized in figure 6.6. In each of these cases, signals can be acoustic, articulatory, or some combination thereof.

Common underlying abstractions

In order to test our hypothesis that both dysarthric and control speakers share a common high-level abstraction of the vocal tract that is in both cases distorted during articulation, we incor-

porate the theory of *task dynamics* (Saltzman and Munhall, 1989) as descried above in section 6.1.

The open-source TADA system (Nam and Goldstein, 2006) estimates the positions of various articulators during speech according to parameters that have been carefully tuned by the authors of TADA according to a generic, speaker-independent representation of the vocal tract (Saltzman and Munhall, 1989). Given a word sequence and a syllable-to-gesture dictionary, TADA produces the continuous tract variable paths that are necessary to produce that sequence. This takes into account various physiological aspects of human speech production, such as interarticulator co-ordination and timing (Nam and Saltzman, 2003).

In this study, we use TADA to produce estimates of a global, high-level representation of speech common to both dysarthric and non-dysarthric speakers alike. Given a word sequence uttered by both types of speaker, we produce five continuous curves prescribed by that word sequence in order to match our available EMA data. Those curves are lip aperture and protrusion (LA and LP), tongue tip constriction location and degree (TTCL and TTCD, representing front-back and top-down positions of the tongue tip, respectively), and lower incisor height (LIH). These curves are then compared against actually observed EMA data, as described below.

Experiments

Our task is to determine whether dysarthric speech is best represented as a distorted version of typical speech, or if both dysarthric and typical speech ought to be viewed as distortions of a common abstract representation. To explore this question, we design a transformation system that produces the most likely observation in one data space given its counterpart in another and the statistical relationship between the two spaces. This transformation in effect implements the noisy channel itself.

To accomplish this, we learn probability distributions over our EMA data. First, we collect all dysarthric data together and all non-dysarthric data together. We then consider the acoustic (Ac) and articulatory (Ar) subsets of these data. In each case, we train Gaussian mixtures,

each with 60 components, over 90% of the data in both dysarthric and non-dysarthric speech. Here, each of the 60 phonemes in the data is represented by one Gaussian component, with the weight of that component determined by the relative proportion of 10 ms frames for that phoneme. Similarly, all training word sequences are passed to TADA, and we train a mixture of Gaussians on its articulatory output.

Across all Gaussian mixtures, we end up with 5 Gaussians tuned to various aspects of each phoneme p: its dysarthric acoustics and articulation ($\mathbf{N}_p^{Ac}(Y_d)$ and $\mathbf{N}_p^{Ar}(Y_d)$), its control acoustics and articulation ($\mathbf{N}_p^{Ac}(Y_d)$ and $\mathbf{N}_p^{Ar}(Y_d)$), and its prescribed articulation from TADA ($\mathbf{N}_p^{Ar}(X)$). Each Gaussian $\mathbf{N}_p^{A}(B)$ is represented by its mean $\mu_p^{(A,B)}$ and its covariance, $\Sigma_p^{(A,B)}$. Furthermore, we compute the cross-covariance matrix between Gaussians for a given phoneme (e.g., $\Sigma_p^{(Ac,Y_c)\rightarrow(Ac,Y_d)}$ is the cross-covariance matrix of the acoustics of the control (Y_c) and dysarthric (Y_d) speech for phoneme p). Given these parameters, we estimate the most likely frame in one domain given its counterpart in another. For example, if we are given a frame of acoustics from a control speaker, we can synthesize the most likely frame of acoustics for a dysarthric speaker, given an application of the noisy channel proposed by Hosom et al. (2003) used to transform dysarthric speech to make it more intelligible. Namely, given a frame of acoustics y_c from a control speaker, we can estimate the acoustics of a dysarthric speaker y_d with:

$$\begin{aligned} f_{Ac}(y_c) =& E(y_d \,|\, y_c) \\ =& \sum_{i=1}^{P} h_i(y_c) \left[\mu_i^{(Ac,Y_d)} + \right. \\ & \Sigma_i^{(Ac,Y_c)\rightarrow(Ac,Y_d)} \cdot \left(\Sigma_i^{(Ac,Y_c)}\right)^{-1} \cdot \\ & \left. \left(y_c - \mu_i^{(Ac,Y_c)}\right)\right], \end{aligned} \tag{6.10}$$

where

$$h_i(y_c) = \frac{\alpha_i N\left(y_c; \mu_i^{(Ac,Y_c)}, \Sigma_i^{(Ac,Y_c)}\right)}{\sum_{j=1}^{P} \alpha_j N\left(y_c; \mu_j^{(Ac,Y_c)}, \Sigma_j^{(Ac,Y_c)}\right)},$$

where α_p is the proportion of the frames of phoneme p in the data. Transforming between different types and sources of data is accomplished merely by substituting in the appropriate Gaussians above.

We now measure how closely the transformed data spaces match their true target spaces. In each case, we transform test utterances (recorded, or synthesized with TADA) according to functions learned in training (i.e., we use the remaining 10% of the data for each speaker type). These transformed spaces are then compared against their target space in our data. Table 6.5 shows the Gaussian mixture phoneme-level Kullback-Leibler divergences given various types of source and target data, weighted by the relative proportions of the phonemes. Each pair of N-dimensional Gaussians ($\mathbf{N}_i$ with mean μ_i and covariance Σ_i) for a given phone and data type is compared with

$$D_{KL}(\mathbf{N}_0 \,||\, \mathbf{N}_1) = \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \text{trace}(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - N\right).$$

Our baseline shows that control and dysarthric speakers differ far more in their acoustics than in their articulation. When our control data (both acoustic and articulatory) are transformed to match the dysarthric data, the result is predictably more similar to the latter than if the conversion had not taken place. This corresponds to the noisy channel model of figure 6.6(a), whereby dysarthric speech is modelled as a distortion of non-dysarthric speech. However, when we model dysarthric and control speech as distortions of a common, abstract representation (i.e., task dynamics) as in figure 6.6(b), the resulting synthesized articulatory spaces are more similar to their respective observed data than the articulation predicted by the first noisy channel model. Dysarthric articulation predicted by transformations from task-dynamics space differ significantly from those predicted by transformations from control EMA data at the 95% confidence interval.

Type 1	Type 2	KL divergence (10^{-2} nats) Acous.	Artic.
Ctrl.	Dys.	25.36	3.23
Ctrl. $\rightarrow$ Dys.	Dys.	17.78	2.11
TADA $\rightarrow$ Ctrl.	Ctrl.	N/A	1.69
TADA $\rightarrow$ Dys.	Dys.	N/A	1.84

Table 6.5: Average weighted phoneme-level Kullback-Leibler divergences of acoustic and articulatory spaces given transformed and untransformed control and dysarthric models, weighted by the relative proportions of the phoneme.

6.3.3 Summary of entropy in dysarthric speech

These experiments demonstrate a few acoustic and articulatory features in speakers with cerebral palsy. First, these speakers are likely to mistakenly voice unvoiced plosives, and to delete fricatives regardless of their word position. We suggest that it might be prudent to modify the vocabularies of ASR systems to account for these expected mispronunciations. Second, dysarthric speakers produce sonorants significantly slower than their non-dysarthric counterparts. This may present an increase in insertion errors in ASR systems (Rosen and Yampolsky, 2000).

Although not quantified here, we detect that a lack of articulatory control can often lead to observable acoustic consequences. For example, our dysarthric data contain considerable involuntary types of velopharyngeal or glottal noise (often associated with respiration), audible swallowing, and involuntary repetition. We intend to work towards methods of explicitly identifying regions of non-speech noise in our ASR systems for dysarthric speakers.

We have considered the amount of statistical disorder (i.e., entropy) in both acoustic and articulatory data in dysarthric and non-dysarthric speakers. The use of articulatory knowledge

reduces the degree of this disorder significantly for dysarthric speakers (18.3%, relatively), though far less than for non-dysarthric speakers (86.2%, relatively). In real-world applications we are not likely to have access to measurements of the vocal tract; however, many approaches exist that estimate the configuration of the vocal tract given only acoustic data (Richmond, King, and Taylor, 2003; Toda, Black, and Tokuda, 2008), often to an average error of less than 1 mm. The generalizability of such work to new speakers (particularly those with dysarthria) without training is an open research question.

We have argued for noisy channel models of the neuro-motor interface assuming that the pathway of motor command to motor activity is a linear sequence of dynamics. The biological reality is much more complicated. In particular, the pathway of verbal motor commands includes several sources of sensory feedback (Seikel, King, and Drumright, 2005) that modulate control parameters during speech (Gracco, 1995). These senses include exteroceptive stimuli (auditory and tactile), and interoceptive stimuli (particularly proprioception and its kinesthetic sense) (Seikel, King, and Drumright, 2005), the disruption of which can lead to a number of production changes. For instance, Abbs, Folkins, and Sivarajan (1976) showed that when conduction in the mandibular branches of the trigeminal nerve is blocked, the resulting speech has considerably more pronunciation errors, although is generally intelligible. Barlow (1989) argues that the redundancy of sensory messages provides the necessary input to the motor *planning* stage, which relates abstract goals to motor activity in the cerebellum.

As we continue to develop our articulatory ASR models for dysarthric speakers, one potential avenue for future research involves the incorporation of feedback from the current state of the vocal tract to the motor planning phase. This would be similar, in premise, to the DIVA model (Guenther and Perkell, 2004). There is also ample evidence for a 'phonological store' or 'phonological loop' in the cerebellum in which articulatory rehearsals and their expected acoustic consequences are stored for between 1.5 and 2.0 seconds during speech comprehension and production (Baddeley, Gathercole, and Papagno, 1998; Beaman, 2007). Under this model, speakers with dysarthria show a normal capacity for articulatory rehearsal, which sug-

gests that distortions occur after the planning stage but before motor execution (Baddeley and Wilson, 1985).

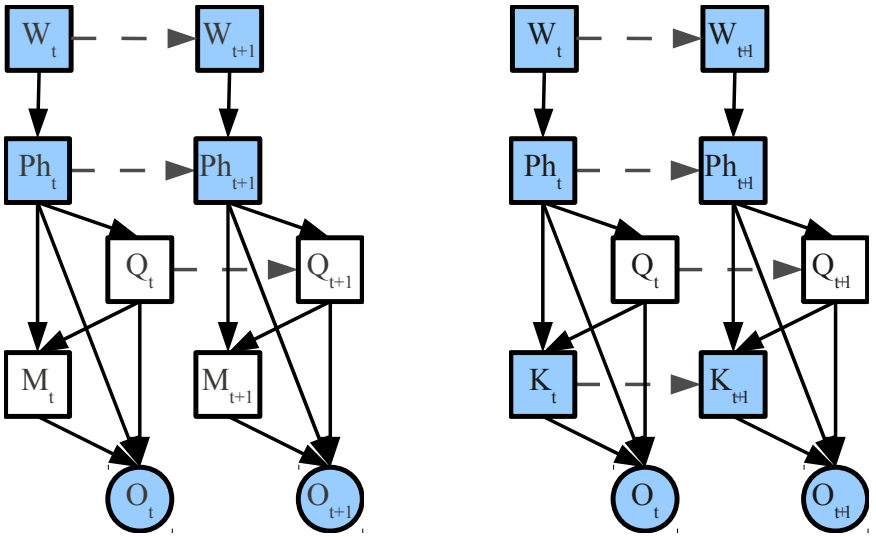
6.4 Correcting errors in ASR with articulatory dynamics

This section describes an integration of task-dynamics theory into an ASR system for word recognition on non-dysarthric data. In section 6.4.1, we augment traditional models of ASR with probabilistic relationships between acoustics and articulation learned from appropriate data. This leads to the incorporation of a high-level, goal-oriented, and control-based theory of speech production within a novel ASR system in section 6.4.3. Experiments with these models are described in section 6.4.4 and summarized in section 6.4.5

6.4.1 Baseline systems

We examine two baseline systems. The first is the standard acoustic hidden Markov model (HMM) augmented with a bigram language model, as shown in figure 6.7(a). Here, $W_t \rightarrow W_{t+1}$ represents word transition probabilities, learned by maximum likelihood estimation, and $Ph_t \rightarrow Ph_{t+1}$ represents phoneme transition probabilities whose order is explicitly specified by the relationship $W_t \rightarrow Ph_t$. Likewise, each phoneme Ph conditions the sub-phoneme state, Q_t, whose transition probabilities $Q_t \rightarrow Q_{t+1}$ describe the dynamics within phonemes. The variable M_t refers to hidden Gaussian indices so that the likelihoods of acoustic observations, O_t, are represented by a mixture of 4, 8, 16, or 32 Gaussians for each state and each phoneme. See Murphy (2002) for a further description of this representation.

The second baseline model is the articulatory dynamic Bayes network (DBN-A). This augments the standard acoustic HMM by replacing hidden indices, M_t, with discrete observations of the vocal tract, K_t, as shown in figure 6.7(b). The pattern of acoustics within each phoneme is dependent on a relatively restricted set of possible articulatory configurations (Roweis, 1999). To find these discrete positions, we obtain k vectors that best describe the articulatory data



(a) HMM (b) DBN-A

Figure 6.7: Baseline systems in evaluating the correction of errors with articulatory dynamics: (a) acoustic hidden Markov model and (b) articulatory dynamic Bayes network. Node W_t represents the current word, Ph_t is the current phoneme, Q_t is that phoneme's dynamic state, O_t is the acoustic observation, M_t is the Gaussian mixture component, and K_t is the discretized articulatory configuration. Filled nodes represent observed variables during training, although only O_t is observed during recognition. Square nodes are discrete variables while circular nodes are continuous variables.

according to k-means clustering with the sum-of-squares error function. During training, the DBN variable K_t is set explicitly to the *index* of the mean vector nearest to the current frame of EMA data at time t. In this way, the relationship $K_t \rightarrow O_t$ allows us to learn how discretized articulatory configurations affect acoustics. The training of DBNs involves a specialized version of expectation-maximization, as described in the literature (Murphy, 2002; Ghahramani, 1998). During inference, variables W_t, Ph_t, and K_t become hidden and we marginalize over their possible values when computing their likelihoods. Bigrams are computed by maximum likelihood on lexical annotations in the training data.

6.4.2 Switching Kalman filter

Our first experimental system attempts speech recognition given only articulatory data. The true state of the tract variables at time $t-1$ constitutes a 9-dimensional vector, $\mathbf{x}_{t-1}$, of continuous values. Under the task dynamics model of section 6.1, the motions of these tract variables obey critically damped second-order oscillatory relationships. We start with the simplifying assumption of linear dynamics here with allowances for random Gaussian *process noise*, $\mathbf{v}_t$, with variance σ_{v_t} since articulatory behaviour is non-deterministic. Moreover, we know that EMA recordings are subject to some error (usually less than 1 mm (Yunusova, Green, and Mefferd, 2009)), so the actual observation at time t, $\mathbf{y}_t$, will not in general be the true position of the articulators. Assuming that the relationship between $\mathbf{y}_t$ and $\mathbf{x}_t$ is also linear, and that the *measurement noise*, $\mathbf{w}_t$, is also Gaussian with variance σ_{w_t}, then the dynamical articulatory system can be described by

$$\begin{aligned} \mathbf{x}_t &= D_t \mathbf{x}_{t-1} + \mathbf{v}_t \\ \mathbf{y}_t &= C_t \mathbf{x}_t + \mathbf{w}_t. \end{aligned} \tag{6.11}$$

Equations 6.11 form the basis of the Kalman filter which allows us to use EMA measurements directly, rather than quantized abstractions thereof as in the DBN-A model. Obviously, since articulatory dynamics vary significantly for different goals, we replicate eq. (6.11) for each phoneme and connect these continuous Kalman filters together with discrete conditioning variables for phoneme and word, resulting in the switching Kalman filter (SKF) model. Here, parameters D_t and $\mathbf{v}_t$ are implicit in the relationship $\mathbf{x}_t \rightarrow \mathbf{x}_{t+1}$, and parameters C_t and $\mathbf{w}_t$ are implicit in $\mathbf{x}_t \rightarrow \mathbf{y}_t$. Each of these parameters depends on the hidden discrete state and switches with the state. In this model, observation $\mathbf{y}_t$ is the instantaneous measurements derived from EMA, and $\mathbf{x}_t$ is their true hidden states.

In order to train the SKF model, we perform a specialized expectation-maximization over its parameters, assuming that the conditioning state is S_t at time t and that it has Markovian dynamics with state transition matrix $Z(S_{t-1}, S_t)$, initial state distribution $\pi_{\mathbf{1}}$ (sequences are 1-

indexed here), mean vectors μ_t, and covariance Σ_t. The complete log likelihood of all training data (of length T) in the SKF model is

$$\begin{aligned} L = \log P(\mathbf{x}_{1:T}, S_{1:T}, \mathbf{y}_{1:T}) &= -\frac{1}{2}\sum_{t=1}^{T}\left([\mathbf{y}_t - C_t\mathbf{x}_\mathbf{t}]^\top \sigma_{w_t}^{-1}[\mathbf{y}_t - C_t\mathbf{x}_t]\right) - \frac{1}{2}\sum_{t=1}^{T}\log\|\sigma_{w_t}\| \\ &- \frac{1}{2}\sum_{t=2}^{T}\left([\mathbf{x}_t - D_t\mathbf{x}_{t-1}]^\top \sigma_{v_t}^{-1}[\mathbf{x}_t - D_t\mathbf{x}_{t-1}]\right) - \frac{1}{2}\sum_{t=2}^{T}\log|\sigma_{v_t}| \\ &- \frac{1}{2}[\mathbf{x}_1 - \mu_1]^\top \Sigma_1^{-1}[\mathbf{x}_1 - \mu_1] - \frac{1}{2}\log\|\Sigma_1\| - \frac{T(n+m)}{2}\log 2\pi \\ &+ \log\pi_\mathbf{1} + \sum_{t=2}^{T}\log Z(S_{t-1}, S_t). \end{aligned} \tag{6.12}$$

During expectation-maximization training, the quantity we maximize is

$$\begin{aligned} \hat{L} &= E_{P(\mathbf{x}_{1:T}, S_{1:T}, \mathbf{y}_{1:T})}[L] \\ &= E_{P(S_{1:T}, \mathbf{y}_{1:T})}\left[E_{P(\mathbf{x}_{1:T}\,|\,S_{1:T}, \mathbf{y}_{1:T})}[L]\right] \\ &\approx E_{P(S_{1:T}, \mathbf{y}_{1:T})}\left[E_{P(\mathbf{x}_{1:T}\,|\,\mathbf{y}_{1:T})}[L]\right] \\ &= P(\mathbf{y}_{1:T})\sum_{t=2}^{T}\sum_{S_t}\left(\sum_{\{S_\tau, \tau\neq t\}} P(S_{1:T}\,|\,\mathbf{y}_{1:T})\right)\hat{E}\left[\log P(\mathbf{x}_t\,|\,\mathbf{x}_{t-1}, S_t)\right] + \ldots, \end{aligned} \tag{6.13}$$

where $\hat{E}[\cdot] = E[\cdot\,|\,\mathbf{y}_{1:T}]$. The approximation is used because $E[\mathbf{x}_t\,|\,\mathbf{y}_{1:T}]$ does not result in an exponential expansion as does $E[\mathbf{x}_t\,|\,\mathbf{y}_{1:T}, S_{1:T}]$. If $\mathbf{x}_{t\,|\,T}$ is the expected value of $\mathbf{x}_t$ given $\mathbf{y}_{1:T}$, $\mathbf{V}_{t\,|\,T}$ is the covariance of $\mathbf{x}_t$ given $\mathbf{y}_{1:T}$, and $\mathbf{V}_{t,t-1\,|\,T}$ is the cross covariance of $\mathbf{x}_t$ with $\mathbf{x}_{t-1}$ given $\mathbf{y}_{1:T}$, then given a set of N independent and identically distributed sequences indexed by ℓ and the definitions

$$\begin{aligned} W_t^j &= P(S_t = j\,|\,\mathbf{y}_{1:T}) \\ \hat{\mathbf{x}}_t &= \hat{E}[\mathbf{x}_t] \\ P_t &= \hat{E}[\mathbf{x}_t\,|\,\mathbf{x}_t^\top] = V_{t\,|\,T} + \mathbf{x}_{t\,|\,T}\mathbf{x}_{t\,|\,T}^\top \\ P_{t,t-1} &= \hat{E}[\mathbf{x}_t\,|\,\mathbf{x}_{t-1}^\top] = V_{t,t-1\,|\,T} + \mathbf{x}_{t\,|\,T}\mathbf{x}_{t-1\,|\,T}^\top, \end{aligned} \tag{6.14}$$

where W_t^j is analogous to the weight component of a Gaussian mixture, then we can set the

derivative of eq. 6.13 with respect to our desired parameters to zero and solve, giving

$$\begin{aligned}
D_i &= \left(\sum_\ell \sum_{t=2}^{T} W_t^i P_{t,t-1}\right)\left(\sum_\ell \sum_{t=2}^{T} W_t^i P_{t-1}\right)^{-1} \\
\sigma_{v_t} &= \left(\frac{1}{\sum_\ell \sum_{t=2}^{T} W_t^i}\right)\left(\sum_\ell \sum_{t=2}^{T} W_t^i P_t - D_i \sum_\ell \sum_{t=2}^{T} W_t^i P_{t,t-1}^\top\right) \\
C_i &= \left(\sum_\ell \sum_{t=1}^{T} W_t^i \mathbf{y}_t \hat{\mathbf{x}}_t^\top\right)\left(\sum_\ell \sum_{t=1}^{T} W_t^i P_t\right)^{-1} \\
\sigma_{w_t} &= \left(\frac{1}{\sum_\ell \sum_{t=1}^{T} W_t^i}\right)\sum_\ell \sum_{t=1}^{T} W_t^i (\mathbf{y}_t \mathbf{y}_t^\top - C_i \hat{\mathbf{x}}_t \mathbf{y}_t^\top)
\end{aligned} \tag{6.15}$$

and

$$\begin{aligned}
\mu_i &= \frac{\sum_\ell W_1^i \hat{\mathbf{x}}_1}{\sum_\ell W_1^i} \\
\Sigma_i &= \frac{\sum_\ell W_1^i (\hat{\mathbf{x}}_1 - \mu_1)(\hat{\mathbf{x}}_1 - \mu_1)^\top}{\sum_\ell W_1^i} \\
Z(i,j) &= \frac{\sum_\ell \sum_{t=2}^{T} P(S_{t-1} = i, S_t = j \,|\, \mathbf{y}_{1:T})}{\sum_\ell \sum_{t=1}^{T-1} W_t^i} \\
\pi_i &= \frac{1}{N} \sum_\ell W_1^i.
\end{aligned} \tag{6.16}$$

Further details are described in the literature (Murphy, 1998; Deng, Bouchard, and Yeap, 2005).

6.4.3 Recognition with task dynamics

Our goal is to integrate task dynamics within an ASR system for continuous sentences called TD-ASR. Our approach is to re-rank an N-best list of sentence hypotheses according to a weighted likelihood of their articulatory realizations. For example, if a word sequence W_i : $w_{i,1}\ w_{i,2}\ \ldots\ w_{i,m}$ has likelihoods $L_X(W_i)$ and $L_\Lambda(W_i)$ according to purely acoustic and articulatory interpretations of an utterance, respectively, then its overall score would be

$$L(W_i) = \alpha L_X(W_i) + (1-\alpha) L_\Lambda(W_i) \tag{6.17}$$

given a weighting parameter α set manually, as in section 6.4.4. Acoustic likelihoods $L_X(W_i)$ are obtained from Viterbi paths through relevant HMMs in the standard fashion.

The TADA component

In order to obtain articulatory likelihoods, $L_\Lambda(W_i)$, for each word sequence, we first generate articulatory realizations of those sequences according to task dynamics. To this end, we use components from the open-source TADA system (Nam and Goldstein, 2006), which is a complete implementation of task dynamics. From this toolbox, we use the following components:

- A syllabic dictionary supplemented with the International Speech Lexicon Dictionary (Hasegawa-Johnson and Fleck, 2007). This breaks word sequences W_i into syllable sequences S_i consisting of onsets, nuclei, and coda and covers all of MOCHA and TORGO.

- A syllable-to-gesture lookup table. Given a syllabic sequence, S_i, this table provides the gestural goals necessary to produce those syllables. For example, given the syllable *pub* in figure 6.2, this table provides the targets for the GLO, VEL, TBCL, and TBCD tract variables, and the parameters for the second-order differential equation, eq. 6.1, that achieves those goals. These parameters have been empirically tuned by the authors of TADA according to a generic, speaker-independent representation of the vocal tract (Saltzman and Munhall, 1989).

- A component that produces the continuous tract variable paths that produce an utterance. This component takes into account various physiological aspects of human speech production, including intergestural and interarticulator co-ordination and timing (Nam and Saltzman, 2003; Goldstein and Fowler, 2003), and the neutral ("schwa") forces of the vocal tract (Saltzman and Munhall, 1989). This component takes a sequence of gestural goals predicted by the segment-to-gesture lookup table, and produces appropriate paths for each tract variable.

The result of the TADA component is a set of N 9-dimensional articulatory paths, $\mathbf{TV}_i$, necessary to produce the associated word sequences, W_i for $i = 1..N$. Since task dynamics is a prescriptive model and fully deterministic, $\mathbf{TV}_i$ sequences are the *canonical* or default articulatory realizations of the associated sentences. These canonical realizations are independent

of our training data, so we transform them in order to more closely resemble the observed articulatory behaviour in our EMA data. Towards this end, we train a switching Kalman filter identical to that in section 6.4.2, except the hidden state variable $\mathbf{x}_t$ is replaced by the observed instantaneous *canonical* TVs predicted by TADA. In this way we are explicitly learning a relationship between TADA's task dynamics and human data. Since the lengths of these sequences are generally unequal, we align the articulatory behaviour predicted by TADA with training data from MOCHA and TORGO using standard dynamic time warping (Sakoe and Chiba, 1978). During run-time, the articulatory sequence $\mathbf{y}_t$ most likely to have been produced by the human data given the canonical sequence $\mathbf{TV}_i$ is inferred by the Viterbi algorithm through the SKF model with all other variables hidden. The result is a set of articulatory sequences, $\mathbf{TV}_i^*$, for $i = 1..N$, that represent the predictions of task dynamics that better resemble our data.

Acoustic-articulatory inversion

In order to estimate the articulatory likelihood of an utterance, we need to evaluate each transformed articulatory sequence, $\mathbf{TV}_i^*$, within probability distributions ranging over all tract variables. These distributions can be inferred using acoustic-articulatory inversion with mixture density networks as described in section 6.2.2. We choose the MDN here, rather than the KCCA approach, because of the relatively demanding computational requirements of the latter and the desire to have complex probability distributions over the range of possible articulatory positions. Our networks are trained with acoustic and EMA-derived data from TORGO and MOCHA, as described in chapter 4.

Recognition by reranking

During recognition of a test utterance, a standard acoustic HMM produces word sequence hypotheses, W_i, and associated likelihoods, $L(W_i)$, for $i = 1..N$. The expected canonical motion of the tract variables, $\mathbf{TV}_i$ is then produced by task dynamics for each of these word sequences and transformed by an SKF to better match speaker data, giving $\mathbf{TV}_i^*$. The likelihoods of these

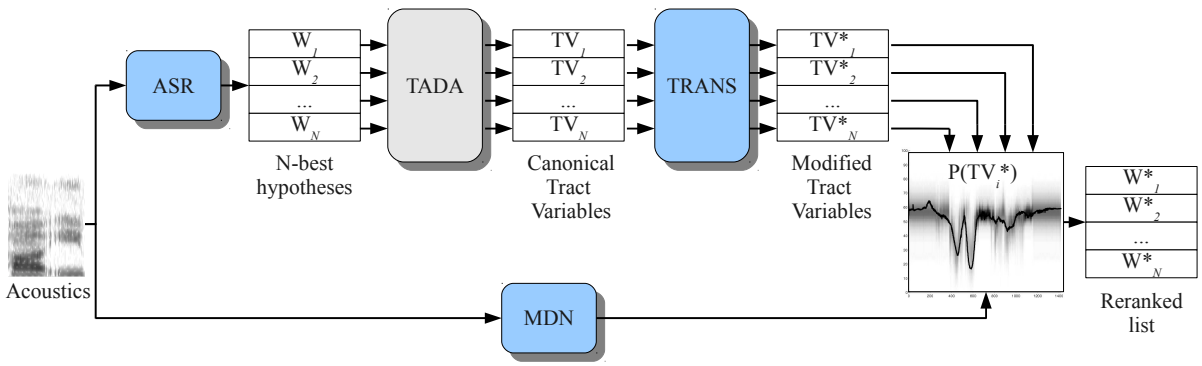


Figure 6.8: The TD-ASR mechanism for deriving articulatory likelihoods, $L_\Lambda(W_i)$, for each word sequence W_i produced by standard acoustic techniques.

paths are then evaluated within probability distributions produced by an MDN. The mechanism for producing the articulatory likelihood is shown in figure 6.8. The overall likelihood, $L(W_i) = \alpha L_X(W_i) + (1-\alpha)L_\Lambda(W_i)$, is then used to produce a final hypothesis list for the given acoustic input.

6.4.4 Experiments

Experimental data is obtained from two sources, as described in chapter 4. We procure 1200 sentences from Toronto's TORGO database, and 896 from Edinburgh's MOCHA. In total, there are 460 total unique sentence forms, 1092 total unique word forms, and 11065 total words uttered. Except where noted, all experiments randomly split the data into 90% training and 10% testing sets for 5-cross validation. MOCHA and TORGO data are never combined in a single training set due to differing EMA recording rates. In all cases, models are database-dependent (i.e., all TORGO data is conflated, as is all of MOCHA).

For each of our baseline systems, we calculate the phoneme-error-rate (PER) and word-error-rate (WER) after training. The phoneme-error-rate is calculated according to the proportion of frames of speech incorrectly assigned to the proper phoneme. The word-error-rate is calculated as the sum of insertion, deletion, and substitution errors in the highest-ranked hypothesis divided by the total number of words in the correct orthography. The traditional

System	Parameters	PER (%)	WER (%)
HMM	$\|M\| = 4$	29.3	14.5
	$\|M\| = 8$	27.0	13.9
	$\|M\| = 16$	26.1	10.2
	$\|M\| = 32$	25.6	9.7
DBN-A	$\|K\| = 4$	26.1	13.0
	$\|K\| = 8$	25.2	11.3
	$\|K\| = 16$	24.9	9.8
	$\|K\| = 32$	24.8	9.4

Table 6.6: Phoneme- and Word-Error-Rate (PER and WER) for different parameterizations of the baseline HMM and DBN-A systems.

HMM is compared by varying the number of Gaussians used in the modelling of acoustic observations. Similarly, the DBN-A model is compared by varying the number of discrete quantizations of articulatory configurations, as described in section 5.1.5. Results are obtained by direct decoding. The average results across both databases, between which there are no significant differences, are shown in table 6.6. In all cases the DBN-A model outperforms the HMM, which highlights the benefit of explicitly conditioning acoustic observations on articulatory causes.

Efficacy of TD-ASR components

In order to evaluate the whole system, we start by evaluating its parts. First, we test how accurately the mixture-density network (MDN) estimates the position of the articulators given only information from the acoustics available during recognition. Table 6.7 shows the average log likelihood over each tract variable across both databases. These results are consistent with the state-of-the-art (Toda, Black, and Tokuda, 2008). In the following experiments, we use MDNs that produce 4 Gaussians.

	No. of Gaussians			
	1	2	3	4
LTH	−0.28	−0.18	−0.15	−0.11
LA	−0.36	−0.32	−0.30	−0.29
LP	−0.46	−0.44	−0.43	−0.43
GLO	−1.48	−1.30	−1.29	−1.25
TTCD	−1.79	−1.60	−1.51	−1.47
TTCL	−1.81	−1.62	−1.53	−1.49
TBCD	−0.88	−0.79	−0.75	−0.72
TDCL	−0.22	−0.20	−0.18	−0.17

Table 6.7: Average log likelihood of true tract variable positions in test data, under distributions produced by mixture density networks with varying numbers of Gaussians.

We evaluate how closely transformations to the canonical tract variables predicted by `TADA` match the data. Namely, we input the known orthography for each test utterance into `TADA`, obtain the predicted canonical tract variables **TV**, and transform these according to our trained SKF. The resulting predicted and transformed sequences are aligned with our measurements derived from EMA with dynamic time warping. Finally, we measure the average difference between the observed data and the predicted (canonical and transformed) tract variables. Table 6.8 shows these differences according to the phonological manner of articulation. In all cases the transformed tract variable motion is more accurate, and significantly so at the 95% confidence level for nasal and retroflex phonemes, and at 99% for fricatives. The practical utility of the transformation component is evaluated in its effect on recognition rates, below.

Recognition with TD-ASR

With the performance of the components of TD-ASR better understood, we combine these and study the resulting composite TD-ASR system. Figure 6.9 shows the WER as a function of

Manner	Canonical	Transformed
approximant	0.19	0.16
fricative	0.37	0.29
nasal*	0.24	0.18
retroflex	0.23	0.19
plosive	0.10	0.08
vowel	0.27	0.25

Table 6.8: Average difference between predicted tract variables and observed data, on $[0,1]$ scale. (*) Nasals are evaluated only with MOCHA data, since TORGO data lacks velum measurements.

α with TD-ASR and $N = 4$ hypotheses per utterance. Recall that the overall likelihood of a word sequence hypothesis W is $L(W) = \alpha L_X(W) + (1-\alpha)L_\Lambda(W)$ (higher α signifies higher weight to the acoustic likelihood L_X relative to the articulatory likelihood L_Λ). The effect of α is clearly non-monotonic, with articulatory information clearly proving useful. Although systems whose rankings are weighted solely by the articulatory component perform better than the exclusively acoustic systems, the lists available to the former are procured from standard acoustic ASR. Interestingly, the gap between systems trained to the two databases increases as α approaches 1.0. Although this gap is not significant, it may be the result of increased inter-speaker articulatory variation in the TORGO database, which includes more than twice as many speakers as MOCHA.

Figure 6.10 shows the WER obtained with TD-ASR given varying-length N-best lists and $\alpha = 0.7$. TD-ASR accuracy at $N = 4$ is significantly better than both TD-ASR at $N = 2$ and the baseline approaches of table 6.6 at the 95% confidence level. However, for $N > 4$ there is a noticeable and systematic worsening of performance.

The optimal parameterization of the TD-ASR model results in an average word-error-rate of 8.43%, which represents a 10.3% relative error reduction over the best parameterization

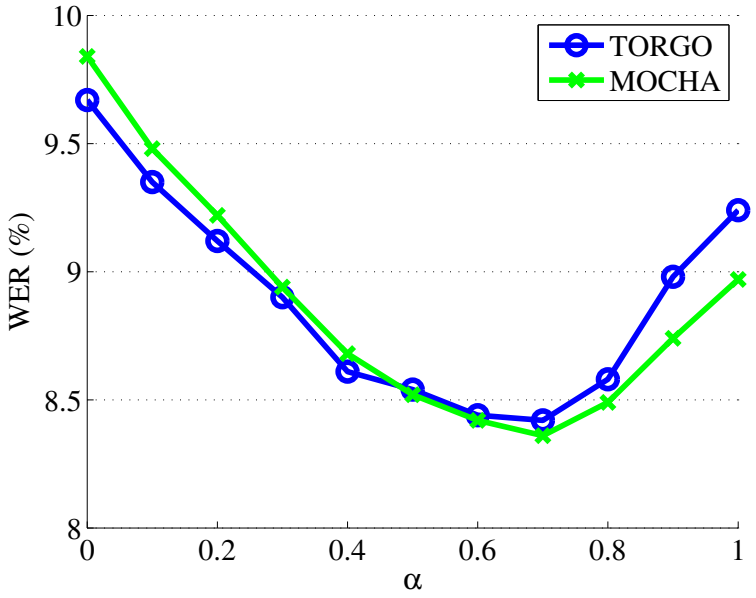


Figure 6.9: Word-error-rate according to varying α, for both TORGO and MOCHA data.

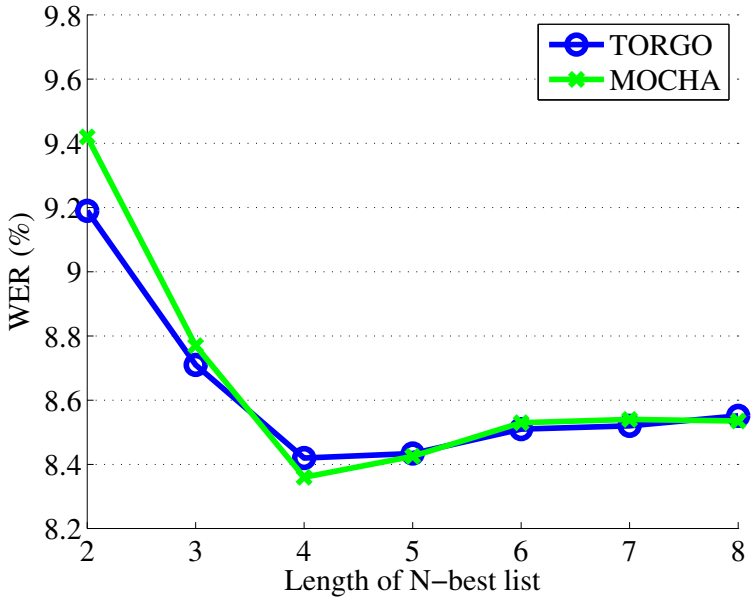


Figure 6.10: Word-error-rate according to varying lengths of *N*-best hypothesis lists used, for both TORGO and MOCHA data.

of our baseline models. The SKF model of section 6.4.2 differs from the HMM and DBN-A baseline models only in its use of continuous (rather than discrete) hidden dynamics and in its articulatory observations. However, its performance is far more variable, and less conclusive. On the MOCHA database the SKF model had an average of 9.54% WER with a standard deviation of 0.73 over 5 trials, and an average of 9.04% WER with a standard deviation of 0.64 over 5 trials on the TORGO database. Despite the presupposed utility of direct articulatory observations, the SKF system does not perform significantly better than the best DBN-A model.

Finally, the experiments of figures 6.9 and 6.10 are repeated with the canonical tract variables passed untransformed to the probability maps generated by the MDNs. Predictably, resulting articulatory likelihoods L_Λ are less representative and increasing their contribution α to the hypothesis reranking does not improve TD-ASR performance significantly, and in some instances worsens it. Although `TADA` is a useful prescriptive model of generic articulation, its use must be tempered with knowledge of inter-speaker variability.

6.4.5 Summary of integrating task-dynamics into ASR

We have demonstrated that the use of direct articulatory knowledge can substantially reduce phoneme and word errors in speech recognition, especially if that knowledge is motivated by high-level abstractions of vocal tract behaviour. Task-dynamic theory provides a coherent and biologically plausible model of speech production with consequences for phonology (Browman and Goldstein, 1986), neurolinguistics (Guenther and Perkell, 2004), and the evolution of speech and language (Goldstein, Byrd, and Saltzman, 2006). We have shown that it is also useful within speech recognition.

We have overcome a conceptual impediment in integrating task dynamics and ASR, which is the former's deterministic nature. This integration is accomplished by stochastically transforming predicted articulatory dynamics and by calculating the likelihoods of these dynamics according to speaker data. However, there are several new avenues for exploration. For example, task dynamics lends itself to more general applications of control theory, including

automated self-correction, rhythm, co-ordination, and segmentation (Friedland, 2005). Other high-level questions also remain, such as whether discrete gestures are the correct biological and practical paradigm, whether a purely continuous representation would be more appropriate, and whether this approach generalizes to other languages.

In general, our experiments have revealed very little difference between the use of MOCHA and TORGO EMA data. An *ad hoc* analysis of some of the errors produced by the TD-ASR system found no particular difference between how systems trained to each of these databases recognized nasal phonemes, although only those trained with MOCHA considered velum motion. Other errors common to both sources of data include phoneme insertion errors, normally vowels, which appear to co-occur with some spurious motion of the tongue between segments, especially for longer *N*-best lists. Despite the relative slow motion of the articulators relative to acoustics, there remains some intermittent noise.

6.5 Identifying articulatory goals using principal differential analysis

Previous sections applied task-dynamics given fixed parameterizations as obtained by the `TADA` component, namely the coefficients in equation 6.1. In order to generalize some of those findings and to extend task-dynamics in ASR, one requires a mechanism for obtaining these parameters directly from data. Typically, however, second-order equations of this type can only be solved given known parameterizations. This section describes a new approach to obtaining the parameters in task-dynamics from data.

Here, we examine a subset of the discrete articulatory features shown in table 2.1 of section 2.4. Of these representations, we are interested in the *Front/Back* and *High/Low* AFs, the values of which are derived directly from phonemic annotations as described in previous work (Wester, 2003; Scharenborg, Wan, and Moore, 2007), and as shown in table 6.9. Furthermore, we are interested in the binary features *Voiced/Unvoiced* and *Bilabial/Non-bilabial*. The former

Front/Back	Front	Central	Back
	/ae, aw, ay, eh, ey, ix, iy, ih/	*/ax, ah/*	*/ao, ow, oy, uh, uw, aa, ux/*
High/Low	High	Mid	Low
	/ix, iy, uh, uw, ih, ux/	*/ax, eh, ey, ow, ah/*	*/ae, ao, aw, ay, oy, aa/*

Table 6.9: Annotated phonemes used to derive specific AF classes, after Wester (Wester, 2003).

distinguishes all voiced sounds (i.e., vowels and sonorant consonants) from non-voiced sounds. The *Bilabial/Non-bilabial* AF has the value *bilabial* during phonemes */m/*, */em/* (i.e., an */m/* preceded by a vowel mora), */p/*, and */b/*, and the value *non-bilabial* otherwise.

Articulatory and acoustic data in this study are derived from the public MOCHA database from the University of Edinburgh (Wrench, 1999) as described in chapter 4. We use eight of the male speaker's articulatory parameters, namely the upper lip, lower lip, upper incisor, lower incisor, tongue tip, tongue blade, tongue dorsum, and velum. Each parameter is measured in the two dimensions of the midsaggital plane.

6.5.1 Principal differential analysis

The term *principal differential analysis* (PDA) is derived from principal components analysis (PCA) (Ramsay and Silverman, 2002; Ramsay and Silverman, 2005). PCA can also be applied to functional data, by treating each corresponding set of frames across the training sequences as measurements of an independent random variable. This is called *functional PCA* (FPCA), as explained by Ramsay and Silverman (2002).

Articulators are mechanical systems and, as such, are constrained in ways not captured by FPCA but which are expressible in terms of differential equations. *Principal differential analysis* (PDA) (Ramsay and Silverman, 2002) is similar to FPCA, but aimed at optimizing the parameters of a linear differential operator that hypothetically constrains a function from which

multiple noisy samples are available. Let L be a second order differential operator defined as

$$Lx_i(t) = \beta_0(t)x_i(t) + \beta_1(t)x_i'(t) + x_i''(t) = f_i(t), \tag{6.18}$$

where $x_i(t)$ is the functional observation from the i^{th} sample at time t, $x_i'(t)$ and $x_i''(t)$ are its first and second derivatives, β_j are the coefficients to be estimated, and $f_i(t)$ is the forcing function of the i^{th} sample at time t. If no forcing function has been observed then we make a simplifying assumption that all $f_i(t)$ are 0, giving us a linear homogeneous differential equation. In this case, PDA finds values of the coefficients $\beta_0(t)$ and $\beta_1(t)$ that minimize the residual $Lx_i(t)$, which can be obtained by Gaussian elimination. On this basis we can build a classifier for functional data by looking at the residuals that result from applying the learned coefficients of a given class to a new sequence.

6.5.2 PDA Classifier

We assume that we have functional observations on an arbitrary number of independent tracts, and that we wish to classify an unseen sequence as having an *articulatory value* or class c from the set of possibilities C for one articulatory feature.

The training procedure begins by normalizing the length of training sequences within each class, which is necessary in order to use PDA. We experimented with several normalization methods, and settled on finding the maximal sequence length within the class (according to the annotation), then shifting the end frame of all other training sequences so as to extend them to that length without distorting those sequences through time warping. This preserves all of the useful information from every sequence, at the cost of introducing some noise in later frames. Next, all tracts of all sequences are smoothed using a set of b-spline basis functions and by penalizing high fourth derivatives. Finally, for each $c \in C$ we run PDA, as described in the previous section, on the aggregated training sequences for c. For each tract, this gives us two coefficient vectors, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$.

In order to classify a new sequence, we compute its first and second derivatives on all tracts

by the method of central finite differences. Then for each $c \in C$ we find a residual vector on each tract t using the differential operator learned on t. Now we can calculate coefficients of determination R_t^2 as

$$R_t^2 = \frac{SSY_t - SSE_t}{SSY_t}, \tag{6.19}$$

where SSY_t is the sum of squared second derivatives on tract t and SSE_t is the sum of squared residuals. The resulting value is less than or equal to 1, with 1 indicating a perfect fit. Finally, we generate a score for c by averaging the coefficients of determination across all tracts t. The sequence is classified as having the articulatory value that assigns it the highest score.

Frame Weighting (FW)

One side-effect of the method that we chose to normalize sequence lengths is that the performance of PDA degrades in later frames of the training sequences, in the sense that the residuals it yields grow larger. This is due to some examples that were annotated as ending earlier having moved into irrelevant or possibly contradictory territory. To counteract this effect, we weight each frame according to the inverse of the squared residual that PDA yields on training data for that frame. During classification, we can multiply the residuals of the unknown sequence by the frame weights for the class in question, which generally places more emphasis on earlier frames.

6.5.3 Experiments with PDA

EMA data from MOCHA are first transformed to an approximation of the tract variable space as described in section 6.1. Our dataset consisted of $15,243$ phoneme instances with aligned acoustic and articulatory data. For all experiments, the data were randomly segregated into a training set and a held-out evaluation set. For each articulatory feature we limited our training and testing data to a subset of the available tracts. For the bilabial AF we used only the lip aperture tract, LA. For the high-low and front-back AFs, we used all of the tongue tip tracts - TTCL, TTCD, TBCL, and TBCD. For the voice AF we used only the glottis tract, GLO.

Articulatory domain

Our first set of experiments compares classifiers using only articulatory data. The baseline is a 5-state left-to-right HMM with observation likelihoods at each state computed over mixtures of 8 Gaussians. Training is performed with Baum-Welch expectation-maximization, and evaluation is performed by Viterbi decoding (Huang, Acero, and Hon, 2001). Each HMM is trained given observation sequences of a particular AF value (e.g., *non-bilabial*) and each Gaussian mixture in these HMMs is initialized given k-means clustering with full covariance over all data of the associated AF value. Table 6.10 shows the results of these experiments. We compared these with a most-frequent baseline classifier in which the most frequent class is blindly chosen for each test sequence, which averaged 67% accuracy. This naïve classifier obtained 87.2% accuracy on the bilabial AF, 62.8% on the high-low AF, 70.2% on the voicing AF, and 47.6% on the front-back AF. On average, PDA significantly outperforms HMMs.

Acoustic domain

We also compare the proposed PDA method given articulatory data against HMM and neural network (NN) baselines given acoustic data, which is a more common scenario, on the task of AF classification. In these experiments we use the full range of articulatory values for each articulatory feature. Specifically, the high-low feature has 5 classes (adding nil and silence), and the front-back feature has 4 classes (adding nil).

Here, the HMM baseline consists of tristate ergodic HMMs with 16 Gaussians per state. These take observations which are 42-dimensional MFCCs that include δ and $\delta\delta$ coefficients, and all models are initialized using k-means clustering on acoustic data. The NN baseline is based on similar work by Kirchhoff (1999) and Frankel, Wester, and King (2007). Each NN has three layers with full feed-forward connections, and is trained by resilient backpropagation. Input layers consist of 42 units, and output layers consist of one unit per class. The size of the hidden layers are dependent on the AF being recognized. The NNs that recognize the high-low and voicing features have 100 hidden units each, while the front-back feature has 200 units, as

	HMM	PDA	PDA+FW
Bilabial	94.5	87.8	96.7
Non-bilabial	74.6	94.6	93.3
All	76.1	93.8	93.8
High	53.2	47.6	44.9
Mid	28.7	43.1	100.0
Low	85.9	71.7	67.7
All	45.2	50.1	84.7
Voiced	98.1	98.0	99.8
Unvoiced	99.8	74.0	86.8
All	99.0	90.9	95.9
Front	22.4	46.1	39.3
Central	47.3	48.0	100.0
Back	62.6	43.8	65.0
All	43.5	46.6	74.9
Average	**66.0**	**70.4**	**87.3**

Table 6.10: Accuracy (%) of articulatory-domain classifiers, including principal differential analysis (PDA) with and without frame weighting (FW) across articulatory features.

	Acoustic HMM	NN	PDA+FW
High-low	48.6	64.8	67.4
Voice	71.6	83.3	95.9
Front-back	49.0	66.1	68.9
Average	**56.4**	**71.4**	**77.4**

Table 6.11: Average accuracies (%) of AF-recognition for HMM and NN classifiers as compared with the PDA approach given acoustic information only.

determined empirically in the literature (Frankel, Wester, and King, 2007).

Table 6.11 shows the results of the acoustic-domain experiments. Once again PDA is a clear winner. These results also demonstrate that the PDA classifier's performance holds up well as the number of classes increases.

The PDA classifier presented here offers a substantial improvement over the baselines. This may suggest that discrimination of speech sounds is highly dependent on how they are produced. Speech production can be modelled as a mechanical process, and the resulting models can be used to constrain our interpretations of articulatory motion in a very natural way. In the acoustic domain, on the other hand, it is very difficult to account for mechanical constraints.

Chapter 7

Speech transformation and synthesis

In previous chapters our focus has been on improving the rates of recognition of speech recognition for dysarthric speakers, given knowledge of the vocal tract. Despite advances made in this area, unrestricted large-vocabulary ASR remains a difficult problem for severely dysarthric speech. While this investigation must continue, we can already make use of our discoveries related to dysarthric speech production to develop applied software that can benefit such speakers. Our goal in this chapter is to establish and evaluate a set of techniques that modify dysarthric input acoustics to produce a more intelligible equivalent of that speech. These techniques include simple acoustic transformations, some of which are dependent on an annotation of the input. This exercise is meant to precede the eventual creation of application software for human-human interaction. In order to isolate the evaluation of these transformations from the ambiguities that can arise from imperfect dependencies such as speech recognition, we will assume that perfect phonemic annotations are available *a priori*.

Background on speech transformation and synthesis is provided in section 7.2 before the TORGOMorph transformations are proposed in section 7.3. This is followed by a brief series of experiments on the results of certain types of modification to the intelligibility of speech in section 7.4 and a discussion in section 7.5.

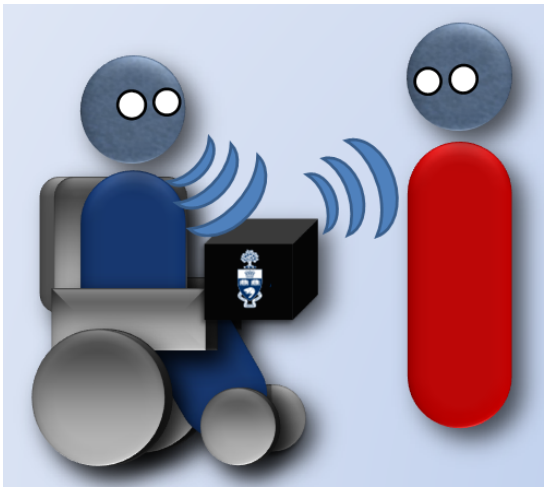


Figure 7.1: Hypothetical conversation between a speaker with dysarthria (left) and a member of the general population (right). The black box mounted on the wheel chair accepts spoken input and transmits a transformed version of that speech which is more intelligible to the typical hearer.

7.1 Usage scenario

Consider a dysarthric individual who is traveling into a city by transit to attend an appointment. This might involve purchasing tickets, asking for directions, or indicating their presence or intention to fellow passengers, which must often be done in a noisy and crowded environment. A personal portable communication device in this scenario (either held or attached to a wheelchair) would transform relatively unintelligible speech spoken into a microphone and play the results of that transformation over a set of speakers so that it could be better understood by a listener in that environment. Such a system could facilitate interaction and overcome difficult or failed attempts at communication in daily life. Such an interaction is represented in figure 7.1.

Many aspects of daily living are duplicated in order to provide access to individuals with disabilities. The Wheel-Trans service in Toronto, for example, provides accessible transit to persons with physical disabilities who cannot use regular transit and is operated by the city's general transit commission. However, this type of service can be more expensive than the

generic systems and it requires booking trips well in advance which can be inconvenient. In some cities, this type of service does not exist at all. Furthermore, such services still require the communication of one's destination or route in which augmentative systems would still be applicable or necessary.

There are augmentative communication devices that employ synthetic text-to-speech in which messages can be written on a specialized keyboard or played back from a repository of pre-recorded messages. However, the type of acoustics produced by such systems often lacks a sufficient degree of individual affectation or natural expression that one might expect in typical human speech (Kain et al., 2007). The use of prosody to convey personal information such as one's emotional state is simply not supported by such systems but is part of a general communicative ability. Transforming one's speech in a way that preserves the natural prosody will therefore also preserve extra-linguistic information, such as emotions, and is therefore a pertinent and crucial response to the limitations of current technology.

7.2 Background on speech transformation and synthesis

We often modify an input signal $x(t)$ (or its spectral envelope $X(f)$) into an output signal $y(t)$ (or its spectral envelope $Y(f)$) by means of a transfer function $H(\cdot)$. This transfer function can operate on several domains, such as short-term frequency characteristics of a signal. In equation 2.3 of section 2.3.1 we introduced the Fourier transform, which determines the amplitude of an arbitrary component sinusoid with frequency f in a signal. However, those component sinusoids can also be described in terms of their *phases* which are their respective offsets in time. In this chapter, the component phases are important in the accurate definition of transformation functions $H(\cdot)$.

In order to obtain a more general parameterization of a signal to include its component phases, we define the two-dimensional complex space $s = \sigma + j2\pi f$, where f is the frequency as before, σ is the phase, and j is the imaginary unit $j = \sqrt{-1}$. The Laplace transform gener-

alizes the Fourier transform for continuous signals as

$$X(s) = \int_{t=-\infty}^{\infty} x(t)e^{-st}\delta t. \tag{7.1}$$

Given the Laplace transform of a signal and the complex space s, a transfer function relating an input signal $X(s)$ and output $Y(s)$ is

$$H(s) = \frac{Y(s)}{X(s)} = \frac{\sum_{k=0}^{M} b_k s^k}{\sum_{k=0}^{N} a_k s^k} \tag{7.2}$$

where the roots of the numerator and denominator polynomials are the *zeros* and *poles* of the signal, respectively, and N and M are arbitrary orders of those polynomials [1] (O'Shaughnessy, 2000). When given only a discrete sampling of the signal, Laplace is replaced by the z-transform

$$X[z] = \sum_{n=-\infty}^{\infty} x[n]z^{-n} \tag{7.3}$$

where z is a complex frequency variable analogous to s in equation 7.1. Since equation 7.3 only sums to a finite value on circles in the complex domain, it is normally described in polar coordinates $z = \|z\| \exp(j2\pi f/F_S)$ where F_S is the sampling frequency (O'Shaughnessy, 2000).

7.2.1 Concatenative and articulatory synthesis

In order to produce speech that is as human-like as possible, a naïve approach is to record, split, and re-assemble actual human utterances. While individual segments can be highly intelligible and natural, their concatenation often results in an unnatural-sounding discord, especially when adjacent phonemes are incompatible (Huang, Acero, and Hon, 2001). To avoid discontinuities of this type, vast corpora of speech segments reflecting various phonetic and emotional contexts are often stored in order to maintain some continuity. Moreover, the boundaries between sonorants are often blended using a technique called time-domain pitch-synchronous overlap-add (Moulines and Charpentier, 1990) in which signals are reconstructed by positioning adjacent segments so that they overlap according to the estimated glottal closure (Schroeter, 2008).

[1] An 'all-pole' model defines only coefficients in the denominator.

The vocal tract is often modelled as a concatenation of many idealized cylindrical tubes aligned at their centers where the k^{th} tube has a cross-sectional area of A_k, as shown in figure 7.2(a). Here, the volume beyond the lips is typically modelled as a tube with an infinite width. In the simplest realization of this model, the glottis produces an oscillating volume velocity, $u_G(t)$ as a function of time t, such as the spline[2] in figure 7.2(b). Here, the wave produced by the glottis is often assumed to be planar and propagated along the axis of the tubes without loss due to viscosity or thermal conduction along the walls[3] (Huang, Acero, and Hon, 2001). If the area of a tube A is fixed, $\rho \approx 1.2$ kg/m^3 is the density of air, and $c \approx 344$ m/s is the speed of sound in a human mouth, the sound waves in this model satisfy

$$\begin{aligned} -\frac{\delta p(x,t)}{\delta x} &= \frac{\rho}{A}\frac{\delta u(x,t)}{\delta t} \\ -\frac{\delta u(x,t)}{\delta x} &= \frac{A}{\rho c^2}\frac{\delta p(x,t)}{\delta t} \end{aligned} \tag{7.4}$$

where $u(x,t)$ and $p(x,t)$ are the volume velocity (in m/s) and pressure (in kg/m^3) at position x (the glottis is the origin) in the tube at time t (Quatieri, 2002). The pressure and volume of the k^{th} tube is then

$$\begin{aligned} u_k(x,t) &= u_k^+(t-x/c) - u_k^-(t+x/c) \\ p_k(x,t) &= \frac{\rho c}{A_k}\left[u_k^+(t-x/c) - u_k^-(t+x/c)\right] \end{aligned} \tag{7.5}$$

where $u_k^+(\cdot)$ and $u_k^-(\cdot)$ are are the waves travelling towards the lips and glottis, respectively, and x is measured from the left-most point in the k^{th} tube (Huang, Acero, and Hon, 2001). The shaping of the sound spectrum occurs because of the changes in the areas of adjacent tubes. At the junction between the k^{th} and $k+1^{st}$ tubes, part of the outgoing wave is reflected back into its originating tube by the reflection coefficient

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{7.6}$$

with larger differences in tube areas reflecting more energy (Deller, Hansen, and Proakis, 2000). The transfer function between the z-transforms of the wave velocities at the lips u_L

[2]Generally, a spline is a piecewise function composed of polynomials.

[3]Not all models are so naïve. The Hagen-Poiseuille flow model, for instance, assumes a parabolic acoustic wave whose velocity is maximal at the axis in the direction of motion and zero at the walls (Boersma, 1998).

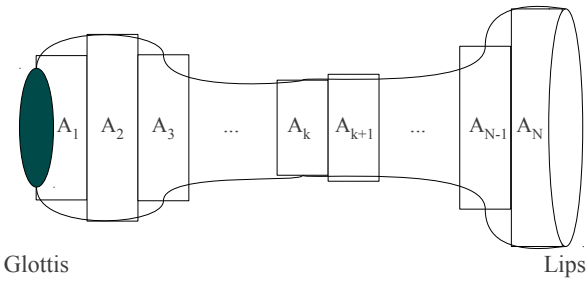
and the glottis u_G given N concatenated idealized tubes is

$$V(z) = \frac{U_L(z)}{U_G(z)} = \frac{0.5z^{-N/2}(1+r_G)\prod_{k=1}^{N}(1+r_k)}{[1-r_G]\left(\prod_{k=1}^{N}\begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix}\right)\begin{bmatrix} 1 \\ 0 \end{bmatrix}} \tag{7.7}$$

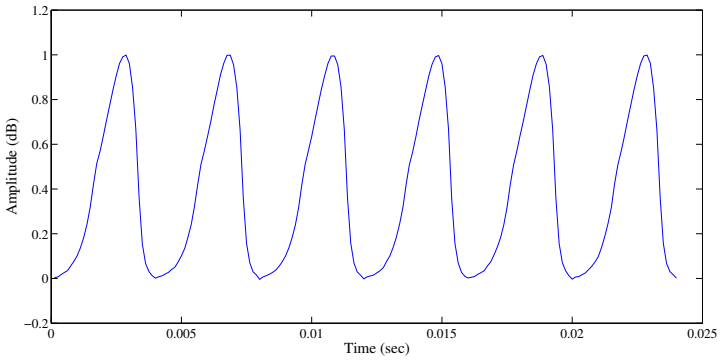
where r_G and $r_N = r_L$ are the reflection coefficients of the glottis and lips, respectively (Deller, Hansen, and Proakis, 2000). In practice, these reflection coefficients are functions of frequency so, for example, $r_L = 1$ when measuring lower frequencies of z so that all energy is transmitted, but $r_L < 1$ at higher frequencies (Huang, Acero, and Hon, 2001). In this model, equation 7.7 describes the spectrum of speech at the lips given knowledge of the produced and reflected waves in the leftmost tube. To account for non-sonorant phonemes such as plosives, fricatives, and affricates, the glottal pulse train is typically replaced with a low-amplitude white-noise generator whose signal passes through $V(z)$ as before (Quatieri, 2002). Extensions exist that allow for nasals by introducing a three-way boundary at a tube midway along the simulated vocal tract to emulate the lowering of the velum (usually with an associated closing of the rightmost tube representing the lips) (Boersma, 1998; Huang, Acero, and Hon, 2001).

Dynamic models of articulation

Over the past several decades, a number of instantiations of the basic principles outlined above have arisen that implement control by an active speaker of the basic uniform-tube model through the exertion of simulated muscles. Coker (1968) proposed a model that shaped the uniform-tube model to a more realistic arrangement mimicking the midsagittal plane of a human vocal tract. Crucially, this allowed for the articulators to be explicitly identified and moved during synthesis, so that particular phonemes could be associated with desired configurations of these articulators, which were described in a code book (Coker, 1976). The configurations of these articulators would warp the physical model of the uniform tube along horizontal, vertical, and radial dimensions, as indicated in figure 7.3. The timing and motion between articulatory positions were in some cases explicitly defined and in others interpolated automatically through



(a) Uniform tube model



(b) Glottal pulse train

Figure 7.2: Acoustical model of speech production. (a) Uniform-tube model and (b) six pitch periods of the glottal pulse at F0 = 250 Hz.

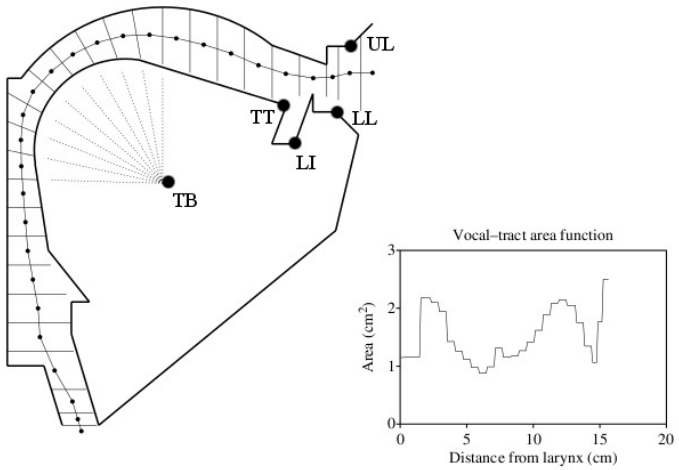


Figure 7.3: Coker model of the vocal tract for speech synthesis. Movable articulators include the tongue body (TB), tongue tip (TT), lower incisor (LI), and upper and lower lips (UL, LL), which move in tandem. Articulators move in dimensions determined by the shape of the vocal tract, adapted from Boersma (1998).

simple linear transformations. This model is in almost all important respects reflected in the independently-developed model of Mermelstein (1973), which was later extended by Rubin, Baer, and Mermelstein (1981) and would come to be known as the CASY (Configurable Articulatory Synthesis) model (Iskarous et al., 2003) used in the `TADA` system discussed in section 6.3, for instance. Other re-implementations of this same underlying model would later include low-level compensation for allophones in different articulatory contexts (Maeda, 1990), encoding of myoelastic effects between artificial muscles and their elastic connected tissues by means of spring-mass systems (Boersma, 1998), and a full complement of articulatory muscles that deform the vocal tract tube model, including the stylo-, genio-, and hyo-glossus tongue muscles and sternohyoid, for instance (Boersma, 1999).

The Klatt synthesizer

Synthesis-by-rule is a less biologically plausible approach to speech synthesis that nevertheless focuses on the realistic acoustic properties of the generated speech. Here, a formant resonance

can be generated at a specified frequency F_i and bandwidth B_i with

$$H_i(z) = \frac{1}{1 - 2e^{-\pi B_i/s_r}\cos(2\pi F_i/s_r)z^{-1} + e^{-2\pi B_i/s_r}z^{-2}} \tag{7.8}$$

where s_r is the sampling rate (Huang, Acero, and Hon, 2001). Klatt (1980) proposes a model which independently simulates acoustic resonances of this type given parameters determined 'by hand' for various parts of speech. For vowels, a bank of six of these resonators is activated in parallel and their outputs are summed together. For nasals, similar resonances are summed together, although the zeros between resonances are also specified (McLennan, 2000). This basic approach is one of the most popular in rule-based synthesis, and a number of derivative implementations have refined the specification of parameter values according to human data (O'Shaughnessy, 2000). In particular, in the following sections we assume the formant parameters for frequency and bandwidth for a stereotypical male speaker as determined by Allen et al. (1987) and listed in appendix C.

7.2.2 Measuring intelligibility

The intelligibility of both purely synthetic and modified speech signals can be measured objectively by simply having a set of participants transcribe what they hear from a selection of word, phrase, or sentence prompts (Spiegel et al., 1990), although no single standard has emerged as pre-eminent (Schroeter, 2008). Occasionally, ASR systems are used to approximate intelligibility (see section 3.1). Hustad (2006) suggests that orthographic transcriptions provide a more accurate predictor of intelligibility among dysarthric speakers than the more subjective estimates used in clinical settings, e.g., Enderby (1983) and Yorkston and Beukelman (1981). That study had 80 listeners who transcribed audio (which is atypically large for this task). It showed that intelligibility increases from 61.9% given only acoustic stimuli to 66.75% given audiovisual stimuli on the transcription task in normal speech. In the current work, we modify only the acoustics of dysarthric speech; however future work might consider how to prompt listeners in a more multimodal context.

7.2.3 Acoustic transformation

Kain et al. (2007) propose the voice transformation system shown in figure 7.4. This system produces output speech by concatenating together original unvoiced segments with synthesized voiced segments that consist of a summation of the original high-bandwidth signal with synthesized low-bandwidth formants. These synthesized formants are produced by modifications to input energy, F0 generation, and formant modifications. Modifications to energy and formants are performed by Gaussian mixture mapping, as described below, in which learned relationships between dysarthric and target acoustics are used to produce output closer to the target space. This process was intended to be automated, but Kain et al. (2007) performed extensive hand-tuning and manually identified formants in the input. This will obviously be impossible in a real-time system, but these processes can to some extent be automated. For example, voicing boundaries can be identified by the weighted combination of various acoustic features (e.g., energy, zero-crossing rate, first LPC coefficient) (Kida and Kawahara, 2005; Hess, 2008), and formants can be identified by the Burg algorithm (Press et al., 1992) or through simple LPC analysis (see section 2.3.1) with continuity constraints on the identified resonances between adjacent frames (O'Shaughnessy, 2008).

Spectral modifications traditionally involve spectral filtering or amplification methods such as spectral subtraction or harmonic filtering (O'Shaughnessy, 2000), but these are not useful for dealing with more serious mispronounciations (e.g., /t/ for /n/). Hosom et al. (2003) show that Gaussian mixture mapping can be used to transform from one set of spectral acoustic features to another space. During analysis, context-independent frames of speech are analyzed for bark-scaled energy and their 24^{th} order cepstral coefficients with

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \\ \hat{X}(k) &= \log |X\left(Bark(k)\right)| \\ c[n] &= \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}(k) e^{j\frac{2\pi}{N}kn}. \end{aligned} \tag{7.9}$$

For synthesis, a cepstral analysis approximates the original spectrum, and a high-order

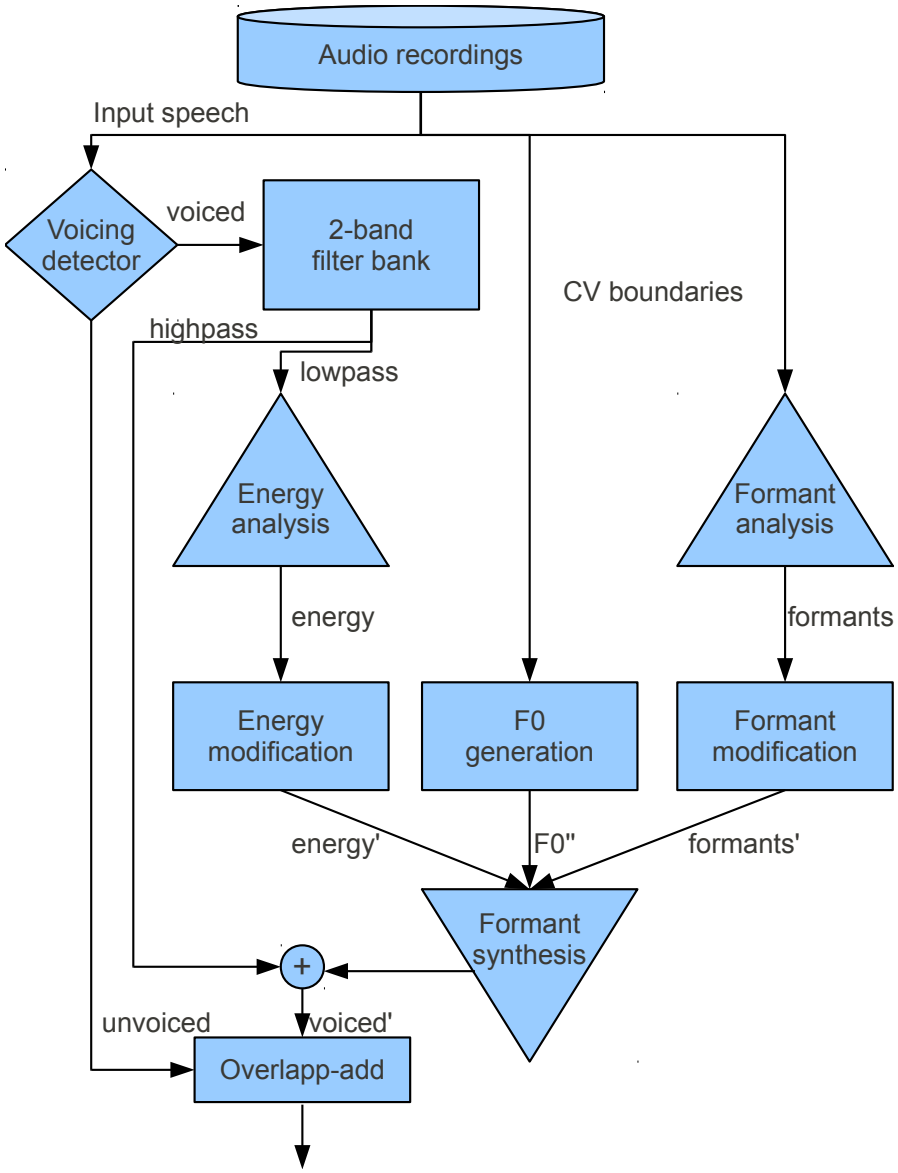


Figure 7.4: Voice transformation system proposed by Kain et al. (2007).

LPC filter is applied to each frame, and excited by impulses or white noise (for voiced and unvoiced segments). Hosom et al. show that given 99% human accuracy in recognizing normal speech data, this method of reconstruction gives 93% accuracy on the same data. They then trained a transformative model between dysarthric and regular speech using aligned, phoneme-annotated, and orthographically identical sentences spoken by dysarthric and regular speakers, and a Gaussian Mixture Model (GMM) to model the probability distribution of the dysarthric source spectral features x as the sum of D normal distributions with mean vector μ, diagonal covariance matrix Σ, and prior probability α:

$$p(x) = \sum_{d=1}^{D} \alpha_d \mathbf{N}(x; \mu_d, \Sigma_d). \tag{7.10}$$

The GMM parameters were trained in an 'unsupervised' mode using the EM algorithm and 1, 2, 4, 8, and 16 mixture components, with $D = 4$ apparently being optimal. A probabilistic least-squares regression mapped the source features x onto the target (regular speaker) features y, producing the model $W_d(x) + b_d$ for each class, and a simple spectral distortion is performed to produce regularized versions of dysarthic speech $\hat{y}$:

$$\hat{y}(x) = \sum_{d=1}^{D} h_d(x) \left(W_d(x) + b_d\right) \tag{7.11}$$

for posterior probabilities $h_d(x)$. This model is interesting in that it explicitly maps the acoustic differences for different features between disordered and regular speech[4]. Reconstructing the dysarthric spectrum in this way to sound more 'normal' while leaving F_0, timing and energy characteristics intact resulted in a 59.4% relative error rate reduction (68% to 87% accuracy) among a group of 18 naive human listeners for a total of 206 dysarthric test words each (Hosom et al., 2003).

[4]This model can also be used to measure the difference between any two types of speech.

7.3 The TORGOMorph transformations

TORGOMorph encapsulates of a number of transformations of the acoustics uttered by a speaker with dysarthria. Each modification is implemented to counter a particular effect of dysarthria on intelligibility as determined by observations on the TORGO data described in section 4.4. Currently, these modifications are uniformly preceded by noise reduction by spectral subtraction (see section 4.2.4) and either phonological or phonemic annotations. This latter step is currently necessary, since certain modifications require either knowledge of the manner of articulation or the identities of the vowel segments, as explained in the subsections below. The purpose of this exercise is to determine which modifications result in the most significant improvements to intelligibility, so the correct annotation sequence is vital to avoid the introduction of an additional dimension of error. Therefore, the annotations used below are extracted directly from the professional markup in the TORGO database. In practice, however, phonemic annotations determined automatically by speech recognition would be imperfect, which is why investigations of this type often forgo that automation altogether (e.g., see Kain et al. (2007) in section 7.2.3). Possible alternatives to full ASR are discussed in section 7.5.

In some cases, the dysarthric speech must be compared or supplemented with another vocal source. Here, we synthesize segments of speech using a text-to-speech application developed by Black and Lenzo (2004). This system is based on the University of Edinburgh's Festival tool and synthesizes phonemes using a standard LPC-based method introduced above with a pronunciation lexicon and part-of-speech tagger that assists in the selection of intonation parameters (Taylor, Black, and Caley, 1998). This system is invoked by providing the expected text uttered by the dysarthic speaker. In order to properly combine this purely synthetic signal and the original waveforms we require identical sampling rates, so we resample the former by a rational factor using a polyphase filter with low-pass filtering to avoid aliasing (Hayes, 1999). Since the discrete phoneme sequences themselves can differ, we find an ideal alignment between the two by the Levenshtein algorithm (Levenshtein, 1966), which is similar to dynamic time warping (see algorithm 1 in section 2.3.2) and which provides the total number

of insertion, deletion, and substitution errors.

The following sections detail the components of TORGOMorph, which is outlined in figure 7.5. These components allow for a cascade of one transformation followed by another, although in the experiments in section 7.4 these are performed independently to isolate their effects. In all cases, the spectrogram is derived with the fast Fourier transform given 2048 bins on the range of 0–5 kHz. Voicing boundaries are extracted in a unidimensional vector aligned with the spectrogram using the method of Kida and Kawahara (2005) which uses GMMs trained with zero-crossing rate, amplitude, and the spectrum as input parameters. A pitch contour is also extracted from the source by the method proposed by Kawahara et al. (2005), which uses a Viterbi-like potential decoding of F_0 traces described by cepstral and temporal features. That work showed an error rate of less than 0.14% in recognizing F_0 curves as compared with simultaneously-recorded electroglottograph data. Similar methods exist that use GMMs and neural networks for global optimization (Ewender, Hoffmann, and Pfister, 2009). This pitch contour is not in general modified by the methods proposed below, since Kain et al. (2007) showed that using original F_0 curves results in the highest intelligibility among alternative systems. Over a few segments, however, this curve can sometimes be decimated in time during the modification proposed in section 7.3.3 and in some cases removed entirely (along with all other acoustics) in the modification proposed in section 7.3.2.

7.3.1 High-pass filter on unvoiced consonants

The first acoustic modification is based on the observation that unvoiced consonants are improperly voiced in up to 18.7% of plosives (e.g. */d/* for */t/*) and up to 8.5% of fricatives (e.g. */v/* for */f/*) in dysarthric speech in the TORGO database (see section 4.4). Voiced consonants are typically differentiated from their unvoiced counterparts by the presence of the *voice bar*, which is a concentration of energy below 150 Hz indicative of vocal fold vibration that often persists throughout the consonant or during the closure before a plosive (Stevens, 1998). Empirical analysis of TORGO data suggests that for at least two male dysarthric speakers this

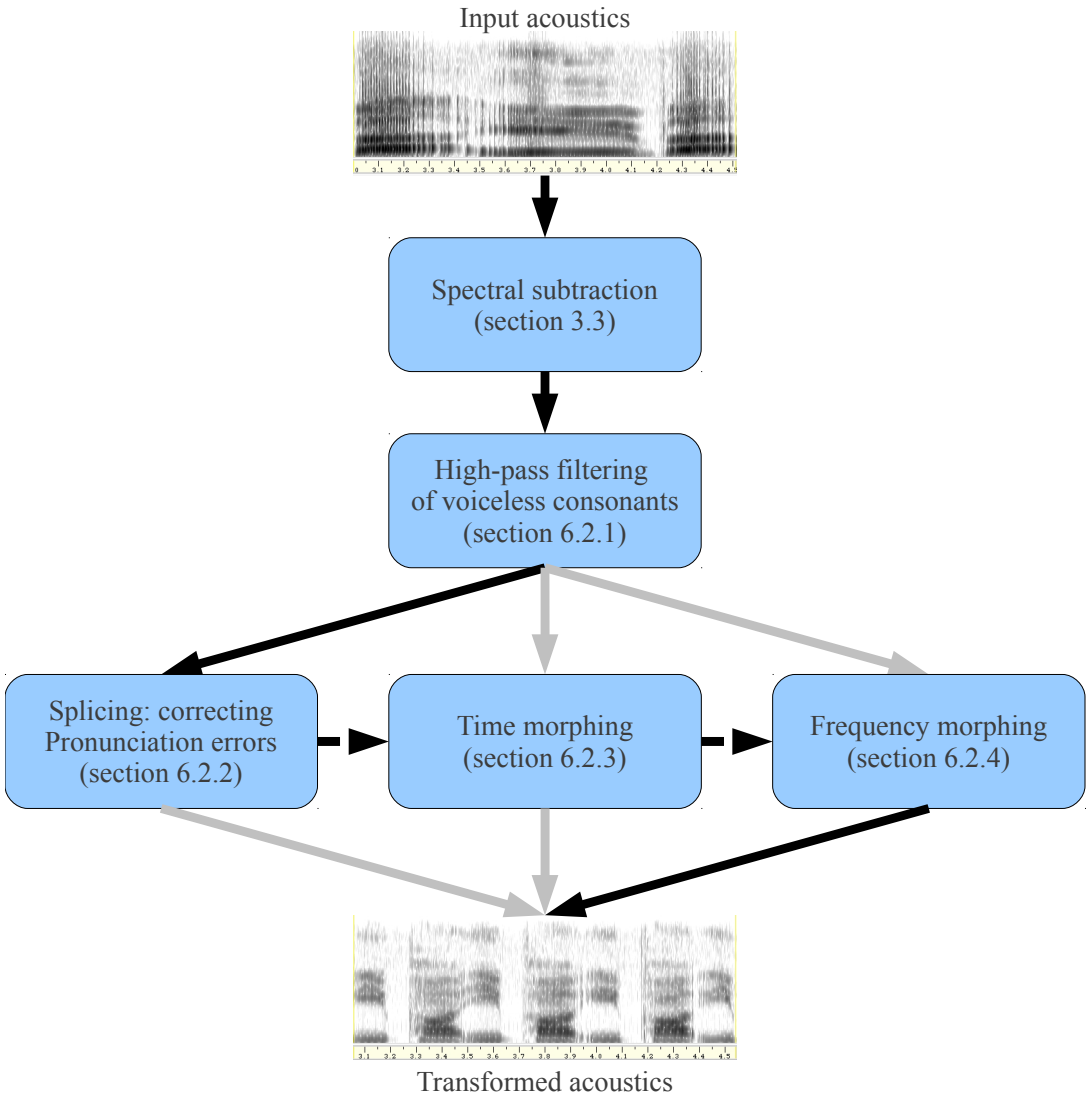


Figure 7.5: Outline of the TORGOMorph transformations. Black path indicates hypothesized cascade system to be used in practice. Solid arrows indicate paths taken during evaluation.

voice bar extends considerably higher, up to 250 Hz.

In order to correct these mispronunciations, the voice bar is filtered out of all acoustic subsequences annotated as unvoiced consonants. For this task we use a high-pass Butterworth filter, which is "maximally flat" in the passband[5] and monotonic in magnitude in the frequency domain (Butterworth, 1930). This is in contrast to the popular Chebyshev filter, for instance, which contains distorting ripples in the passband (Parks and Burns, 1987). Here, this filter is computed on a normalized frequency range respecting the Nyquist frequency, so that if a waveform's sampling rate is 16 kHz, the normalized cutoff frequency for this component is $f^*_{Norm} = 250/(1.6 \times 10^4/2) = 3.125 \times 10^{-2}$. The effect of this cutoff frequency is shown in the magnitude response in figure 7.6. The Butterworth filter is an all-pole transfer function between signals as described in section 7.2. Here, we use the 10^{th} order low-pass Butterworth filter whose magnitude response is

$$|\mathscr{B}(z;10)|^2 = |H(z;10)|^2 = \frac{1}{1+\left(jz/jz^*_{Norm}\right)^{2\times 10}} \tag{7.12}$$

where z is the complex frequency in polar coordinates and z^*_{Norm} is the cutoff frequency in that domain (Hayes, 1999). This translates into the transfer function

$$\mathscr{B}(z;10) = H(z;10) = \frac{1}{1+z^{10}+c_1z^9+\ldots+c_9z+c_{10}} \tag{7.13}$$

whose poles occur at known symmetric intervals around the unit complex-domain circle (Butterworth, 1930). These poles are then transformed by the Matlab function `zp2ss`, which produces the state-space coefficients α_i and β_i that describe the output signal resulting from applying the low-pass Butterworth filter to the discrete signal $x[n]$. These coefficients are further converted by

$$\begin{aligned} \vec{a} &= z^*_{Norm}\vec{\alpha}^{-1} \\ \vec{b} &= -z^*_{Norm}\left(\vec{\alpha}^{-1}\vec{\beta}\right) \end{aligned} \tag{7.14}$$

[5]The passband is the frequency range in which the component magnitudes in the original signal should not be changed.

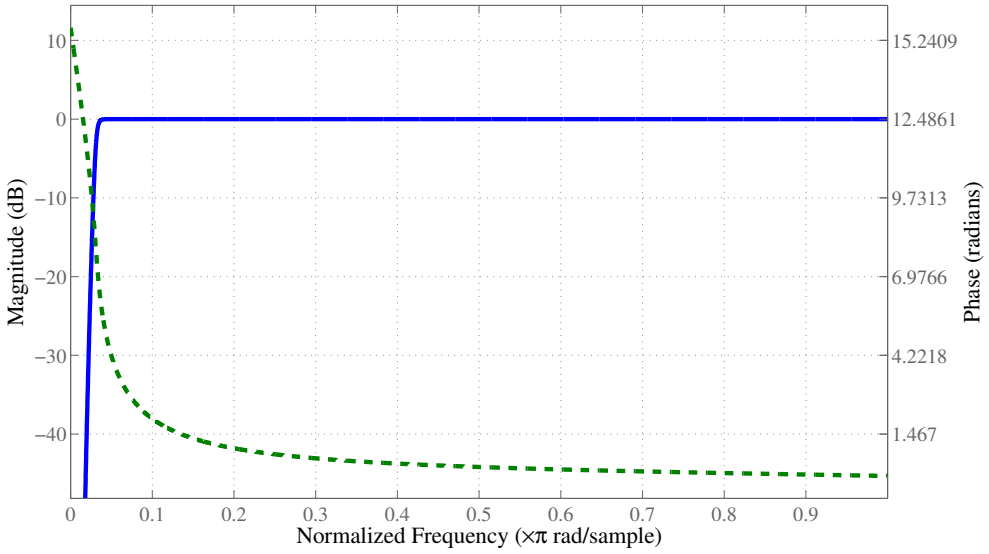


Figure 7.6: The Butterworth high pass filter with a cutoff frequency of 250 Hz in speech sampled at 16 kHz. The solid curve indicates the magnitude response in decibels relative to the original signal over the normalized frequency range, and the dashed curve represents the frequency response which measures the amount of phase shift in the resulting signal in radians.

giving the high-pass Butterworth filter with the same cutoff frequency of z^*_{Norm}[6]. This continuous system is converted to the discrete equivalent through the impulse-invariant discretization method and is implemented by the difference equation

$$y[n] = \sum_{k=1}^{10} a_k y[n-k] + \sum_{k=0}^{10} b_k x[n-k]. \tag{7.15}$$

as shown in figure 7.6. As previously mentioned, this equation is applied to each acoustic subsequence annotated as unvoiced consonants, thereby smoothly removing the energy below 250 Hz.

[6]See the Matlab function `lp2hp`.

7.3.2 Splicing: correcting dropped and inserted phoneme errors

The Levenshtein algorithm mentioned above finds a best possible alignment of the phoneme sequence in actually uttered speech and the expected phoneme sequence, given the known word sequence. Isolating phoneme insertions and deletions are therefore a simple matter of iteratively adjusting the source speech according to that alignment. There are two cases where action is required:

insertion error In this case a phoneme is present where it ought not be. In every such case for speaker M04, these insertion errors are repetitions of phonemes occurring in the first syllable of a word, according to the International Speech Lexicon Dictionary (Hasegawa-Johnson and Fleck, 2007). When an insertion error is identified the entire associated segment of the signal is simply removed. In the case that the associated segment is not surrounded by silence, adjacent phonemes can be merged together with time-domain pitch-synchronous overlap-add (Moulines and Charpentier, 1990), although in the data used here this is unnecessary, since inserted phoneme subsequences are always surrounded by stop-gaps.

deletion error Speaker M04 follows the same general behaviour as outlined in table 4.1 of section 4.4 in that the vast majority of accidentally deleted phonemes are fricatives, affricates, and plosives. Often, this involves not pluralizing nouns (e.g., *book* for *books*). Given the preponderance of error with these phonemes, these are the only classes we insert into the dysarthric source speech. Specifically, when the deletion of a phoneme is recognized, we simply extract the associated segment from the aligned synthesized speech and insert it into the appropriate spot in the dysarthric speech. For all unvoiced fricatives, affricates, and plosives no further action is required. When these phonemes are voiced, however, we first extract and remove the F_0 curve from the synthetic speech, linearly interpolate the F_0 curve from adjacent phonemes in the source dysarthric speech, and resynthesize with the synthetic spectrum and interpolated F_0. If interpolation is

not possible (e.g., the synthetic voiced phoneme is to be inserted beside an unvoiced phoneme), we simply generate a flat F_0 equal to the nearest natural F_0 curve.

7.3.3 Morphing in time

Figures 4.9, 4.10, and 4.11 in section 4.4 show that vowels uttered by a dysarthric speaker are significantly slower than those uttered by a typical speaker and can be up to twice as long, on average. In this modification, phoneme sequences identified as sonorant are simply contracted in time in order to be equal in extent to the greater of half their original length or the equivalent synthetic phoneme's length. In all cases this involved shortening the dysarthric source sonorant.

Since we wish to contract the length of a signal segment here without affecting its pitch or frequency characteristics, we use a phase vocoder based on digital short-time FFT analysis (Portnoff, 1976). Here, Hamming-windowed segments of the source phoneme are analyzed with a z-transform giving both frequency and phase estimates for up to 2048 frequency bands. During pitch-preserving time-scaled warping, we specify the magnitude spectrum directly from the input magnitude spectrum with phase values chosen to ensure continuity (Sethares, 2007). Specifically, for the frequency band at frequency F and frames j and $k > j$ in the modified spectrogram, the phase is predicted by

$$\theta_k^{(F)} = \theta_j^{(F)} + 2\pi F(j-k). \tag{7.16}$$

In our case the discrete warping of the spectrogram involves simple decimation by a constant factor. The spectrogram is then converted into a time-domain signal modified in tempo but not in pitch relative to the original phoneme segment. This conversion is accomplished simply through the inverse Fourier transform in which the transformation in equation 2.3 of section 2.3.1 is inverted with

$$w_s[n] = \frac{1}{N}\sum_{k=0}^{K-1} X[k]e^{\frac{j2\pi nk}{N}}, \tag{7.17}$$

where $w[n]$ is the generated waveform at discreet time n, k is a frequency band of the spectrum X, and K is the total number of such bands (Press et al., 1992).

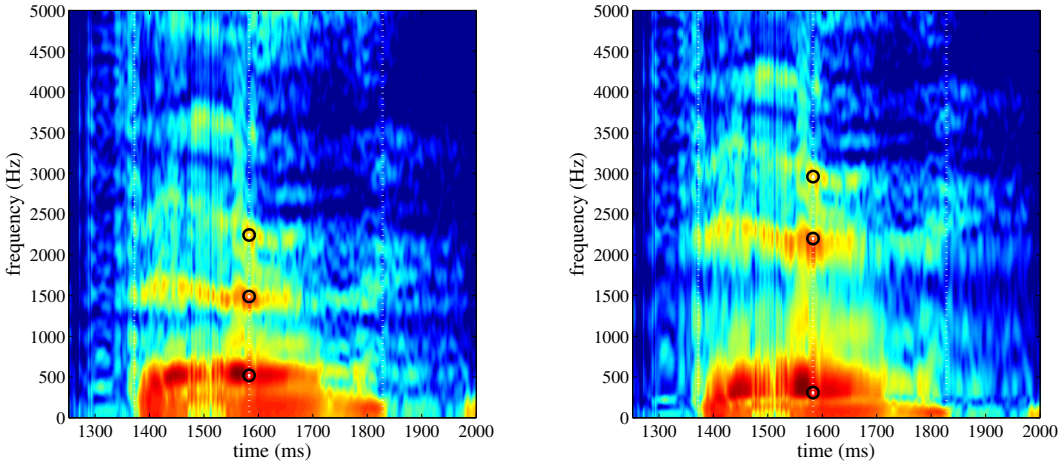
7.3.4 Morphing in frequency

Formant trajectories inform the listener as to the identities of vowels, but the vowel space of dysarthric speakers tends to be constrained (see section 5.5.2). In order to improve a listener's ability to differentiate between the vowels, this modification component identifies formant trajectories in the acoustics and modifies these according to the known vowel identity of a segment. Here, formants are identified through a simple LPC analysis with a 14^{th} order linear-predictive coder with continuity constraints on the identified resonances between adjacent frames (Snell and Milinazzo, 1993; O'Shaughnessy, 2008). Bandwidths are determined by the negative natural logarithm of the pole magnitude, as implemented in the STRAIGHT analysis system (Banno et al., 2007; Kawahara, 2006).

For each identified vowel in the dysarthric speech[7], formant candidates are identified at each frame in time up to 5 kHz. Only those time frames having at least 3 such candidates within 250 Hz of expected values are then considered. The expected values of formants are derived from analyses performed by Allen et al. (1987) and are shown in appendix C. Given these subsets of candidate time frames in the vowel, the one having the highest spectral energy within the middle 50% of the length of the vowel is established as the *anchor position*, and the three formant candidates within the expected ranges are established as the *anchor frequencies* for formants F_1 to F_3. If more than one formant candidate falls within expected ranges, the one with the lowest bandwidth becomes the anchor frequency.

Given identified anchor points and target sonorant-specific frequencies and bandwidths (see appendix C), there are several methods to modify the spectrum. The most common appears to be to learn a statistical conversion function based on Gaussian mixture mapping, as described elsewhere in this dissertation (e.g., sections 5.5.2, 6.2, and 7.2.3), typically preceded by alignment of sequences using dynamic time warping (Stylianou, 2008). Here, we use the STRAIGHT morphing implemented by Kawahara and Matsui (2003), among others. Here,

[7] Accidentally inserted vowels are also included here, unless previously removed by the splicing technique in section 7.3.2.



(a) (b)

Figure 7.7: Spectrograms for (a) the dysarthric original and (b) the frequency-modified renditions of the word *fear*. Circles represent indicative formant locations.

the transformation of a frame of speech x_A for speaker A is performed with a multivariate frequency-transformation function $T_{A\beta}$ given known targets β using

$$\begin{aligned} T_{A\beta}(x_A) &= \int_0^{x_A} \exp\left(\log\left(\frac{\delta T_{A\beta}(\lambda)}{\delta\lambda}\right)\right)\delta\lambda \\ &= \int_0^{x_A} \exp\left((1-r)\log\left(\frac{\delta T_{AA}(\lambda)}{\delta\lambda}\right) + r\log\left(\frac{\delta T_{A\beta}(\lambda)}{\delta\lambda}\right)\right)\delta\lambda \\ &= \int_0^{x_A} \left(\frac{\delta T_{A\beta}(\lambda)}{\delta\lambda}\right)^r \delta\lambda, \end{aligned} \tag{7.18}$$

where λ is the frame-based time dimension and where $0 \leq r \leq 1$ is an interpolative rate at which to perform morphing (i.e., $r = 1$ implies complete conversion of the parameters of speaker A to parameter set β and $r = 0$ implies no conversion.) (Kawahara et al., 2009). An example of the results of this morphing technique is shown in figure 7.7 in which the three identified formants are shifted to their expected frequencies.

This method tracks formants and warps the frequency space automatically, whereas Kain et al. (2007) perform these functions manually. A future implementation may use Kalman filters to reduce the noise inherent in trajectory tracking. Such an approach has shown significant improvements in formant tracking, especially for F_1 (Yan et al., 2007).

7.4 Intelligibility experiments with individual transformations in TORGOMorph

In order to gauge the intelligibility of our modifications, we designed a simple experiment in which human listeners attempt to identify words in sentence-level utterances under a number of acoustic scenarios. Sentences are either uttered by a speaker with dysarthria, modified from their original source acoustics, or manufactured by a text-to-speech synthesizer. Each participant is seated at a personal computer with a simple graphical user interface with a button which plays or replays the audio (up to 5 times), a text box in which to write responses, and a second button to submit that response. Audio is played over a pair of headphones. The participants are told to only transcribe the words with which they are reasonably confident and to ignore those that they cannot discern. They are also informed that the sentences are grammatically correct but not necessarily semantically coherent, and that there is no profanity. Each participant listens to 20 sentences selected at random with the constraints that at least two utterances are taken from each category of audio, described below, and that at least five utterances are also provided to another listener, in order to evaluate inter-annotator agreement. Participants are self-selected to have no extensive prior experience in speaking with individuals with dysarthria, in order to reflect the general population. Although dysarthric utterances are likely to be contextualized within meaningful conversations in real-world situations, such pragmatic aspects of discourse are not considered here in order to concentrate on acoustic effects alone. No cues as to the topic or semantic context of the sentences is given, as there is no evidence that such aids to comprehension affect intelligibility (Hustad and Beukelman, 2002). In this study we use sentence-level utterances uttered by speaker M04 from the TORGO database.

Baseline performance is measured on the original dysarthric speech. Two other systems are used for reference:

Synthetic Word sequences are produced by the Cepstral commercial text-to-speech system using the U.S. English voice ‘David’. This system is based on Festival in almost every

respect, including its use of linguistic pre-processing (e.g., part-of-speech tagging) and rule-based generation (Taylor, Black, and Caley, 1998). This approach has the advantage that every aspect of the synthesized speech (e.g., the word sequence) can be controlled although here, as in practice, synthesized speech will not mimic the user's own acoustic patterns, and will often sound more 'mechanical' due to artificial prosody (Black and Lenzo, 2007).

GMM This system uses the Gaussian mixture mapping type of modification suggested by Toda, Black, and Tokuda (2005) and Kain et al. (2007) and discussed in section 7.2.3. Here, we use the FestVox implementation of this algorithm, which includes F0 extraction, some phonological knowledge (Toth and Black, 2005), and a method for resynthesis. Parameters for this model are trained by the FestVox system using a standard expectation-maximization approach (Reynolds and Rose, 1995) with 24^{th} order cepstral coefficients and 4 Gaussian components. The training set consists of all vowels uttered by speaker M04 and their synthetic realizations produced by the method above.

Performance is evaluated on the three other acoustic transformations, namely those described in sections 7.3.2, 7.3.3, and 7.3.4 above. Tables 7.1 and 7.2 respectively show the percentage of words and phonemes correctly identified by each listener relative to the expected word sequence under each acoustic condition. In each case, annotator transcriptions were aligned with the 'true' or expected sequences using the Levenshtein algorithm described in section 7.3. Plural forms of singular words, for example, are considered incorrect in word alignment although one obvious spelling mistake (i.e., 'skilfully') is corrected before evaluation. Words are split into component phonemes according to the CMU dictionary, with words having multiple pronunciations given the first decomposition therein.

In these experiments there is not enough data from which to make definitive claims of statistical significance, but it is clear that the purely synthetic speech has a far greater intelligibility than other approaches, more than doubling the average accuracy of the TORGOMorph modifications. The GMM transformation method proposed by Kain et al. (2007) gave very poor

performance, although our experiments are distinguished from theirs in that our formant traces are detected automatically, rather than by hand. The relative success of the synthetic approach is not an argument against the type of modifications proposed here and by Kain et al. (2007), since our aim is to avoid the use of impersonal and invariant utterances. Indeed, future study in this area should incorporate subjective measures of 'naturalness'. Further uses of acoustic modifications not attainable by text-to-speech synthesis are discussed in section 7.5.

In all cases, the splicing technique of removing accidentally inserted phonemes and inserting missing ones gives the highest intelligibility relative to all modifications in TORGOMorph and the Gaussian mixture mapping method. Although more study is required, this result emphasizes the importance of a lexically correct phoneme sequence. In the word-recognition experiment, there were an average of 5.2 substitution errors per sentence in the unmodified dysarthric speech against 2.75 in the synthetic speech. There were also 2.6 substitution errors on average per sentence for the speech modified in frequency, but 3.1 deletion errors, on average, against 0.24 in synthetic speech. No correlation was found between the 'loudness' of the speech (determined by the overall energy in the sonorants) and intelligibility results, although this might change with the acquisition of more data. Neel (2009), for instance, found that loud or amplified speech from individuals with Parkinson's disease was more intelligible to human listeners than quieter speech.

Our results are comparable in many respects to the experiments of Kain et al. (2007), although they only looked at simple consonant-vowel-consonant stimuli and had 64 listeners annotate the center phoneme. Their results showed an average of 92% correct synthetic vowel recognition (compared with 94.2% phoneme recognition in table 7.2) and 48% correct dysarthric vowel recognition (compared with 52.9% in table 7.2). Our results, however, show that modified timing and modified frequencies do not actually benefit intelligibility in either the word or phoneme cases whereas slight improvement in vowel recognition, up to 54%, with modified durations and formants. This disparity may in part be due to the fact that our stimuli are much more complex (quicker sentences do not necessarily improve intelligibility).

Listener	Original	GMM	Synthetic	Splice	Timing	Frequency
L01	22.1	15.6	82.0	40.2	34.7	35.2
L02	27.8	12.2	75.5	44.9	39.4	33.8
L03	38.3	14.8	76.3	37.5	12.9	21.4
L04	24.7	10.8	72.1	32.6	22.2	18.4
Average	28.2	13.6	76.5	38.8	27.3	27.2

Table 7.1: Percentage of words correctly identified by each listener relative to the expected word sequence under each acoustic condition.

Listener	Original	GMM	Synthetic	Splice	Timing	Frequency
L01	52.0	43.1	98.2	64.7	47.8	55.1
L02	57.8	38.2	92.9	68.9	50.6	53.3
L03	50.1	41.4	96.8	57.1	30.7	46.7
L04	51.6	33.8	88.7	51.9	43.2	45.0
Average	52.9	39.1	94.2	60.7	43.1	50.0

Table 7.2: Percentage of phonemes correctly identified by each listener relative to the expected word sequence under each acoustic condition.

7.5 Discussion

Tolba and Torgoman (2009) claimed that significant improvements in recognition of dysarthric speech are attainable by modifying formants F1 and F2 to be more similar to expected values. In that study, formants were identified using standard LPC-based techniques, although no information is provided as to how these formants were modified nor how their targets were determined. However, they claimed that modified dysarthric speech results in 'recognition rates' (by which they presumably meant word-accuracy) of 71.4% in the HMM-based HTK ASR system, as compared with 28% on the unmodified dysarthric speech from 7 individuals. The results in section 7.4 show that human listeners are more likely to correctly identify utterances in which phoneme insertion and deletion errors are corrected than those in which formant frequencies are adjusted. Therefore, one might hypothesize that such pre-processing might provide even greater gains than those reported by Tolba and Torgoman (2009). Ongoing work ought to confirm or deny this claim.

A prototypical client-based application based on this research for unrestricted speech transformation of novel sentences is currently in development. Such work would involve improving factors such as accuracy and accessibility for individuals whose neuro-motor disabilities limit the use of modern ASR, and for whom alternative interaction modalities are insufficient. This application is being developed for Google's Android platform under the assumption that it will be used in a mobile device embeddable within a wheelchair. If word-prediction is to be incorporated, the predicted continuations of uttered sentence fragments can be synthesized without requiring acoustic input, as in section 7.2.3.

In practice, the modifications presented above will have to be based on automatically-generated annotations of the source audio. This is especially important to the 'splicing' module in which word-identification is crucial. There are a number of techniques that can be exercised in this area. Czyzewski, Kaczmarek, and Kostek (2003) apply both a variety of neural net-

works and 'rough sets'[8] to the task of classifying segments of speech according to the presence of stop-gaps, vowel prolongations, and incorrect syllable repetitions. In each case, input includes source waveforms and detected formant frequencies. They found that stop-gaps and vowel prolongations could be detected with up to 97.2% accuracy and that vowel repetitions could be detected with up to 90% accuracy using the rough set method. Accuracy was similar although slightly lower using traditional neural networks (Czyzewski, Kaczmarek, and Kostek, 2003). These results appear generally invariant even under frequency modifications to the source speech. Nakatani (1993), Plauché and Shriberg (2007), and Arbisi-Kelm (2010), for example, also suggest that disfluent repetitions can be identified reliably through the use of pitch, duration, and pause detection (with precision up to 93% (Nakatani, 1993)). If more traditional models of speech recognition such as HMMs are to be employed in order to identify vowels, the probabilities that they generate across hypothesized words might be used to weight the manner in which acoustic transformations are made, although the exact mechanism of this weighting is yet to be determined. For example, modifications to the formant frequencies might only be performed if the identity of the phoneme involved can be estimated above some minimum threshold of likelihood.

[8]Rough sets in this context are sets of items whose memberships in subsets are 'fuzzy' according to the indistinguishability of their parameter sets (Pawlak, 1982).

Chapter 8

Concluding remarks

The purpose of this thesis was to determine whether *classification systems built using empirical and theoretical models of speech production can significantly improve recognition accuracy for dysarthric speakers*. This has been answered affirmatively through the experiments performed in chapters 5 and 6 that are based on the new data described in chapter 4. Despite the advances described in this dissertation, there yet remains much work to follow. This chapter summarizes contributions made by this work in section 8.1, projects future work that follows naturally from this thesis in section 8.2 and concludes with a closing thought in section 8.3.

8.1 Summary of contributions

A number of contributions have originated from the research described in this thesis. The most significant of these are the following:

- The design and construction of TORGO, one of the first and most extensive databases of dysarthric articulation (Rudzicz et al., 2008; Rudzicz, Namasivayam, and Wolff, 2010), described in chapter 4.

- The construction of various generative models (HMM, DBN) and discriminative models (CRF, neural network,SVM) that recognize discrete articulatory features and replace

standard acoustic models in ASR systems. This work showed that the performance of generative methods with articulatory knowledge is similar to that of advanced discriminative methods (Rudzicz, 2009a; Rudzicz, 2009b; Rudzicz, 2010a), as described in chapter 5.

- The introduction of an acoustic-to-articulatory inversion system based on Hammerstein systems that significantly outperforms the state-of-the art on the task of inferring the positions of task-dynamics parameters, described in section 6.2.

- The demonstration in section 6.3 that articulatory data removes ambiguity from the acoustic data in the TORGO database, even if the mutual information between these spaces is not as great for dysarthric speakers as for non-dysarthric speakers (Rudzicz, 2010d; Rudzicz, 2010c).

- The construction of an ASR reranking system that reduces word-error rate significantly over acoustic-only baselines by incorporating information present in task-dynamics modelling, described in section 6.4. This represents the first incorporation of the long-term dynamics of task-dynamics into ASR (Rudzicz, 2010b).

This research has also made a number of secondary contributions, including a comparison of adaptive and dependent speaker modelling in HMMs for speakers with dysarthria (section 5.2.2), an investigation into a noisy-channel model of dysarthric speech (section 6.3.2), and the use of principal differential analysis in the recognition of non-dysarthric speech (section 6.5). We have also designed a number of automatic transformations to improve the intelligibility of dysarthric speech, as described in chapter 7. As previously discussed, it is possible for automatic methods to outperform human rates of recognition on disordered speech (Jayaram and Abdelhamied, 1995), but it is not yet clear whether there is some fundamental limit to the achievable accuracy of automatic methods. We are currently performing intelligibility assessments with naïve human listeners which may help to answer this question.

In general, this dissertation represents a novel confluence of disparate disciplines and related research areas within speech recognition. In our opinion, there has not hitherto been sufficient interaction between engineering (signal processing and artificial intelligence, specifically) and speech science (i.e., speech language pathology) despite a tremendous amount of knowledge and perspective that each could provide the other. This dissertation provides a biological basis for the models used in classification, and applies advanced machine learning to ecologically interesting data.

8.2 Future work

This section outlines three research avenues for exploration that the work described in this dissertation has opened.

8.2.1 Dysarthria in task-dynamics

Several high-level models of speech production have been discussed throughout this thesis, including task-dynamics and DIVA. A commonality among these models is that they represent the synthetic behaviour of an idealized average human speaker, which avoids certain biological realities. Future work should be based upon the study of the dysarthric data collected for TORGO within the framework of task-dynamics. Here, articulatory behaviour of six or seven of the dysarthric speakers for whom there is enough data should be compared against the behaviour of control speakers by applying and extending methods introduced in section 6.5 that learn the parameters of second-order differential equations with principal differential analysis. In practice, however, there are several other aspects of task-dynamics that are not represented by the fundamental underlying spring-mass equation. For each speaker and each linguistic unit (i.e., syllable), several parameters can be derived. First, the *damping* and *stiffness* coefficients of tract variables can be derived using PDA as described earlier for each tract variable available in the data, namely tongue tip constriction degrees and locations, tongue

body constriction degrees and locations, lip protrusion and aperture, and lower tooth height. The *target* (i.e., equilibrium) position (z^* in equation 6.1 of section 6.1) of the relevant gesture can be derived from data by finding zeros in second-order regression analysis. More subtly, the *articulatory weights* set the effectiveness of the associated tract variables in the production of a gesture. A high articulatory weight reduces the amount of motion associated with a given articulator as if it is 'heavier'. This parameter is similar to the *mass* coefficient of the spring-mass equation and may be derived through a combination of examining the average amplitude of articulatory motion during production along with the relative entropy of the acoustics, given the articulation (see section 6.3). Other parameters in `TADA` may be carried over from the defaults, such as α which specifies the strength of a gesture in the presence of other gestures on the same tract variable. If $\alpha = 0$, for example, the associated gesture participates in additive rather than averaging blending (Nam and Goldstein, 2006).

Additionally, the geometry of each vocal tract in the data should be measured, along with the 'natural attractor' position (Saltzman and Munhall, 1989), which is normally associated with the schwa. These parameters modulate the behaviour of the `TADA` system for task-dynamics, which is utilized in the speech recognition system described in section 6.4. By adapting the parameters of this system and repeating experiments in section 6.4, future work could measure the usefulness of task-dynamics in speech recognition for dysarthric speakers. This system could then be compared against further baselines, including ergodic hidden Markov Models designed to capture involuntary repetition, which is more common among dysarthric speakers (Sharma and Hasegawa-Johnson, 2010b).

8.2.2 Discriminative training of language models

Discriminative training based on generalized probabilistic descent and the minimum classification error criterion can overcome some of the limitations of maximum likelihood estimation with acoustically confusable speech signals by increasing the discriminative power of $P(W)$ (Gopalakrishnan et al., 1991; Katagiri, Juang, and Lee, 1998). The score applied to a hypothe-

sized word sequence W_i is usually some perturbation of

$$g(W_i,X) = \alpha \log P(X|W_i) + (1-\alpha)\log P(W_i) \tag{8.1}$$

for some parameter α. Classification errors depend on the relative scores of the top N hypotheses, $g(W_i,X)$ for $i = 1..N$, where the correct hypothesis is W_C. The combined score of the best incorrect hypotheses of this list is modelled by the anti-discriminant function

$$G(X,W_1,...,W_N,c) = \log\left[\frac{1}{N}\sum_{h=1,h\neq c}^{N}\exp\left(g(W_h,X)\eta\right)\right]^{\eta}, \tag{8.2}$$

where the weight given to the best of the N alternatives relative to the worst increases with parameter η. The misclassification function is then just

$$d(X) = -g(W_c,X) + G_c(X,W_1,...,W_N,c) \tag{8.3}$$

The idea is for $d(X)$ to be negative with all correct classifications. The concept of 'loss' incurred by misclassification is neatly encapsulated with

$$l\left(d(X)\right) = \frac{1}{1+\exp(-\rho d(X)+\theta)} \tag{8.4}$$

whose derivative is also continuous and in the range $[0..1]$. Using generalized probabilistic descent, the language model Γ can be updated iteratively by a factor of ε

$$\Gamma_{t+1} = \Gamma_t - \varepsilon \bigtriangledown l\left(d(X)\right) \tag{8.5}$$

where

$$\begin{aligned} \bigtriangledown l &= \frac{\delta l_i}{\delta d_i}\frac{\delta d(X_i)}{\delta \Gamma} \\ &= \rho l(d_i)(1-l(d_i))\frac{\delta d(X_i)}{\delta \Gamma}. \end{aligned} \tag{8.6}$$

If $p_{x,y} = \log P(w_y|w_x)$ is a bigram parameterizing Γ, and $n(W,w_x,w_y)$ is the number of times $w_x w_y$ appears in W, then

$$\frac{\delta d(X_i)}{\delta p_{x,y}} = \left[-n(W_c,w_x,w_y) + \sum_{r=1}^{N} C_r n(W_r,w_x,w_y)\right] \tag{8.7}$$

where

$$C_r = \frac{\exp(g(W_r, X_i)\eta)}{\sum_{j=1}^{N} \exp(g(W_j, X_i)\eta)}. \tag{8.8}$$

Intuitively, the likelihoods of bigrams in the correct string but not in competing hypotheses is increased, and diminished for those only in the chief competitors. Using this approach, Na et al. were able to improve accuracy on isolated Korean digits from 92.7% to 94.5% using a training set of only 50 utterances (Na, Rheem, and Ann, 1994), eliminating all substitution errors between the most confusable pairs of words. Similarly, Kuo et al. (2002) were able to reduce error by as much as 15% relative to an MLE-trained language model. Discriminative training has also been applied to training acoustic (Sandness, 2000) models using essentially the same algorithm and discriminant functions. Discriminative training of language models does not yet appear to have been applied to dysarthric speech. We can therefore include some of this work in augmenting language models used in our previous experiments.

8.2.3 Multimodal interaction for individuals with special needs

Despite the relative speed and ease with which information can be conveyed verbally, especially by speakers with neuromotor disorders (see section 2.3.3), certain kinds of information may best be communicated by other means. Deixis, for example, is a phenomenon in which contextual information is required to resolve semantic ambiguity in an utterance. To fully understand the phrase *put that there*, for example, requires that the demonstrative pronouns *that* and *there* resolve to a specific object and a specific location in the world, respectively. In this example and others like it, arm gestures can be used naturally to provide information otherwise absent in the referring words, as in the seminal work of Bolt (1980) and in similar research (Cohen et al., 1989; Cheyer and Julia, 1998). Multimodal technology has since diversified considerably as new hardware platforms permit profuse permutations of physical interaction, from portable devices to large-screen environments (Kettebekov et al., 2002; Sharma et al., 2003). There is also a practical benefit to co-ordinating multiple concurrent streams, with error suppression in certain contexts in excess of 40% relative to unimodal counterparts (Oviatt,

2003; Tamura, Iwano, and Furui, 2004; Motlcek, Burget, and Cernock, 2005).

The research described in this dissertation and its predecessor (Rudzicz, 2006) has involved the co-ordination of various technologies that are commonly combined in multimodal human-computer interaction, especially video and audio. A natural progression of this work, therefore, would use these technologies in an applied domain to improve the expressive abilities of end-users. For example, the use of video to determine head pose, in addition to lip-reading, has been successfully used as a method of hands-free interaction with graphical user interfaces (Karpov et al., 2004). Further research should compare the EMA and video data collected in TORGO to determine whether observable features in the latter are suitably informative substitutes for those in the former. Speech recognition is also being developed as a mode of communication with automated emergency response systems in home environments in which video tracking of humans is already being performed for people with special needs (Hamill et al., 2009). The continued development of word-prediction and speech transformation described respectively in section 2.3.3 and chapter 7 may eventually be inclined towards potential commercial products.

8.3 Closing thought

Our species distinguishes itself by its exceptional capacity to understand and to overcome limitations presented by nature. Our technology has given us abilities that we once thought impossible and our science has raised questions that were once unfathomable. Future technological responses to the limitations of communication should continue to expose latent abilities both individually and collectively.

We shape our tools and thereafter our tools shape us.

– Marshall McLuhan (1964)

Appendix A

Articulatory contrasts

Front-back vowel	knew/knee	pat/pot	him/hum	shoot/sheet	beet/boot
	geese/goose	feed/food	air/are	chop/chap	fill/full
High-low vowel	knew/know	knew/gnaw	him/hem	him/ham	shoot/shot
	geese/gas	geese/guess	pit/pet	pit/pat	feet/fat
	heat/hate	had/hid			
Vowel duration	beat/bit	slip/sleep	leak/lick	knot/nut	read/rid
	ship/sheep	feet/fit	lip/leap	ease/is	reap/rip
Voicing, initial consonants	pat/bat	bad/pad	pit/bit	sip/zip	coat/goat
	dug/tug	cash/gash	tile/dial	bunch/punch	
Voicing, final consonants	feet/feed	bad/bat	leak/league	knot/nod	write/ride
	side/sight	coat/code	dug/duck	ate/aid	at/add
Alveolar-palatal	sip/ship	shoot/suit	shy/sigh	sell/shell	sin/shin
	sew/show	see/she	sheet/seat		
Consonant place	bug/dug	tile/pile	cake/take	meat/neat	bill/dill
	bill/gill	ache/ape	ache/ate	lip/lit	

Table A.1: Articulatory contrasts, after Kent et al. (1989).

Other fricative	sheet/feet	sigh/thigh	hill/fill	hand/sand	sew/foe
	see/he	nice/knife	hat/fat	sell/fell	feet/heat
	hat/that	hold/fold	hail/sail	harm/farm	seed/feed
Fricative-affricate	chair/share	wish/witch	much/mush	ship/chip	chop/shop
	cash/catch	sheer/cheer	hash/hatch	harm/charm	
Stop-affricate	chair/tear	much/mut	chop/top	witch/wit	much/muck
Stop-nasal	beat/meat	knot/dot	side/sign	nice/dice	steak/snake
	bill/mill	dock/mock	dock/knock	bunch/munch	tile/mile
Initial glottal-null	air/hair	ate/hate	at/hat	hand/and	hold/old
	heat/eat	hash/ash	harm/arm	had/add	hail/ail
Initial consonant-null	air/fair	ate/fate	at/at	sin/in	sheet/eat
	chair/air	spit/it	blend/end	ease/peas	ease/cheese
	sink/ink	cake/ache	rise/eyes	row/ow	
Final consonant-null	feed/fee	side/sigh	blow/bloat	fork/four	rake/ray
	leak/lee	meat/me	bunch/bun	seed/see	
Initial cluster-initial singleton	slip/sip	slip/lip	spit/pit	spit/sit	blend/bend
	blend/lend	sticks/six	sticks/ticks	steak/take	steak/sake
	blow/low	blow/bow			
Final cluster-final singleton	sticks/stick	rock/rocks	seed/seeds	sink/sing	cake/cakes
	meat/meats	fork/forks	rake/rakes	leak/leaks	ache/aches
	wax/wack	docks/dock			
/r/-/l/	read/lead	write/light	leak/reek	rock/lock	rake/lake
	lip/rip	reap/leap	rise/lies	row/low	racks/lax
/r/-/w/	read/weed	write/white	rich/witch	rock/walk	reap/weep
	rise/wise	row/woe	racks/wax		

Table A.2: Articulatory contrasts, after Kent et al. (1989) *continued*.

Appendix B

Frenchay Assessment in TORGO

Category	Test	Observation	Males $\mu(\sigma)$	Females $\mu(\sigma)$	All $\mu(\sigma)$
Reflex	Cough	Presence of cough during eating and drinking.	6(2.45)	7.3(0.96)	6.6(1.85)
	Swallow	Speed and ease of swallowing liquid.	7(2.0)	8(0.0)	7.5(1.41)
	Dribble	Presence of drool generally.	6.5(2.38)	7.5(1.0)	7(1.77)
Lips	At rest	Asymetry of lips during rest.	6.3(2.36)	8(0.0)	7.1(1.81)
	Spread	Distortion during smile.	6(2.31)	8(0.0)	7(1.85)
	Seal	Ability to maintain pressure at lips over time.	3.3(3.4)	7(2)	5.1(3.27)
	Alternate	Variability in repetitions of *"oo ee"*.	3.8(2.87)	7(2)	5.4((2.88)
	In speech	Excessive briskness or weakness during regular speech.	4.3(1.89)	6.5(1.91)	5.4(2.13)

Table B.1: Frenchay Dysarthria Assessment dimensions (Enderby, 1983), each on a scale of 0 (no function) to 8 (normal function).

Category	Test	Observation	Males $\mu(\sigma)$	Females $\mu(\sigma)$	All $\mu(\sigma)$
Respiration	At rest	Ability to control breathing during rest.	4(2.71)	8(0.0)	6(2.78)
	In speech	Breaks in fluency caused by poor respiratory control.	4(2)	6.5(3)	5.3(2.71)
Jaw	At rest	Hanging open of jaw at rest.	7(1.15)	8(0.0)	7.5(0.93)
	In speech	Fixed position or sudden jerks of jaw during speech.	5.8(2.63)	6.3(2.36)	6.1(2.42)
Velum	Fluids	Liquid passing through velum while eating.	7(2.0)	8(0.0)	7.5(1.41)
	Maintenance	Elevation of palate in repetitions of *"ah ah ah"*.	5.8(2.06)	7.5(1.0)	6.6(1.77)
	In speech	Hypernasality or imbalanced nasal resonance in speech.	6.3(2.36)	6(2.83)	6.1(2.42)
Laryngeal	Time	Sustainability of vowels in time.	5.3(2.5)	7.5(1.0)	6.4(2.13)
	Pitch	Ability to sing a scale of distinct notes.	2(2.16)	5.3(2.5)	3.6(2.77)
	Volume	Ability to control volume of voice.	3.5(3.11)	4.8(3.2)	4.1(3.0)
	In speech	Phonation, volume, and pitch in conversational speech.	3.3(2.87)	6(2.83)	4.6(3.02)
Intelligibility	Words	Interpretability of 10 isolated spoken words from a closed set.	4(2.94)	4.5(2.52)	4.3(2.55)
	Sentences	Interpretability of 10 spoken sentences from a closed set.	3.5(3.32)	5.3(3.4)	4.4(3.25)
	Conversation	General distortion or decipherability of speech in casual conversation.	4.5(2.38)	6.5(1.91)	5.5(2.27)

Table B.2: Frenchay Dysarthria Assessment dimensions (Enderby, 1983), each on a scale of 0 (no function) to 8 (normal function) *continued*.

Category	Test	Observation	Males $\mu(\sigma)$	Females $\mu(\sigma)$	All $\mu(\sigma)$
Tongue	At rest	Deviation of tongue to one side, or involuntary movement.	5.5(2.08)	5.5(1.73)	5.5(1.77)
	Protrustion	Variability, irregularity, or tremor during repeated tongue protrusion and retraction.	3.8(3.1)	5.3(1.5)	4.5(2.39)
	Elevation	Laboriousness and speed of repeated motion of tongue tip towards nose and chin.	3.3(3.2)	4.3(1.71)	3.7(2.43)
	Lateral	Laboriousness and speed of repeated motion of tongue tip from side to side.	3.8(3.1)	3.5(1.91)	3.6(2.39)
	Alternate	Deterioration or variability in repetitions of phrase *"ka la"*.	4(2.71)	5.3(1.91)	4.9(2.9)
	In speech	Correctness of articulation points and laboriousness of tongue motion during speech generally.	4(2.71)	6(2.83)	5(2.78)

Table B.3: Frenchay Dysarthria Assessment dimensions (Enderby, 1983), each on a scale of 0 (no function) to 8 (normal function) *concluded*.

Appendix C

Formant targets in synthesis

	F1	F2	F3	BW1	BW2	BW3
el	450	800	2850	65	60	80
en	200	900	2100	120	60	70
em	200	1600	2700	120	70	110
l	330	1050	2800	50	100	280
n	480	1400	2700	40	300	260
m	480	1050	2100	40	175	120
ng	480	1460	2050	160	150	100
r	330	1060	1380	70	100	120
w	285	610	2150	50	80	60

Table C.1: Formant target frequencies (F1–3) and bandwidths (BW1–3) in Hz for synthesis in sonorant consonants for a male speaker of English, after Allen et al. (1987).

	F1	**F2**	**F3**	**BW1**	**BW2**	**BW3**
aa	700	1220	2600	130	70	160
ae	620	1660	2430	70	130	300
ah	620	1220	2550	80	50	140
ao	600	990	2570	90	100	80
aw	640	1230	2550	80	70	110
ax	550	1260	2470	80	50	140
axr	680	1170	2380	60	60	110
ay	660	1200	2550	100	120	200
eh	530	1680	2500	60	90	200
er	470	1270	1540	100	60	110
ey	480	1720	2520	70	100	200
ih	400	1800	2670	50	100	140
ix	420	1680	2520	50	100	140
iy	310	2200	2960	50	200	400
ow	540	1100	2300	80	70	70
oy	550	960	2400	80	120	160
uh	450	1100	2350	80	100	80
uw	350	1250	2200	65	110	140

Table C.2: Formant target frequencies (F1–3) and bandwidths (BW1–3) in Hz for synthesis in vowels for a male speaker of English, after Allen et al. (1987).

Appendix D

Electrical synchronization in TORGO

Audio in the TORGO database is recorded simultaneously by a head-worn microphone connected to the AG500 EMA machine and by a directional microphone connected to the system that randomizes and presents prompts graphically to the participant. An electronic circuit was devised to communicate between these two systems. The AG500 system includes a breakout box called the Sybox-Opto4, shown in figure D.1, which allows access to some of the system's low-level signals. Specifically, there are 4 DB9 RS-232 connectors on the front of this box, each of which operates at +5V and 50 mA. The circuitry of these connectors is shown in figure D.3. On each of these connectors, pin 4 is the *sweep* signal, which is in the active state as long as speech is being recorded by the EMA system and has a timing precision of < 50 ns.

A laptop is required to connect to the presentation monitor, to manage prompt lists, and to record directional audio for each prompt. The *sweep* signal from the Sybox-Opto4 is used to trigger audio recording on this laptop synchronously. The presentation software used to prompt the speaker includes a threaded component which monitors the system's serial bus for a binary change of the *sweep* signal which indicates either the start or end of recording, depending on its direction. Since most laptop RS-232 connectors[1] operate at $\pm$ 12V, an amplifier circuit had to be constructed to bring the *sweep* signal up to a voltage that could be read by the second

[1] Some laptops include embedded serial connectors; however, often this has to be replicated via a USB/RS-232 dongle (e.g., a KeySpan serial adapter) and the laptop's USB system.

Figure D.1: The Sybox-Opto4 synchronization device for the AG500. Image taken from documentation from Carstens Medizinelektronik GmbH, Lenglern, Germany.

laptop, as shown in figure D.2. This system was later abandoned because acoustic alignment with cross-correlation would be performed regardless, but this circuitry can be replicated in situations where no such software method is available.

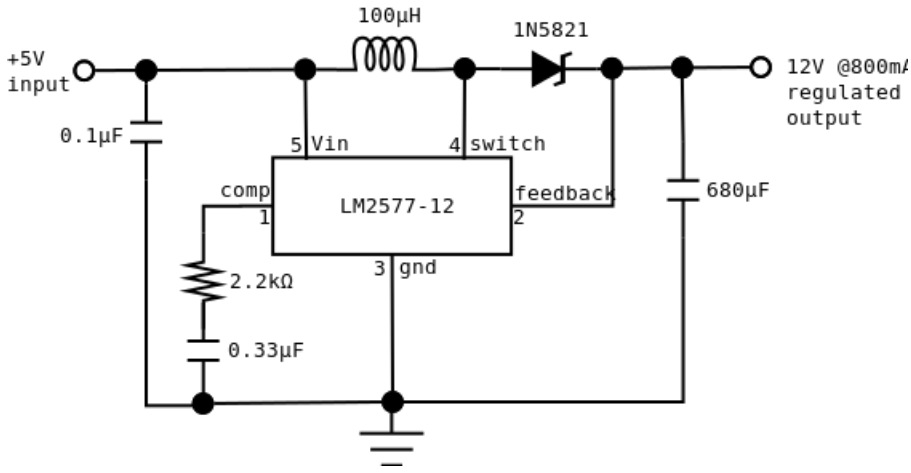


Figure D.2: Circuitry to amplify the 5V *sweep* signal from the AG500 to the 12V required by the PC serial bus. The component '1N5821' is a Schottky barrier diode which converts alternating current to direct current and the component 'LM2577-12' is a switching regulator.

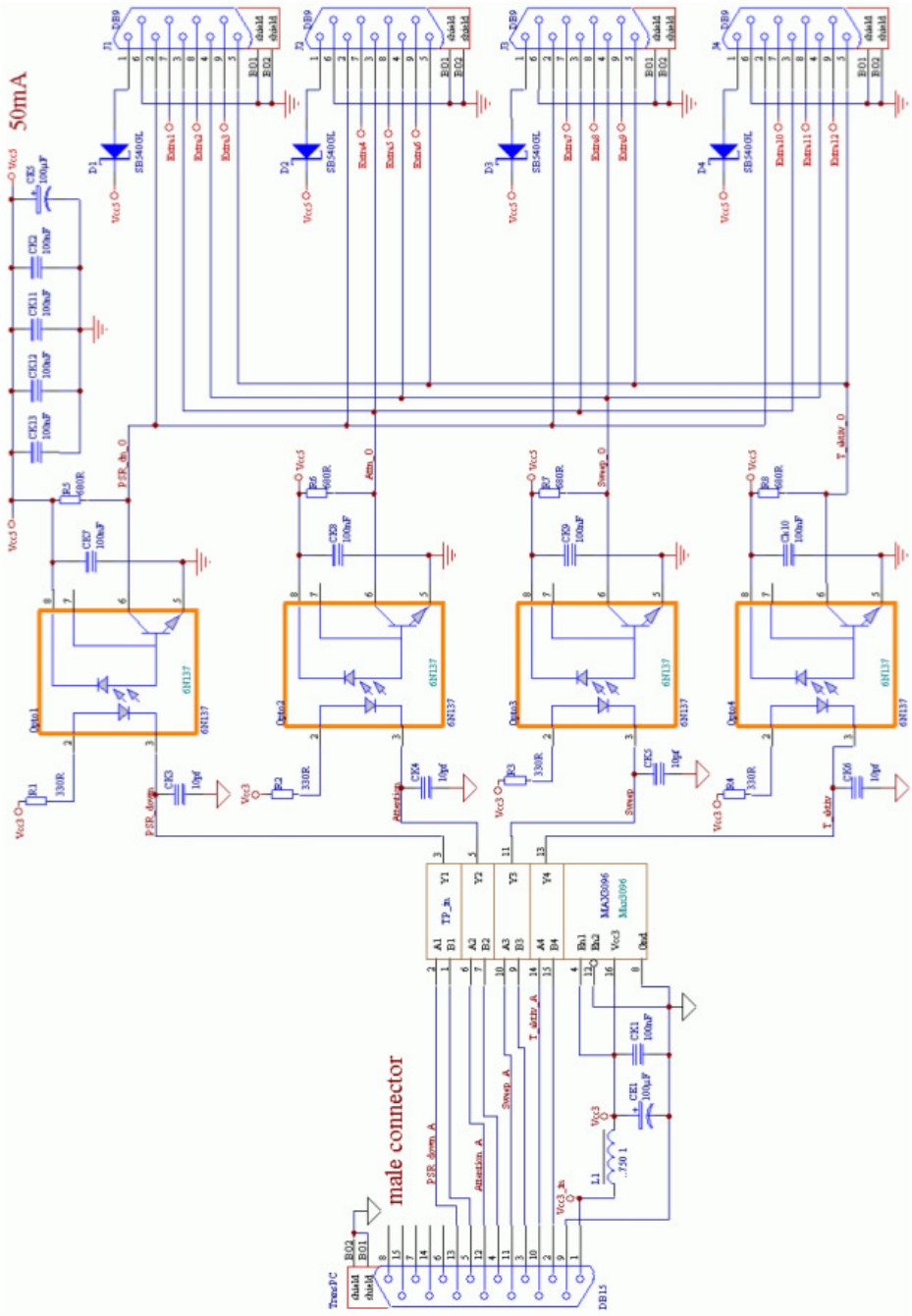


Figure D.3: The internal wiring of the Sybox-Opto4 synchronization device. Image taken from documentation from Carstens Medizinelektronik GmbH, Lenglern, Germany.

References

Aarabi, Parham and Guangji Shi. 2004. Phase-based dual-microphone robust speech enhancement. *IEEE Transactions on Systems, Man, and Cybernetics Part B*, 34(4):1763–1773, August.

Abbs, James H., John W. Folkins, and Murali Sivarajan. 1976. Motor Impairment following Blockade of the Infraorbital Nerve: Implications for the Use of Anesthetization Techniques in Speech Research. *Journal of Speech and Hearing Research*, 19(1):19–35.

Adjoudani, A and C. Benoit. 1995. On the integration of auditory and visual parameters in an hmm-based asr. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pages 461–472, September.

Ahadi-Sarkani, S.M. 1996. *Bayesian and predictive techniques for speaker adaptation*. Ph.D. thesis, Cambridge University.

Allen, Jonathan, M. Sharon Hunnicutt, Dennis H. Klatt, Robert C. Armstrong, and David B. Pisoni. 1987. *From text to speech: the MITalk system*. Cambridge University Press, New York, NY, USA.

Alm, Norman, John L. Arnott, and Alan F. Newell. 1992. Prediction and conversational momentum in an augmentative communication system. *Communications of the ACM*, 35(5):46–57.

Ananthakrishnan, G., Daniel Neiberg, and Olov Engwall. 2009. In search of non-uniqueness in the acoustic-to-articulatory mapping. In *Proceedings of Interspeech 2009*, Brighton UK.

Arbisi-Kelm, Timothy. 2010. Intonation structure and disfluency detection in stuttering. *Laboratory Phonology 10*, 4:405–432.

Aschbacher, Ernst and Markus Rupp. 2005. Robustness analysis of a gradient identification method for a nonlinear Wiener system. In *Proceedings of the 13th Statistical Signal Processing Workshop (SSP)*, Bordeaux, France, July.

Augmentative Communication Incorporated (ACI). 2007. Section 3: Clinical Aspects of AAC Devices.

Baddeley, Alan, Susan Gathercole, and Costanza Papagno. 1998. The phonological loop as a language learning device. *Psychological Review*, 105(1):158–173, January.

Baddeley, Alan and Barbara Wilson. 1985. Phonological coding and short-term memory in patients without speech. *Journal of Memory and Language*, 24(4):490 – 502.

Bahr, Ruth H. 2005. Differential diagnosis of severe speech disorders using speech gestures. *Topics in Language Disorders. Clinical Perspectives on Speech Sound Disorders*, 25(3):254–265.

Banno, Hideki, Hiroaki Hata, Masanori Morise, Toru Takahashi, Toshio Irino, and Hideki Kawahara. 2007. Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation. *Acoustical Science and Technology*, 28(3):140–146.

Barlow, H.B. 1989. Unsupervised learning. *Neural Computation*, 1(3):295–311.

Beaman, C. Philip. 2007. Modern cognition in the absence of working memory: Does the working memory account of Neandertal cognition work? *Journal of Human Evolution*, pages 702–706.

Bennett, Janice W., Pascal van Lieshout, and Catriona M. Steele. 2007. Tongue control for speech and swallowing in healthy younger and older subjects. *International Journal of Orofacial Myology*, 33:5–18.

Black, Alan W. and Kevin A. Lenzo. 2004. Multilingual text-to-speech synthesis. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*.

Black, Alan W. and Kevin A. Lenzo. 2007. Building synthetic voices.

Blanz, Volker and Thomas Vetter. 2003. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 25(9), September.

Bodt, Marc S. De, Maria E. Hernandez-Diaz Huici, and Paul H. Van De Heyning. 2002. Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, 35:283–292.

Boersma, Paul. 1998. *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. Ph.D. thesis, Universiteit van Amsterdam, September.

Boersma, Paul. 1999. Optimality-theoretic learning in the PRAAT program . In *Proceedings of the Institute of Phonetic Sciences*, volume 23, pages 17–35.

Bolt, Richard A. 1980. "put-that-there": Voice and gesture at the graphics interface. In *SIGGRAPH 80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA. ACM Press.

Bouguet, Jean-Yves. 1999. *Visual methods for three-dimensional modeling*. Ph.D. thesis, California Institute of Technology, Pasadena, California.

Bouguet, Jean-Yves. 2010. Camera Calibration Toolbox for Matlab.

Browman, Catherine P. and Louis M. Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252.

Butterworth, Stephen. 1930. On the theory of filter amplifiers. *Experimental Wireless and the Wireless Engineer*,

7:536–541.

Cambridge. 2007. Htk speech recognition toolkit, *http://htk.eng.cam.ac.uk/*. WWW, March.

Campbell, Jonathan M., Stephen K. Bell, and Lori K. Keith. 2001. Concurrent Validity of the Peabody Picture Vocabulary Test-Third Edition As an Intelligence and Achievement Screener for Low SES African American Children. *Assessment*, 8(1):85–94.

Castleman, Kenneth R. 1996. *Digital Image Processing*. Prentice-Hall, Englewood Cliffs NJ.

Chaudhari, Upendra V. and Michael Picheny. 2009. Articulatory feature detection with support vector machines for integration into asr and phone recognition. In *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 93–98, Merano Italy, November.

Chen, Fangxin and Aleksandar Kostov. 1997. Optimization of dysarthric speech recognition. In *Proceedings of the 19th Annual International Conference of the IEEE*, volume 4, pages 1436–1439. Engineering in Medicine and Biology society, November.

Chen, Helen and Kenneth N. Stevens. 2001. An acoustical study of the fricative /s/ in the speech of individuals with dysarthria. *Journal of Speech, Language, and Hearing Research*, 44:1300–1314, December.

Chesta, Christina, Olivier Siohan, and Chin-Hui Lee. 1999. Maximum a posteriori linear regression for hidden Markov model adaptation. In *Proceedings of EUROSPEECH'99*, pages 211–214.

Cheyer, Adam and Luc Julia. 1998. Multimodal maps: An agent-based approach. In *Multimodal Human-Computer Communication, Systems, Techniques, and Experiments*, pages 111–121, London, UK. Springer-Verlag.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

Chowdhury, Amit K. Roy and Rama Chellappa. 2003. Face reconstruction from monocular video using uncertainty analysis and a generic model. *Computer Vision and Image Understanding*, 91(1-2):188 – 213. Special Issue on Face Recognition.

Chu, Stephen M. and Thomas S. Huang. 2000. Bimodal speech recognition using coupled hidden Markov models. In *Proceedings of the IEEE International Conference on Spoken Language Processing*, pages 747–750, Beijing China.

Clear, Jeremy H. 1993. The British national corpus. In *The digital word: text-based computing in the humanities*. MIT Press, Cambridge, MA, USA, pages 163–187.

Clements, G. N. 1985. The geometry of phonological features. *Phonology Yearbook*, 2:225–252.

Cohen, P. R., M. Dalrymple, D. B. Moran, F. C. Pereira, and J. W. Sullivan. 1989. Synergistic use of direct manipulation and natural language. *SIGCHI Bull.*, 20(SI):227–233.

Coker, Cecil H. 1968. Speech synthesis with a parametric articulatory model. In *Proceedings of the Speech*

Symposium, Kyoto, Japan.

Coker, Cecil H. 1976. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, April.

Coleman, Colette and Lawrence Meyers. 1991. Computer recognition of the speech of adults with cerebral palsy and dysarthria. *Augmentative & Alternative Communication*, 7(1):34–42, March.

Constantinescu, Gabriella, Deborah Theodoros, Trevor Russell, Elizabeth Ward, Stephen Wilson, and Richard Wootton. 2010. Assessing disordered speech and voice in Parkinson's disease: a telerehabilitation application. *International Journal of Language & Communication Disorders*, 45(6):630–644, November.

Craig, Matthew, Pascal van Lieshout, and Willy Wong. 2007. Suitability of a UV-based video recording system for the analysis of small facial motions during speech. *Speech Communication*, 49(9):679–686, September.

Czyzewski, Andrzej, Andrzej Kaczmarek, and Bozena Kostek. 2003. Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems*, 21(2):143–171.

D'Ausilio, Alessandro, Friedemann Pulvermuller, Paola Salmas, Ilaria Bufalari, Chiara Begliomini, and Luciano Fadiga. 2009. The motor somatotopy of speech perception. *Current Biology*, 19(5):381–385, February.

Deller, J.R., J.H.L. Hansen, and J.G. Proakis. 2000. *Discrete-Time Processing of Speech Signals*. IEEE Press.

Deller, J.R. and R.K. Snider. 1990. 'quantized' hidden markov models for efficient recognition of cerebral palsy speech. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 2041–2044, New Orleans LA, May.

Deng, Jianping, M. Bouchard, and Tet Yeap. 2005. Speech Enhancement Using a Switching Kalman Filter with a Perceptual Post-Filter. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 1121–1124, 18-23,.

Deng, Li. 2000. Switching dynamic system models for speech articulation and acoustics. In *Proceedings of the IMA Workshop*, September.

Deng, Li. 2006. *Dynamic speech models: theory, algorithms, and applications*. Morgan & Claypool Publishers.

Deng, Li and D. Sun. 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, 85(5):2702–2719.

Dogil, Grzegorz and Jorg Mayer. 1998. Selective phonological impairment: a case of apraxia of speech. *Phonology*, 15(2).

Doyle, Philip C., Herbert A. Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, Charlene O'Neill, Marie-Claire Dylke, and Katherine Rolls. 1997. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development*, 34(3):309–316, July.

Duffy, Joseph R. 2005. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Mosby Inc.

Dupont, Stéphane and Juergen Luettin. 2000. Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on*, 2(3):141–151, September.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Enderby, Pamela M. 1983. *Frenchay Dysarthria Assessment*. College Hill Press.

Ephraim, Yariv and David Malah. 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, April.

Erdogan, Hakan, Ruhi Sarikaya, Stanley F. Chen, Yuqing Gao, and Michael Picheny. 2005. Using semantic analysis to improve speech recognition performance. *Computer Speech and Language*, 19:321–343.

Erler, K. and Li Deng. 1993. Hidden markov model representation of quantized articulatory features for speech recognition. *Computer Speech and Language*, 7(3):265–282.

Ewender, Thomas, Sarah Hoffmann, and Beat Pfister. 2009. Nearly Perfect Detection of Continuous F_0 Contour and Frame Classification for TTS Synthesis. In *Proceedings of INTERSPEECH 2009*, Brighton, UK.

Fazly, Afsaneh and Graeme Hirst. 2003. Testing the efficacy of part-of-speech information in word completion. In *Proceedings of the EACL 2003 Workshop on Language Modeling for Text Entry Methods*.

Felber, Philip. 2001. Speech recognition ; report of an isolated word experiment. Technical report, Illinois Institute of Technology, April.

Ferrier, Linda J., Howard C. Shane, Holly F. Ballard, Tyler Carpenter, and Anne Benoit. 1995. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative & Alternative Communication*, 11(3):165–175, January.

Frankel, Joe, Mirjam Wester, and Simon King. 2007. Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language*, 21:620–640.

Freund, Hans-Joachim, Marc Jeannerod, Mark Hallett, and Ramón Leiguarda. 2005. *Higher-order motor disorders: From neuroanatomy and neurobiology to clinical neurology*. Oxford University Press.

Friedland, Bernard. 2005. *Control System Design: An Introduction to State-Space Methods*. Dover.

Fujimura, Osamu. 1986. Relative invariance of articulatory movements: An icerberg model. In J.S. Perkell and D. Klatt, editors, *Invariance and Variability of Speech Processes*. Erlbaum, Hillsdale, NJ, chapter 11, pages 226–242.

Fukuda, Takashi and Tsuneo Nitta. 2003. Noise-robust automatic speech recognition using orthogonalized distinctive phonetic feature vectors. In *Proceedings of EUROSPEECH-2003*, pages 2189–2192.

Fukuda, Takashi, Wataru Yamamoto, and Tsuneo Nitta. 2003. Distinctive phonetic feature extraction for robust

speech recognition. In *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 2, pages 25–28, Hong Kong, April.

Garay-Vitoria, Nestor and Julio Abascal. 2006. Text prediction systems: a survey. *Universal Access in the Information Society*, 4(3):188–203.

Gauvain, Jean-Luc and Chin-Hui Lee. 1994. Maximimum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. 2:291–298.

Ghahramani, Zoubin. 1998. Learning dynamic Bayesian networks. In *Adaptive Processing of Sequences and Data Structures*, pages 168–197. Springer-Verlag.

Gluck, Mark A. and Catherine E. Myers. 1999. *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus and Learning*. MIT Press.

Goldstein, Louis, Dani Byrd, and Elliot Saltzman. 2006. The role of vocal tract gestural action units in understanding the evolution of phonology. In M.A. Arib, editor, *Action to Language via the Mirror Neuron System*. Cambridge University Press, Cambridge, UK, pages 215–249.

Goldstein, Louis M. and Carol Fowler. 2003. Articulatory phonology: a phonology for public language use. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*.

Goozee, Justine V., Bruce E. Murdoch, and Deborah G. Theodoros. 2001. Physiological assessment of tongue function in dysarthria following traumatic brain injury. *Logopedics Phoniatrics Vocology*, 26(2):51–65.

Gopalakrishnan, P.S., Dimitri Kanevsky, Arthur Nádas, and David Nahamoo. 1991. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory*, 37(1):107–113.

Gotoh, Yoshihoko, Michael M. Hochberg, Daniel J. Mashao, and Harvey F. Silverman. 1995. Incremental map estimation of hmms for efficient training and improved performance. In *Proceedings of 1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 457–460, May.

Gracco, Vincent L. 1995. Central and peripheral components in the control of speech movements. In Fredericka Bell-Berti and Lawrence J. Raphael, editors, *Introducing Speech: Contempory Issues, for Katherine Safford Harris*. American Institute of Physics press, chapter 12, pages 417–431.

Green, Phil, James Carmichael, Athanassios Hatzis, Pam Enderby, Mark Hawley, and Mark Parker. 2003. Automatic speech recognition with sparse training data for dysarthric speakers. In *Proceedings of Eurospeech 2003*, pages 1189–1192, Geneva.

Gruen, Armin and Thomas S. Huang, editors. 2001. *Calibration and orientation of cameras in computer vision*. Springer-Verlag, Berlin.

Guenther, Frank H. and Joseph S. Perkell. 2004. A neural model of speech production and its application to stud-

ies of the role of auditory feedback in speech. In Ben Maassen, Raymond Kent, Herman Peters, Pascal Van Lieshout, and Wouter Hulstijn, editors, *Speech Motor Control in Normal and Disordered Speech*. Oxford University Press, Oxford, chapter 4, pages 29–49.

Hamill, Melinda, Vicky Young, Jennifer Boger, and Alex Mihailidis. 2009. Development of an automated speech recognition interface for personal emergency response systems. *Journal of Neuroengineering and Rehabilitation*, 6(26), July.

Hammen, Vicki L., Kathryn M. Yorkston, and Fred D. Minifie. 1994. Effects of Temporal Alterations on Speech Intelligibility in Parkinsonian Dysarthria. *Journal of Speech and Hearing Research*, 37:244–253.

Hartley, Richard and Andrew Zisserman. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, April.

Hasegawa-Johnson, Mark and Margaret Fleck. 2007. International Speech Lexicon Project.

Hasegawa-Johnson, Mark, Jon Gunderson, Adrienne Perlman, and Thomas Huang. 2006a. HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, volume 3, pages 1060–1063, May.

Hasegawa-Johnson, Mark, Jon Gunderson, Adrienne Perlman, and Thomas S. Huang. 2006b. Audiovisual phonologic-feature-based recognition of dysarthric speech. abstract.

Havstam, Christina, Margret Buchholz, and Lena Hartelius. 2003. Speech recognition and dysarthria: a single subject study of two individuals with profound impairment of speech and motor control. *Logopedics Phoniatrics Vocology*, 28:81–90(10), August.

Hawley, Mark S., Pam Enderby, Phil Green, Stuart Cunningham, Simon Brownsell, James Carmichael, Mark Parker, Athanassios Hatzis, Peter ONeill, and Rebecca Palmer. 2007. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593, June.

Hayes, Monson H. 1999. *Digital Signal Processing*. Schaum's Outlines. McGraw Hill.

Heikkilä, Janne and Olli Silvén. 1997. A four-step camera calibration procedure with implicit image correction. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR97)*.

Herndon, Robert M. 1997. *Handbook of Neurologic Rating Scales*. Demos Medical Publishing, 1st edition.

Hess, Wolfgang J. 2008. Pitch and voicing determination of speech with an extension toward music signal. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Hill, Anne J., Deborah G. Theodoros, Trevor G. Russell, Louise M. Cahill, Elizabeth C. Ward, and Kathy M. Clark. 2006. An internet-based telerehabilitation system for the assessment of motor speech disorders: A pilot study. *American Journal of Speech-Language Pathology*, 15:45–56, February.

Hogden, John, Anders Lofqvist, Vince Gracco, Igor Ziokarnik, Philip Rubin, and Elliot Saltzman. 1996. Accurate

recovery of articulator positions from acoustics: New conclusions based on human data. *Journal of the Acoustical Society of America*, 100(3):1819–1834.

Hogden, John, Philip Rubin, Erik McDermott, Shigeru Katagiri, and Louis Goldstein. 2007. Inverting mappings from smooth paths through r^n to paths through r^m: A technique applied to recovering articulation from acoustics. *Speech Communication*, 49(5):361–383.

Hogden, John E. 1996. A maximum likelihood approach to estimating speech articulator positions from speech acoustics. *The Journal of the Acoustical Society of America*, 100(4):2663–2664.

Hoole, Philip and Andreas Zierdt. 2010. Five-dimensional articulography. In Ben Maassen and Pascal H.H.M. van Lieshout, editors, *Speech motor control: New developments in basic and applied research*. Oxford University Press, Oxford, chapter 20, pages 331–349.

Hoole, Philip, Andreas Zierdt, and Christian Geng. 2003. Beyond 2D in articulatory data acquisition and analysis. In *Proceedings of the Fifteenth International Congress of Phonetic Sciences*, pages 265–268, Barcelona.

Hosom, John-Paul, Alexander B. Kain, Taniya Mishra, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2003. Intelligibility of modifications to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 924–927, April.

Huang, Xuedong, Alex Acero, and Hsiao-Wuen Hon. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, April.

Huang, Xuedong and Kai-Fu Lee. 1993. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1:150–157, April.

Huber, Marco F., Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. 2008. On entropy approximation for Gaussian mixture random vectors. In *Proceedings of the 2008 IEEE International Conference on In Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188, Seoul, South Korea.

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

Hustad, Katherine C. and David R. Beukelman. 2002. Listener comprehension of severely dysarthric speech: Effects of linguistic cues and stimulus cohesion. *Journal of Speech, Language, and Hearing Research*, 45:545–558, June.

Hux, Karen, Joan Rankin-Erickson, Nancy Manasse, and Elizabeth Lauritzen. 2000. Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication (AAC)*, 16(3):186 –196, January.

Iskarous, Khalil, Louis M. Goldstein, D.H. Whalen, Mark K. Tiede, and Philip E. Rubin. 2003. CASY: The

Haskins Configurable Articulatory Synthesizer. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 185–188, Barcelona, Spain, August.

Jamieson, D.G., L. Deng, M. Price, V. Parsa, and J. Till. 1996. Interaction of speech disorders with speech coders: effects on speech intelligibility. In *Proceedings of Fourth International Conference on Spoken Language, 1996. ICSLP 96.*, volume 2, pages 737–740.

Jayaram, Gowtham and Kadry Abdelhamied. 1995. Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development*, 32(2):162–169.

Jensen, J.L.W.V. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193. 10.1007/BF02418571.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2 edition.

Kaburagi, Tokihiko, Kohei Wakamiya, and Msaaki Honda. 2005. Three-dimensional electromagnetic articulography: A measurement principle. *Journal of the Acoustical Society of America*, 118(1):428–443, July.

Kain, Alexander B., John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, September.

Karpov, Alexey A., Andrey L. Ronzhin, Alexander I. Nechaev, and Svetlana E. Chernakova. 2004. Assistive multimodal system based on speech recognition and head tracking. In *Proceedings of SPECOM'2004*, St. Petersburg Russia, September.

Katagiri, Shigeru, Biing-Hwang Juang, and Chin-Hui Lee. 1998. Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. *Transactions of the IEEE*, 86(11):2345–2373, November.

Kawahara, H. and H. Matsui. 2003. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, pages I–256 – I–259 vol.1, April.

Kawahara, H., R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno. 2009. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 3905–3908, April.

Kawahara, Hideki. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353.

Kawahara, Hideki, Alain de Cheveigné, Hideki Banno, Toru Takahashi, and Toshio Irino. 2005. Nearly Defect-Free F0 Trajectory Extraction for Expressive Speech Modifications Based on STRAIGHT. In *Proceedings of INTERSPEECH 2005*, pages 537–540, September.

Kent, Ray D. 2000. Research on speech motor control and its disorders: a review and prospective. *Journal of Communication Disorders*, 33(5):391–428.

Kent, Ray D. and Kristin Rosen. 2004. Motor control perspectives on motor speech disorders. In Ben Maassen, Raymond Kent, Herman Peters, Pascal Van Lieshout, and Wouter Hulstijn, editors, *Speech Motor Control in Normal and Disordered Speech*. Oxford University Press, Oxford, chapter 12, pages 285–311.

Kent, Ray D., Gary Weismer, Jane F. Kent, and John C. Rosenbek. 1989. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499.

Kettebekov, Sanshzar, Mohammed Yeasin, Nils Krahnstoever, and Rajeev Sharma. 2002. Prosody based co-analysis of deictic gestures and speech in weather narration broadcast. In *Workshop on Multimodal Resources and Multimodal System Evaluation. (LREC 2002)*, pages 57–62.

Kida, Yusuke and Tatsuya Kawahara. 2005. Voice activity detection based on optimally weighted combination of multiple features. In *Proceedings of INTERSPEECH-2005*, pages 2621–2624.

Kim, Heejin, Mark Hasegawa-Johnson, and Adrienne Perlman. 2010. Acoustic cues to lexical stress in spastic dysarthria. In *Proceedings of Speech Prosody 2010*.

Kim, Heejin, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech '08)*, pages 1741–1744, Brisbane, Australia, September.

Kim, Heejin, Panying Rong, Torrey M. Loucks, and Mark Hasegawa-Johnson. 2010. Kinematic analysis of tongue movement control in spastic dysarthria. In *Proceedings of Interspeech 2010*.

King, Simon, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester. 2007. Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121(2):723–742, February.

King, Simon and Paul Taylor. 2000. Detection of phonological features in continuous speech using neural networks. *Computer Speech & Language*, 14(4):333–353, October.

Kirchhoff, Katrin. 1999. *Robust Speech Recognition Using Articulatory Information*. Ph.D. thesis, University of Bielefeld, Germany, July.

Klatt, Dennis H. 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995.

Kroos, Christian. 2008. Measurement Accuracy in 3D Electromagnetic Articulography (Carstens AG500). In *Proceedings of the 8th International Seminar on Speech Production*, pages 61–64.

Krose, Ben J. A. and Patrick van der Smagt. 1996. An introduction to neural networks. Technical report, University of Amsterdam.

Kubrick, Stanley. 1968. 2001: A space odyssey. Motion picture. Distributed by Metro-Goldwyn-Mayer.

Kuo, Hong-Kwang Jeff, Eric Fosler-Lussier, Hui Jiang, and Chin-Hui Lee. 2002. Discriminative training of language models for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '02)*, pages I–325–I–328, Orlando, USA.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Lai, Pei Ling and Colin Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377.

Lamere, Paul, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, M. Warmuth, and Peter Wolf. 2003. The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, Apr.

Lazo, Aida C. G. Verdugo and Pushpa N. Rathie. 1978. On the entropy of continuous probability distributions. *IEEE Transactions on Information Theory*, 23(1):120–122, January.

Lease, Matthew, Mark Johnson, and Eugene Charniak. 2006. Recognizing disfluencies in conversational speech. *IEEE transactions on Audio, Speech and Language Processing*, 14(5):1566–1573.

Lee, Leo J., Paul Fieguth, and Li Deng. 2001. A functional articulatory dynamic model for speech production. In *in Proceedings of ICASSP*, pages 797–800, Salt Lake City, USA.

Levelt, Willem J. M., Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–75, February.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710.

Li, Jianhua and Graeme Hirst. 2005. Semantic knowledge in word completion. In *Assets '05: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pages 121–128, New York, NY, USA. ACM Press.

Liberman, A. M., F.S. Cooper, D.P. Shankweiler, and M Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review*, 74:431–461.

Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith. 1957. The discrimi-

nation of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), November.

Liberman, Alvin M. and Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition*, 21:1–36.

Little, William John. 1861. On the influence of abnormal parturition, difficult labour, premature birth, and asphyxia neonatorum on the mental and physical condition of the child, especially in relation to deformities. *Transactions of the Obstetrical Society of London*, 3:293–344.

Livescu, Karen, Ozgur Cetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Haggerty, and Bronwyn Woods. 2007. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, April.

Löfqvist, Anders, Nancy S. McGarr, and Kiyoshi Honda. 1984. Laryngeal muscles and articulatory control. *The Journal of the Acoustical Society of America*, 76(3):951–954.

Maeda, Shinji. 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*. Kluwer, pages 131–149.

Maier, A., T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nth. 2009. PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425 – 437.

Markov, Konstantin, Jianwu Dang, and Satoshi Nakamura. 2006. Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. *Speech Communication*, 48(2):161–175, February.

Martin, Rainer. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512, July.

Matiasek, Johannes, Marco Baroni, and Harald Trost. 2002. FASTY : A Multi-lingual Approach to Text Prediction. In *ICCHP '02: Proceedings of the 8th International Conference on Computers Helping People with Special Needs*, pages 243–250, London, UK. Springer-Verlag.

Matsumasa, Hironori, Tetsuya Takiguchi, Yasuo Ariki, I-Chao Li, and Toshitaka Nakabayashi. 2009. Integration of metamodel and acoustic model for dysarthric speech recognition. *Journal of Multimedia*, 4(4):254–261, August.

Matsumasa, Hironori, Tetsuya Takiguchi, Yasuo Ariki, Ichao Li, and Toshitaka Nakabayashi. 2007. PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders. In *Proceedings of Interspeech 2007*, pages 1150–1153.

McDermott, Erik and Atsushi Nakamura. 2006. Production-oriented models for speech recognition. *IEICE - Trans. Inf. Syst.*, E89-D(3):1006–1014.

McGurk, Harry and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264:746–748, December.

McHenry, Monica A. and Julie M. Liss. 2006. The impact of stimulated vocal loudness on nasalance in dysarthria. *Journal of Medical Speech-Language Pathology*, 14(3):197–205, September.

McLennan, Sean. 2000. Klatt Synthesizer in Simulink. Technical report, Indiana University, April.

McLuhan, Marshall. 1964. *Understanding Media: The Extensions of Man*. McGraw Hill, New York NY.

Melen, T. 1994. *Geometrical Modelling and Calibration of Video Cameras for Underwater Navigation*. Ph.D. thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Menendez-Pidal, Xavier, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzjo, and H.T. Bunnell. 1996. The Nemours Database of Dysarthric Speech. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA, USA, October.

Mermelstein, P. 1973. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070–1082.

Messina, James J. and Constance M. Messina. 2007. Description of AAC devices. http://www.coping.org/specialneeds/assistech/aacdev.htm, April.

Metze, Florian. 2007. Discriminative speaker adaptation using articulatory features. *Speech Communication*, 49(5):348–360.

Miyazawa, Yasunaga. 1993. An all-phoneme ergodic hmm for unsupervised speaker adaptation. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 574 –577 vol.2, Aoruk.

Moore, Keith L. and Arthur F. Dalley. 2005. *Clinically Oriented Anatomy, Fifth Edition*. Lippincott, Williams and Wilkins.

Morales, Santiago Omar Caballero and Stephen J. Cox. 2009. Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing*.

Morency, Louis-Philippe, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June.

Mori, Hiroki, Yasunori Kobayashi, Hideki Kasuya, Noriko Kobayashi, and Hajime Hirose. 2005. Evaluation of fundamental frequency (f0) characteristics of speech in dysarthrias: A comparative study. *Acoustical Science and Technology*, 26(6):540–543.

Motlcek, Petr, Luks Burget, and Jan Cernock. 2005. Visual features for multimodal speech recognition. In

Proceedings of Radioelektronika 2005. Fakulta elektrotechniky a komunikacnch technologi VUT.

Moulines, Eric and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.*, 9:453–467, December.

Murphy, Kevin Patrick. 1998. Switching Kalman Filters. Technical report.

Murphy, Kevin Patrick. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California at Berkeley.

Na, KyungMin, JaeYeol Rheem, and SouGuil Ann. 1994. A discriminative training algorithm for predictive neural network models. In *1994 IEEE International Symposium on Circuits and Systems*, pages 431–434, London, England, June.

Nakatani, Christine. 1993. A speech-first model for repair detection and correction. In *Proceedings of the 31 th Annual Meeting of the Association for Computational Linguistics*, pages 46–53.

Nam, Hosung and Louis Goldstein. 2006. TADA (TAsk Dynamics Application) manual.

Nam, Hosung and Elliot Saltzman. 2003. A competitive, coupled oscillator model of syllable structure. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pages 2253–2256, Barcelona, Spain.

Namasivayam, Aravind Kumar and Pascal van Lieshout. 2008. Investigating speech motor practice and learning in people who stutter. *Journal of Fluency Disorders*, 33:32–51.

Neel, Amy T. 2009. Effects of loud and amplified speech on sentence and word intelligibility in parkinson disease. *Journal of Speech, Language, and Hearing Research*, 52:1021–1033, August.

Nefian, Ara V., Luhong Liang, Xiaoxing Liu, Xiaobo Pi, and Kevin Murphy. 2002a. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP, Journal of Applied Signal Processing*.

Nefian, Ara V., Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. 2002b. A coupled hmm for audio-visual speech recognition. In *in International Conference on Acoustics, Speech and Signal Processing (CASSP02*, pages 2013–2016.

Neti, Chalapathy, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, Azad Mashari, and Jie Zhou. 2000. Audio-visual speech recognition, final workshop 2000 report. Technical report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, October.

Nishio, Masaki and Seiji Niimi. 2001. Speaking rate and its components in dysarthric speakers. *Clinical Linguistics & Phonetics*, 15(4):309–317, June.

Niyogi, Partha and Chris Burges. 2002. Detecting and interpreting acoustic features with support vector machines. Technical Report TR-2002-02, University of Chicago.

Noyes, Jan M. and Clive R. Frankish. 1992. Speech recognition technology for individuals with disabilities. *Augmentative and Alternative Communication (AAC)*, 8(4):297–303.

Nuffelen, Gwen Van, Catherine Middag, Marc De Bodt, and JeanPierre Martens. 2009. Speech technologybased assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders*, 44(5):716–730.

O'Shaughnessy, Douglas. 2000. *Speech Communications – Human and Machine*. IEEE Press, New York, NY, USA.

O'Shaughnessy, Douglas. 2008. Formant estimation and tracking. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Ostendorf, Mari. 2000. Incorporating linguistic theories of pronunciation variation into speech-recognition models. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 358(1769):1325–1338.

Oviatt, Sharon. 2003. Multimodal interfaces. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, pages 286–304.

Ozawa, Yoshiaki, Osamu Shiromoto, Fumiko Ishizaki, and Toshiko Watamori. 2001. Symptomatic differences in decreased alternating motion rates between individuals with spastic and with ataxic dysarthria: An acoustic analysis. *International Journal of Phoniatrics, Speech Therapy and Communication Pathology*, 53(2).

Palmer, Rebecca, Pam Enderby, and Mark Hawley. 2007. Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared. *International Journal of Language & Communication Disorders*, 42:61–79.

Papandreou, George, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. 2009. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, March.

Park, Unsang and Anil K. Jain. 2006. 3D Face Reconstruction from Stereo Video. In *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV06)*.

Parks, T. W. and C. S. Burns. 1987. *Digital Filter Design*. John Wiley & Sons.

Patel, Rupal. 1998. Control of prosodic parameters by an individual with severe dysarthria. Technical report, University of Toronto, December.

Patel, Rupal. 2002a. Phonatory control in adults with cerebral palsy and severe dysarthria. *AAC Augmentative and Alternative Communication*, 18:2–10.

Patel, Rupal. 2002b. Prosodic control in severe dysarthria: Preserved ability to mark the question-statement contrast. *Journal of speech, language, and hearing research*, 45(5):858–870.

Patterson, E. K., S. Gurbuz, Z. Tufekci, and J. N. Gowdy. 2002. CUAVE: A new audio-visual database for

multimodal human-computer interface research. In *Proceedings of ICASSP*, pages 2017–2020.

Pawlak, Zdzislaw. 1982. Rough sets. *International Journal of Information and Computer Sciences*, 11(5):341–356.

Platt, John C., Nello Cristianini, and John Shawe-Taylor, 2000. *Advances in Neural Information Processing Systems*, chapter Large Margin DAGS for Multiclass Classification. MIT Press, 12 edition.

Plauché, Madelaine C. and Elizabeth E. Shriberg. 2007. Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proceedings of the International Congress of Phonetic Sciences*.

Polur, Prasad D. and Gerald E. Miller. 2006. Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical Engineering and Physics*, 28(8):741–748, October.

Portnoff, Michael R. 1976. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):243–248.

Posey, William Campbell. 1923. Some ocular phases of Litte's disease (congenital spastic rigidity of the limbs. *Journal of the American Medical Association*, 80(2):80–82.

Potamianos, Gerasimos and Hans Peter Graf. 1998. Discriminative training of hmm stream exponents for audio-visual speech recognition. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pages 3733–3736.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: the art of scientific computing*. Cambridge University Press, second edition.

Quatieri, Thomas F. 2002. *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice Hall.

Raghavendra, Parimala, Elisabet Rosengren, and Sheri Hunnicutt. 2001. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication (AAC)*, 17(4):265–275, December.

Ramsay, Jim O. and Bernard W. Silverman. 2002. *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. Springer-Verlag.

Ramsay, J.O. and B.W. Silverman, 2005. *Fitting differential equations to functional data: Principal differential analysis*, pages 327–348. Springer.

Reynolds, Douglas A. and Richard C. Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January.

Richardson, Matt, Jeff Bilmes, and Chris Diorio. 2000. Hidden-articulator markov models: Performance improvements and robustness to noise. In *Proceedings of the ICSLP*, pages 131–134.

Richmond, Korin, Simon King, and Paul Taylor. 2003. Modelling the uncertainty in recovering articulation from

acoustics. *Computer Speech and Language*, 17:153–172.

Robinson, Tony, Mike Hochberg, and Steve Renals, 1996. *The use of recurrent neural networks in continuous speech recognition in* Automatic Speech and Speaker Recognition - Advanced Topics, chapter 10, pages 233–258. Kluwer Academic Publishers.

Rodman, R.D., T.S. Moody, and J.A. Price. 1985. Speech recognizer performance with dysarthric speakers: A comparison of two training procedures. *Speech Technology*, 1:65–71.

Rosen, Kristin and Sasha Yampolsky. 2000. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative & Alternative Communication*, 16(1):48–60, Jan.

Roweis, Sam T. 1999. *Data Driven Production Models for Speech Processing*. Ph.D. thesis, California Institute of Technology, Pasadena, California.

Roy, Nelson, Herbert A. Leeper, Michael Blomgren, and Rosalea M. Cameron. 2001. A description of phonetic, acoustic, and physiological changes associated with improved intelligibility in a speaker with spastic dysarthria. *American Journal of Speech-Language Pathology*, 10:274–290, August.

Rubin, Philip, Thomas Baer, and Paul Mermelstein. 1981. An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, 70(2):321–328.

Rudzicz, Frank. 2006. Clavius: Understanding language understanding in multimodal interaction. Master's thesis, McGill University, Dept. of Electrical and Computer Engineering.

Rudzicz, Frank. 2009a. Applying discretized articulatory knowledge to dysarthric speech. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09)*, Taipei, Taiwan, April.

Rudzicz, Frank. 2009b. Phonological features in discriminative classification of dysarthric speech. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09)*, Taipei, Taiwan, April.

Rudzicz, Frank. 2010a. Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing,* (in press).

Rudzicz, Frank. 2010b. Correcting errors in speech recognition with articulatory dynamics. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala Sweden, July 12-14.

Rudzicz, Frank. 2010c. Learning mixed acoustic/articulatory models for disabled speech. In *The NIPS-10 Workshop on Machine Learning for Assistive Technologies*.

Rudzicz, Frank. 2010d. Towards a noisy-channel model of dysarthria in speech recognition. In *Proceedings of the First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT) at the 11th*

Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010), pages 80–88, Los Angeles California, June 2-6.

Rudzicz, Frank, Aravind Kumar Namasivayam, and Talya Wolff. 2010. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation,* (in press).

Rudzicz, Frank, Pascal van Lieshout, Graeme Hirst, Gerald Penn, Fraser Shein, and Talya Wolff. 2008. Towards a comparative database of dysarthric articulation. In *Proceedings of the eighth International Seminar on Speech Production (ISSP'08)*, Strasbourg France, December.

Russell, Stuart and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Saenko, K. and K. Livescu. 2006. An Asynchronous DBN for Audio-Visual Speech Recognition. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 154–157, December.

Sakoe, Hiroaki and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-26, February.

Salleh, Sheikh Husssain Shaikh, Ahmad Zuri, Zulkarnian Yusoff, Syed Rahman, and Lim Soon Chieh. 2000. Implementation of speaker identification system by means of personel (sic) computer. In *Proceedings of IEEE TENCON 2000*, pages 43–48, September.

Saltzman, Elliot L. and Kevin G. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382.

Saltzman, Elliot M., 1986. *Task dynamic co-ordination of the speech articulators: a preliminary model*, pages 129–144. Springer-Verlag.

Sanders, Eric, Marina Ruiter, Lilian Beijer, and Helmer Strik. 2002a. Automatic Recognition of Dutch Dysarthric Speech: a Pilot Study. In *7th International Conference on Spoken Language Processing*, Denver, Colorado, September.

Sanders, Eric, Marina Ruiter, Lilian Beijer, and Helmer Strik. 2002b. Automatic recognition of Dutch dysarthric speech: A pilot study. In *Proceedings of 7th International Conference on Spoken Language Processing*, pages 661–664, September.

Sandness, Eric D. 2000. Discriminative training of acoustic models in a segment-based speech recognizer. Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science.

Sawhney, Nitin and Sean Wheeler. 1999. Using phonological context for improved recognition of dysarthric speech. Technical Report 6345, MIT Media Lab.

Scharenborg, Odette, Vincent Wan, and Roger K. Moore. 2007. Towards capturing fine phonetic variation in speech using articulatory features. *Speech Communication*, 49(10-11):811–826, October-November.

Schneiderman, Carl R. and Robert E. Potter. 2002. *Speech-language pathology : a simplified guide to structures,*

functions, and clinical implications. Academic Press, San Diego, CA.

Schroeter, Juergen. 2008. Basic principles of speech synthesis. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Schwander, Teresa and Harry Levitt. 1987. Effect of two-microphone noise reduction on speech recognition by normal-hearing listeners. *Journal of Rehabilitation Research and Development*, 24(2):87–92.

Seikel, J. Anthony, Douglas W. King, and David G. Drumright, editors. 2005. *Anatomy & Physiology: for Speech, Language, and Hearing*. Thomson Delmar Learning, third edition.

Sethares, William Arthur. 2007. *Rhythm and Transforms*. Springer.

Shannon, Claude E. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Sharma, Harsh Vardhan and Mark Hasegawa-Johnson. 2010a. State transition interpolation and map adaptation for hmm-based dysarthric speech recognition. In *Proceedings of the First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT) at the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, Los Angeles California, June 2-6.

Sharma, Harsh Vardhan and Mark Hasegawa-Johnson. 2010b. State-transition interpolation and map adaptation for hmm-based dysarthric speech recognition. In *First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*.

Sharma, R., M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, A. MacEachren, K. Sengupta, and I. Brewer. 2003. Speech-gesture driven multimodal interfaces for crisis management. In *Proceedings of IEEE special issue on Multimodal Human-Computer Interface*.

Shi, Guangji, Parham Aarabi, and Hui Jiang. 2007. Phase-based dual-microphone speech enhancement using a prior speech model. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):109–118, January.

Slama, Chester C., editor. 1980. *Manual of Photogrammetry*. American Society of Photogrammetry and Remote Sensing, 4 edition.

Smith, Anne and Lisa Goffman, 2004. *Interaction of motor and language factors in the development of speech production*, chapter 10, pages 227–252. Speech Motor Control in Normal and Disordered Speech. Oxford University Press, Oxford.

Snell, Roy C. and Fausto Milinazzo. 1993. Formant Location from LPC Analysis Data. *IEEE Transactions on Speech and Audio Processing*, 1(2), April.

Solomon, Nancy P., Donald A. Robin, and Erich S. Luschei. 2000. Strength, Endurance, and Stability of the Tongue and Hand in Parkinson Disease. *Journal of Speech, Language, and Hearing Research*, 43:256–267.

Spiegel, Murray F., Mary Jo Altom, Marian J. Macchi, and Karen L. Wallace. 1990. Comprehensive assessment

of the telephone intelligibility of synthesized and natural speech. *Speech Communication*, 9(4):279 – 291.

Stephenson, Todd A., Mathew Magimai-Doss, and Hervé Bourlard. 2004. Speech recognition with auxiliary information. *IEEE Transactions on Speech and Audio Processing*, 12(3):189–203.

Stevens, Kenneth N. 1998. *Acoustic Phonetics*. MIT Press, Cambridge, Massachussetts.

Stevens, Kenneth Noble. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In Peter B. Denes and Edward E. David Jr., editors, *Human communication: A unified view*. McGraw Hill, New York, pages 51–66.

Stevens, Kenneth Noble and Samuel Jay Keyser. 2010. Quantal theory, enhancement and overlap. *Journal of Phonetics*, 38(1):10 – 19. Phonetic Bases of Distinctive Features.

Stevens, S.S., J. Volkman, and E.B. Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, January.

Stork, David G. and Marcus E. Hennecke, editors. 1996. *Speechreading by Humans and Machines: Models, Systems, and Applications*. Springer.

Stylianou, Yannis. 2008. Voice transformation. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Summerfield, Quentin. 1992. Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences*, 335(1273):71–78, January.

Sun, Jiping and Li Deng. 2002a. An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition. *The Journal of the Acoustical Society of America*, 111(2):1086–1101.

Sun, Jiping and Li Deng. 2002b. An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition. *Journal of the Acoustical Society of America*, 111(2), February.

Sun, Jiping, Xing Jing, and Li Deng. 2000. Data-driven model construction for continuous speech recognition using overlapping articulatory features. In *Proceedings of the International Conference on Spoken Language Processing*, October.

Sundberg, Johan. 1977. The acoustics of the singing voice. *Scientific American*, 234:82–91.

Swiffin, Andrew, John Arnott, J. Adrian Pickering, and Alan Newell. 1987. Adaptive and predictive techniques in a communication prosthesis. *Augmentative & Alternative Communication*, 3(4):181–191, December.

Tamura, Satoshi, Koji Iwano, and Sadaoki Furui. 2004. Multi-modal speech recognition using optical-flow analysis for lip images. *J. VLSI Signal Process. Syst.*, 36(2-3):117–124.

Taylor, Paul, Alan W. Black, and Richard Caley. 1998. The architecture of the Festival speech synthesis system. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 147–151, Jenolan Caves, Australia.

Tebelskis, Joe. 1995. *Speech Recognition using Neural Networks*. Ph.D. thesis, Carnegie Mellon University, School of Computer Science, May.

Thomas-Stonell, Nancy, Ava-Lee Kotler, Herbert A. Leeper, and Philip C. Doyle. 1998. Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative & Alternative Communication*, 14(1):51–56, March.

Thompson, E.C., B.E. Murdoch, and P.D. Stokes. 1995. Lip function in subjects with upper motor neuron type dysarthria following cerebrovascular accidents. *European Journal of Disorders of Communication*, 30:451–466.

Thubthong, Nuttakorn, Prakasith Kayasith, Sriwimon Manochiopinig, Wisit Leelasiriwong, and Onwadee Rukkharangsarit. 2005. Articulation analysis of Thai cerebral palsy children with dysarthric speech. In *Proceedings of the 6th Symposium on Natural Language Processing*.

Tiede, Mark. 2008. MVIEW: Multi-channel visualization application for displaying dynamic sensor movements. in development.

Titze, Ingo R. 1994. *Principles of Voice Production*. Prentice-Hall, Englewood Cliffs, NJ.

Toda, Tomoki, Alan W. Black, and Keiichi Tokuda. 2005. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, Pennsylvania.

Toda, Tomoki, Alan W. Black, and Keiichi Tokuda. 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3):215–227, March.

Tolba, Hesham and Ahmed S. El Torgoman. 2009. Towards the improvement of automatic recognition of dysarthric speech. In *International Conference on Computer Science and Information Technology*, pages 277–281, Los Alamitos, CA, USA. IEEE Computer Society.

Toth, Arthur R. and Alan W. Black. 2005. Cross-speaker articulatory position data for phonetic feature prediction. In *Proceedings of Interspeech 2005*, Lisbon, Portugal.

Tsai, Roger Y. 1987. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, August.

Tsao, Ying-Chiao, Gary Weismer, and Kamran Iqbal. 2006. The effect of intertalker speech rate variation on acoustic vowel space. *The Journal of the Acoustical Society of America*, 119(2):1074–1082, February.

Umapathi, T., N. Venketasubramanian, K. J. Leck, C.B. Tan, W.L. Lee, and H. Tjia. 2000. Tongue deviation in acute ischaemic stroke: a study of supranuclear twelfth cranial nerve palsy in 300 stroke patients. *Cerebrovascular Diseases*, 10:462–465.

Vaerenbergh, Steven Van, Javier Via, and Ignacio Santamaria. 2008. Adaptive kernel canonical correlation analysis algorithms for nonparametric identification of Wiener and Hammerstein systems. *EURASIP Journal on Advances in Signal Processing*, 8(2):1–13, January.

Vaerenbergh, Steven Van, Javier Via, and Ignatio Santamaria. 2006a. A sliding-window kernel RLS algorithm and its application to nonlinear channel identification. In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May.

Vaerenbergh, Steven Van, Javier Via, and Ignatio Santamaria. 2006b. Online kernel canonical correlation analysis for supervised equalization of Wiener systems. In *Proceedings of the 2006 International Joint Conference on Neural Networks*, pages 1198–1204, Vancouver, Canada, July.

van Lieshout, Pascal, Wouter Hulstijn, Peter J. Alfonso, and Herman F.M. Peters. 1997. Higher and lower order influences on the stability of the dynamic coupling between articulators. In Wouter Hulstijn, Herman F.M. Peters, and Pascal van Lieshout, editors, *Speech production: Motor control, brain research and fluency disorders*. Elsevier Science Publishers, Amsterdam, pages 161–170.

van Lieshout, Pascal, Gwen Merrick, and Louis Goldstein. 2008. An articulatory phonology perspective on rhotic articulation problems: A descriptive case study. *Asia Pacific Journal of Speech, Language, and Hearing*, 11(4):283–303.

van Lieshout, Pascal H. H. M., Arpita Bose, Paula A. Square, and Catriona M. Steele. 2007. Speech motor control in fluent and dysfluent speech production of an individual with apraxia of speech and Broca's aphasia. *Clinical Linguistics & Phonetics*, 21(3):159–188, March.

van Lieshout, Pascal H.H.M. and Wassim Moussa. 2000. The assessment of speech motor behavior using electromagnetic articulography. 81:9–22.

Vapnik, Vladimir Naumovich. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.

Wan, Vincent and James Carmichael. 2005. Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2005)*, September.

Webber, Sharon G. 2005. Webber photo cards: Story starters.

Weijnen, F.G., J.B.M. Kuks, A. van der Bilt, H.W. van der Glas, M.W. Wassenberg, and F. Bosman. 2000. Tongue force in patients with myasthenia gravis. *Acta Neurologica Scandinavica*, 102(5):303–308.

Westbury, John R., 1994. *X-ray microbeam speech production database user's handbook*. Waisman Center on Mental Retardation & Human Development.

Wester, Mirjam. 2003. Syllable classification using articulatory - acoustic features. In *Proceedings of Eurospeech 2003*, pages 233–236, Geneva, Switzerland.

Wester, Mirjam, Joe Frankel, and Simon King. 2004. Asynchronous articulatory feature recognition using dynamic Bayesian networks. In *Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop*, volume 104, pages 37–42, Kyoto, Japan.

Wilson, William H. 1995. Stability of learning in classes of recurrent and feedforward networks. In *Proceedings of the Sixth Australian Conference on Neural Networks (ACNN 95)*, pages 142–145, February.

Woodland, P.C. 2001. Speaker adaptation for continuous density HMMs: a review. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, August.

Wrench, Alan. 1999. The MOCHA-TIMIT articulatory database, November.

Wrench, Alan and Korin Richmond. 2000. Continuous speech recognition using articulatory data. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China.

Wu, Ting-Fan, Chih-Jen Lin, and Ruby C. Weng. 2003. Probability estimates for multi-class classification by pairwise coupling. In *Proceedings of Neural Information Processing Systems 2003*.

Yan, Qin, Saeed Vaseghi, Esfandiar Zavarehei, Ben Milner, Jonathan Darch, Paul White, and Ioannis Andrianakis. 2007. Formant tracking linear prediction model usng hmms and kalman filters for noisy speech processing. *Computer Speech and Language*, 21:543–561.

Yehia, Hani Camille. 2002. *A Study On The Speech Acoustic-To-Articulatory Mapping Using Morphological Constraints*. Ph.D. thesis, Nagoya University, Graduate School of Engineering.

Yorkston, Kathryn M. and David R. Beukelman. 1981. *Assessment of Intelligibility of Dysarthric Speech*. C.C. Publications Inc., Tigard, Oregon.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2006. The HTK Book (version 3.4).

Yunusova, Yana, Jordan R. Green, and Antje Mefferd. 2009. Accuracy Assessment for AG500, Electromagnetic Articulograph. *Journal of Speech, Language, and Hearing Research*, 52:547–555, April.

Yunusova, Yana, Gary Weismer, John R. Westbury, and Mary J. Lindstrom. 2008. Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research*, 51:596–611, June.

Zheng, Wenming, Xiaoyan Zhou, Cairong Zou, and Li Zhao. 2006. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE Transactions on Neural Networks*, 17(1):233–238.

Ziegler, Wolfram and Ben Maassen, 2004. *The role of the syllable in disorders of spoken language production*, chapter 16, pages 415–447. Speech Motor Control in Normal and Disordered Speech. Oxford University Press, Oxford.

Zierdt, Andreas, Philip Hoole, and Hans G. Tillmann. 1999. Development of a system for three-dimensional fleshpoint measurement of speech movements. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, page 3.

Zierdt, Andreas, Phillip Hoole, Masaaki Honda, Tokihiko Kaburagi, and Hans G. Tillmann. 2000. Extracting tongues from moving heads. In *Proceedings of the 5th Speech Production Seminar*, pages 313–316.

Zue, Victor, Stephanie Seneff, and James Glass. 1989. Speech Database Development: TIMIT and Beyond. In *Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases (SIOA-1989)*, volume 2, pages 35–40, Noordwijkerhout, The Netherlands.

Zweig, Geoffrey G. 1998. *Speech Recognition with Dynamic Bayesian Networks*. Ph.D. thesis, University of California, Berkeley, Department of Computer Science.