USC-SIPI REPORT #415

Structure and Function in Speech Production

by

Adam Lammert

February 2014

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

STRUCTURE AND FUNCTION IN SPEECH PRODUCTION

Ph.D. Dissertation Submitted by Adam C. Lammert

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE)

February 2014

GUIDANCE COMMITTEE

Dr. Shrikanth S. Narayanan (Chairperson)

Dr. Gérard G. Medioni

Dr. Louis M. Goldstein (Outside Member)

To my family:

Tina, Henry and Kia

Mom and Dad

Ann, Fred and Freddie

Nicki and Lars

Acknowledgments

I would first like to acknowledge the unwavering support and encouragement of my mentors, Shrikanth Narayanan, Louis Goldstein and Michael Proctor. I gratefully acknowledge the constructive feedback from my qualifying exam and defense committee members, Gerard Medioni, Gaurav Sukhatme and Stefan Schaal. I would also like to acknowledge Pierre Divenyi, for introducing me to speech and hearing research and for allowing me the freedom to explore its many aspects.

I could not have completed this experience without Vikram Ramanarayanan and Dirk Hovy, my counterparts and friends. My work has benefitted enormously from the insight of Khalil Iskarous, Dani Byrd, Sungbok Lee, Krishna Nayak, Richard Leahy, Athanasios Katsamanis, Panayiotis Georgiou and Rachel Walker, as well as many helpful discussions with Erik Bresch, Prasanta Ghosh, Asterios Toutios, Christina Hagedorn, Ben Parrell, Yoon Kim, Jangwon Kim, Yinghua Zhu, Daniel Bone, Colin Vaz. I am indebted to my mentees Simon Berman, Li Hsuan Lu and Nishit Malde for helping me to see old ideas in a new light. I am very appreciative of the many interesting conversations I have had with Naveen Kumar, Kartik Audhkhasi, Emily Mower, Matt Black, Carlos Busso, Joe Tepperman, Viktor Rozgic, Zisis Skordilis, Joseph Crew and Justin Aronoff. My breadth of knowledge and experience was greatly expanded by Sarah Bottjer, Neil Segil and all the faculty in the Hearing and Communication Neuroscience program at the University of Southern California. I would also like to acknowledge

Thomas Quatieri, Nicholas Malyska and Andrew Dumas at MIT Lincoln Laboratory for their encouragement and for our many lively discussions over common interests.

Funding for my work has come largely from research grants from the National Institutes of Health. My graduate studies have been supported by a graduate fellowship from the Annenberg Foundation, a graduate training fellowship from the National Institutes of Health, a scholarship from the Achievement Rewards for College Scientist Foundation, research assistantships from the Signal and Image Processing Institute at the University of Southern California, and by the Raymond H. Stetson Scholarship in Phonetics and Speech Science from the Acoustical Society of America.

Contents

List of Tables

List of Figures

Preface

A key axiom of the modern perspective on cognition is that its various forms – perception and action, learning and memory, language, for example – can be accurately characterized as computational/informational processes (e.g., [Wiener, 1948, McCulloch, 1949, Newell and Simon, 1972, Pfeifer, 2001]). The goals of the present research program stem from the diversity of my research experiences, all of which are unified by an underlying interest in understanding the computational mechanisms of cognition in the natural domain and an interest in applying those mechanisms in the technological domain. Some of my early work focused on nature's elegant solutions to seemingly impossible challenges in directed locomotion [Long et al., 2004, 2006]. Modeling and implementing these biologically-inspired control architectures in simulation and in mobile robots provided a detailed understanding of their precise functioning and validation of their robustness. Conversely, I have seen how biologically-inspired genetic algorithms can be harnessed for extremely practical purposes, namely to improve the efficiency of digital logic circuits [Lammert, 2006].

I have a particular interest in the mechanisms that drive perceptual and motor processes, which represent a fundamental aspect of intelligence. Appropriate interaction with the environment is essential for intelligence, whether by transforming raw sensory data into meaningful percepts, or by actively utilizing motor systems to affect some physical/informational change on the environment. Without the ability to interact with

the world, higher-level processes – e.g., complex reasoning and language – run the risk of being disconnected and irrelevant. Moreover, it has been often noted that, contrary to expectations, low-level sensorimotor mechanisms are actually among the most computationally challenging problems faced by intelligent systems (sometimes known as Moravec's paradox [Moravec, 1988]). Perception-action mechanisms related to speech are some of the most crucial that humans posses because the ability to produce and perceive speech forms much of the basis for human communication, expression and social interaction.

Just as perception and action are at the core of intelligence, physical structure is at the core of how embodied perception-action systems function. Structure is an especially important consideration in speech production, because of the highly complex and unusual physical structure upon which speech production is built. Speech production involves the rapid actions of a highly redundant and diverse set of articulators within a confined space (i.e., the vocal tract) which, despite their essential role as part of a communicative system, are largely apparent to others only by their complex acoustic consequences. The basic premise of this thesis is that mechanical knowledge of the speech production system's structure is essential to understanding its function. I will attempt to show how structure itself can be analyzed, as well as ways in which that knowledge can be used for analyzing the mechanisms of speech motor control and for explaining speech production behavior.

Introduction

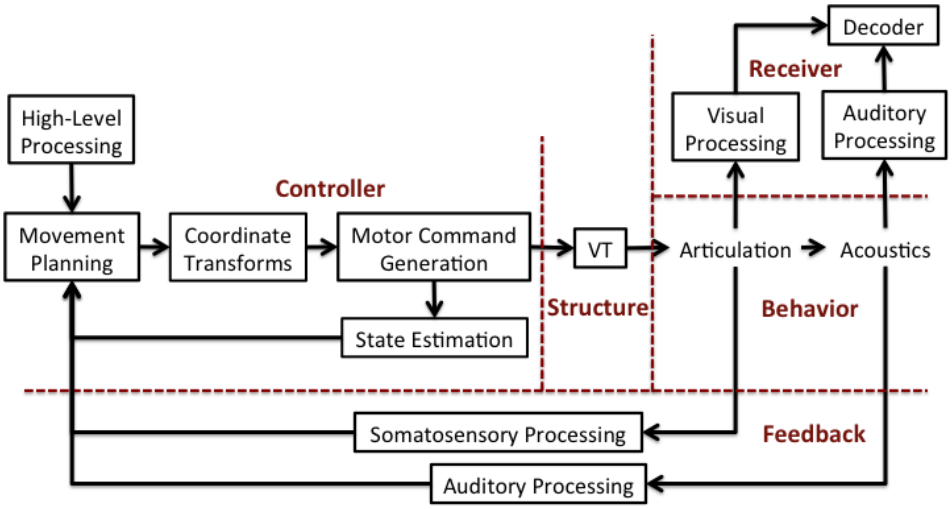


Figure 1: Control-theoretic view of speech production – somewhat detailed.

Figure 1 shows a somewhat detailed control-theoretic diagram of speech production. Speech motor control is complex, presumably involving many distinct processes from planning to articulation, as well as multiple feedback and output signals. To facilitate discussion, we can divide this diagram it into a smaller number of general components, including the controller, the structure itself (i.e., the plant), behavior (e.g., articulation and acoustics), feedback (i.e., sensory processing) and the receiver.

In the more generalized diagram (Figure 2), it becomes clearer that the physical structure is at the center of things. It is upon the physical structure that the controller imposes its commands, and it is the physical structure that subsequently generates

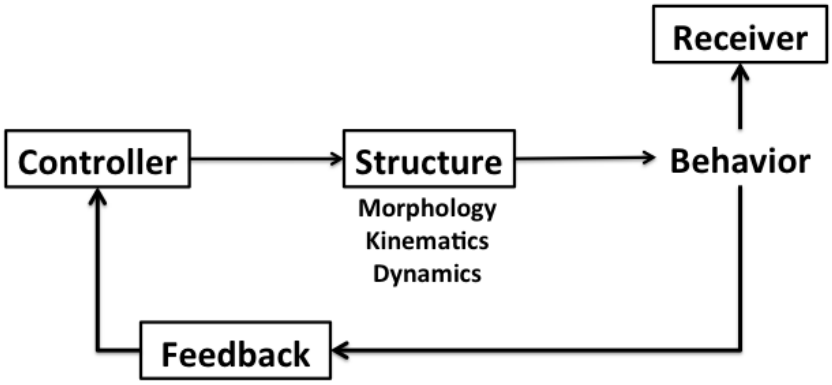


Figure 2: Control-theoretic view of speech production – generic.

behavior. Behavior essentially constitutes a way of describing the structure, namely its actions across time. However, there is another description of the structure which is equally important: a mechanical description, including descriptions that are morphological (i.e., the size and shape and structural components), kinematical (i.e., the configuration of structural components and its impact on their motion and interactions) and dynamical (i.e., the inertial properties of the system's components and, consequently, their reaction to forces acting on the system). A mechanical description of the physical structure is crucial because it constitutes a set of constraints by which control is transformed into behavior. It also provides the necessary context for interpreting behavior and inferring control from behavior. Indeed, the structure of any motor apparatus is a central consideration for understanding its control and behavior, but it is especially true in the domain of speech production, which is a central, motivating factor behind this thesis.

The speech production apparatus has many special considerations that make structure essential to understanding its function. Most speech articulation takes place in a confined environment (i.e., the vocal tract), and that environment and the articulators within it vary widely across speakers (see Figure 3). Moreover, the shape of the vocal tract determines much of the acoustic properties of the system, a crucial consideration

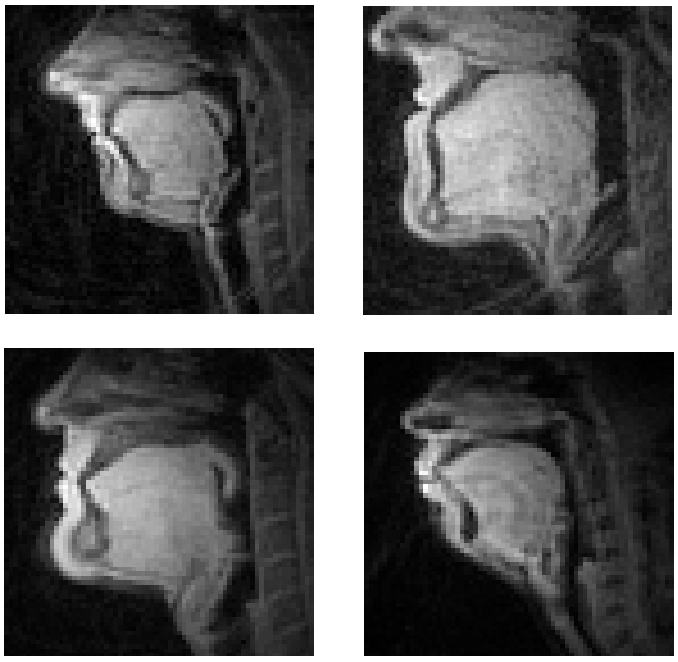


Figure 3: Midsagittal view of four speakers with wide differences in the morphology of their speech production apparatus.

in speech. The system itself is also highly articulated with many degrees of freedom and a diversity of articulators. These special considerations set up a complex interplay between structure and function, which is represented in Figure 4. It is well-known that the acoustic speech signal is a result of the vocal tract shape, and the speech motor controller is free to change the vocal tract shape via speech articulation. However, the structure of the speech production apparatus has a direct influence on vocal tract shape, because it is the size and shape of the articulators – whether passive or active – in combination with articulation that actually determines the overall vocal tract shape. Structure also has an indirect influence on control because it circumscribes the space of possible vocal tract shapes, and also constrains the ways in which those shapes are achieved.

With this in mind, the **guiding statement** of this thesis is that, in order to understand behavior of the speech production apparatus and the mechanisms of its control, it is crucial to incorporate a three-part approach. The first component is empirical and

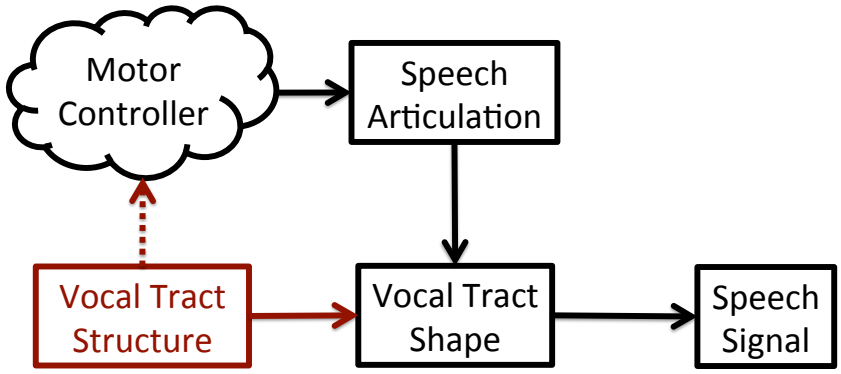


Figure 4: Diagrammatic representation of the interplay between structure of the speech production apparatus and its functional aspects.

involves collecting and processing lots of high-quality production data, of which real-time magnetic resonance imaging is the most useful and important. The second component is computational, and involves methods for data processing, analysis and modeling. The third component is a theoretical perspective that is grounded in understanding and exploiting the crucial structure-function interplay in speech production.

Aspects of *control and behavior* that are of interest relate to variance and invariance in motor action, both within and across speakers. Issues of variance and invariance have long taken center stage in speech production because variance appears ubiquitous and invariant aspects of behavior are elusive. In terms of scientific accounts, these issues make it difficult to define the goals of production, to describe the spatiotemporal form of behavior and to model control mechanisms. These issues also impede technological progress, making it difficult to perform speaker normalization for automatic speech recognition, to accurate design biometrics and speaker identification, and to design realistic and flexible speech synthesis. The hope is that this three-part approach provides some leverage on the questions of interest which are, in many ways, classic problems in speech production research.

The work described here comprises four studies that lay a foundation for this research program. Chapter 1 develops a methodology for quantification and analysis of morphological variation from real-time MRI data, and then applies that methodology to analysis of variation in two key vocal tract structures: the hard palate and the posterior pharyngeal wall. The knowledge gained in the first chapter regarding variation in hard palate morphology is subsequently utilized in Chapter 2 to explore the complex interplay between hard palate morphology, lingual articulation and acoustics during vowel articulation. Chapter 3 describes a different approach to understanding the structure-function interplay, namely through inverting the acoustic signal to predict structural characteristics from their acoustic consequences. This approach motivates a deeper understanding of, and a general theoretical result concerning speech production physics. Finally, in Chapter 4, attention is turned toward kinematics, which is motivated by the potential utility of kinematic knowledge for a variety of applications in system characterization and analysis. The work is toward developing and validating statistical methodologies for quantifying the direct and differential kinematic maps of the vocal tract.

Chapter 1

Morphological Variation in the Adult Hard Palate and Posterior Pharyngeal Wall

1.1 Abstract

Adult human vocal tracts display considerable morphological variation across individuals, but the nature and extent of this variation has not been extensively studied for many vocal tract structures. There exists a need to analyze morphological variation and, even more basically, to develop a methodology for morphological analysis of the vocal tract. Such analysis will facilitate fundamental characterization of the speech production system, with broad implications from modeling to explaining inter-speaker variability. A data-driven methodology to automatically analyze the extent and variety of morphological variation is proposed and applied to a diverse subject pool of 36 adults. Analysis is focused on two key aspects of vocal tract structure: the midsagittal shape of the hard palate and the posterior pharyngeal wall. Palatal morphology varies widely in its degree of concavity, but also in anteriority and sharpness. Pharyngeal wall morphology, by contrast, varies mostly in terms of concavity alone. The distribution of morphological characteristics is complex, and analysis suggests that certain variations may be categorical in nature. Major modes of morphological variation are identified, including their

relative magnitude, distribution and categorical nature. Implications of these findings for speech articulation strategies and speech acoustics are discussed.

1.2 Introduction

Vocal tract morphology is a fundamental consideration in characterizing the human speech production system because, as with any motor system, the physical size and shape of structures that comprise the vocal tract underlie many aspects of articulation and control. Morphology has additional importance for the speech production system due to its role in shaping speech sounds. The vocal tract's acoustical properties (e.g., resonant characteristics) are determined by its shape, which is determined not only by *active* shaping and articulation, but also by the vocal tract's *inherent* morphology. At the same time, morphology varies widely across individuals, which has at least two major implications. First, morphological variation is a potential source of variability in both the articulatory and acoustic domains. Second, a detailed understanding of morphological differences between individuals can facilitate fresh insights into many aspects of inter-speaker variability, speech motor control and speech production modeling.

Many studies (e.g., [Hiki and Itoh, 1986]) have observed differences in palatal concavity (i.e., whether the palate is flat or has a high, domed shape), but little beyond concavity has been noted or quantified. Even with this basic understanding of morphological differences, it has become clear that palate shape influences many aspects of speech production, particularly for coronal consonants [Fuchs et al., 2006]. Many aspects of sibilant fricative articulation are related to palate shape, including laminal versus apical articulation [Dart, 1991], medial groove formation [McCutcheon et al., 1980] and tongue placement strategies [Toda, 2006, Weirich and Fuchs, 2011]. Moreover, when palate shape is artificially altered, articulation of sibilant fricatives has been

shown to adapt over time [Baum and McFarland, 1997, Honda et al., 2002, Thibeault et al., 2011]. Sonorant articulation is also variable depending on whether the palate is domed or flat in shape. [Tiede et al., 2005] demonstrated that altering palate shape with a prosthesis can switch subjects from producing "bunched" to "retroflex" American English /r/. Speakers with flat palates have been shown to exhibit less articulatory variability during vowel production than speakers with domed palates [Perkell, 1997, Mooshammer et al., 2004, Brunner et al., 2005, 2009]. Vowel production also adapts over time to artificial changes in palate shape [Brunner et al., 2007]. These changes are likely due to the fact that palate shape alters the resonant properties of the vocal tract, particularly for high front vowels [Lammert et al., 2011a].

Most attention toward morphological variation in the vocal tract has been focused on overall length and proportions along a single dimension defined by the midsagittal vocal tract midline, from the lips to the glottis. Overall length of the vocal tract varies significantly through development and between adult individuals [Fant, 1960, Vorperian et al., 2005, 2009]. Proportions of the vocal tract also vary, particularly the relative length of the oral and pharyngeal cavities [Chiba and Kajiyama, 1941, King, 1952, Fitch and Giedd, 1999, Vorperian et al., 1999, Arens et al., 2002, Boë et al., 2006, 2008, Lammert et al., 2011b]. The purpose of this study is to provide an in-depth quantification and analysis of key morphological variations orthogonal to the midline in adult speakers. The width of the vocal tract orthogonal to its midline is central to articulatory descriptions of phonetic segments (e.g., vocal tract area functions, manner of articulation, constriction degree), yet morphological variation in this direction has not been extensively investigated. This study focuses specifically on the hard palate and the posterior pharyngeal wall, which determine much of the morphology orthogonal to the midsagittal midline. The hard palate, because it is immovable, constitutes a cornerstone of the articulatory environment in which speech production takes place. The pharyngeal wall is movable,

but is similarly important because of its large size and because its movements during speech are small relative to its size.

Despite several studies showing that hard palatal morphology impacts speech production, very little is known about the extent and variety of morphological variation in that structure. Even less is known about morphological variation of the posterior pharyngeal wall, which may have a related influence on speech production. The current investigation aims to address this gap in knowledge by developing and applying a methodology to automatically determine the principal varieties of shape variation in the hard palate and posterior pharyngeal wall across individuals, referred to here as *modes*, along with the proportion of total observed variance explained by each of these modes. As an illustration, consider the differences in palatal concavity that have been previously observed by researchers (see above). Differences in palatal concavity constitute one possible mode of variation in palate shape, but may not constitute the most prominent mode, and there may be other prominent modes to consider. The proposed methodology addresses these issues, and it does so in data-driven fashion, rather than imposing prior notions regarding what kinds of variation are expected.

Given the modes of shape variation, it is also possible to examine whether categorical distinctions are suggested by the data, independent of any known groups. For instance, do individuals exhibit a tight, unimodal distribution with regard to palatal concavity? If they do, then one can reasonably say what the 'typical' shape is. If, however, speakers exhibit a more complex, multimodal distribution, then it might be better to say that they fall into distinct categories (i.e., that they form clusters). A second set of statistical analysis aims to automatically estimate the distribution of individuals according to the major modes of shape variation, and whether any clusters are indicated by the data.

Because speech is the primary interest of this study, a group of speakers who have no history of speech, language, or hearing pathology is investigated. Any subject who

met this criterion was included in the study, regardless of factors such as race and language background. The motivation for assembling a diverse group was to understand the extent and variety of morphological variations that can still result in normal speech. Many factors influence cranio-facial morphology, including sex [Xue and Hao, 2006], dental pathology [Ishii et al., 2002], race [Morgan et al., 1995, Everekliogl u et al., 2002, Xue et al., 2006, Wu et al., 2010, Wamalwa et al., 2011, Gu et al., 2011], and history of mouthbreathing [Gross et al., 1994, Harari et al., 2010], but normal speech can result in any of these conditions. The primary interest motivating this study is not the sources and correlates of morphological variability, but rather the breadth of morphological variation that exists in a normal-speaking group of individuals, and particularly those morphological variations that may impact speech production.

1.3 Methods

1.3.1 Subjects

A group of 36 healthy adult subjects with no reported history of speech, language, or hearing pathology were considered. The average age of subjects was 27.0 years with a standard deviation of 4.3 years (range between 19 and 37). Subjects included 30 individuals who self-identified their race as White, Non-Hispanic, and 6 Asians. One subject exhibited a Class III malocclusion (a white male speaker of German), and all other subjects showed normal dental occlusion patterns. Subjects were from diverse language backgrounds, including 22 native speakers of American English, 8 native German speakers, 5 native Mandarin speakers, and 1 native speaker of Hindi.

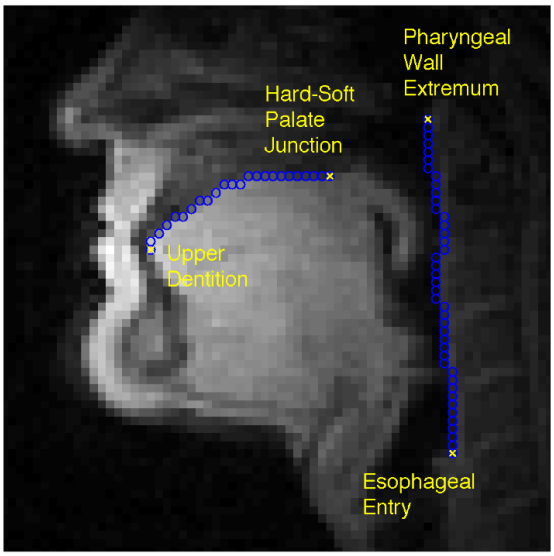


Figure 1.1: Midsagittal image of a male subject used in the analysis. The image shows the subject at rest, with mouth closed and breathing through the nose. Automatically-derived traces of the hard palate and posterior pharyngeal wall have been overlaid, along with anatomical landmarks used to delimit those structures.

1.3.2 Image Acquisition

Midsagittal vocal tract images of all subjects were collected using real-time magnetic resonance imaging (rtMRI) as part of a larger study assessing the explicit connection between variation in the morphological, articulatory and acoustic domains. The use of rtMRI reflects the goals of this larger study, which will require imaging techniques that capture articulatory dynamics and the corresponding acoustic signal in conjunction with each subject's morphological features.

Image acquisition was performed at Los Angeles County Hospital on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha, WI) with gradients capable of 40 mT/m amplitude and 150 mT/m/ms slew rate. A custom 4-channel upper airway receiver coil array, with two anterior coil elements was used for radio frequency (RF) signal reception. A 13-interleaf spiral gradient echo pulse sequence (TR = 6.164 msec, FOV = 200

mm $\times$ 200 mm, flip angle = 15°) was used. The scan slice had a thickness of approximately 5 mm. Resolution of reconstructed images was 68 pixels $\times$ 68 pixels, which equates to a pixel width of approximately 3.0 mm $\times$ 3.0 mm. New image data were acquired at a rate of 12.5 frames per second, and reconstructed using a sliding window technique to produce a video rate of 23.18 frames per second. Further details about the rtMRI image acquisition protocol can be found in [Narayanan et al., 2004]. Data considered in this work were acquired over scan sessions starting in March of 2006 and ending in December of 2011. All work was approved by the University of Southern California institutional review board prior to acquisition.

Images used in this study showed subjects at rest with mouths closed, breathing through the nose. All subjects were instructed to lie comfortably in the scanner in supine position. Subjects' heads were oriented along the midline of the body and padded in place to prevent lateral motion during the scan. The midsagittal plane was localized by real-time examination of slices orthogonal to the midsagittal plane (e.g., an oblique axial slice) [Santos et al., 2004]. By visualizing these planes, localization markers could be placed over landmarks such as the nose tip and the pharyngeal cross-sectional airway and iteratively refined to ensure accurate localization.

A potential confound in studying morphology of the posterior pharyngeal wall is that is can deform somewhat by active articulation and passive conditions in the pharynx. The posterior pharyngeal wall can be actively recruited for swallowing and other functions of the vocal tract [Magen et al., 2003]. It can also deform due to extreme flexion/extension of the neck [Penning, 1988] and due to pressure buildup in the pharynx [Proctor et al., 2010]. To ensure an accurate reflection of the posterior pharyngeal wall's inherent morphology, these sources of deformation were controlled for in several ways. Subjects were imaged during rest position for breathing in order to avoid effects from active articulation or pharyngeal pressure buildup. Flexion/extension of the neck was

controlled for by instructing subjects to lie comfortably in the scanner. Subject comfort has previously been used to define a natural reference position for flexion/extension of the head and neck for studying the shape of the pharynx [Mohammed et al., 1994]. Note that asking subjects to assume a pre-defined amount of flexion/extension (e.g., in terms of degrees) is problematic because (a) it may be uncomfortable for some subjects to hold the pre-defined position and (b) it may not reflect a subjects' natural posture and thereby potentially violate ecological validity. The results section of this paper presents further statistical analysis, related to this point, with the aim of identifying any significant relationship between flexion/extension of the head and the modes of pharyngeal wall deformation.

1.3.3 Image Processing

Five images were identified for each subject, capturing rest position during breathing and with the tongue pressed against the teeth and hard palate. These images were averaged to improve the signal-to-noise ratio and to ensure a representative rest position. Canny edge detection [Canny, 1986] was used with manual linking and correction to trace the hard palate and posterior pharyngeal wall (see Figure 2.2). Traces of the hard palate began at the upper dentition and extended along the palate to the posterior nasal spine (i.e., hard-soft palate junction). Pharyngeal wall traces extended from its highest point in the nasopharynx,down to the entry of the esophagus, a reliable anatomical landmark.

Traces of each structure were aligned at their end-points through rotation, translation and uniform scaling. This allowed each contour to be regarded as a single vector of distance measurements along and perpendicular to the line defined by its end-points. All vectors were subsequently resampled to 100 elements and compiled into the sets $\mathbf{x^{pal}} = \{x_{i=1}^{pal}, \ldots, x_{36}^{pal}\}$ and $\mathbf{x^{phar}} = \{x_{i=1}^{phar}, \ldots, x_{36}^{phar}\}$, for each of the 36 subjects, i.

1.3.4 Analysis

The analysis was designed to be as data-driven as possible, providing a description of the statistical aspects of shape variations present in the data with minimal assumptions and maximum generality. Analysis was aimed at the following aspects of the data: (1) the principal modes of shape variation in the hard palate and posterior pharyngeal wall, and what proportion of the total observed variance can be explained by each of these modes, (2) the distribution of individuals according to the modes of shape variation and (3) any general categorical distinctions in shape (i.e., clusters of speakers) suggested by the data, independent of any known groups.

To address the first question, Principal Component Analysis (PCA) was applied. Given a set of observations, $\mathbf{x}^{\mathbf{phar}}$ and $\mathbf{x}^{\mathbf{pal}}$, PCA finds the orthogonal modes of variation present in the data, and also numerical values representing an individual's shape according to those modes, often called scores. The analysis is defined such that each successive mode accounts for as much of the variance as possible, and such that the proportion of the variance accounted for by each mode can be calculated. Moreover, because the largest few modes account for most of the variance in the data, one can describe complex shapes using the scores from only a small number of modes. Thus, PCA directly addresses the first question, and also facilitates further analyses by consideration of the scores.

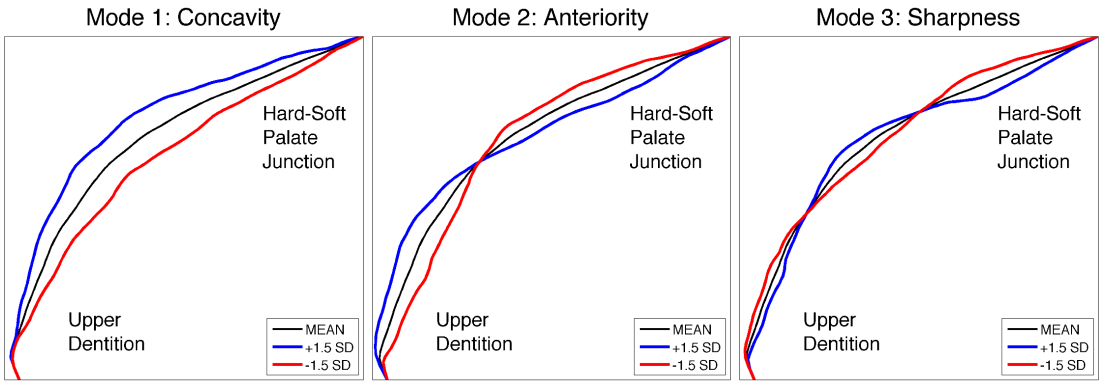
The second question requires accurate estimation of the probability distribution of individuals according to the largest modes of variation. Distributions were estimated by employing Kernel Density Estimation, using a Gaussian kernel to estimate the probability density at 100 points. The width of the Gaussian kernel, corresponding to the standard deviation of the Gaussian, was set to 0.3σ, where σ is the standard deviation of the specific feature in question.

The third question is best addressed through cluster analysis. Clusters were found by applying the K-means algorithm to the PCA scores. All cluster optimizations were done with random centroid initializations and 100 repetitions to avoid convergence to a local minimum (the lowest-cost solution was selected). In choosing the number of clusters, a size constraint was imposed such that all clusters were required to have more than four individuals (i.e., 10% of the subject pool). Individuals were clustered into the largest number of clusters that did not violate this size constraint. For both palate shapes and pharyngeal wall shapes, the appropriate number of clusters, according to these criteria, was precisely three.

1.4 Results

1.4.1 Hard Palate

The major modes of hard palate variation, as suggested by the data, can be seen in Figure 2.1. The three largest modes are shown, which together account for over 85% of the variance in the data. Moreover, these modes seem to have easy interpretations in terms of their physical meaning. The first mode, accounting for 51% of the variance in the data, represents the degree of concavity of the palate (i.e., whether it is flat or domed). The second mode, which accounts for another 25% of the variance, is related to the anteriority of the palate: whether the apex of the dome is positioned toward the anterior or posterior portion of the oral cavity. An additional 10% of the variance can be attributed to the sharpness/flatness of the palate at its apex. These modes will be referred to, respectively, as concavity, anteriority and sharpness for the remainder of the discussion of palatal variation. Figure 1.3 shows images of individuals who represent the extremes of these three modes.



(a) 51% of variance (b) 25% of variance (c) 10% of variance

Figure 1.2: The three largest modes of variation in hard palate shape, determined in completely data-driven fashion, without imposing any prior notions about expected shape variations, by applying PCA to the observed hard palate shapes from the subject pool. Modes reflect differences in concavity, anteriority of the apex, and sharpness of the palate around the apex. The overall mean hard palate shape is shown in black, and the blue and red lines show the nature of deviations from the mean according to each mode. The magnitude of the deviations shown reflect the magnitude of variations seen in the subject pool, at precisely +/- 1.5 standard deviations from the mean shape. Because these modes account for over 85% of the overall variance, arbitrary hard palate shapes may be well-represented using only these three modes.

The distribution of individuals according to concavity, anteriority and sharpness can be seen in Figure 1.4. All three distributions appear to be bimodal, with both modes being approximately equally likely. Subjects are distributed most broadly according to palatal concavity, slightly less broadly according to anteriority and even less broadly according to the palatal sharpness. This pattern corresponds closely to the proportion of variance accounted for by each kind of variation. Moreover, the presence of multiple modes exhibited by all three distributions indicates that hard palate shapes may naturally separate into categories, which can be found by applying cluster analysis.

Cluster analysis revealed three categories of palate shapes. The mean shapes for all individuals in each cluster are visualized in Figure 1.5. These clusters can be interpreted as comprising individuals with (1) concave palates, (2) flat, anterior palates and (3) flat,

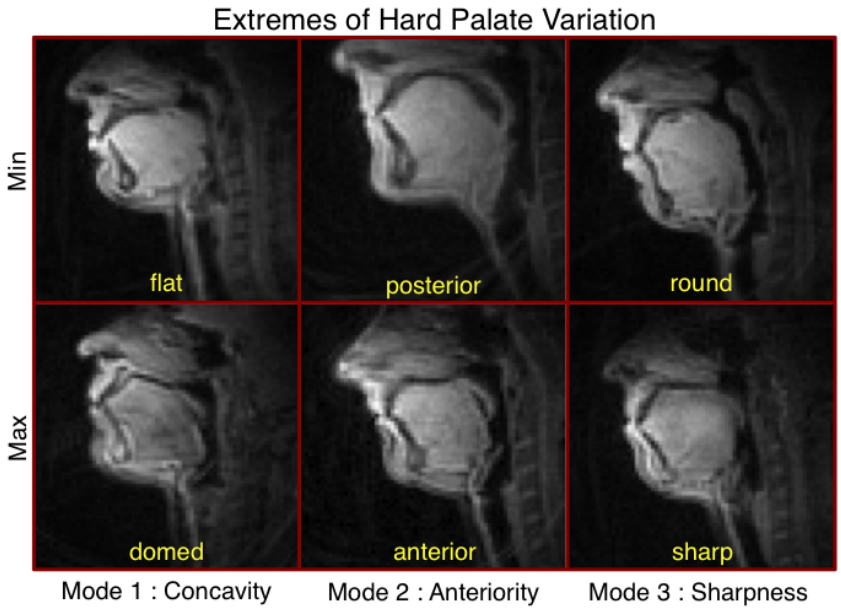


Figure 1.3: Midsagittal images of subjects representing the extremes of each mode of variation in hard palate shape. Modes reflect differences in palatal concavity, anteriority of the palatal dome's apex, and sharpness of the palate around its apex.

posterior palates. Approximately half of the subjects fell into the first cluster, containing concave palates. Remaining subjects were split between the other two clusters.

1.4.2 Posterior Pharyngeal Wall

The major modes of posterior pharyngeal wall variation can be seen in Figure 1.6. The two largest modes are shown, which together account for over 82% of the variance in the data. Similar to the palate shapes, the largest mode of variation in the pharyngeal wall is related to the degree of concavity (75% of the total variance). A much smaller second mode, accounting for an additional 7% of the variance, reflects differences in the inclination of the pharyngeal wall, from fairly vertical to forward-leaning. These modes will be referred to, respectively, as concavity and inclination for the remainder of the discussion of pharyngeal wall variation. Images of individuals who represent the extremes of these two modes can be seen in Figure 1.7.

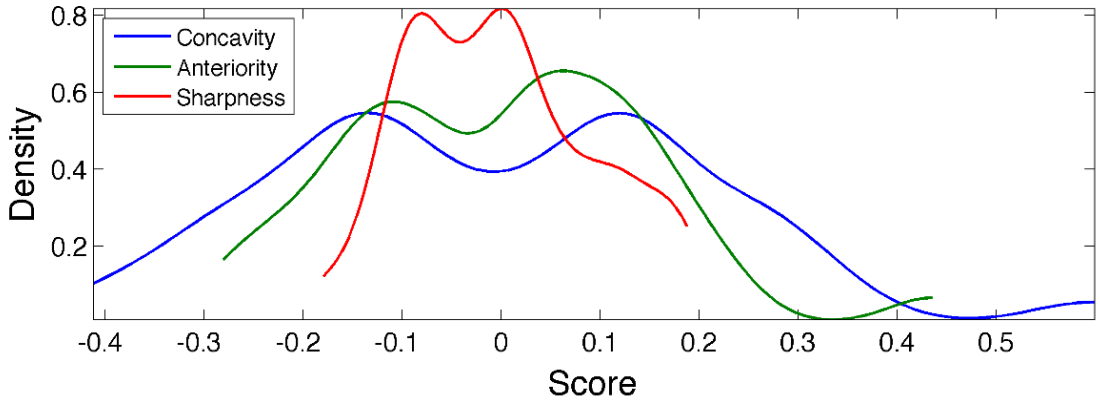


Figure 1.4: Distribution of hard palates according to the three largest modes of shape variation. Abscissa values represent scores derived from PCA, with a value of 0 representing the hard palate mean shape. Modes accounting for more of the variance in the data (e.g., concavity) display a broader distribution. The presence of multiple modes, exhibited by all three distributions, indicates that hard palate shapes may naturally separate into categories.

Further analyses were run to establish that the observed differences in pharyngeal wall shape reflected inherent morphological differences and not differences due to neck flexion/extension. Neck extension was estimated by drawing one line each through the palate and pharyngeal wall endpoints and calculating the angle between those lines. Previous research has indicated that flexing/extending the head across a wide range (40 degrees, centered on a comfortable, neutral posture) has little effect on key upper airway dimensions [Mohammed et al., 1994]. It was found that rest positions varied by only 21 degrees (from 64 to 85 degrees) across subjects in this study. Correlation coefficients were calculated between neck extension and pharyngeal wall shape (i.e., the first two principal modes). Correlation between neck extension and pharyngeal wall concavity was not statistically significant (Pearson's $r = -0.01$, $p = 0.96$), making it very likely that this mode of variation reflects inherent differences in morphology. Correlation with inclination, on the other hand, approached significance (Pearson's $r = 0.31$, $p = 0.07$).

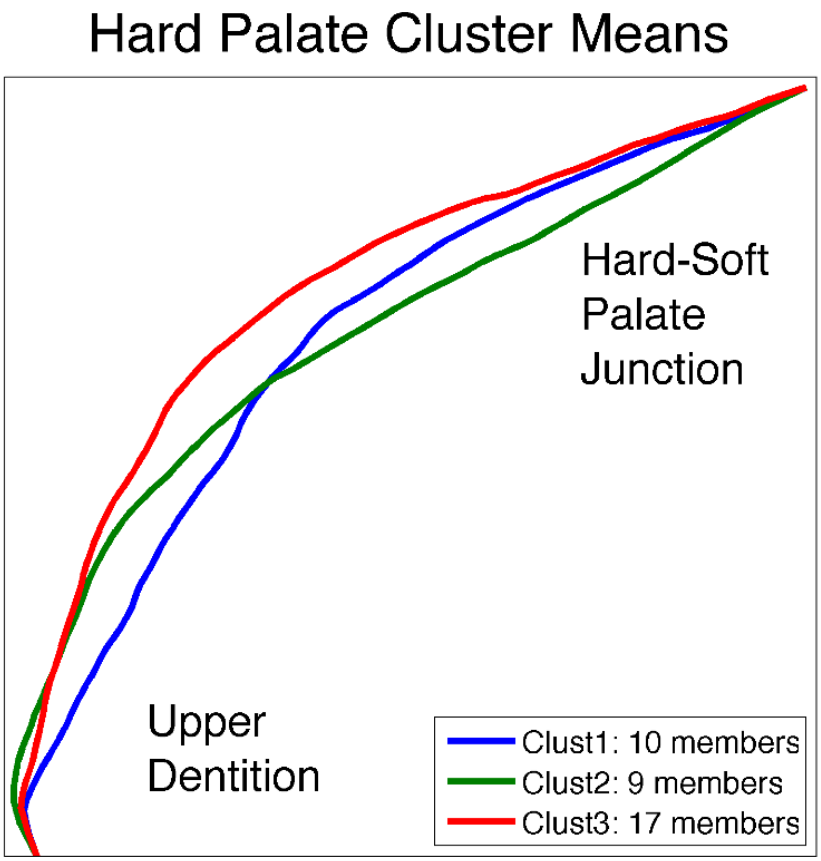
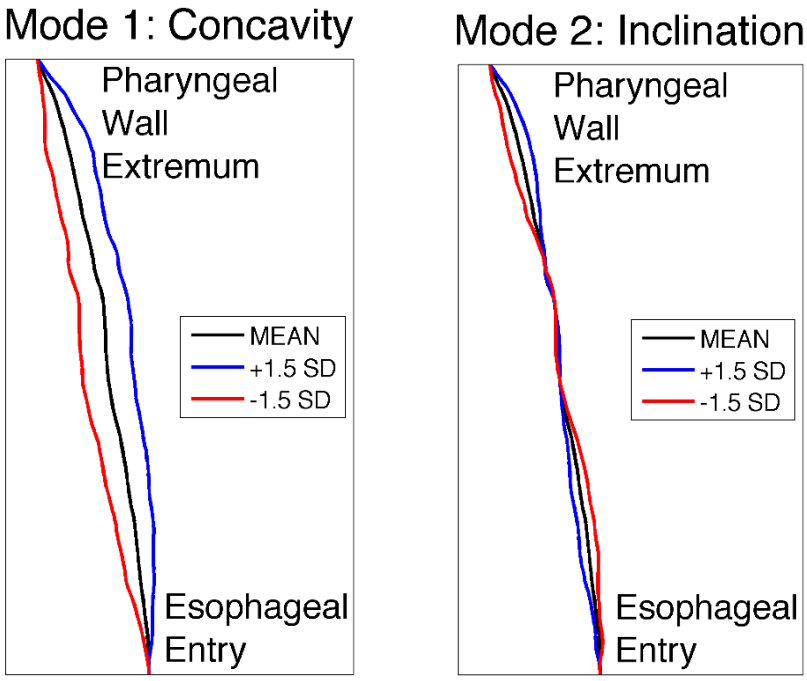


Figure 1.5: Hard palate shapes representing the three categories of hard palate shape, determined in completely data-driven fashion, by applying K-means cluster analysis to the observed hard palate shapes from the subject pool. The displayed hard palates reflect the mean shape of all hard palates contained within one cluster. Clusters can be interpreted as comprising (1) concave palates, (2) flat, anterior palates and (3) flat, posterior palates.

Based on this result, it is conceivable that the observed differences in posterior pharyngeal wall inclination were, at least in part, caused by neck flexion/extension.

The distribution of individuals according to concavity and inclination can be seen in Figure 1.8. Concavity exhibits a very broad distribution, which appears to be trimodal. Moreover, the left two modes of this distribution are both more likely than the rightmost mode. Inclination, on the other hand, displays a highly peaked, unimodal distribution,



(a) 75% of variance

(b) 7% of variance

Figure 1.6: The two largest modes of variation in posterior pharyngeal wall shape, determined in completely data-driven fashion, without imposing any prior notions about expected shape variations, by applying PCA to the observed pharyngeal wall shapes from the subject pool. Modes reflect differences in concavity and inclination of the pharyngeal wall. The overall mean pharyngeal wall shape is shown in black, and the blue and red lines show the nature of deviations from the mean according to each mode. The magnitude of the deviations shown reflect the magnitude of variations seen in the subject pool, at precisely +/- 1.5 standard deviations from the mean shape. Because these modes account for over 82% of the overall variance, arbitrary pharyngeal wall shapes may be well represented using only these two modes.

centered about the mean. The presence of multiple modes, exhibited by concavity, indicates that hard palate shapes may naturally separate into categories, which can be found by applying cluster analysis.

Cluster analysis revealed three categories of pharyngeal wall shapes. The mean shapes for all individuals in each cluster are visualized in Figure 1.9. These clusters can be interpreted as comprising individuals at various levels of concavity, from very straight, to slightly concave, to extremely concave. Approximately 45% of the subjects

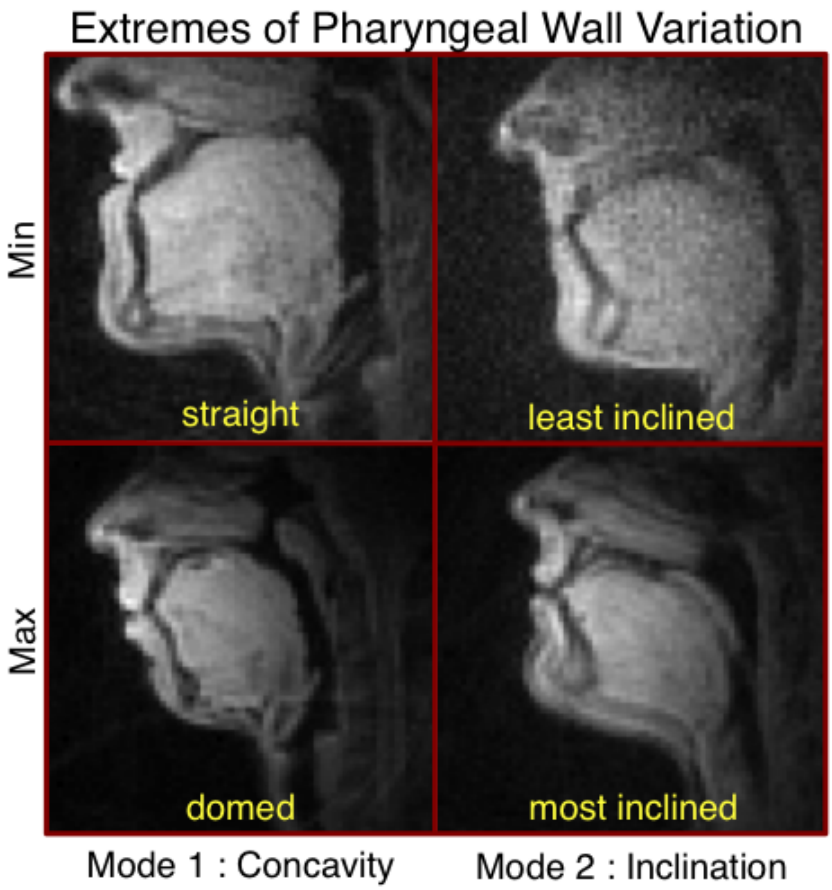


Figure 1.7: Midsagittal images of subjects representing the extremes of each mode of variation in posterior pharyngeal wall shape. Modes reflect differences in pharyngeal wall concavity and inclination of the pharyngeal wall.

had very straight pharyngeal walls, while fewer were slightly concave (33%), and even fewer had extremely concave pharyngeal walls (22%).

1.5 Discussion

A methodology has been proposed for detailed statistical analysis of morphological differences in the vocal tract in which analysis is largely automatic and data-driven. The advantage of a data-driven approach is that it allows the data to directly express the

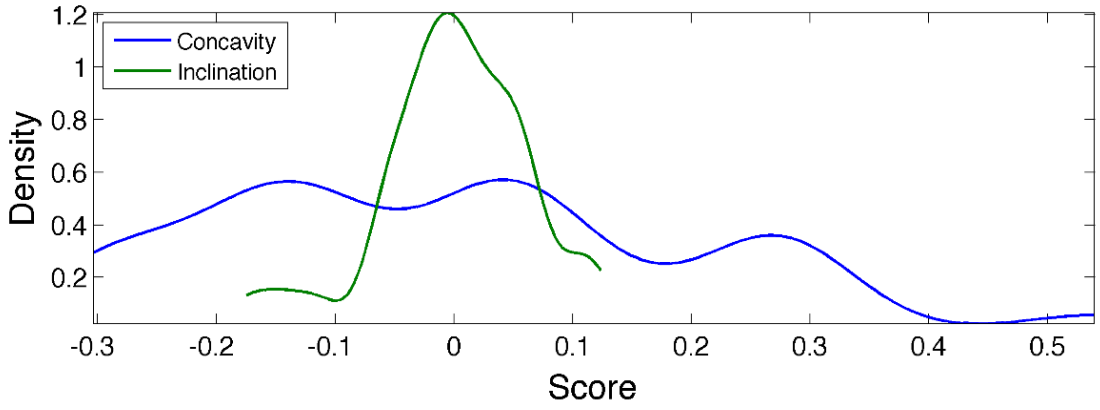


Figure 1.8: Distribution of posterior pharyngeal walls according to the three largest modes of shape variation. Abscissa values represent scores derived from PCA, with a value of 0 representing the pharyngeal wall mean shape. Modes accounting for more of the variance in the data (e.g., concavity) display a broader distribution. The presence of multiple modes, such as that exhibited by concavity, indicates that pharyngeal wall shapes may naturally separate into categories.

variety and extent of variation, rather than imposing prior notions of expected variations. It is notable that the analyses here revealed highly interpretable structure in the data because the results of a data-driven approach can sometimes suffer from poor interpretability. For instance, with respect to the major modes of shape variation in the hard palate and posterior pharyngeal wall, a large majority of the variance can be cleanly interpreted.

Substantial variation was observed in the degree of concavity of the hard palate, which accounted for more of the variance than any other single mode. This reinforces the observations of previous studies (e.g., [Hiki and Itoh, 1986, Brunner et al., 2009]), which noted that palatal concavity is a major source of morphological variation. Two additional modes of variation were found orthogonal to concavity that accounted for substantial amounts of variability in the data. Additional dimensions include the anterior-posterior position of the apex of the palatal dome, and the sharpness/flatness of the palatal dome shape around that apex. The diversity of hard palate morphology

observed in these data may have important implications for articulation strategies across individuals. Anteriority of the palatal inflection has the potential to affect place of articulation for all coronal segments. Roundness or sharpness of the palate could affect the details of tongue shaping for production of coronal fricatives. Future work will focus on these kinds of effects.

Concavity differences also constitute the largest mode of morphological variation in the posterior pharyngeal wall. Unlike the hard palate, however, these concavity differences account for the vast majority of the observed variance, with much less contribution from additional modes. The next largest mode – vertical inclination of the pharyngeal wall – accounts for an order of magnitude less variation compared to concavity, and may simply be related to differences in head flexion/extension. Differences in pharyngeal wall concavity have been shown to impact vowel production by determining the width of the pharynx, and consequently the resonant properties of the vocal tract, especially for low back vowels [Lammert et al., 2011a]. There may be additional consequences for maintaining pressure gradients in the pharynx which would affect voicing, particularly for voiced fricatives where pressure buildup must be carefully controlled. For languages with pharyngeal and emphatic consonants (e.g., Semitic, Afro-Asiatic), place of articulation may also be impacted.

The current data set suggests that variations in shape may be categorical, tending to cluster into specific shape classes. For instance, hard palate shapes reliably cluster into three categories: (1) highly domed palates, (2) flatter palates, for which the small dome is more anterior, and (3) flatter palates, for which the small dome is more posterior. Posterior pharyngeal wall shapes also cluster into three categories, mostly related to concavity: (1) very straight pharyngeal walls, (2) slightly concave pharyngeal walls and (3) extremely concave pharyngeal walls. Some of these categorical differences in

morphology may be accompanied by categories in the articulatory and acoustic domains, which will be investigated in future work.

As previously mentioned, most attention toward morphological variation in the vocal tract has been focused on overall length and proportions along a single dimension defined by the vocal tract midline. The interplay of acoustical and articulatory variability with respect to these differences has been of value in the domain of speech research for studying longstanding questions related to inter-speaker variability [Vilain et al., 1999, Fuchs et al., 2008, Nissen and Fox, 2009], goals of speech production [Ménard et al., 2007], speech acquisition [Ananthakrishnan, 2011] and motor control [Winkler et al., 2006, 2011a]. The current analysis may facilitate similar investigations into the impact of palate and pharyngeal wall structure on articulation and acoustics.

The ultimate goal of this line of research is to assess the impact of morphological variation on speech articulation and acoustics. Examining the relationships between variations in morphology, articulation and acoustics holds promise for explaining inter-speaker variability in production patterns. It should be possible to predict production patterns from observations about an individual's palatal and pharyngeal morphology. Moreover, a fundamental analysis of the speech production system's physical structure (e.g., morphology) can act as a foundation for understanding many aspects of speech motor control. Effective control demands detailed knowledge of structure, which implies that modeling of control would benefit from such knowledge, as well. Many additional questions may also be examined using morphological knowledge, such as the longstanding debate over the nature of speech production goals. The extent to which speakers minimize differences in the articulatory versus acoustic domains can offer insight into the goals of production. By finding ways to quantify and analyze morphological differences, the current study constitutes an important step toward achieving these larger goals. Future research will address those goals by combining the current

analyses with the articulatory information afforded by rtMRI and the noise mitigated audio that was recorded in synchrony with the articulatory data [Bresch et al., 2006].

There are many aspects of morphological variation remaining to be studied in more detail. The morphology of movable structures (e.g., the tongue and lips) should be of particular importance for patterns of articulation. Studying these structures poses serious practical and theoretical challenges, including the need to define a reference posture as a basis for comparing morphology. Detailing morphology off the midsagittal plane – and especially 3-dimensional morphology – is also of major importance. Studying the connection between 3D morphology and articulation will be crucial, but also poses practical challenges given the limitations of current real-time imaging. Work also remains in terms of identifying systematic correlates of morphological variation, both ontogenetic and hereditary. This kind of understanding may make it possible to robustly predict production patterns and explain inter-speaker variability. Finally, future study could benefit from an even more diverse subject pool, to more accurate estimate the full range and variety of morphological differences that can result in normal speech patterns.

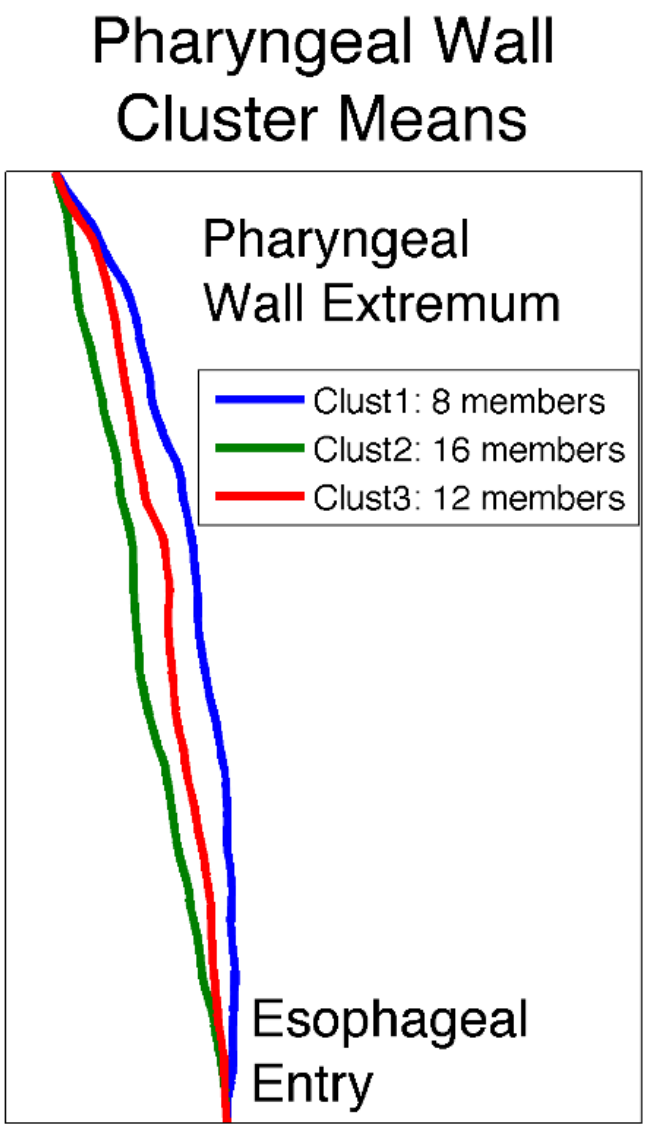


Figure 1.9: Posterior pharyngeal wall shapes representing the three categories of pharyngeal wall shape, determined in completely data-driven fashion, by applying K-means cluster analysis to the observed pharyngeal wall shapes from the subject pool. The displayed pharyngeal walls reflect the mean shape of all pharyngeal walls contained within one cluster. Clusters can be interpreted as comprising shapes of increasing concavity, from very straight, to slightly concave, to extremely concave.

Chapter 2

Interspeaker Variability in Hard Palate Morphology and Vowel Production

2.1 Abstract

Differences in vocal tract morphology have the potential to explain interspeaker variability in speech production. The potential acoustic impact of hard palate shape is examined in simulation, in addition to the interplay between morphology, articulation and acoustics in real vowel production data. High front vowel production from five speakers of American English is examined using midsagittal real-time magnetic resonance imaging data with synchronized audio. Relationships between hard palate morphology, tongue shaping and formant frequencies are analyzed. Simulations are performed to determine the acoustical properties of vocal tracts whose area functions are altered according to prominent hard palate variations. Simulations reveal that altering the height and position of the palatal dome alters formant frequencies. Examinations of real speech data show that palatal morphology is not significantly correlated with any formant frequency, but is correlated with major aspects of lingual articulation. Certain differences in hard palate morphology can substantially affect vowel acoustics, but those effects are not noticeable in real speech. Speakers adapt their lingual articulation to accommodate palate shape differences with the potential to substantially affect formant frequencies, while ignoring palate shape differences with relatively little acoustic impact, lending support for acoustic goals of vowel production.

2.2 Introduction

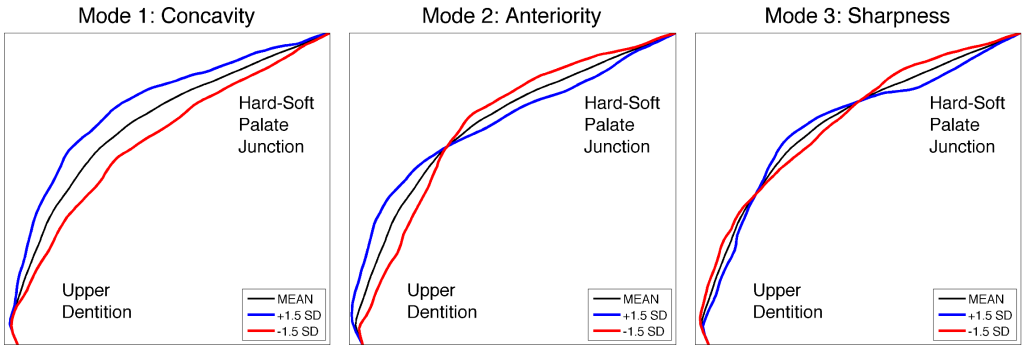
Speech production research has long been concerned with explaining variability in the acoustic and articulatory domains, both within and across speakers. An essential consideration for explaining this variability is vocal tract morphology. The morphology of vocal tract structures is fundamentally linked to speech articulation and acoustics through a complex interplay. The overall shape of the vocal tract – so crucial for determining its acoustical properties (e.g., resonant characteristics) – is determined not only by actively controlled shaping mechanisms, but also by the inherent shape of vocal tract components. Elements of vocal tract morphology that vary across speakers should then result in some combination of articulatory and acoustic differences across those same speakers, even when producing identical segments. Achieving the same acoustic output implies articulatory differences in compensation for morphological variation, unless those specific articulatory differences happen to not result in acoustic differences. Conversely, if speakers do not differ in their articulations, then their morphological differences will be manifested in the acoustics. Several key questions arise from this interplay. First, do certain active articulatory strategies reflect morphological characteristics? Also, which morphological differences have the potential to affect the acoustics? Finally, which aspects of morphology are evident in the acoustic signal?

Perhaps the most extensively studied aspect of the structure-function interplay concerns the role of vocal tract length in vowel production variability. Vocal tract length varies considerably through ontogenesis, where its average length doubles from around 8 cm to 16 cm [Fitch and Giedd, 1999, Vorperian et al., 2005, 2009]. Adult vocal tracts also vary substantially across individuals, ranging from approximately 13 cm to as many as 20 cm. Lengthening the vocal tract has the potential to lower all formant frequencies [Fant, 1960, Stevens, 1998] and this effect is observed in real vowel acoustics [Peterson

and Barney, 1952, Lee et al., 1999]. Differences in articulation cannot easily compensate for this acoustic effect, should speakers try, except to a limited extent through lip protrusion and laryngeal raising. In addition, articulatory strategies may be influenced by the proportional length of the oral to pharyngeal cavities. The pharynx becomes proportionally longer through the course of development [Chiba and Kajiyama, 1941, King, 1952, Arens et al., 2002] and adult speakers display sexual dimorphism, such that the pharynx of males is proportionally longer [Vorperian and Kent, 2007, Vorperian et al., 2011]. Differences in proportions can influence articulatory strategies by forcing speakers into specific production patterns [Winkler et al., 2006, Ménard et al., 2007, Fuchs et al., 2008, Nissen and Fox, 2009, Winkler et al., 2011a], and are also evident in the acoustics as vowel-specific scaling of acoustics with vocal tract length [Fant, 1966, Nordström, 1975, Fant, 1975].

Studies have investigated the role of hard palate morphology in speech production, as well, which is also the focus of this work. Hard palate shape varies in three prominent ways: the height of the palatal dome, the position of the dome's apex in the oral cavity and the angularity of the dome around the apex, as shown in Figure 2.1 [Lammert et al., 2013]. Several studies have shown that the first of these variations – whether the palate is highly domed or relatively flat – has a substantial impact on articulatory strategies. Speakers with flat palates have been shown to exhibit less articulatory variability during vowel production than speakers with domed palates [Perkell, 1997, Mooshammer et al., 2004, Brunner et al., 2005, 2009]. Apical vs. laminal articulation of sibilant fricatives also depends on palate shape [Dart, 1991]. Furthermore, artificially flattening palate shape forces corresponding changes in jaw height and the positioning of the tongue during coronal fricative production [Honda et al., 2002, Thibeault et al., 2011], which quickly minimizes acoustic differences [Baum and McFarland, 1997]. Brunner et al. [2007] showed that vowel articulation also adapts over time to artificial changes

in hard palate shape, suggesting that palatal morphology accounts for some vowel articulation variability across speakers. However, the specific aspects of vowel articulation and acoustics that are affected by morphology have not been rigorously investigated. It is clear that differences in palatal morphology have the potential to affect the resonant properties of the vocal tract and thereby alter the acoustic output [Lammert et al., 2011a], but whether articulation varies in compensation for palatal differences has not been well studied.



(a) 51% of variance (b) 25% of variance (c) 10% of variance

Figure 2.1: The three largest modes of variation in hard palate shape, previously determined. Modes reflect differences in concavity, anteriority of the apex, and sharpness of the palate around the apex. The overall mean hard palate shape is shown in black, and the blue and red lines show the nature of deviations from the mean according to each mode. The magnitude of the deviations shown reflect the magnitude of variations seen in the subject pool, at precisely +/- 1.5 standard deviations from the mean shape. Because these modes account for over 85% of the overall variance, it is possible to well-represent arbitrary hard palate shapes using only these three modes. This figure was reproduced from Lammert et al. (2013).

The present paper is focused on the interplay between hard palate morphology and vowel production behavior. Specifically, interspeaker variations in hard palate morphology are considered toward explaining variability in vowel articulation and acoustics. The investigation involves two parts that provide complementary insights: (1) analysis of real speech data to assess the extent of systematic relationships between morphological variation and vowel production variation, and (2) acoustic vocal tract simulations to

assess the potential of different morphological variations to impact the acoustics. The first part is essential for analyzing the interplay in question, but it does not provide a complete picture. With real speech data, it is not possible to factor out the effect of articulation, which creates ambiguity in the interpretation of some results. For instance, articulatory variability might be viewed as an attempt by speakers to minimize acoustic variability in the face of morphological differences. But, without knowing what acoustic variability results from those morphological differences *irrespective* of the accompanying articulatory variability, it is not possible to draw that conclusion with certainty. It may simply be that the particular morphological differences between those speakers have no acoustic impact. Scenarios like this make it necessary to consider simulations alongside analysis of real data.

The present paper reflects an expansion of initial work to quantify morphological variations in the hard palate across speakers using real-time magnetic resonance imaging (rtMRI) data [Lammert et al., 2011b]. Subsequent work has examined the nature of morphological variation in this crucial structure [Lammert et al., 2013], and the theoretical consequences of those variations on vowel acoustics [Lammert et al., 2011a]. The current work comprises an investigation of the palate's impact on speech production using rtMRI paired with complementary acoustic simulations. By combining parametric analysis of palatal morphology with acoustic simulations of different vocal tract shapes, the effect of palatal variation on vocal tract resonances is studied. Palate shapes, lingual articulation and formant frequencies of several American English speakers are also examined to examine their interplay. Finally, the relevance of this work to current understanding of morphological variation, interspeaker variability and goals of speech production is discussed.

2.3 Method

2.3.1 Speech Data

Articulatory and acoustic data were taken from five male speakers in the recently-collected MRI-TIMIT corpus [Narayanan et al., 2011]. The MRI-TIMIT corpus is a collection of real-time magnetic resonance imaging (rtMRI) data [Narayanan et al., 2004] of continuous read speech from ten native speakers of American English. Images were reconstructed using a sliding-window procedure with a step size of one TR (= 6.164 msec), resulting in an effective frame rate of 162.23 f.p.s. with a spatial resolution of 68 $\times$68 pixels over 20$\times$20 cm (approximately 2.9 cm pixel width). These images show full midsagittal views of the subjects' upper airways, including articulatory dynamics and morphological characteristics in the midsagittal plane. Speech acoustics were simultaneously recorded using an optical microphone, and subsequently processed according to the method described by Bresch et al. [2006] in order to remove scanner-generated audio noise.

For each of the five subjects, five tokens were selected of the high-front vowel /i/ as spoken during the word *people* [1]. Tokens produced in interlabial contexts were chosen in order to minimize lingual co-articulatory effects on the vowel of interest. Use of high-front vowels was motivated by previous modeling work, which suggested that high-front vowels emphasize any potential effect of palate morphology on formant frequencies differences [Lammert et al., 2011a], presumably because of the relative narrowing of the vocal tract in the palatal region. Each vowel token was considered to extend from the first and last peak of periodicity in the acoustic signal. Three points in each vowel were marked for further analysis, corresponding to 25%, 50% and 75% of the total vowel interval.

[1] Subject mm2 had only three usable tokens, due to problems at acquisition time.

Formant frequencies were estimated from the LPC spectrum, calculated over a 25 ms window centered at each specified time point. Positions of the three lowest peaks in the spectrum were identified as formants one through three. The mean formant values across all measurement points in a given vowel were used to represent the acoustics for that token. Linear prediction order was initialized to 14 and, in situations where this order failed to return formants in the broad frequency ranges expected for a high-front vowel (200–500 for F1, 1800–2500 for F2 and 2300–4000 for F3, with no bandwidth criteria), the order of analysis was refined using LPC models varying in order between 12 and 18 until a set of three formants could be robustly identified. The order number closest to 14 that returned formant values in the specified ranges was taken to be the optimal one.

For each token, a composite semi-polar analysis grid was superimposed onto the midsagittal plane, extending from the glottis to the lips with gridlines spaced at approximately 5 mm intervals [Öhman, 1967, Maeda, 1979]. An example grid for one subject can be seen in Figure 2.2. The grid was manually positioned relative to four anatomical landmarks: the glottis, the highest point on the palate, the alveolar ridge and the lips. Proctor et al. [2010] described this placement method, along with a technique for automatically tracing the vocal tract outlines in rtMRI data by identifying air-tissue boundaries intersecting with the gridlines. This tracing method was used to produce midsagittal vocal tract outlines, which were subsequently inspected for accuracy and manually corrected when necessary.

Using the midsagittal vocal tract outlines, palate traces, tongue traces and midsagittal distance functions were extracted. Palate traces were defined as the segment of the upper vocal tract outlines extending from the gridline passing closest to the upper dentition to the gridline passing closest to the hard-soft palate junction (i.e., the posterior

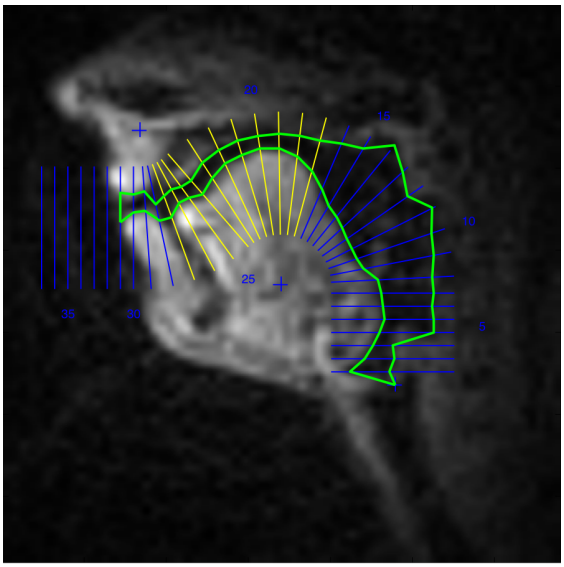


Figure 2.2: Midsagittal image of one male subject that was used in the analysis. The image shows the subject during production of a high-front vowel. Automatically-derived traces of the vocal tract outlines have been overlain, along with the gridlines used for analysis.

nasal spine). This definition was used to be consistent with the traces identified by Lammert et al. [2013]. These same gridlines bounding the palate were also used to delimit the relevant tongue traces for articulatory analysis. Tongue traces were extracted from three-frame intervals centered at the time points marked for acoustic analysis. This provided a total of nine tongue traces for each vowel token, the means of which was used to represent the tongue trace for that token. The overall mean tongue trace, across all five tokens, was used to represent the tongue trace for each subject. Midsagittal distance functions were extracted from the same three-frame intervals as the distance functions, and the means were taken in the same manner to provide a distance function for each subject. These midsagittal distance functions were calculated by finding the distance between the upper and lower outlines along each gridline. The resulting distance functions can be seen in Figure 2.3.

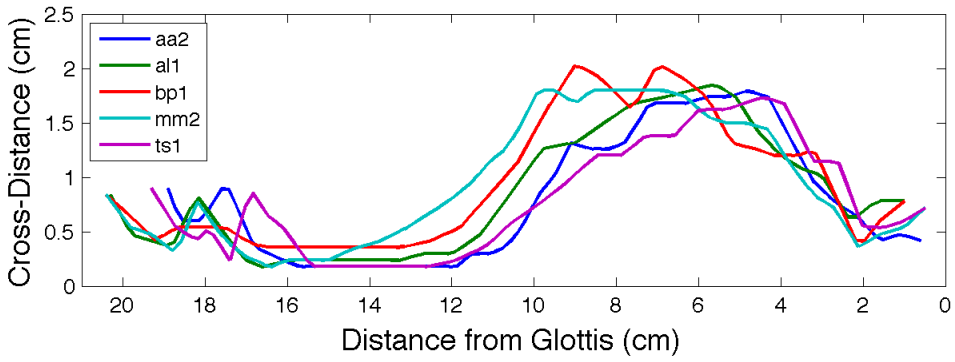


Figure 2.3: Midsagittal distance functions, taken from the five subjects producing high-front vowels. These five distance functions were also used, in addition to one of uniform diameter, as template vectors for in the acoustic simulation experiments, where their palate regions were deformed according to the three major modes of palate shape variation.

Vocal tract length and hard palate length were also measured for each token using the midsagittal vocal tract outlines. Gridlines that intersected with the glottis and the most anterior protrusion of the lips were identified and used to delimit the extent of the vocal tract. The vocal tract midline was then defined as the series of line segments whose endpoints lie along neighboring gridlines, halfway between the outer and inner vocal tract outlines. Vocal tract length was calculated as the distance along the vocal tract midline between these delimiting gridlines. Hard palate length was taken to be the distance, along the vocal tract midline, between the gridlines delimiting the hard palate. Table 2.1 shows the mean vocal tract length and palate length measurements for each subject in this study. These measurements facilitate normalization of formant frequencies for acoustic analysis, as well as subject-specific acoustic simulations, discussed in *Methods:Simulations*.

Measurements of vocal tract length were used to normalize formant frequencies, which otherwise may vary with vocal tract length in addition to lingual articulation. Meaningful comparison of articulation-relevant formant variation is possible only after

Subject ID	Mean Vocal Tract Length (cm)	Palate Limits (cm)	Palate Length (cm)
aa2	18.9	11.1 – 15.7	4.6
al1	20.3	12.0 – 17.0	5.0
bp1	20.4	11.1 – 17.0	5.9
mm2	20.4	11.5 – 17.1	5.6
ts1	19.3	10.5 – 15.4	4.9

Table 2.1: Lengths of relevant vocal tract structures for each subject. Distances were calculated along the midsagittal vocal tract midline in all cases. Limits of the palate are given relative to the position of the glottis, and refer to the position of the posterior nasal spine (posterior) and the upper dentition (anterior), respectively.

factoring out the effect of vocal tract length variation. Having vocal tract length measurements for each token affords very detailed normalization of formant frequencies, accounting for changes in vocal tract length due to, for instance, differences in lip rounding or vocal tract posture. Given measurements of vocal tract length, L_{obs}, and observed formant frequencies, F_{obs}, the normalized formant frequencies are obtained as follows: $F_{norm} = (L_{obs}F_{obs})/17.5$. This procedure scales all formant frequencies to those produced by a 17.5 cm vocal tract.

Descriptions of hard palate and tongue shape were based on the parameterization of hard palate morphology described by Lammert et al. [2013]. A key result of that work was that most of the variation (approximately 85%) observed across subjects could be represented by a small number of modes in shape variation. The three largest modes are shown in Figure 2.1. The first mode, accounting for 51% of the variance in the data, represents the degree of concavity of the palate (i.e., whether it is flat or domed). The second mode, which accounts for another 25% of the variance, is related to the anteriority of the palate: whether the apex of the dome is positioned toward the anterior or posterior portion of the oral cavity. An additional 10% of the variance can be attributed to the sharpness/flatness of the palate at its apex. These modes will be referred to as concavity, anteriority and sharpness, respectively, for the remainder of the present paper.

The first step in this shape parameterization is to align the palate and overall mean tongue shapes of the subjects by their end-points through rotation, translation and uniform scaling (as shown in Figure 2.4). This allows each trace to be regarded as a single vector of distance measurements, along the line defined by its delimiting points (i.e., the perpendicular distance). The resulting vectors for each speaker i are referred to as $\mathbf{P}_i$ and $\mathbf{T}_i$, respectively. A given speaker's palate trace, $\mathbf{P}_i$, can then be approximated as follows:

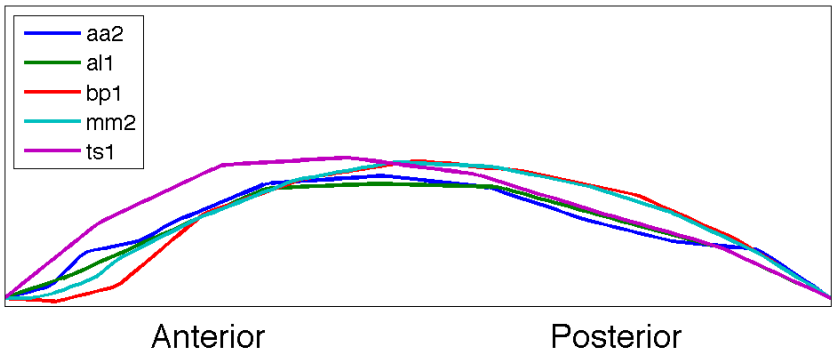
$$\hat{\mathbf{P}}_i = c_i\mathbf{C} + a_i\mathbf{A} + s_i\mathbf{S} + \mathbf{P}_\mu \tag{2.1}$$

where $\mathbf{P}_\mu$ is the overall mean palate shape of all individuals. The vectors $\mathbf{C}$, $\mathbf{N}$ and $\mathbf{S}$ represent the modes of palate shape variation in terms of palatal concavity, anteriority and sharpness, respectively. That is, they represent unit deviations in palate shape according to those three modes. The coefficients, c_i, a_i, s_i, reflect the contribution of each mode of shape variation to the specific palate under consideration. These modes are shown in Figure 2.1, where the coefficient are manipulated independently. Crucially, the coefficients themselves can be used as a low-dimensional parameterization (i.e., three parameters for each individual) of a palate's shape.

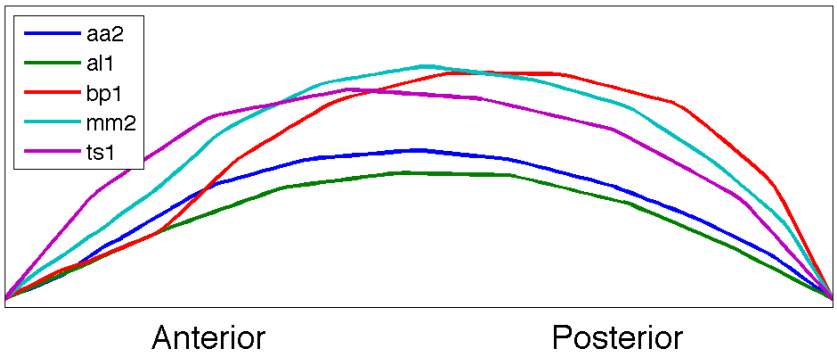
For an novel and arbitrary palate trace, shape parameterization can be done by projecting individual palate shapes into the vector space defined by the modes of variation. For instance, the coefficient representing concavity, c_i, can be calculated for palate i in the following way:

$$c_i = \mathbf{C}^\mathbf{T}(\mathbf{P}_i - \mathbf{P}_\mu) \tag{2.2}$$

This same form of projection can be done to obtain a_i or s_i by substituting either $\mathbf{A}$ or $\mathbf{S}$, respectively, for $\mathbf{C}$ in Equation 2.2. Moreover, this parameterization can be obtained for a speaker's tongue trace in a similar way, as well, by substituting $\mathbf{T}_i$ for $\mathbf{P}_i$. The result of this quantification is three coefficients for each subject that represent the concavity,



(a) Palate Shapes



(b) Tongue Shapes

Figure 2.4: Midsagittal hard palate (a) and tongue shapes (b) of the five subjects, from the upper dentition to the hard-soft palate junction. Hard palates vary substantially in terms of the height of the palatal apex (concavity), the anterior-posterior position of the apex, as well as sharpness of the palate around the apex. Tongue shapes also vary, primarily in terms of concavity and anteriority.

anteriority and sharpness of the palate, and can be used to represent or reconstruct the palate, as shown in Equation 2.1.

To ensure that this parameterization did, in fact, accurately represent palate and tongue traces, the proportion of variance explained by application of the parameterization among the subjects was calculated. It was found that 85.75% of palate shape variance was explained by these three modes of variation, which is highly consistent with the findings presented in Lammert et al. [2013]. Moreover, it was found that 95.13 % of tongue shape variance was explained by the same representation, with 77.83%

represented by concavity, 13.45% by anteriority and 3.86% by sharpness. Note that this high degree of explanatory power for tongue shape is not necessarily expected, since these modes were developed to represent palate shapes. The primary objective of parameterizing tongue contours using this method is simply to facilitate a meaningful comparison between palate shape and tongue shape, but this level of explanatory power indicates that the comparison will be relatively complete, as well.

In order to observe the articulatory and acoustic effects of palate shape, it is necessary that the speakers under consideration exhibit substantial variation along the major modes of palatal variation. The palate shape parameter values were used to quantify the amount of variation in the present sample, and the range of variation in the present sample was compared to the range of values in the much larger and more diverse sample analyzed in Lammert et al. [2013]. It was observed that concavity values in the current sample ranged across 18% of the larger sample's range (0.83 SD), while anteriority ranged across 46% of the range exhibited by that larger sample (2.16 SD), and sharpness ranged across 13% (0.52 SD). These data indicate that palate morphology variation in the present sample is sufficiently large to allow examination of its influence on vowel production within the present speaker population.

2.3.2 Simulations

Vocal tract simulations were based on modeling the vocal tract as a series of lossless, cylindrical, concatenated tubes, which has been extensively studied with respect to vocal tract modeling and synthesis [Fant, 1960, Kelly and Lochbaum, 1962, Rabiner and Schafer, 1978, Stevens, 1998]. Building such an acoustic model can begin by defining a vector, **D**, representing the midsagittal distance at equally-spaced intervals along the length of vocal tract. In this case, the equally-spaced intervals or, equivalently, the

lengths of the tubes were fixed to 2.5 mm in length. Longer vocal tracts were represented by adding more tubes to the model.

For the purposes of acoustic simulation experiments, the tubes can subsequently be deformed according to the modes of palate shape variation utilized here for analysis. In particular, midsagittal distance vector can be represented in the following way:

$$\mathbf{D}_i = \mathbf{M} + \mathbf{F}_i \tag{2.3}$$

where $\mathbf{M}$ is a template vector of midsagittal distances which is fixed for a particular deformation experiment, and $\mathbf{F_i}$ is a deformation vector that can be made to represent arbitrary changes to the template shape. Assuming that the mean palate shape, $\mathbf{P}_\mu$, is already reflected in the template vector, palate shape can be represented as $\hat{\mathbf{P}}_i - \mathbf{P}_\mu$ and the deformation vector can subsequently be represented as $\mathbf{F_i} = [\mathbf{0}, \hat{\mathbf{P}}_i - \mathbf{P}_\mu, \mathbf{0}]$, with zero vectors used to pad the deformation vector on either side of the palate. In the present experiments, each shape coefficient from Equation 2.1 was varied independently, while setting the other two equal to zero. This allowed the acoustic impact of each mode of shape variation to be examined individually.

A total of six template vectors were utilized in the simulation experiments. The first of these vectors simply assumes a uniform vocal tract diameter of 1 cm – i.e., $\mathbf{M} = 1$. This uniform template vector further assumes a 17 cm vocal tract, making 68 total tubes from the glottis to the lips, with the hard palate extending 6 cm in length from 1 cm behind the open end of the tube (i.e., the lips). The other five template vectors were based on the midsagittal distance functions of subjects producing high-front vowels, already discussed in *Methods:Speech Data* and shown in Figure 2.3. These latter five template vectors further utilize the vocal tract lengths and palate positions described in Table 2.1.

In order to calculate the resonance frequencies of this multi-tube model, it is necessary to convert the midsagittal distances into cross-sectional areas. This conversion was done according to the following formula:

$$\mathbf{A_i} = \pi(\mathbf{D_i}/2)^2 \tag{2.4}$$

The assumptions behind this conversion are coarse compared to methods used in recent efforts to accurately model vowel acoustics (e.g., Jackson and McGowan [2012]). However, there are at least two reasons to believe that such assumptions are more appropriate for this study. First, since morphology is the focus here, conversion techniques based on information from other speakers with different morphology may confound the simulations with shape information that is not appropriate to the subjects considered here. Second, assumptions behind any conversion technique must be simultaneously coherent for both the hard palate and the tongue. Even if certain conversion techniques are more accurate for the overall area function, assuming a similar conversion for the hard palate by itself may not be appropriate because parametric descriptions of the hard palate's three-dimensional morphology are not well-known. Therefore, this study remains neutral about three-dimensional hard palate morphology by assuming a uniform projection from midsagittal profile to a normalized acoustic tube model of each speaker's vocal tract.

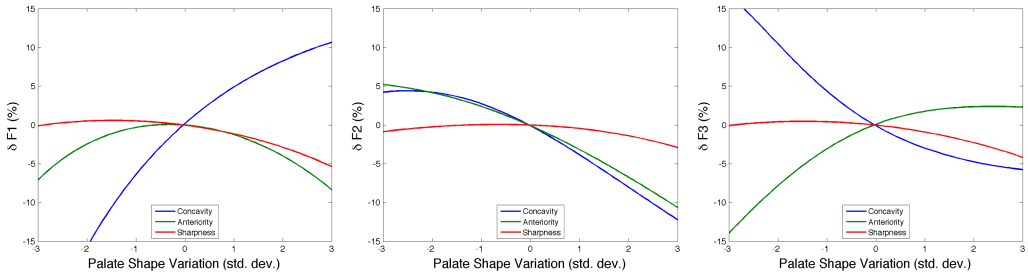
From the area vectors, the formants can easily be computed by first calculating the reflection coefficients between each adjacent tubes: $\Gamma_j = (\alpha_{j+1} - \alpha_j)(\alpha_{j+1} + \alpha_j)$, where α_j is the value of element j in A. Reflection coefficients are then used to compute the coefficients of the prediction filter polynomial. This is done using Levinson's recursion, as described in Kay [1988] and implemented in the Matlab® Signal Processing

Toolbox™ (The MathWorks Inc., version 7.8.0). Finally, the formants can be found by taking the roots of the prediction filter polynomial.

2.4 Results

2.4.1 Simulation

Figure 2.5 shows the results from acoustic simulations using a uniform template vector. Shown are the effects on the first three formant frequencies of varying each mode of palatal variation. Palatal variations along the three principal modes are plotted in terms of standard deviations from the mean shape, where the standard deviations are defined over the distributions described in Lammert et al. [2013]. Specifically, the SD = 0.219 for concavity, SD = 0.154 for anteriority and SD = 0.095 for sharpness. Formant frequencies are plotted in terms of percent frequency change in Hz from the frequency at the mean shape.

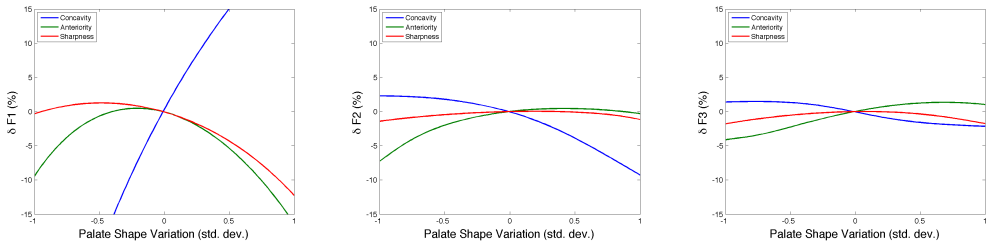


(a) F1 (b) F2 (c) F3

Figure 2.5: The acoustic impact of varying palate shape in a tube of uniform width and area. Palates were varied according to the three major modes of variation: concavity, anteriority and sharpness. Change in frequencies of the first three formants are plotted as a function of variation along these three modes. Palate shape changes are represented in terms of the standard deviations.

Figure 2.6 shows the results from acoustic simulation using a template vector representing the high-front vowel posture from the five subjects in this study. Shown in

the figure is the mean effect – across all five subjects – of varying each mode of palatal variation on the first three formant frequencies. Palatal variations along the three principal modes are plotted in terms of standard deviations from the mean shape, again using the values described in Lammert et al. [2013]. Note that it was not possible to deform the tube through as large a range when beginning with this template vector because deformations near the already narrow constriction in the palatal region quickly lead to negative midsagittal distances. Thus, only +/- 1 standard deviation of deformation are shown. Formant frequencies are plotted in terms of percent frequency change in Hz from the frequency at the mean shape, corresponding to the observed high-front vowel shape.



(a) F1 (b) F2 (c) F3

Figure 2.6: The acoustic impact of varying palate shape in tubes of nonuniform width, averaged across subjects, corresponding to the articulation of a high-front vowel from the five speakers in this study. Palates were varied according to the three major modes of variation: concavity, anteriority and sharpness. Change in frequencies of the first three formants are plotted as a function of variation along these three modes. Palate shape changes are represented in terms of the standard deviations.

2.4.2 Speech Data

Table 2.2 shows Pearson's correlation coefficient between each mode of palate variation and each of the first three normalized formant frequencies. The statistical significance of these values was tested with a two-tailed hypothesis test (n=5) based on Student's t distribution at a a significance level of p = 0.05. None of the correlations are significant

and the correlation values are all rather small. The largest observed correlation value is between palatal anteriority and F3, which displays a negative correlation of -0.80 (p = 0.10).

	F1	F2	F3
P1: Concavity	0.38	-0.41	-0.19
P2: Anteriority	-0.40	0.09	-0.80
P3: Sharpness	-0.05	0.47	0.53

Table 2.2: Correlation values between palate shapes (P) and formant frequencies (F). Palate shapes were parameterized according to the major modes of shape variation. High values reflect systematic relationships between hard palate shape variation and formant frequency variation across subjects. No correlations were statistically significant, as determined using a two-tailed hypothesis test at the p = 0.05 level (n=5).

Table 2.3 shows Pearson's correlation coefficient between each mode of palate variation and each mode of tongue shape variation. The statistical significance of these values was tested with a two-tailed hypothesis test (n=5) based on Student's t distribution at a a significance level of p = 0.05. Results indicate that three correlation coefficients that are largest in magnitude are statistically significant. First, there is a positive correlation between concavity of the palate and concavity of the tongue, and there is also a positive correlation between anteriority of the palate and anteriority of the tongue. In addition, a negative correlation exists between concavity of the tongue and sharpness of the tongue. All other correlations were nonsignificant.

2.5 Discussion

Without any articulatory compensation, variations in palatal concavity and anteriority have the potential to substantially alter the resonant properties of the vocal tract. Simulations indicate that F1 increases with palatal concavity, and that F2 and F3 decrease with concavity. By comparing Figure 2.5 and Figure 2.6, one can see that the magnitude

	T1	T2	T3
P1: Concavity	**0.96**	-0.03	**-0.95**
P2: Anteriority	-0.25	**0.91**	0.18
P3: Sharpness	-0.55	-0.54	0.57

Table 2.3: Correlation values between palate (P) and tongue (T) shapes, according to the major modes of shape variation (numbered 1–3, and labeled at each row). High values reflect systematic relationships between hard palate shape variation and tongue shape variation across subjects. Boldface text indicates statistical significance of a two-tailed hypothesis test at the p = 0.05 level (n=5).

of this effect is somewhat amplified in a high-front vowel posture, when compared to a more uniform vocal tract shape. Any variation in anteriority from the mean palate shape causes F1 to decrease. Anteriority also has a substantial impact on F2 which is posture-dependent: F2 is an increasing function of anteriority for high-front vowel postures and a decreasing function for a neutral posture. Furthermore, F3 increases with anteriority, which is nearly the mirror image of concavity's effect. Note that the effect of sharpness is generally marginal. An effect of sharpness can only be observed on F1 in high-front vowel posture, and only at extremely sharp palate shapes.

Despite the predicted acoustic impact, differences in palatal morphology are not reflected in the human vowel acoustics in any statistically significant way. This apparent discrepancy would appear to be a direct result of the articulatory compensation observed during vowel production. Subjects appear to adapt their lingual contours to emulate the specific concavity and anteriority of their palates, resulting in midsagittal distance functions for all subjects that are relatively uniform throughout the palate region (see Figure 2.7). This similarity of vocal tract shapes across subjects leads to similar formant frequencies, as well. Note that, even with these findings, it is still not possible to say whether the production goal is to achieve a particular constriction shape or specific resonant characteristics, since the two are likely isomorphic in this situation and subjects appear to be doing both in the situation considered.

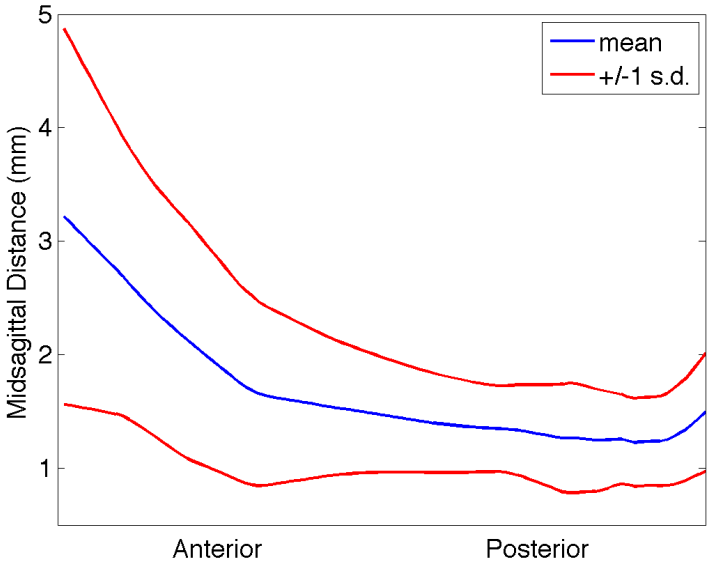


Figure 2.7: Mean midsagittal distance function across the five subjects, from the upper dentition to the hard-soft palate junction. Standard deviations (+/- 1 s.d.) along each line of the analysis grid are also shown.

Deeper insight into the question of goals comes from considering the case of palatal sharpness more closely. Subjects do not emulate palatal sharpness in their tongue shapes, nor are variations in palatal sharpness evident in acoustic variations. These findings are consistent with the assessment that palatal sharpness has marginal potential to affect the acoustics. Furthermore, they indicate that subjects will not account for certain morphological variations in their articulation if those variations do not alter the acoustics, which implies an acoustic production goal. This interpretation is reinforced by the significant negative correlation between palatal concavity and tongue sharpness, which implies that when subjects are faced with the choice between making a concave tongue shape and a sharp tongue shape, they choose to make concave tongue shapes, which has a more substantial impact on the acoustics.

2.5.1 Conclusions & Future Work

Examining the interplay between variations in articulation, acoustics and morphology holds promise for explaining interspeaker variability. More comprehensive examination of these complex interrelationships will require detailed knowledge of: (1) the variety and extent of morphological variations, (2) the theoretical acoustic impact of these variations, (3) the observed variation in articulation, and (4) the observed variation in acoustics. This work represents a first attempt to leverage recent findings regarding the first point to gain insights into the latter three points with respect to vowel production. Results indicate that, although palatal morphology has the potential to substantially affect vowel acoustics, this effect of palatal morphology may not be observed because speakers appear to adjust their articulations to match the key components of their palatal morphology and minimize potential acoustic variation.

The analysis presented here will be extended further in an ongoing study as more subjects with more diverse vocal tract morphologies are included in the subject pool. Future work will investigate the interplay of other morphological characteristics with speech production behavior, including the posterior pharyngeal wall, which has been parameterized using a similar method to that presented here. Detailing morphological characteristics off the midsagittal plane is also of major importance for analysis of speech behavior and for acoustic vocal tract simulations. Three-dimensional aspects of morphology are assumed to be particularly important both for analyzing real speech data and for accurately modeling vocal tract acoustics, but also pose substantial practical challenges given the current limitations of real-time MRI acquisition. In addition, investigations are ongoing regarding the impact of structural variation on other classes of speech sounds as part of a broader research program into the influence of morphological variation on all aspects of speech production. The ultimate goal of this work is

to predict individual speakers' production patterns, both articulatory and acoustic, from their morphological characteristics.

Another important avenue of inquiry is inverting the acoustic effect of vocal tract morphology. That is, assessing the extent to which information about a speaker's vocal tract morphology can be recovered from the speech signal. Results of the current study cast doubt on the feasibility of accurately predicting hard palate morphology from vowel acoustics, but do not entirely eliminate the possibility that accurate statistical methods for prediction may be developed. For instance, the correlation value between palatal anteriority and F3 ($r = -0.80$), though non-significant, still indicates that 64% of the variance in F3 is explained by palatal anteriority in the current data set (i.e., $r^2 = 0.64$). Several other correlation values are large enough in magnitude to indicate that they may be useful, as well. Many morphological characteristics are inherently more difficult to compensate for with articulatory changes, as well (e.g., vocal tract length), which may facilitate their accurate prediction from the acoustic signal.

Chapter 3

On Short-Time Estimation of Vocal Tract Length from Formant Frequencies

3.1 Abstract

Vocal tract length is both highly variable across speakers and determining of many aspects of the acoustic speech signal, making it an essential consideration for explaining behavioral variability. A method for accurate estimation of vocal tract length from formant frequencies would afford normalization of interspeaker variability and facilitate acoustic comparisons across speakers. A framework for considering estimation methods is developed from the basic principles of vocal tract acoustics, and an estimation method is proposed that follows naturally from this framework. The proposed method is evaluated using acoustic characteristics of simulated vocal tracts ranging from 14 to 19 cm in length, as well as real-time magnetic resonance imaging data with synchronous audio from five speakers whose vocal tracts range from 13.2 to 16.4 cm in length. Evaluations show improvements in accuracy over all previously proposed methods, with 0.631 and 1.159 cm root mean square error on simulated and human speech data, respectively. Empirical results show that the effectiveness of the proposed method is based on emphasizing higher formant frequencies, which seem less affected by speech

articulation. Theoretical predictions of formant sensitivity reinforce this empirical finding. Moreover, theoretical insights are explained regarding the reason for differences in formant sensitivity.

3.2 Introduction

Vocal tract length is an essential consideration for explaining behavioral variability. This structural characteristic of the speech production apparatus determines many aspects of the acoustic speech signal and, at the same time, is highly variable across speakers. The vocal tract lengthens throughout development, from an average length of approximately 8 cm at birth, up to 16 cm in adulthood [Fitch and Giedd, 1999, Vorperian et al., 2005, 2009]. Even between adults, vocal tracts vary from approximately 13 cm, to as many as 20 cm. These differences are particularly significant when one considers the relatively limited ability of most individuals to modulate vocal tract length, mainly through lip protrusion and laryngeal height. The role of vocal tract length in vowel production variability has been extensively studied and modeled, particularly with regard to the position and spacing of formant frequencies. It has been well-established that longer vocal tracts are associated with lower formant frequencies. This effect is supported both theoretically [Fant, 1960, Stevens, 1998] and has been repeatedly confirmed empirically [Peterson and Barney, 1952, Lee et al., 1999]. Recent work to quantify vocal tract length as a source of acoustic variability suggests that it is the second largest source of formant frequency variability overall after phonemic identity, accounting for up to 18% [Turner et al., 2009].

Given that information about vocal tract length is available in the acoustic signal, it is reasonable to expect that accurate predictions of vocal tract length should be possible from acoustic information alone. Developing a method to produce vocal tract length

estimates is easily motivated by the many practical applications that could be found for such a method. For instance, accurate vocal tract length estimation would afford the ability to normalize the acoustic characteristics of vowels for meaningful comparison across speakers. Indeed, vocal tract length normalization (VTLN) is already commonly used in automatic speech recognition (ASR) applications and has been shown to provide significant gains in system performance [Eide and Gish, 1996, Lee and Rose, 1996, Wegmann et al., 1996]. VTLN techniques typically seek a frequency scale transformation that allows for optimal comparison of spectral features extracted from different speakers. Appropriate transformations are commonly found by optimizing some maximum likelihood criterion over a set of acoustic data [Lee and Rose, 1998, Pitz et al., 2001], although formant alignment has also been investigated [Gouvea and Stern, 1997, Claes et al., 1998]. However, because the end goal of such techniques is the improvement of ASR performance, their ability to accurately estimate vocal tract length as a physical quantity has not been rigorously validated on data sets with careful vocal tract length measurements.

Accurate estimation of vocal tract length from the acoustic signal has been addressed in a handful of studies. In developing estimation techniques, the focus has been on the utility of vocal tract resonant characteristics. Almost all such studies have incorporated formants (i.e., the odd resonances, or poles) as features. Using formants is attractive because they are readily available in the speech signal – especially lower formants – and because the physical relationship between length and formant frequencies is fairly well-understood. One difficulty with relying on formants is that vocal tracts of different lengths can be made to produce the same formant frequencies through deformation of the area function, and formants alone are not enough to recover the area function and normalize out its effects [Mermelstein, 1967]. Some studies have also incorporated the zeros (i.e., even resonances) as additional features, which are sufficient to recover

an approximation to the area function when combined with formants [Schroeder, 1967], although they are not directly available in the acoustic signal. Other studies have utilized resonance bandwidths in addition to formant frequencies. Bandwidths are available in the acoustic signal, but they can be difficult to measure robustly. The most substantial differences, though, between the various proposed techniques has been in the form of the models used for estimation, and whether those models lends themselves to closed-form or iterative solutions.

Paige and Zue [Paige and Zue, 1970] develop a closed-form estimator based on an approximate relationship between resonant frequencies and the parameters of a band-limited approximation to the area function. Length was determined by minimizing a criterion that identifies the most uniform tube that might have produced a given formant structure. This estimator was initially tested using the first three poles and zeros as input, and was found to produce absolute errors in the range of 3.5 to 11.5% of total vocal tract length on a small data set. Using an iterative method that effectively increases the number of poles and zeros by estimating their frequencies, errors dropped down to the range of 0.6 to 4.9%. Subsequent studies have found that, when either of these these estimators is applied to only a small number of formants, the performance of drops dramatically [Kirlin, 1978, Necioğlu et al., 2000].

Wakita [Wakita, 1977] used the same uniformity criterion to develop an iterative algorithm that takes a set of formant frequencies and bandwidths as input. This method was able to produce absolute errors in the range of 1.6 to 8.6% of total vocal tract length on a small data set of full vowels. It should be noted that the literature repeatedly remarks that these is no compelling theoretical reason behind the uniformity criterion, despite the fact that it provides good performance in practice. Indeed, additional experiments by Necioglu [Necioğlu et al., 2000] have provided additional empirical evidence that this criterion, and several variants of it, result in reasonable estimation performance.

Wakita [Wakita, 1977] also suggested that several closed-form estimation techniques are possible to build, using only formant frequency information as input features. These suggestions were based on the observation that higher formants tend to be less affected by speech articulation (e.g., changes in the area) and therefore reflect vocal tract length more reliably. Specifically mentioned were using the fourth formant alone, and taking the mean length estimate from the fourth and higher formants. Adding successively higher formants provided increasing performance. In order for these closed-form estimators to be competitive with the iterative algorithm, however, it was necessary to utilize the fourth through eighth formants, which would be very difficult to estimate in general.

Kirlin [Kirlin, 1978] presented a probabilistic formulation of estimating vocal tract length from formant frequencies. By treating formant frequencies as erroneous measurements of a uniform vocal tract's resonances, Kirlin was able to effectively find closed-form expressions for both the maximum likelihood estimate and the maximum a posterior estimate of vocal tract length given some formant frequency measurements. When combined with data from the study by Wakita [Wakita, 1977], this formulation resulted in perhaps the first estimator designed in a data-driven fashion. Performance of this estimator was shown to be highly competitive with the previously proposed iterative methods when tested on a small data set, which clearly demonstrates the power of a statistical, data-driven approach in designing a closed-form estimator.

Fitch [Fitch, 1997] also proposed an estimator based on the spacing of successive formant pairs. As is consistent with theory, if vocal tract length is increased while the area function is held constant, all formant frequencies should lower and the spacing between them should decrease. This estimator has been shown to correlate well with both vocal tract length and body size in human males [Rendall et al., 2005]. Indeed, this estimator has been used extensively in the literature as a predictor of body size in a variety of other animals, as well, including domestic dog [Riede and Fitch, 1999], rhesus

macaques [Fitch, 1997], red deer [Reby and McComb, 2003], and colobus monkeys [Harris et al., 2006].

The present study further examines the possibility of accurate vocal tract length estimation from formant frequencies. Exclusive consideration is given to estimators that use only formant frequencies as input because of their ready availability in the acoustic signal. In keeping with previous work on this topic, it will be assumed that length estimation must be performed on a single short-time analysis window, and that information about speech articulation and phonemic identity is completely unknown. Furthermore, the focus here is on the design and evaluation of closed-form estimators because of the practical (e.g., computational) advantages associated with having a close-form solution, and because the various closed-form estimators mentioned above have not been rigorously evaluated and compared on larger data sets.

In examining vocal tract length estimation with these constraints, the specific goals of this work are as follows: (1) to develop a general framework from the basic principles of vocal tract acoustics for describing estimation methods, (2) to propose a new estimation method that follows naturally from the developed framework, (3) to evaluate this new method using simulated vocal tract data and real human speech data from real-time magnetic resonance imaging (rtMRI), and (4) to provide a theoretical justification for the proposed estimation method based on an examination of the relative sensitivity of different formants to changes in the vocal tract area function.

It is important to note from the outset that there are two perspectives regarding the concept of vocal tract length. The perspective taken here is that vocal tract length is a static characteristic that is inherent to a given speaker. Such a characteristic which might be measured, for instance, from a neutral or overall average vocal tract configuration. As such, the present work focuses on the development and evaluation of methods for accurate estimation a single length parameter per speaker. Vocal tract length defined in

this way could be used directly for speech normalization in ASR applications, or could function more generally as a physiologically and perceptually meaningful normalization factor for speech. It could also be used as an invariant speaker characteristic for applications in speaker modeling and identification, or as a correlate of other physiological characteristics (e.g., body size, as in the work by Fitch [Fitch, 1997]). The alternative perspective regarding vocal tract length of a dynamic characteristic that changes over short timescales (e.g., from phonetic segment to phonetic segment) due to lip rounding, larynx height and even tongue shape. Indeed, this was the perspective taken by Paige and Zue [Paige and Zue, 1970] and Wakita [Wakita, 1977]. Although the dynamic perspective is not taken in the present work, and accuracies in that regard are not evaluated here, it should be noted that all the methods discussed here will operate in the context of either perspective, and the relationship between both perspectives is discussed at various points throughout the present paper.

Section 3.3 of this paper describes our methodology, including the framework of vocal tract length estimators and the proposed method, as well as the details of acoustic modeling, rtMRI acquisition and experimental setup. In Section 4.4, the results of the various experiments are described. Section 4.5 provides a discussion of the experimental results and theoretical insights, and concluding remarks can be found in Section 4.6.

3.3 Method

3.3.1 Vocal Tract Length Estimation Framework

The current approach to length estimation proceeds, as many others have, from the well-known resonant properties of a tube which is assumed to be lossless, with an idealized radiation impedance, and uniform in cross-sectional area along its length. These simplifying assumptions are widely made in first-order examinations of vocal tract acoustic

characteristics, particularly the assumption of losslessness. As such, they will allow for the development of a general framework for vocal tract length estimation into which previous estimators and the presently proposed estimator will be placed. Assumptions related to the radiation impedance and cross-sectional area are justified by the motivation to create a purely acoustic method of estimation, where no knowledge of the area function can be assumed. Under such constraints, perhaps the most neutral assumption about the area function is that it is uniform. The radiation impedance is assumed to be zero for similar reasons, namely that it depends on the effective radius of the area function at the lips, which is here assumed to be unknown.

Under these assumptions, the length of the vocal tract has a simple relationship with vowel formant frequencies of the form:

$$L = \frac{c}{4\Phi} \tag{3.1}$$

where L is the length of the vocal tract, c is the speed of sound. The parameter Φ is defined as the lowest – or, first – resonance frequency of a lossless uniform vocal tract of length L. The term *length* will be used throughout the manuscript to refer to distance along the longitudinal axis of the vocal tract, extending from the glottis and the lips. This term contrasts with the terms *width* and, inversely, *constriction* and *narrowing*, which will be used to refer to distance perpendicular to the longitudinal axis.

For a lossless uniform vocal tract, the parameter Φ is related to formant frequencies by:

$$\Phi = \frac{F_n}{(2n-1)}, \quad n = 1, 2, 3, \ldots \tag{3.2}$$

where n represents the integer label of the formant frequency. Thus one can easily calculate length of a lossless, uniform vocal tract from any formant frequency using Equations 3.1 and 3.2. In the case of speech, however, the strict assumptions behind

Equations 3.1 and 3.2 are not generally applicable, and the relationship expressed in Equation 3.2 is only approximate. In particular,

$$\hat{L} = \frac{c}{4\hat{\Phi}}, \tag{3.3}$$

where $\hat{L}$ and $\hat{\Phi}$ are approximations to L and Φ.

Any calculation of vocal tract length using Equation 3.3 with formant frequencies from human speech data is an estimate, and each formant frequency can be considered as a feature that has the potential to provide some information about vocal tract length. It becomes of interest to determine the usefulness of these features and the accuracy of estimates that can be obtained using this model. To that end, it is possible to generalize the linear relationship between Φ and F_n expressed in Equation 3.2, allowing one to incorporate all formant frequencies as features into a linear combination. In particular,

$$\hat{\Phi} = \frac{\beta_1 F_1}{1} + \frac{\beta_2 F_2}{3} + \frac{\beta_3 F_3}{5} + \cdots + \frac{\beta_m F_m}{2m-1} \tag{3.4}$$

up to the highest possible integer formant number, m, that can reliably be estimated from the frequency spectrum. Note that, if the assumptions behind Equations 3.1 and 3.2 are correct, then Equation 3.4 provides a precise value for Φ if any single coefficient $\beta_n = 1$ and all others are zero. The same is also true if all coefficients $\beta_{1\ldots m} = 1/m$.

This linear model will serve as the basic, generalized framework for estimating Φ from formant frequency measurements. It is generalized because it allows for information about Φ contained in any formant frequency to be incorporated into the estimate. It is basic because there are several obvious ways to extend the model, including by the addition of terms that raise formant frequencies to higher powers (e.g., $\beta_{n_2} F_n^2$), which would make this a nonlinear model. Perhaps the most obvious extension of this basic model would be to add an constant offset term, β_0, in order to make this a full multiple

linear regression model (i.e, one that need not pass through the origin). Extending the model by adding an offset term has a specific motivation stemming from Kirlin's [Kirlin, 1978] probabilistic formulation of vocal tract length estimation. This specific extension will be discussed in greater detail in Section 4.5.

For present purposes the discussion will be confined to the basic, general linear model in Equation 3.4, which is sufficient to describe a variety of estimation schemes, including many previously proposed estimators of length. In fact, one can incorporate several previously proposed estimators into this framework. It was suggested by Wakita[Wakita, 1977] that higher formant frequencies are less sensitive to speech articulation, and would therefore provide more robust estimates of vocal tract length. Using a single formant frequency, F_n, to estimate Φ from Equation 3.4 can be represented by setting coefficient $\beta_n = 1$ and all others to zero. Wakita [Wakita, 1977] also suggested that averaging together the p highest formants might provide an even more robust estimate, which can be represented by setting all coefficients $\beta_1 \ldots \beta_{m-p}$ to zero and those $\beta_{m-p+1} \ldots \beta_m$ to $1/p$.

Fitch [Fitch, 1997] proposed an estimator called Frequency Dispersion, based on the average spacing between successive formant pairs:

$$\hat{\Phi}_{FD} = \frac{F_2 - F_1}{2(m-1)} + \frac{F_3 - F_2}{2(m-1)} + \cdots + \frac{F_m - F_{m-1}}{2(m-1)}. \tag{3.5}$$

Note that Equation 3.5 can be considerably simplified, because many of the terms ultimately cancel out. The equation can be eventually re-written as:

$$\hat{\Phi}_{FD} = -\frac{F_1}{2(m-1)} + \frac{F_m}{2(m-1)}. \tag{3.6}$$

Equation 3.6 can be made consistent with the general form specified above by setting all coefficients in Equation 3.4 to zero, except for $\beta_1 = -\frac{1}{2m-2}$ and $\beta_m = \frac{2m-1}{2m-2}$.

Kirlin [Kirlin, 1978] provided a probabilistic formulation that allowed for explicit maximization of the conditional probability of Φ given some measured formant frequencies: $p(\Phi|F_n)$. The maximum likelihood estimate of Φ given Kirlin's formulation is:

$$\hat{\Phi}_{MLE} = \frac{\sum_{n=1}^{m}(2n-1)F_n/\sigma_n^2}{\sum_{n=1}^{m}(2n-1)/\sigma_n^2}, \tag{3.7}$$

where $\sigma_n^2 = \frac{1}{m}\sum_{n=1}^{m}(F_n - \Phi(2n-1))^2$. To place this estimator in the current framework, one can easily manipulate the terms to find that $\beta_n = \frac{(2n-1)^2}{\sigma_n^2 \sum_{i=1}^{m}(2i-1)^2/\sigma_i^2}$.

Working within the framework presented above raises an obvious question: what is the optimal set of coefficients for predicting $\hat{\Phi}$ – for instance, what values minimize the least-squared error criterion? One answer to this question can be obtained by setting the coefficients in a data-driven fashion by treating this question as a regression problem. In order to do that, one must have a data set containing a matrix, $\mathbf{M}$, where each row contains a set of normalized formant frequencies, that is $F_n/(2n-1)$ for $n = 1 \ldots m$. It is also necessary to have a vector, $\mathbf{\Phi}$, with the same number of rows as $\mathbf{M}$, containing corresponding values of Φ. The desired vector of coefficients, $\boldsymbol{\beta}$, are then found according to:

$$\boldsymbol{\beta} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{\Phi} \tag{3.8}$$

which is a solution to ordinary least-squares regression. The data utilized for this purpose should ideally be gathered from a variety of vocal tract configurations that are representative of speech articulation. Two methods of gathering such data are explored here: through acoustic modeling and acquisition of human speech data.

3.3.2 Acoustic Modeling

Modeling of vocal tract acoustics is done in two ways, as described in the following sections: 3.3.2 and 3.3.2. The first model is based on treating the vocal tract as a series of

concatenated tubes [Kelly and Lochbaum, 1962]. This model, which will be henceforth referred to as the 'multi-tube' model, has been well-studied with respect to vocal tract modeling and synthesis [Fant, 1960, Rabiner and Schafer, 1978, Stevens, 1989, 1998]. The second method is based on perturbation theory, which has a similarly long history in speech production research [Chiba and Kajiyama, 1941, Stevens, 1998]. These models are utilized for two purposes in the present study. The multi-tube model is used to generate synthetic formant frequencies for experiments in vocal tract length estimation, while both models are used to provide theoretically-motivated explanations for the form of the estimation models. The specifications of these models are explained in the following subsections.

It should be noted that the primary purpose of utilizing these models is not to provide highly faithful speech synthesis. Rather, these models are meant to provide correspondence to essential aspects of human vocal tract acoustics, but with certain simplifying assumptions. To reiterate from above, these assumptions include idealized geometry, lack of loss and lack of consideration for the radiation impedance. It is important to make these assumptions for a number of reasons. First, they are intended to be consistent with previous work on vocal tract length estimation (e.g., by Wakita[Wakita, 1977] and Fitch[Fitch, 1997]), which can then be unified into the proposed estimation framework. Because this framework subsequently forms the basis the proposed estimator, the performance of all the considered estimators can then be meaningfully compared. These assumptions also facilitate the theoretically-motivated insight into the performance of the proposed estimator. Because the assumptions are somewhat strong, the present study includes a set of parallel estimation experiments on real human speech data. Indeed, comparing estimation accuracies on real and synthetic data provides an opportunity – discussed later in Section 3.3.3 – to assess estimation accuracy as a function of potential mismatch between modeling assumptions and characteristics of the

data. It is for all these reasons that the assumptions in the acoustic models are consistent with the assumptions behind the proposed estimator.

Multi-Tube Model

The multi-tube model treats the vocal tract as a series of lossless, cylindrical tubes that are concatenated end-to-end. For a given area function, A, the formant frequencies can easily be computed by first calculating the reflection coefficients between each pair of adjacent tubes:

$$\Gamma(x) = \frac{A(x+1) - A(x)}{A(x+1) + A(x)}, \tag{3.9}$$

where $A(x)$ is the cross-sectional area of the vocal tract at distance x from the glottis. Reflection coefficients are then used to compute the coefficients of the prediction filter polynomial. This is done using Levinson recursion, as described in [Kay, 1988] and implemented in the Matlab® Signal Processing Toolbox™ (The MathWorks Inc., version 7.8.0). Finally, the formant frequencies can be found by taking the roots of the prediction filter polynomial (in radians), and converting to Hertz. Formant frequencies are then sorted and assigned integer labels.

Perturbation Theory

This section provides a brief review of perturbation theory as presented by Stevens [Stevens, 1998], which is used in the present study to model and explain the effect of vocal tract constrictions on formant frequencies. A more complete exposition of perturbation theory can be found in the reference provided. Only the details that are important in leading up to the presentation of Equation 3.15 are presented here.

Standing wave patterns in a completely uniform vocal tract produce a pressure value at x, a location along the length of the vocal tract from the lips to the glottis, of the following form:

$$p_n(x) = P_m sin \frac{2\pi F_n x}{c} \tag{3.10}$$

where P_m is the maximum pressure possible, F_n is the n^{th} natural frequency of the vocal tract and c is the speed of sound inside the vocal tract. The volume velocity profile also has a sinusoidal shape, following:

$$U_n(x) = jP_m \frac{A}{\rho c} cos \frac{2\pi F_n x}{c} \tag{3.11}$$

where A is the cross-sectional area of the uniform tube under consideration, and ρ is the ambient density of air.

When an initially-uniform tube is subsequently constricted at some location along its length, the amount of stored energy in the system (W) is changed as function of both the stored potential (V) and stored kinetic energy (T) in the system:

$$\Delta W_n = \Delta V_n + \Delta T_n \tag{3.12}$$

where n is an integer label referring to a specific natural frequency. Changes in stored energy cause a shift in natural frequencies $dF_n = -\Delta W_n F_n$. Changes in stored potential energy can be expressed as

$$\Delta V_n = \frac{1}{4} |p_n(x)|^2 \frac{\Delta l \Delta A}{\rho c^2} \tag{3.13}$$

and changes in stored kinetic energy can be expressed as

$$\Delta T_n = -\frac{1}{4} |U_n(x)|^2 \frac{\rho \Delta l \Delta A}{A^2} \tag{3.14}$$

for small changes in the cross-sectional area of the tube, ΔA, at location x and over a short length of the tube, Δl. This model assumes that the pressure and volume velocity profiles are not substantially altered by small perturbations of the vocal tract area or length. It has been repeatedly shown that this assumption is reasonable for relatively unconstricted vocal tract shapes, as in vowels (e.g., as described by Mrayati [Mrayati et al., 1988]). By extending this assumption to several adjacent short area perturbations, the acoustic consequences of vocal tract constrictions with a longer spatial extent can be modeled by summing in the following way:

$$\Delta W_n = \sum_{a=-t}^{t} \Delta V_n(x_a) + \Delta T_n(x_a) \tag{3.15}$$

where $x_a = x_0 + \frac{a\Delta l}{2}$, x is perturbation location variable defined with respect to Equations 3.10 and 3.11, and x_0 is the spatial center of the long constriction. Thus, this relatively long constriction is composed of t adjacent short perturbations, and the value $t\Delta l$ is the overall length of the constriction. It may seem necessary to employ more sophisticated models for calculating changes to the pressure and volume velocity profiles (e.g., as utilized by Story [Story, 2005]). However, it will be shown empirically later in the present paper that perturbation theory produces highly meaningful results when extended in this way, as least for the relatively simple constriction shapes considered here.

3.3.3 Data for Estimation Experiments

To conduct experiments regarding the estimation of vocal tract length from formant frequencies, two data sets were gathered: one composed of simulated speech data from the multi-tube acoustic model and one composed of real human speech data obtained from rtMRI. The following sections, 3.3.3 and 3.3.3, describe the details of how those data

sets were gathered. Consideration of resonances was limited to the first four formant frequencies in all cases, due to the difficulty associated with estimating higher formants on human speech data.

Having two data sets also provides an opportunity to analyze whether the modeling assumptions behind the proposed estimator framework, which incorporates the previous and proposed estimators, are reasonable. The synthetic data maintain the assumptions of losslessness and of a idealized radiation impedance, but not of uniformity of the area function. The human speech data, by contrast, are likely to violate all of these assumptions to some degree. Thus, the differential accuracy of the proposed estimator on the two data sets should provide an indication of how much estimator error can be attributed to normal articulation of the vocal tract versus modeling assumptions behind the estimator, keeping in mind that the human speech data may also contain measurement error.

Simulated Speech Data

Simulated formant frequencies were gathered from a set of randomly-generated vocal tract area functions. Area functions were parameterized using the spatial discrete Fourier transform of vocal tract area functions developed by Schroeder [Schroeder, 1967] and Mermelstein [Mermelstein, 1967], and later utilized by Iskarous [Iskarous, 2010]. This parameterization is based on representing an arbitrary area function as a linear combination of spatial sinusoids defined along the vocal tract's length. In this work, a spatial half-cosine and its first five integer harmonics were used as the basis for representation. Specifically, at x, a location along the length of the vocal tract from the glottis to the lips, the area function's deviation from a uniform shape is:

$$\delta A(x) = A_0 \sum_{n=1}^{6} a_n cos(\frac{\pi n x}{L}), \tag{3.16}$$

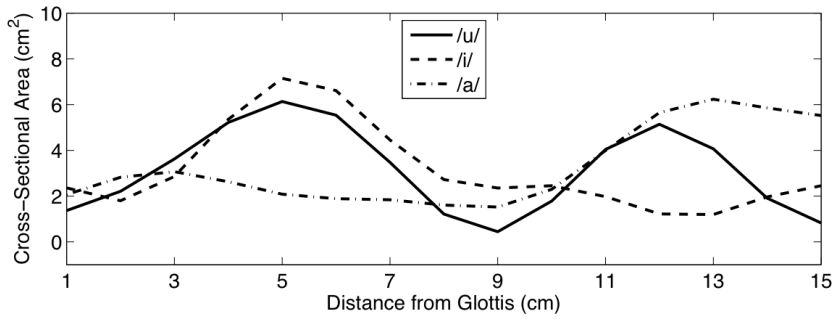
where n is the integer label of the sinusoidal harmonic component and A_0 is the cross-sectional area of a uniform vocal tract in cm^2. In this case, $A_0 = \pi$, corresponding to a midsagittal distance of 2 cm. The area function at this location is then,

$$A(x) = A_0 + \delta A(x) \tag{3.17}$$

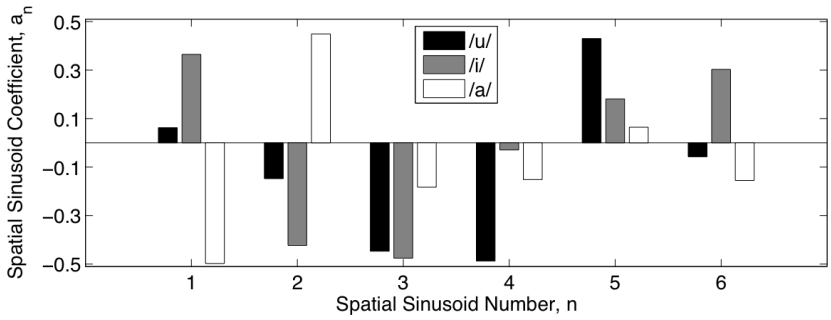
where coefficients $a_{1...6}$ determine the contribution of each sinusoid to the overall shape. The harmonic series of six sinusoidal components can also be regarded as three components that are symmetric along the vocal tract (i.e., the even harmonics), and three that are anti-symmetric (i.e., the odd harmonics).

Although originally developed to represent known area functions, this area function parameterization was used in the present study to generate a total of 3,200 area functions for six simulated "speakers", each with a different vocal tract length, ranging from 14 to 19 cm with 1 cm intervals. Area functions were generated by random selection of the coefficients, $a_{1...6}$, from a uniform distribution with range (-1,1). Random selection of coefficients in this way produces area functions with very small or negative values – here defined as areas less than 0.1 cm^2 – approximately 30% of the time, at locations distributed along the entire length of the area function. During the generation process, any area functions displaying an area less than this pre-specified amount were discarded. This process of random coefficient generation and checking for sub-threshold values was repeated for a given simulated speaker until 3,200 area functions were compiled. Repeating this for all six speakers amounted to a total of 19,200 total area functions in the data set.

Using this area function generation procedure, a wide variety of vocal tract shapes can be generated automatically. Figure 3.1 shows three area functions extracted from the data which were found to be closest, in terms of mean squared error, to the area



(a) Area Functions



(b) Coefficients

Figure 3.1: Three randomly-generated area functions (a) from the simulated speech data set and their corresponding coefficient values (b) that define their shape according to Equations 3.16 and 3.17. These example area functions were found to be closest, in terms of RMS error, to the area functions presented by Wood[Wood, 1979] for the English vowels /u/, /i/ and /a/.

functions presented by Woods[Wood, 1979] for the English vowels /u/, /i/ and /a/. The overall range of cross-sectional area in the entire data set was 0.1 cm^2 – 11.3 cm^2, corresponding to an overall range of midsagittal distances of 0.36 – 3.79 cm. Each whole-centimeter location along the length of the area function displayed a standard deviation of approximately 1.5 across the entire data set, indicating that no single locations exhibited diminished area variation compared to the others.

Formant frequencies were generated from these area functions using the multi-tube model. This necessitated the conversion of continuous area functions into vectors of finite area measurements, with each measurement representing the cross-sectional area of one tube in the model. All tubes were considered to have a constant length of 0.25

cm, making the number of tubes change with different overall vocal tract lengths, from 56 (14 cm) to 76 (19 cm).

Human Speech Data

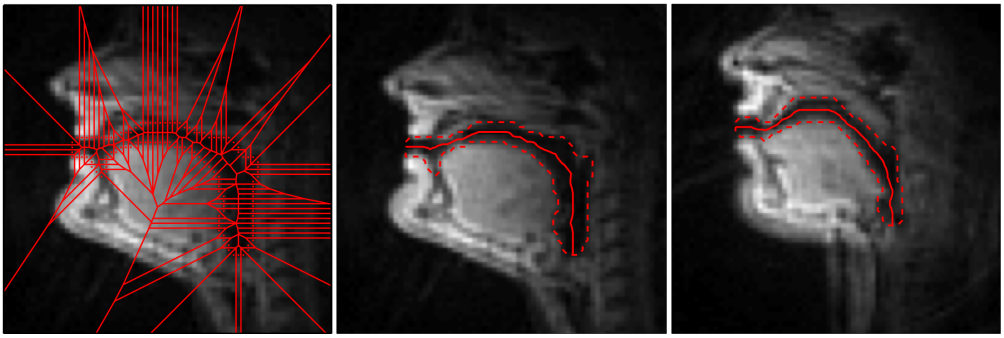
Human speech data were collected from five native speakers of American English (two male, three female). Each subject spoke five sentences from the MOCHA-TIMIT corpus (see Appendix C). Data were acquired at Los Angeles County Hospital on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha WI). A custom 4-channel upper airway receiver coil array was used for RF signal reception, with two anterior coil elements and two coil elements posterior to the head and neck. A 13-interleaf spiral gradient echo pulse sequence was used (T_R = 6.164 msec, FOV = 200$\times$ 200 mm, flip angle = 15 degrees, receiver bandwidth = $\pm$125 kHz). Scan plane localization of the 5 mm mid-sagittal slice was performed using RTHawk (HeartVista, Inc., Los Altos, CA), a custom real-time imaging platform [Santos et al., 2004]. Images were reconstructed at a rate of 23.33 frames/second. Image resolution after reconstruction was 68$\times$ 68 pixels at 2.9$\times$ 2.9 mm pixel width. More details about the rtMRI acquisition can be found in [Narayanan et al., 2004, Bresch et al., 2008, Kim et al., 2011]. Synchronous audio recordings of the subjects' speech were also acquired using an optical microphone. Audio were subsequently denoised according to the protocol described by Bresch [Bresch et al., 2006]. This denoising technique promises nearly 30dB noise suppression during speech and, in general, minimal errors are expected in terms of spectral distortion as a result of applying this method. However, one may expect, due to the nature of the recording environment inside the scanner bore, some reverberation and background noise caused by the cryogen pump and ventilation system.

Audio was analyzed using Praat [Boersma, 2001] for both formant and pitch tracking. Formant tracking was configured to find six formants in the range 0–5500Hz with

a window length of 25ms. These settings provide superfluous and potentially spurious formant tracks, but also provided the most accurate formant tracking of the formants frequencies of interest when overlaid on a spectrogram of the same utterance and visually inspected. Formants of interest were identified by this overlay procedure, and irrelevant formant values were discarded. Pitch tracking was configured to find pitch in the range 75–500Hz using the autocorrelation method with a window length of 10 ms. Pitch measurements were used to remove non-sonorant sounds from further analysis. Formant analysis frames were eliminated from further consideration if a reliable pitch could not be found by the pitch tracking algorithm in the temporally nearest pitch analysis frame. In total, this resulted in 3,870 frames of data, or approximately 775 frames per subject.

Toward measuring vocal tract length for each subject, vocal tract images for each subject were extracted from the rtMRI video sequences at times when formant measurements were also being considered. From this subset of video frames, an overall mean image was formed for each subject by taking the pixel-wise mean intensity value across all the images in the subset. This overall mean image is a representation of the average vocal tract posture assumed by the subject during production of the formants in the data set. However, calculating a mean image over such a large number of images (approximately 200 frames per speaker) results in a mean image which is blurred and unsuitable for finding vocal tract outlines/midlines. Therefore, one additional mean image was calculated for each subject from the ten images that were closest to the overall mean image in the terms of having the smallest sum of squared pixel intensity differences across all pixels in the image plane. From this final, unblurred mean image, the vocal tract outlines were traced.

Vocal tract outlines were found using Canny edge detection [Canny, 1986] with manual linking and correction. For all five subjects, this procedure resulted in an outer vocal tract contour following the contour of the upper lip, alveolar ridge, hard palate, velum



(a) al1: Voronoi diagram (b) al1: 16.4 cm (c) al3: 14.9 cm

(d) ch1: 13.2 cm (e) mf2: 15.8 cm (f) ts1: 16.0 cm

Figure 3.2: Midsagittal rtMR image of subject al1 (a) with vocal tract outlines and Voronoi diagram overlaid. Mean images of the five subjects (b–f), a representation of the average vocal tract posture assumed by the subject during production of the formants in the data set. Vocal tract outlines are overlaid (dashed lines) with the vocal tract midlines (solid lines) calculated from those outlines using Voronoi skeletons. The length of these midlines was used as the measure of vocal tract length for each subject. The specific values of vocal tract length are given in centimeters next to the subjects' identifiers.

and posterior pharyngeal wall. The lower vocal tract contour followed the lower lip and the tongue from its tip to below its attachment point of the epiglottis. The vocal tract endpoints were identified by inspection, with the lower point being the upper edge of the false vocal folds. The Voronoi skeleton [Preparata and Shamos, 1990] was subsequently calculated from these outlines. Voronoi vertices were selected that were equidistant from vocal tract outlines, forming the shortest path between the end vertices. The vocal

tract outlines and resulting midlines for all five subjects can be seen in Figure 3.2. The mean vocal tract length across these five subjects was 15.3 cm.

3.3.4 Length Estimation Experiments

Using both the simulated and human speech data that were collected, a series of experiments were carried out to evaluate the performance of several vocal tract length estimators described in Section 3.3.1. The evaluated estimators included Fitch's [Fitch, 1997] Frequency Dispersion (Equation 3.6), Wakita's [Wakita, 1977] proposal of using the highest formant (i.e., F4) and mean of the two highest formants (i.e., F3 and F4), Kirlin's [Kirlin, 1978] maximum likelihood estimator (Equation 3.7, and the proposed regression-based method with coefficients determined as in Equation 3.8. The precise specifications of these estimators are given in Tables 3.2 and 3.3.

Experiments were conducted using a repeated resampling procedure with heldout data. Separate subsets of both full data sets were utilized for training and testing. Half of the data points were randomly assigned to a training set, used to obtain the regression coefficients, and the other half were assigned to a test set, used to evaluate the accuracy of the estimates. Using a heldout data procedure provides a better indication of how the results will generalize to other data sets. This procedure of random assignment, model fitting and evaluation was repeated 1000 times. Resampling the data in this way should provide a more robust estimate of the overall accuracy by examining the mean accuracy over all repetitions. Resampling also affords an estimate of the stability of the model coefficients by examining their standard deviation over all repetitions.

3.3.5 Formant Sensitivity Experiments

Wakita's [Wakita, 1977] idea to use higher formants, alone or in combination, to predict vocal tract length was based on the suggestion that higher formants are less sensitive to

Table 3.1: Empirical comparison of the sensitivity formant frequencies to speech articulation in the simulated and human speech data sets. The measure presented is normalized standard deviation: $\frac{\sigma_{F_n}}{2n-1}$, where n is the formant number. Lower numbers indicate less sensitivity to articulation. Note that sensitivity decreases as higher formants are considered.

Formant #	**Simulated Speech Data**	**Human Speech Data**
1	104.0	198.0
2	98.3	102.9
3	97.2	74.7
4	65.1	55.0

speech articulation. The *sensitivity* of a formant, in this context, refers to the amount of variation in that formant's observed frequency, expressed as a proportion of its expected frequency during a neutral area function shape. Decreased sensitivity of certain formants across a variety of area function shapes would make them better predictors of vocal tract length because their observed frequencies, which are always determined by some combination of length and area function shape, would be determined relatively more by length alone. It was observed that the design of the proposed, data-driven estimator, as shown in Tables 3.2 and 3.3, places differentially more weight on higher formants. This finding is consistent with the idea that higher formants are more reliable as features, perhaps because they are less sensitive. Indeed, higher formants appear to vary less across the vocal tract configurations contained in the collected data sets. Table 3.1) shows that the normalized standard deviation of formant frequencies decreases monotonically as higher formants are considered. Further empirical verification for this suggestion was sought.

Experiments were performed using a simulated vocal tract of 17 cm length with uniform cross-sectional area of π cm^2. Vocal tract constrictions of uniform 0.5 cm^2 narrowing, with lengths ranging from 0 to 8.5 cm (i.e., 0 to 50% of total vocal tract length), were generated such that the central point of their spatial extent – the constriction location – was placed at every possible location along the length of the tract. Note

that, although a single value of constriction degree was used, according to perturbation theory (see Equations 13 and 14), changing the constriction degree should only act as a scaling factor on the sensitivity of all formants, and therefore not affect comparisons of sensitivity across formants. Formant frequencies were calculated for vocal tracts of the specified parameters using both modeling techniques described in Section 3.3.2. Perturbation theory would suffice for these sensitivity experiments, and can also provide deeper insights into the reasons for the formant sensitivity differences, as will be discussed later. However, it is not immediately clear that the assumptions made in extending the theory from a single constriction to multiple ones (see Equation 3.15) are reasonable, nor is it clear whether a perturbation of the size assumed here is reasonably small. Concern over these assumptions compelled additional simulations using the multi-tube model, to provide corroborating evidence for their adoption. It will be shown that the simulations results using either method match very closely, indicating that these assumptions are, indeed, reasonable.

3.4 Results

3.4.1 Length Estimation Experiments

Accuracies of various estimators on the simulated data set are shown in Table 3.2. Results on human speech data are shown in Table 3.3. Accuracies are presented in terms of the root mean squared error (RMS error) across all test data. The specific estimator coefficients, corresponding to those in Equation 3.4, are also listed. RMS error of the proposed estimator on simulated data was 0.630 cm, which is 3.82% of the mean vocal tract length in the data set. The proposed estimator achieved an RMS error of 1.159 cm on the human speech data, which is 7.58% of the mean vocal tract length in that data set.

Table 3.2: Estimation accuracies of several estimators on simulated speech data in terms of RMS error. The specific coefficients ($\beta_1 \ldots \beta_4$) that define the estimators are also shown. Note that the coefficient values for the maximum likelihood estimator and the proposed estimator are mean values across all resamplings. Standard deviations of $\beta_{1\ldots4}$ the maximum likelihood estimator were 0.001, 0.001, 0.002 and 0.003, respectively. Standard deviations of $\beta_{1\ldots4}$ for the proposed estimator $\beta_{1\ldots4}$ were all equal to 0.002.

Estimator	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	**RMSE (cm)**
Freq. Dispersion	-0.167	0.000	0.000	1.167	1.692
F_4 only	0.000	0.000	0.000	1.000	1.206
Mean F_3 & F_4	0.000	0.000	0.500	0.500	1.194
Max. Likelihood	0.083	0.122	0.192	0.604	0.683
Proposed	0.089	0.102	0.121	0.669	0.631

Table 3.3: Estimation accuracies of several estimators on human speech data in terms of RMS error. The specific coefficients ($\beta_1 \ldots \beta_4$) that define the estimators are also shown. Note that the coefficient values for the maximum likelihood estimator and the proposed estimator are mean values across all resamplings. Standard deviations of $\beta_{1\ldots4}$ the maximum likelihood estimator were 0.004, 0.006, 0.006 and 0.005, respectively. Standard deviations of $\beta_{1\ldots4}$ for the proposed estimator $\beta_{1\ldots4}$ were 0.008, 0.009, 0.020 and 0.024, respectively.

Estimator	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	**RMSE (cm)**
Freq. Dispersion	-0.167	0.000	0.000	1.167	5.356
F_4 only	0.000	0.000	0.000	1.000	3.911
Mean F_3 & F_4	0.000	0.000	0.500	0.500	3.243
Max. Likelihood	0.082	0.345	0.348	0.225	2.034
Proposed	0.024	0.151	0.277	0.702	1.159

3.4.2 Formant Sensitivity Experiments

Results of the formant sensitivity experiments can be summarized in two ways. First, formant sensitivity can be shown as a function of the constriction location along the length of the vocal tract. In order to effectively visualize these *sensitivity functions*, a particular constriction length must be chosen. Figure 3.3 shows the sensitivity functions for the first four formants when a vocal tract constriction that is 25% of the total vocal tract length is chosen. The range of these sensitivity functions can also be calculated and used as a indication of overall formant sensitivity as a function of constriction length,

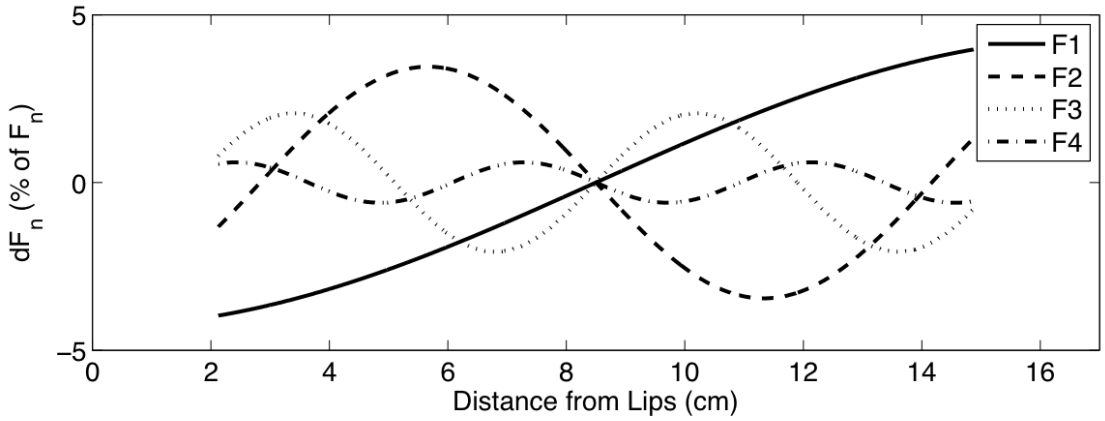


Figure 3.3: Sensitivity functions predicted by perturbation theory, showing the sensitivity the first four formants to a vocal tract constriction whose center is placed at all possible locations along the length of the vocal tract. This example uses a uniform vocal tract constriction with a length that is 25% of total vocal tract length. Note that the range of the sensitivity functions, in terms of dF_n, decreases as higher formants are considered. The range of sensitivity functions can be used as a measure of formant sensitivity to constrictions of a given length, regardless of their location, as in the sensitivity range functions shown in Figures 3.4 and 3.5.

without regard to the specific location of the constriction. Figure 3.4 shows these *sensitivity range functions*, as calculated using perturbation theory, across all constriction lengths considered in the formant sensitivity experiments. Figure 3.5 shows the sensitivity range functions calculated using the multitube model. The fact that Figures 3.4 and 3.5 are strikingly consistent can be taken as evidence that perturbation theory can meaningful be extended in a way consistent with Equation 3.15.

3.5 Discussion

3.5.1 Optimal Estimation of Vocal Tract Length

The proposed vocal tract length estimator displays the lowest estimation error compared to previously proposed estimators on both simulated and human speech data. This

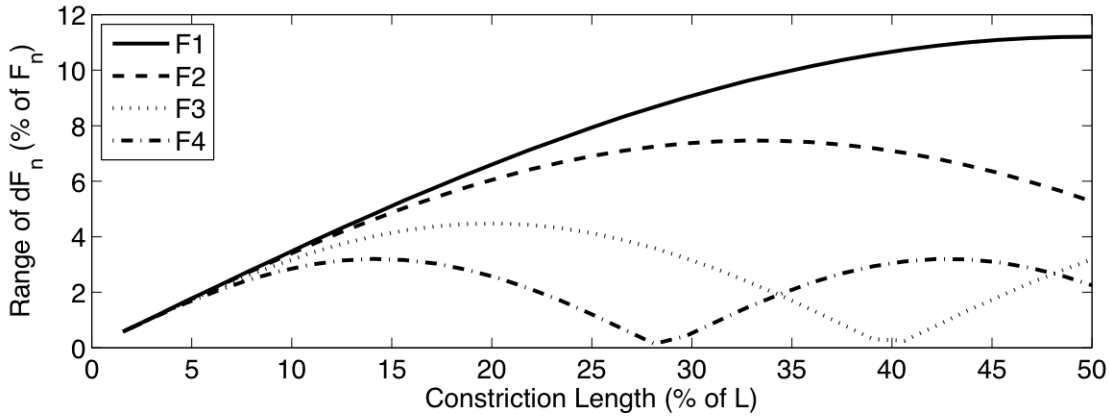


Figure 3.4: Sensitivity range functions predicted by perturbation theory, showing the range of sensitivity of the first four formants to vocal tract constrictions of different lengths, regardless of their location. This example uses uniform vocal tract constrictions with lengths varying from 0 to 50% of total vocal tract length. Note that, in general, the sensitivity range decreases as higher formants are considered, although there are some exceptions to this general trend. When constrictions are very small (i.e., less than 5% of total vocal tract length), there is not much difference between the sensitivity ranges of different formants. There are also some exceptions to the decreasing sensitivity trend, as can be seen in the reversal sensitivity range for F3 and F4 when constrictions are between approximately 34% and 48% of total vocal tract length.

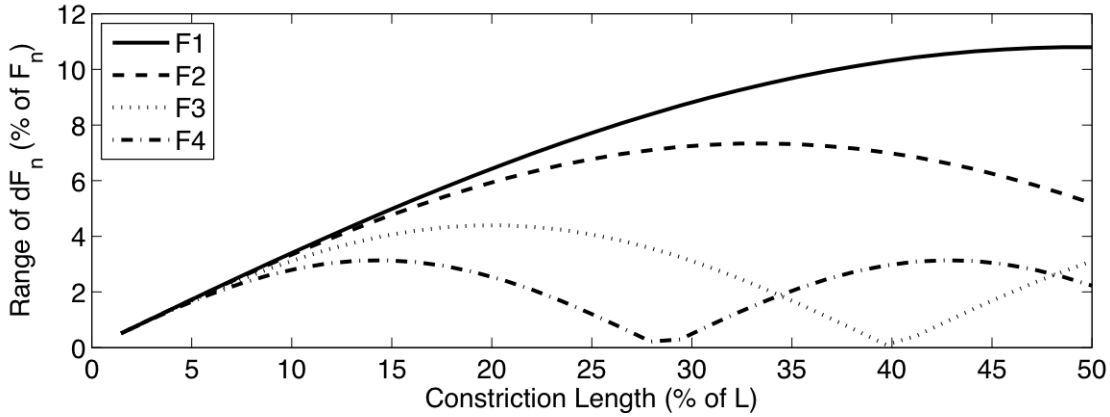


Figure 3.5: Sensitivity range functions predicted by the multitube model, showing the range of sensitivity of the first four formants to vocal tract constrictions of different lengths, regardless of their location. This example uses uniform vocal tract constrictions with lengths varying from 0 to 50% of total vocal tract length. Note the striking similarity to the results predicted by perturbation theory, as shown in Figure 3.4, which provides corroborating evidence that the proposed assumption behind and extension to perturbation theory are reasonable.

result is perhaps not surprising, given that the proposed estimator is optimal in the least-squares sense among estimators belonging to the framework developed here on the data sets used in the present paper. It should be noted, however, that this optimality is only guaranteed on the training data, and not necessarily on the test data. The heldout data validation scheme employed here lends confidence that the proposed estimator's performance is generalizable. Kirlin's [Kirlin, 1978] maximum likelihood estimator – the only other estimator developed in a data-driven fashion – displays the second best performance on both data sets, and is very similar in terms of accuracy on the simulated data set. The remaining estimators display substantially lower performance on the data sets used in the present study. The rank order of methods by accuracy is precisely the same on both data sets.

The proposed estimator, whether it is trained on simulated or human speech data, exhibits coefficients that increase in value as higher formants are considered. The maximum likelihood estimator, which provides the second best performance overall, also has coefficients of this form. This shows that all formants provide some information about vocal tract length, and it strongly implies that formants provide increasingly reliable information as higher formants are considered. As such, it is consistent with the idea that higher formants are less sensitive to speech articulation. Further empirical evidence and a theoretical justification for this idea are presented in Section IV.B.

Accuracies on human speech data are generally lower than accuracies on simulated data. Using the proposed estimator, for example, the difference is approximately 0.5 cm, or between 3 and 4% of average vocal tract length. This difference highlights errors that result from the simplifying assumptions made in developing the previous and proposed estimators, and in developing the general estimation framework of Equation 3.4. Those assumptions include idealized geometry and radiation impedance, and lack of loss, all of which are also true of the simulated speech data set, but not the human speech data.

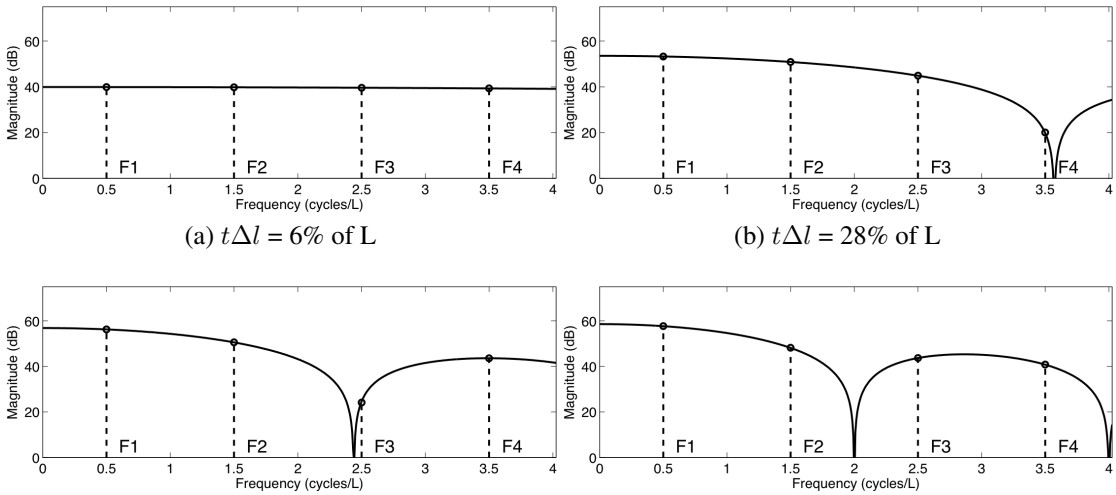
An example of this can be seen very clearly by considering the performance of the maximum likelihood estimator, which was derived from a probabilistic formulation that follows a similar set of simplifying assumptions. As a result, this estimator performs close to optimal on the simulated data, and approximately three times worse on the human speech data. Errors resulting from these assumptions can be addressed by incorporating more physical knowledge into the estimator framework, with the goal of eliminating certain assumptions altogether. Such changes will be most feasible for assumptions that do not depend critically on knowledge of speech articulation (e.g., lossy sidewalls). Assumptions that do depend on articulation knowledge (e.g., the radiation impedance) may still be refined, however, by modifying or weakening the assumptions. Other obvious sources of error for estimates on human speech data include measurement noise introduced during formant tracking or when measuring vocal tract length. The human speech data set is also considerably smaller than the simulated data set, and increasing the amount of human speech data would likely lead to small improvements.

Eliminating any remaining estimation error may require a more complex model than stated in Equation 3.4. Model complexity can be increased in the most straightforward way by either adding model parameters or adding input features. One possible extension of the model by adding parameters is explored in Section IV.C. Input features that might be added to the model include acoustical features, such as formant bandwidths. However, it seems likely that some knowledge of speech articulation must be incorporated to make this problem well-posed, because changes in the area function are such a large – even dominant [Turner et al., 2009] – source of formant variation. Even if a purely acoustical method of estimation is desired, some knowledge of the area function could still be utilized if that knowledge was obtained through acoustic-to-articulatory inversion, although inverted features would likely eliminate the possibility of a closed-form estimator.

3.5.2 Formant Sensitivity Differences

Several empirical results in this study reinforce the idea that higher formants are relatively less sensitive to speech articulation as compared to lower formants. Direct examination of formant frequencies in the data sets presented here reveals that each formant varies proportionally less than the formant just below it. Moreover, the final design of the proposed estimator indicates that higher formants provide more reliable information about vocal tract length because their observed frequencies are determined relatively less by area function shape. This evidence is further corroborated by formant sensitivity experiments presented here, which indicate that both perturbation theory and the widely-used multitube model predict similar effects. Formant sensitivity decreases as higher formants are considered, in general. However, these sensitivity differences vary as a function of constriction length. For constrictions between 0% and approximately 33% of the vocal tract length, this characterization is accurate, with the reduction in sensitivity becoming most dramatic for constrictions just below 30% of vocal tract length. On the other hand, this effect is very small for constrictions less than 10% of vocal tract length. For constrictions approximately between 33% and 48% of overall vocal tract length, this effect still holds for F_1 and F_2, but there is a reversal for F_3 and F_4. Using the perturbation theory model, deeper insights into the reasons for this effect can also be found.

One interpretation of this effect can be stated precisely by observing that Equation 3.15, which represents the sensitivity of some formant frequency to a relatively long constriction, can be interpreted as a summation over some section of the sensitivity function defined for single, short perturbations. Applying this summation across all constriction locations along the vocal tract is equivalent to applying a finite impulse response filter to the single-perturbation sensitivity function, where the filter, which is



(a) $t\Delta l$ = 6% of L (b) $t\Delta l$ = 28% of L

(c) $t\Delta l$ = 41 % of L (d) $t\Delta l$ = 50 % of L

Figure 3.6: Frequency response of the filter specified by Equation 3.18, which reflects an interpretation of vocal tract constrictions, presented in Equation 3.15, as a summation over some section of the sensitivity function defined for single, short perturbations. The transfer function depends on the length of the constriction, $t\Delta l$, which is why the frequency response for different constrictions lengths are shown. Overlaid are lines indicating the fundamental frequencies of single-perturbation sensitivity functions corresponding to the first four formant frequencies. When constriction length is long (d), magnitude of the frequency response generally decreases as the spatial frequency of the sensitivity function increases, corresponding to the overall reduction in the sensitivity of higher formants. For very short constrictions (a), this sensitivity difference is marginal. It is also possible to get very dramatic reduction of sensitivity for higher formants (b), or to most attenuate the sensitivity of non-highest formant (c). These results are highly consistent with the sensitivity range functions from the formant sensitivity experiments, as presented in Figures 3.4 and 3.5.

non-causal in the spatial domain, can be represented with numerator coefficients equal to one. In particular, the output of the filter is defined as:

$$Y[x] = S[x + \frac{t\Delta l}{2}] + \ldots + S[x + \Delta l] + S[x] + S[x - \Delta l] + \ldots + S[x - \frac{t\Delta l}{2}], \quad (3.18)$$

where x is the spatial location along the vocal tract, $S[x]$ are the input samples of the single-perturbation sensitivity function, and $Y[x]$ is the filtered long-perturbation sensitivity function. The transfer function of this filter depends heavily on the length of the constriction, $t\Delta l$. Figure 3.6 shows the frequency response of the filter for a variety of constriction lengths. Overlaid are lines indicating the fundamental frequencies of single-perturbation sensitivity functions corresponding to the first four formant frequencies. These frequencies can be determined by examining Equations 3.10 and 3.11, which show that the pressure and volume velocity profiles vary according to sine and cosine of the position along the length of the vocal tract. If perturbations of a constant area along their length are considered, as before, then the pressure and volume velocity profiles will be sinusoids with spatial frequencies (i.e., wavenumber) equal to F_n/c. The short-perturbation sensitivity function is formed by the sum of these two sinusoids after they have been full-wave rectified, resulting in a fundamental frequency of $2F_n/c$.

Examining the frequency response for the longest constriction length (d), it is possible to see that the magnitude of the frequency response generally decreases as the spatial frequency of the sensitivity function increases, corresponding to the overall reduction in the sensitivity of higher formants. For constrictions that are quite short (a), this sensitivity difference is marginal. Depending on the specific length of the constriction, it is possible to get very dramatic reduction of sensitivity for higher formants (b). In certain cases, however, it is possible to most attenuate the sensitivity of non-highest formant (c). These results are highly consistent with the sensitivity range functions from the formant sensitivity experiments.

3.5.3 Framework Extension

Perhaps the most obvious way to extend the basic estimation framework of Equation 3.4 is to add a constant offset term, β_0, which would make this a full multiple linear regression model, of the following form:

$$\hat{\Phi} = \beta_0 + \frac{\beta_1 F_1}{1} + \frac{\beta_2 F_2}{3} + \frac{\beta_3 F_3}{5} + \cdots + \frac{\beta_m F_m}{2m-1}. \tag{3.19}$$

Kirlin [Kirlin, 1978] showed that the addition of an offset term can be motivated in a probabilistic way. Specifically, whereas the maximum likelihood estimate, already discussed, can be represented in the basic framework by appropriate selection of coefficients, a maximum *a posteriori* (MAP) estimator requires the addition of an offset term to properly represent the prior. The equation for such an estimator was given by Kirlin as:

$$\hat{\Phi}_{MAP} = \frac{\sum_{n=1}^{m}(2n-1)F_n/\sigma_n^2 + \mu_0/\sigma_0^2}{\sum_{i=1}^{m}(2i-1)/\sigma_i^2 + 1/\sigma_0^2}. \tag{3.20}$$

The variables μ_0 and σ_0^2, in this case, represent the mean and variance of Φ, respectively. After manipulating the terms, it can be found that $\beta_n = \frac{(2n-1)^2}{\sigma_n^2 \sum_{i=1}^{m}(2i-1)^2/\sigma_i^2 + 1/\sigma_0^2}$, and the offset term can be expressed is $\beta_0 = \frac{\mu_0/\sigma_0^2}{\sum_{i=1}^{m}(2i-1)^2 + 1/\sigma_0^2}$.

The experiments on human speech data described in Section 3.3.4 were repeated using this extended model. During these experiments, the mean value of $\beta_{1\ldots4} = 0.036, 0.149, 0.150, 0.097$ across all resamplings, with the standard deviation < 0.005 for all coefficients, and the mean value of $\beta_0 = 329$ with a standard deviation $= 3.64$. This estimator was found to provide an RMS error of 0.946 cm, which is superior to the performance of any of the previously described estimators. It is expected, in general, that a more complex model will be able to provide better performance by providing a better fit to the data. In addition, it was noted that Equation 3.19 constitutes a multiple linear regression model that is only marginally more difficult to fit than the model presented

in Equation 3.4. Therefore, it is also possible to stay within the spirit of the proposed estimator from this work, and to fit this more complex linear model by finding the least squares solution. The experiments on human speech data described in Section 3.3.4 were repeated again using the extended model and finding the least squares solution. During these experiments, the mean value of $\beta_{1...4} = 0.033, 0.090, 0.136, 0.390$ (with standard deviations $= 0.006, 0.006, 0.014, 0.015$, respectively) and the mean value of $\beta_0 = 252$ with a standard deviation $= 5.58$. This estimator was found to provide an RMS error was 0.763 cm, which is the best performance observed in the course of the present work. Note, also, the ascending quality of the coefficient values in this estimator, which is again consistent with the relative insensitivity of higher formants.

3.6 Conclusion

A general framework for designing vocal tract length estimators was developed, beginning from the basic principles of vocal tract acoustics. It was shown that several previously proposed estimators fit within this framework. Moreover, it was shown that an estimator which is optimal in the least-squares sense can be developed in a statistical fashion with a sufficient amount of data using multiple linear regression. The proposed estimator was evaluated on both simulated and human speech data sets and was shown to outperform previously proposed estimators. A key characteristic of the proposed estimator was the increasing weight placed on higher formants, suggesting that higher formants are more reliable features for estimating vocal tract length. Direct examination of formant frequency variation in the current data sets further reinforced the idea that higher formants are less sensitive to speech articulation. These empirical findings lead to a theoretical examination of formant sensitivity, where corroborating predictions from two methods of vocal tract acoustic modeling were found. Insights from perturbation theory

further revealed that vocal tract constrictions can be interpreted as filtering the formant sensitivity functions, and the frequency response of this filter generally decreases with frequency.

Vocal tract length estimation is an instance of morphological inversion of speech – that is, attempting to predict inherent speaker-specific characteristics about the size and shape of the speech production apparatus from the acoustic signal. Morphological inversion holds promise for many technological applications where knowledge of morphological characteristics would be advantageous in analyzing the speech signal. In addition to normalizing acoustic characteristics of different speakers, accurate length estimation could lead to biometric applications, because vocal tract length may represent an individual speech characteristic that cannot be easily forged or altered. An interesting extension of the present work will be to further examine vocal tract length estimation from the dynamic perspective concerning vocal tract length. Data from rtMRI will allow us to measure vocal tract length at finer temporal scales in the future, and to do so with higher accuracy than previously possible. It would be interesting to see how the various models discussed here will perform when data concerning vocal tract length comprise precise, short-time length measurements. It will also be interesting to examine the perceptual aspects of vocal tract length estimation in humans to provide deeper insights into the success of length normalization in technological applications, and to examine situations where perception of vocal tract length can apparently be misleading or inaccurate. For instance, Black-and-White Colobus monkeys produce vocalizations that come from apparently much longer vocal tracts that they actually possess [Harris et al., 2006], and it has been suggested that human males have vocal tracts optimized to give the impression of size [de Boer, 2007], but presumably without compromising intelligibility.

Chapter 4

Statistical Methods for Estimation of Direct and Differential Kinematics of the Vocal Tract

4.1 Abstract

We present and evaluate two statistical methods for estimating kinematic relationships of the speech production system: Artificial Neural Networks and Locally-Weighted Regression. The work is motivated by the need to characterize this motor system, with particular focus on estimating differential aspects of kinematics. Kinematic analysis will facilitate progress in a variety of areas, including the nature of speech production goals, articulatory redundancy and, relatedly, acoustic-to-articulatory inversion. Statistical methods must be used to estimate these relationships from data since they are infeasible to express in closed form. Statistical models are optimized and evaluated – using a heldout data validation procedure – on two sets of synthetic speech data. The theoretical and practical advantages of both methods are also discussed. It is shown that both direct and differential kinematics can be estimated with high accuracy, even for complex, nonlinear relationships. Locally-Weighted Regression displays the best overall performance, which may be due to practical advantages in its training procedure. Moreover, accurate estimation can be achieved using only a modest amount of training data, as judged by convergence of performance. The algorithms are also applied to

real-time MRI data, and the results are generally consistent with those obtained from synthetic data.

4.2 Introduction

The kinematics of complex motor systems can be described at different levels of abstraction [Bernstein, 1967, Hollerbach, 1982, Saltzman and Kelso, 1987]. One classic example is the arm – whether human or robotic – which can variously be described as a collection joint angles or by the spatial coordinates of the end-effector. Similarly, the speech production system can be described at several levels, including muscle activations, constriction degrees/locations or formant frequencies. Thus, the choice of variables for describing a system can be low-level (i.e., close to the articulatory substrate) or high-level (i.e., removed from the articulators). This multi-level view of motor systems is common to many studies of biological motor activity [Soechting, 1982, Mottet et al., 2001], and has also been extensively studied in robotic control [Khatib, 1987, Nakanishi et al., 2008]. In order to completely characterize a motor system, it is crucial to understand the maps that relate these different kinematic levels. A fundamental understanding of the speech production system can also be built on an understanding these relationships.

The importance of these maps in characterizing motor systems is underscored by the fact that the goals of movement are often defined in terms of relatively higher-level variables, rather than at the level of the articulatory substrate. The variables which are used to define these goals can be called *task variables*, and the space defined by those variables is known as *task space*. Similarly, the lowest level of description is defined by *articulator variables* in *articulator space*. It is often convenient and desirable for a system to be controlled in task space. There is strong empirical evidence supporting the

idea that control of speech production is done at a higher level than muscle activations. For instance, it has been shown that the low-level articulators are kept compliant during speech production, allowing the achievement of higher-level tasks even despite perturbation [Abbs and Gracco, 1984, Kelso et al., 1984, Guigon et al., 2007]. Knowledge of the map between articulator space and task space is a prerequisite for being able to accomplish this.

A significant challenge for system characterization, then, is posed by the fact that these kinematic relationships are complex, nonlinear and infeasible to express in closed form for complex motor systems [Sciavicco and Siciliano, 2005]. This is certainly true of the speech production system. For instance, if one considers articulatory variables to be individual muscle activations and the task variables to be more abstract quantities like constriction degrees or formant frequencies, then the map will represent a variety of physical processes between those levels. There have been, of course, painstaking efforts to develop such models for both research and speech synthesis, built upon knowledge of vocal tract geometry [Rubin et al., 1996, Iskarous et al., 2003, Nam et al., 2004, 2006] and biomechanical knowledge [Perrier et al., 1996, Payan and Perrier, 1997, Perrier et al., 2003, Gérard et al., 2003, Fels et al., 2005, Vogt et al., 2005, 2006, Gérard et al., 2006, Winkler et al., 2011c,b]. Fortunately, it is possible to directly and statistically estimate the maps from data when they cannot be expressed succinctly. Building models in this way should be especially useful and timely in light of the accelerating availability of rich, complete kinematic data of speech (e.g., [Wrench and Hardcastle, 2000, Narayanan et al., 2004, 2011]).

Attempts have been made to statistically estimate the *direct kinematics*, which express task variables as a function of articulator variables. There has long been a need to statistically examine the direct kinematics of robotic systems for the purposes of calibration. By admitting that any idealized mathematical description of a motor system

will differ from its actual physical instantiation, there arises a need to tune the parameters of the system to ensure proper control. Traditionally, the functional form of the relationship is known *a priori*, the problem consists of estimating only the parameters of that form. This can be done statistically, using data from the system in operation. This kind of kinematic calibration is common practice in experimental robotics, and of critical importance in industrial robotic situations [Sklar, 1989, Mooring et al., 1991, Bennet and Hollerbach, 1991, Hollerbach and Wampler, 1996].

Statistical methods have also been employed to estimate the entire functional form of the forward map for a variety of motor systems. Artificial Neural Networks (ANNs) have been used to learn the direct kinematic relationships in the context of articulated robotic arms [Jordan, 1992, Jordan and Rumelhart, 1992] and simulated biological arms [Bullock et al., 1993]. Locally-linear techniques have also been employed for estimating robotic forward models [D'Souza et al., 2001, Ting et al., 2008].

Estimation of direct kinematics has also been demonstrated in the domain of speech production. Purely codebook-driven techniques have been utilized for this purpose [Kaburagi and Honda, 1998]. Clustering techniques [Shiga and King, 2004] including, perhaps most notably, Gaussian Mixture Models have been used to learn the forward map to a high degree of accuracy [Toda et al., 2004, 2008]. ANNs have also been used to estimate the direct kinematics of the speech production system [Bailly et al., 1991, Kello and Plaut, 2004], most prominently as part of developing the DIVA model [Guenther, 1994, 1995, Guenther et al., 1998]. Hidden Markov Models have also been used, usually for applications in speech synthesis [Hiroya and Honda, 2002a,b, 2003, 2004, Nakamura et al., 2006]. Locally-linear techniques have also been recently used for estimating the map between fleshpoints on the tongue and the formant frequencies of vowels [McGowan and Berger, 2009].

This work is aimed at identifying reliable algorithms that hold promise for estimating the direct and, crucially, the differential kinematics of speech production from speech articulation data. *Differential kinematics* relate velocities in task space with velocities in articulator space (i.e., the first-order partial derivatives of task variables with respect to articulator variables). These relationships have been very well studied in the robotics community as a key aspect of system characterization [Sciavicco and Siciliano, 2005]. However, differential kinematics are largely unstudied with respect to speech motor control. It was suggested by [Saltzman et al., 2006] that differential kinematics could be used to quantify the debate over the nature of speech production goals. However, we are not aware of any studies which specifically attempt to model the differential kinematics of speech production, nor do we know of any that utilize the differential kinematics for the purposes of characterizing the speech production system. We demonstrate the accuracy of Artificial Neural Networks (ANNs) and Locally-Weighted Linear Regression (LWR) by evaluating them on data relevant to speech production. We also argue for the utility of looking at speech data in this way.

Differential kinematics offer an exciting new way of looking at speech data, with the potential to offer many insights and to be useful for many applications. They form a rigorous mathematical framework for exploring systematic characterization of the speech production system in several respects. For instance, they allow for comparison of the degrees of freedom in the articulator space versus the task space (i.e., redundancy). They can facilitate the identification of localized reductions in task-space degrees of freedom (i.e., singular postures) which may occur, for instance, when articulators become perfectly aligned. They also provide a basis for inverse kinematic algorithms, deriving the equations of motion for a system, looking at the force control and interface interaction (e.g., interaction between the tongue and palate), and designing and evaluating task-space control schemes.

Differential kinematics may be of particular use in studying speech due to the long-standing debate over the nature of speech tasks – i.e., whether they are acoustical (e.g., [Guenther, 1994]) or articulatory (e.g., [Saltzman and Munhall, 1989]). This debate has been encapsulated in specific models of motor coordination, including the DIVA model [Guenther, 1995, Guenther et al., 1998] and the Task Dynamics account [Saltzman and Munhall, 1989, Saltzman and Byrd, 2000], which are based on different assertions about how best to describe the task space. Using differential kinematics, one can apply computational methods such as the Uncontrolled Manifold Method (UCM) [Scholz and Schöner, 1999], recently suggested by [Saltzman et al., 2006] as a way to quantify the debate concerning the nature of the speech production tasks. The UCM takes advantage of the fact that, using the differential kinematics, one can divide the observed kinematic variability into that which is relevant to a given task, and that which is not. Given two competing task descriptions, the better one will show a higher proportion of variability in the task-relevant portion.

The potential also exists for differential kinematics to inform new methods for acoustic-to-articulatory inversion – i.e., the problem of recovering vocal tract configurations given only speech acoustics. Early, analytical approaches to this problem showed the apparent redundancy in the system, which implies that the forward map is non-invertible [Mermelstein and Schroeder, 1965, Wakita, 1973]. Later techniques attempted to resolve this ambiguity, either by using statistical properties of the map [Atal and Rioul, 1989, Papcun et al., 1992, Hogden et al., 1996, Qin and Cerreira-Perpiñán, 2007, Lammert et al., 2008, Ananthakrishnan et al., 2009, Qin and Cerreira-Perpiñán, 2010, Ghosh and Narayanan, 2010] or through analysis-by-synthesis [Atal et al., 1978, Boë and Bailly, 1992, Schroeter and Sondhi, 1994, Panchapagesan and Alwan, 2011]. Statistical methods similar to those explored here have been utilized to learn the inverse

map directly [Rahim et al., 1991, Richmond, 2010, Al Moubayed and Ananthakrishnan, 2010].

While these inversion efforts have yielded substantial progress, differential kinematics can offer new insight by way of identifying and quantifying redundancy in the system. More importantly, new inversion techniques should be possible which confront the non-uniqueness problem in novel ways. Many iterative, computational methods have been developed in the robotics community for finding a reasonable path through articulator space to reach a position in task space, even despite nonuniqueness. Some common solutions utilize the differential kinematics to accomplish this, including Jacobian pseudoinverse methods [Whitney, 1969], the Jacobian transpose method [Balestrino et al., 1984, Wolovich and Elliot, 1984] and damped least squares methods [Nakamura and Hanafusa, 1986, Wampler, 1986].

In order to estimate the direct and differential kinematics of the speech production system, one must choose a statistical method that is capable of (1) estimating complex, nonlinear relationships to a high degree of accuracy, and (2) facilitating the extraction of partial derivatives that make up the differential kinematics, preferably directly and without appealing to numerical approximations. These are the key technical challenges for any candidate method. In addition, it is desirable to have a method which is easily designed and trained. These practical challenges are of equal importance for the utility of a chosen method. We explore the use of ANNs and LWR, two methods which represent drastically different underlying assumptions in terms of model fitting and prediction.

ANNs were recently suggested for this purpose by [Saltzman et al., 2006]. Their abilities as universal function approximators allows them to learn complex maps [Cybenko, 1989, Hornik et al., 1989]. However, ANNs have some practical drawbacks, in the sense that they are notoriously difficult to design and slow to train [Bishop, 2006,

Wilamowski et al., 2008]. Training is computationally expensive, involving optimization of a non-convex objective function, and it can be difficult to determine training convergence in practice. If new data arrive, training must be repeated.

LWR offers a complementary set of advantages to ANNs. It has many practical advantages, most notably efficiency of training and the presence of fewer free parameters. The method is powerful enough to approximate very complex functions, even despite assumptions of local linearity. Indeed, local linearizations are ubiquitous in kinematics and control applications because they are entirely appropriate, practical approximations to the kinds of nonlinearities seen in many motor systems. The relationship between velocities in articulator and task space is often expressed mathematically as a linear transformation (see Equation 4.3, below).

We have previously reported on an initial effort to estimate kinematic relationships from data [Lammert et al., 2010]. This paper constitutes an expansion of that work, providing a substantial refinement of the techniques previously presented. To that end, we have refined our previous formulation of LWR. We have also implemented a heldout data validation scheme for parameter optimization with respect to both algorithms. This has resulted in more confidence about the generalizability of the results and, crucially, in superior estimation accuracies according to our previous evaluation metrics. We have also included additional evaluations to assess our modeling of kinematic aspects which were neglected in previous work. Additionally, we have tested our methods on an additional data set, which was designed to more accurately reflect data which might be acquired in real speech studies. Finally, we present a more thorough discussion of the motivations for kinematic analysis, as well as its potential applications for studying speech production.

Section 4.3 will serve to explain our methodology, including a formal description of direct and differential kinematics in Section 4.3.1, the creation of our data sets in

Sections 4.3.2 and 4.3.3, as well as a review of ANNs and LWR in Sections 4.3.4 and 4.3.5. An explanation of our model optimization procedure is in Section 4.3.6 and an overview of our experiments for evaluation is in Section 4.3.7. In Section 4.4, we provide the results of our experiments: Section 4.4.1 provides the results on two synthetic data sets in terms of accuracies in estimating direct and differential kinematic relationships, while Section 4.4.2 provides direct kinematic accuracies on a real speech data set. In Section 4.5, we discuss the performance of ANNs and LWR, and in Section 4.6, we present our concluding remarks and remaining challenges for this line of work.

4.3 Methods

Broadly stated, our methodology involves the generation of two large corpora of parallel articulator and task vectors using an articulatory model of the vocal tract. We then employ both ANNs and LWR to estimate the direct and differential kinematics of this model from the data. Parameter tuning was done by examining accuracy of the direct kinematic estimation on a development set of heldout data. Evaluation was then performed by examining accuracy of both the direct and differential kinematics on a heldout test set. In this section, we provide detail about the theory and implementation of our methods, as well as the specific rationale for our choices.

We would like to be especially clear regarding the motivation for using synthetic data, given that this study is ultimately aimed at developing methods which can be used on real data. Indeed, we intend to apply these methods to real data, which we discuss in our concluding remarks (see Section 4.6). The use of synthetic data is driven by the need to evaluate the differential kinematic estimates. These cannot be evaluated on real data since the relevant relationships are not directly observable, in contrast to direct kinematics which can be precisely measured and for which the accuracy can be

evaluated by calculating the residual error in task space. For synthetic data, on the other hand, the differential kinematics are known *a priori* and can be used as a standard for evaluation.

With this in mind, we also develop and demonstrate methods for parameter tuning (i.e., model selection) and model training that do not depend on knowing the differential kinematics, and therefore can be performed on real speech data where knowledge of these aspects is not available. We also wish to emphasize that deriving estimates of the direct and differential kinematics in the way we demonstrate is completely repeateable on real articulatory data.

4.3.1 Direct and Differential Kinematics

Given a vector q, representing n low-level articulator variables of the system, and a vector x, representing m high-level task variables of the system, the relationship between them is commonly expressed by the direct kinematics equation, of the form:

$$x = k(q) \tag{4.1}$$

where the function $k(\cdot)$ represents the forward map which is assumed to be complex and nonlinear. It is the forward map that we are attempting to learn from a representative corpus of parallel example pairs of q and x.

We are particularly interested in modeling $k(\cdot)$ so as to facilitate derivation of the Jacobian matrix:

$$J(q) = \begin{pmatrix} \partial x_1/\partial q_1 & \cdots & \partial x_1/\partial q_n \\ \vdots & \ddots & \vdots \\ \partial x_m/\partial q_1 & \cdots & \partial x_m/\partial q_n \end{pmatrix} \tag{4.2}$$

The Jacobian is a compact representation of knowledge regarding the posture-specific 1^{st}-order partial derivatives of tasks with respect to articulators. It allows us to write the differential kinematics equation, which relates articulator velocities and task velocities in the following way:

$$\dot{x} = J(q)\dot{q} \tag{4.3}$$

Note that this equation expresses the relationship between $\dot{q}$ and $\dot{x}$ as a linear transformation. This kind of approximation allows for such an elegant mathematical formulation of the kinematics. It also highlights why locally-linear methods, such as LWR, are appropriate for estimating kinematic relationships.

It is possible that further approximations, if appropriate, may allow additional simplification of the mathematics. In certain applications (e.g., [Cootes et al., 2001]), it is appropriate to assume that the differential kinematics are globally constant. This means that the Jacobian is no longer a function of the pose (i.e., $J(q) \approx J$ and $\dot{x} = J\dot{q}$). We do not believe that this approximation is appropriate for the speech production system, but future work may help determine this (see Section 4.6).

As mentioned above (see Section 4.2), an understanding of differential kinematics allows one to address fundamental issues of system characterization. Indeed, the Jacobian is one of the most useful tools for characterizing systems. For example, postural singularities can easily be identified by examining the rank of the Jacobian. If it is rank-deficient, then there is a singularity at the current posture. Similarly, an examination of the range and null of the Jacobian provides a formal analysis of articulatory redundancy. The Jacobian also represents sufficient information about the kinematics to apply the Uncontrolled Manifold Method.

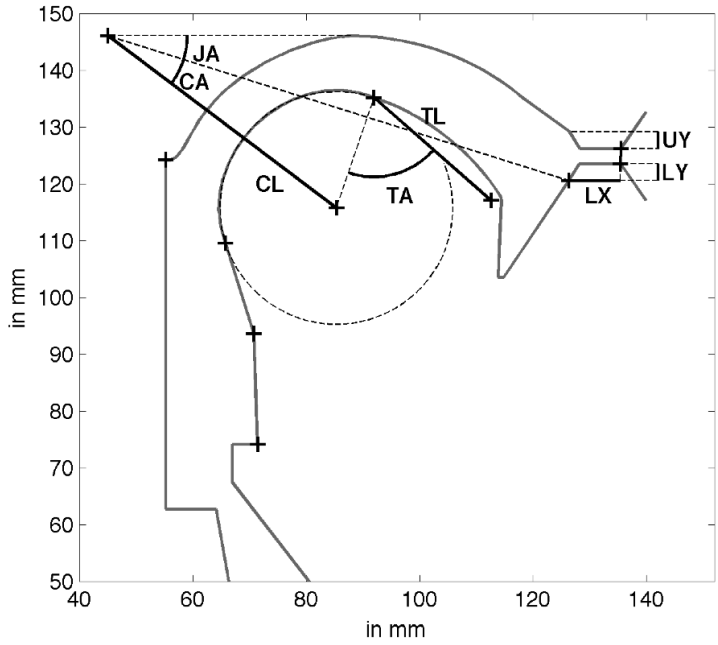


Figure 4.1: A visualization of the Configurable Articulatory Synthesizer (CASY) in a neutral position, showing the outline of the vocal tract model (gray line). Overlain are the key points (black crosses) and geometric reference lines (dashed lines) used to define the articulatory parameters (black lines and angles), which are also labeled.

4.3.2 Kinematic Model

The model we utilize in generating our data is the TAsk Dynamic Application (TADA) [Nam et al., 2004, 2006], an implementation of the Task Dynamic model of articulatory control and coordination [Saltzman and Munhall, 1989]. The specific articulatory model at the core of TADA is the Configurable Articulatory Synthesizer (CASY) [Rubin et al., 1996, Iskarous et al., 2003], which specifies geometric articulator variables.

As outlined at the beginning of this section, we are using synthetic articulatory data to facilitate evaluation of the differential kinematics. We will not make any bold claims about the realism of TADA as a model for speech production, although we believe that is remains faithful to real speech articulation in certain key ways. Evidence for this assertion comes from work using TADA for analysis-by-synthesis [Mitra et al., 2009,

Nam et al., 2010, Mitra et al., 2010, 2011]. Articulatory time series extracted from this process have been shown to improve automatic speech recognition, which would be unlikely if correspondence was poor between TADA's articulation and real speech.

We note that CASY/TADA is only one possible implementation of a kinematic model for speech production. The crucial aspect for our purposes is only that we use a model that is well-informed about the kinds of nonlinearities that are relevant for speech. The forward model incorporated in TADA is such a model. It is based on the geometrical equations of [Mermelstein, 1973] for describing vocal tract configurations and deriving vocal tract outlines from articulator positions. Assuming a simple parameterized shape for the palate and rear pharyngeal wall, the constriction task variables can also be derived from the outline.

The forward model implemented in CASY uses geometric articulator variables which are shown in Figure 4.1. These include lip protrusion (LX) and vertical displacements of the upper lip (UY) and lower lip (LY) relative to the teeth. The jaw angle (JA) and tongue body angle (CA) are defined relative to a fixed point above and behind the velum, as is the tongue body length (CL). The tongue tip is defined by its length (TL) from the base to the tip, as well as its angle (TA) relative to the tongue body center and the posterior base of the tongue blade.

Using the vocal tract outline, it is possible to calculate a vector of task variables for each articulator configuration. These tasks are articulatory in nature, but are high-level relative to the articulator variables. The tasks are as follows: lip aperature (LA) and protrusion (PRO), tongue body constriction degree (TBCD) and location (TBCL), as well as tongue tip constriction degree (TTCD) and location (TTCL).

The complexity of the model varies depending on the task variable in question. All tongue-related tasks are quite complex and nonlinear, with the most prominent nonlinearities conforming to trigonometric and polynomial functions. The tongue tip tasks

are the most complex, due to many nonlinearities and to having the greatest articulatory redundancy (five different articulators contribute to them). At the other extreme, the lip protrusion task variable depends only on the lip protrusion articulatory variable in a linear way (i.e., a linear, non-redundant relationship). All other task variables are related to the articulators by at least two nonlinearities. The full list of equations is given in A.

TADA is an implementation of precisely those dynamical equations described by [Saltzman and Munhall, 1989] in their dynamical approach to control of speech production gestures. A full exposition of the model can be found in referring to Appendix 2 of that publication. In brief, it states that the equation of motion for the actively controlled articulators is:

$$\ddot{q}_A = J * (M^{-1}[-BJ\dot{q} - K\Delta x(q)]) - (J * \dot{J}\dot{q} + (I_n - J * J))\ddot{q}_d \tag{4.4}$$

where $\ddot{q}_A$ represents the articulatory acceleration vector, which encapsulates the active driving influences on the model articulators. Also, $\ddot{q}_d = B_n\dot{q}$ is the acceleration damping, where B_n is a diagonal matrix of damping constants. The other variables include M, a diagonal matrix of inertial coefficients, B, the diagonal matrix of damping coefficients for the task variables, K, the stiffness coefficients for the task variables, and J, the Jacobian matrix.

4.3.3 Data Sets

We generated two large codebooks of data using the CASY/TADA forward model. These two codebooks provide complementary ways of populating the articulator space. We first uniformly populated the articulator space (i.e., data points conformed to a grid

Articulatory Variable	min	max	range	units
Lip Protrusion (LX)	9.11	12.00	2.89	mm
Jaw Angle (JA)	1.11	1.41	0.31	rad
Upper Lip Displacement (UY)	-4.78	0.93	5.70	mm
Lower Lip Displacement (LY)	-8.57	19.95	28.52	mm
Tongue Body Length (CL)	68.59	83.62	15.03	mm
Tongue Body Angle (CA)	-0.36	0.04	0.40	rad
Tongue Tip Length (TL)	6.50	44.38	37.88	mm
Tongue Tip Angle (TA)	-0.24	1.40	1.67	rad

Table 4.1: Ranges of the various CASY articulator variables, as observed during synthesis of the speech-relevant data set described in Section 4.3.3.

within the space). The filled space was bounded by the articulator ranges obtained from the speech-relevant data (see below and Table 4.1). Within those bounds, we defined a rectangular grid with 5 evenly spaced points along each of the 8 dimensions. This created $5^8 = 390{,}625$ total articulator vectors and their accompanying $390{,}625$ task vectors of length 6. This uniformly-distributed codebook represents all possible vocal tract configurations with equal density of coverage. Building this data set provides a way of testing our estimation methods that is independent of TADA's ability to move CASY's articulators in a realistic way. It also provides broad coverage of the articulator space, which should lead to thorough testing of the modeling methods.

It seems very unlikely, however, that real articulatory recordings of natural speech would reflect this kind of data distribution. We would expect that real speech would fill only a subspace of the full articulator space due to inter-articulator correlations and the likelihood that certain configurations are not used for speech. Thus, it is equally important to evaluate estimation accuracy on a codebook that is representative of real speech data which might be acquired from speaking subjects.

For this reason, we used CASY/TADA to synthesize a set of English sentences. We used TADA to synthesize 30 English sentences taken from the MOCHA-TIMIT corpus [Wrench and Hardcastle, 2000] at an effective sampling rate of 200 Hz. The specific

sentences constitute the first 30 sentences of that corpus, and have been reproduced in C. For reference, the resulting articulator ranges are shown in Table 4.1. This synthesis provided us with speech-relevant codebook containing 17,198 total articulator vectors of length 8. These were accompanied by the same number of task vectors of length 6.

From these two large codebooks, we created data sets for evaluation by randomly sampling vectors without replacement. Data sets were of sizes 78, 156, 312, 625, 1250, 2500 and 5000. Having data sets of standard sizes facilitated a direct comparison between both algorithms. The reason for producing data sets of varying sizes was to determine (1) whether the amount of training data affected the accuracy of estimation (i.e., training effects) and (2) whether the overall amount of training data was sufficient (i.e., convergence of performance). Although the largest data set seems modest in size, we were able to show that it is sufficiently large (see Section 4.5).

4.3.4 Artificial Neural Networks

In keeping with the suggestion of [Saltzman et al., 2006], we implemented a directed, multilayer, feedforward neural network, otherwise known as a multilayer perceptron (MLP) [Bishop, 2006]. We acknowledge that many alternative ANN architectures and topologies exist, from mixture of experts networks [Jacobs et al., 1991, Jordan and Jacobs, 1995] to fully connected networks [Wilamowski et al., 2008]. These are equally capable of learning arbitrarily complex functions. However, the advantage of using the MLP architecture is in the ability to easily and analytically (i.e., without the need for numerical methods) extract the Jacobian. The method for doing this is described below.

Our network hidden nodes employed the commonly-used sigmoidal transfer functions. It is possible to tailor ANNs to a specific estimation task by choosing different nonlinear transfer functions, especially if the functional form of the map is known *a priori*. Many different functions can be used while still maintaining the universal

approximation power of MLPs [Duch and Jankowski, 1999]. Moreover, as long as these nonlinearities are differentiable, it is still possible to easily extract the Jacobian. Our intention here was to choose a transfer function that offers the most generality, so as to make minimal assumptions about the structure of real speech data. When it comes to applying these techniques on real speech data, this kind of generality and flexibility will hopefully be a benefit.

We trained the ANNs with the standard error backpropagation [Rumelhart et al., 1986b,a, Jordan, 1992]. We acknowledge that more efficient training algorithms have been suggested for ANNs with an MLP-type topology. The Levenberg-Marquardt algorithm [Hagan and Menhaj, 1994, Toledo et al., 2005, J.-M., 2008] is widely considered to be the most efficient algorithm for training MLP architectures. Still, there is no guarantee that these algorithms provide greater accuracy after training. Since our primary concern is accuracy and not speed of training, we did not implement these methods.

The error function that backpropagation attempts to minimize through gradient descent is a standard least squares function of the form:

$$E_n = \frac{1}{2} \sum_k (\hat{x}_{nk} - x_{nk})^2 \tag{4.5}$$

where the quantity $\hat{x}_{nk}$ is the output of the network at unit k when presented with input data point n.

Our network topology is represented in Figure 4.2. This network had linear input and output nodes, corresponding to each of the articulatory and task variables, respectively. Between were two layers of hidden units, all with sigmoidal activation functions:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{4.6}$$

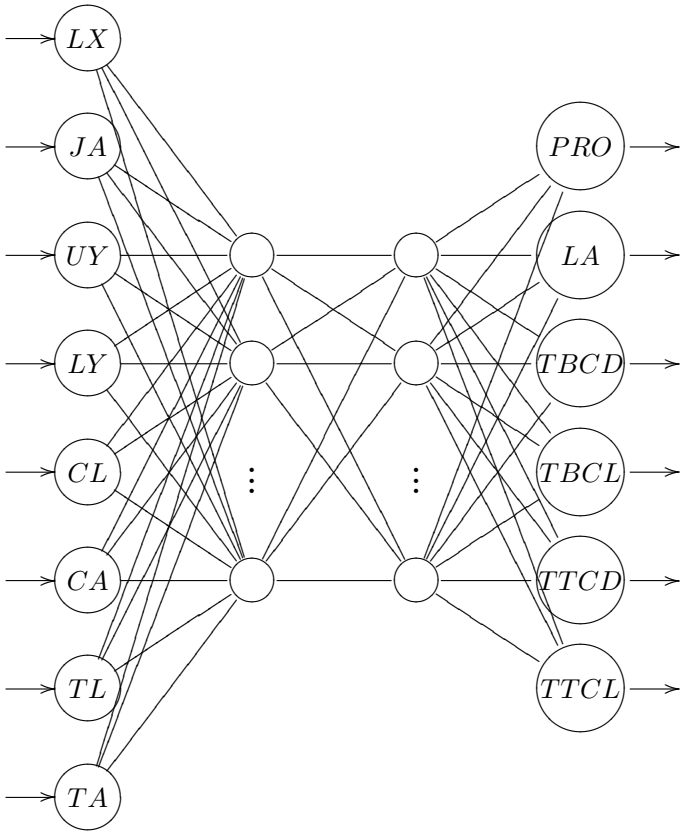


Figure 4.2: Our ANN topology, an instance of a fully-connected, feedforward Multilayer Perceptron with two hidden layers. The number of nodes in the hidden layers is a free parameter which determines the complexity of the model, after training.

Upon completion of training, a pose-specific Jacobian matrix can be obtained for any articulatory input vector. This can be done with the use of numerical methods [Bishop, 2006]. However, the Jacobian can also be obtained analytically for a network of this kind using the feedforward formalism [Jordan and Rumelhart, 1992, Saltzman et al., 2006]. Note that each hidden node has a sigmoidal activation function, we can write the derivative of each node's activation with respect to its input as follows:

$$\delta = z(1 - z) \tag{4.7}$$

Then, we can arrange these values in a diagonal matrix, denoted Δ_i. The Jacobian for a given input posture is then

$$J = \Delta_{out} W_{out,H2} \Delta_{H2} W_{H2,H1} \Delta_{H1} W_{H1,in} \tag{4.8}$$

where W_{ij} is the weight matrix connecting layer i to layer j.

General design parameters of networks with this architecture are many, and include the learning rate (α), the number of training iterations (num_{iter}), the number of hidden layers and the number of nodes in each hidden layer. It is possible in theory to tune each of these parameters to optimal values, but the tuning procedure becomes intractable with so many free parameters. The complexity of the model is essentially determined by the number of nodes in each hidden layer, which we denote p_{ANN}. Thus, we chose to treat only this as a free parameter within the scope of parameter tuning for ANNs. Other design parameters were fixed, so that $num_{iter} = 300$, α = 0.001, and the number of hidden layers was always 2. Limiting the free parameters in this way still allows for proper tuning to be performed, so as to promote generalizability of the statistical model (see Section 4.3.6), while keeping this procedure tractable.

4.3.5 Locally-Weighted Regression

Locally-weighted linear regression is one outcome of a long line of research into non-parametric methods which use locally-defined, low-order polynomials to approximate globally nonlinear functional relationships. Much of this early work is contained in the Statistics literature, where these techniques have a long successful history [Cleveland, 1979, Cleveland and Devlin, 1988, Cleveland et al., 1988]. [Atkeson et al., 1997] surveyed much of the early work on this topic from the Statistics literature, and also provided a unifying view of these techniques for the Machine Learning community.

LWR is a memory-based, lazy learning method, which means that it keeps the entire data set in memory and uses it directly at prediction time in order to calculate the parameters of interest. Formulation of this technique begins by assuming that the data were generated by a model following

$$x_i = k(q_i) + \epsilon \tag{4.9}$$

where k is a function which can be nonlinear, in general. The value ϵ represents the noise which is assumed to follow a Gaussian distribution

$$\epsilon \sim N(0, \sigma^2) \tag{4.10}$$

a normal distribution with mean 0 and variance σ^2.

We would like to fit the data in a local region defined by the data point q_i. The measure of locality K is taken to be a Gaussian kernel function

$$K(q_i, q_j, h) = exp\{-(q_i - q_j)^T H(q_i - q_j)\} \tag{4.11}$$

although any such kernel can be utilized. H is a positive semi-definite diagonal matrix, with diagonal elements equal to $1/2h^2$. The value h is a free parameter with a straightforward interpretation: it is the standard deviation of the Gaussian kernel. If h has the same value in each column in H (i.e., in all directions in articulatory space), the kernel will be spherical. Since the articulators in this case have a variety of ranges and units, we chose to set h differently in each direction, which gives one value for each articulator variable.

We assume that a linear model is an appropriate approximation to the forward map within the local region. Thus, the model we would like to fit locally is of the form

$$x_i = \beta_i^T q_i \tag{4.12}$$

where β is the vector of regression coefficients.

The error function that needs to be minimized is an extension of the standard weighted least squares function of the form:

$$E_i = \frac{1}{2}\sum_j [(\beta_i^T q_i - x_i)^2 K(q_i, q_j, h)] + \frac{\lambda}{2}\|\beta\| \tag{4.13}$$

The second term is a regularization term, which contains the ridge regression parameter γ. In cases when there are very few data points near q_i, a danger is that the regression matrix may become nearly singular and numerical issues will arise in computing the solution. Adding the regularization term prevents this problem at the expense of biasing the solution very slightly. In practice, the parameter γ can be effective even when $\ll 1$, which ensures a marginal bias of the solution.

An analytical solution can be found for β with ridge regression as follows:

$$\beta_i = (Q^T W_i Q + \gamma I)^{-1} Q^T W_i X \tag{4.14}$$

The matrix, W_i, is a diagonal weight matrix, formed from the outputs of the kernel function with a fixed q_i.

An illustrative example is shown in Figure 4.3 for a toy set of low-dimensional non-linear data. The Gaussian kernel is visualized along with a locally-linear model for one point in articulator space A global fit line, also visualized, can be obtained as a agglomeration of many locally-linear models. This is done by simply identifying an arbitrary

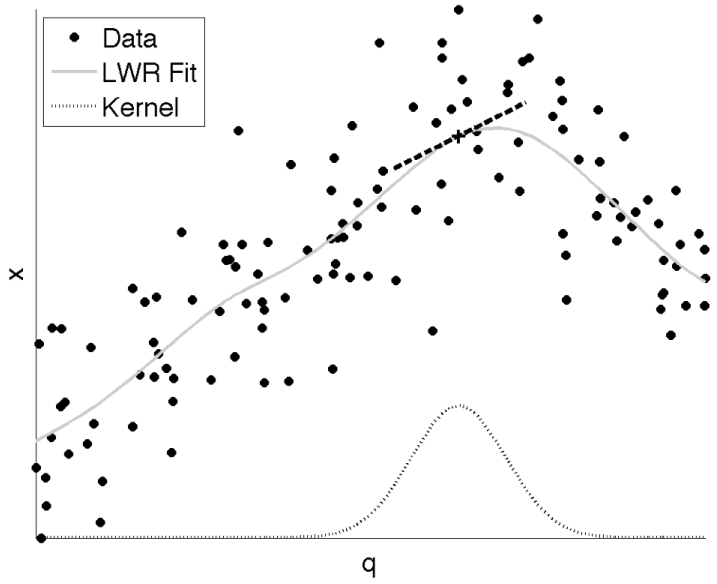


Figure 4.3: An illustration of modeling with LWR. For a particular point (black cross) a local region is defined in articulator space by a Gaussian-shaped kernel (gray dashed curve). A line is fit in the local region using a weighted least-squares solution, indicated by the black dashed line. The global fit is generated by repeating this procedure at a large number of local regions. The resulting fit can be quite complex (gray curve), and depends on the width of the kernel.

number of points in articulator space and solving for β_i and x_i at each of them. The locally-linear solutions can be said to constitute a global fit to the data.

Obtaining the Jacobian from this model is trivial, since it is already linear. The regression vector β contains the locally-relevant partial derivatives. In other words, the values of this vector are the elements of the Jacobian.

The only design parameters for LWR are the kernel width parameters, h. The values determine the complexity of the resulting model, much like the number of hidden nodes in the ANN. Therefore, these parameters must be subject to careful tuning in the evaluation experiments. While in principle, h can vary freely in each direction of articulator space, this would make the tuning search space quite large. At the same time, we can assume that the kernel width will be similar in each direction if only it is normalized by the articulator range. Thus, we define a hyperparameter p_{LWR}, such that, for articulator

j, the kernel width is $h_j = p_{LWR} \cdot range(Q_{*,j})$. This heuristic approach allows us to collapse the parameter tuning process into a 1-dimensional problem of tuning the hyper-parameter. Note that a more complex model is indicated by *smaller* values of p_{LWR}, the opposite of p_{ANN}.

4.3.6 Model Selection

An important practical challenge in applying these techniques is determining good values for the free parameters p_{ANN} and p_{LWR}. This is the problem of model selection, which is common to many statistical analyses or model fitting applications. For instance, if we are planning to use LWR on a given data set in order to estimate the Jacobian at a posture of interest, it is necessary to select a value for p_{LWR} which provides a close fit, but which promotes generalizability by avoiding overfitting. Finding this critical value is commonly known as the bias-variance tradeoff.

Optimization of the parameter value can be done using a heldout data validation procedure, whereby the data set is partitioned into a training and development set. Using a wide range of values for the free parameter, the training set is used to provide an estimate for each data point in the development set. The parameter value which provides the highest accuracy is selected as the optimal value. Heldout data validation is a principled method for model selection. However, we note that a rotating heldout validation procedure (e.g., n-fold or leave-one-out cross-validation), while deemed excessive in this situation, would be more robust on real data.

We implemented this kind of model selection procedure by randomly assigning 90% of the vectors in a given data set to a training subset and 5% to a development subset. The remaining 5% were set aside in a separate test set, to function as our postures of interest for the purposes of evaluation later on. Performance on the development set, given a particular parameter value, was determined by calculating the mean across

all tasks of the normalized root mean squared error (RMSE). Note that the differential kinematics are not used at all in the model selection process. Model selection can be done in terms of the direct kinematics. There is, therefore, no reason why this procedure cannot be repeated on entirely real speech data, when the Jacobian is not available for the purposes of development.

For the uniformly-distributed data set, it was determined that the optimal network topology had hidden layers containing 200 nodes each (i.e., $p_{ANN} = 200$). The optimal value for p_{LWR} was determined to be 0.300, corresponding to a Gaussian kernel with standard deviation equal to 30.0% of each articulator's range. For the speech-relevant data set, the optimal network had $p_{ANN} = 250$, and LWR was optimal with $p_{LWR} = 0.067$, corresponding to a Gaussian kernel with standard deviation equal to 6.7% of each articulator's range. It is notable that there was a shift in parameter values between the two data sets, such that models were allowed to become more complex on the speech-relevant data. This is likely due to the speech-relevant data being more densely packed in the articulatory space. This allows more detail about the kinematics to be obtained. The relatively sparser uniformly-distributed data requires more generalization in order to avoid overfitting.

To illustrate the model selection procedure, the performance of both algorithms are shown in Figure 4.4 over a range of parameter values on a speech-relevant development set. It should be noted that the axis values for p_{ANN} and p_{LWR} are reversed with respect to each other. This was done to consistently indicate more complex models (represented by large values of p_{ANN} and small values of p_{LWR}) toward the right-hand side of the plot. For LWR, the errors form a smooth and apparently convex function with a unique minimum over the range of explored values. This inspires confidence that the parameter value selected is the optimal one. The errors for the ANNs suggest a similar trend, but with a much more jagged appearance. This most likely reflects the difficulties associated

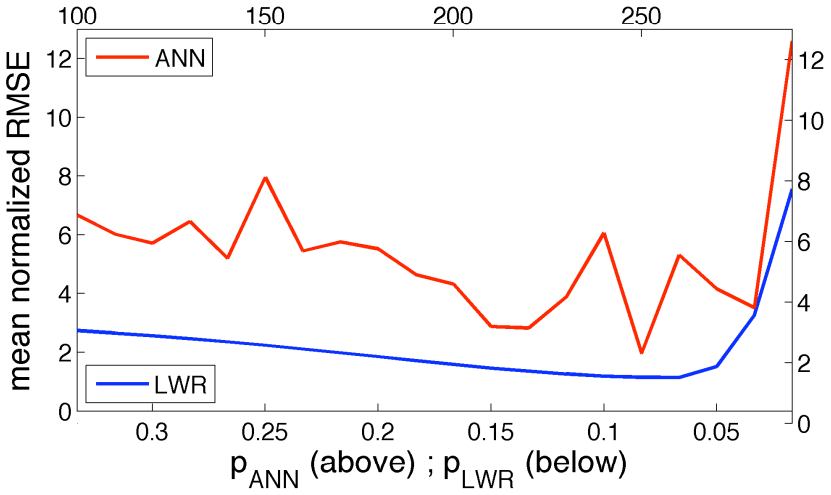


Figure 4.4: A comparison of LWR and ANN performance on held-out speech-relevant data as a function of their free parameter values. Each displays a minimum point, which represents the optimal value for that parameter. Choosing other values runs the risk of either overfitting or underlearning. Note that the axis values for p_{LWR} are reversed, so that the right-hand side indicates more complex models for both parameters.

with training ANNs. That is to say, backpropagation likely got stuck in a local minimum and did not successfully converge on the global minimum of the objective function. Still, the overall shape of the tuning curve is highly consistent with that displayed by LWR.

4.3.7 Evaluation

After performing model selection, we used the optimal parameter values to perform a final evaluation on the test set (see Section 4.3.6 for a description of the development and test sets). The entire procedure, including model selection and evaluation, was done identically on uniformly-distribute and speech-relevant data sets of all sizes.

Direct Kinematic Accuracy – Uniform Data

Task Variable	RMSE$_{ANN}$	RMSE$_{LWR}$
PRO	2.78 mm (92.6 %)	0.06 mm (2.1 %)
LA	3.13 mm (4.8 %)	0.13 mm (0.2 %)
$TBCL$	24.14 deg (12.8 %)	20.37 deg (10.8 %)
$TBCD$	4.38 mm (13.3 %)	2.01 mm (6.1 %)
$TTCL$	12.85 deg (4.1 %)	20.69 deg (6.6 %)
$TTCD$	4.34 mm (4.5 %)	2.03 mm (2.1 %)

Direct Kinematic Accuracy – Speech-Relevant Data

Task Variable	RMSE$_{ANN}$	RMSE$_{LWR}$
PRO	0.40 mm (13.8%)	0.05 mm (1.6%)
LA	0.38 mm (1.2%)	0.31 mm (1.0%)
$TBCL$	2.00 deg (2.3%)	1.08 deg (1.2%)
$TBCD$	0.36 mm (2.0%)	0.18 mm (1.0%)
$TTCL$	0.67 deg (0.9%)	0.42 deg (0.5%)
$TTCD$	0.67 mm (2.0%)	0.40 mm (1.2%)

Table 4.2: Accuracies of the direct kinematic estimates for both algorithms on each data set. Displayed are the root mean squared error (and RMSE as a percentage of the task range) across all vectors in the test sets.

4.4 Results

4.4.1 Synthetic Data

Our experimental results were evaluated with respect to the accuracy of modeling the direct and differential kinematics. Methods for quantifying this accuracy can be found by inspecting the direct and differential kinematics equations, as written in Equations 4.1 and 4.3, though assessment of each is done differently.

Since an accurate model of direct kinematics should constitute the forward map, $k(q)$, accuracy can be defined in terms of its prediction of the task vector, x, given an articulator vector, q. This can be quantified in terms of the root mean square error (RMSE) for predicting all task vectors in the test set. The task-wise accuracies are shown in Table 4.2 for both algorithms on the two 5000-point data sets. Also displayed

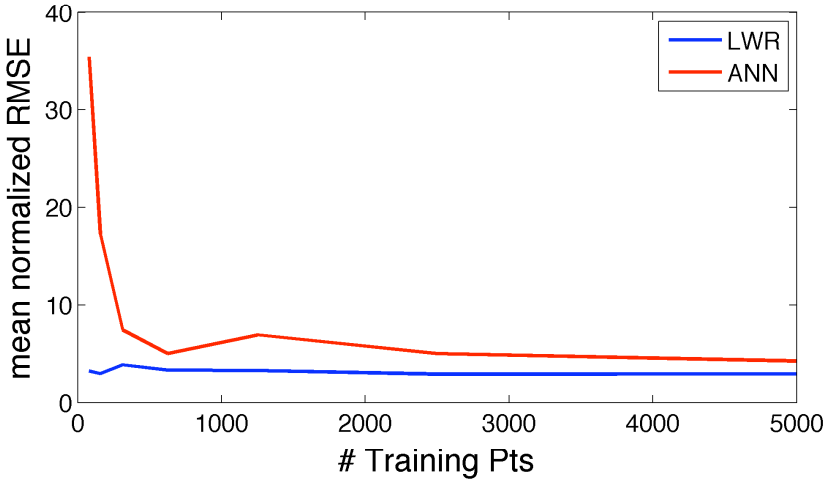


Figure 4.5: A comparison of LWR and ANN performance across a variety of data set sizes. It is clear that LWR outperforms ANN no matter the quantity of training data. Moreover, LWR is very insensitive to the amount of training data, unlike the ANN. Moreover, both algorithms seem to plateau around 3000 data points, indicating that the amount of training data we used is sufficient.

are the normalized RMSE values, reflecting the errors as a percentage of the observed task range. Figure 4.5 shows a comparison of LWR and ANN performance on data sets of varying sizes drawn from the uniformly-distributed codebook.

Accuracy in estimating the differential kinematics can be assessed directly in terms of the Jacobian matrix, $J(q)$, for any articulatory configuration, q. In turn, this can be judged by comparing the relative contributions of all articulators to each task. Mathematically, this is done by calculating the vector angle between each row of the estimated Jacobian with its corresponding row in the analytically-derived Jacobian from CASY/TADA. Relatedly, and perhaps more interpretably, one can quantify this using Pearson's r. Tables 4.3 and 4.4 display the average angle and correlation values for both algorithms operating on both 5000-point data sets. Also displayed are the row-wise Euclidean vector norms of the estimated Jacobian rows and the analytically-derived Jacobian. Comparing these values gives an indication of how well the overall magnitude of the articulator contributions was estimated for each task.

Diff Kin Accuracy – ANN, Uniform Data

J_i	r	$\angle(J_{A,i}, J_{N,i})$	$\| J_{A,i} \|$	$\| J_{N,i} \|$
J_{PRO}	0.35	64	1.00	1.22
J_{LA}	0.89	27	1.77	2.08
J_{TBCL}	0.65	39	9.21	5.66
J_{TBCD}	0.74	42	1.04	1.11
J_{TTCL}	0.74	40	4.66	3.80
J_{TTCD}	0.71	37	1.77	1.87

Diff Kin Accuracy – LWR, Uniform Data

J_i	r	$\angle(J_{A,i}, J_{R,i})$	$\| J_{A,i} \|$	$\| J_{R,i} \|$
J_{PRO}	1.00	0	1.00	0.88
J_{LA}	1.00	1	1.77	1.75
J_{TBCL}	0.70	35	9.21	6.19
J_{TBCD}	0.90	23	1.04	0.84
J_{TTCL}	0.76	35	4.66	3.63
J_{TTCD}	0.95	14	1.77	1.56

Table 4.3: The mean correlation coefficient (r) between the rows of the analytically-derived Jacobian (J_A) and the Jacobian estimated with the ANN (J_N) and LWR (J_R) from the uniformly-distributed data set. Also shown are the mean angle in degrees, the norm of each vector. The abbreviations that instantiate the subscript i of J correspond to the task variables described in Section 4.3.2

4.4.2 Real Speech Data

In preliminary experiments, we have applied the proposed methods to real speech data [Lammert et al., 2011c]. In particular, we applied these methods to a subset of the real-time MRI (rtMRI) data presented in [Narayanan et al., 2011] from one speaker of American English (5000 data points = 65 sentences at 23.33 frame/second). We extracted articulator and task variables, corresponding to those described above (see Section 4.3.2), from midsagittal vocal tract outlines automatically fit to the rtMRI images [Bresch and Narayanan, 2009]. Outlines fit with this method are pre-segmented into anatomical structures (e.g., tongue, lips, etc.) that facilitate the extraction of articulator and task

Diff Kin Accuracy – ANN, Speech Data

J_i	r	$\angle(J_{A,i}, J_{N,i})$	$\parallel J_{A,i} \parallel$	$\parallel J_{N,i} \parallel$
J_{PRO}	0.76	43	1.00	0.29
J_{LA}	0.79	38	1.78	1.45
J_{TBCL}	0.95	15	9.33	4.31
J_{TBCD}	0.94	20	1.03	1.04
J_{TTCL}	0.95	15	4.86	2.50
J_{TTCD}	0.89	22	2.09	1.70

Diff Kin Accuracy – LWR, Speech Data

J_i	r	$\angle(J_{A,i}, J_{R,i})$	$\parallel J_{A,i} \parallel$	$\parallel J_{R,i} \parallel$
J_{PRO}	0.84	27	1.00	0.09
J_{LA}	0.84	32	1.78	1.14
J_{TBCL}	0.89	25	9.33	3.81
J_{TBCD}	0.94	18	1.03	0.76
J_{TTCL}	0.96	15	4.86	2.12
J_{TTCD}	0.88	23	2.03	1.27

Table 4.4: The mean correlation coefficient (r) between the rows of the analytically-derived Jacobian (J_A) and the Jacobian estimated with the ANN (J_N) and LWR (J_R) from the speech-relevant data set. Also shown are the mean angle in degrees, the norm of each vector. The abbreviations that instantiate the subscript i of J correspond to the task variables described in Section 4.3.2

variables. Articulator variables were extracted using geometrical relationships between these outlines. For instance, jaw angle was defined as the angle between the jaw outline and pharyngeal wall outlines. Constriction degree task variables were extracted by calculating the Euclidean distance between the closest points on the relevant outlines. For example, lip aperture was defined as the minimum distance between the upper and lower lip outlines.

Direct kinematic accuracies were quantitatively evaluated as before, and the results are presented in Table 4.5. Results show that constriction degrees were accurate to 2.7 mm on average. This level of accuracy is close to the 2.9 mm pixel width (i.e., spatial

Direct Kinematic Accuracy – Real Data

Task Variable	$\textbf{RMSE}_{ANN}$	$\textbf{RMSE}_{LWR}$
LA	2.32 mm (13.7 %)	1.66 mm (9.8 %)
$TTCD$	3.04 mm (9.6 %)	2.96 mm (9.4 %)
$TBCD$	2.61 mm (14.8 %)	2.33 mm (13.2 %)

Table 4.5: Accuracies of the direct kinematic estimates for both algorithms on a vocal tract data set acquired using rtMRI from an American English speaker reading 65 TIMIT sentences. Displayed are the root mean squared error (and RMSE as a percentage of the task range) across all vectors in the test sets.

resolution) of the rtMRI data set. Moreover, results are consistent with the results on synthetic data, in that LWR outperforms ANN on this real speech data, as well.

4.5 Discussion

The overall best estimation was observed by LWR operating on the speech-relevant data set. Errors in estimating the direct kinematic relationships in this situation are approximately 1% of the task range, with only slight variability across tasks. The estimates of the differential kinematics in this situation are quite accurate, with all Jacobian rows displaying correlation coefficients of approximately 0.90 with the ground truth, again with slight variation across tasks. We believe that these accuracies are sufficiently good as to be useful in facilitating a variety of kinematic analyses for characterizing speech production.

We observe that LWR outperforms the ANN in terms of accuracy, for the map explored here and for these particular parameters values. We must note that this is not the first time that locally-linear models have been shown to outperform ANNs in practice [Lawrence et al., 1996]. The difference in accuracy most likely reflects practical difficulties associated with selecting the appropriate parameters for the ANN, and the

uncertainty associated with the outcome of iterative, error backpropagation for optimization. These practical challenges are well-known and may prevent ANNs from realizing their full theoretical potential.

There were a few situations (e.g., the estimate of J_{TBCL} for the speech-relevant data) where the ANN did perform better than LWR. Indeed, the ANN displayed a comparable mean normalized RMSE in estimating the direct kinematics on the speech-relevant development data (i.e.,during the parameter optimization process, see Figure 4.4). This accuracy did not generalize to the test set, however.

The ANN appears to struggle with modeling lip protrusion (PRO), which represents the simplest articulator-task relationship in the model. PRO is only dependent on the articulator variable LX, and the relationship is linear. This may be due to the built-in, static nonlinearities in the ANN, which are biased away from learning perfectly linear relationships. By that same token, the ANN does well on the highly nonlinear tasks related to the tongue body and tongue tip. LWR, on the other hand, performs well for all tasks, in spite of its assumptions of linearity.

In most cases, estimating the kinematic relationships from the speech-relevant data is more accurate and more consistent across tasks. Differences are particularly dramatic for the direct kinematic estimates. The overall higher accuracies in estimating from the speech-relevant data may be due to data sparsity issues. The speech-relevant data do not fill the entire articulator space, but rather are confined to certain regions of that space. Consequently, a smaller amount of data is needed to ensure dense coverage of the relevant space and accurate modeling. To increase the data density of uniformly-distributed data, an enormous number of data points are needed since the total number increases exponentially with the points along any single dimension. These trends are encouraging for future work on characterizing the kinematics of speech production, as opposed

to vocal tract kinematics more generally. Attempts to estimate kinematic relationships from real speech-relevant data will require less data overall.

Finally, we observe that LWR outperforms ANN regardless of the quantity of training data. Moreover, LWR is very insensitive to the amount of training data, unlike the ANN which suffers from very high errors when the amount of training data is small. We also note that the performance of both algorithms seems to plateau around 3000 data points, indicating that the 5000-point data sets used for full evaluation were of sufficient size to provide stable results. This result also implies that very modest amounts of data are sufficient for learning the kinematic relationships with these methods, especially LWR.

Results from our preliminary experiments on real speech data were promising. Although observed estimation errors were consistently higher than errors on synthetic data, they were near the spatial resolution of our rtMRI data. It is difficult to interpret the specific implications of these errors because it depends on how these estimates get used downstream (e.g., features for an automatic speech recognizer [Ghosh and Narayanan, 2011]). Still, we can identify several sources of error that cause the discrepancy in errors between synthetic and real speech data. First, the simplified geometry and relatively fewer degrees of freedom in CASY/TADA versus a real vocal tract artificially reduce the error on synthesized data, which is generally to be expected when using synthetic data. Errors may also reflect limitations in the acquisition and processing of rtMRI data. Those data are noisy, and the spatial and temporal resolution give only a limited view of the phenomena in question. Moreover, extraction of articulatory and task parameters (such as TTCD) from these data introduces noise through measurement uncertainty and approximation errors. Finally, our estimation techniques can likely be refined in the way they handle the variability and uncertainty in measurements and representations being

used. Ongoing and future work will be aimed at improving upon these limitations, in particular the estimation techniques and methods for fitting them appropriately.

4.6 Conclusion

We have described and evaluated two statistical models which can estimate the direct and the differential kinematics of a complex, nonlinear motor system from data. Both Artificial Neural Networks and Locally-Weighted Regression were trained, optimized and tested on several data sets of synthesized speech production data. Accuracies were appear high enough to facilitate further use of these methods for estimating the kinematic relationships of real speech production data.

Although synthesized data was used to facilitate evaluation of differential kinematic estimation accuracy, we have taken care to develop a methodology which is entirely repeatable on real speech production data in future extensions of this work. The estimation and evaluation of differential kinematics is a crucial aspect of this work, and the potential applications for these kinematic relationships for characterization of the speech production system was reviewed.

It was observed that the accuracies of both statistical methods were high, even with a relatively modest amount of data. The best accuracies resulted from LWR. Moreover, LWR appears relatively insensitive to the amount of training data available. LWR also has many desirable qualities from a practical standpoint, such as few free parameters and a training procedure with an analytical solution. The assumptions of local linearity might be viewed as limiting, but are seen here as entirely appropriate in context of similar assumptions that are widely made on motor control formulations and applications. Indeed, these assumptions did not seem to be any hindrance in the experiments presented here. Still, it must be noted that the performance of ANNs was very close to

LWR in many aspects of the evaluation. ANNs are very powerful in theory, but may suffer from the practical difficulties associated with iterative training procedures like error backpropagation.

Technical challenges still remain in terms of improving estimation accuracy even further. Alternative estimation techniques could be employed to that end. One possibility would be the Bayesian formulation of LWR developed by [Ting et al., 2008]. Based on a probabilistic formulation of regression, this technique allows for automatic optimization of the locality parameters. More advanced Neural Network architectures, such as Deep Belief Networks [Hinton et al., 2006], are also promising candidates.

Additional challenges remain in assessing whether the differential kinematics need to be estimated to the level of detail espoused here, or whether further approximations are appropriate for speech. For instance, it is not entirely clear to what degree the Jacobian is actually pose-dependent, even though the mathematics presented here express it as such. Given the nature of the expected nonlinearities, we suspect that the Jacobian will indeed be quite dependent on the pose. This claim should be assessed empirically, however, by inspecting the pose-by-pose changes to Jacobians estimated from real speech data.

The ultimate goal of this work is to utilize knowledge of kinematic relationships in order to gain insight into interesting, longstanding and unaddressed problems in the study of speech production. As such, a key extensions of this work will be to apply these estimation methods to real articulatory and acoustic data. The ability to estimate the differential kinematics, in particular, of real speech production data will provide insight and facilitate characterization of the speech production system. Notable applications will include the ability to ascertain the nature of speech production goals and to gain insight into acoustic-to-articulatory inversion.

It is important to note that the utility of these methods depends heavily on obtaining high-quality speech production data. The recently collected MRI-TIMIT database [Narayanan et al., 2011] may provide a useful platform for many applications of these methods. However, the addition of muscle activation data for speech would enhance the possibilities even further, as would the acquisition of 3-dimensional data at high frame rates and the ability to gather clean audio from MRI.

Appendix A

CASY Equations

The articulatory variables are denoted q_i and the task variables are denoted by x_j. Constants include $l_{ut} = 1.1438$, $a_{ut} = -0.1888$, $l_{lt} = 1.1286$, $o_x = 0.7339$, $o_y = -0.4562$, $r_{ts} = 0.4$, $r_{tb} = 0.02$, $a_{tc} = 1.7279$, $l_{tb} = 0.8482$ and $s_{tb} = 4.48$.

The equations for calculating the lip-related tasks from the articulatory variables are:

$$x_{PRO} = q_{lx} \tag{A.1}$$

$$x_{LA} = l_{ut}sin(a_{ut}) + l_{lt}cos(q_{ja}) + q_{uy} - q_{ly} \tag{A.2}$$

The equations for calculating the tongue body tasks from the articulatory variables are:

$$a = q_{cl}sin(q_{ja} + q_{ca}) \tag{A.3}$$

$$b = -q_{cl}cos(q_{ja} + q_{ca}) \tag{A.4}$$

$$x_{TBCL} = cos^{-1}\frac{a - o_x}{\sqrt{(a - o_x)^2 + (b - o_y)^2}} \tag{A.5}$$

$$x_{TBCD} = r_{ts} - \sqrt{(a - o_x)^2 + (b - o_y)^2} + r_{tb} \tag{A.6}$$

The equations for calculating the tongue-tip-related tasks from the articulatory variables are:

$$c = q_{ja} + q_{ta} + s_{tb}(q_{cl} - l_{tb}) \tag{A.7}$$

$$d = a + r_{tb}sin(q_{ja} + a_{tc}) + q_{tl}sin(c) \tag{A.8}$$

$$e = b - r_{tb}cos(q_{ja} + a_{tc}) - q_{tl}cos(c) \tag{A.9}$$

$$x_{TTCL} = cos^{-1}\frac{d - o_x}{\sqrt{(d - o_x)^2 + (e - o_y)^2}} \tag{A.10}$$

$$x_{TTCD} = r_{tb} - \sqrt{(d - o_x)^2 + (e - o_y)^2} \tag{A.11}$$

Appendix B

Stimuli

We synthesized the following sentences using CASY/TADA to create the speech-relevant data set, including the articulator vectors and their accompanying task vectors. These 30 sentences represent the first 30 sentences of the MOCHA-TIMIT corpus, as developed by Wrench and Hardcastle [2000].

1. This was easy for us.
2. Is this seesaw safe?
3. Those thieves stole thirty jewels.
4. Jane may earn more money by working hard.
5. She is thinner than I am.
6. Bright sunshine shimmers on the ocean.
7. Nothing is as offensive as innocence.
8. Why yell or worry over silly items?
9. Where were you while we were away?
10. Are your grades higher or lower than Nancy's?
11. He will allow a rare lie.
12. Will Robin wear a yellow lily?
13. Swing your arm as high as you can.
14. Before Thursday's exam, review every formula.
15. The museum hires musicians every evening.
16. A roll of wire lay near the wall.

17. Carl lives in a lively home.
18. Alimony harms a divorced man's wealth.
19. Aluminium cutlery can often be flimsy.
20. She wore warm, fleecy, woolen overalls.
21. Alfalfa is healthy for you.
22. When all else fails, use force.
23. Those musicians harmonize marvellously.
24. Although always alone, we survive.
25. Only lawyers love millionaires.
26. Most young rabbits rise early every morning.
27. Did dad do academic bidding?
28. Beg that guard for one gallon of petrol.
29. Help Greg to pick a peck of potatoes.
30. Get a calico cat to keep the rodents away.

Appendix C

TIMIT Sentences

1. This was easy for us.
2. Is this seesaw safe?
3. Those thieves stole thirty jewels.
4. Jane may earn more money by working hard.
5. She is thinner than I am.

Reference List

J.H. Abbs and V.L. Gracco. Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology*, 51:705–723, 1984.

S. Al Moubayed and G. Ananthakrishnan. Acoustic-to-articulatory inversion based on local regression. In *Proceedings of INTERSPEECH*, pages 937–940, 2010.

G. Ananthakrishnan. Imitating adult speech: An infant's motivation. In *Proceedings of the International Seminar on Speech Production*, pages 361–369, Montreal, 2011.

G. Ananthakrishnan, D. Neiberg, and O. Engwall. In search of non-uniqueness in the acoustic-to-articulatory mapping. In *Proceedings of INTERSPEECH*, pages 2799–2802, 2009.

R. Arens, J.M. McDonough, A.M. Corbin, M.E. Hernandez, G. Maislin, R.J. Schwab, and A.I. Pack. Linear dimensions of the upper airway structure during development assessment by Magnetic Resonance Imaging. *American Journal of Respiratory and Critical Care Medicine*, 165(1):117–122, 2002.

B.S. Atal and O. Rioul. Neural networks for estimating articulatory positions from speech. *Journal of the Acoustical Society of America*, 86, 1989.

B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63:1535–1555, 1978.

C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *AI Review*, 11:11–73, 1997.

G. Bailly, R. Laboissière, and J. L. Schwartz. Formant trajectories as audible gestures: an alternative to speech synthesis. *Journal of Phonetics*, 19:9–23, 1991.

A. Balestrino, G. De Maria, and L. Sciavicco. Robust control of robotic manipulators. In *Proceedings of the 9^{th} IFAC World Congress*, volume 5, pages 2435–2440, 1984.

S.R. Baum and D.H. McFarland. The development of speech adaptation to an artificial palate. *Journal of the Acoustical Society of America*, 102(4):2353–2359, 1997.

D.J. Bennet and J.M. Hollerbach. Autonomous calibration of single-loop closed kinematic chains formed by manipulators with passive end-point constraints. *IEEE Transactions on Robotic Automation*, 7:597–606, 1991.

N.A. Bernstein. *The coordination and regulation of movements*. Pergamon Press, 1967. originally published in 1947.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

L.-J. Boë and G. Bailly. The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1):27–38, 1992.

L.-J. Boë, J. Granat, P. Badin, D. Autesserre, D. Pochic, N. Zga, N. Henrich, and L. Menard. Skull and vocal tract growth from newborn to adult. In *Proceedings of the International Seminar on Speech Production*, pages 75–82, Ubatuba, 2006.

L.-J. Boë, G. Captier, J. Granat, M. Deshayes, J. Heim, P. Birkholz, P. Badin, N. Kielwasser, and T Sawallis. Skull and vocal tract growth from fetus to 2 years. In *Proceedings of the International Seminar on Speech Production*, pages 157–160, Strasbourg, 2008.

B. de Boer. Investigating the acoustic effect of the descended larynx with articulatory models. *ACLC Working Papers*, 2(2):61–86, 2007.

P. Boersma. Praat, a system for doing phonetics by computer. *Glot Int.*, 5(9):341–345, 2001.

E. Bresch and S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time mri. *IEEE Transactions in Medical Imaging*, 28(3):323–338, 2009.

E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan. Synchronized and noise-robust audio recordings during realtime MRI scans. *Journal of the Acoustical Society of America*, 120(4):1791–1794, 2006.

E. Bresch, Yoon-Chul Kim, K. Nayak, D. Byrd, and S. Narayanan. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. *IEEE Signal Process. Mag.*, 25(3):123–132, May 2008.

J. Brunner, S. Fuchs, and P. Perrier. The influence of the palate shape on articulatory token-to-token variability. In C. Geng, J. Brunner, and D. Pape, editors, *ZAS Papers in Linguistics*, volume 42, pages 43–67. 2005.

J. Brunner, P. Hoole, and P. Perrier. Articulatory optimisation in perturbed vowel articulation. In *Proceedings of the International Congress of Phonetic Sciences*, pages 497–500, Saarbrücken, 2007.

J. Brunner, S. Fuchs, and P. Perrier. On the relationship of palate shape and articulatory behavior. *Journal of the Acoustical Society of America*, 125(6):3936–3949, 2009.

D. Bullock, S. Grossberg, and F.H. Guenther. A self-organizing neural network model for redundant sensory-motor control, motor equivalence, and tool use. *Journal of Cognitive Neuroscience*, 5:408–435, 1993.

J. Canny. Computational approach to edge detection. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

T. Chiba and M. Kajiyama. *The Vowel – Its Nature and Structure*. Kaiseikan, Tokyo, 1941. Chap. 11.

T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle. A novel feature transformation for vocal tract length normalization in automatic speech recognition. *IEEE Trans. Acoust. Speech Sig. Proc.*, 6(6):549–557, 1998.

W. S. Cleveland. Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*, 74:829–836, 1979.

W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83: 596–610, 1988.

W. S. Cleveland, S. J. Devlin, and E. Grosse. Regression by local fitting: Methods, properties, and computational algorithms. *Journal of Econometrics*, 37:87–114, 1988.

T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

G. Cybenko. Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

S. Dart. Articulatory and acoustic properties of apical and laminal articulations. In I. Maddieson, editor, *UCLA Working Papers in Phonetics*, volume 79. 1991.

A. D'Souza, S. Vijayakumar, and S. Schaal. Learning inverse kinematics. In *Proceedings of CIRAS*, 2001.

W. Duch and N.J. Jankowski. Survey of neural transfer functions. *Neural Computing Surveys*, 2:163–212, 1999.

E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pages 346–348, 1996.

C. Everekliogˇlu, S. Doganay, H. Er, A. Gunduz, M. Tercan, A. Balat, and T. Cumurcu. Craniofacial anthropometry in a Turkish population. *Cleft Palate-Craniofacial Journal*, 39(2):208–218, 2002.

G. Fant. *Acoustic Theory of Speech Production*. Mouton & Co., The Hague, 1960. Chap. 1.

G. Fant. A note on vocal tract size factors and non-uniform f-pattern scalings. *Speech Transmission Laboratory Quarterly Progress Status Report*, 7(4):22–30, 1966.

G. Fant. Non-uniform vowel normalization. *Speech Transmission Laboratory Quarterly Progress Status Report*, 16(2–3):1–19, 1975.

S. Fels, F. Vogt, K. van den Doel, J. Lloyd, and O. Guenter. Artisynth: Towards realizing an extensible, portable 3d articulatory speech synthesizer. In *Proceedings of the International Workshop on Auditory Visual Speech Processing*, pages 119–124, July 2005.

W.T. Fitch. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America*, 102(2):1213–1222, 1997.

W.T. Fitch and J. Giedd. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 106(3):1511–1522, 1999.

S. Fuchs, P. Perrier, C. Geng, and C. Mooshammer. What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents. In J. Harrington and M. Tabain, editors, *Towards a better understanding of production processes*, pages 149–164. Psychology Press, New York, 2006.

S. Fuchs, R. Winkler, and P. Perrier. Do speakers' vocal tract geometries shape their articulatory vowel space? In *Proceedings of the International Seminar on Speech Production*, pages 333–336, Strasbourg, 2008.

J.M. Gérard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan. A 3d dynamical biomechanical tongue model to study speech motor control. *Recent Research Development in Biomechanics*, 1:49–64, 2003.

J.M. Gérard, P. Perrier, and Y. Payan. 3d biomechanical tongue modeling to study speech production. In J. Harrington and M. Tabain, editors, *Speech Production: Models, Phonetic Processes, and Techniques*, pages 85–102. Psychology Press, 2006.

P. Ghosh and S. Narayanan. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *J. Acoust. Soc. Am. Express Letters*, 130(4):EL251–EL257, 2011.

P.K. Ghosh and S. Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 128(4):2162–2172, 2010.

E.B. Gouvea and R.M. Stern. Speaker normalization through formant-based warping of the frequency scale. In *EUROSPEECH*, pages 1139–1142, 1997.

A. M Gross, G. D. Kellum, D. Franz, K. Michas, M. Walker, M. Foster, and F. W. Bishop. A longitudinal evaluation of open mouth posture and maxillary arch width in children. *Orthodontist*, 64(6):419–424, 1994.

Y. Gu, Jr. McNamara, J. A., L. M. Sigler, and T. Baccetti. Comparison of craniofacial characteristics of typical Chinese and Caucasian young adults. *European Journal of Orthodontics*, 33(2):205–211, 2011.

F. Guenther. A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, 72:43–53, 1994.

F. Guenther. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102:594–621, 1995.

F. Guenther, M. Hampson, and D. Johnson. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105:611–633, 1998.

E. Guigon, P. Baraduc, and M. Desmurget. Computational motor control: Redundancy and invariance. *Journal of Neurophysiology*, 97(1):331–347, 2007.

M. T. Hagan and M. Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994.

D. Harari, M. Redlich, S. Miri, T. Hamud, and M. Gross. The effect of mouth breathing versus nasal breathing on dentofacial and craniofacial development in orthodontic patients. *Laryngoscope*, 120(10):2089–2093, 2010.

T.R. Harris, W.T. Fitch, L.M. Goldstein, and P.J. Fashing. Black and white colobus monkey (colobus guereza) roars as a source of both honest and exaggerated information about body mass. *Ethology*, 112:911–920, 2006.

S. Hiki and H. Itoh. Influence of palate shape on lingual articulation. *Journal of Speech Communication*, 5(2):141–158, 1986.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

S. Hiroya and M. Honda. Determination of articulatory movements from speech acoustics using an hmm-based speech production model. In *Proceedings of ICASSP*, pages 437–440, 2002a.

S. Hiroya and M. Honda. Acoustic-to-articulatory inverse mapping using an hmm-based speech production model. In *Proceedings of ICSLP*, pages 2305–2308, 2002b.

S. Hiroya and M. Honda. Speech inversion for arbitrary speaker using a stochastic speech production model. In *Proceedings of the Interdisciplinary Workshop on Speech Dynamics by Ear, Eye, Mouth and Machine*, pages 9–14, 2003.

S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an hmm-based speech production model. *IEEE Trans Speech and Audio Processing*, 12(2):175–185, 2004.

J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. Accurate recovery of articulator positions from acoustics: new conclusions based on human data. *Journal of the Acoustical Society of America*, 100:1819–1834, 1996.

J.M. Hollerbach. Computers, brains and the control of movement. *Trends in Neurosciences*, 5:189–192, 1982.

J.M. Hollerbach and C.W. Wampler. The calibration index and taxonomy of robot kinematic calibration methods. *The International Journal of Robotics Research*, 15:573, 1996.

M. Honda, A. Fujino, and T. Kaburagi. Compensatory responses of articulators to unexpected perturbation of the palate shape. *Journal of Phonetics*, 30(3):281–302, 2002.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal function approximators. *Neural Networks*, 2, 1989.

N. Ishii, T. Deguchi, and N. P. Hunt. Craniofacial differences between Japanese and British Caucasian females with a skeletal Class iii malocclusion. *Orthodontics*, 24 (5):494–499, 2002.

K. Iskarous. Vowel constrictions are recoverable from formants. *Journal of Phonetics*, 38:375–387, 2010.

K. Iskarous, L. Goldstein, D.H. Whalen, M. Tiede, and P. Rubin. Casy: The haskins configurable articulatory synthesizer. In *Proceedings of ICPhS*, 2003.

Wu. J.-M. Multilayer potts perceptrons with levenberg-marquardt learning. *IEEE Transactions on Neural Networks*, 19(12):2032–2043, 2008.

M.T.-T. Jackson and R.S. McGowan. A study of high front vowels with articulatory data and acoustic simulations. *Journal of the Acoustical Society of America*, pages 3017–3035, 2012.

R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

M. Jordan. Constrained supervised learning. *Journal of Mathematical Psychology*, 36: 396–425, 1992.

M. Jordan and D. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

M.I. Jordan and R.A. Jacobs. Modular and hierarchical learning systems. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

T. Kaburagi and M. Honda. Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database. In *Proceedings of ICSLP*, 1998.

S.M. Kay. *Modern Spectral Estimation*. Prentice-Hall, Englewood Cliffs, NJ, 1988. Chap. 11.

C. Kello and D.C. Plaut. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *Journal of the Acoustical Society of America*, 116(4):2354–2364, 2004.

J.L. Kelly and C.C. Lochbaum. Speech synthesis. In *Proc. Int. Conf. Acoust.*, pages 1–4, 1962.

S. Kelso, B. Tuller, E. Vatikiotis-Bateson, and C. Fowler. Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology*, 10(6):812–832, 1984.

O. Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal of Robotics and Automation*, 3(1):43–53, 1987.

Yoon-Chul Kim, Shrikanth S. Narayanan, and Krishna S. Nayak. Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order. *Magn. Reson. Med.*, 65(5):1365–1371, 2011.

E.W. King. A roentgenographic study of pharyngeal growth. *Angle Orthodontist*, 22(1): 23–37, 1952.

R.L. Kirlin. *A Posteriori* estimation of vocal tract length. *IEEE Trans. Acoust. Speech Sig. Proc.*, 26(6):571–574, 1978.

A. Lammert. Searching for better logic circuits: Using artificial intelligence techniques to automate digital design. Master's thesis, North Carolina State University, Raleigh, NC, 2006.

A. Lammert, L. Goldstein, and K. Iskarous. Locally-weighted regression for estimating the forward kinematics of a geometric vocal tract model. In *Proceedings of INTERSPEECH*, 2010.

A. Lammert, M. Proctor, A. Katsamanis, and S. Narayanan. Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact. In *Proceedings of INTERSPEECH*, pages 2813–2816, Florence, 2011a.

A. Lammert, M. Proctor, and S. Narayanan. Morphological variation in the adult vocal tract: A study using rtMRI. In *Proceedings of the International Seminar on Speech Production*, Montreal, 2011b.

A. Lammert, V. Ramanarayanan, L. Goldstein, K. Iskarous, E. Saltzman, H. Nam, and S. Narayanan. Statistical estimation of speech kinematics from real-time mri data. *Journal of the Acoustical Society of America*, 130(4):2549, 2011c.

A. Lammert, M. Proctor, and S. Narayanan. Morphological variation in the adult hard palate and posterior pharyngeal wall. *Journal of Speech, Language and Hearing Research*, 56:521–530, 2013.

A.C. Lammert, D.P.W. Ellis, and P. Divenyi. Data-driven articulatory inversion incorporating articulatory priors. In *Proceedings of SAPA*, pages 29–34, 2008.

S. Lawrence, A.C. Tsoi, and A.D. Black. Function approximation with neural networks and local methods: Bias, variance and smoothness. In *Proceedings of Australian Conference on Neural Networks*, pages 16–21, 1996.

L. Lee and R. Rose. Speaker normalization using efficient frequency warping procedures. In *IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pages 353–356, 1996.

L. Lee and R. C. Rose. A frequency warping approach to speaker normalization. *IEEE Trans. Acoust. Speech Sig. Proc.*, 6(1):49–60, 1998.

S. Lee, A. Potamianos, and S. Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105(3):1455–1468, 1999.

J.H. Jr. Long, A.C. Lammert, C.A. Pell, M. Kemp, J. Strother, H.C. Crenshaw, and M.J. McHenry. A navigational primitive: biorobotic implementation of cycloptic helical klinotaxis in planar motion. *IEEE Journal of Oceanic Engineering*, 29:795–806, 2004.

J.H. Jr. Long, T.J. Koob, K. Irving, K. Combie, V. Engel, N. Livingston, A.C. Lammert, and J. Schumacher. Biomimetic evolutionary analysis: Testing the adaptive value of vertebrate tail stiffness in autonomous swimming robots. *Journal of Experimental Biology*, 209:4732–4746, 2006.

S. Maeda. Un modèle articulatoire de la langue avec des composantes lineaires. In *10ème Journées d'Etude sur la Parole*, pages 1–9, 1979.

H.S. Magen, A.M. Kang, M.K. Tiede, and D.H. Whalen. Posterior pharyngeal wall position in the production of speech. *Journal of Speech, Language and Hearing Research*, 46(1):241–251, 2003.

W. S. McCulloch. The brain as a computing machine. *Electrical Engineering*, 68: 492–497, 1949.

M. McCutcheon, A. Hasegawa, and S. Fletcher. Effects of palatal morphology on /s, z/ articulation. *Journal of the Acoustical Society of America*, 67(1):94–94, 1980.

R.S. McGowan and M.A. Berger. Acoustic-articulatory mapping in vowels by locally-weighted regression. *Speech Communication*, 126(4):2011–2032, 2009.

L. Ménard, J.-L. Schwartz, L.-J. Boë, and J. Aubin. Articulatory-acoustic relationships during vocal tract growth for French vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, 35(1):1–19, 2007.

P. Mermelstein. Determination of the vocal-tract shape from measured formant frequencies. *Journal of the Acoustical Society of America*, 41(5):1283–1294, 1967.

P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

P. Mermelstein and M. Schroeder. Determination of smoothed cross-sectional area functions of the vocal tract from formant frequencies. *Journal of the Acoustical Society of America*, 37:1186, 1965.

V. Mitra, Y. Ozbek, H. Nam, X. Zhou, and C.Y. Espy-Wilson. From acoustics to vocal tract time functions. In *Proceedings of ICASSP*, 2009.

V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein. Retrieving tract variables from acoustics: a comparison of different machine learning strategies. *IEEE Journal of Selected Topics on Signal Processing*, 4:1027–1045, 2010. Sp. Iss. on Statistical Learning Methods for Speech and Language Processing.

V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein. Tract variables for noise robust speech recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 2011. To Appear.

A.J. Mohammed, I. Marshall, and N. Douglas. Effect of posture on upper airway dimensions in normal human. *American Journal of Respiratory and Critical Care Medicine*, 149(1):145–148, 1994.

B.W. Mooring, Z.S. Roth, and M.R. Driels. *Fundamentals of Manipulator Calibration*. Wiley Interscience, 1991.

C. Mooshammer, P. Perrier, C. Geng, and D. Pape. An EMMA and EPG study on token-to-token variability. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, 36:47–63, 2004.

H. Moravec. *Mind Children*. Harvard University Press, 1988.

N. J. Morgan, F. B. MacGregor, M. A. Birchall, V. J. Lund, and Y. Sittampalam. Racial differences in nasal fossa dimensions determined by acoustic rhinometry. *Rhinology*, 33(4):224–228, 1995.

D. Mottet, Y. Guiard, T. Ferrand, and R. Bootsma. Two-handed performance of a rhythmical fitts task by individuals and dyads. *Experimental Psychology: Human Perception and Performance*, 27:1275–1286, 2001.

M. Mrayati, R. Carre, and B. Guerin. Distinctive regions and modes: A new theory of speech production. *Journal of Speech Communication*, 7:257–286, 1988.

K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda. On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum. In *Proceedings of ICASSP*, 2006.

Y. Nakamura and H. Hanafusa. Inverse kinematics solutions with singularity robustness for robot manipulator control. *Journal of Dynamic Systems, Measurement, and Control*, 108:163–171, 1986.

J. Nakanishi, M. Mistry, J. Peters, and S. Schaal. Operational space control: A theoretical and empirical comparison. *International Journal of Robotics Research*, 27(6): 737–757, 2008.

H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. Tada: An enhanced, portable task dynamics model in matlab. *Journal of the Acoustical Society of America*, 115(5): 2430–2430, 2004.

H. Nam, L. Goldstein, C. Browman, P. Rubin, M. Proctor, and E. Saltzman. *TADA (TAsk Dynamics Application) manual*, 2006.

H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C.Y. Espy-Wilson, and M. Hasegawa-Johnson. A procedure for estimating gestural scores from natural speech. In *Proceedings of ICASSP*, 2010.

S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, 115(4):1771–1776, 2004.

S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu. A multimodal real-time mri articulatory corpus for speech research. In *Proceedings of INTERSPEECH*, pages 837–840, 2011.

B.F. Necioğlu, M.A. Clements, and T.P. Barnwell III. Unsupervised estimation of the human vocal tract length over sentence level utterances. In *IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pages 1319–1322, 2000.

A. Newell and H. Simon. *Human Problem Solving*. Prentice-Hall, 1972.

S. Nissen and R.A. Fox. Acoustic and spectral patterns in young children's stop consonant productions. *Journal of the Acoustical Society of America*, 126(3):1369–1378, 2009.

P.-E. Nordström. Attempts to simulate female and infant vocal tracts from male area functions. *Speech Transmission Laboratory Quarterly Progress Status Report*, 16 (2–3):20–33, 1975.

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

A. Paige and V.W. Zue. Calculation of vocal tract length. *IEEE Trans. Audio Electroacoust.*, 18(3), 1970.

S. Panchapagesan and A. Alwan. A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model. *Journal of the Acoustical Society of America*, 129(4):2144–2162, 2011.

G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network train on x-ray microbeam data. *Journal of the Acoustical Society of America*, 92:688–700, 1992.

Y. Payan and P. Perrier. Synthesis of v–v sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Communication*, 22: 187–205, 1997.

L. Penning. Radioanatomy of upper airways in flexion and retroflexion of the neck. *Neuroradiology*, 30(1):17–21, 1988.

J.S. Perkell. Articulatory processes. In W.J. Hardcastle and J. Laver, editors, *The Handbook of Phonetic Sciences*, pages 333–370. Blackwell, Cambridge, MA, 1997.

P. Perrier, H. Lœvenbruck, and Y. Payan. Control of tongue movements in speech: The equilibrium point hypothesis perspective. *Journal of Phonetics*, 24:53–75, 1996.

P. Perrier, Y. Payan, M. Zandipour, and J. Perkell. Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *Journal of the Acoustical Society of America*, 114(3):1582–1599, 2003.

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

R. Pfeifer. *Understanding Intelligence*. The MIT Press, Cambridge, MA, 2001.

M. Pitz, S. Molau, R. Schlueter, and H. Ney. Vocal tract normalization equals linear transformation in cepstral space. In *EUROSPEECH*, pages 721–724, 2001.

F.P. Preparata and M.I. Shamos. *Computational Geometry*, pages 185–223. Springer-Verlag, New York, 1990.

M. Proctor, C.H. Shadle, and K. Iskarous. Pharyngeal articulation in the production of voiced and voiceless fricatives. *Journal of the Acoustical Society of America*, 127(3): 1507–1518, 2010.

C. Qin and M.Á. Cerreira-Perpiñán. An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. In *Proceedings of INTERSPEECH*, 2007.

C. Qin and M.Á. Cerreira-Perpiñán. Articulatory inversion of american english /r/ by conditional density modes. In *Proceedings of INTERSPEECH*, 2010.

L. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978. Chap. 3.

M.G. Rahim, W.B. Kleijn, J. Schroeter, and C.C. Goodyear. Acoustic-to-articulatory parameter mapping using an assembly of neural networks. In *Proceedings of ICASSP*, pages 485–488, 1991.

D. Reby and K. McComb. Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Anim. Behav.*, 65:519–530, 2003.

D. Rendall, S. Kollias, C. Ney, and P. Lloyd. Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice acoustic allometry. *Journal of the Acoustical Society of America*, 117:944–955, 2005.

K. Richmond. A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Proceedings of INTERSPEECH*, pages 577–580, 2010.

T. Riede and W. T. Fitch. Vocal tract length and acoustics of vocalization in the domestic dog (canis familiaris). *J. Exp. Biol.*, 202:2859–2867, 1999.

P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede, and C. Browman. Casy and extensions to the task-dynamic model. In *Proceedings of the 1st ETRW on Speech Production Modeling, Autrans, France*, 1996.

D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press, 1986a.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088), 1986b.

E. Saltzman and D. Byrd. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19(4):499–526, 2000.

E. Saltzman and J.A.S. Kelso. Skilled actions: A task dynamic approach. *Psychological Review*, 94:84–106, 1987.

E. Saltzman and K.G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.

E. Saltzman, M. Kubo, and C.C. Tsao. Controlled variables, the uncontrolled manifold, and the task-dynamic model of speech production. In Divenyi et al., editor, *Dynamics of Speech Production and Perception*. IOS Press, 2006.

J.M. Santos, Wright G.A., and J.M. Pauly. Flexible real-time magnetic resonance imaging framework. In *Conference Proceedings of IEEE Engineering in Medicine and Biology Society*, pages 1048–1051, San Francisco, 2004.

J.P. Scholz and G. Schöner. The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental Brain Research*, 126:189–306, 1999.

M.R. Schroeder. Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of the Acoustical Society of America*, 41(4):1002–1010, 1967.

J. Schroeter and M.M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech Audio Processing*, 2:133–150, 1994.

L. Sciavicco and B. Siciliano. *Modelling and Control of Robot Manipulators*. Springer, 2005.

Y. Shiga and S. King. Estimating detailed spectral envelopes using articulatory clustering. In *Proceedings of INTERSPEECH*, 2004.

M.E. Sklar. Geometric calibration of industrial manipulators by circle point analysis. In *Proceedings of the 2nd Conference on Recent Advances in Robotics*, pages 178–202, 1989.

J.F. Soechting. Does position sense at the elbow joint reflect a sense of elbow joint angle or one of limb orientation? *Brain Research*, 248:392–395, 1982.

K. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA., 1998. Chap. 3.

K.N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17:3–45, 1989.

B. Story. Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions. *Journal of the Acoustical Society of America*, 119(2):715–718, 2005.

M. Thibeault, L. Ménard, S.R. Baum, G. Richard, and D.H. McFarland. Articulatory and acoustic adaptation to palatal perturbation. *Journal of the Acoustical Society of America*, 129(4):2112–2120, 2011.

M.K. Tiede, V.L. Gracco, D.M. Shiller, and S.E Espy-Wilson, C.E. Boyce. Perturbed palatal shape and North American English /r/ production. *Journal of the Acoustical Society of America*, 117(4):2568–2569, 2005.

J.A. Ting, A. D'Souza, S. Vijayakumar, and S. Schaal. A bayesian approach to empirical local linearization for robotics. In *Proceedings of ICRA, Pasadena, CA.*, 2008.

M. Toda. Deux stratégies articulatoires pour la réalisation du contraste acoustique des sibilantes /s/ et /S/ en français. In *XXVI ès Journées d'Étude de la Parole*, pages 65–68, 2006.

T. Toda, A.W. Black, and K. Tokuda. Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pages 31–36, 2004.

T. Toda, A.W. Black, and K. Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3), 2008.

A. Toledo, M. Pinzolas, J.J. Ibarrola, and G Lera. Improvements of the neighborhood based levenberg-marquardt algorithm by local adaptation of the learning coefficient. *IEEE Transactions on Neural Networks*, 16(4):988–992, 2005.

R.E. Turner, T.C. Walters, J.J.M. Monaghan, and R.D. Pattern. A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *Journal of the Acoustical Society of America*, 125(4):2374–2386, 2009.

A. Vilain, C. Abry, S. Brosda, and P. Badin. From idiosyncratic pure frames to variegated babbling: Evidence from articulatory modeling. In *Proceedings of the International Congress of Phonetic Sciences*, pages 2497–2500, San Francisco, 1999.

F. Vogt, O. Guenther, A. Hannam, K. van den Doel, J. Lloyd, L. Vilhan, R. Chander, J. Lam, C. Wilson, K. Tait, D. Derrick, I. Wilson, C. Jaeger, B. Gick, E. Vatikiotis-Bateson, and S. Fels. Artisynth designing a modular 3d articulatory speech synthesizer. *Journal of the Acoustical Society of America*, 117(4):2542, May 2005.

F. Vogt, J.E. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S.S. Fels. An efficient biomechanical tongue model for speech research. In *Proceedings of ISSP 06*, pages 51–58, 2006.

H.K. Vorperian and R.D. Kent. Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language and Hearing Research*, 50:1510–1545, 2007.

H.K. Vorperian, R.D. Kent, L.R. Gentry, and B.S. Yandell. Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: preliminary results. *International Journal of Pediatric Otorhinolaryngology*, 49:197–206, 1999.

H.K. Vorperian, R.D. Kent, M. J. Lindstrom, C. M. Kalina, L. R. Gentry, and B. S. Yandell. Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, 117(1):338–350, 2005.

H.K. Vorperian, S. Wang, M. Chung, E. Schimek, R. Durtschi, R. Kent, A. Ziegert, and L. Gentry. Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of the Acoustical Society of America*, 125(3): 1666–1678, 2009.

H.K. Vorperian, S. Wang, E.M. Schimek, B.D. Reid, R.D. Kent, L.R. Gentry, and M.K. Chung. Developmental sexual dimorphism of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of Speech, Language and Hearing Research*, 54:995–1010, 2011.

H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, 21(5):417–427, 1973.

H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. Acoust. Speech Sig. Proc.*, 25(2):183–192, 1977.

P. Wamalwa, S. K. Amisi, Y. Wang, and S. Chen. Angular photogrammetric comparison of the soft-tissue facial profile of Kenyans and Chinese. *Journal of Craniofacial Surgery*, 22(3):1064–1072, 2011.

C.W. Wampler. Manipulator inverse kinematic solutions based on vector formulations and damped least squares methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 16:93–101, 1986.

S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pages 339–341, 1996.

M. Weirich and S. Fuchs. Vocal tract morphology can influence speaker specific realisations of phonemic contrasts. In *Proceedings of the International Seminar on Speech Production*, pages 251–259, Montreal, 2011.

D.E. Whitney. Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man-Machine Systems*, 10:47–53, 1969.

N. Wiener. *Cybernetics: Or control and communication in the animal and the machine*. The MIT Press, Cambridge, MA, 1948.

B.M. Wilamowski, N.J. Cotton, O. Kaynak, and G. Dündar. Computing gradient vector and jacobian matrix in arbitrarily connected neural networks. *IEEE Trans on Industrial Electronics*, 55(10):3784–3790, 2008.

R. Winkler, S. Fuchs, and P. Perrier. The relation between differences in vocal tract geometry and articulatory control strategies in the production of French vowels: Evidence from MRI and modelling. In *Proceedings of the International Seminar on Speech Production*, Ubatuba, 2006.

R. Winkler, S. Fuchs, P. Perrier, and M. Tiede. Speaker-specific biomechanical models: From acoustic variability via articulatory variability to the variability of motor commands in selected tongue muscles. In *Proceedings of the International Seminar on Speech Production*, pages 219–226, Montreal, 2011a.

R. Winkler, S. Fuchs, P. Perrier, and M. Tiede. Biomechanical tongue models: An approach to studying inter-speaker variability. In *Proceedings of INTERSPEECH*, 2011b.

R. Winkler, L. Ma, and P. Perrier. A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing. In *Proceedings of ISSP*, 2011c.

W.A. Wolovich and H. Elliot. A computational technique for inverse kinematics. In *Proceedings of the 23rd IEEE Conference on Decision and Control*, pages 1359–1363, 1984.

S. Wood. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics*, 7:25–43, 1979.

A. Wrench and W. Hardcastle. A multichannel articulatory speech database and its application for automatic speech recognition. In *Proceedings of the 5th Seminar on Speech Production*, pages 305–308, 2000.

J. Y. Wu, U. Hagg, H. Pancherz, R. W. Wong, and C. McGrath. Sagittal and vertical occlusal cephalometric analyses of pancherz: Norms for Chinese children. *American Journal of Orthodontics and Dentofacial Orthopedics*, 137(6):816–824, 2010.

S. A. Xue and J. G. Hao. Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *Journal of Voice*, 20(3):391–400, 2006.

S. A. Xue, G. J. P. Hao, and R. Mayo. Volumetric measurements of vocal tracts for male speakers from different races. *Clinical Linguistics & Phonetics*, 20(9):691–702, 2006.