ABSTRACT

Title of dissertation: AN MRI-BASED ARTICULATORY AND ACOUSTIC STUDY OF AMERICAN ENGLISH LIQUID SOUNDS /R/ AND /L/

Xinhui Zhou, Doctor of Philosophy, 2009

Dissertation directed by: Professor Carol Y. Espy-Wilson
Department of Computer and Electrical Engineering

In American English, the liquid sounds /r/ and /l/ are the most articulatorily variable and complex sounds. They can be produced by several distinct types of tongue configurations and are the most troublesome sounds for children and nonnative English-speakers to learn. Better understanding of this many-to-one mapping between articulation and acoustics would be beneficial to other areas such as speech pathology, speaker verification, speech recognition and speech synthesis.

In this dissertation, two articulatory configurations for each liquid sound were studied (a "retroflex" /r/ vs. a "bunched" /r/, and a light /l/ vs. a dark /l/). Different from previous work on liquids, finite element analysis has been performed to obtain the acoustic responses of the three-dimensional (3-D) vocal tract models, which are based on volumetric magnetic resonance (MR) imaging. Area function models were derived based on the wave propagation property inside the vocal tract.

The retroflex /r/ and the bunched /r/ show similar patterns of F1-F3 but very different spacing between F4 and F5. The results from the formant acoustic

sensitivity functions and simple-tube vocal tract models suggested that this F4/F5 difference can be explained largely by differences in whether the long cavity behind the palatal constriction acts as a half- or a quarter-wavelength resonator. For both the retroflex /r/ and the bunched /r/, F4 and F5 (along with F3 for the particular speakers studied in this research) come from the long back cavity. However, these formants are half wavelength resonances for the retroflex /r/, but quarter wavelength resonances for the bunched /r/.

While both the dark /l/ and the light /l/ have a linguo-alveolar contact and two lateral channels, they differ in the length of the linguo-alveolar contact and in the presence of the linguopalatal contacts caused by raising the sides of the tongue. Both have similar patterns in F1-F3, but differ in the number and locations of zeros in spectrum. For the dark /l/, only one zero occurs below 6 kHz and it is produced by the cross mode posterior to the linguo-alveolar contact. For the light /l/, three zeros below 6 kHz are produced by the asymmetrical channels, the supralingual cavity and the cross mode posterior to the linguo-alveolar contact. The results from two simple vocal tract models show that the lateral channels have to be asymmetrical with an effective length between 3-6 cm to get a zero in the region of F3-F5.

Based on the Buckeye database, the acoustic variability and discriminative power of liquids were studied with the mel-frequency band energy coefficients as acoustic parameter. Analysis of variance shows that the inter-speaker variability of /r/ is larger than any other phonemes except /sh/, /s/ and /zh/. On average, /r/ and /l/ have larger inter-speaker variability than any other broad phonetic class. The F-ratio averages of liquids are larger than glides, fricatives, affricates and

stops, but smaller than nasals. The speaker identification experiments show that the ranking of the average discriminative power for liquids and other broad phonetic classes is: /r/ > Glides > /l/ > Affricates > Fricatives > Stops > Nasals > Vowels.

An MRI-based articulatory and acoustic study of American English liquid sounds /r/ and /l/

by

Xinhui Zhou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Carol Y. Espy-Wilson, Chair/Advisor
Professor Shihab Shamma
Professor William S. Levine
Professor Maureen Stone
Professor Elias Balaras (Dean's Representative)

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I'd like to thank my advisor, Professor Carol Espy-Wilson, for giving me an invaluable opportunity working in her group. It would be impossible for me to continue my PhD study without her financial support. Thanks for her guidance and patience when I was stuck in my research and for her letting me explore my research ideas with freedom.

Thanks to National Health Institute and National Science Foundation for the financial aid without which this thesis would have been impossible. Thanks also to Dr. Chiman Kwan at Signal Processing Inc. at Rockville, Maryland, for his financial support during the last year of my PhD study.

Thanks to my thesis committee members Professor Shihab Shamma, Professor William S. Levine, Professor Maureen Stone, and Professor Elias Balaras for their helpful comments and encouragement.

Thanks to our collaborators Professor Suzanne Boyce at University of Cincinnati, and Mark Tiede at Haskins Laboratories for their support, guidance, and encouragement.

Thanks to all my labmates including Daniel Garcia-Romero, Vikramjit Mitra, Srikanth Vishnubhotla, Tarun Pruthi, Om Deshmukh, Sandeep Manocha, Zhaoyan Zhang and Gongjun Li, for their support and help in my research project. I benefited a lot from those stimulating discussions with them. Special thanks to Daniel for his

insightful comments on my work, for his help in acoustics and speaker recognition experiment, and for all the lunches and discussions we had together; to Vikramjit for his warm-hearted help in many aspects of my study, and for his inspiration of being hard-working and focused; to Tarun for his inspiration of being organized and his great programming style.

Thanks to the ISR technical support for their computer help and their professional attitude, including Peggy Jayant, Carlos Luceno, Jeff McKinney and others.

I would like to thank my parents for their hard-work spirits, their valuing advanced education and their continuing encouragement.

I am profoundly indebted to my dear wife Ling for her love, encouragement, support during my PhD study and for putting up with frequent disturbance caused by many unexpected circumstances, even though she needs to work hard on her own PhD at University of Maryland. Thanks also to her help on my presentation skill and her high expectation on me.

Last but not least, I would like to thank all of my friends during my school years, who shared their ideas and time with me and encouraged me during my hard time. Thanks to all of my friends in the basketball court. Playing with them each week brought a lot of fun to my life, and made me refreshed and energetic for my research.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3-D	Three Dimensional
ANOVA	Analysis of Variance
FEM	Finite Element Method
FMPSPGR	Fast MultiPlanar SPoiled GRadient echo
LPC	Linear Prediction Coefficient
LPCC	Linear Prediction Cepstral Coefficient
RC	Reflection Coefficient
MFCC	Mel-Frequency Cepstrum Coefficient
MFB	Mel-Frequency filter Bank
MRI	Magnetic Resonance Imaging
STL	STereoLithography
VTAR	Vocal Tract Acoustic Response

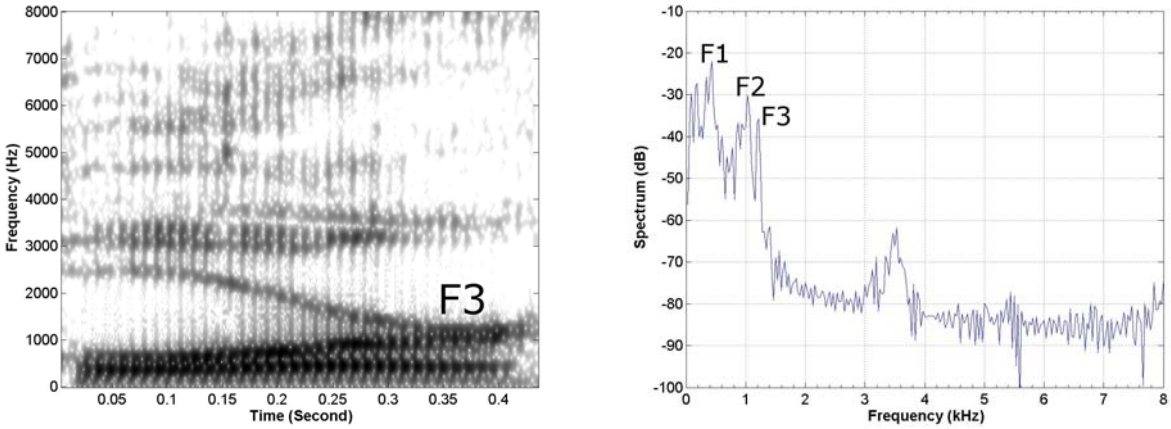
Chapter 1

Introduction

1.1 What are liquid sounds in American English ?

In American English, the consonant phonemes /r/ as in "read" or "poor" and /l/ as in "lead" or "pool" are called liquids. The /r/ sound is also called a rhotic sound, and the /l/ sound is also called the lateral sound. The term "liquid" originates from a Latin (mis)translation of a Greek technical term (Allen, 1965, page 32). The Greek grammarians used `hygros_` ("fluid") for /r/, /l/, /n/ and /m/. This term was translated into Latin as `_liquidus_` ("liquid"). In Latin, however, the term "liquid" has been restricted to /r/ and /l/. Roach (2002, page 47) claims that "liquid" is an old-fashioned phonetic term without any scientific definition.

A chart for American English consonants is shown in Table A.1 in Appendix A. The liquids and glides (/w/ and /j/) are also called semivowels. There are constrictions along the vocal tract for semivowels, however the constrictions are not sufficiently narrow to cause a significant pressure drop due to the glottal air flow, or to cause turbulence in the vicinity of the constriction like fricatives. They are also called sonorant sounds along with nasals and vowels since the radical constrictions along the vocal tracts for these sounds do not inhibit spontaneous voicing. Glides must involve a continuous movement from one sound to another (e.g. /j/ as in "yet" and /w/ as in "wet"). Liquids are different from glides in that they can be



(a) Spectrogram of word "pour"

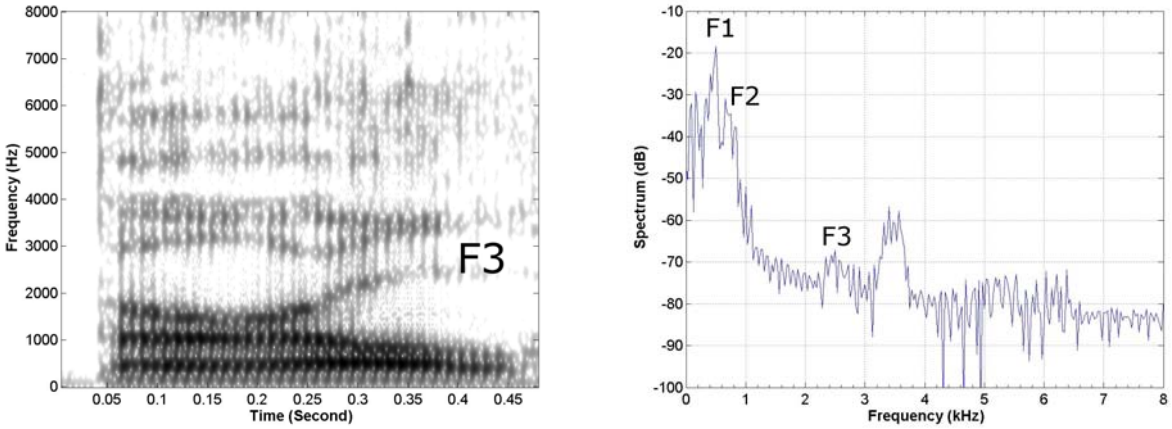
(b) Spectrum of /r/ in word "pour"

Figure 1.1: Examples of spectrogram of word "pour" and /r/ spectrum.

maintained as steady sounds. In some other languages there are liquid consonants for which turbulence noise is produced at the constriction formed by the tongue blade, but only sonorant liquids exist in American English.

Liquids possess spectral characteristics similar to vowels, but normally they have weaker energy than most vowels due to their more constricted vocal tracts (Deller et al., 2000, page 129). Usually /r/ and /l/ have similar pattern in the first two formants (F1 and F2). However /r/ has a low third formant (F3 gets below 2000 Hz) and /l/ has a relatively high third formant (F3 usually at or above 2500 Hz), as shown in spectrograms and spectra in Figure 1.1 and 1.2 respectively.

Flanagan (1972) shows that, in standard prose, the relative occurrence frequency of /r/ is 6.88% and the relative occurrence frequency of /l/ is 3.74%. Comparatively, the relative occurrence frequencies of sonorant sounds /m/, /n/, and /ŋ/ are 2.78%, 7.24%, and 0.96% respectively.



(a) Spectrogram of word "berle"

(b) Spectrum of /l/ in word "berle"

Figure 1.2: Examples of spectrogram of word "berle" and /l/ spectrum.

1.2 Why study liquid sounds ?

We study liquids for several reasons.

- Liquids are considered to be the most difficult sounds to learn (Shriberg and Kent, 1982), both for children and adult English learners. Clinically, cases of "resistant" /r/ and /l/ are regularly seen, translating into significant levels of frustration for patients and therapists. According to some reports, problems with /r/ alone can account for as much as 60% of the typical school-based clinician's caseload (Creaghead et al., 1989).

- Words containing /r/ and /l/ are frequently the source of errors in automatic recognition system (Espy-Wilson, 1992).

- Compared to other sounds, /r/ and /l/ have much more complex and also more variable articulatory configurations across speakers. The articulation

variability across speakers might lead to some individual speaker's information in the acoustic signal (Eatock and Mason, 1994; Goldstein, 1976; Nolan, 1983; Westbury et al., 1998), which is perceptually unimportant but useful in speech technology such as speaker verification.

- Compared to vowels or obstruent consonants, the acoustics and vocal tract models of /r/ and /l/ are less studied, except some idealized articulation (Stevens, 1998).

- In articulatory speech synthesis, one of the articulatory configurations for /r/and /l/ might be preferable to others to produce natural dynamic speech.

1.3 Challenges in studying liquid sounds

Liquid sounds exemplify the non-uniqueness problem, in that speakers show a remarkable variability in articulatory configurations while producing stable acoustic profiles to be perceived as liquids. The acoustics of /r/ and /l/ is one of the most outstanding incompletely-solved problems in phonetics.

The variability is accompanied with the complexity of the vocal tract configuration producing /r/ or /l/. Different from vowels and other semivowels, the tongues for liquids in the oral cavity are shaped in such a way that there might be split or bifurcation in the air flow in the vocal tracts. The split or bifurcation has certain acoustic consequences in the middle or higher frequencies in speech. The geometry in the vocal tract for split or bifurcation may involve large front cavity or lateral channels which may not be observed from traditional midsagittal X-ray or magnetic

resonance (MR) images. A comprehensive three-dimensional vocal tract model integrated with a tongue model is needed to understand their acoustic consequences of different components along the vocal tract for various articulatory configurations of liquids.

1.4 Objectives of this study

In phonetics, there are always three levels to consider and keep separated: the articulatory level, the acoustic level, and the perceptual or auditory levels. The first two levels are objective and quantitative, and the last one is subjective only. Only the articulatory and acoustic levels about liquid sounds /r/ and /l/ in American English are within the scope of this dissertation.

There are two main objectives in this dissertation.

- To better understand the acoustics and articulation of the liquid sounds in American English. Particularly, to understand how to model typical articulatory configurations for /r/ and /l/, and to understand the major articulatory and acoustical differences among them.

- To study the acoustic variability and the speaker discriminative power of the liquids, i.e., to study if the variability in articulation across speakers make the liquid sounds have more inter-speaker acoustic variability and, thereby, have more discriminative power in speaker recognition relative to other sounds (vowels, nasals, glides, fricatives, affricates and stops).

1.5 Organization of the dissertation

`Chapter 1` describes what the liquid sounds are, why we study them, the challenges of studying them, and the scope and objectives of this dissertation.

`Chapter 2` presents an extensive literature survey on acoustics, articulation and vocal tract modeling of /r/ and /l/, and on acoustic variability study for phonemes and phoneme-based speaker recognition. This chapter also points out some problems in the vocal tract modeling of liquids in previous studies.

`Chapter 3` describes the databases, tools and methodologies used in this dissertation.

`Chapter 4` presents the study of acoustics, articulation, and vocal tract modeling of retroflex /r/ and bunched /r/. This chapter describes the results of three dimensional (3-D) reconstructions of the vocal tracts, and the results of 3-D finite element analysis and area function based vocal tract models. It analyzes the similarities and differences in articulation and acoustics between retroflex /r/ and bunched /r/, and analyzes their underlying difference in vocal tract modeling.

`Chapter 5` presents the study of acoustics, articulation, and vocal tract modeling of lateral sound /l/, including one light /l/ and one dark /l/. As in chapter 4, this chapter also describes the results of three dimensional (3-D) reconstructions of the vocal tracts, and the results of 3-D finite element analysis and area function based vocal tract models. It analyzes the similarities and differences in articulation and acoustics between the light /l/ and the dark /l/. Details are given on how to obtain the area function based vocal tract models in order to explain the zero

sources(s) in /l/ spectrum.

`Chapter 6` presents the study of acoustic variability and speaker discriminative power of liquids along with other sounds. This study is based on the Buckeye database (Pitt et al., 2005).

`Chapter 7` summarizes the work in this dissertation and presents some future research topics as extensions from this dissertation.

1.6 Conventions used

In this dissertation, the labels in the TIMIT (TIMIT, 1990) database are used to represent different phonemes. For convenience, Appendix B gives the correspondence between the TIMIT label and the International Phonetic Alphabet (IPA). The label for each phoneme is enclosed within the forward slashes (for example, /r/).

Chapter 2

Literature survey

2.1 /r/ acoustics, articulation and modeling

2.1.1 /r/ acoustics

American English /r/ occurs both as a syllable nucleus (as in "burr") and in consonantal position (as in "read" or "dear"). The most salient acoustic property of American English /r/ is a very low third formant frequency (F3) which often comes close to F2 (Dalston, 1975; Espy-Wilson, 1987; Lehiste, 1964). In a study of 15 subjects, Hagiwara (1995) found that, for any one speaker, F3 for /r/ was between 60% and 80% of the average F3 for that speaker's vowels. A major focus in the vocal tract modeling of /r/ is accounting for the low F3.

The characteristic formant pattern in F1 and F2 of American English /r/ is similar to that of a canonical central and rounded vowel (Espy-Wilson, 1992). The range of formant values reported in the literature for the first three formants of /r/ is approximately 250-550 Hz for F1, 900-1500 Hz for F2, and 1300-1950 Hz for F3 (Delattre and Freeman, 1968; Espy-Wilson, 1992; Espy-Wilson et al., 2000; Westbury et al., 1998) (note that Hagiwara (1995) found a higher range for some female subjects). There is a tendency for F3 values to be lower or higher according to /r/'s position in the word (Delattre and Freeman, 1968; Lehiste, 1964).

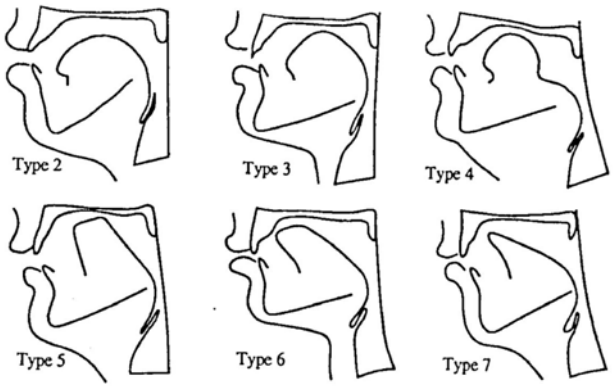


Figure 2.1: Tongue configuration types for American English /r/ as identified by Delattre and Freeman (1968) (types 1 and 8 exist in British English). From adapted figure in (Hagiwara, 1995).

There are few statistical studies of F4 and F5 of /r/, probably due to the insignificance of F4 and F5 in perception and also due to the weaker energy in F4 and F5 range at dynamic speech. Espy-Wilson (1992) reported the average F4 of prevocalic /r/ is 3350 Hz, average F4 of intervocalic /r/ is 3433 Hz, and average F4 of postvocalic /r/ is 3391 Hz.

2.1.2 /r/ articulation

It is well known that different speakers may use very different tongue configurations for producing American English /r/ (Alwan et al., 1997; Delattre and Freeman, 1968; Espy-Wilson et al., 2000; Hagiwara, 1995; Tiede et al., 2004; Westbury et al., 1998). Traditionally, phoneticians have classified the tongue shapes for American English /r/ into two maximally distinct types: "retroflex" (with a raised tongue tip and a lowered tongue dorsum) and "bunched" (with a lowered tongue

tip and a raised tongue dorsum). However, the classification as "retroflex" and "bunched" are only two extremes in a continuum with many incremental variants (Alwan et al., 1997; Delattre and Freeman, 1968; Espy-Wilson et al., 2000; Tiede et al., 2004; Westbury et al., 1998), and it understates the degree of variability found across speakers. Based on X-ray motion pictures, Delattre and Freeman (1968) divided the tongue shapes in American English /r/ into six types, as shown in tracings from X-rays representatives in Figure 2.1. Usually these shapes have three supraglottal constrictions along the vocal tract (except Type 2): a constriction narrowing the pharynx, a constriction along the palatal vault and a constriction at the lips. Overall, articulatory configurations differ most in the palatal region, i.e., how the palatal constriction is formed.

- Type 2 does not have a palatal constriction.
- Type 3 and 4 form the constriction at palatovelar regions by the raised tongue dorsum with a lowered tongue tip.
- Type 5 and 6 have palatal constrictions at both the alveolar and palatovelar regions by the raised tongue tip and blade.
- Type 7 has a palatal constriction at the alveolar ridge formed solely by the raised tongue tip.

Another common characteristic of all these types of /r/ shapes is that all the shapes have a large front cavity (between the lips and the palatal constriction) inside the vocal tract which has the effect of lowering F3 directly or indirectly.

In a study of Westbury et al. (1998), they used X-ray microbeam fleshpoint measures of prevocalic /r/ for five test words spoken by 53 normal young adult speakers of American English. They found that tongue shapes for /r/ vary widely across speakers within any single phonetic context, more continually than categorically across the representational space. Tongue shapes also vary by context in ways that are similar across most speakers. The tongue shapes for American English /r/ do not seem to be reliably linked to gender, measures of oral cavity size, or formant frequency measurements.

In a study of Alwan et al. (1997), magnetic resonance imaging (MRI) of the vocal tract during sustained production of /r/ by four native American English speakers was employed for measuring the vocal tract dimensions and for morphological study of the vocal tract and tongue shapes. All the four speakers in this study showed a large volume in the front cavity anterior to the palatal constriction, which was the result of an inward-drawn tongue body which is characterized by convex cross sections at the anterior part and concave cross sections at the posterior part. No systematic differences were found between the 3-D vocal tract and tongues shapes of word-initial /r/ and syllabic /r/s.

Recently, Tiede et al. (2004) collected a large database with more types of tongue shapes. Part of this database was used for this dissertation, and all the 22 speakers' midsagittal MR images for producing sustained /r/are shown in Figure 3.1 of Chapter 3 (page 32). In this series, it is easily seen that while some speakers use the classic "retroflex" configuration and some use the classic "bunched" configuration, there are a number of subjects whose /r/ configuration appears to be

intermediate between them. It can be seen that the shape of the tongue behind the palatal constriction can also vary considerably across speakers. The degree of variability illustrated in this figure has not been matched by any other sound.

Delattre and Freeman (1968) found that the positional distribution of the six types of tongues shapes is not clear among their 46 speakers, 32 words for each speaker. Speakers who use type 2 after vowels normally use type 7 before vowels. But those who use types 3, 4, or 5 after vowels are equally likely to use types 3, 4, 5 or types 6 and 7 before vowels. Many American speakers use type 3 or 4 in all syllable positions. One speaker used type 7 at all syllable positions. So these different types of tongue shapes occur both within and across speakers. While some speakers may use one type of tongue shape exclusively, other speaker may switch between two or three types of tongue shapes across phonetic contexts (Zawadzki and Kuehn, 1980). However, using an electromagnetic midsagittal articulometer (EMMA) system to track the movements of six small points on the tongues of speakers, Guenther et al. (1999) showed that the tongue configurations still have a lot of similarity across contexts in nonsense words such as "warav", "wabrav","wavrav", "wadrav", "wagrav".

Given the large degree of articulatory differences between "bunched" /r/ and "retroflex" /r/, it might be expected that the two would be acoustically distinct. There have been several attempts to correlate particular tongue configurations and acoustic differences across different types of /r/ using F1, F2, and F3. However, no consistent pattern has emerged (Delattre and Freeman, 1968; Westbury et al., 1998). In recent years, Espy-Wilson and colleagues have suggested that the higher

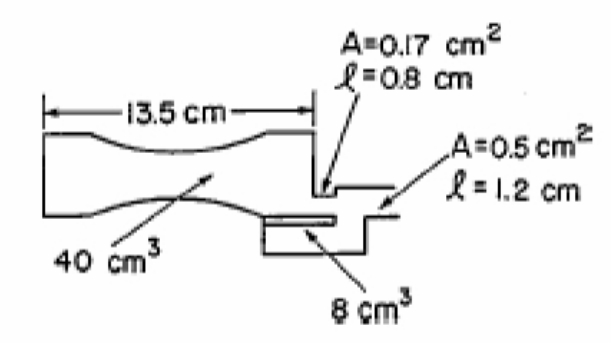


Figure 2.2: Simple-tube model for a tip-up retroflex /r/ (Stevens, 1998). The symbol "A" stands for area and the symbol "l" stands for length. The orientation of this model is such that the glottis is at the left edge and the lips are at the right edge.

formants may contain clues to the tongue configuration and vocal tract dimensions (Espy-Wilson, 2004; Espy-Wilson and Boyce, 1999). Our recent study (Zhou et al., 2008) shows that the difference in F4 and F5 in the retroflex /r/ is much larger than it is for the bunched /r/ (see Chapter 4 for details).

2.1.3 /r/ vocal tract modeling

Studies of vocal tract modeling of /r/ have been focused on how to explain the source of the low third formant F3. There are two types of proposed models for American English /r/. One is the perturbation theory (Johnson, 2003; Ohala, 1985), and the other is the decoupling account (Alwan et al., 1997; Espy-Wilson et al., 2000; Fant, 1970; Narayanan et al., 1999; Stevens, 1998; Zhang et al., 2003).

The perturbation theory account of /r/ is based on a general principle of uniform tube acoustics. The maximum volume velocity in a quarter-wavelength tube happens to be around the three constrictions of the vocal tract, that have

the effect of lowering F3. However, Espy-Wilson et al. (2000) concluded that the perturbation theory can not account adequately for /r/'s low F3, which is due to the ideal initial uniform tube assumption in this theory.

On the contrary, decoupling accounts of /r/ assume that the vocal tract is divided into several different tubes. There is some decoupling or coupling between them, depending on the degree of the constrictions. Fant (1970) stated that the acoustics of /r/ can be treated as in a vowel since there is no side branch which introduces antiresonance. He extracted the area function for a "retroflex" /r/ in Russian and found the F3 is produced by the front cavity which is anterior to the palatal constriction.

However, Stevens (1998) explicitly pointed out that there is a split of air flow inside the vocal tract for American English /r/ and, as a result, the vocal tract configuration for /r/ can not be approximated by a simple tube. A detailed model of the acoustics of a "retroflex" /r/ is found in Stevens (1998). A diagram is also shown in Figure 2.2. In this model, the palatal constriction is narrow enough to decouple the back cavity and the front cavity. The back cavity produces F2 and the front cavity, including the sublingual space, produces F3 which is close in frequency to F2. The sublingual space is part of the large front cavity volume that lowers F3. In addition, the sublingual space is regarded as a side branch to produce a zero in the /r/ spectrum around 2 kHz. No detailed acoustic model is given for a "bunched" /r/ in Stevens (1998).

The advent of the volumetric magnetic resonance imaging (MR) technique made it possible to acquire vocal tract data in three dimensions specific to a par-

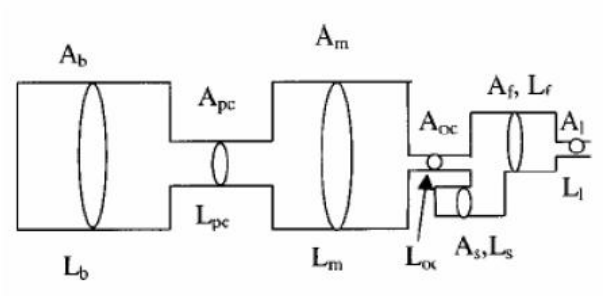


Figure 2.3: Simple-tube model for the bunched /r/s (Espy-Wilson et al., 2000). A_b and L_b correspond to the area and length of the back cavity; A_{pc} and L_{pc} correspond to the area and length of the pharyngeal constriction; A_m and L_m correspond to the area and length of the midcavity between the pharyngeal constriction and the oral constriction; A_{oc} and L_{oc} correspond to the area and length of the oral-palatal constriction; A_f and L_f correspond to the area and length of the front cavity between the oral constriction and the lip constriction; and A_l and L_l correspond to the area and length of the lip constriction.

ticular individual (Alwan et al., 1997; Baer et al., 1991; Espy-Wilson et al., 2000; Narayanan et al., 1997; Ong and Stone, 1998; Story et al., 1996). This advance gives us more exact specifications for variables such as constriction location, cavity length, and constriction area. It has allowed us to model vocal tract acoustics more accurately and to investigate individual variation. Based on the dimensions of two speakers from magnetic resonance imaging of vocal tracts obtained by Alwan et al. (1997), Espy-Wilson et al. (2000) constructed acoustic models for /r/. They found that by using the sublingual space as a side branch or as an increment to the dimension of the front cavity, the F3 range can be matched very well with measurements from the acoustic signal. They also developed simple-tube models (See Figure 2.3) to account for the formant cavity affiliation and confirmed that F3 is the resonance

of the front cavity and F1, F2 and F4 are from the back cavity geometry.

However, Zhang et al. (2003) suggested that, for some speakers the front cavity volume is so large that there is a switch in formant-cavity affiliation, i.e., the front cavity resonance is so low that it becomes F2 and the first resonance of the cavity posterior to the palatal constriction which typically produces F2 becomes F3.

2.2 /l/ acoustics, articulation and modeling

2.2.1 /l/ acoustics

In the /l/ acoustic spectrum, F1 is low, although higher than a high vowel, and F2 is barely separated from F1. F3 in /l/ is higher in frequency than in most vowels (Dalston, 1975). A high F3 is the major acoustic cue which makes /l/ differ from /r/. Espy-Wilson (1992) reported that the average formant frequencies of prevocalic /l/ are: F1 399 Hz, F2 1074 Hz, F3 2533 Hz, F4 3767 Hz. The average formant frequencies of intervocalic /l/ are: F1 445 Hz, F2 1060 Hz, F3 2640 Hz, F4 3762 Hz. Finally the average formant frequencies of postvocalic /l/ are: F1 465 Hz, F2 898 Hz, F3 2630 Hz, F4 3650 Hz. However, it is very difficult to characterize the /l/ sound since it has a large variation in spectrum among different speakers and contexts (Espy-Wilson, 1992; Nolan, 1983). The /l/ sound has both formants and antiformants and, therefore, is similar to nasal sounds.

Traditional phonetics distinguishes between two types of /l/: "light" /l/ and "dark" /l/ (Shriberg and Kent, 1982). Which type occurs in speech varies according to syllable positions (with dark /l/ occurring finally as in "bell" and light /l/

occurring initially as in "luck") and phonetic context (dark /l/ next to back vowels and light /l/ next to front vowels) (Lehman and Swartz, 2000; Sproat and Fujimura, 1993). Acoustically, light /l/ has a relatively lower F1 and higher F2 (Espy-Wilson, 1992; Lehiste, 1964; Lehman and Swartz, 2000). Lehman and Swartz (2000) reported that, for light /l/ but not for dark /l/, the F2 and F3 were often weak or absent and the vowel context had a great acoustic effect. However, acoustic and articulatory properties which are intermediate to those of dark /l/ and light /l/ have been known (Narayanan et al., 1997; Sproat and Fujimura, 1993). Sproat and Fujimura (1993) argued that the dark /l/ and the light /l/ were not two distinct elements. Instead, the /l/ is phonetically implemented as a lighter or darker variant depending on factors such as the /l/'s position within the syllable and the duration of the prosodic context containing /l/.

2.2.2 /l/ articulation

The /l/ sound is typically produced with linguo-alveolar contact along the midsagittal line such that air flows along one or both sides of the tongues. The space behind the linguo-alveolar contact is called the supralingual cavity, and the flow channels along the sides of the tongue are called lateral channels. In most cases, air flow above the tongue is occluded at the lingual-alveolar contact around 1-2.5 cm behind the lips (Panchapagesan, 2003).

The /l/ involves the bifurcation of the air flow around the linguo-alveolar contact, which allows the sound to radiate from the opening at the sides, and it was

reported that this bifurcation causes the zero(s) in the spectrum (Narayanan et al., 1997; Stevens, 1998; Zhang and Espy-Wilson, 2004).

The articulation variability in /l/ production was not described as much as in the case of /r/. The number of lateral channels, linguo-alveolar contact and the tongue shape are the main concerns in articulatory configurations. Articulatory studies have shown differences in the production of the /l/ sound (Giles and Moll, 1975; Narayanan et al., 1997). Giles and Moll (1975) reported that linguo-alveolar contact is not often observed in the dark /l/ in American English. Narayanan et al. (1997) did an articulatory study of /l/ based on MRI from four subjects, two females and two males and their midsagittal sketches are shown in Figure 2.4. They found that primary tongue-shape mechanisms for /l/ are responsible for the linguo-alveolar contact, inward-lateral compression and convex shaping of the middle and back tongue body. The flattening or grooving of the tongue body immediately behind the linguo-alveolar contact is a secondary feature, but it varies. For the light /l/ of AK as shown in Figure 2.4, the tongue tip is lowered, the mid part of the tongue is raised and the tongue back is lowered to form a concave shape. For the light /l/ of MI, the tongue tip is raised, the mid part of the tongue is lowered and the tongue back is raised to form a convex shape. For some speakers such as PK, the tracing of the tongue shape is more or less flat. The dark /l/ in PK does not have a linguo-alveolar contact.

The main articulatory difference between light /l/ and dark /l/ are the greater retraction of the anterior tongue body in dark /l/ (Narayanan et al., 1997) and larger linguo-alveolar contact in light /l/ (Panchapagesan, 2003). Dark /l/ shows smaller

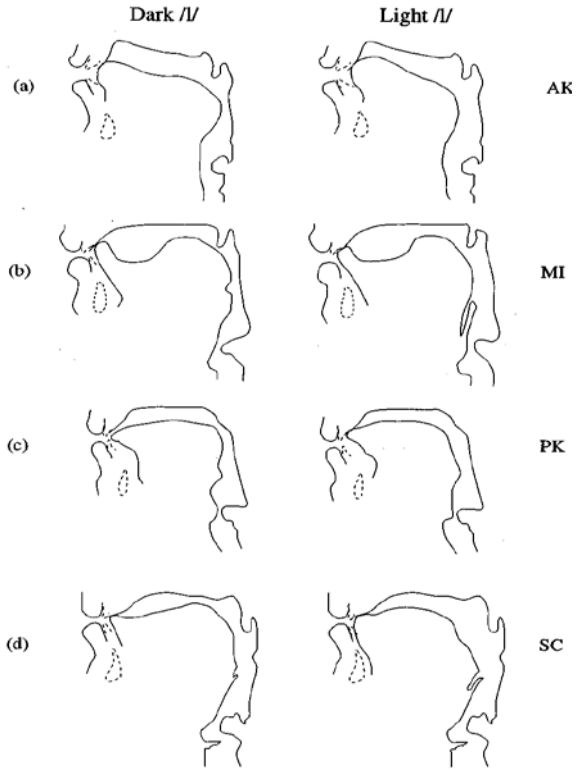


Figure 2.4: Tracings of the midsagittal profiles of the vocal tracts for different subjects during /l/ production (four subjects AK, MI, PK, SC) (Narayanan et al., 1997).

pharyngeal areas than light /l/ due to the more retracted tongue body and its possible raising towards the velum (Narayanan et al., 1997). In some cases, dark /l/ was found to have little or no linguo-alveolar contact (Narayanan et al., 1997).

2.2.3 /l/ vocal tract modeling

Early work in /l/ vocal tract modeling was done by Fant (1970) for Russian. He classified /l/ into two varieties, the "palatalized" /l/ and the "non-palatalized" /l/. The "non-palatalized" /l/ has a constriction in the uvular region. His model has a

supralingual cavity along a combined channel, and the area function was estimated from an X-ray midsagittal image. He explained the formant-cavity affiliations as follows. F1 was produced by a Helmholtz resonator due to the alveolar constriction, and F2 was produced by the half-wavelength resonator formed by the back cavity. In the non-palatalized /l/, F2 can be explained by the perturbation theory. F3 was associated with the oral cavity anterior to the occlusion. F4 in the "non-palatalized" /l/ comes from the cavity between the uvular and the occlusion, and F4 in the "palatalized" /l/ comes from the cavity between the larynx and the occlusion. The zero was caused by the supralingual cavity behind the tongue occlusion. He suggested that the effect of the pole-zero pair is to make F4 take the role of F3. However, he claimed that the pole-zero pair is not as important as the pole-zero pair of the nasal sounds.

Stevens (1998) used a model similar to Fant (1970), as shown in Figure 2.5. Stevens (1998) considered one articulatory configuration where there is contact between only one of the lateral edges of the tongue and the palate, and therefore there is only one channel. In Steven's model, F1 and F2 have the same formant-cavity affiliations as in Fant's model. There are two half-wavelength resonance frequencies from the back cavity, around 2.8 kHz and 3.9 kHz, and there is a resonance around 3.5 kHz from the front cavity anterior to the occlusion. The zero for the side branch is in the range of 2.2-4.4 kHz. Hence in the range of 1.5 to 4 kHz, the /l/ spectrum has a cluster of three formants and one zero. The variability of this cluster pattern increases the complexity of the /l/ spectrum. Stevens (1998) explained the effect of the possible asymmetry between the lateral channels that are formed along the

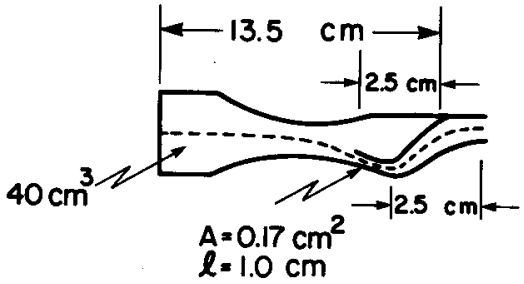


Figure 2.5: Stylized model of /l/ (Stevens, 1998).

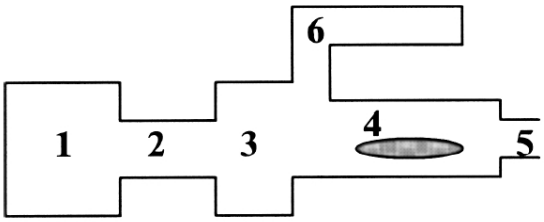


Figure 2.6: Simple-tube model of the vocal tract for /l/ sound production (Zhang and Espy-Wilson, 2004) (1,2,3- back cavity, 4-lateral channel(s), 5-lips, 6-supralingual cavity).

sides of the linguo-alveolar occlusion and this asymmetry will lead to additional zeroes in spectrum, which may cause double peaks or a formant that is split into a pole-zero-pole cluster. Stevens (1998) also pointed out that the zeros not only shift the formants of /l/, but also modify the overall spectrum between 2.5-4 kHz.

Narayanan et al. (1997) collected MRI data of /l/ from four speakers and those data can be used for vocal tract modeling of /l/. Narayanan et al. (1999) studied a Tamil /l/ using MRI from one male subject , but they did not study the effect of two lateral channels.

Much of the work on vocal tract modeling of /l/ was focused on the sources of zeros in the spectrum. Prahler (1998) studied the detailed effects of the two lateral channels with asymmetry, but the supralingual cavity was not considered. Prahler (1998) used a simple tube model for /l/ and was able to determine the zero in the spectrum for different areas and lengths of each channel. The first zero was found to be at $c/(l_1+l_2)$, where c is the speed of sound in air, and l_1 and l_2 are the lengths of each channel. Prahler (1998) has shown that if two uniform channels are of the same length, the pole-zero pair will be at the same frequency and, as a result, cancel each other. However, if two uniform channels are not of the same length, the poles and zeros will be at different frequencies. Prahler (1998) also has shown that, in order to produce a zero at 2-3 kHz, the combined length of the channels needs to be around 16 cm long, which is larger than that measured from MRI data (Narayanan et al., 1997; Zhang and Espy-Wilson, 2004) or predicted by Fant (1970).

Zhang and Espy-Wilson (2004) developed a vocal tract model with parallel lateral channels and a supralingual cavity, as shown in Figure 2.6. It was found that, with the dimensions estimated from MR images of a male speaker, the lateral channels produced a pole-zero pair in the frequency range of 2-5 kHz, and the supralingual cavity produced an additional pole-zero pair in the same frequency range. These two types of pole-zero pairs result in a low-amplitude and relatively flat spectral shape in the F3-F5 region. The subject's axial linguo-alveolar contact in this study was very small, about 1-2 cm long. But the effective lengths of the two lateral channels were longer due to the flow split property in the vocal tract. The vocal tract cross sections in the region immediately posterior to the midsagittal

contact were divided artificially into three regions which consist of one supralingual cavity and two lateral channels. The two lateral channels are longer than the actual length of the linguo-alveolar contact, with 5.0 cm for the channel 1 and 3.4 cm for the channel 2. Panchapagesan (2003) did a similar study using area functions extracted from MR images from two speakers, one female and one male. He found that the first major zero occurred around 1.5-3 kHz due to the supralingual cavity, and the second zero is around 2.5-4 kHz due to the asymmetry in the lateral channels. However, there is no work on a vocal tract model for the case where the linguo-alveolar contact is not complete so that there is no occlusion, but rather a constriction.

2.3 Problems with the current vocal tract models of liquids

There are several problems in the current vocal tract models of liquids:

- Most existing models of vocal tract acoustics for liquids have been developed based on an idealized vocal tract (Stevens, 1998) and some of the data were from midsagittal X-ray images (Delattre and Freeman, 1968; Fant, 1970). Only a few studies of liquids are carried out using MR images for area function data (Alwan et al., 1997; Espy-Wilson et al., 2000; Narayanan et al., 1997; Ong and Stone, 1998; Panchapagesan, 2003; Prahler, 1998; Story et al., 1996).

- There were very few subjects in the MRI studies of liquids. There were only one to four subjects in each of the previous studies (Alwan et al., 1997; Espy-Wilson et al., 2000; Narayanan et al., 1997; Ong and Stone, 1998; Story et al., 1996).

- Area function extraction of the liquids was based on an assumed vocal tract model configuration and planar wave propagation was assumed. For example, treating the front cavity for /r/ as a side branch, or treating the "cul-de-sac" supralingual space as a side branch for /l/ assumes a vocal tract model with a side branch. Three dimensional (3-D) acoustic analysis of the vocal tract is needed to validate these assumptions and to provide guidance on how to obtain the area functions for each channel or side branch.

- The acoustic response of the area function vocal tract model may not be able to estimate the formants and zero frequencies accurately since it is a simplification of the 3-D vocal tract. Finite element method (FEM) based acoustic analysis (Burnett, 1988; Motoki, 2002) is a standard procedure for studying the 3-D vocal tract acoustics and it can give us the ground truth for the acoustic response if the geometry reconstruction is accurate. Also, cross modes in the vocal tract may produce zeros (Motoki, 2002) which has to be revealed by the 3-D FEM analysis. Given the complexity of the vocal tract geometry for liquids, 3-D FEM should be employed in the vocal tract modeling of liquids.

- Due to the slow scanning speed of the MR machine, MRI's application is limited to sustained sound production, which means only static vocal tracts can be imaged. Generally, dynamic MRI is not readily available for studying the dynamics of the vocal tract, although Narayanan et al. (2004) got one MR slice at each instant for dynamic speech production and Takemoto et al.

(2006b) used cine-MRI to study temporal changes of the vocal tract area function.

- The ideal way of vocal tract modeling is to integrate a 3-D tongue model into the vocal tract model, so that any kind of articulatory configurations can be simulated by manipulating the tongue model. However, the current available tongue models are not good enough for this purpose (Badin and Serrurier, 2006; Dang and Honda, 2004; Engwall, 2003; Gerard et al., 2003; Stone, 1990; WilhelmsTricarico, 1995, 1996).

2.4 Acoustic feature variability and the speaker-discriminating property of liquid sounds

2.4.1 Acoustic feature variability

In addition to factors such as gender, dialect, vocal tract length, and speaking styles, diversity of the tongue shape for liquids might be another factor that increases the inter-speaker variability in speech. Given the various articulatory configurations produced for liquids across speakers, the acoustic inter-speaker variability of liquids might be relatively larger than it is for other sounds. This inter-speaker variability may be beneficial in distinguishing one speaker from another. This idea motivated Goldstein (1976) to look at features based on formants track, and statistics F-ratios were calculated and speaker identification tasks were performed in his study. The formant structures of three diphthongs, four tense vowels and three retroflex sounds

were examined for possible speaker-identifying features. The inter-speaker variability of about 200 measures made on these formant tracks was compared initially with the intra-speaker variability through the calculation of F-ratios. The two features that were most effective in identifying speakers were the minimum second formant value in /r/ after the vowel /aa/ and the maximum first formant of /r/ after the vowel /aa/. The drawback of this study is that the database included only ten speakers of American English with ten sentences and ten repetitions. Nolan (1983) studied the intra- and inter-speaker variabilities of /r/ and /l/ in terms of F1-F3. In his study, fifteen speakers were used and each one read fifteen words including /r/ and fifteen words including /l/. He found that the liquids /r/ and /l/ provide moderate performance in speaker identification, and pointed out they are less useful than the nasal sounds. In Nolan's study, /r/ F-ratios in terms of F1-F3 were found to be larger than the corresponding F-ratios in /l/. This might be because of the less intra-speaker variability in /r/ and greater degree of coarticulation of /l/. However, the speakers spoke British English in Nolan's study and British English has less variety of tongue shapes for /r/ than what occurs in American English (Delattre and Freeman, 1968).

Although there are some studies on analyzing variability in speech in terms of phonetic, contextual, channel, and speaker variability (Kajarekar, 2002; Kajarekar et al., 1999), there is no thorough study on phoneme variability in American English. Sun and Li (1995) performed ANOVA (Analysis of Variance) analysis for individual phonemes, using mel-frequency cepstrum coefficients (MFCC) extracted from the TIMIT database (TIMIT, 1990). However, the results for liquids in terms of speaker

variability were not reported.

2.4.2 Speaker-discriminating ability

Without studying acoustic feature variability of phonemes, Eatock and Mason (1994) studied the relative speaker discriminating properties of phonemes in British English by performing a speaker verification test. The database used is of telephone quality, comprising 125 speakers each uttering six sentences from a pool of 201 sentence-texts. The input features consisted of 12th order cepstral coefficients. The nasal sounds and the vowels were found to provide the best performance, followed by fricative, affricates and liquids. Similar results were found in studying Dutch phonemes (Heuvel and Rietveld, 1992).

Phonetic class-based speaker verification is a natural refinement of the traditional single Gaussian mixture model (GMM) scheme. Its objective is to model the voice characteristics at the level of the phoneme (Antal and Toderean, 2006; Auckenthaler et al., 1999; Faltlhauser and Ruske, 2001; Hebert and Heck, 2003; Kajarekar and Hermansky, 2001). It is desirable to find optimal class-specific acoustic features for modeling each phoneme. Using the NIST (The National Institute of Standards and Technology) speaker verification evaluation data which is sampled at 8 kHz, Kajarekar and Hermansky (2001) observed that vowel, diphthongs, nasals and fricatives are the most important sounds for speaker verification. However liquids were not included in the GMM model. Faltlhauser and Ruske (2001) used the German Verbmobil database and included liquids as a class in the GMM model.

Auckenthaler et al. (1999) found that while there is a strong correlation between performance and the amount of training data, there is also an obvious difference in discriminating ability among phonemes with the same amount of data. Antal and Toderean (2006) used the TIMIT database and built pure phonetic GMMs. It was observed that the discriminating power of phonemes was ranked in the order of vowels, nasals, fricative and semivowels. However, if the training and test data for each phoneme is made equal, the order was changed to nasals, vowel, semivowels, and fricatives.

Hayakawa and Itakura (1994) found that the wider the frequency range is, the higher the recognition rate is, and some speakers show significantly better performances using the higher frequency band than using the lower one, so it is concluded that there is a rich amount of speaker information contained in the higher frequency band. Lin et al. (1996) did a study on high frequency performance in speaker identification task in TIMIT using MFCCs and found that the high frequencies band, 3.5-7 kHz, contain more reliable idiosyncratic information about the speaker.

2.5 Chapter summary

This chapter presents a literature survey on acoustics, articulation and vocal tract modeling of /r/ and /l/ in American English. It also presents past work on acoustic variability studies for phonemes and phoneme-based speaker recognition.

Previous studies of vocal tract modeling of liquids had limited articulatory data from very few speakers. The vocal tract models of liquids in past literature

were based on vocal tract area function which assumes planar wave propagation and neglects the 3-D property of acoustics. A comprehensive study is needed on typical articulatory configurations of liquids with MR images and acoustic data from more speakers. Medical image processing software and the 3-D FEM tool are useful to facilitate the detailed acoustic analysis on the complex vocal tract geometries for liquid sounds. Based on MR images and 3-D FEM, the studies in Chapters 4 and 5 of this dissertation attempt to shed some light on vocal tract acoustics for some typical articulatory configurations of liquids, i.e., retroflex /r/ vs. bunched /r/, and light /l/ vs. dark /l/.

There are few studies on the acoustic feature variation and speaker discriminative power of /r/ and /l/ in American English. The databases in previous studies were limited to read words or sentences, or limited to telephone quality speech. No study has been done on spontaneous speech sampled at 16 kHz. Such studies might be helpful in finding optimal class-specific acoustic features for modeling the liquids in a phonetic class-based speaker verification system. The study in Chapter 6 of this dissertation attempts to find out the acoustic variability and speaker discriminative power of liquids in a relatively large database which contains spontaneous speech. For comparison, the results for liquids will be presented along with the results for other sounds (vowels, nasals, glides, fricatives, affricates and stops).

Chapter 3

Databases, tools and methodologies

This chapter describes the groundwork that has been done for the vocal tract modeling of the liquids in American English and for their acoustic variability and discriminative power study. This includes the details of two databases used in this dissertation, the tools used or developed to process and analyze image and acoustic data along with the details of the methodologies for 3-D vocal tract reconstruction, 3-D finite element analysis, the FEM-based area function extraction and the other techniques necessary for this dissertation.

3.1 Databases

3.1.1 UC Database

the UC database was collected at the University of Cincinnati, USA by our collaborators. This database was created for articulatory and acoustic studies of liquids in American English and includes the MR images of the vocal tracts for sustained American English /r/ and /l/ with a variety of tongue shapes from different subjects (Tiede et al., 2004). Those subjects are from many different states, with age ranging from 21 to 48. Their midsagittal MR images are shown in Figures 3.1 and 3.2. The subjects were instructed to pronounce sustained sounds (/r/ as in "pour", and /l/ as in "pole") while they were being scanned by the MR machine.

For each subject, MR images from coronal, axial and sagittal orientations were obtained. Detailed MR scanning information is in Section 3.1.1.1. In addition to the MR images, dental casts and CT (Computed Tomography) images of the dental casts were obtained for each speaker for teeth compensation in the vocal tract segmentation. This procedure was used because the MR machine cannot image the teeth. Acoustic recordings of the sustained sounds, the nonsense words, and real words were collected for further acoustic analysis.

There are 22 subjects (13 males and 9 females) in the UC database. Among them, subjects 22 and 5 are selected for the comparison study of the vocal tract modeling of liquids in this dissertation. These two subjects have similar vocal tract dimensions and produce a typical retroflex /r/ and a typical bunched /r/, respectively. Subject 5 can produce both a sustained light /l/ and a sustained dark /l/. For convenience, subject 22 is renamed S1 and subject 5 is renamed S2 in the remaining chapters. The details about S1 and S2 are presented in Section 4.2 on page 49.

3.1.1.1 Image acquisitions

MR imaging in the UC database was performed on a 1.5 Tesla G.E. `Echospeed` MR scanner with a standard phased array neurovascular coil at the University Hospital of the University of Cincinnati, USA. Subjects were positioned in supine posture, with their heads supported by foam padding to minimize movement. The subjects were instructed to remain motionless to the extent possible during and between

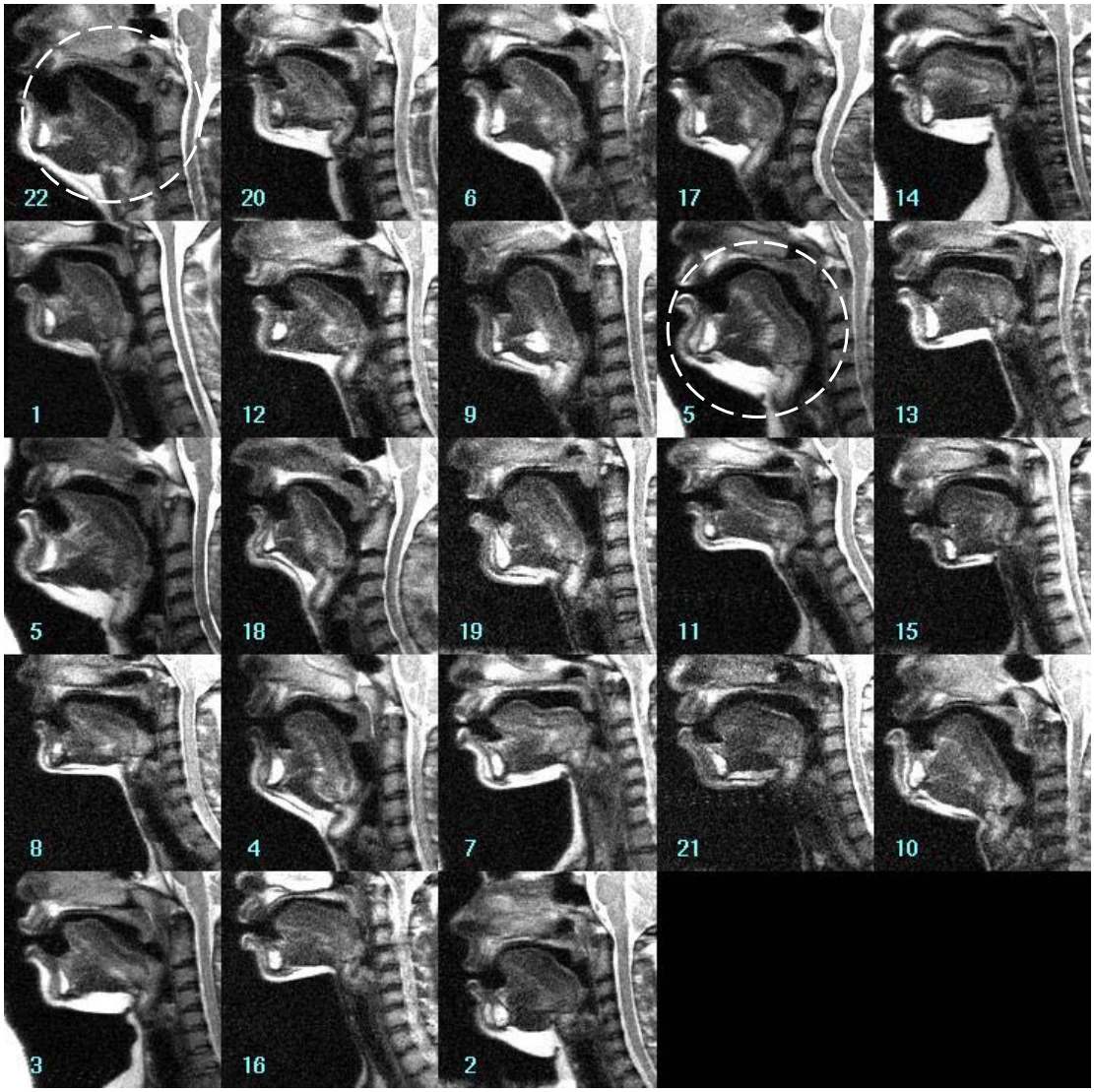


Figure 3.1: Midsagittal MR image of /r/ for all 22 subjects in the UC database (Tiede et al., 2004). (Subjects with circle on images were studied in this dissertation, subjects 22 and 5 are renamed as S1 and S2 respectively in the remaining chapters.)

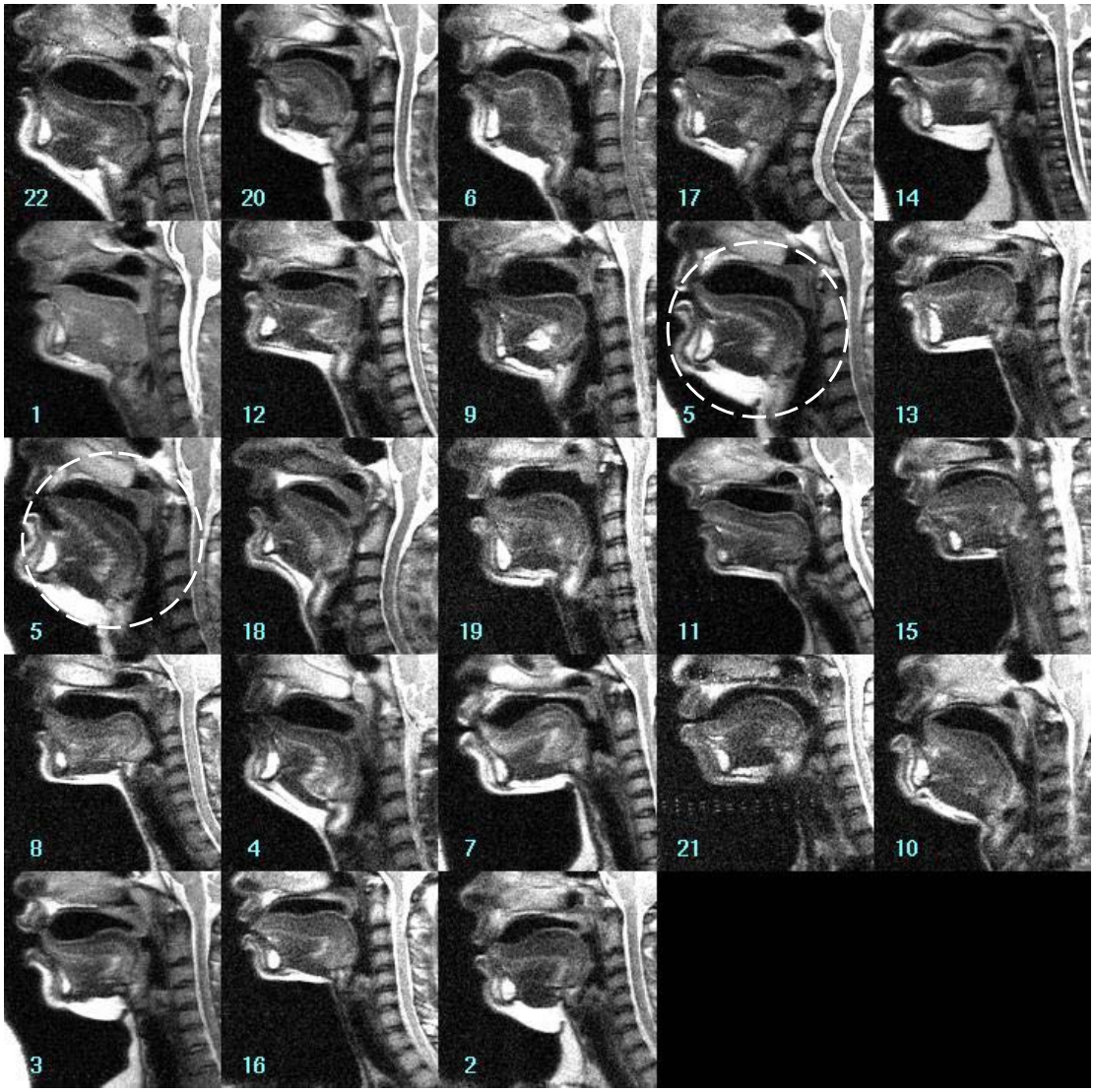


Figure 3.2: Midsagittal MR image of /l/ for all 22 subjects in UC database (Tiede et al., 2004). (Speakers with circle on images were studied in this dissertation, subjects 22 and 5 are renamed as S1 and S2 respectively in the remaining chapters.)

scans. For hearing protection and comfort, subjects wore earplugs during the entire session. In addition, subjects' ears were covered by padded earphones. Localization scans were performed in multiple planes to determine the optimal obliquities for orthogonal imaging. A midsagittal plane was identified from brain morphology. Axial and coronal planes were then oriented to this midsagittal plane. During each subsequent scan, the subject was instructed to produce sustained /r/ as in "pour" or sustained /l/ as in "pole"for a defined period of time (between 5 and 25 seconds depending on the sequence). T2 weighted 5 mm single shot fast spin echo images were obtained in the midline sagittal plane with two parasagital slices. T1 weighted fast multiplanar spoiled gradient echo (FMPSPGR) images (TR 100-120 ms, TE 4.2 ms, 75 degree flip angle) were obtained in the coronal and axial planes with a 5 mm slice thickness. There was no gap between adjacent slices. The scanning regions for the coronal and axial planes include the region from the surface of the vocal folds to the velopharyngeal port and the region from the rear wall of the velopharynx to the outside edge of the lips. Depending on the dimensions of the subjects' vocal tract, the data set comprised 24 to 33 images in the axial and coronal planes. For all images, the field of view was 240 mm by 240 mm with an imaging matrix of 256 x 256 to yield an in-plane resolution of 0.938 mm per pixel. The MR imaging technique used does not distinguish between bony structures such as teeth and air, due to the low levels of imageable hydrogen. Thus, to avoid overestimation of oral tract air space, CT scans of each subject's dental cast were acquired on a GE Lightspeed Ultra multidetector scanner with a slice thickness of 1.25 mm, subsequently superimposed on the volumes derived from MRI as described below. Images were

resampled to 1.25 mm at 0.625 mm intervals to optimize 3-D modeling. The field of view was 120 mm with an imaging matrix of 512 x 512 to yield an in-plane image resolution of 0.234 mm per pixel.

3.1.1.2 Acoustic signal recording

During the MRI sessions, the subject's phonation in the supine position was recorded using a custom-designed microphone system (Resonance Technology Inc.), and continuously monitored by a trained phonetician to ensure that the production of /r/ remained consistent over the course of the experiment. Subjects were instructed to begin phonation prior to the onset of scanning, and to continue to phonate for a period after scanning was complete. A full audio record of the session was preserved using a portable DAT tape recorder (SONY TD-800). Due to the noise emitted by the scanner during the scans, the only portions of the subject's productions of /r/ or /l/ that can be reliably analyzed occur in the 500 ms after phonation began, and before the scanner noise commenced, and in the 500 ms after the scanner noise ceased while the subject continued to phonate. The recordings are still quite noisy, but it was possible to measure F1-F3 with reasonable accuracy during most scans. Subjects were also recorded acoustically in separate sessions in a sound-treated room, using a Sennheiser headset microphone and a portable DAT tape recorder (SONY TD-800). Subjects recorded a set of utterances encompassing sustained productions of /r/ or /l/ plus a number of real and nonsense words containing /r/ and /l/. As in the MR condition, subjects were instructed to produce

/r/ as in “pour” or /l/ as in “pole”. In addition, they recorded sustained /r/ as in “right”, “read”, “role”, “feel”, “light” and “lee”. For the sustained productions, subjects were recorded in both upright and supine postures. The nonsense words were “warav”, “wavrav”, “wadrav”, and “wagrav”, repeated with stress either on the first syllable or the second syllable. The real words included /r/ and /l/ in word-initial, word-final and intervocalic positions. For the real and nonsense words, subjects were recorded in the upright posture. Acoustic data recorded in the sound-proofed room are referred to as sound booth acoustic data. Recording conditions were such that, in addition to F1-F3, F4 and F5 could be measured reliably.

3.1.2 Buckeye database

The Buckeye database (Pitt et al., 2005) is a free available corpus of spontaneous speech in American English. The purpose of creating this database was to study phonological variation and its effects on speech recognition by human and machines. It includes conversational speech of 40 speakers from central Ohio, USA (half male and half female). The duration for each speaker’s conversation is about 30-60 minutes, sampled at 48 kHz. It has 307,000 words which are phonemically labeled, so that this database can be used to analyze the acoustic variability for each phoneme, and to carry out the phoneme-based speaker identification task in this dissertation. It has more words phonemically labeled than some other similar databases such as the TIMIT database (TIMIT, 1990) (6,300 words) and a subset of the Switchboard database (Greenberg, 1997) (35,000 words of bandlimited

telephone speech).

3.2 Tools

3.2.1 MIMICS

The medical image processing software package MIMICS (Materialise, 2007) was used to segment the vocal tract from MR images and to obtain a 3-D reconstruction of the vocal tract. This software has been widely employed in the medical imaging field for MRI and CT image processing, for rapid prototyping, and for 3-D reconstruction in surgery.

3.2.2 COMSOL MULTIPHYSICS

The FEM software COMSOL MULTIPHYSICS package (Comsol, 2007) was used in 3-D acoustic analysis. In addition to supporting importation of CAD geometries such as the STL file format, this software is also capable of solving problems which have several partial differential equations (PDEs) coupled together. This multiphysics modeling feature is very useful in the case where the wall compliance, viscosity and heat conduction in the 3-D vocal tract are included in the modeling.

3.2.3 VTAR

VTAR (Vocal tract acoustic response) is a Matlab-based computer program for vocal tract acoustic response calculation (Zhou et al., 2004). Based on a frequency-domain model (Zhang and Espy-Wilson, 2004), VTAR is able to model various com-

plex tube configurations such as a side branch and a two-channel module along with sets of area functions. It considers the acoustic effect of radiation, viscosity, heat conduction, and wall property. With input in the form of vocal tract cross-sectional area functions, VTAR calculates the vocal tract acoustic response and the formant frequencies and bandwidths. The user-friendly interface allows directed data input. The program also provides an interface for input and modification of arbitrary vocal tract geometry configurations, which is ideal for research applications. In addition to the vocal tract acoustic response, VTAR also provides modules for format sensitivity functions, susceptance plots, area function modification for targeted formant patterns, and sound synthesis based on the vocal tract acoustic response.

3.2.4 MIT Lincoln lab speaker recognition system

The MIT Lincoln Lab's speaker recognition system (Reynolds et al., 2000) was used for testing the discriminating ability of liquids and other sounds in this dissertation. This system uses a GMM-UBM model for an identification task (GMM:Gaussin Mixture Model, UBM: Uniform Background Model). A UBM model is constructed based on all of the training data, and then the UBM model is adapted for each speaker's GMM model. For the identification task, the scores for individual's speaker models are compared, and the speaker whose model has the highest score is the hypothesized speaker.

3.3 Methodologies

3.3.1 Image processing and 3-D vocal tract reconstruction

The reconstruction of 3-D vocal tracts using MIMICS proceeded in four steps. Step (1) involved the segmentation between the tissue of the vocal tract and the air space inside the vocal tract for each MR image slice in the coronal and axial sets. Because the cross-section of the oral cavity is best represented by the coronal slices, and the cross-section of the pharyngeal and laryngeal cavities are best represented by the axial slices, the following procedure was used to weight them approximately. First, the segmented axial slices were transformed into a 3-D model. Then, the coronal slices were overlapped with the axial-derived model. As in Takemoto et al. (2006b), we extended the cross-sectional area of the last lip slice with a closed boundary halfway to the last slice in which the upper and lower lip are still visible. The coronal slice segmentation in the pharyngeal and laryngeal cavities was then corrected by reference to the axial slice 3-D model. Step (2) involved compensation for the volume of the teeth using the CT scans, which were made in the coronal plane. The CT images were segmented to provide a 3-D reconstruction of the mandible and the maxillae with the teeth. This process was considerably easier than for the MR slices described above, given the straightforward nature of the air/tissue boundary in that imaging modality. The 3-D reconstruction of the dental cast was then overlapped with the MRI coronal slices. The reconstruction of the maxilla cast was positioned on the MR images by following the curvature of the palate. The reconstruction of the mandible cast was positioned with reference to the boundary

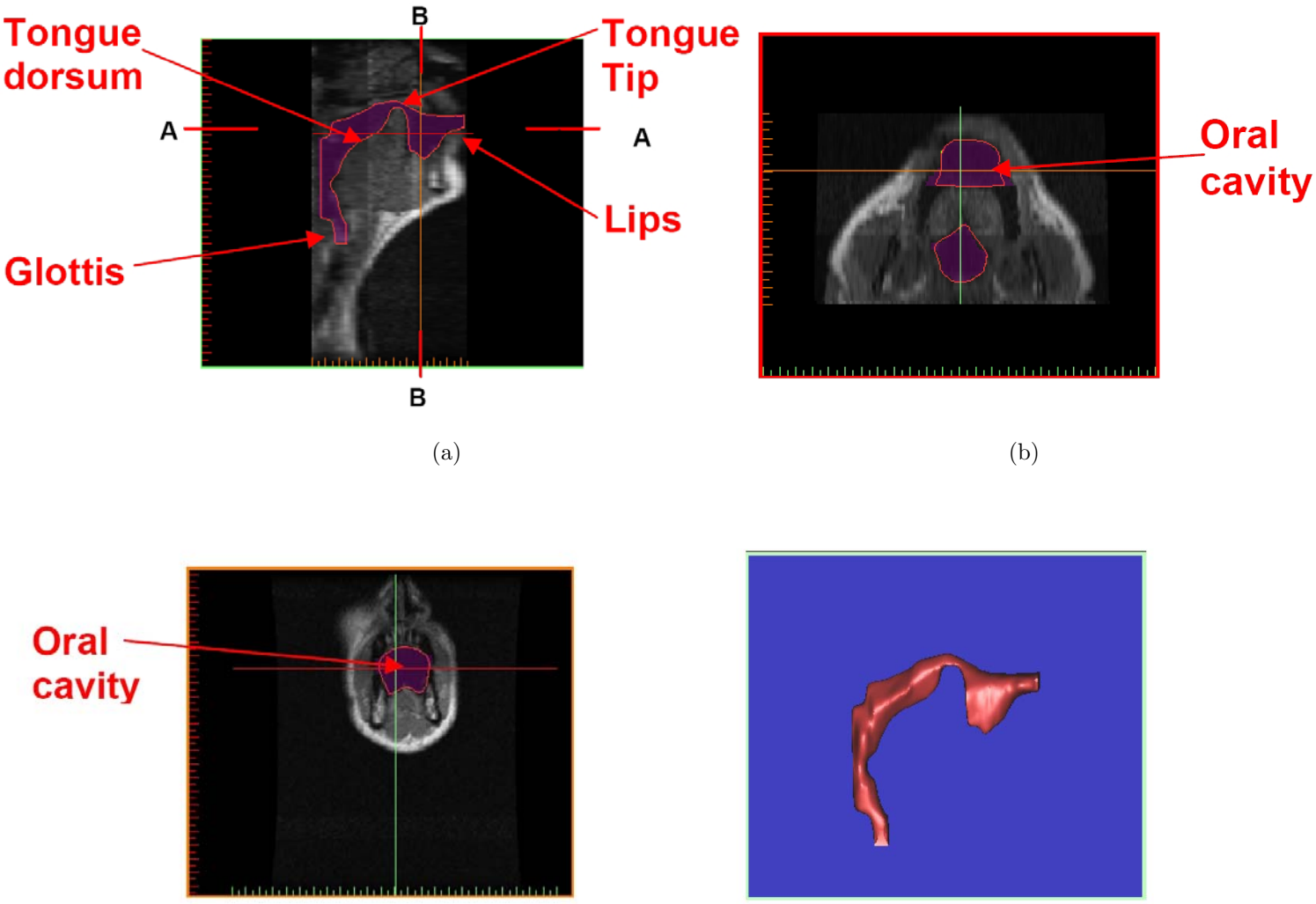
provided by the lips. In Step (3), the final segmentation was translated into a surface model in STL (STereoLithography) format (Lee, 1999). Finally, the 3-D geometry surface was smoothed using the MAGICS software package (Materialise, 2007). The validity of the reconstructed 3-D vocal tract geometry was evaluated by comparing midsagittal slices created from the reconstructed 3-D geometry with the original midsagittal MR images. This method was also used to check for the possibility that subjects had changed their vocal tract configuration for sustained /r/ or /l/ across scans. The data sets of all the subjects in this study show very good consistency, and overall boundary continuity between the tissue and the airway was achieved successfully. Figure 3.3 shows the MR images in three views with the overlapped 3-D reconstruction of the vocal tract (dental casts are not displayed here).

3.3.2 3-D finite element analysis

The finite element method (FEM) was used in the acoustic simulation to obtain the acoustic response of the 3-D vocal tract and to study the wave propagation at different frequencies. The pressure isosurfaces at low frequency were used to extract area functions. The governing equation for this harmonic analysis is the Helmholtz equation,

$$\nabla . (\frac{1}{\rho} \nabla p) + \frac{\omega^2 p}{\rho c^2} = 0 \tag{3.1}$$

where p is the acoustic pressure, ρ (1.14 kg/m3) is the density of air at body temperature, c (350 m/s) is the speed of sound, and ω is the angular frequency ($\omega = 2\pi f$, where f is the vibration frequency in Hz and the highest frequency in the harmonic



(a) (b)

(c) (d)

Figure 3.3: Segmentation of the 3-D vocal tract from MR images. (a) midsagittal view, (b) axial view (A-A), (c) coronal view (B-B), (d) reconstructed 3-D vocal tract.

analysis is 8000 Hz). The boundary conditions for the 3-D finite element analysis are as follows, Glottis: Normal velocity profile as sinusoidal signal at various frequencies Wall: Rigid Lips: The radiation impedance Z of an ideal piston in an infinitely flat baffle (Morse and Ingard, 1968).

$$Z = \rho c(1 - J_1(2k\alpha)/(k\alpha))_j K_1(2k\alpha)/(2k\alpha)) \tag{3.2}$$

where $k = 2\pi f/c$, $\alpha = \sqrt{A_1/\pi}$, (A_1 is the area of the lips opening), J_1 is the Bessel function of order 1, K_1 is the Struve function of order 1. The volume velocity at the lips is measured by velocity integration over the cross section at the lips, and the acoustic response of the vocal tract is defined as the volume velocity at the lips divided by the volume velocity at the glottis. Note that for the purpose at hand, the ideal piston model has been shown to be computationally equivalent to a 3-D radiation model at the lips (Matsuzaki et al., 1996). The mesh for FEM was created using tetrahedral elements as in the STL format.

3.3.3 Area function extraction

Area functions were generated by treating the vocal tract as a series of uniform tubes with varying areas and lengths. The extraction of area functions from imaging data is typically an empirical process. Baer et al. (1991), Narayanan et al. (1997) and Ong and Stone (1998) based their area function extractions on a semi-polar grid (Heinz and Stevens, 1964). In contrast, Chiba and Kajiyama (1941), Story (2006) and Takemoto et al. (2006b) extracted area functions by computing a centerline in air space and then evaluating the cross-sectional areas within planes chosen to

be perpendicular to the centerline extending from the glottis to the mouth. Our area functions were derived from the 3-D FEM model, so it might be expected that the area function simulation and the simulated acoustic response from the 3-D model should be the same. However, it should be noted that area function extraction, by transforming the bent 3-D geometry of the vocal tract into a straight tube with varying cross-sectional areas (Chiba and Kajiyama, 1941; Fant, 1970), necessarily involves considerable simplification. An additional and related problem is that it assumes planar wave propagation and thus tends to neglect cross-mode wave propagation and potential anti-resonances or zeros. Thus, we expect some small differences between the simulation results using area function analysis and planar wave propagation from simulation results obtained directly from the corresponding 3-D geometry (Sondhi, 1986).

In this study, the low-frequency wave propagation properties resulting from the 3-D finite element analysis were used to guide the area function extraction from the reconstructed 3-D geometry. This approach is quite similar to the centerline approach. The logic of this procedure was as follows. As noted above, area function-based vocal tract models assume planar wave propagation. Finite element analysis at low frequencies such as 400 Hz (around F1 for /r/) produces pressure isosurfaces that indicate approximate planar acoustic wave propagation. Thus, a tube model derived from area functions whose cutting plane follows these pressure isosurfaces should constitute a reasonable 1D model for the 3-D vocal tract. In this study, as the curvature of the vocal tract changes, the cutting orientation in our method was adjusted to be approximately parallel to the pressure isosurface at 400 Hz. This

procedure was performed by recording the coordinates of the isosurfaces. Those coordinates are then used to determine the cutting planes. The distance between two sampling planes was set to be the distance between their centroids. Vocal tract length was estimated as the cumulative sum of the distance between the centroids. The cutting plane gap was about 3 mm. Since this method was based on the 3-D reconstructed geometry instead of sets of MR images, pixel counting and other manipulations such as reslicing of images are not needed. The area calculation is based on the geometric coordinates of the reconstructed vocal tract. As noted above, the reduction of a vocal tract 3-D model to area functions requires considerable simplification. To assess the degree to which our area function extraction preserved essential aspects of the vocal tract response, we compared the simulation output from the 3-D FEM model to the acoustic response of VTAR. The vocal tract response from the 3-D model and from VTAR were, in turn, evaluated by comparison with formant measurements from real speech produced by the subjects, as described in Section 4.5 on page 55.

3.3.4 Formant measurement of acoustic data

Formants from both sound booth and MR acoustic recordings were measured by an automatic procedure that computed 24th order linear prediction coefficients (LPC) over a 50 ms window from a stable section of the sustained production. The 50 ms window for the MR acoustic data was taken from the least noisy segment of the approximately 500 ms production preceding the onset of MR scanning noise.

Only F1-F3 were measured in the MR acoustic recording because the noise in the high frequency region masked the higher formants very effectively. To maximize the comparability of the MR and sound booth acoustic measures, the latter were measured from productions recorded when the subjects were in supine posture. The formant values of the sustained /r/ or /l/ in MRI sessions are the average of the measurements from all the scans including midsagittal, axial and coronal scans.

3.3.5 Formant sensitivity functions

The acoustic sensitivity of one specific formant frequency to change of the vocal tract area function is used to analyze the formant-cavity affiliations. If one formant is only sensitive to a certain part of the vocal tract, that means that formant is produced by that part of the vocal tract.

The acoustic sensitivity function of the formants is defined as the difference between the kinetic energy (KE) and potential energy (PE) as a function of distance starting from the glottis, divided by the total energy (TE) (sum of kinetic and potential energy in the system)(Fant and Pauli, 1974; Story, 2006). The sensitivity function is written as

$$S_n(i) = \frac{KE_n(i) - PE_n(i)}{TE_n(i)} \quad \text{n=1-5 and i=1-N} \tag{3.3}$$

where n is the formant number, and i is the section number of the vocal tract area function. Section 1 is the first section starting from the glottis, and N is the last section at the lips, and

$$TE_n = \sum_{i=1}^{N} (KE_n(i) + PE_n(i)) \tag{3.4}$$

$$KE_n(i) = \frac{1}{2}\frac{\rho l(i)}{\alpha(i)}|U_n(i)|^2 \tag{3.5}$$

$$PE_n(i) = \frac{1}{2}\frac{\alpha(i)l(i)}{\rho c^2}|P_n(i)|^2 \tag{3.6}$$

where $\alpha(i)$ and $l(i)$ are the cross section area and length of section i of the vocal tract area function respectively. $U(i)$ and $P(i)$ are the volume velocity and pressure at section i. ρ is the density of air and c is the speed of sound. The relative formant change corresponding to the area function change is described by Equation 3.7 where F_n is the *nth* formant, $\triangle F_n$ is the change of the *nth* formant, A_i is the area of the *ith* section and $\triangle A_i$ is the area change of the *ith* section.

$$\frac{\triangle F_n}{F_n} = \sum_{i=1}^{N} S_n(i)\frac{\triangle A_i}{A_i} \tag{3.7}$$

3.4 Chapter summary

In this chapter, the details of two databases used in this dissertation are presented along with the tools used or developed to process and analyze image and speech data. Details of the methodologies used for 3-D vocal tract reconstruction, 3-D FEM, area function extraction and other techniques necessary for this dissertation are also presented. All the findings described in chapters 4, 5 and 6 are based on the databases and techniques presented here.

Chapter 4

Acoustic modeling of retroflex /r/ and bunched /r/

4.1 Introduction

This chapter presents the study of two subjects who have similar vocal tract anatomy but produce very different bunched and retroflex tongue shapes for /r/. The vocal tract midsagittal MR images of these two subjects are shown in the top panel of Figure 4.1. As the middle panel of Figure 4.1 shows, the subjects' acoustic profiles resemble those discussed in Delattre and Freeman (1968) and Westbury et al. (1998) in that their F1, F2 and F3 values are similar. However, the two subjects also show very different patterns for F4 and F5. In particular, the distance between F4 and F5 for the retroflex /r/ is double that for the bunched /r/. The lower panel of Figure 4.1 shows examples of the same F4/F5 pattern drawn from running speech from production of the nonsense word "warav". The question of whether different patterns of the higher formants are a consistent feature of bunched vs. retroflex tongue shape has been asked. If so, this difference in acoustic signatures may be useful for a number of purposes that involve the mapping between articulation and acoustics, i.e. speaker recognition, articulatory training, speech synthesis, etc. Alternatively, the different patterns of F4 and F5 may derive from structures independent of tongue shape: for instance, additional cavities in the vocal tract such as the laryngeal vestibule (Kitamura et al., 2006; Takemoto et al., 2006a), or the

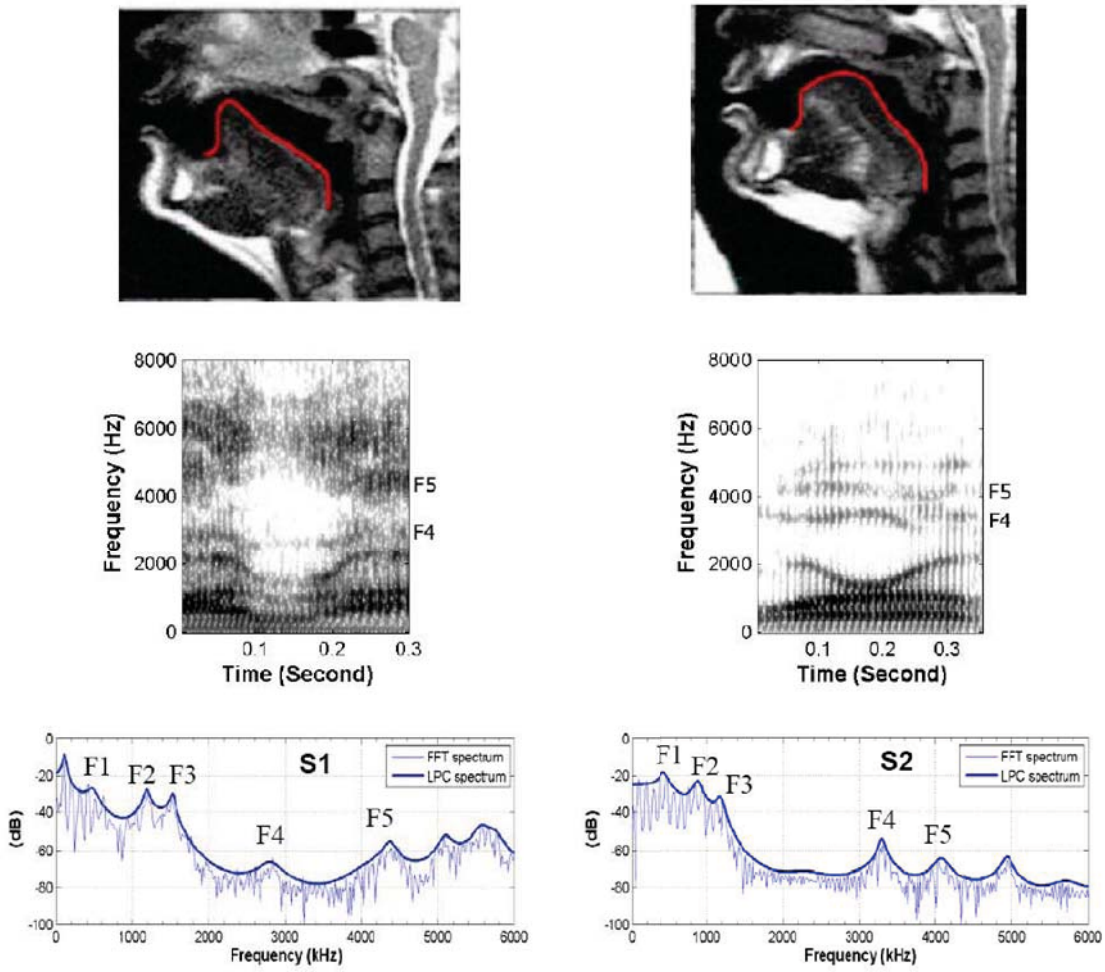


Figure 4.1: Top panel: Midsagittal MR images of two tongue configurations for American English /r/. Middle panel: Spectrograms for nonsense word "warav". Lower panel: Spectra of sustained /r/ utterance. The left side is for S1 and the right side is for S2.

piriform sinuses (Dang and Honda, 1997). The key piece of evidence is whether such structures differ in such a way as to explain the F4/F5 patterns across /r/ types.

The task of understanding this difference in formant pattern has been approached in the following way. First, magnetic resonance imaging (MRI) was used to acquire a detailed three-dimensional geometric reconstruction of the vocal tract. Second, finite element analysis has been performed to simulate the acoustic response of the 3-D vocal tract and to study the wave propagation properties at different frequencies. Third, area function models were obtained from the FEM analysis of 3-D geometry. The resulting simulated acoustic response was verified against the 3-D acoustic response. The area function models were then used to isolate the effects of formant cavity affiliations by formant sensitivity functions and simple-tube models. The results of the simulation were compared to actual formant values from the subjects.

4.2 Subjects

As described in Section 3.1.1, subjects 22 and 5 in the UC database were used to study /r/, and they are renamed as S1 and S2 respectively. As Figure 4.1 shows, S1 produces a retroflex /r/ and S2 produces a bunched /r/. Both subjects are male. S1 was 48 years old and S2 was 51 years old at the time the data were collected. S1 had lived in California, Minnesota and Connecticut and S2 had lived in Texas, Massachusetts and Southwestern Ohio. Both spoke a rhotic dialect of American English. The subjects were similar in palate length, palate volume, overall stature,

Table 4.1: Dimensions of S1 and S2 in overall height, and volume, length, depth, and width of the palate. The measurements of the palate are based on the dental casts of the subjects. The width of the palate is the distance between edges of the gum between the second premolar and the first molar on both sides of the upper jaw. The length of the palate is the distance of the edges of the gum between the upper middle two incisors and the cross section of the posterior edge of the back teeth. The depth of the palate is the distance from the floor of the mouth to the cross section with the lateral plane. The volume of the palate is the space surrounded by the margin between the teeth and gums, the posterior edge of the back teeth, and the lateral plane. Several techniques have been used to calculate the volume, all of which gave the same answer within a certain range, and the average volume as a matter of displacement in water is reported here. That measure was done three times.

	S1	S2
Height of subject	188 cm	188 cm
Length of palate	35.8 mm	33.6 mm
Depth of palate	16.1 mm	13.2 mm
Width of palate	25.5 mm	25.0 mm
Av. volume of palate	29.1 mm^3	29.1 mm^3
Maxillary teeth volume	3.4 mm^3	3.3 mm^3

and vocal tract length (see Table 4.1). The data from S1 and S2 are also compared to that from other subjects with similar retroflex or bunched tongue shapes for /r/ collected in the larger study. These subjects are referred to as S3, S4, S5 and S6 and they are subjects 1, 20, 17, 19 in the UC database respectively. As described in Section 3.1.1, the articulatory data collected for all subjects includes MRI scans of the vocal tract for sustained natural /r/ or /l/, dental cast measurements and Computed Tomography (CT) scans of the dental casts, and acoustic recordings made at various points in time.

4.3 Reconstructed 3-D vocal tract geometries

The reconstructed 3-D vocal tract shapes for the retroflex /r/ of S1 and the bunched /r/ of S2 are shown in Figure 4.2. The two shapes are significantly different in several dimensions that are likely to cause differences in cavity affiliations. First, S1's retroflex /r/ has a shorter and more forward palatal constriction, leading to a slightly smaller front cavity. At the same time, the lowered tongue dorsum of the retroflex /r/ leads to a particularly large volume of the mid cavity between the palatal and pharyngeal constrictions. Further, the transition between the front and mid cavities is sharper for the retroflex /r/. This difference makes it more likely that the front and mid cavities are decoupled for the retroflex /r/ of S1 than for the bunched /r/ of S2. Unlike the speakers analyzed in Alwan et al. (1997) and Espy-Wilson et al. (2000), neither S1 nor S2 shows a sublingual space whose geometry is clearly a side branch to the front cavity. However, the two subjects' overall vocal

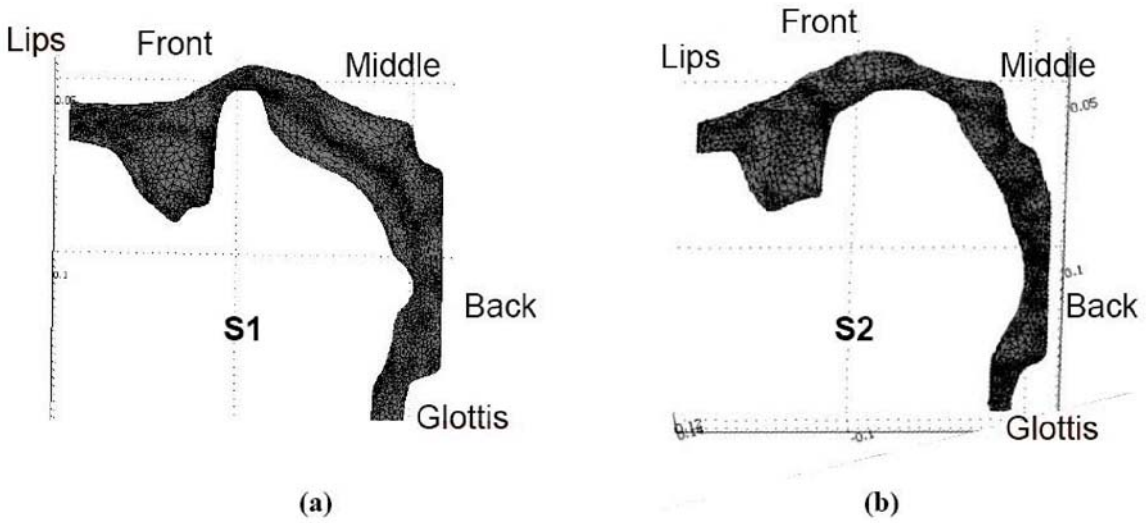


Figure 4.2: FEM mesh of the reconstructed 3-D vocal tract. (a) the retroflex tongue shape, (b) the bunched tongue shape.

tract dimensions from the 3-D model are very similar. These dimensions are shown in Table 4.2.

As noted above, the difference in the F4/F5 formant pattern between S1 and S2 must be derived from a difference in vocal tract dimensions, either in small structures such as the piriform sinuses and laryngeal vestibule (Dang and Honda, 1997; Kitamura et al., 2006; Takemoto et al., 2006a) or in tongue shape differences. The laryngeal vestibule cavities were included in the 3-D model, but given the resolution of the MR data, the representation is relatively crude. The dimensions of the piriform sinuses were measured and found to be similar to the range in length of 16 to 20 mm, and in volume of 2-3 cm^3 reported in Dang and Honda (1997). Because no significant differences were found between the subjects for either structure it is concluded that the tongue shape differences between S1's retroflex and S2's bunched

Table 4.2: Measurements on the reconstructed 3-D vocal tract in surface model (STL file format).

	S1	S2
X dimension	51 mm	46 mm
Y dimension	106 mm	107 mm
Z dimension	106 mm	100 mm
Volume	62 909 mm^3	48 337 mm^3
Surface area	14 394 mm^2	12 243 mm^2

/r/ are likely the major factor determining their differences in F4/F5 pattern. Possibly these cavities at the glottal end of the vocal tract are less influential for /r/ than for vowels due to the greater number, length, and narrowness of constrictions involved.

4.4 FEM-based acoustic analysis and the derived area function vocal tract models

In previous work, FEM analysis has been used to study the acoustics of the vocal tract for open vocal tract sounds, i.e. vowels (Matsuzaki et al., 2000; Miki et al., 1996; Motoki, 2002; Thomas, 1986). Zhang et al. (2005) applied this approach to a 2-D vocal tract for a schematized geometry based on a single subject producing /r/. In this study, the work by Zhang et al. (2005) has been extended by computing the pressure isosufaces at various frequencies to 3-D vocal tract shapes based on S1's retroflex and S2's bunched /r/. As Figure 4.3 shows, the retroflex and bunched /r/

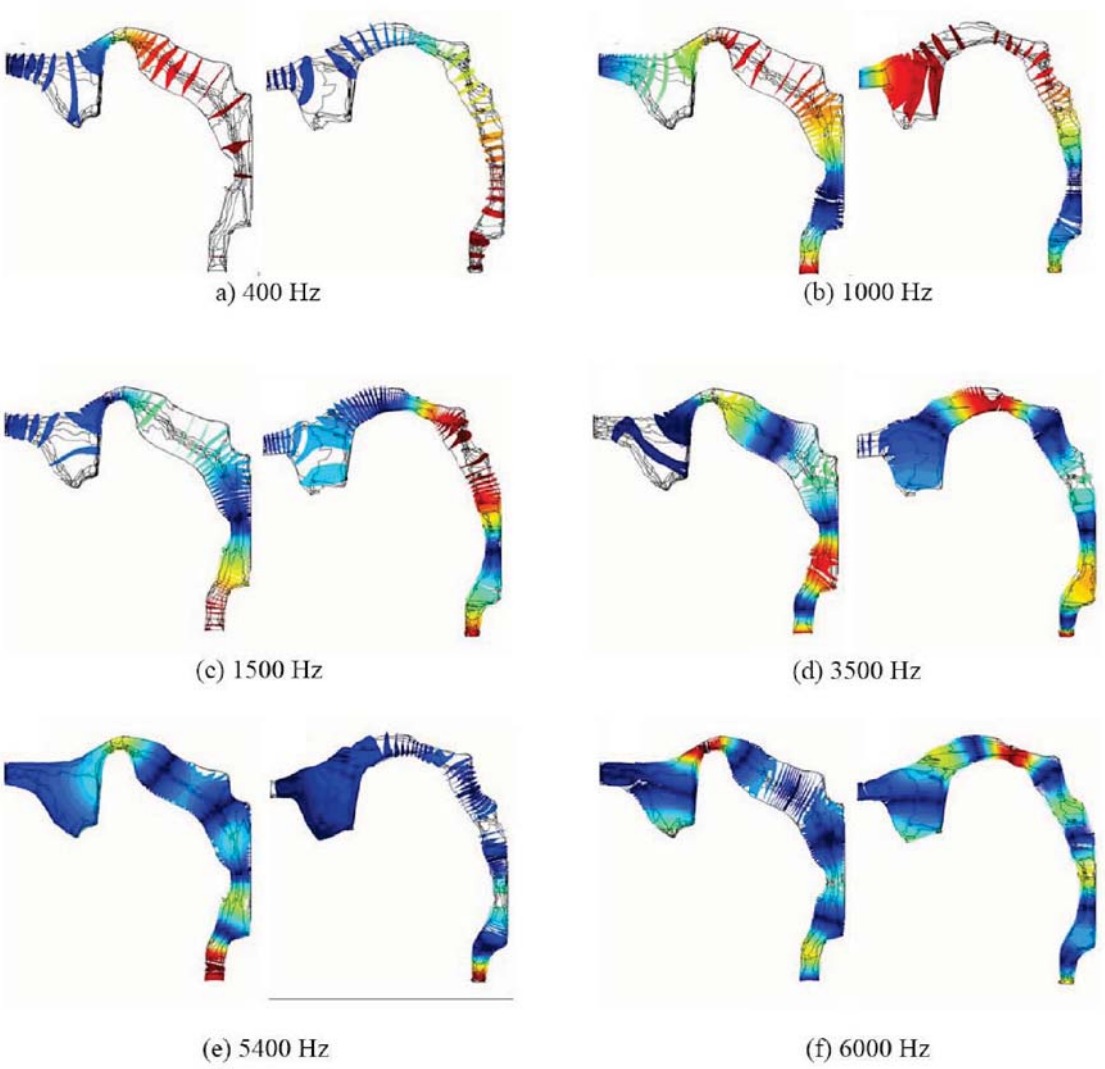


Figure 4.3: Pressure isosurface plots of wave propagation inside the vocal tracts of the retroflex /r/ (S1 on the right side) and the bunched /r/ (S2 on the right side) at different frequencies. (Pressure isosurfaces are coded by color: the red color stands for high amplitude and the blue color stands for low amplitude.) (a) 400 Hz , (b) 1000 Hz , (c)1500 Hz , (d)3500 Hz , (e) 5400 Hz , (f) 6000 Hz.

shapes have similar wave propagation. For both, as expected, the wave propagation is almost planar up to about 1000 Hz. Between 1500 and 3500 Hz, a second wave propagates almost vertically to the bottom of the front cavity. Above 4500 Hz, the isosurface becomes more complex and part of the acoustic wave propagates to the two sides of the front cavity. The results show that the wave propagation property should be kept in mind when assuming planar wave propagation along the vocal tract, particularly for antiresonances. Note that for both subjects, F4 and F5 occur in the region below 4500 Hz.

The cutting orientations for the area functions based on the pressure isosurfaces are shown in the upper panel of Figure 4.4 as grid lines. The area functions themselves are shown in the lower panel of Figure 4.4. Spectra generated from 3-D FEM and area functions are shown in Figure 4.5c and d. Formant values generated are shown in Tables 4.3 and 4.4. Both comparisons show that the results from the two methods match within 5 percent with each other. Note, however, that although the FEM model produces zeros above 5000 Hz, they are not produced by the area function vocal tract model because it does not contain side branches and is based on only plane wave propagation.

4.5 Comparisons between vocal tract acoustic response and measured spectra

This section compares the results of calculations to acoustic spectra from actual productions by the subjects during (a) MR and (b) sound booth acoustic ses-

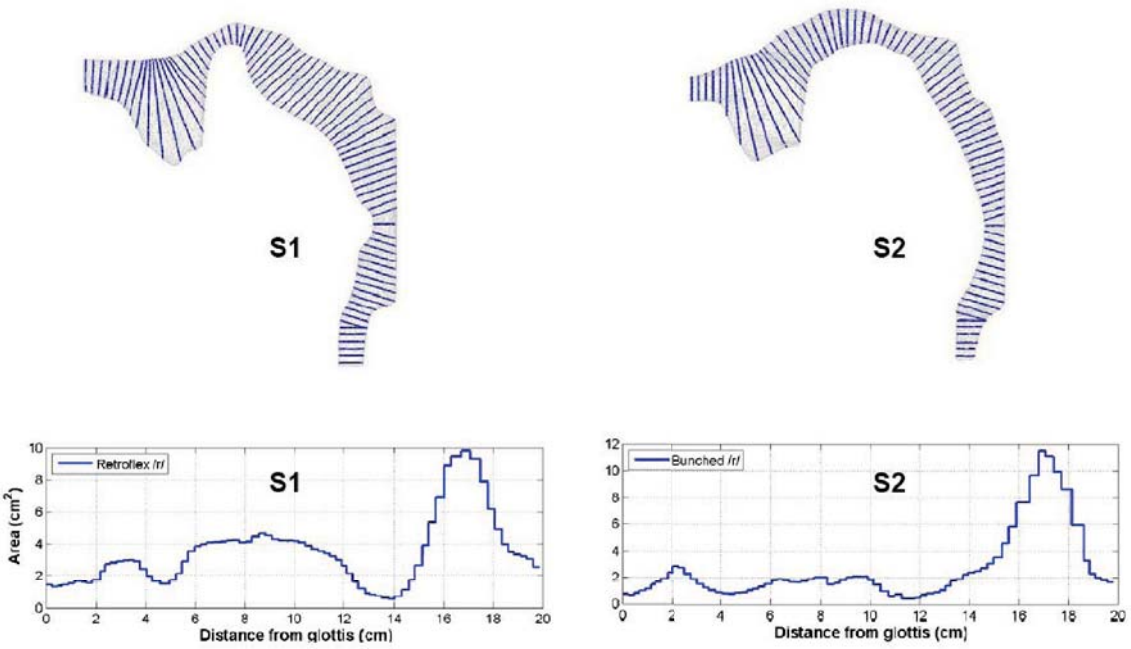


Figure 4.4: Top panel: Grid lines for area function extraction inside the vocal tract. Lower panel: Area function based on the grid lines. (In each panel, left side is for S1 and right side is for S2.)

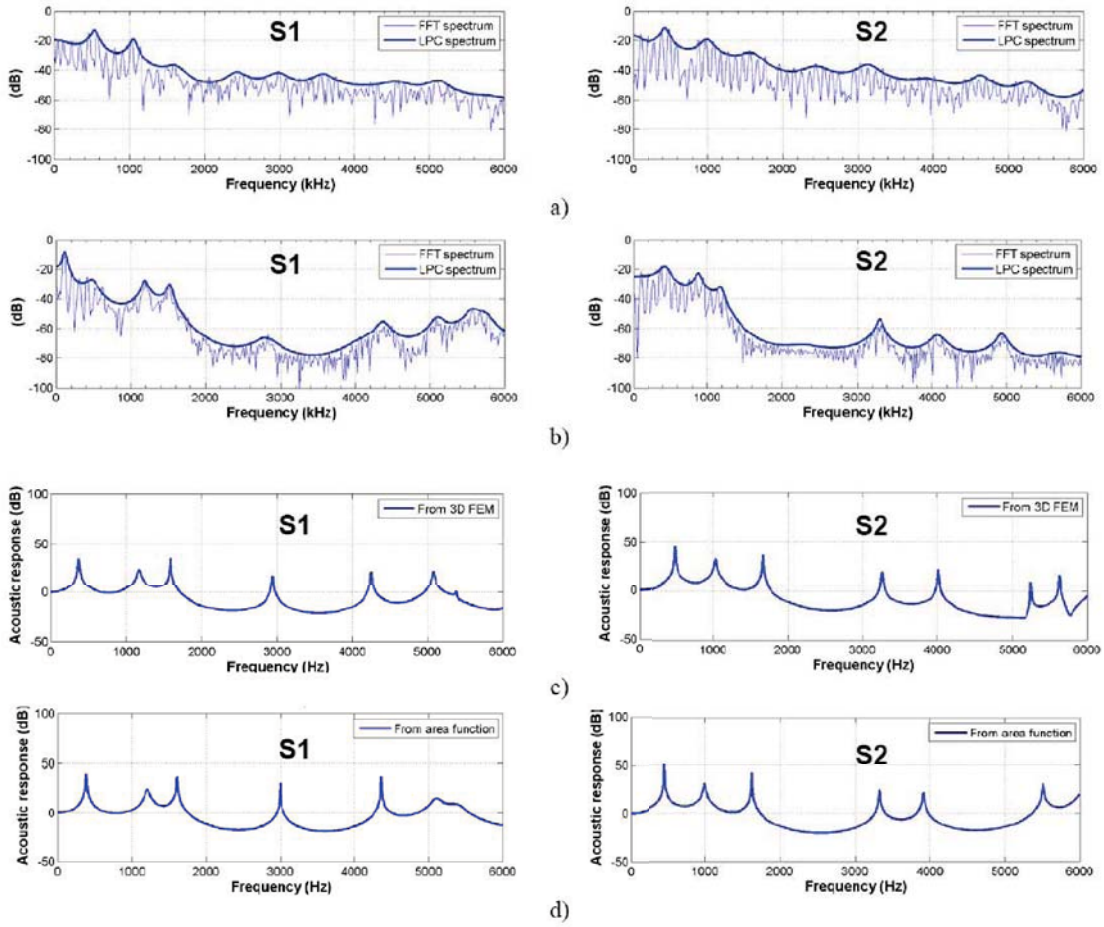


Figure 4.5: For S1 and S2: (a) spectrum of sustained /r/ utterance in MRI session, (b) spectrum of sustained /r/ utterance in the sound booth acoustic data, (c) the acoustic response based on 3-D FEM , (d) the acoustic response based on the area function.

Table 4.3: Formants measured from S1's retroflex /r/ compared with calculated values from the 3-D FEM, tube model with area function model, and simple-tube model, respectively (Unit: Hz). The percentage difference between the FEM formant values and the actual subject formant values from MR ($\Delta 1$) and sound acoustic ($\Delta 2$) sessions are also given. Note that due to background noise, only F1−F3 could be consistently measured from the MRI acoustic data.

	Retroflex /r/ (S1)							
	MRI acoustic data	Second both supine position	Sound booth upright position	3D FEM			Area function tube model	Simple tube model
				Formant	$\Delta 1$ (%)	$\Delta 2$ (%)		
F1	522	391	438	380	27.2	2.81	383	418
F2	1075	1234	1188	1160	7.91	6.0	1209	1262
F3	1534	1547	1563	1580	3.0	2.13	1609	1660
F4		2797	2828	2940		5.11	3002	2936
F5		4328	4234	4280		1.11	4366	4233
F5-F4		1531	1406	1340			1364	1297

sions, respectively. The calculated results include (c) acoustic response from the FEM analysis based on the 3-D model, (d) acoustic response from the VTAR computational model using FEM-derived area functions. The FEM analysis makes no assumptions regarding planar wave propagation, whereas, the area functions are derived from cutting planes determined by the FEM at low frequency.

4.5.1 MR vs. sound booth acoustic data

Because the FEM analysis and area functions are both based on MR data, the F4/F5 patterns would ideally have been extracted from the simultaneously recorded

Table 4.4: Formants measured from S2's retroflex /r/ compared with calculated values from the 3-D FEM, tube model with area function model, and simple-tube model, respectively (Unit: Hz). The percentage difference between the FEM formant values and the actual subject formant values from MR ($\Delta 1$) and sound acoustic ($\Delta 2$) sessions are also given. Note that due to background noise, only F1−F3 could be consistently measured from the MRI acoustic data.

	Bunched /r/ (S2)							
	MRI acoustic data	Second both supine position	Sound booth upright position	3D FEM			Area function tube model	Simple tube model
				Formant	$\Delta 1$ (%)	$\Delta 2$ (%)		
F1	445	453	391	480	7.87	5.96	457	472
F2	1008	906	891	1040	3.17	14.79	998	1047
F3	1469	1203	1219	1660	13.0	37.99	1626	1680
F4		3313	3281	3260		1.60	3330	3190
F5		4109	4016	4000		2.65	3912	3841
F5-F4		796	735	740			582	651

acoustic signal ("MR acoustic data"). As noted previously, however, F4 and F5 are masked in the MRI condition by the noise of the scanner. Hence, acoustic data recorded in a sound booth (from the supine posture) were used for comparisons with the calculated acoustic response results. Comparison between the MR and sound booth acoustic data for the first three formants show that subjects' productions are, for the most part, highly similar as shown in Tables 4.3 and 4.4. There are notable deviations in the F1 and F2 produced by S1 and in the F3 produced by S2. While these differences probably indicate a slight difference in articulatory configuration for sustained /r/, this same alternation between formant values can also be seen in their running speech for both real and nonsense words . In all cases, the characteristic F4/F5 pattern is maintained. The difference in F4/F5 pattern between the retroflex configuration of S1 and the bunched configuration of S2 is also observed when subjects produce /r/ in the upright posture. This is shown for running speech in Figure 4.1. In addition, the formant values from sound booth acoustic sustained productions recorded in upright posture are reported in Tables 4.3 and 4.4, for comparison to the values recorded in supine posture.

4.5.2 Comparison of actual formants to acoustic response from FEM and area function

In Figure 4.5, spectra from subjects' actual productions are shown along with acoustic responses from the models for S1 and S2. As shown in Figure 4.5a and c (in addition to Tables 4.3 and 4.4), the FEM method provides formant values for F1,

F2, and F3 similar to those measured from actual productions in MRI sessions by each subject. The percentage differences (between modeled and measured acoustics) are also given in Tables 4.3 and 4.4. As Figure 4.5b, and Tables 4.3 and 4.4 also show, the spacing between F4 and F5 in the sound booth data for actual speaker production is much larger for the retroflex /r/ than for the bunched /r/ (a difference of 1531 Hz vs. 796 Hz for the supine position, and 1469 Hz vs. 651 Hz for the upright position). Notably, the FEM method also replicates this pattern of different spacing between F4 and F5. A similar difference in spacing is also predicted by the VTAR computer model using the extracted area functions (see Tables 4.3 and 4.4). Thus these results support our methods for deriving a 3-D model. They also suggest that the source of the differences in the F4/F5 pattern between the bunched and retroflex /r/ follows from their respective differences in overall tongue shape.

4.6 Analysis based on vocal tract area function models

To gain insight into formant-cavity affiliations, the area function models were used to obtain sensitivity functions for F1-F5. Additionally, the area function models were simplified to arrive at models consisting of 3 to 8 sections (as opposed to about 70 sections), in order to gain insight into the types of resonators from which the formants originate and the effects of area perturbations of these resonators. These will be referred to as simple-tube models.

4.6.1 Sensitivity functions of F1-F5

The definition of the formant sensitivity function is described in Section 3.3.5. The calculated sensitivity functions S_n for F1-F5 are shown in Figure 4.6 (the left panel is for S1 and the right panel is for S2). At a point where a curve for a given formant passes through zero, a perturbation in the cross-sectional area will cause no shift in the formant frequency. Otherwise, the curve shows how the formant will change if the area is increased at that point. If S_n is positive at a certain point, increasing the area at that point will increase the value of the nth formant. If S_n is negative at a certain point, increasing the area at that point will decrease the value of the nth formant. The number of such zero crossings on a curve is equal to 2N-1 (1, 3, 5 , 7, and 9 for F1-F5 respectively) as stated by Mrayati et al. (1988), where N is the formant number for that curve.

As shown in Figure 4.6, the sensitivity functions for F1, F2 and F3 have some similarities in their patterns for both the retroflex /r/ and the bunched /r/. In both cases, F2 is mainly affected by the front cavity where the lip constriction with small area and the large posterior volume between the lip constriction and the palatal constriction act as a Helmholtz resonator. The frequency of a Helmholtz resonator is given by

$$F_H = \frac{c}{2\pi}\sqrt{\frac{A_1}{l_1 A_2 l_2}} \tag{4.1}$$

where A_1 and l_1 are the area and length of the lip constriction and A_2 and l_2 are the area and length of the large volume behind the lip constriction. From this equation, F_H will increase if the area of the lip constriction increases, or if the area of the large

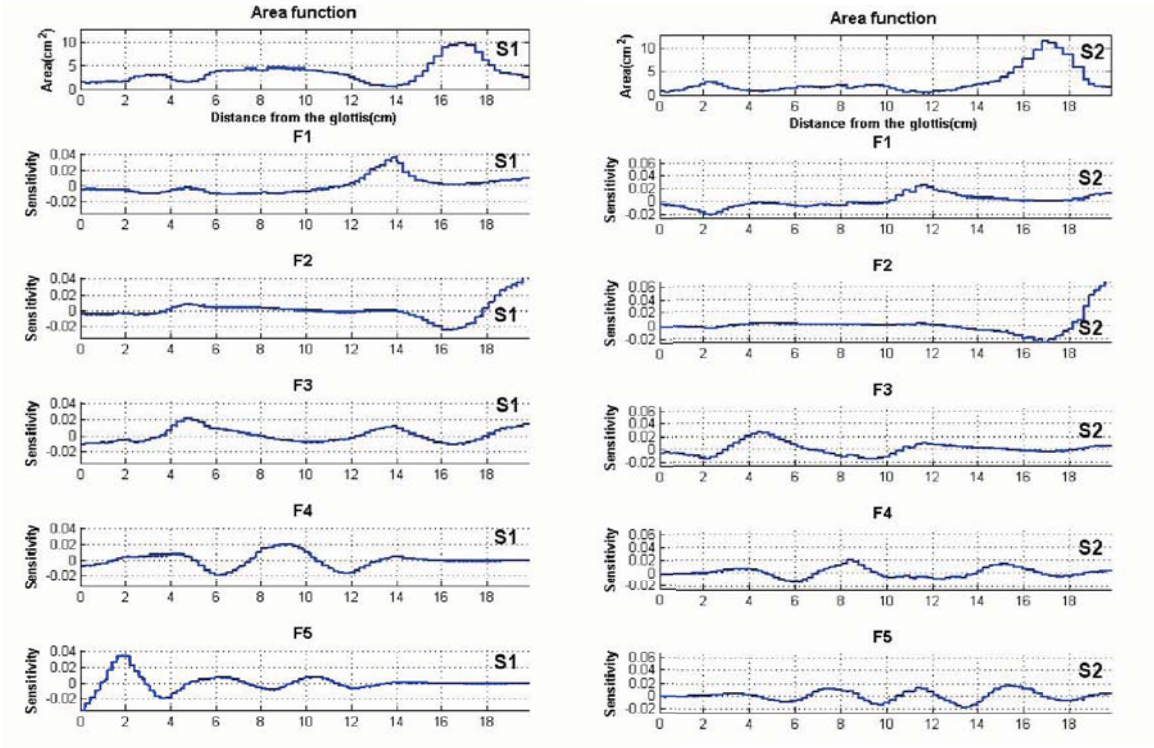


Figure 4.6: Acoustic sensitivity functions of F1-F5 for the retroflex /r/ of S1 and S2.

volume behind the lip constriction decreases. The sensitivity functions for F2 show this behavior since it is significantly positive during the portion of the tube that corresponds to the lip constriction and, conversely, significantly negative during the portion of the tube that corresponds with the large volume.

This conclusion is supported by the spectra in Figures 4.7 and 4.8. Figures 4.7 and 4.8 compare the spectra from the full vocal tract model with the spectra from the shortened vocal tract that includes only the front cavity as highlighted (acoustic responses were calculated with radiation at the lips) and the spectra from the shortened vocal tract that includes only the back cavity as highlighted (pressure on the front side is assumed to be zero). As can be seen, the first resonance of the front cavity is F2 from the full vocal tract for both subjects.

Based on the area function data of S1, Figure 4.9 shows how the F2/F3 cavity affiliations switch when the front cavity volume is changed by varying its length. When the front cavity volume exceeds about 17 cm^3, there is a switch in formant-cavity affiliation between F2 and F3. The front cavity resonance is so low that it becomes F2 and the resonance of the cavity posterior to the palatal constriction becomes F3. It seems that the front cavity resonance may be F2 or F3 depending upon the size of the volume of the Helmholtz resonator. This conclusion is supported by the findings from two different subjects showing bunched configurations discussed in Espy-Wilson et al. (2000). In that study, F3 was clearly derived from the Helmholtz front cavity resonance. However, the subjects in that study had much smaller front cavity volumes (of 5 cm^3 and 8 cm^3) relative to those of the current subjects S1 and S2 (of 24 cm^3 and 27 cm^3) , respectively.

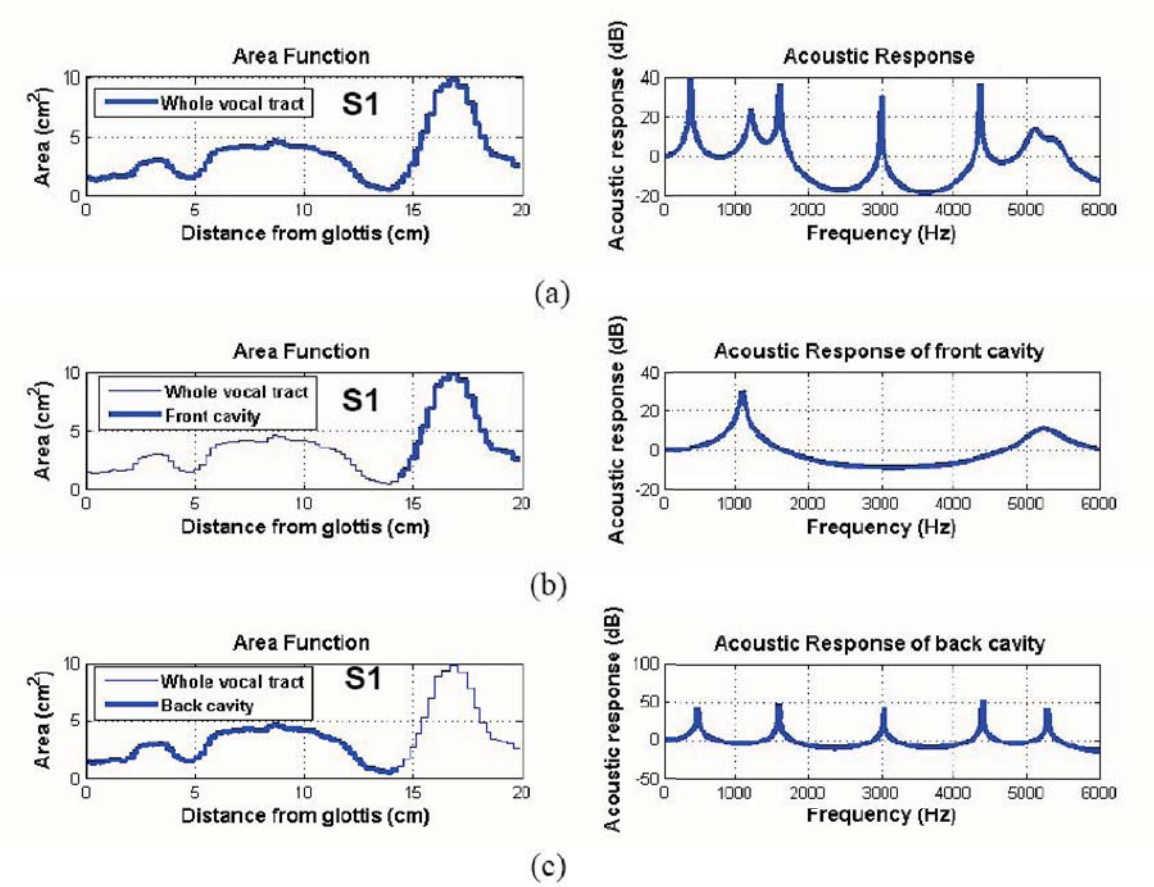


Figure 4.7: Acoustic response of S1's retroflex /r/ area function with front and back cavities separately modeled. (Left side is the area function and the right side is the corresponding acoustic response.) (a) area function of the whole vocal tract and its corresponding acoustic response, (b) area function of the front cavity and its corresponding acoustic response, (c) area function of the back cavity and its corresponding acoustic response.

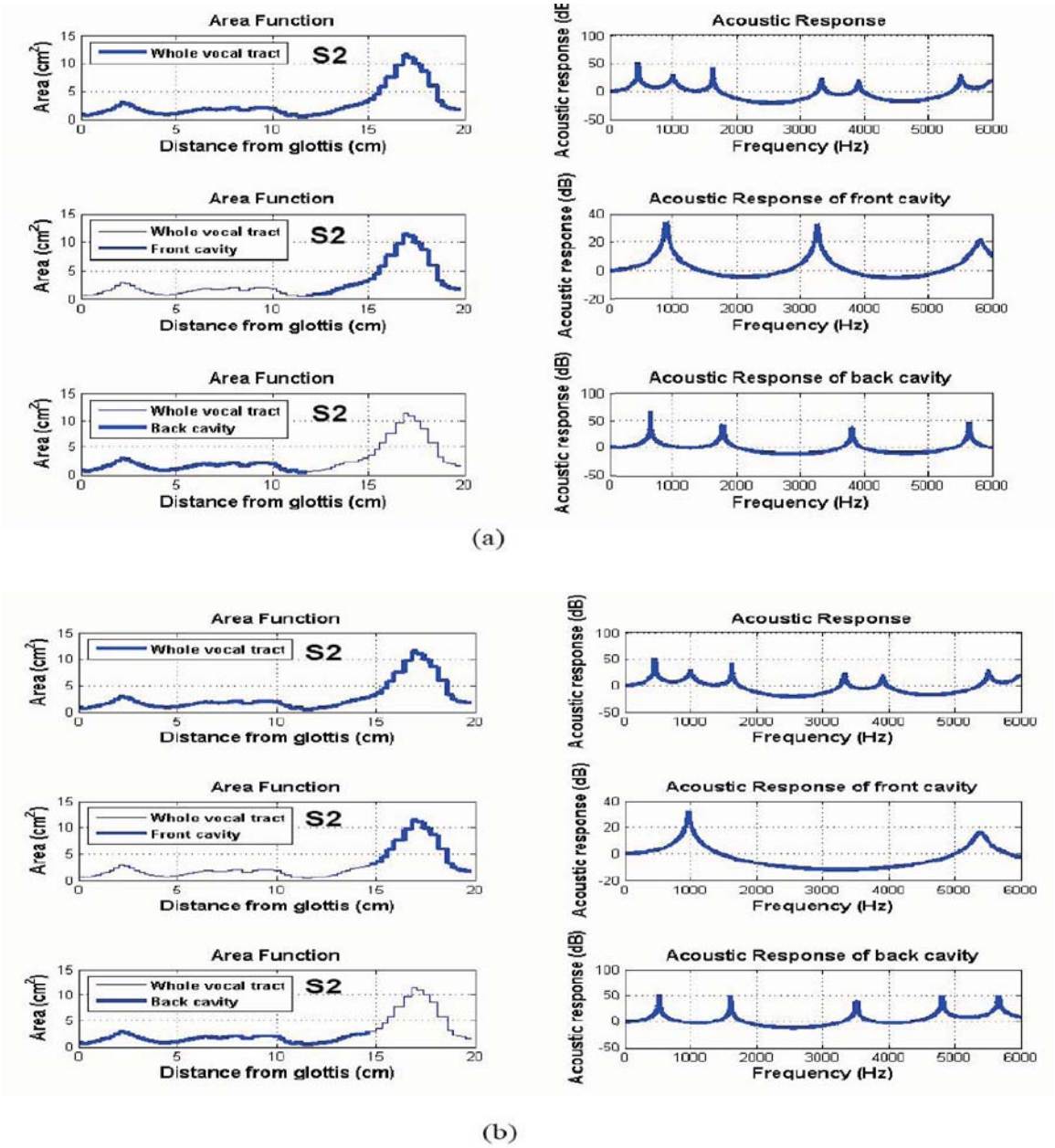


Figure 4.8: Acoustic response of S2's bunched /r/ area function with front and back cavities separately modeled. (Left side is the area function and the right side is the corresponding acoustic response.) (a) the dividing point between the front cavity and the back cavity at about 12 cm, (b) the dividing point between the front cavity and the back cavity at about 15 cm.

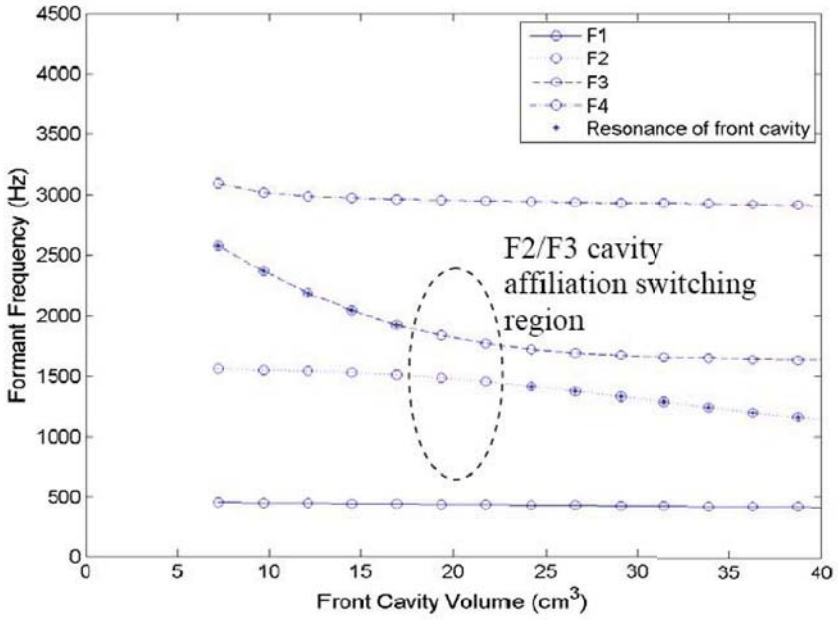


Figure 4.9: F2/F3 cavity affiliation switching with the change of the front cavity volume by varying its length (based on the area function data of S1).

Due to coupling between cavities along the vocal tract, F1 and F3 of both retroflex and bunched /r/ can be affected by area perturbation along much of the vocal tract. However, there are differences. The F1 sensitivity function for S1's retroflex /r/ shows a prominent peak in the region of the palatal constriction (between 12.6 cm and 14.6 cm), whereas the F1 sensitivity function for S2's bunched /r/ shows a prominent peak in the region of the palatal constriction (between 10.7 cm and 12.3 cm) and also a prominent dip in the region posterior to the pharyngeal constriction (between 1.6 cm and 2.8 cm). This difference in the F1 sensitivity functions of the retroflex and bunched /r/s is due to the differences in the area functions posterior to the front cavity. In the retroflex /r/, the areas of the palatal constriction are much smaller than the areas of the back cavity posterior to the

palatal constriction. This shape is more like a Helmholtz resonator for F1. In the bunched /r/, the overall shape of the area function posterior to the front cavity is similar to that of the retroflex /r/. However, the areas are more uniform so that F1 is the first resonance of a uniform tube (see discussion of simple tube modeling in Section 4.6.2).

As the sensitivity functions indicate, F3 can be decreased by narrowing at each of the three constriction locations along the vocal tract. Note, however, that in both of these cases, F3 is most sensitive to the perturbation of the pharyngeal constriction. It is relatively much less sensitive to the palatal constriction and even less to the lip constriction. This result confirms the finding of Delattre and Freeman (1968) that the percept of /r/ depends strongly on the existence of a constriction in the pharynx.

Sensitivity functions for F4 and F5 have very different patterns for the retroflex /r/ and the bunched /r/. In the retroflex /r/, F4 and F5 are affected only minimally by the area perturbation of the front cavity, starting at the location about 14.8 cm from the glottis, which means that they are resonances of the cavities posterior to the palatal constriction. This conclusion is supported by the spectra in Figure 4.7 which shows that the first four resonances of that part of the vocal tract behind the palatal constriction are close to F1, F3, F4 and F5. In the bunched /r/, F4 and F5 are not sensitive to the area perturbation of the cavity posterior to the pharyngeal constriction and they are affected to some extent by the front cavity. Again, this sensitivity to the front cavity is probably due to a higher degree of coupling between the back and front cavities for the bunched /r/ relative to the retroflex /r/. Given

the more gradual transition between the back and front parts of the vocal tract for the bunched /r/, Figure 4.8 shows two possible divisions. In one case, the front cavity is assumed to start at 11.8 cm from the glottis. In the other case, it starts 2.9 cm further forward, at 14.7 cm from the glottis. In both cases, the first resonance (a Helmholtz resonance formed by the lip constriction and the large volume behind it) of the front cavity is around 1000 Hz, the frequency of F2 in the spectrum derived from the full vocal tract. However, this choice of a division point has a significant effect on the location of the second resonance (a half-wavelength resonance of the large volume between the lip constriction and the palatal constriction) from the front cavity. If the front cavity starts at 11.8 cm, the second resonance is around 3300 Hz, the region of F4 from the full vocal tract spectrum. If the front cavity starts around 14.7 cm, the second resonance of the front cavity is around 5500 Hz, which corresponds to the region around F6 in the spectrum derived from the full vocal tract.

4.6.2 Simple-tube modeling

Figure 4.10 shows simple-tube models for the retroflex and bunched /r/s along with the original area functions and the corresponding acoustic responses. In the first case of the retroflex /r/, as shown in Figure 4.10a, the simple model consists of four tubes: a lip constriction, a large volume behind the lip constriction, a palatal constriction and a long tube posterior to the palatal constriction (see Figure 4.10a). Henceforth, the area forward of the palatal constriction will be referred to as the

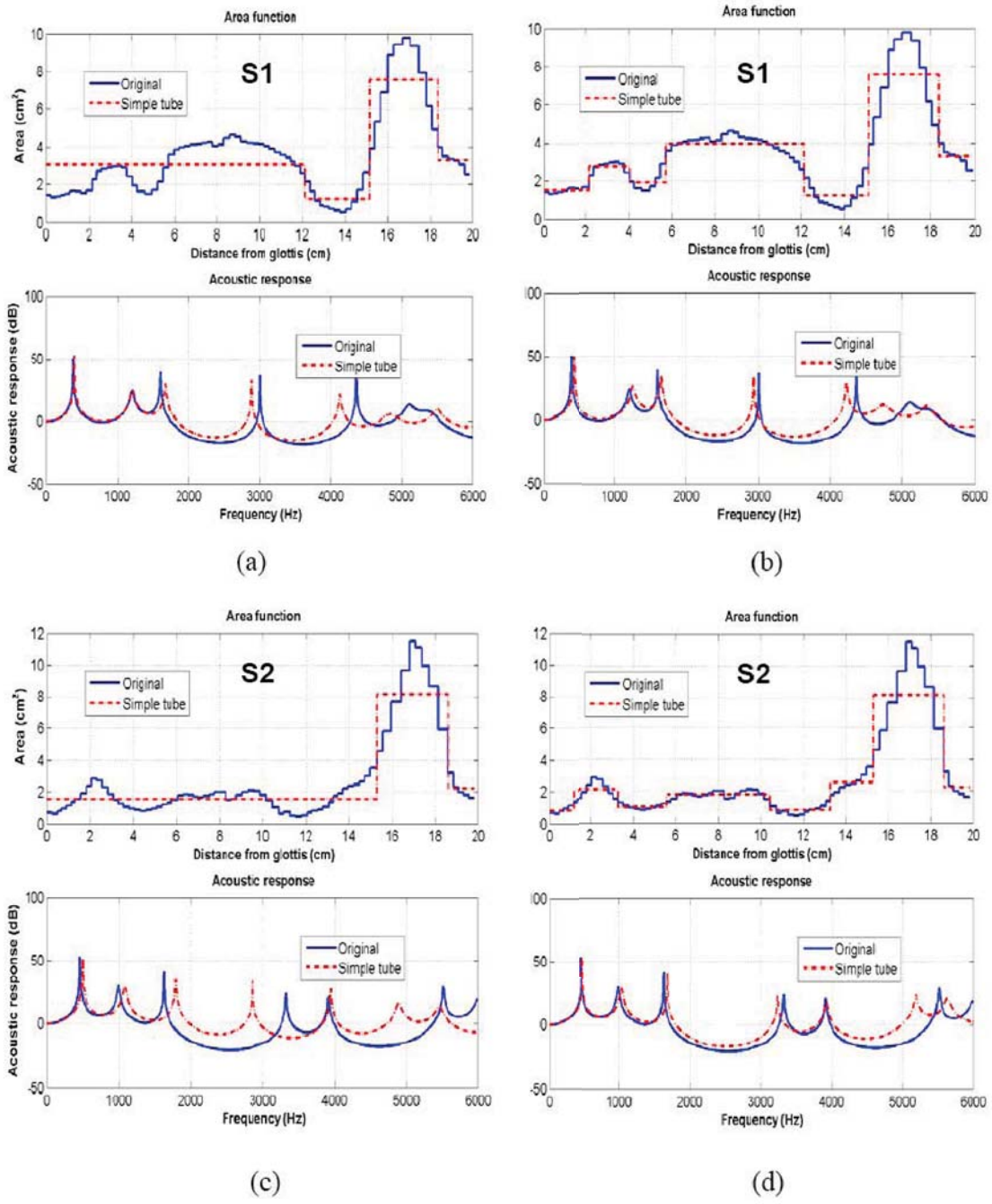


Figure 4.10: Simple-tube models overlaid on FEM-derived area functions at top panel, and corresponding acoustic responses at bottom panel. (a) four element simple-tube model of the retroflex /r/ of S1, (b) Seven element simple-tube model of the retroflex /r/ of S1, (c) three element simple-tube model of the bunched /r/ of S2, (d) eight element simple-tube model of the bunched /r/ of S2.

front cavity, while the area from the palatal constriction backward to the glottis will be referred to as the long back cavity. As we saw from the sensitivity functions, F2 comes from the front cavity, acting like a Helmholtz resonator at low frequencies. F1 comes from the long back cavity plus the palatal constriction, which together act as a Helmholtz resonator at low frequencies. F3, F4 and F5 are half-wavelength resonances of the long back cavity. The fact that the three formants are fairly evenly spaced (see Figure 4.10a and b) is thus explained. Refinement of the simple tube, by allowing additional discrete sections as in Figure 4.10b, indicates that if we include the pharyngeal narrowing in our model, F3 is further lowered in frequency. In addition, if we include the narrowing in the laryngeal region above the glottis, F4 and F5 rise in frequency. The net results from these perturbations can be seen in Figure 4.10b. These formant-cavity affiliations agree well with our understanding from the sensitivity functions. Further, Tables 4.3 and 4.4 show that there is close agreement between the formant frequencies measured from the actual acoustic data, and those predicted both by the FEM-derived area functions and the simple-tube model. In the case of the bunched /r/, the long back cavity has a wide constriction in the pharynx and is more uniform overall, so that we model it initially as a quarter-wavelength tube (see Figure 4.10c). If we then account for the pharyngeal narrowing, F3 is lowered and F5 is raised. If we include the palatal constriction itself, F4 is raised and F5 is lowered. Finally, including the laryngeal narrowing in the model raises F4 and (to a lesser extent) F5. The net results of these manipulations are shown in Figure 4.10d. Again, Tables 4.3 and 4.4 show that there is close agreement between the formant frequencies predicted both by the FEM-derived area functions

(a)

(b)

Figure 4.11: Midsagittal MR images of the vocal tracts for retroflex and bunched shapes (a subset of the UC database (Tiede et al., 2004)). (a) retroflex /r/s (Left: S1, Middle: S3, Right: S4), (b) bunched /r/s (Left: S2, Middle: S5, Right: S6).

and the simple-tube model and measured from the actual acoustic data.

4.7 Formants in acousitc data of sustained /r/ and nonsense word "warav"

As a partial confirmation of the hypothesis that the F4/F5 pattern shown by S1 and S2 is a function of their retroflex and bunched tongue shapes, four extra subjects's acoustic data were studied. The midsagittal MR images of the four subjects (S3, S4, S5, S6) are displayed along with S1 and S2 in Figure 4.11. S3 and S4 have retroflex /r/ tongue shapes similar to S1, and S5 and S6 have bunched /r/ tongue shapes similar to S2. The averaged spectra (from a 300 ms segment of sound booth acoustic recordings) of the sustained /r/ sounds produced by the six subjects in the

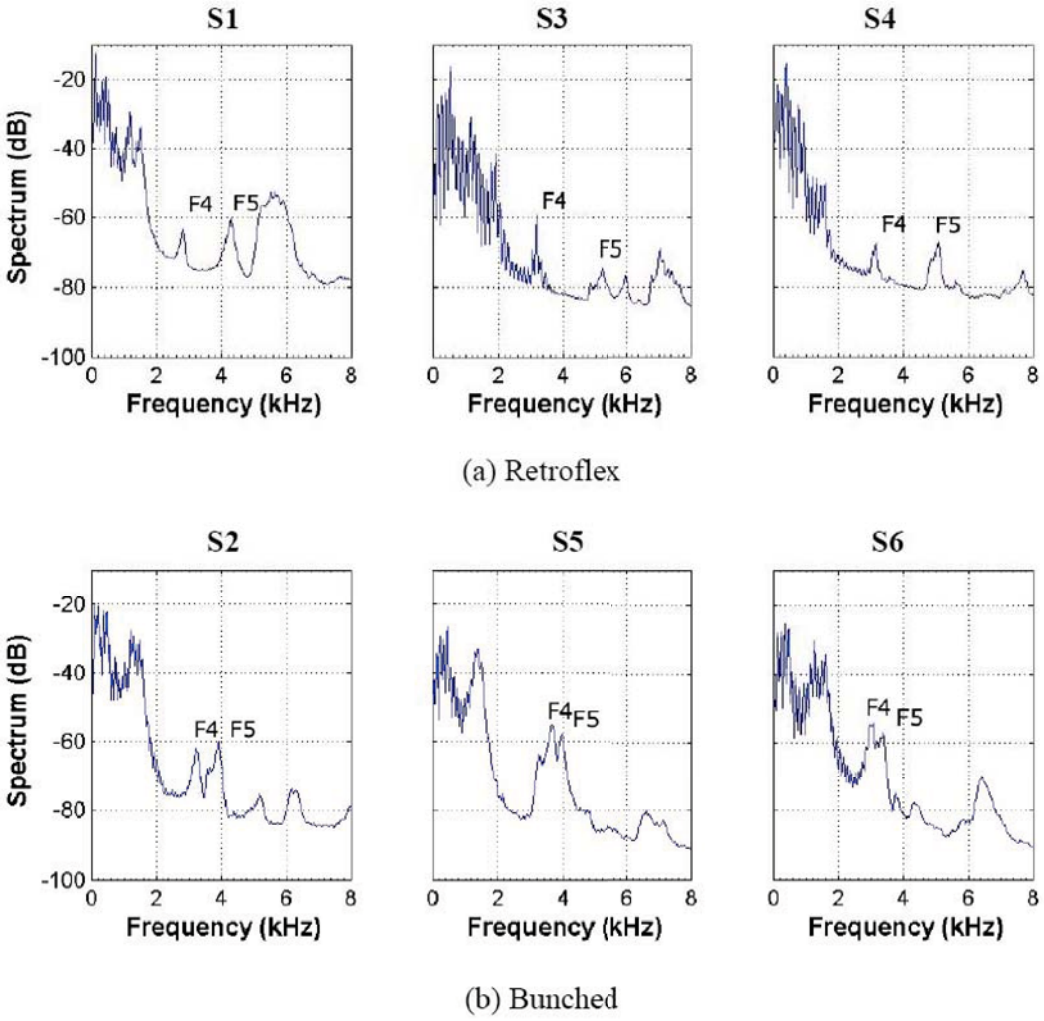


Figure 4.12: Spectra of sustained /r/ utterances from 6 speakers (3 retroflex /r/s and 3 bunched /r/s). (a) retroflex /r/s (Left: S1, Middle: S3, Right: S4), (b) bunched /r/s (Left: S2, Middle: S5, Right: S6).

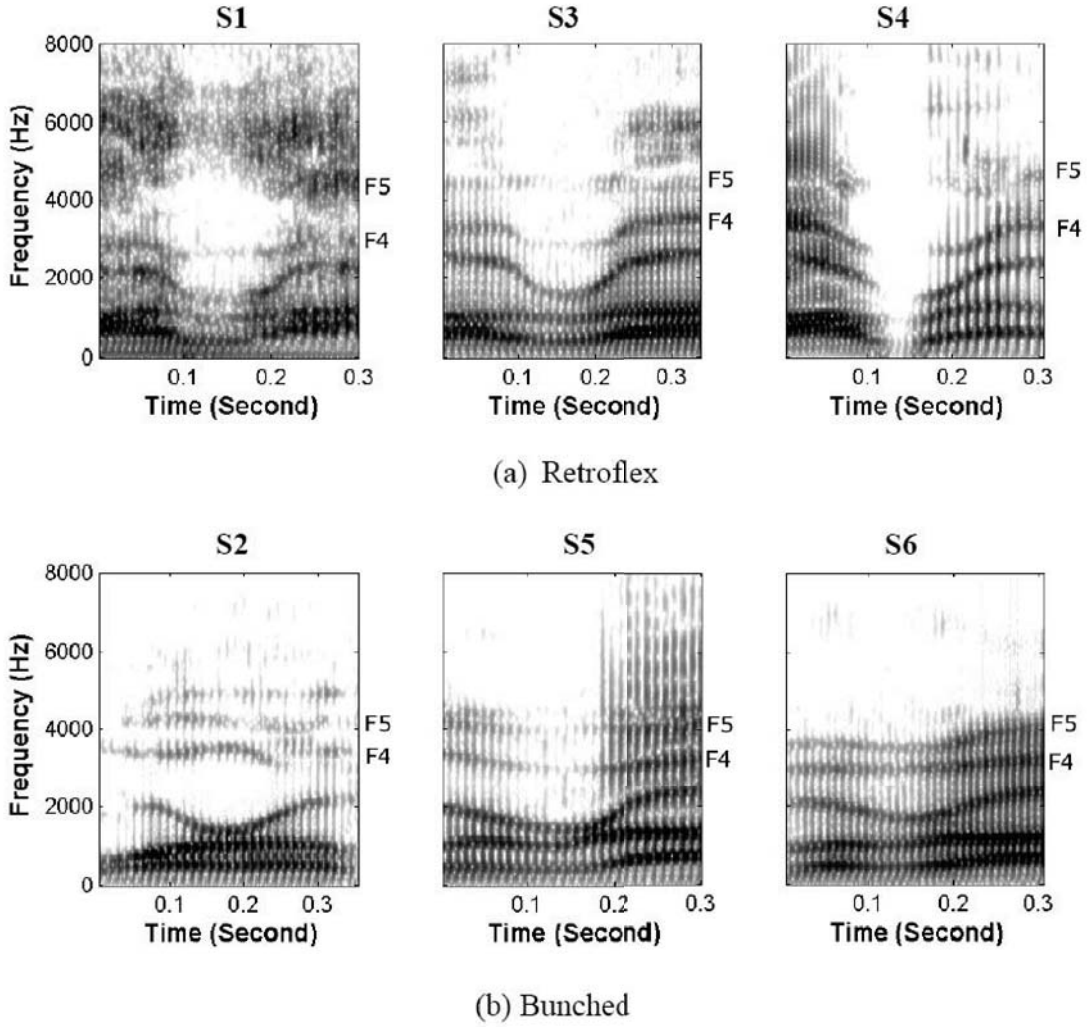


Figure 4.13: Spectrograms for nonsense word "warav" from 6 speakers (3 retroflex /r/s and 3 bunched /r/s, only portion of spectrograms are shown in the figure with /r/ in the middle). (a) retroflex /r/s (Left: S1, Middle: S3, Right: S4), (b) bunched /r/s (Left: S2, Middle: S5, Right: S6.

upright position are shown in Figure 4.12. As can be seen, the retroflex /r/s have a larger difference in F4 and F5 than the bunched /r/s. The difference between F4 and F5 for S3 and S4 is about 1900 Hz and 2000 Hz, respectively, while the difference between F4 and F5 for S5 and S6 is about 500 Hz and 600 Hz, respectively. These results are consistent with the results obtained from S1 and S2 in that the spacing between F4 and F5 is larger for the retroflex /r/ than for the bunched /r/.

In addition, the formant trajectories of the nonsense word "warav" for all the six subjects are shown in Figure 4.13 (note that the spectrograms of Figure 4.1 are repeated here for comparison). The difference between F4 and F5 of /r/ at the lowest point of F3 for S1, S3, and S4 is about 2100 Hz, 1500 Hz, and 1600 Hz, respectively, while the differences between F4 and F5 of /r/ at the lowest point of F3 for S2, S5, and S6 are about 700 Hz, 900 Hz, and 600 Hz, respectively. These results indicate that, for these subjects, the difference between F4 and F5 for the retroflex /r/ in dynamic speech is relatively larger than that in the bunched /r/ and provides additional support for the simulation result from the 3-D FEM and computer vocal tract models based on the area functions.

4.8 Discussion

In this chapter, the relationship between acoustic patterns in F4 and F5 and articulatory differences in tongue shape between subjects has been investigated. The primary data come from S1 and S2, who produce sharply different bunched and retroflex variants of /r/ associated with different patterns of F4 and F5. S1

and S2 are particularly comparable because they resemble each other in terms of vocal tract length and oral tract dimensions. The results suggest that bunched and retroflex tongue shapes differ in the frequency spacing between F4 and F5. Further, the F4/F5 patterns produced by S1 and S2 can be derived from a very simple aspect of the difference between the two vocal tract shapes. For both S1's retroflex /r/ and S2's bunched /r/, F4 and F5 (along with F3) come from the long back cavity. However, for S1, these formants are half wavelength resonances, while for S2, these formants are quarter wavelength resonances of the cavity. Additionally, the finding of an F4/F5 difference in pattern is replicated in the acoustic data from an additional set of four subjects, two with bunched and two with retroflex tongue shapes for /r/. These results suggest that acoustic cues based on F4-F5 spacing may be robust and reliable indicators of tongue shape, at least for the classic (tongue tip down) bunched and (tongue dorsum down) retroflex shapes discussed here.

It appears that this spacing between F4 and F5 is due to the difference in the long back cavity dimension/shape. In the case of the retroflex /r/, there is one long back cavity posterior to the palatal constriction. Our simple-tube modeling and the sensitivity functions show that F4 and F5 are half-wavelength resonances of the back cavity. In fact, F4 and F5 are the second and third resonances of the back cavity (F3 is the first resonance of this cavity). For S1, this half-wavelength cavity is about 12 cm long which gives a spacing between the resonances of about 1460 Hz. The narrowing in the laryngeal regions shifts F4 and F5 upwards by different amounts so that the spacing changes to about 1300 Hz. This spacing agrees well with the 1469-1531 Hz measured from S1's sustained /r/. For the bunched /r/,

the back cavity can be modeled as a quarter-wavelength tube. Our simple-tube modeling shows that F4 and F5 are the third and fourth resonances of this cavity. The sensitivity functions, on the other hand, show that F4 and F5 are influenced by the front cavity. This is probably due to the higher degree of coupling between the front and back cavities for the bunched /r/ of S2. The length of the back cavity for S2 is about 15 cm. Thus, the spacing between F4 and F5 for the bunched /r/ should be about 1150 Hz. However, the narrowing in the laryngeal, pharyngeal and palatal regions decreases this difference to about 650 Hz as seen in Figure 4.10(d). This formant difference agrees well with the value of 651-796 Hz measured from S2's sustained /r/. As a point of interest, the spacing between F4 and F5 in the spectrograms of Figure 4.13 is generally greater across all of the consonants and vowels for the speakers who produce the retroflex tongue shape for /r/ than it is in the spectrograms for the speakers who produce the bunched tongue shape for /r/. However, the difference does appear to be considerably enhanced during the /r/ sounds with the lowering of F4 and the slight rising of F5 during the retroflex /r/s, and the rising of F4 for S2 during the bunched /r/.

The relation of tongue shapes for /r/ to specific acoustic properties as found in this study may be useful for development of speech technologies such as speaker and speech recognition. For example, knowledge-based approaches to speech recognition rely heavily on acoustic information to infer articulatory behavior (Hasegawa-Johnson et al., 2005; Juneja and Espy-Wilson, 2008; Kinga et al., 2006). In addition, speakers appear to use tongue shapes in very consistent ways (Guenther et al., 1999). Thus, the use of a particular tongue shape for /r/ may produce acoustic character-

istics that are indicative of a speaker's identity, even if these characteristics are not relevant to the phonetic content.

4.9 Chapter summary

In this chapter, two subjects whose productions of "retroflex" /r/ and "bunched" /r/ show similar patterns of F1-F3 but very different spacing between F4 and F5 are contrasted. Using finite element analysis and area functions based on magnetic resonance images (MRI) of the vocal tract for sustained productions, the results of computer vocal tract models are compared to actual speech recordings. In particular, formant cavity affiliations are explored using formant sensitivity functions and vocal tract simple-tube models. The difference in F4/F5 pattern between the subjects is confirmed for several additional subjects with "retroflex" and "bunched" vocal tract configurations. Results suggest that the F4/F5 differences between the variants can be explained largely by differences in whether the long cavity behind the palatal constriction acts as a half- or a quarter-wavelength resonator.

Chapter 5

Acoustic modeling of lateral /l/

5.1 Introduction

The production of /l/ generally involves a linguo-alveolar contact and one or two lateral channels along the parasagittal sides of the tongue blade. This is shown in the midsagittal profile of Figure 5.1. The effect of these geometric features on the acoustics of the vocal tract are not clearly understood. Particulary, /l/'s spectrum has relatively weak energy in the F3-F5 region, as shown in the spectrograms of Figure 5.2. It has been proposed that this weak energy in F3-F5 region was due to the pole-zero clusters produced by the lateral channels and/or the supralingual space (Fant, 1970; Prahler, 1998; Stevens, 1998; Zhang and Espy-Wilson, 2004), and the complexity of /l/ spectrum was caused by the variability of zero's frequency. However these explanations were generally based on an assumed area function vocal tract model, not based on the acoustic analysis of a 3-D vocal tract geometry. A 3-D acoustic study of the vocal tract of /l/ may provide extra insights on the /l/ production, and it may also give guidance on how to build an area function vocal tract model of /l/.

This chapter presents a 3-D vocal tract acoustic study on two tongue shapes of /l/ production in American English. One produced a sustained dark /l/, and the other produced a sustained light /l/. Both tongue shapes are produced by the

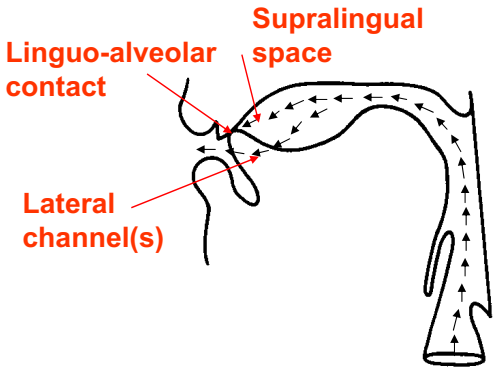
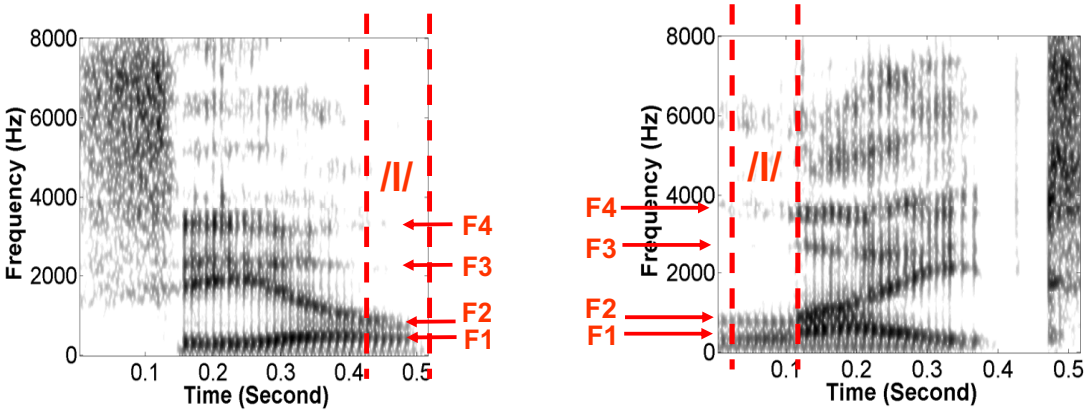


Figure 5.1: Midsagittal profile of the vocal tract producing /l/, adapted from Stevens (1998).



(a) "feel"

(b) "light"

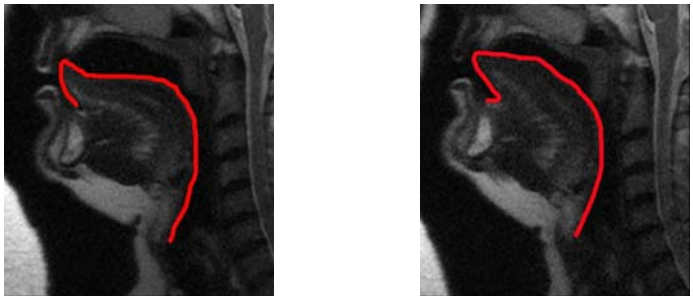
Figure 5.2: Spectrograms of word "feel" and word "light". (a) "feel" (dark /l/, syllable final), and (b) "light" (light /l/, syllable initial).

same subject. Based on reconstructed 3-D vocal tract geometries, the effects of the lateral channel(s) and the linguo-alveolar contact on the vocal tract acoustics were studied. The zero sources in the /l/ spectrum were the main focus in this study. The format cavity affiliations were not studied here, because there is no sharp area transition in /l/ production to acoustically decouple the vocal tract.

The task of understanding the zero sources in /l/ production has been approached in the following way. First, magnetic resonance images were used to acquire a detailed 3-D geometric reconstruction of the vocal tract. Second, finite element analysis has been performed to simulate the acoustic response of the 3-D vocal tract. The wave propagation property at different frequencies has been studied to understand the zero sources in the acoustic response. Third, area function models were obtained from the FEM analysis of 3-D geometry and the resulting acoustic response was verified against the 3-D acoustic response. Fourth, two simple 3-D vocal tract models are studied to gain additional insights on the acoustic effects of the lateral channels and the linguo-alveolar contact.

5.2 Subject

S2 in the UC database was selected for this /l/ production study. He produced both a sustained dark /l/ and a sustained light /l/ with MR images acquired. But the tongues shapes for these two /l/s are different, which is shown in the midsagittal MR images of Figure 5.3. The dark /l/ was produced as /l/ in "pole", and the light /l/ was produced as /l/ in "lee". The advantage of using one subject for



(a) The dark /l/ (b) The light /l/

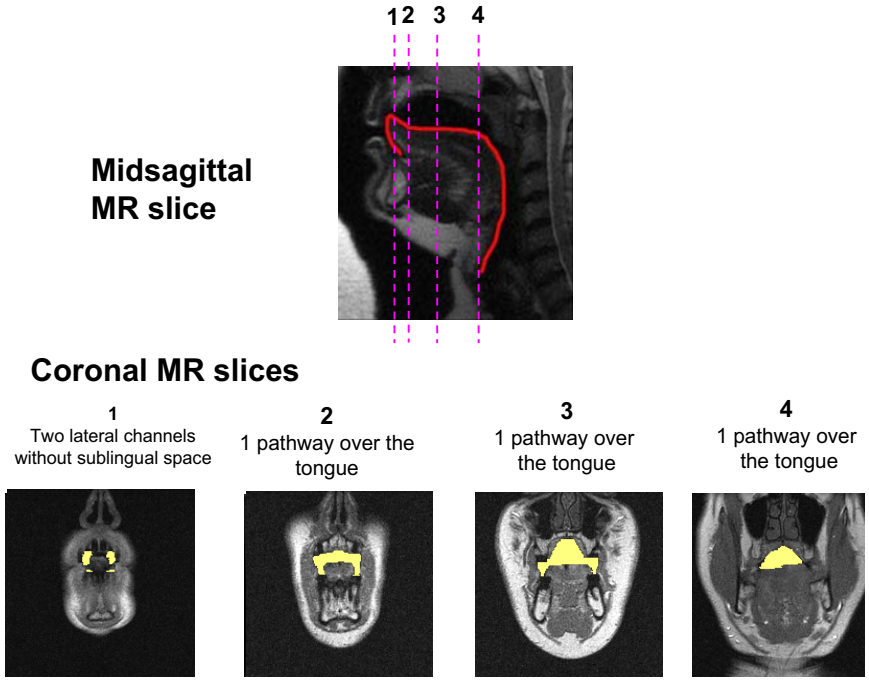
Figure 5.3: Midsagittal MR images of two tongue configurations of S2 for American English /l/. (a) the dark /l/, and (b) the light /l/.

both /l/s is that the vocal tract anatomy difference between two subjects can be avoided. Therefore, the acoustic effects caused by the tongues shape difference can be identified. Remember that both of the tongue shapes are just two examples for the /l/ production, and they are not exclusively for a light /l/ or a dark /l/'s production. As shown in Figure 5.3, the dark /l/'s linguo-alveolar contact is established with the tongue tip, and the light /l/'s is established with the tongue blade. The tongue dorsum is lowered for the dark /l/, whereas the tongue dorsum is raised for the light /l/. These articulation differences lead to the differences in the geometry of the vocal tract. The linguo-alveolar contact for the dark /l/ is relatively shorter than it is for the light /l/, so are the lateral channels around the contact. For the light /l/, due to its raised tongue dorsum, there are lateral linguopalatal contacts which separate the supralingual space as a side branch and also makes the lateral channels longer. In the midsagittal slices, the boundary of the tongue was manually drawn in red color for a better visualization of its shape.

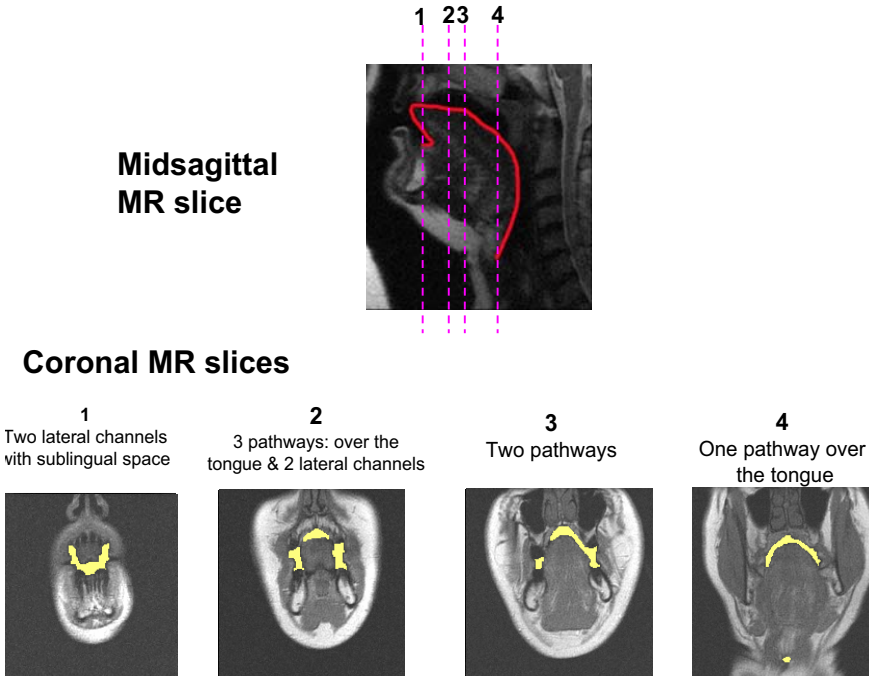
Figures 5.4a and 5.4b show the coronal MR slices at four different locations for both the dark /l/ and the light /l/ respectively.

For the dark /l/, due to its shorter linguo-alveolar contact, there is only one coronal MR slice with two lateral channels around the tongue, and this slice is located at position 1. The coronal MR slices at three other positions (2, 3 and 4) have only one pathway for the air flow. However, the missing teeth in the MR image make the cross-section areas look larger than their actual areas.

For the light /l/, the coronal slice at position 1 in Figure 5.4b shows a cross section of the two lateral channels around the tongue and also a sublingual space. Due to the lateral linguopalatal contacts, the coronal slice at position 2 has three pathways of air flow which include two lateral channels and one supralingual space. In contrast, the dark /l/ does not have lateral linguopalatal contacts. In the light /l/'s coronal slice at position 3, there is only one lateral linguopalatal contact on the left side of the image (or the right side of the subject), so the supralingual space is connected to one lateral channel on the right side. This means that the light /l/ has asymmetrical lateral linguopalatal contacts and, therefore, two asymmetrical lateral channels with different lengths. The coronal slice at position 4 shows a much smaller cross section area than the corresponding slice has for the dark /l/. This is caused by the raised tongue dorsum in the light /l/.



(a) The dark /l/



(b) The light /l/

Figure 5.4: Midsagittal and coronal MR images at different locations of S2. (The boundary of the tongue in the midsagittal slice is manually drawn for better visualization of its shape. The airways in the coronal slices are filled in yellow color) (a) the dark /l/, and (b) the light /l/.

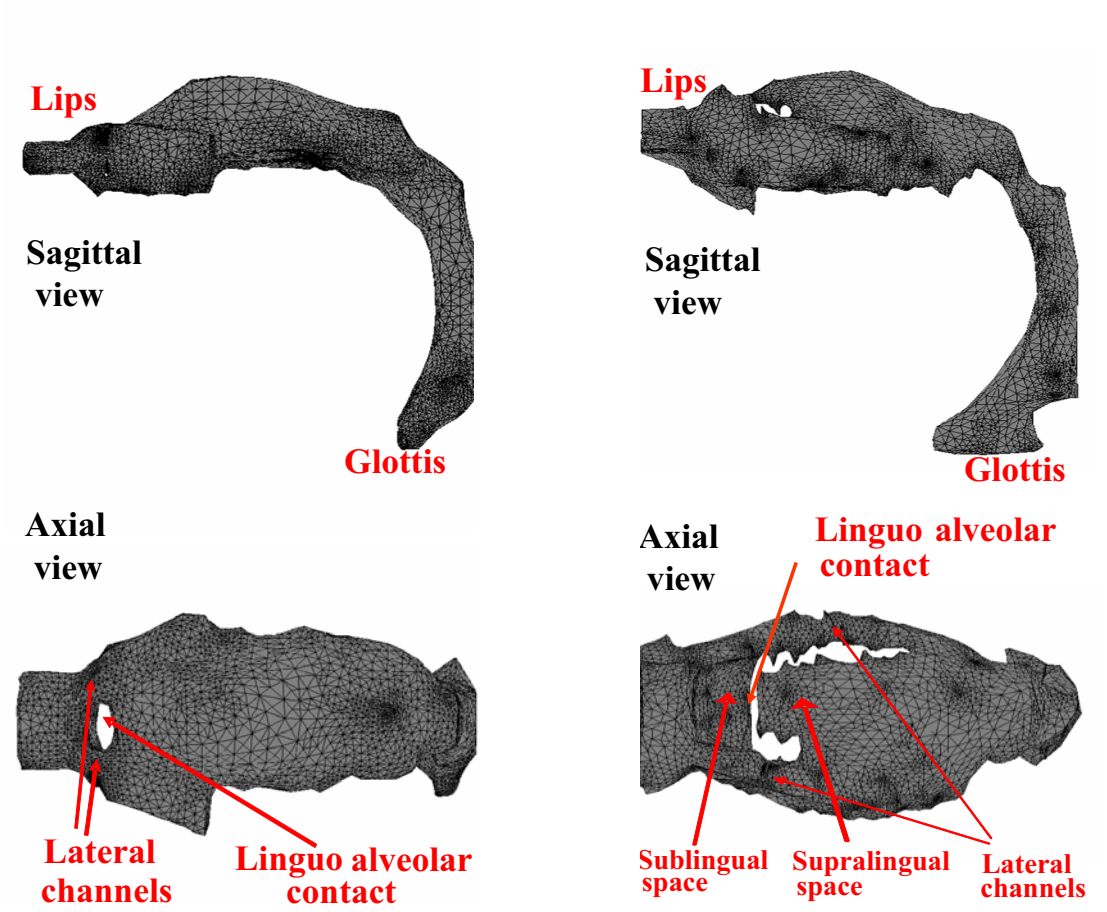
5.3 Reconstructed 3-D vocal tract geometries

Figure 5.5 shows the sagittal and axial views of the 3-D reconstructed geometries of the vocal tracts for the dark /l/ and the light /l/ respectively. In Figure 5.5a, the axial view of the dark /l/ shows a short linguo-alveolar contact which is about 0.8 cm long. Therefore, the lateral channels are short too. In Figure 5.5b, the linguo-alveolar contact for the light /l/ is about 1.7 cm long. This measurement is consistent with the result from Narayanan et al. (1997) where the length of the linguo-alveolar contact was found to be less than 2 cm. The axial view of the light /l/ shows two asymmetrical lateral channels (about 4.9 cm long on the right side vs. 2.1 cm long on the left side) and a separate supralingual space like a side branch. These two asymmetrical lateral channels are created by the combination of the linguo-alveolar contact and asymmetrical lateral linguopalatal contacts.

5.4 FEM-based acoustic analysis

5.4.1 Acoustic responses of 3-D FEM

Based on the reconstructed 3-D geometries in Figure 5.5, 3-D FEM analysis has been performed for the dark /l and the light /l/. Instead of the ideal piston radiation model, the pressure release boundary condition was applied at the lips in the FEM analysis. The reason for this is twofold. First, the pressure release avoids the radiation loss. Therefore, the pole/zero pair in the acoustic response is more prominent than it is in the case of a radiation model. This is particulary important



(a) The dark /l/

(b) The light /l/

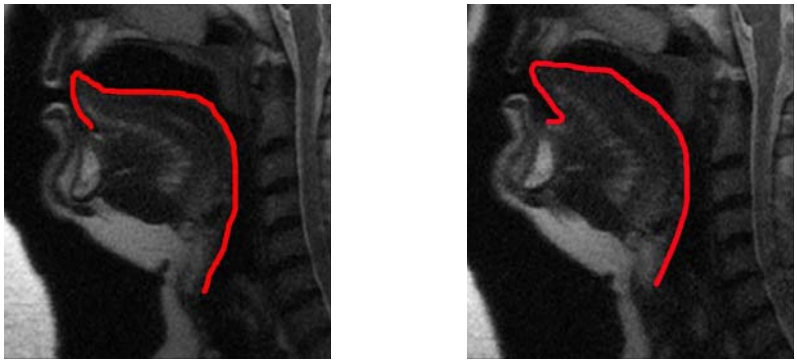
Figure 5.5: FEM meshes of the reconstructed 3-D vocal tracts of S2. (a) the dark /l/, and (b) the light /l/.

when the pole and the zero come close to each other. Second, both the boundary conditions give very similar acoustic responses in the 3-D FEM model, or in the area function vocal tract model which will be described in Section 5.5 (page 94). It is indicated that the choice of the boundary condition at the lips does not greatly affect the acoustic response of /l/.

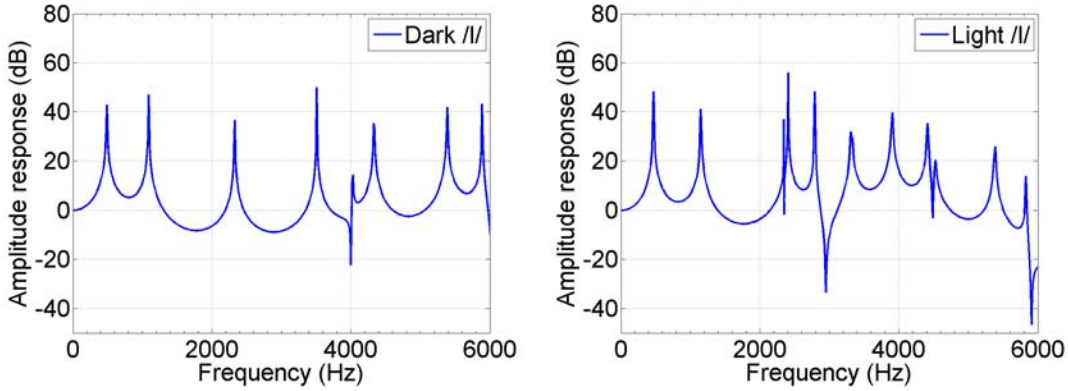
Figure 5.6 shows the midsagittal MR images, the acoustic responses of 3-D FEM, and the spectra of booth acoustic data for both the dark /l/ and the light /l/, respectively. Table 5.1 lists the measurements of F1-F3 from 3-D FEM and booth acoustic spectra. Zeros from 3-D FEM are also listed. However, there is no systematic method for detecting zeros in the spectra of the acoustic booth data and the values presented here are manually measured by locating the frequencies of deep valleys in the spectra.

It can be seen in Figure 5.6 and Table 5.1 that the acoustic responses of the dark /l/ and the light /l/ and the spectra derived from the booth acoustic data have very similar patterns in F1, F2 and F3. However, they have different zeros. In the 3-D FEM, the dark /l/ has a zero at 4000 Hz, whereas the light /l/ has zeros at 2350 Hz, 2950 Hz, and 4490 Hz. The zero at 2350 Hz in the light /l/ is hard to detect, because the pole-zero pair are very close to each other.

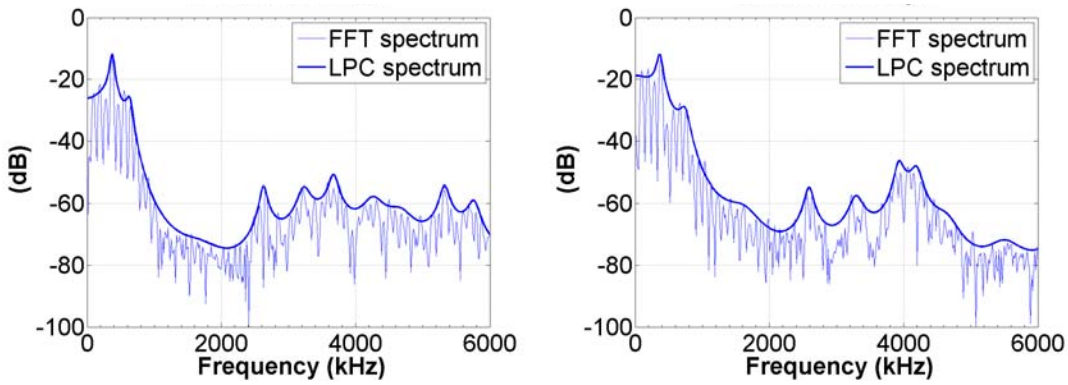
There are some discrepancies in F1-F3 between the acoustic response of the 3-D FEM and the spectra of the booth acoustic data. For example, the difference in F2 is more than 400 Hz, in both the dark /l/ and the light /l/. There are some reasons for this discrepancy, including the coarse MR image quality and subject articulation inconsistency in two different environments (MR room vs. acoustic



(a)



(b)



(c)

Figure 5.6: For S2 (left: the dark /l/; right: the light /l/): (a) midsagittal MR images, (b) acoustic responses based on 3-D FEM, and (c) spectra of sustained /l/ utterance in booth acoustic data (/l/ as in "pole" or /l/ as in "lee").

Table 5.1: Formants and zeros measured from S2's sustained /l/ utterance compared with calculated values from the 3-D FEM (Unit: Hz). (The zeros measured from spectra of acoustic booth data are denoted with symbol '*'. There is no systematic method for detecting zeros in spectra and the values presented here were manually measured by locating the frequency of deep valley in the spectra.)

	Dark /l/ of S2		Light /l/ of S2	
	Booth data in supine position	3-D FEM	Booth data in supine position	3-D FEM
F1	375	490	390	470
F2	625	1090	750	1150
F3	2625	2330	2625	2410
Zero(s)	3980*	4000	2890*, 3560*	2350, 2950, 4490

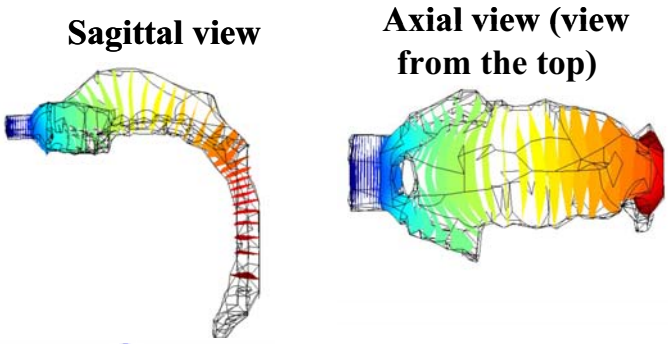
booth). Considering that both the dark /l/ and the light /l/ have the similar trend in this discrepancy, the subject may have used a different articulatory configuration during the booth data recording. Surprisingly, the zeros match well between the 3-D-FEM and the booth acoustic data.

5.4.2 Wave propagation at different frequencies

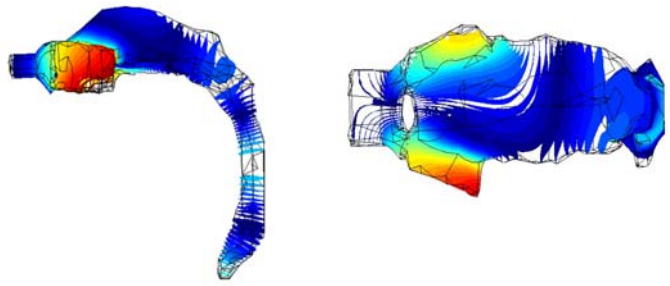
In order to understand why the zeros differ between the dark /l/ and the light /l/ and how the zeros are related to the articulatory configurations, particularly the linguo-alveolar contact and the supralingual space, the wave propagation properties at different frequencies inside the vocal tract have been studied.

Figure 5.7 shows the pressure isosurfaces for the dark /l/ at 500 Hz and 4000 Hz. The wave propagation at 500 Hz, which is indicated by the pressure isosurfaces, is approximately planar, even in the region immediately posterior to the linguo-alveolar contact. However, at 4000 Hz, a cross mode appears in the region posterior to the contact. When the wave propagates as a cross mode, it propagates towards the two sides of the vocal tract, and hardly comes out from the lips. Therefore, the volume velocity at the lips is extremely small and a zero is produced.

The lateral channels for the dark /l/ are about 1 cm long. It will be shown in Section 5.6 (page 106) that two lateral channels with one or two cm long are too short to produce a zero at 4000 Hz. The two lateral channels described in Zhang and Espy-Wilson (2004) are 3.4 cm and 5.0 cm long respectively, and they are much longer than the lateral channels for the dark /l/ here. In Zhang and Espy-Wilson

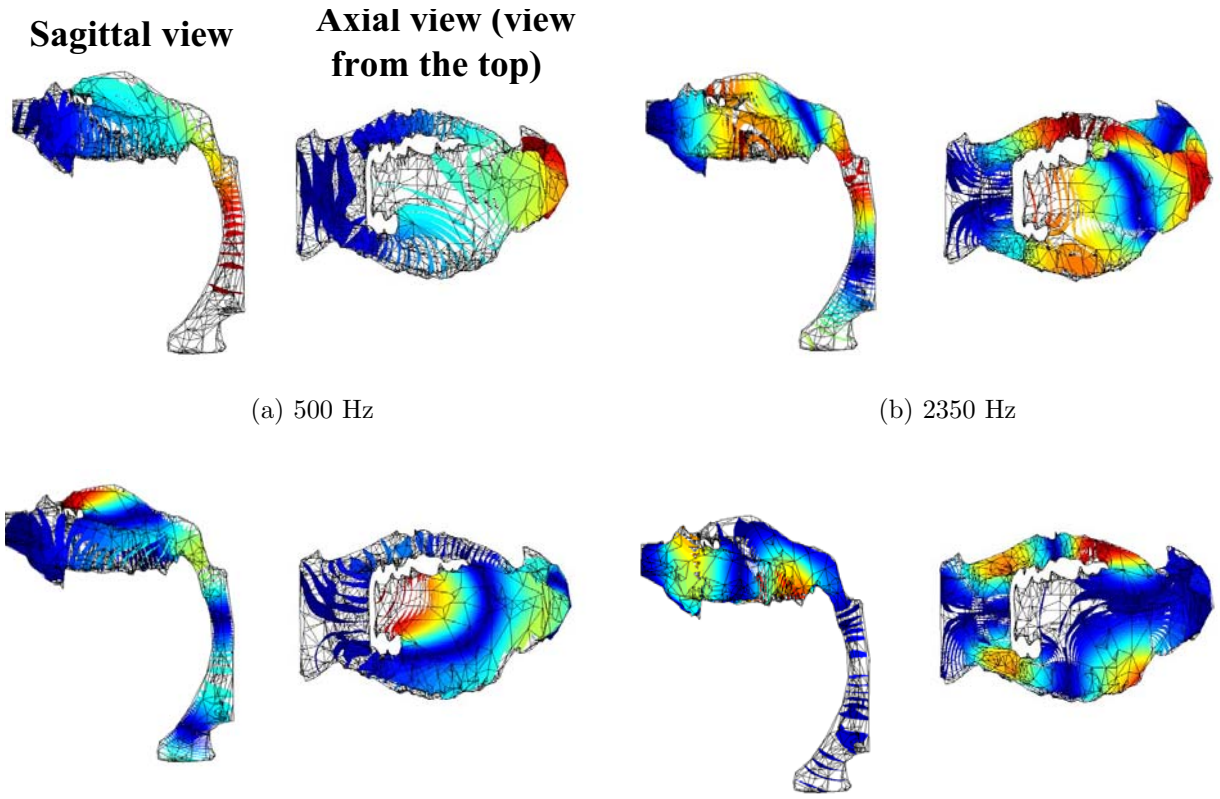


(a) 500 Hz



(b) 4000 Hz

Figure 5.7: Pressure isosurface plots of wave propagation inside the vocal tracts of the dark /l/ of S2 at different frequencies. (Pressure isosurfaces are coded by color: the red color stands for high amplitude and the blue color stands for low amplitude.) (a) 500 Hz, and (b) 4000 Hz.



(a) 500 Hz

(b) 2350 Hz

(c) 2950 Hz

(d) 4490 Hz

Figure 5.8: Pressure isosurface plots of wave propagation inside the vocal tracts of the light /l/ of S2 at different frequencies. (Pressure isosurfaces are coded by color: the red color stands for high amplitude and the blue color stands for low amplitude.) (a) 500 Hz, (b) 2350 Hz, (c) 2950 Hz, and (d) 4490 Hz.

(2004), the supralingual cavity is another source of zeros, but the dark /l/ we studied here does not have a separated supralingual cavity.

Another alternative explanation of the zero source in the dark /l/ is the asymmetry of the two effective lateral channels. The cross mode changes the approximate planar wave propagation into a more complex propagation, and the effective lateral lengths are modified to be longer than the real lateral channel length. Thus a zero around 4000 Hz is produced.

Figure 5.8 shows the pressure isosurfaces for the light /l/ at frequencies 500 Hz, 2350 Hz, 2950 Hz and 4490 Hz. The asymmetry of the vocal tract due to the linguo-alveolar contact and the linguopalatal contact make the wave propagation more complex than it is in the case of dark /l/. At 500 Hz, the wave propagation in each branch (the right lateral channel or the left lateral channel plus the supralingual space) is approximately planar. At the first zero 2350 Hz, both branches have approximately planar wave propagation. The zero is attributed to the asymmetry between the two lateral channels. It is produced when the volume velocity output of the two lateral channels are 180 degree out of phase. At 2950 Hz, the pressure isosurfaces in the supralingual space makes it like a separate side branch of the vocal tract. The side branch has zero impedance and traps all of the energy at this frequency. Therefore, a zero is produced by the supralingual cavity. At 4490 Hz, the cross mode appears just as it does in the dark /l/, which produces the third zero.

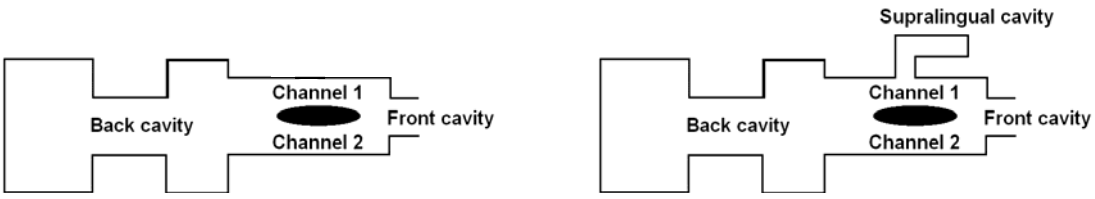
5.5 Area function based vocal tract modeling of /l/

Based on the wave propagation properties, area function vocal tract models of /l/ can be obtained from the 3-D vocal tract geometry, as done in Chapter 4. However, given its complex articulatory configuration, a more detailed vocal tract model as in Zhang and Espy-Wilson (2004) (see Figure 2.6 in Chapter 2) is needed. The complex geometry of /l/ production makes the area function extraction process difficult. Dividing the vocal tract into different components and assigning an area function to each component is not straightforward. Translating a 3-D geometry into a set of area functions is a simplification process. The objective here is to make the area function vocal tract model reproduce the main characteristics of the 3-D acoustic response from 3-D FEM, including the formants and zeros. This might have to be done based on a trial-and-error process.

5.5.1 Area functions of the dark /l/

Figure 5.9a shows a schematic of area function based vocal tract model for the dark /l/ of S2. The 3-D geometry of the dark /l/ does not have a supralingual space as a separate side branch. Thus, unlike the /l/ model in Zhang and Espy-Wilson (2004), there is no supralingual cavity. Based on the wave propagation properties shown in Figure 5.7, a set of grid lines is created for the area function extraction, and those gird lines are shown in Figure 5.10a. The resulting area functions, based on the grid lines, are shown in Figure 5.10b.

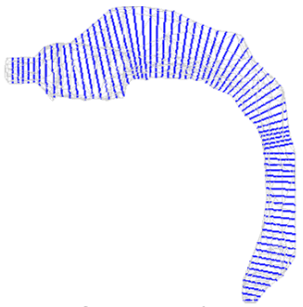
Figure 5.10c shows the acoustic responses. The resulting acoustic response



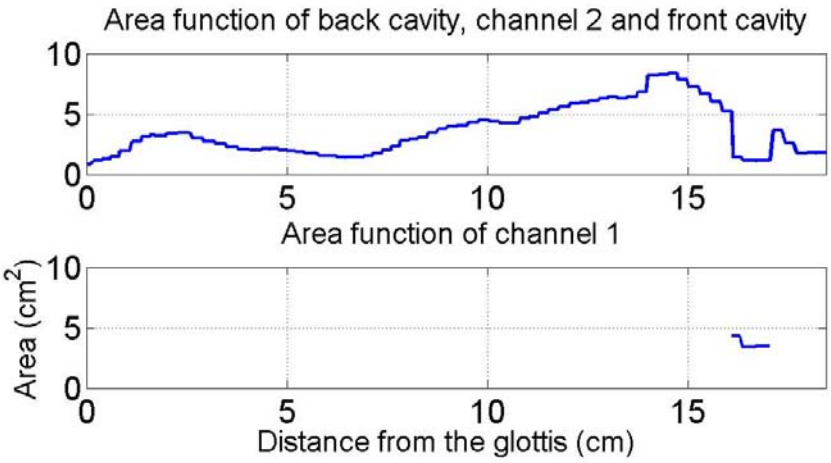
(a) The dark /l/ (b) The light /l/

Figure 5.9: Schematics of area function vocal tract models for /l/ production of S2. (Each component consists of an area function.) (a) the dark /l/, and (b) the light /l/.

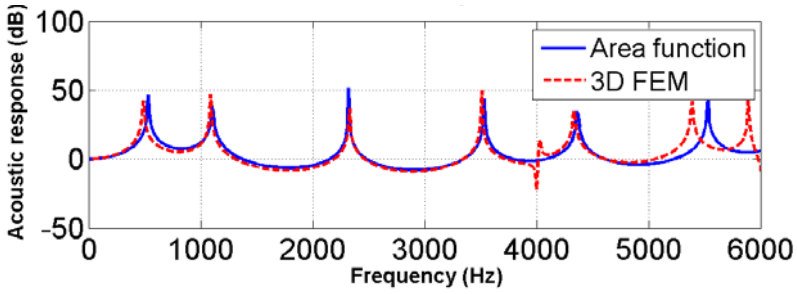
of the area function vocal tract model matches well the acoustic response of the 3-D FEM in F1-F4. Among F1-F4, F1 has the largest difference of 50 Hz. Figure 5.11 shows that combining the two channels into one does not change the acoustic response much. The largest difference of 60 Hz is in F5. However, the zero at 4000 Hz due to the cross mode can not be reproduced in this area function vocal tract model since there are no side branches. A more complicated way of extracting area functions is not explored here. It is feasible to get a matched zero by using other sets of grid lines. For example, the lateral channels can be artificially extended to the region posterior to the contact. Figure 5.12a shows a set of area functions with the two lateral channels lengthened to be 3.7 cm long by assigning the cross section area posterior to the contact equally into the two channels. This set of area functions did produce a zero at about 4 kHz, as shown in Figure 5.12b. But it will be very difficult to generalize this lengthening process for creating a zero in other cases, because this is a trial-and-error process. The challenge lies in that the area function vocal tract



(a)



(b)



(c)

Figure 5.10: For the dark /l/ of S2: (a) grid lines for area function extraction inside the vocal tract, (b) area function based on the grid lines, and (c) acoustic responses from 3-D FEM and area functions.

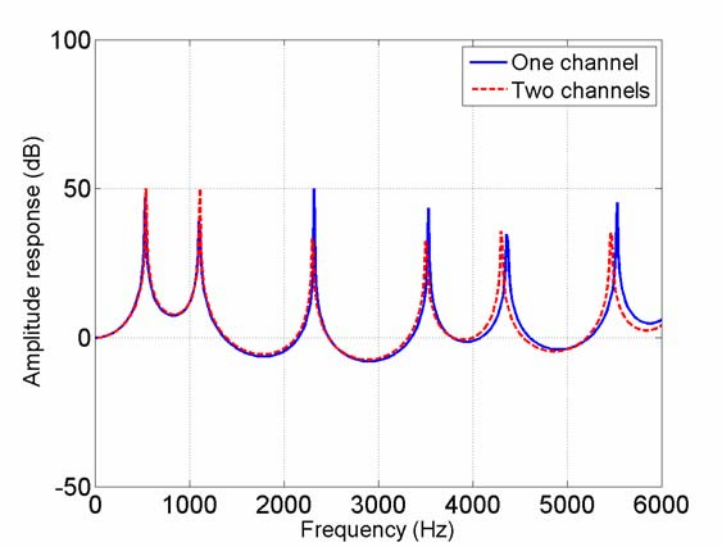
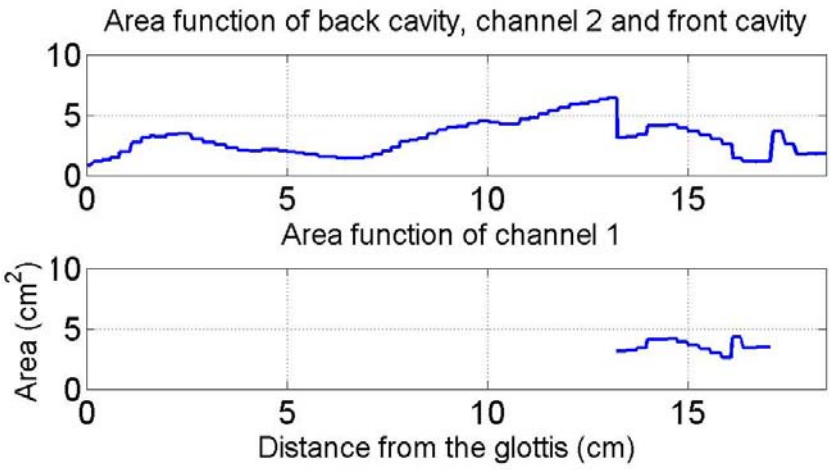


Figure 5.11: Acoustic response comparisons between the model with two lateral channels and the model with one combined channel for the dark /l/ of S2.

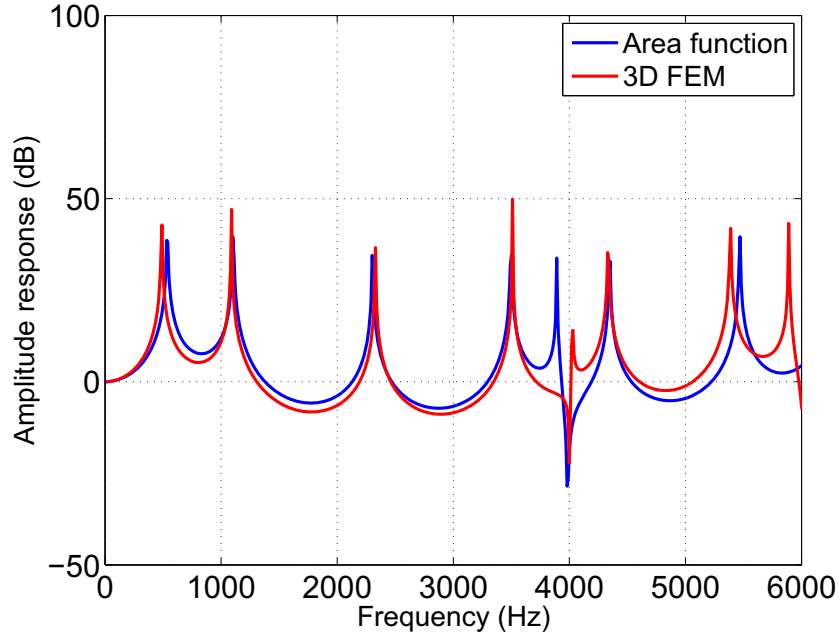
model is based on the planar wave propagation assumption, whereas the real vocal tract's acoustics has a 3-D property.

5.5.2 Area functions of the light /l/

Figure 5.9b shows a schematic of the area function based vocal tract model for the light /l/ of S2. In this model, channel 2 is the right lateral channel of the light /l/, and channel 1 consists of the left lateral channel of S2 plus part of the supralingual space. There is a supralingual cavity as a side branch of channel 1. This is different from the /l/ model in Zhang and Espy-Wilson (2004) where the supralingual cavity and the two lateral channels started from the same location. For the rest of this section, 'the channel' means 'channel 1' or 'channel 2' unless 'lateral' is mentioned. The reason for modeling the supralingual cavity as a side



(a)



(b)

Figure 5.12: For the dark /l/ of S2 with the lengthened lateral channels: (a) area functions, and (b) acoustic responses from 3-D FEM and area functions.

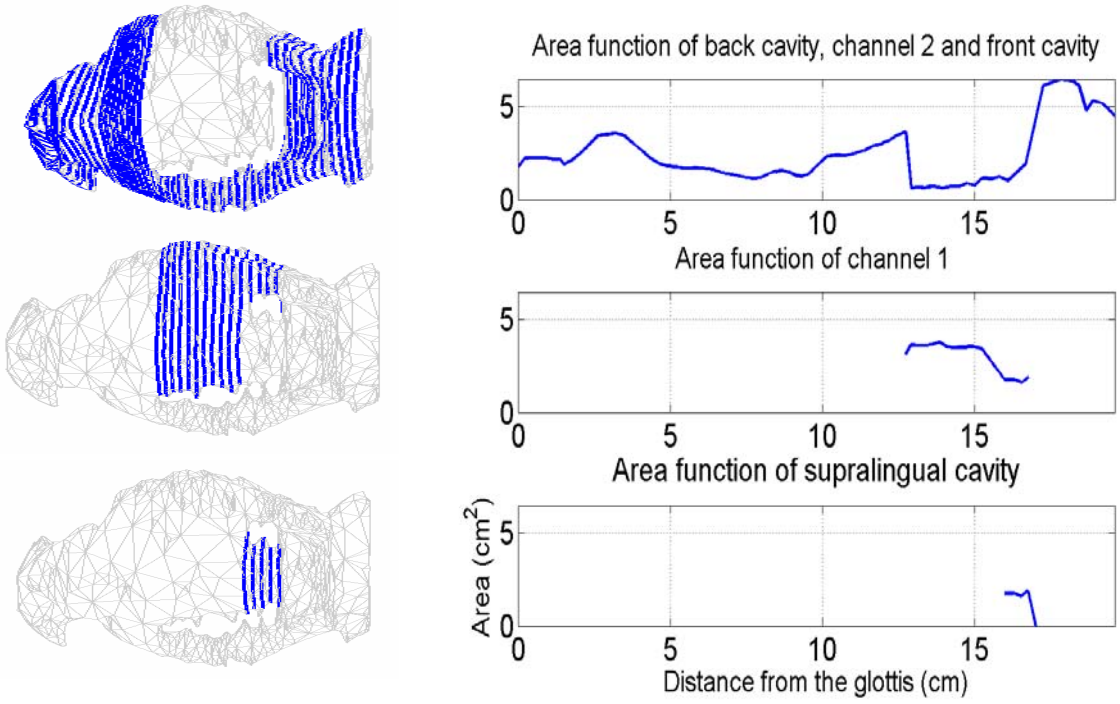
branch attached to channel 1 is because of the strong coupling the supralingual cavity and channel. Based on the wave propagation as shown in Figure 5.8, a set of grid lines is created for the area function extraction of the light /l/. However, it is not straightforward to extract each component in the schematic from the 3-D geometry. Two methods have been applied to get the area functions for each component. Both methods can reproduce F1-F3 from the 3-D FEM, but the second method results in a better match of the zeros.

5.5.2.1 The first method of area function extraction

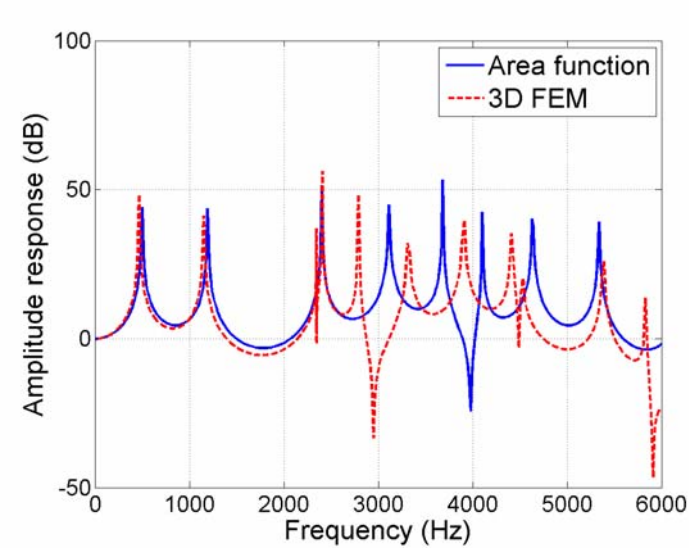
Figure 5.13a shows the grid lines for the area function extraction. The top plot of grid lines is for the back cavity, channel 2 (the right lateral channel in the 3-D geometry) and the front cavity. The middle plot of grid lines is for channel 1 (channel 1 starts at the same location as the right lateral channel, but it essentially includes the left lateral channel plus part of the supralingual space). The bottom plot of grid lines is for the supralingual cavity. This method of dividing the geometry into individual components is intuitive since the supralingual cavity specified here is a natural side branch to channel 1.

Figures 5.13b shows the area functions for each component in the model. The two channels are about 4.5 cm long, and the supralingual cavity is short, only 1 cm.

Figures 5.13c shows that the resulting acoustic response of the area function vocal tract model matches the acoustic response of the 3-D FEM in F1-F3. Among the first three formants, F1 has the largest difference of 50 Hz. However, the zeros

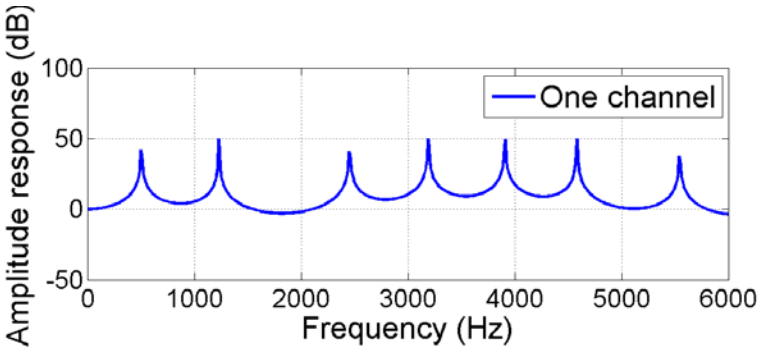


(a) (b)

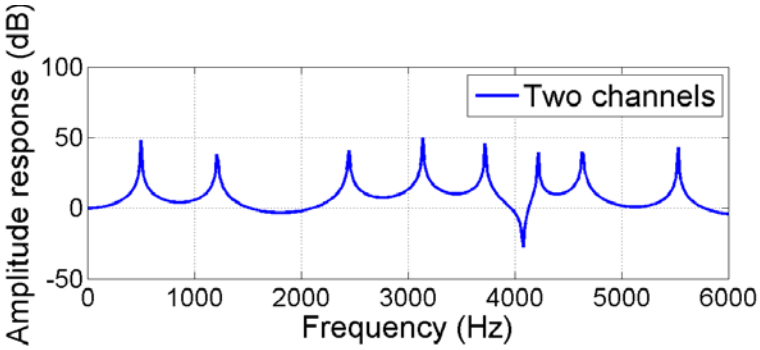


(c)

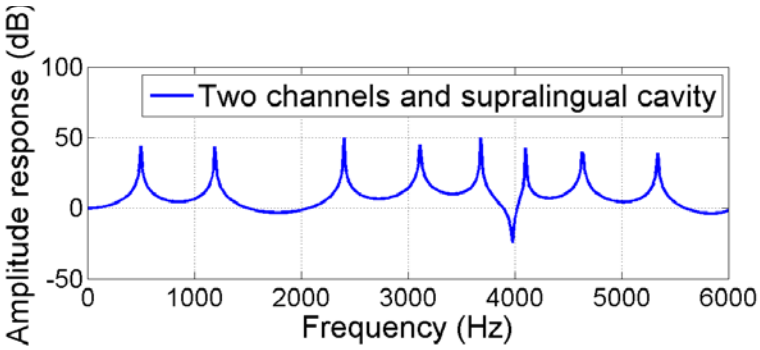
Figure 5.13: For the light /l/ of S2 with the first method of area function extraction: (a) grid lines for area function extraction inside the vocal tract, (b) area functions based on the grid lines, and (c) acoustic responses from 3-D FEM and area functions.



(a) One channel without supralingual cavity



(b) Two channels without supralingual cavity



(c) Two channels with supralingual cavity

Figure 5.14: For method 1: acoustic response comparisons among the different models by removing supralingual cavity and/or combining two channels into one channel for the light /l/ of S2. (a) one channel (combining two channels into one), (b) two channels without supralingual cavity, and (c) two channels with supralingual cavity.

from the 3-D FEM can not be reproduced in the area function vocal tract model. A zero appears at 3980 Hz for the area function vocal tract model.

Figure 5.14 shows how the acoustic response changes while adding two channels and supralingual cavity in the area function model. The purpose of this manipulation is to study the sources of the zeros in the acoustic response. Figure 5.14a is the acoustic response when the two channels are combined into one channel and the supralingual cavity is excluded. It can be seen that there is no zero in the acoustic response. Figure 5.14b shows that adding a two-channel module will produce a zero at 4080 Hz in the acoustic response. Figure 5.14c shows that the further addition of the supralingual cavity does not produce an extra zero. So it can be concluded that the zero produced in the area function model is from the two channels. Intuitively, the supralingual cavity is too short to produce a zero below 6000 Hz. Its length is about 1 cm. Based on the acoustic theory, the first zero produced by the supralingual cavity should be at 8750 Hz (calculated by the equation $c/(4L)$, where c is the sound speed, and L is the length of the cavity).

5.5.2.2 The second method of area function extraction

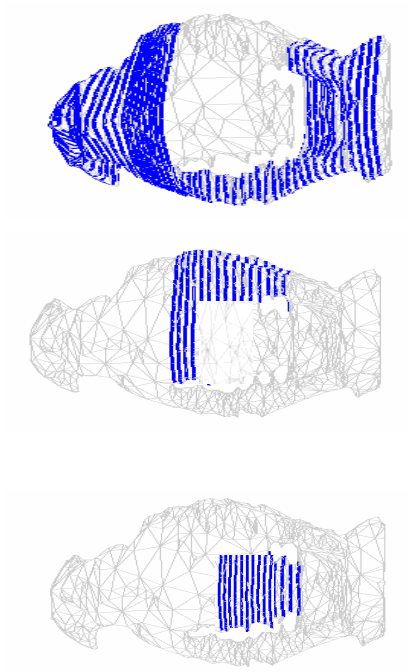
As observed from the first method, the supralingual cavity is too short to produce a zero. But the wave propagation in Figure 5.8 clearly shows that this cavity functions as a side branch. This indicates that the effective length of this cavity is longer than 1 cm. Another method is applied to determine the division of grid lines as shown in Figure 5.15a. In this method, the supralingual cavity has been

lengthened to 3.0 cm. In the meantime, the area of channel 1 has been reduced to half of the area as obtained in method 1 to make the total area of the supralingual cavity and the left channel invariant for each cross section.

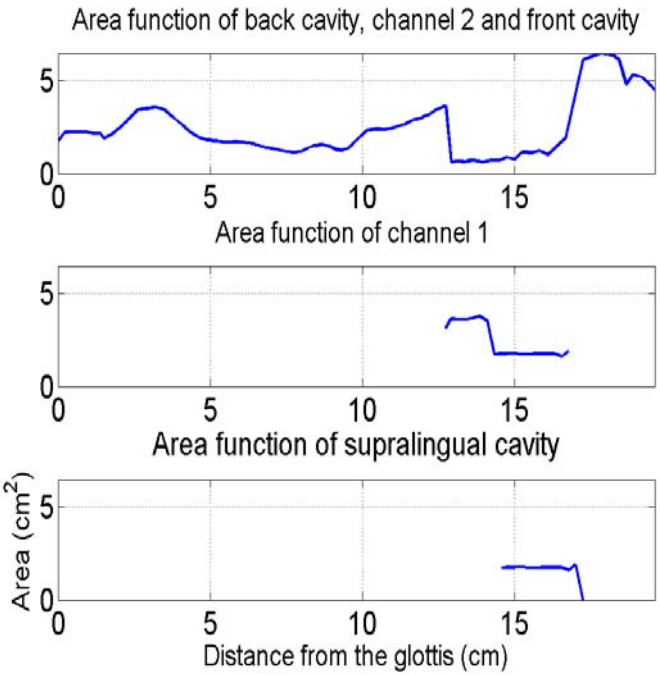
The resulting area functions are shown in Figures 5.15b. It can be seen that the area function of channel 1 has an abrupt change due to this new area function extraction strategy, and the length of the supralingual cavity is changed to 3 cm.

Figures 5.15c shows the acoustic response from the area functions. It can be seen that the acoustic response from the area functions matches the acoustic response of 3-D FEM in F1-F3. Among F1-F3, F3 has the largest difference of 70 Hz. There are two zeros produced, one is at 2910 Hz, and the other one is at 4600 Hz. These two zeros are close to the second and third zeros from 3D FEM, which are 2950 Hz and 4490 Hz. However this method of extracting the area functions is a trial-and-error process, because there is no systematic way to determine the length of the supralingual cavity. Intuitively, the first zero produced by a 3 cm long supralingual cavity should be at around 2920 Hz (calculated by the equation $c/(4L)$, where c is the sound speed, and L is the length of cavity). Remember that the first zero at 2350 Hz from 3-D FEM could not be reproduced in the area function vocal tract model.

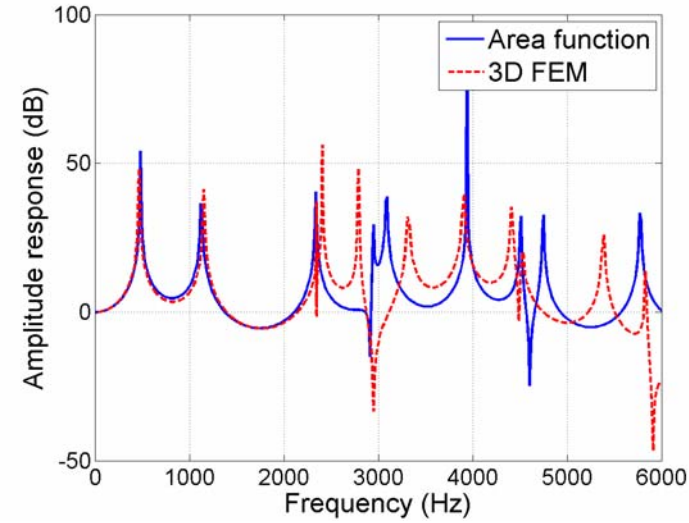
Figure 5.16 shows how the acoustic response changes while adding the two-channel module and the supralingual cavity in the area function model. Figure 5.16a is the acoustic response when the two channels are combined into one channel and the supralingual cavity is excluded. It can be seen that there is no zero in the acoustic response. Figure 5.16b shows that adding a two-channel module will



(a)

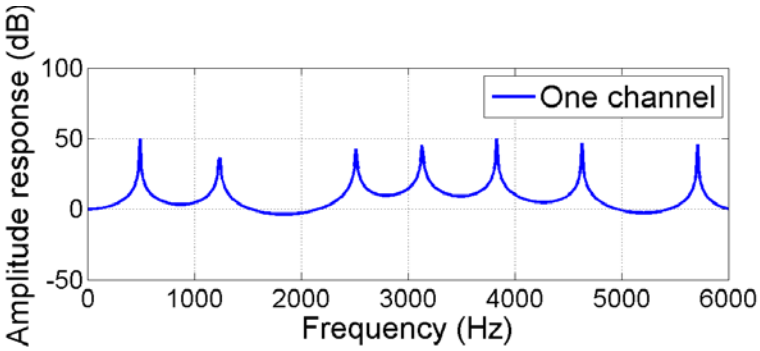


(b)

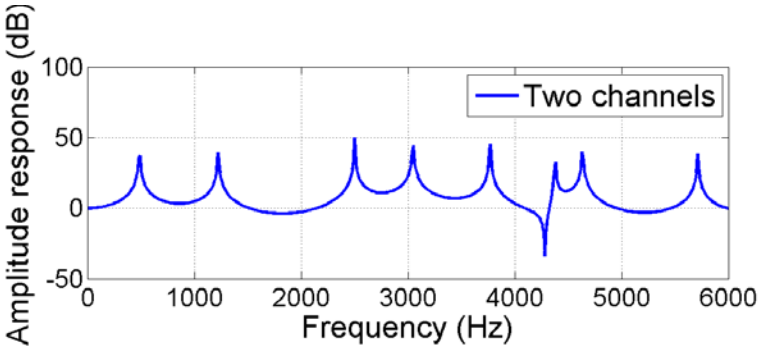


(c)

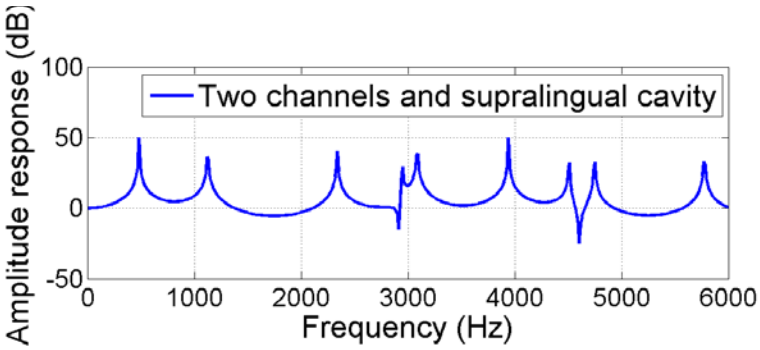
Figure 5.15: For the light /l/ of S2 with the second method of area function extraction: (a) grid lines for area function extraction inside the vocal tract, (b) area functions based on the grid lines, and (c) acoustic responses from 3-D FEM and area functions.



(a) One channel without supralingual cavity



(b) Two channels without supralingual cavity



(c) Two channels with supralingual cavity

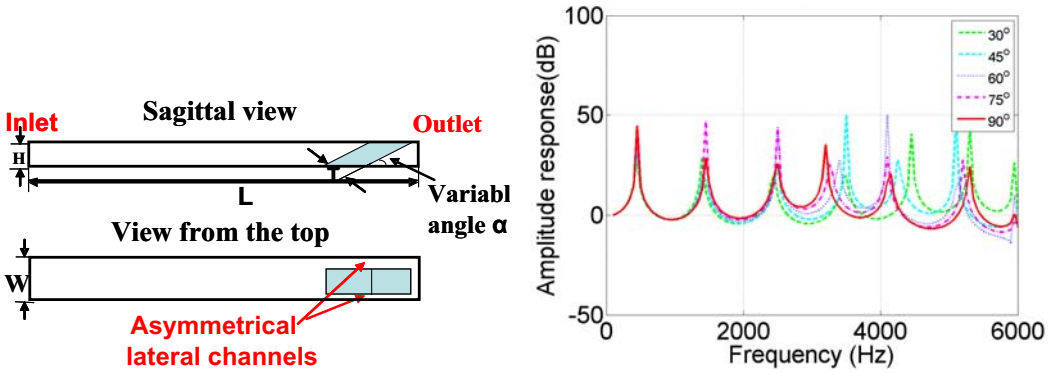
Figure 5.16: For method 2: acoustic response comparisons among the different models by removing supralingual cavity and/or combining two channels into one channel for the light /l/ of S2. (a) one channel (combining two channels into one), (b) two channels without supralingual cavity, and (c) two channels with supralingual cavity.

produce a zero at 4280 Hz in the acoustic response. Figure 5.16c shows that further adding a supralingual cavity produces an extra zero which is at 2910 Hz and the zero at 4280 is changed to 4600 Hz. The addition of the supralingual cavity affects the zero produced by the two lateral channels. This interaction is because the supralingual cavity is connected to channel 1 and it affects the acoustic impedance of this channel. So it can be concluded that the zero at 4600 is produced by the the two channels, and the zero at 2910 Hz is produced by the supralingual cavity.

5.6 The simple 3-D vocal tract models

Without a fully functioning 3-D tongue model, it is difficult to modify the real 3-D vocal tract shape for a different tongue configuration. Therefore it is hard to study how the acoustic response is modified with a change of the articulatory configuration. The reconstructed 3-D geometry in this study is not flexible enough to be changed arbitrarily. In the light /l/ and the dark /l/ we studied, there is always an alveolar contact. The /l/ production without a contact was reported (Narayanan et al., 1997). But there is no work on a vocal tract model for this case where the linguo-alveolar contact is not complete so that there is no occlusion, but rather a constriction.

In order to gain insights on the acoustic effect of the linguo-alveolar contact and the lateral channels, two simple 3-D vocal tract models have been studied.



(a) Geometry (b) Acoustic response

Figure 5.17: The simple 3-D vocal tract model I with two asymmetrical lateral channels. (a) the geometry, and (b) the acoustic responses for different angles α. (H: 1.4 cm, W: 2.8 cm, L: 18 cm, T: 1 cm, block width: 1.4 cm, block starting location: 4.8 cm from the outlet.)

5.6.1 Model I

Figure 5.17a shows the first simple model of a 3-D vocal tract. It is a uniform tube with a rectangular cross section where a block with 1 cm thickness is positioned at different angles in the front (starting location: 4.8 cm from the outlet) to simulate the contact and the two lateral channels in /l/ productoin. Its length of 18 cm is based on the average vocal tract length of a male adult, and its cross section area of 4 cm^2 is based on the average volume of the human vocal tract (Stevens, 1998). The block 's width is half or three-fourths of the vocal tract width. The length of the lateral channels is in the range of 1-2 cm when the angle α is in the range of 30-90 degree.

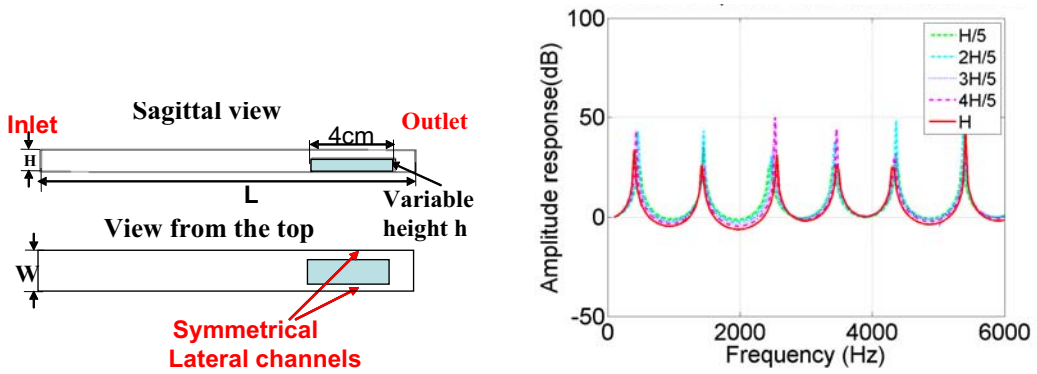
Three configurations have been simulated: symmetrical lateral channels, asym-

metrical lateral channels, and only one lateral channel. These configurations have been realized by shifting the block step by step from the center to the side of the vocal tract. However, zero does not appear below 5000 Hz for any configuration. One example of the vocal tract acoustic response is shown in Figure 5.17b. Each line in the figure stands for the acoustic response at a specific angle α. These results indicate that a 1-2 cm long contact can not produce zeros below 5000 Hz, even if the two lateral channels are asymmetrical. In contrast, the dark /l/ of S2 has a 1 cm long linguo-alveolar contact, and it has a zero at about 4000 Hz in the acoustic response. This further proves that the cross mode posterior to the contact produced the zero around 4000 Hz.

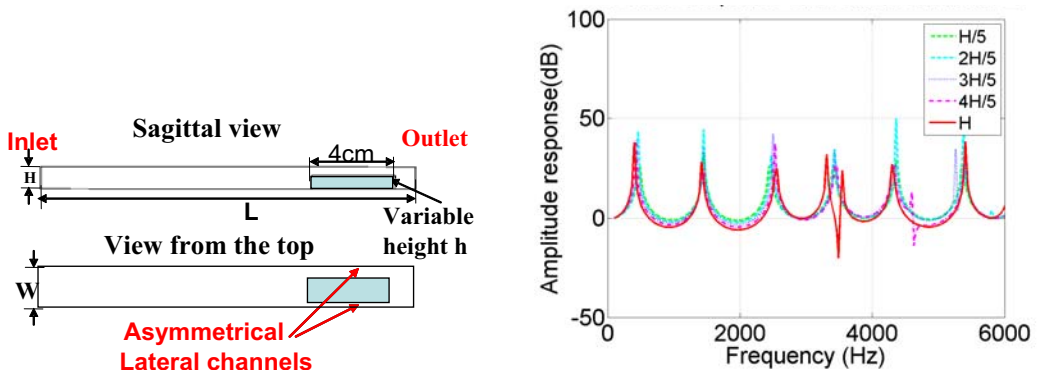
5.6.2 Model II

The left side of Figure 5.18 shows the second simple model of a 3-D vocal tract. The dimensions in model II are the same as in model I. Instead of a block positioned at an angle, a block with certain length and certain height is positioned flat in the front to simulate the two lateral channels. When the height of the block reaches the height of the vocal tract, the two lateral channels are separated by the closure.

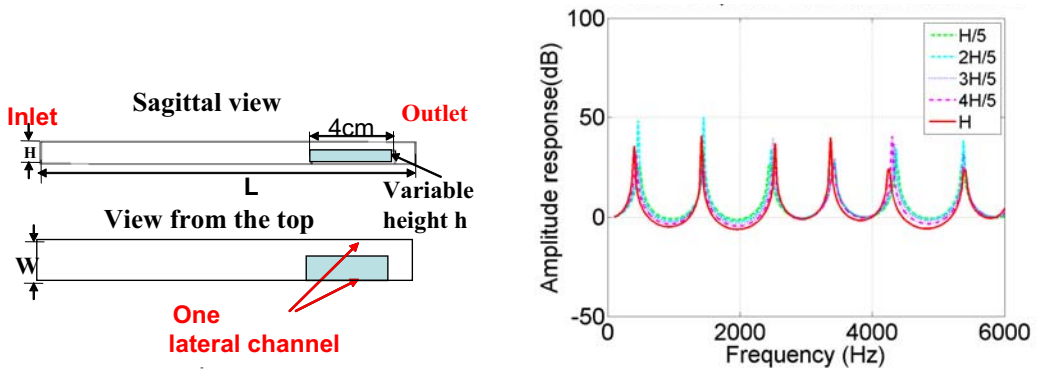
Three configurations have also been simulated: symmetrical lateral channels, asymmetrical lateral channels, and only one lateral channel. These configurations have been realized by shifting the block step by step from the center to the side of the vocal tract.



(a) two symmetrical lateral channels



(b) two asymmetrical lateral channels (the ratio of the two channels cross section areas is 3:5)



(c) one lateral channel

Figure 5.18: The simple 3-D vocal tract model II (the left side is the geometry, and the right side is the acoustic response for different block heights h, H: 1.4 cm, W: 2.8 cm, L: 18 cm, block width: 1.4 cm, block starting location: 5 cm from the outlet). (a) two symmetrical lateral channels, (b) two asymmetrical lateral channels (the ratio of the two channels cross section areas is 3:5.), and (c) one lateral channel.

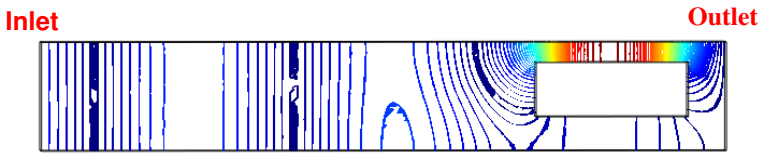


Figure 5.19: Pressure isosurfaces at 3340 Hz (a zero) in the simple 3-D vocal tract model II.

A vocal tract model with two lateral channels that are 4 cm long is shown in Figure 5.18. A zero does not appear below 6000 Hz for the symmetrical and one lateral channel configurations for any block height, as shown in Figures 5.18a and 5.18c. In the case of two asymmetrical channels shown in Figure 5.18b, the acoustic response has a zero at 4630 Hz for h = 4/5H or a zero at 3340 Hz when there is a complete closure. This means that the two lateral channels with lengths of 4 cm can produce a zero below 6000 Hz, but only when there is a closure or a narrow constriction. It is indicated that a closure can lower the frequency of the zero.

In order to understand why the asymmetrical configuration can produce a zero below 6000 Hz, pressure isosurfaces at 3340 Hz have been plotted in Figure 5.19. It can be seen that the geometry asymmetry makes the wave propagation different inside the two lateral channels. Even though the two lateral channels have the same length, the effective lengths are different due to the asymmetry, and therefore a zero is produced.

In order to understand how the lengths of the lateral channels affect the zero, the lengths were varied from 2 cm to 6 cm have been simulated. Figure 5.20 shows the acoustic response with different channel lengths. These simulations are based

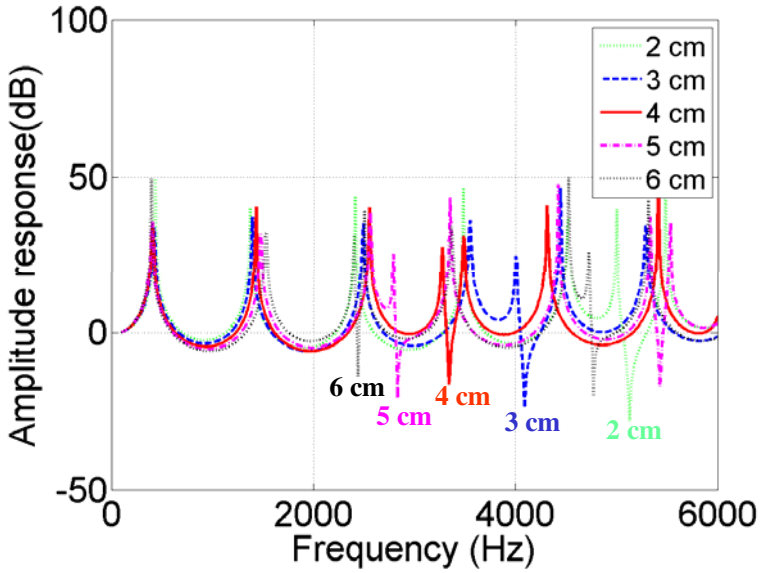


Figure 5.20: Acoustic responses at different lateral lengths in the simple 3-D vocal tract model II with a closure.

on the asymmetrical configurations with a closure. It can be seen that when the length varies from 2 cm to 6 cm, the zeros vary from 5130 Hz to 2440 Hz accordingly. For the geometry with the same asymmetry, the longer channels will produce a zero at lower frequency. This result is also confirmed in VTAR. It shows that, for a fixed length difference between the two channels, the two channels with longer total length produce zeros at lower frequencies.

5.7 Discussion

As described in Section 5.2, the two tongue shapes studied here are just two examples of /l/ production, and they are not exclusively for a dark /l/ or for a light /l/. The two tongue shapes can produce the /l/ sound with similar patterns in F1-F3. But they results in a different number of zeros below 6000 Hz and the frequencies

of the zeros are different. The finite element analysis revealed the acoustic effects of the lateral channels and the supralingual space. The zeros can be produced by the asymmetrical channels, the supralingual cavity as a side branch, and the cross mode posterior to the linguo-alveolar contact.

Furthermore, the simple 3-D vocal tract models have been simulated to show the effect of lateral channels and the linguo-alveolar contact. The results show that zeros can not be produced if the lateral channel length is too short, or the channels are not asymmetrical, or there is no narrow constriction or complete closure. The zero can be changed significantly when the length of the lateral channels ranges from 2-6 cm.

These results show us that the zeros in the spectrum of an /l/ could be produced in different ways, and the frequency of the zeros can vary a lot with the variation of the articulatory configuration. This variability in the zeros should increase the complexity of the /l/ spectrum. It might be part of the reason why the lateral sound is more difficult to characterize than other consonants (Stevens, 1998).

Based on the 3-D FEM, area function vocal tract models have been developed for the dark /l/ and the light /l/. It has been shown that one area function vocal tract model might not be able to accommodate all the articulatory configurations. For example, a component for the supralingual cavity does not exist in the area function vocal tract model for the dark /l/, and its location might vary with the variation of the articulatory configuration .

The area function vocal tract modeling is an empirical process in terms of reproducing the zeros. Zeros could be missed or predicted inaccurately in the area

function model. A caveat here is that the area function vocal tract model is a simplification of a 3-D vocal tract model, and the conclusions from it should be taken cautiously or should be verified with the 3-D acoustic analysis.

5.8 Chapter summary

In this chapter, two tongue shapes for American English /l/ production were studied. One is for producing a sustained dark /l/, and the other is for producing a sustained light /l/. Both the dark /l/ and the light /l/ have similar patterns in F1-F3, but differ in the number of the zeros in the spectrum, and the frequencies of the zeros below 6000 Hz. Using finite element analysis based on magnetic resonance images of the vocal tract for sustained productions, the acoustic effects of the lateral channels and the supralingual space have been investigated, and proper area function vocal tract models have been suggested for both cases. For the dark /l/, the zero below 6000 Hz is produced by the cross mode posterior to the linguo-alveolar contact. For the light /l/, the zeros below 6000 Hz are produced by the asymmetrical channels, the supralingual cavity as a side branch and the cross mode posterior to the linguo-alveolar contact.

Two simple vocal tract models have been simulated to show the effect of lateral channels and the linguo-alveolar contact. The results show that lateral channels that are 1-2 cm long can not produce a zero in the region of F3-F5. In order to get a zero in the region of F3-F5, the lateral channels have to be asymmetrical and 3-6 cm long. In addition, a narrow constriction or a complete closure is also required.

Chapter 6

Acoustic variability and discriminative power analysis

6.1 Introduction

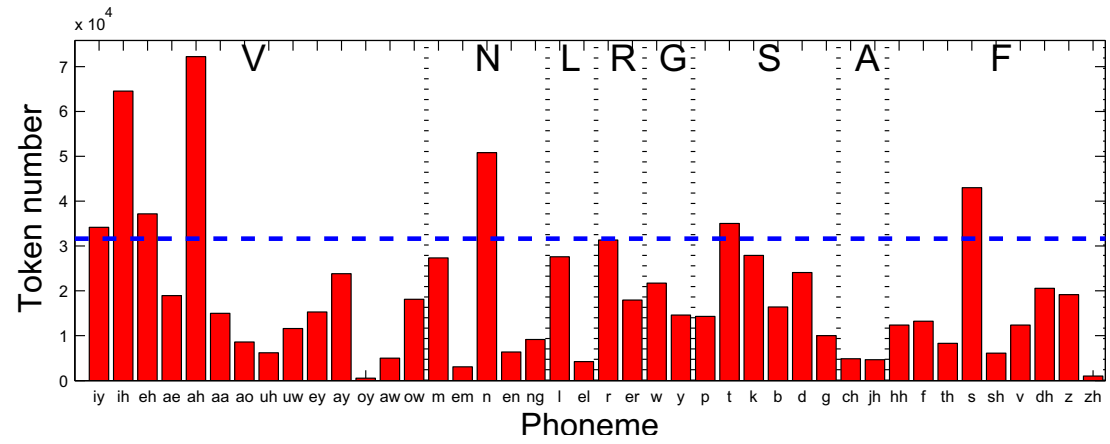
It has been shown in Chapter 4 that the retroflex and bunched tongue shapes for /r/ in American English have different F4 and F5 patterns. It has also been shown in Chapter 5 that the two different tongue shapes for the dark /l/ and the light /l/ have different zero patterns in the F3-F5 region. These different formant or zero patterns across different tongue shapes may increase the spectra variability. The articulatory variability of liquid sounds across the speakers might make them have more inter-speaker acoustic variability and, thereby, have more discriminative power in speaker recognition relative to other sounds (vowels, nasals, glides, fricatives, affricates and stops).

This chapter presents a preliminary study on the acoustic variability and the speaker discriminative power of different phonemes in American English. Different from previous studies (Antal and Toderean, 2006; Kajarekar and Hermansky, 2001) which used either read sentences as in the TIMIT database or speech of telephone quality as in the Switchboard database, this study was based on a conversational speech database sampled at 48 kHz. Analysis of variance (ANOVA) was performed for the acoustic variability. Speaker identification experiments were performed to study the discriminative power of different phonemes.

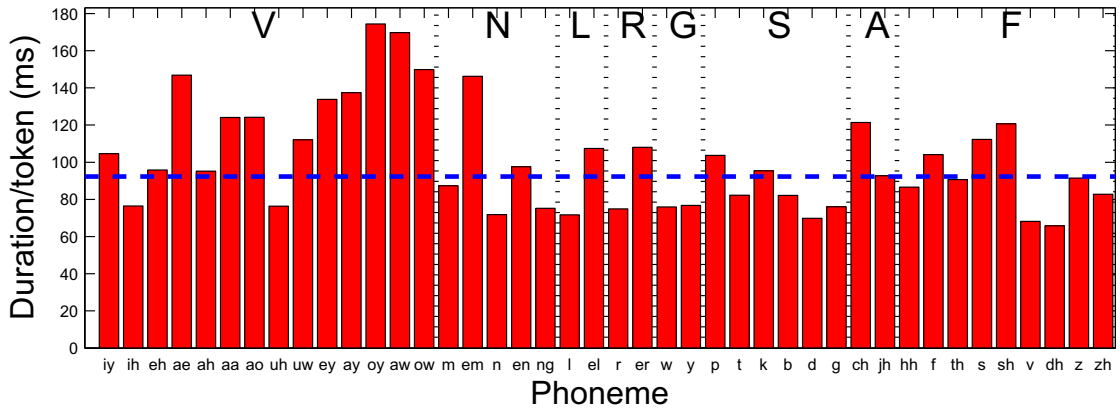
6.2 Database and acoustic parameters

The Buckeye database (Pitt et al., 2005) was used for both the ANOVA analysis and the speaker identification experiments in this study. It was resampled at 16 kHz. In addition to its description in Section 3.1.2 (page 36), the token number of each phoneme in the Buckeye database is shown in Figure 6.1a (only 42 phonemes are listed in the figure). Figure 6.1b shows the average duration for each phoneme. In this database, the syllabic /l/ and the syllabic /r/ are labeled as /el/ and /er/, respectively. In this chapter, the label /l/ or /r/ refers to its consonantal sound as well as its syllabic sound. The occurrence frequency for /r/ (including the syllabic /r/) is 5.8%, whereas the occurrence frequency for /l/ (including the syllabic /l/) is 3.7%. The syllabic /r/ and /l/ have relatively longer durations than the consonantal /r/ and /l/ (108 ms vs. 75 ms for /r/, and 107 ms vs. 72 ms for /l/). It can be seen in Figure 6.1a that /oy/, /em/ and /zh/ have the least numbers of token. Therefore, they do not have enough data for the statistical acoustic model training in the speaker identification experiments. In addition, the occurrence frequencies and the average durations for /r/, /l/ and other broad phonetic classes are listed in Table 6.1.

The acoustic parameters extracted in this study were the mel-frequency filterbank (MFB) energies (31 coefficients). They were computed by a feature extraction routine in the MIT Lincoln lab speaker recognition system (Reynolds et al., 2000). The mel-frequency scale and the center frequencies of the filterbanks are shown in Figure 6.2. The frame size for the feature extraction was 20 ms, and the shift



(a) Token number



(b) Average duration

Figure 6.1: Token information of each phoneme in the Buckeye database (**V**:Vowels, **N**: Nasals, **R**:/r/, **L**:/l/, **G**:Glides, **S**:Stops, **A**:Affricates, **F**:Fricatives, the horizontal dashed blue line is for the average). (a) token number, and (b) average duration.

Table 6.1: Occurrence frequencies and average durations of /r/, /l/ and other broad phonetic classes in the Buckeye database.

	Vowels	Nasals	/l/	/r/	Glides	Stops	Affricates	Fricatives
Frequency	38.8%	11.3%	3.7%	5.8%	4.3%	15.0%	1.1%	15.9%
Avg. duration (ms)	106.9	80.6	76.4	87.0	76.2	84.7	107.4	94.0

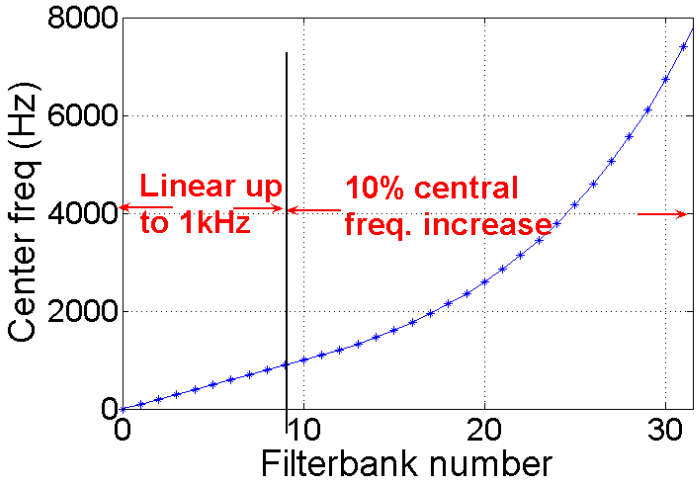


Figure 6.2: Mel frequency scale and center frequency of each filterbank used in MIT Lincoln lab speaker recognition system (Reynolds et al., 2000).

size was 10 ms. The reason for using the MFB coefficients instead of MFCC and others was twofold. First, it is convenient to relate each MFB coefficient to a mel-frequency filterbank, so that the acoustic variabilities at different frequency bands can be analyzed directly. Second, the MFB coefficients have very similar performance as MFCC in terms of speaker recognition accuracy, which was shown in a previous experiment on the Switchboard-I database in our lab.

6.3 Acoustic variability

6.3.1 Definitions

In the ANOVA analysis, the inter-speaker variability and the intra-speaker variability were analyzed. For a speaker identification task, it is desirable to have a large inter-speaker variability and a small intra-speaker variability in the acoustic parameter, so that the acoustic parameters from different speakers will not have much overlapping. The statistical F-ratio of an acoustic parameter is defined as the ratio of the inter-speaker variability to the intra-speaker variability. Intuitively, a high F-ratio means that the speakers are well separated by the corresponding acoustic parameter. Therefore, the acoustic parameter will have a good ability to distinguish a speaker from others.

In the case that the acoustic parameter is a scalar, the F-ratio is calculated as in Equation 6.1, where σ^2_{inter}, as the inter-speaker variability, is the inter-speaker mean variance and σ^2_{intra}, as the intra-speaker variability, is the mean of the intra-speaker variance.

$$F = \frac{\sigma^2_{inter}}{\sigma^2_{intra}} \tag{6.1}$$

In the case that the acoustic parameter is a vector, the F-ratio calculation uses the scatter matrices (Theodoridis and Koutroumbas, 2003). First, the following matrices are defined:

Within-speaker scatter matrix

$$S_w = \sum_{i=1}^{N} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_i)^T \tag{6.2}$$

Between-speaker scatter matrix

$$S_b = \sum_{i=1}^{N} n_i(\overline{X}_i - \overline{\overline{X}}_i)(\overline{X}_i - \overline{\overline{X}}_i)^T \tag{6.3}$$

where n_i is the token number for each speaker, i means the ith speaker, N is the total number of speakers, X_i is the vector of acoustic parameters. $\overline{X}_i$ is the mean vector for the ith speaker, as defined by Equation 6.4. $\overline{\overline{X}}_i$ is the mean vector of all the speakers, as defined by Equation 6.5.

$$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \tag{6.4}$$

$$\overline{\overline{X}}_i = \frac{1}{N} \sum_{j=1}^{N} \overline{X}_i \tag{6.5}$$

In the case that the acoustic parameter is a vector, the F-ratio is calculated as in Equation 6.6 or Equation 6.7.

$$F_1 = trace(S_w^{-1} S_b) \tag{6.6}$$

$$F_2 = trace(S_b)/trace(S_w) \tag{6.7}$$

The inter-speaker variability Σ_{inter} and the intra-speaker variability Σ_{intra} are calculated by Equations 6.8 and 6.9, respectively.

$$\Sigma_{inter} = trace(S_b) \tag{6.8}$$

$$\Sigma_{intra} = trace(S_w) \tag{6.9}$$

6.3.2 Results

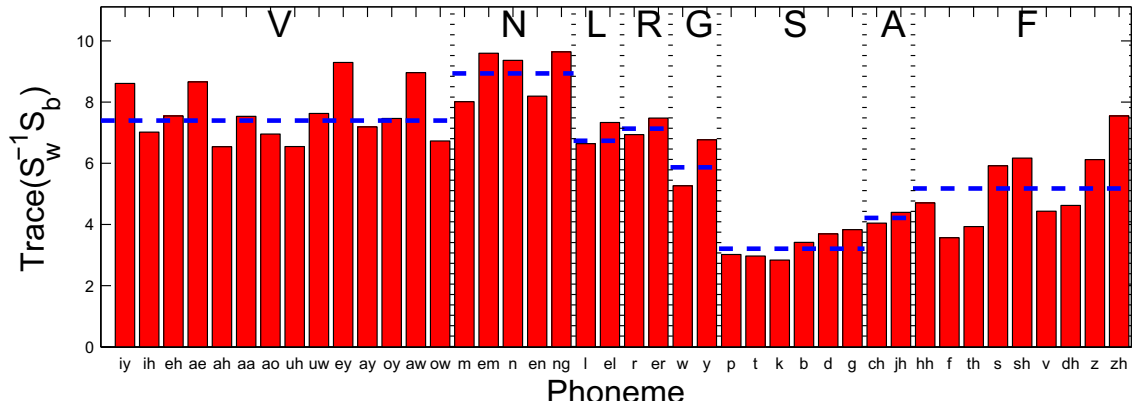
6.3.2.1 F-ratio and acoustic variability based on the 31 MFB coefficients

Figures 6.3a and 6.3b show the F-ratio of each phoneme in terms of the 31 MFB coefficients. Figure 6.3a shows the F-ratios computed by $trace(S_w^{-1}S_b)$, whereas Figure 6.3b shows the F-ratios computed by $trace(S_b)/trace(S_w)$. The dashed lines in the figures specify the average F-ratios for each broad phonetic class, and the averages were weighted by the token number of each phoneme in the broad class. It can be seen that the dynamic range of the F-ratios computed by these two equations are different. Furthermore they did not produce a consistent ranking among the phonemes. In Figure 6.3a, the positions of consonantal /r/, syllabic /r/, consonantal /l/ and syllabic /l/ in the F-ratio ranking are 21st, 15th, 25th and 17th, respectively. But, in Figure 6.3b, their positions in the F-ratio ranking are 14th, 8th, 18th and 1st, respectively. So, Equation 6.7 favors the liquid sounds in terms of the F-ratio ranking.

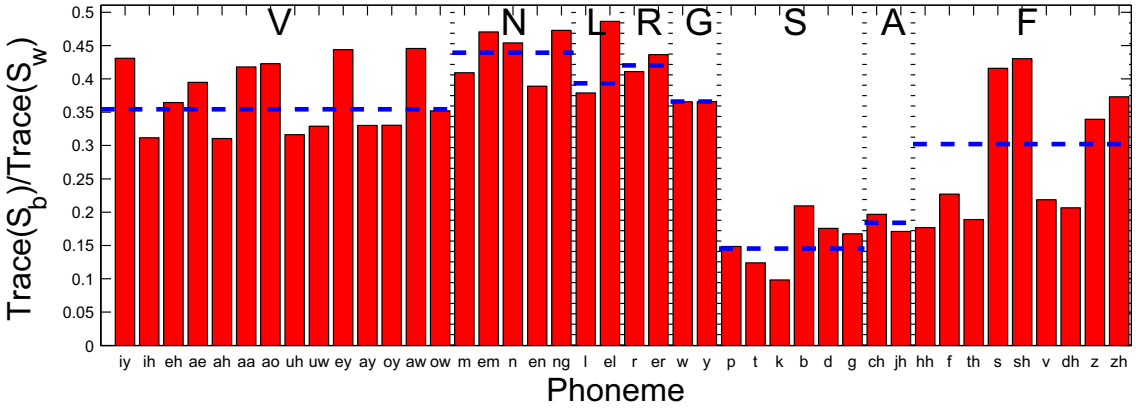
In addition, all the top ten phonemes with high F-ratios are nasals or vowels, except syllabic /l/ and syllabic /r/. For F-ratios computed by $trace(S_w^{-1}S_b)$, the top ten phonemes with high F-ratios are:

/nx/, /ng/, /em/, /n/, /ey/, /aw/, /ae/, /iy/, /en/, /m/.

For F-ratios computed by $trace(S_b)/trace(S_w)$, the top ten phonemes with high F-ratios are:



(a)



(b)

Figure 6.3: F-ratio of each phoneme (**V**:Vowels, **N**: Nasals, **R**:/r/, **L**:/l/, **G**:Glides, **S**:Stops, **A**:Affricates, **F**:Fricatives, the horizontal dashed blue line is for the average). (a) F-ratio in $trace(S_w^{-1}S_b)$, and (b) F-ratio in $trace(S_b)/trace(S_w)$.

Table 6.2: The weighted averages of F-ratio for /r/, /l/ and other broad phonetic classes in the Buckeye database (F_1 and F_2 are computed by $trace(S_w^{-1}S_b)$ and $trace(S_b)/trace(S_w)$, respectively).

	Vowels	Nasals	/l/	/r/	Glides	Stops	Affricates	Fricatives
F_1	7.4	8.9	6.7	7.1	5.9	3.2	4.2	5.2
F_2	0.35	0.44	0.39	0.42	0.37	0.15	0.18	0.30

/el/, /ng/, /em/, /n/, /nx/, /aw/, /ey/, /er/, /iy/, /sh/.

Although the F-ratio rankings from $trace(S_w^{-1}S_b)$ and $trace(S_b)/trace(S_w)$ have some discrepancy, the overall shapes of the two plots in Figures 6.3a and 6.3b are very similar to each other and the cross-correlation coefficient between them is 0.9. Furthermore, in terms of the weighted F-ratio averages, the rankings of the broad phonetic classes computed by both formulas are very similar. They only differ in the vowels' position. The ranking computed from $trace(S_w^{-1}S_b)$, in the descending order, is :

Nasals > Vowels > /r/ > /l/ > Glides > Fricatives > Affricates > Stops

whereas, the ranking computed from $trace(S_b)/trace(S_w)$, in the descent order, is :

Nasals > /r/ > /l/ > Glides > Vowels > Fricatives > Affricates > Stops

The weighted F-ratio averages for /r/, /l/ and the broad phonetic classes are also listed in Table 6.2.

The inter-speaker variability Σ_{inter} for each phoneme is shown in Figure 6.4a. The positions of consonantal /r/, syllabic /r/, consonantal /l/ and syllabic /l/ in the

inter-speaker variability ranking are 5th, 4th, 14th and 7th, respectively. So /r/ has a large inter-speaker variability, only smaller than /sh/, /s/ and /zh/. The averages of the inter-speaker variability for the broad phonetic classes are listed in Table 6.3. For the broad phonetic classes, the ranking of the averages of inter-speaker variability, in the descending order, is:

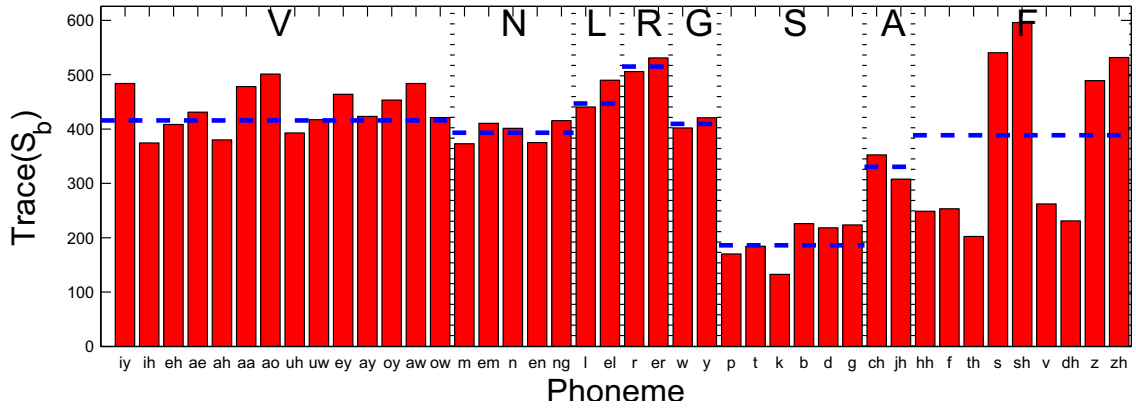
/r/ > /l/ > Vowels > Glides > Nasals > Fricatives > Affricates > Stops

The intra-speaker variability Σ_{intra} for each phoneme is shown in Figure 6.4b. The positions of consonantal /r/, syllabic /r/, consonantal /l/ and syllabic /l/ in the intra-speaker variability ranking are 16th, 18th, 23th, 37th, respectively. The averages of the intra-speaker variability for the broad phonetic classes are listed in Table 6.3. For the broad phonetic classes, the ranking of the average intra-speaker variability, in the descending order, is:

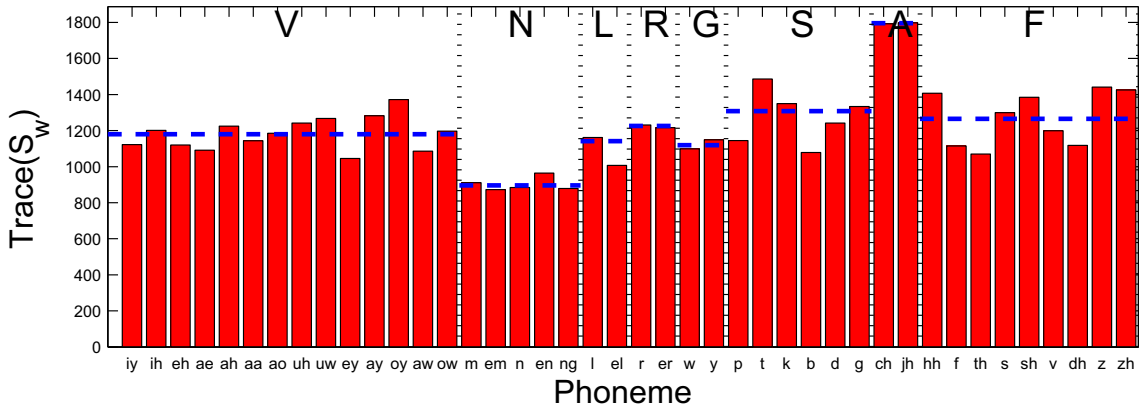
Affricates > Stops > Fricatives > /r/ > Vowels > /l/ > Glides > Nasals

It is not surprising to see that nasals have the smallest intra-speaker variability and the largest F-ratios in the weighted average. The human nasal cavity shape is almost fixed during articulation. Therefore nasal sounds have a very small variability in speech.

It can be seen in Table 6.3 that the inter-speaker variability has a smaller value than the intra-speaker variability. This indicates that the acoustic parameters for different speakers are overlapped with each other to some extent, so that the variance of the mean vectors among all the speakers is smaller than the average variance of the feature vector for each speaker.



(a)



(b)

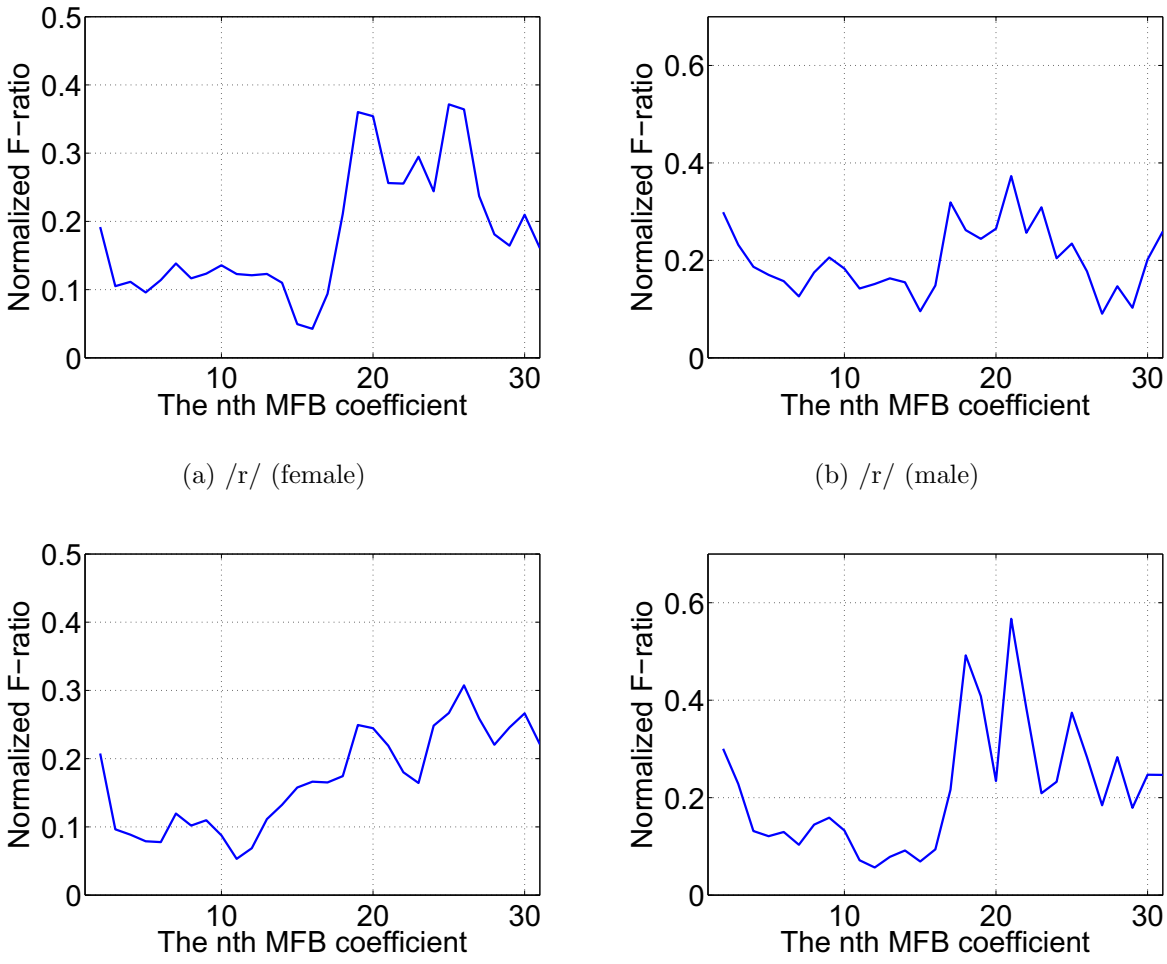
Figure 6.4: Inter-speaker variability and intra-speaker variability of each phoneme (**V**:Vowels, **N**: Nasals, **R**:/r/, **L**:/l/, **G**:Glides, **S**:Stops, **A**:Affricates, **F**:Fricatives, the horizontal dashed blue lines are for the averages). (a) inter-speaker variability Σ_{inter} ($trace(S_b)$), and (b) intra-speaker variability Σ_{intra} ($trace(S_w)$).

Table 6.3: The weighted averages of inter-speaker variability Σ_{inter} and intra-speaker variability Σ_{intra} for /r/, /l/ and other broad phonetic classes in the Buckeye database (Σ_{inter} and Σ_{intra} are computed by Equations 6.8 and 6.9, respectively).

	Vowels	Nasals	/l/	/r/	Glides	Stops	Affricates	Fricatives
Σ_{inter}	416	393	447	515	409	186	331	389
Σ_{intra}	1180	896	1141	1226	1119	1308	1796	1266

6.3.2.2 F-ratio based on each of the 31 MFB coefficients

Based on Equation 6.1, each of the 31 MFB coefficients was used to compute the F-ratio for each phoneme. The purpose of this computation is to see how different is the acoustic variability across different mel-frequency filterbanks. Figure 6.5 shows the F-ratios of each coefficient for /r/ and /l/, respectively. Figures 6.5a and 6.5c are for the female speakers, whereas Figures 6.5b and 6.5d are for the male speakers. It can be seen that /r/ and /l/ have the maximum F-ratios in the range of coefficients 17 to 25, which is between 2 and 4 kHz in frequency (in the region of F3-F5). The male speakers have the F-ratio peaks in a lower frequency range than the female speakers. This is presumably because the male speakers normally have a longer vocal tract length and, therefore, have lower formant values. Furthermore, similar results are observed for many other sounds such as vowels and nasals. It is indicated that the coefficients in high mel-frequency bands might have a better discriminative power than the coefficients in the low mel-frequency bands. This indication will be verified in the speaker identification experiments in Section 6.4.1.



(a) /r/ (female)

(b) /r/ (male)

(c) /l/ (female)

(d) /l/ (male)

Figure 6.5: Normalized F-ratio of each MFB coefficient in /r/ and /l/ (F-ratio is normalized by the largest F-ratio obtained in all the phonemes). (a) /r/ (female), (b) /r/ (male), (c) /l/ (female), and (d) /l/ (male).

6.4 Discriminative power

6.4.1 Speaker identification task

The speaker identification experiment was performed to study how good each phoneme can discriminate the speakers. This is an experiment in a close set, and there are 40 speakers in total. For each phoneme, 75% of the tokens were used in the training set, and 25% of the tokens were used in the test set. The acoustic feature set is the 31 MFB coefficients. The MIT Lincoln lab speaker recognition system was used for both the feature extraction and the speaker identification experiments (Reynolds et al., 2000). The statistical speaker model is a 512 Gaussian mixture model (GMM). A universal background model (UBM) was trained first. Each speaker's model was adapted from the UBM model. The speaker models for /ay/, /em/ and /zh/ can not be built due to the inadequate number of tokens in the database. In addition, the first 11 (from the 1st to the 11th) and the last 11 (from the 21st to the 31st) MFB coefficients have been used for the speaker identification experiments, respectively. The purpose was to compare the discriminative powers between the MFB coefficients in the low mel-frequency range and the MFB coefficients in the high mel-frequency range.

6.4.2 Results

Figure 6.6 shows the speaker identification result for each phoneme with the 31 MFB coefficients as the acoustic parameter. Phonemes /oy/, /em/ and /zh/ do not have identification results, because they do not have enough data for training the

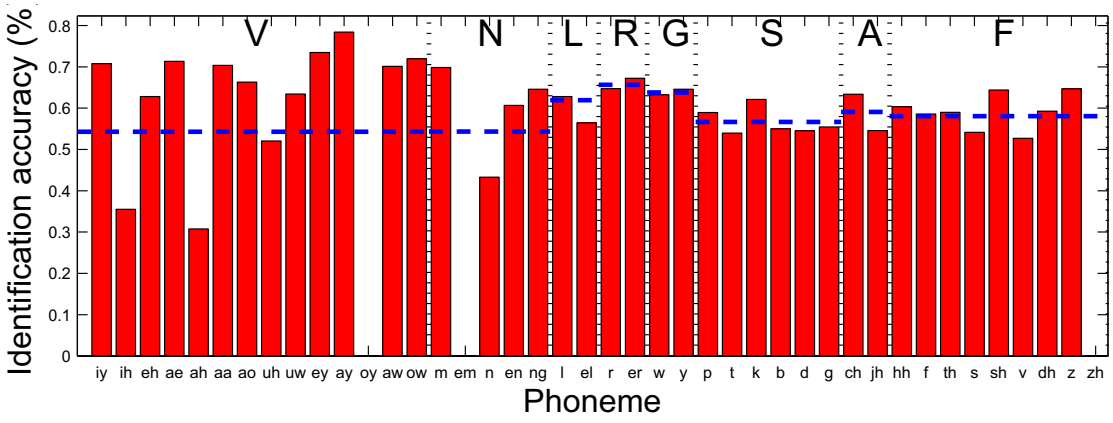


Figure 6.6: Speaker identification results based on the 31 MFB coefficients (**V**:Vowels, **N**: Nasals, **R**:/r/, **L**:/l/, **G**:Glides, **S**:Stops, **A**:Affricates, **F**:Fricatives, the horizontal dashed blue lines are for the weighted averages).

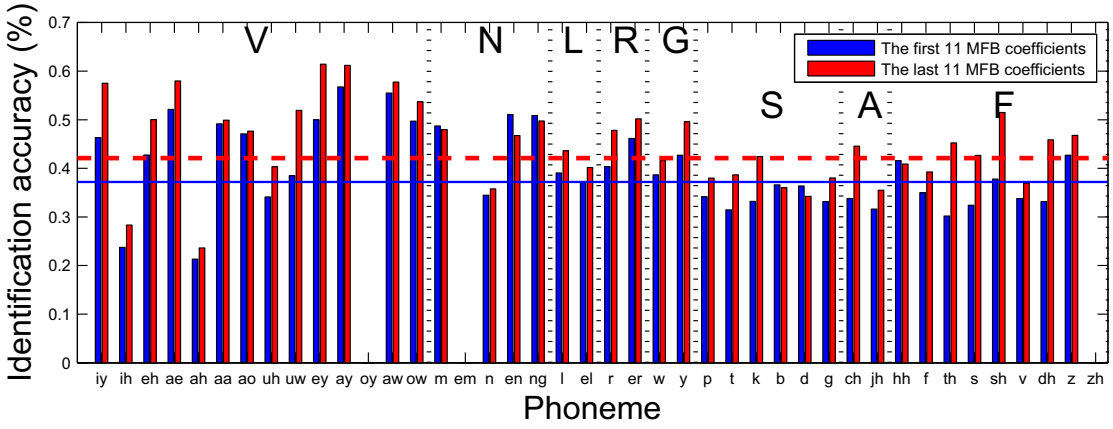


Figure 6.7: Comparison of speaker identification results for the first and the last 11 MFB coefficients (**V**:Vowels, **N**:Nasals, **R**:/r/, **L**:/l/, **G**:Glides, **S**:Stops, **A**:Affricates, **F**:Fricatives, the horizontal solid blue lines are for the weighted averages of the first 11 coefficients, the horizontal dashed red lines are for the weighted averages of the last 11 coefficients).

Table 6.4: The weighted averages of speaker identification accuracy for /r/, /l/ and other broad phonetic classes in the Buckeye database, with the 31 MFB coefficients as the acoustic parameter.

Vowels	Nasals	/l/	/r/	Glides	Stops	Affricates	Fricatives
54.2%	54.3%	61.9%	65.6%	63.8%	56.7%	59.1%	58.1%

Table 6.5: Comparison of the weighted averages of speaker identification accuracies between the first and the last 11 MFB coefficients for /r/, /l/ and other broad phonetic classes in the Buckeye database.

	Vowels	Nasals	/l/	/r/	Glides	Stops	Affricates	Fricatives
The first 11 MFB coef.	37.3%	41.4%	38.8%	42.5%	40.3%	33.9%	32.7%	35.3%
The last 11 MFB coef.	42.7%	41.4%	43.2%	48.7%	44.8%	38.2%	40.1%	43.3%

statistical speaker models in the speaker identification experiment. The positions of consonantal /r/, syllabic /r/, consonantal /l/ and syllabic /l/ in the identification accuracy ranking are 11th, 9th, 20th and 28th, respectively. /r/ performs better than /l/ in this experiment. The top ten best phonemes in speaker identification, in the descending order, are:

/ay/, /ey/, /ow/, /ae/, /iy/, /aa/, /aw/, /m/, /er/, and /ao/.

All of them are vowels or nasals, except syllabic /r/.

Table 6.4 lists the values of the weighted identification accuracy averages for the broad phonetic classes. The ranking in the descending order is:

/r/ > Glides > /l/ > Affricates >Fricatives > Stops > Nasals > Vowels

In addition, the speaker identification result for each phoneme with the first 11 or the last 11 MFB coefficients is shown in Figure 6.7. For most phonemes, the last 11 MFB coefficients produced a better identification accuracy than the first 11 coefficients. As shown in Figure 6.7, the weighted average identification accuracy for all the phonemes is 37.2% for the first 11 MFB coefficients, and 41.4% for the last 11 MFB coefficients. This confirms that the high mel-frequency band energies are more discriminative than the low mel-frequency band energies. Table 6.5 also lists the values of the weighted averages of identification accuracy for the broad phonetic classes.

6.5 Discussion

As shown in Section 6.3.2.1, /r/ and /l/ have very large inter-speaker variability. /r/ has a larger inter-speaker variability than any phoneme in vowels, nasals and glides. It is only smaller than the inter-speaker variability in /sh/, /s/ and /zh/. In addition, on average, /r/ and /l/ have larger inter-speaker variability than any other broad phonetic class such as vowels, glides, nasals, fricatives, affricates and stops. This indicates that the articulatory variability of liquids (particularly /r/) in American English may play an important role in increasing their inter-speaker variability. However, this articulatory variability may also affect the intra-speaker variability. As shown in Section 6.3.2.1, on average, /r/ has a larger intra-speaker variability than vowels, nasals, and glides, whereas /l/ has a larger variability than

nasals and glides. In average, the F-ratios of /r/ and /l/ are larger than the F-ratios of glides, fricatives, affricates and stops, but smaller than the F-ratio of nasals. In average, /r/ and /l/ have larger F-ratio than vowel if it is computed by $trace(S_w^{-1}S_b)$, and have smaller F-ratio than vowel if it is computed by $trace(S_b)/trace(S_w)$. In addition, /r/ has a larger F-ratio average than /l/.

Intuitively, a larger F-ratio means a better ability of discriminating speakers. However, our results showed that the F-ratio measure of each phoneme might not be a consistent indicator of its discriminative power in the speaker recognition experiment. The cross-correlation coefficient between the F-ratio and the identification accuracy is 0.04 for the F-ratio computed by $trace(S_w^{-1}S_b)$, and is 0.13 for the F-ratio computed by $trace(S_b)/trace(S_w)$. This discrepancy is probably caused by the different assumptions for the ANOVA analysis and the speaker identification experiment. In ANOVA analysis, the acoustic feature is assumed to have a Gaussian distribution, whereas in the speaker identification experiments, the acoustic feature is assumed to have a mixture of multiple Gaussian distributions.

It has been shown that all top ten best phonemes in speaker identification are vowels or nasal, except syllabic /r/. However, the ranking of the average identification accuracy for broad phonetic classes, in the descending order, is:

/r/ > Glides > /l/ > Affricates >Fricatives > Stops > Nasals > Vowels

where /r/and /l/ rank 1st and 3rd, respectively. Nasals and vowels have the smallest identification accuracies on average. This ranking is very different from the results in previous studies (Antal and Toderean, 2006; Eatock and Mason, 1994) where nasals, vowels and fricatives performed better than semivowels. This differ-

ence might be caused by the different databases, different classifiers, or different acoustic parameters between them. The speaker identification system used in this study is not optimized. It used the same parameters for all the phonemes regardless of their different durations in the database. The reason for the low average accuracy in vowels and nasals is because the phonemes /ih/, /ah/ and /n/ did not perform well in the speaker identification experiment. But these three phonemes comprise a large percentage of the tokens. The vowel /ih/ accounts for 19.5% of the vowel tokens, /ah/ accounts for 21.8% of the vowel tokens, and /n/ accounts for 52.5% of the nasal tokens. The average is weighted by the token number of each phoneme. This is why the averages of the identification accuracy in vowels and nasals are so low, even though other phonemes in the same broad class perform well in the experiment. However, the reason why these phonemes did not perform well is still unclear and further study is needed. If these three phonemes are excluded from the calculation of the average, the accuracy average will be 69.3% for vowels and 67.4% for nasals. The ranking will be changed to :

Vowels > Nasals > /r/ > Glides > /l/ > Affricates >Fricatives > Stops

6.6 Chapter summary

This chapter presented a preliminary study on the acoustic variability and the discriminative power of liquids along with other sounds. It was based on the Buckeye database, and acoustic parameters consisting of 31 MFB coefficients were extracted. ANOVA analysis showed that the inter-speaker variability of /r/ is larger

than any phoneme in vowels, nasals and glides. It is only smaller than /sh/, /s/ and /zh/. On average, /r/ and /l/ have larger inter-speaker variability than any other broad phonetic class such as vowels, glides, nasals, fricatives, affricates and stops. These results indicate that the variety of the articulatory configurations of liquids may increase the inter-speaker variability. On average, the F-ratios of /r/ and /l/ are larger than glides, fricatives, affricates and stops, but smaller than nasals. The speaker identification experiments showed that the ranking of the discriminative power of /r/, /l/ and other broad phonetic classes is: /r/ > Glides > /l/ > Affricates > Fricatives > Stops > Nasals > Vowels.

Chapter 7

Summary and future work

7.1 Summary

There are two goals in this dissertation. First, we wanted to better understand the acoustics and articulation of the liquid sounds in American English. In particular, we wanted to understand how to model typical articulatory configurations for /r/ and /l/, and we wanted to understand the major articulatory and acoustical differences among them. Second, we wanted to study the acoustic variability and the speaker discriminative power of the liquids, i.e., to study if the variability in articulation across speakers results in the liquid sounds having more inter-speaker acoustic variability and, thereby, more discriminative power in speaker recognition relative to other sounds.

In Chapter 4, a "retroflex" /r/ and a "bunched" /r/, produced respectively by two subjects (S1 and S2), have been studied. The retroflex /r/ and the bunched /r/ show similar patterns of F1-F3 but very different spacing between F4 and F5. Based on magnetic resonance images (MRI) of the vocal tract for sustained /r/ productions, 3-D finite element analysis has been performed to study the acoustic responses and the wave propagations inside the vocal tracts. Area functions were extracted based on the wave propagation property. The results of computer vocal tract models were compared to actual speech recordings. In particular, formant

cavity affiliations were explored using formant sensitivity functions and vocal tract simple-tube models. While both /r/s are produced with constrictions in the palatal, pharyngeal and laryngeal regions, there is a much larger difference in areas between the constricted and unconstricted regions for the retroflex /r/ than for the bunched /r/. This is because the palatal constriction in the retroflex /r/ is formed by the raised tongue tip, whereas the palatal constriction in the bunched /r/ is formed by the raised tongue dorsum. In both cases, F2 is produced by the front cavity, which consists of a lip constriction and a large volume posterior to the lip constriction. For the retroflex /r/, the palatal constriction decouples the vocal tract, and F1, F3, F4 and F5 are mainly produced by the back cavity posterior to the palatal constriction. However, in the bunched /r/, it is difficult to decouple the vocal tract due to the gradual changing of the area function around the palatal constriction. It is suggested that the F4/F5 differences between the variants can be explained largely by differences in whether the long cavity behind the palatal constriction acts as a half- or a quarter-wavelength resonator. For both S1's retroflex /r/ and S2's bunched /r/, F4 and F5 (along with F3) come from the long back cavity. However, for S1, these formants are half wavelength resonances, while for S2, these formants are quarter wavelength resonances of the cavity. Additionally, the finding of an F4/F5 difference in pattern is replicated in the acoustic data from an additional set of four subjects, two with bunched and two with retroflex tongue shapes for /r/. These results suggest that acoustic cues based on F4-F5 spacing may be robust and reliable indicators of tongue shape, at least for the classic (tongue tip down) bunched and (tongue dorsum down) retroflex shapes discussed here.

In Chapter 5, two tongue shapes of one subject (S2) for the /l/ production have been studied. One is for producing a sustained dark /l/, and the other is for producing a sustained light /l/. While both have a linguo-alveolar contact and two lateral channels, they differ in the axial length of the linguo-alveolar contact. In addition, due to the raised tongue dorsum, there are linguopalatal contacts in the light /l/. Both the dark /l/ and the light /l/ have similar patterns in F1-F3, but differ in the number of zeros in the spectrum and the frequencies of zeros. Using finite element analysis based on magnetic resonance images of the vocal tract for sustained productions, the acoustic effects of the lateral channels and the supralingual space have been investigated, and proper area function vocal tract models have been suggested for both cases. For the dark /l/, the zero is produced by the cross mode posterior to the linguo-alveolar contact. For the light /l/, the zeros are produced by the asymmetrical channels, the supralingual cavity as a side branch and the cross mode posterior to the linguo-alveolar contact. Two simple vocal tract models have been simulated to show the effect of lateral channels and the linguo-alveolar contact. The results showed that lateral channels with 1-2 cm long are not able to produce a zero in the region of F3-F5. In order to get a zero in the region of F3-F5, the lateral channels have to be asymmetrical and long enough (3-6 cm). In addition, a narrow constriction or a complete closure is also required. The articulation variability of /l/ production causes the zeros appear at different frequencies, which leads to the complexity of /l/ spectrum.

In Chapter 6, the acoustic variability and the discriminative power of liquids along with other sounds has been studied preliminarily. It was based on the Buckeye

database, and acoustic parameter 31 MFB coefficients were extracted. ANOVA analysis showed that the inter-speaker variability of /r/ is larger than any phoneme in vowels, nasals and glides. It is only smaller than /sh/, /s/ and /zh/. In average, /r/ and /l/ have larger inter-speaker variability than any other broad phonetic classes such as vowels, glides, nasals, fricatives, affricates and stops. These results indicate that the variety of the articulatory configurations of liquids may increase the inter-speaker variability. In average, the F-ratios of /r/ and /l/ are larger than glides, fricatives, affricates and stops, but smaller than nasals. The speaker identification experiments showed that the ranking of the discriminative power of /r/, /l/ and other broad phonetic classes is: /r/ > Glides > /l/ > Affricates > Fricatives > Stops > Nasals > Vowels.

7.2 Future work

There are several topics which can be extended from this dissertation in future. These topics might be in other research areas in speech and are not limited to the study of liquid sounds /r/ and /l/.

1. **Vocal tract modeling:** Even though the vocal tract modeling of liquids presented in this dissertation have made contributions to the knowledge of the /r/ and /l/ productions, there are still many intermediate tongue shapes for /r/ and /l/ in the UC database (Tiede et al., 2004). These tongue shapes are shown in Figures 3.1 (page 32) and 3.2 (page 33), respectively. Applying the same methodologies described in this dissertation, more articulatory configu-

rations of /r/ and /l/ can be studied thoroughly to understand the acoustic effects of this wide variety of tongue shapes.

The data acquisition procedures for MR images and MR acoustic data can be improved for a better quality. The MRI data can be scanned with a better image resolution such as 0.5 mm x 0.5 mm in-plane resolution and 2 mm slice thickness, as done by Kitamura et al. (2006). With higher image quality, the detailed structures such as the laryngeal cavity and the piriform sinuses can be reconstructed with better accuracy. In addition, recording techniques with a noise cancelation feature should be applied for the capturing of MR acoustic data (Narayanan et al., 2004; NessAiver et al., 2006).

2. **3-D tongue model:** It is advantageous to have a 3-D tongue model to simulate the tongue deformation. The tongue model can be integrated with other vocal tract anatomies for simulation of the vocal tract acoustics at different articulatory configurations, mainly at different tongue shapes (Badin and Serrurier, 2006; Dang and Honda, 2004; Engwall, 2003; Gerard et al., 2003; Stone, 1990; WilhelmsTricarico, 1995, 1996). The acoustic effect of the tongue shape can be investigated through the 3-D finite element analysis. In this way, the acoustic effect of the tongue shape will be isolated from the effect caused by the anatomy differences across different speakers.

3. **Vocal tract dynamics based on dynamic MR imaging:** All the MR data used in this dissertation were from static vocal tract shapes for sustained sounds. In order to study how the coarticulation in different contexts influ-

ences the production of liquids, dynamic MR imaging technology can be applied to record the dynamics of the vocal tract (Bresch et al., 2008; Narayanan et al., 2004; Takemoto et al., 2006b).

4. **Automatic segmentation of the 3-D vocal tract:** The segmentation of the 3-D vocal tract geometry from MR images in this dissertation was semi-manually done. This procedure took a lot of human effort. It is desirable to perform automatic segmentation with least possible human and computational work. Applying some automatic segmentation or registration methods on MR images may help get a 3-D vocal tract reconstruction with less manual effort (Vinitski et al., 1995; Zhukov et al., 2002).

5. **Superresolution image processing:** Techniques of superresolution image processing can be applied to get a more accurate reconstruction of the 3-D vocal tract geometry. Usually, the MR data includes three sets of images collected at axial, coronal and sagittal orientations. Better resolution in reconstruction might be achievable by combining slices from different orientations (Carmi et al., 2006; Wood et al., 2006).

6. **3-D speech synthesis:** Speech synthesis based on a 3-D vocal tract model might produce more natural sound because the sound is synthesized by a more realistic vocal tract model. 2-D digital waveguide filter has been used for implementation of a speech synthesizer (Mullen et al., 2006, 2007; Murphy et al., 2007). It has been shown that the 2-D digital waveguide filter has more flexibility in controlling the formant bandwidth. Based on a 3-D vocal tract

reconstruction, a 3-D digital waveguide filter can be implemented for speech synthesis.

7. **Phoneme discriminative power analysis:** This dissertation considered only one type of acoustic parameter. In order to see how the discriminative power of each phoneme varies with different parameters, other acoustic parameters such as the mel-frequency cepstrum coefficients (MFCCs), linear prediction coefficients (LPCs), linear prediction cepstral coefficients (LPCCs), reflection coefficients (RC), and wavelet-based feature (Farooq and Datta, 2003) should be tested in the speaker identification experiment.

Appendix A

Symbols used for American English consonants, by traditional articulatory categories

Table A.1: Symbols used for consonants of American English (Kent and Read, 2002)

	Bilabial	Labio-dental	Inter-dental	Alveolar	Retroflex	Alveo-palatal	Velar	Glottal
Stop	p b			t d			k g	
Fricative		f v	θ ð	s z		ʃ ʒ		h
Affricative						ʧ ʤ		
Nasal	m			n			ŋ	
Liquid				**l**	**r**			
Glide	w					j		

Appendix B

TIMIT and IPA labels

Table B.1: Table of TIMIT and IPA labels

TIMIT	IPA	Example	TIMIT	IPA	Example	Vowel Properties
p	p	pea	iy	i	beet	high front tense
b	b	bee	ih	ɪ	bit	high front lax
t	t	tea	eh	ɛ	bet	middle front lax
d	d	day	ey	e	bait	middle front tense
k	k	key	ae	æ	bat	low front lax
g	ɡ	gay	aa	ɑ	bott	low back lax
dx	ɾ	muddy	aw	aʊ	bout	low central lax
q	ʔ	bat	ay	aɪ	bite	low central tense dip
jh	ʤ	joke	ah	ʌ	but	
ch	ʧ	choke	ao	ɔ	bought	middle back lax rnd
f	f	fin	oy	ɔɪ	boy	middle back tense rnd dip
v	v	van	ow	o	boat	middle back tense rnd
th	θ	thin	uh	ʊ	book	high back lax rnd
dh	ð	then	uw	u	boot	high back tense rnd
s	s	sea	ux	ü	toot	

z	z	zone	er	ɝ	bird	high central lax rcld (str)
sh	ʃ	she	ax	ə	about	middle central lax (unstr)
zh	ʒ	azure	ix	ɨ	debit	
m	m	mom	axr	ɚ	butter	high central lax rcld (unstr)
n	n	noon	ax-h	ə^{h}	suspect	
ng	ŋ	sing				
em	m̩	bottom				
en	n̩	button				
eng	ŋ̩	washington				
nx	ɾ̃	winner				
l	l	lay				
r	r	ray				
w	w	way				
y	j	yacht				
hh	h	hay				
hv	ɦ	ahead				
el	l̩	bottle				

rnd: rounded, rcld: r-colored, str: stressed, unstr: unstressed, dip: diphthong

Bibliography

Allen, W. S., 1965. Vox Latina, a guide to the pronunciation of classical Latin. University Press, Cambridge.

Alwan, A., Narayanan, S., Haker, K., 1997. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. the rhotics. Journal of the Acoustical Society of America 101 (2), 1078–1089.

Antal, M., Toderean, G., 2006. Speaker recognition and broad phonetic groups. In: Proceedings of the 24th IASTED international conference on signal processing, pattern recognition, and applications. pp. 155–159.

Auckenthaler, R., Parris, E. S., Carey, M. J., 1999. Improving a GMM speaker verification system by phonetic weighting. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 313–316.

Badin, P., Serrurier, A., 2006. Three-dimensional linear modeling of tongue: articulatory data and models. In: Proceedings of the 7th international seminar on speech production, ISSP7 (H.C. Yehia, D. Demolin & R. Laboissire, Eds.). pp. 395–402.

Baer, T., Gore, J. C., Gracco, L. C., Nye, P. W., 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging - vowels. Journal of the Acoustical Society of America 90 (2), 799–828.

Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., Narayanan, S., May 2008. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [exploratory dsp]. Signal Processing Magazine, IEEE 25 (3), 123–132.

Burnett, D. S., 1988. Finite element analysis: from concepts to applications. Addison-Wesley Pub. Co., Reading, Massachusetts.

Carmi, E., Liu, S., Alon, N., Fiat, A., Fiat, D., 2006. Resolution enhancement in MRI. Magnetic resonance imaging 24 (2), 133–154.

Chiba, T., Kajiyama, M., 1941. The vowel: its nature and structure. Tokyo-Kaiseikan, Tokyo.

Comsol, 2007. (COMSOL Multiphysics), http://www.comsol.com, accessed 07/10/2007.

Creaghead, N., Newman, P., Secord, w., 1989. Assessment and remediation of articulatory and phonological disorders. Allyn & Bacon, Newton, MA.

Dalston, R. M., 1975. Acoustic characteristics of English /w,r,l/ spoken correctly by young children and adults. Journal of the Acoustical Society of America 57 (2), 462–469.

Dang, J. W., Honda, K., 1997. Acoustic characteristics of the piriform fossa in models and humans. Journal of the Acoustical Society of America 101 (1), 456–465.

Dang, J. W., Honda, K., 2004. Construction and control of a physiological articulatory model. Journal of the Acoustical Society of America 115 (2), 853–870.

Delattre, P., Freeman, D. C., 1968. A dialect study of American English r's by X-ray motion picture. Linguistics 44, 28–69.

Deller, J. R., Hansen, J. H. L., Proakis, J. G., 2000. Discrete-time processing of speech signals. Institute of Electrical and Electronics Engineers, New York.

Eatock, J. P., Mason, J. S., 1994. A quantitative assessment of the relative speaker discrimination properties of phonemes. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 133–137.

Engwall, O., 2003. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. Speech Communication 41 (2-3), 303–329.

Espy-Wilson, C. Y., 1987. An acoustic-phonetic approach to speech recognition: application to the semivowels. Ph.D. thesis, MIT.

Espy-Wilson, C. Y., 1992. Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. Journal of the Acoustical Society of America 92 (2), 736–757.

Espy-Wilson, C. Y., 2004. Articulatory strategies, speech acoustics and variability. In: Proceedings of sound to sense: 50+ years of discoveries in speech communication. MIT, Cambridge.

Espy-Wilson, C. Y., Boyce, S. E., 1999. The relevance of F4 in distinguishing between different articulatory configurations of American English /r/. Journal of the Acoustical Society of America 105 (2), 1400.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., Alwan, A., 2000. Acoustic modeling of American English /r/. Journal of the Acoustical Society of America 108 (1), 343–356.

Faltlhauser, R., Ruske, G., 2001. Improving speaker recognition using phonetically structured gaussian mixture models. In: EUROSPEECH. Scandinavia.

Fant, G., 1970. Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations, 2nd Edition. Description and analysis of contemporary standard Russian. Mouton, The Hague.

Fant, G., Pauli, S., 1974. Spatial characteristics of vocal tract resonance modes. In: Proceedings of the speech communication seminar 74. Stockholm, Sweden, August 1-3, pp. 121–132.

Farooq, O., Datta, S., 2003. Phoneme recognition using wavelet based features. Information Sciences 150 (1-2), 5–15.

Flanagan, J. L., 1972. Speech analysis: synthesis and perception, 2nd Edition. Kommunikation und Kybernetik in Einzeldarstellungen. Springer-Verlag, New York.

Gerard, J.-M., Wilhelms-Tricarico, R., Perrier, P., Payan, Y., 2003. A 3-D dynamical biomechanical tongue model to study speech motor control. Research Developments in Biomechanics 1, 49.

Giles, S. B., Moll, K. L., 1975. Cineflurographic study of selected allophones of English /l/. Phonetica 31, 206–227.

Goldstein, U. G., 1976. Speaker-identifying features based on formant tracks. Journal of the Acoustical Society of America 59 (1), 176–182.

Greenberg, S., 1997. The Switchboard transcription project in research report 24, 1996 large vocabulary continuous speech recognition summer research workshop technical report Series. Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., Perkell, J. S., 1999. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. Journal of the Acoustical Society of America 105 (5), 2854–2865.

Hagiwara, R., 1995. Acoustic realizations of American /r/ as produced by women and men. UCLA Working Papers in Phonetics 90, 1–187.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., Tianyu, W., 2005. Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 213–216.

Hayakawa, S., Itakura, F., 1994. Text-dependent speaker recognition using the information in the higher frequency band. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 137–140.

Hebert, M., Heck, L. P., 2003. Phonetic class-based speaker verification. In: EUROSPEECH. pp. 1665–1668.

Heinz, J. M., Stevens, K. N., 1964. On the derivation of area functions and acoustic spectra from cinradiographic films of speech. Journal of the Acoustical Society of America 36 (5), 1037–1038.

Heuvel, H. v. d., Rietveld, T., 1992. Speaker related variability in cepstral representations of dutch speech segments. In: ICSLP. pp. 1581–1584.

Johnson, K., 2003. Acoustic and auditory phonetics, 2nd Edition. Blackwell Publisher, Malden, Massachusetts.

Juneja, A., Espy-Wilson, C. Y., 2008. Probabilistic landmark detection for automatic speech recognition using acoustic-phonetic information. Journal of the Acoustical Society of America 123 (2), 1154–1168.

Kajarekar, S., 2002. Analysis of variability in speech with applications to speech and speaker recognition. Ph.D. thesis, OGI school of science and engieering.

Kajarekar, S., Hermansky, H., 2001. Speaker verification based on broad phonetic categories. In: A speaker Odyssey, The Speaker Recognition Workshop.

Kajarekar, S., Malayath, N., Hermansky, H., 1999. Analysis of speaker and channel variability in speech. In: Proceedings of Automatic Speech Recognition and Understanding (ASRU), Colorado.

Kent, R. A., Read, C., 2002. Acoustic analysis of speech, 2nd Edition. Thomson Learning, Albany, NY.

Kinga, S., Frankel, J., Livesc, K., McDermott, E., Richmon, K., Wester, M., 2006. Speech production knowledge in automatic speech recognition. Journal of the Acoustical Society of America 121 (2), 723–742.

Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., Honda, K., 2006. Cyclicity of laryngeal cavity resonance due to vocal fold vibration. Journal of the Acoustical Society of America 120 (4), 2239–2249.

Lee, K., 1999. Principles of CAD/CAM/CAE systems. Addison-Wesley, Reading, Massachusetts.

Lehiste, I., 1964. Acoustical characteristics of selected English consonants. Indiana University, Bloomington.

Lehman, M. E., Swartz, B., 2000. Electropalatographic and spectrographic descriptions of allophonic variants of /l/. Perceptual & Motor Skills 90 (1), 47–61.

Lin, Q., Jan, E. E., Che, C., Yuk, D. S., Flanagan, J. L., 1996. Selective use of the speech spectrum and a VQGMM method for speaker identification. In: ICSLP. Vol. 4. pp. 2415–2418.

Materialise, 2007. (Trial versions of Mimics and Magics), http://www.materialise.com, accessed 12/20/2007.

Matsuzaki, H., Miki, N., Ogawa, Y., 2000. 3-D finite element analysis of Japanese vowels in elliptic sound tube model. Electronics and Communications in Japan Part III-Fundamental Electronic Science 83 (4), 43–51.

Matsuzaki, H., Miki, N., Ogawa, Y., Matsuzaki, H., Miki, N., Ogawa, Y., 1996. FEM analysis of sound wave propagation in the vocal tract with 3-D radiational model. Journal of the Acoustical Society of Japan (E) 17 (3), 163–166.

Miki, N., Mastuzaki, H., Aoyama, K., Ogawa, Y., 1996. Transfer function of 3-D vocal tract model with higher mode. In: Proceedings of 4th Speech Production Seminar (Autrans). pp. 211-214.

Morse, P. M., Ingard, K. U., 1968. Theoretical acoustics. International series in pure and applied physics. McGraw-Hill, New York.

Motoki, K., 2002. Three-dimensional acoustic field in vocal-tract. Acoustical Science and Technology 23 (4), 207–212.

Mrayati, M., Carre, R., Guerin, B., 1988. Distinctive regions and modes - a new theory of speech production. Speech Communication 7 (3), 257–286.

Mullen, J., Howard, D. M., Murphy, D. T., 2006. Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality. IEEE Transactions on Audio Speech and Language Processing 14 (3), 964–971.

Mullen, J., Howard, D. M., Murphy, D. T., 2007. Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model. IEEE Transactions on Audio Speech and Language Processing 15 (2), 577–585.

Murphy, D., Kelloniemi, A., Mullen, J., Shelley, S., 2007. Acoustic modeling using the digital waveguide mesh. IEEE Signal Processing Magazine 24 (2), 55–66.

Narayanan, S., Alwan, A. A., Haker, K., 1997. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. the laterals. Journal of the Acoustical Society of America 101 (2), 1064–1077.

Narayanan, S., Byrd, D., Kaun, A., 1999. Geometry, kinematics, and acoustics of Tamil liquid consonants. Journal of the Acoustical Society of America 106 (4), 1993–2007.

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

NessAiver, M., Stone, M., Parthasarathy, V., Kahana, Y., Kots, A., Paritsky, A., 2006. Recording high quality speech during tagged Cine MRI studies using a fiber optic microphone. Journal of Magnetic Resonance Imaging 23, 92–97.

Nolan, F., 1983. The phonetic bases of speaker recognition. Cambridge studies in speech science and communication. Cambridge University Press, Cambridge; New York.

Ohala, J., 1985. Around flat. In: Fromkin, V. (Ed.), Phonetic linguistics: essays in honor of Peter Ladefoged. Academic Press, Orlando, pp. 223–241.

Ong, D., Stone, M., 1998. Three dimensional vocal tract shapes in /r/ and /l/: a study of MRI, ultrasound, electropalatography, and acoustics. Phonoscope 1 (1), 1–13.

Panchapagesan, S., 2003. Modeling the production of /l/ based on MRI data. Master's thesis, University of California, Los Angeles.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., Raymond, W., 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication 45 (1), 89–95.

Prahler, A., 1998. Analysis and synthesis of the American English lateral consonant. Master's thesis, MIT, Cambridge, Massachusetts.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10 (1-3), 19–41.

Roach, P., 2002. A little encyclopaedia of phonetics. www.personal.rdg.ac.uk/~llsroach/encyc.pdf.

Shriberg, L. D., Kent, R. D., 1982. Clinical phonetics. Macmillan, New York.

Sondhi, M. M., 1986. Resonances of a bent vocal-tract. Journal of the Acoustical Society of America 79 (4), 1113–1116.

Sproat, R., Fujimura, O., 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. Journal of Phonetics 21 (3), 291–311.

Stevens, K. N., 1998. Acoustic phonetics. MIT Press, Cambridge, Massachusetts.

Stone, M., 1990. A 3-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. Journal of the Acoustical Society of America 87 (5), 2207–2217.

Story, B. H., 2006. Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions (l). Journal of the Acoustical Society of America 119 (2), 715–718.

Story, B. H., Titze, I. R., Hoffman, E. A., 1996. Vocal tract area functions from magnetic resonance imaging. Journal of the Acoustical Society of America 100 (1), 537–554.

Sun, D. X., Li, D., 1995. Analysis of acoustic-phonetic variations in fluent speech using TIMIT. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 201–204.

Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., Honda, K., 2006a. Acoustic roles of the laryngeal cavity in vocal tract resonance. Journal of the Acoustical Society of America 120 (4), 2228–2238.

Takemoto, H., Honda, K., Masaki, S., Shimada, Y., Fujimoto, I., 2006b. Measurement of temporal changes in vocal tract area function from 3-D cine-MRI data. Journal of the Acoustical Society of America 119 (2), 1037–1049.

Theodoridis, S., Koutroumbas, K., 2003. Pattern recognition. Elsevier Science Publishers.

Thomas, T. J., 1986. A finite element model of fluid flow in the vocal tract. Computer Speech and Language 1, 131–151.

Tiede, M., Boyce, S. E., Holland, C., Chou, A., 2004. A new taxonomy of American English /r/ using MRI and ultrasound. Journal of the Acoustical Society of America 115 (5), 2633–2634.

TIMIT, 1990. TIMIT acoustic-phonetic continuous speech corpus, National Institute of Standards and Technology speech disc 1-1.1, October 1990.

Vinitski, S., Gonzalez, C., Burnett, C., Buchheit, W., Mohamed, F., Ortega, H., Faro, S., 1995. 3-D segmentation in MRI of brain tumors: preliminary results. Engineering in Medicine and Biology Society, 1995., IEEE 17th Annual Conference 1, 481–482.

Westbury, J. R., Hashi, M., Lindstrom, M. J., 1998. Differences among speakers in lingual articulation for American English /r/. Speech Communication 26 (3), 203–226.

WilhelmsTricarico, R., 1995. Physiological modeling of speech production - methods for modeling soft-tissue articulators. Journal of the Acoustical Society of America 97 (5), 3085–3098.

WilhelmsTricarico, R., 1996. A biomechanical and physiologically-based vocal tract model and its control. Journal of Phonetics 24 (1), 23–38.

Wood, S., Lan, H.-B., Christensen, M., Rajan, D., 2006. Edge detection performance in super-resolution image reconstruction from camera arrays. Digital Signal Processing Workshop, 12th - Signal Processing Education Workshop, 4th, 38–43.

Zawadzki, P. A., Kuehn, D. P., 1980. A cineradiographic study of static and dynamic aspects of American English /r/. Phonetica 37, 253–266.

Zhang, Z. Y., C., E.-W., Boyce, S., Tiede, M., 2005. Modeling of the front cavity and sublingual space in American English rhotic sounds. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 893–896.

Zhang, Z. Y., Espy-Wilson, C., Boyce, S., Tiede, M., 2003. Acoustic strategies for production of American English retroflex /r/. In: The 15th International Congress of Phonetic Sciences (ICPhS), Barcelona, Spain.

Zhang, Z. Y., Espy-Wilson, C. Y., 2004. A vocal-tract model of American English /l/. Journal of the Acoustical Society of America 115 (3), 1274–1280.

Zhou, X. H., Espy-Wilson, C., Boyce, S., Tiede, M., Holland, C., Choe, A., 2008. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. Journal of the Acoustical Society of America 123 (6), 4466–4481.

Zhou, X. H., Zhang, Z. Y., Espy-Wilson, C. Y., 2004. VTAR: A Matlab-based computer program for vocal tract acoustic modeling. Journal of the Acoustical Society of America 115 (5), 2543.

Zhukov, L., Bao, Z., Guskov, I., Wood, J., Breen, D., 2002. Dynamic deformable models for 3-D MRI heart segmentation. In: Proceedings of SPIE Medical Imaging 2002 Conference. pp. 1398–1405.