University of California

Los Angeles

On the role of subglottal acoustics in height estimation, and speech and speaker recognition

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Electrical Engineering

by

Harish Arsikere

2014

Abstract of the Dissertation

On the role of subglottal acoustics in height estimation, and speech and speaker recognition

by

Harish Arsikere

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2014

Professor Abeer Alwan, Chair

The subglottal system comprises the trachea, bronchi and their accompanying airways. Its configuration changes very little compared to that of the supraglottal vocal tract, as a result of which its acoustic properties are relatively more stationary and speaker specific. In this dissertation, our knowledge of subglottal acoustics—subglottal resonances (SGRs), most importantly—is leveraged to develop novel solutions to three problems that involve using or estimating speaker-specific characteristics: (1) body height estimation, (2) speaker normalization for automatic speech recognition (ASR), and (3) speaker identification (SID) and verification (SV). The focus is on scenarios where purely statistical methods may be sub-optimal owing to limited and/or noisy speech data.

Simultaneous recordings of speech and subglottal acoustics are collected (using a microphone and an *accelerometer*, respectively) from native American English speakers (50 adults and 43 children) and 6 adult bilingual speakers of Mexican Spanish (first language) and American English. The data are analyzed to understand the relationships between SGRs, and vocal-tract resonances (formants), body height and native language. Results indicate that (1) phonological vowel features (tongue *height* and *backness*) can be characterized via acoustic measures of formants and SGRs, (2) SGRs correlate well with body height, and (3) SGRs are practically independent of language and phonetic content. Based on these findings, algorithms are developed for the automatic estimation of SGRs from speech signals (i.e., without using accelerometer information). The algorithms are found to be effective for both adults and children, in quiet as well as noisy environments; their performance is equally good for native English and bilingual English/Spanish speakers, and does not degrade much with limited data.

Predictive models between body height and SGRs (in conjunction with SGR estimation algorithms) are used to develop an automatic approach to speech-based height estimation for adult speakers. The method is comparable in performance to existing data-driven techniques, but requires less training data, offers better generalization, and is more robust to noise. In the context of ASR for children, SGRs are used for speaker normalization via piece-wise linear frequency warping. On a digit-recognition task, the method achieves lower word error rates than conventional vocal-tract length normalization, in clean as well as noisy environments. The benefit is particularly significant for young speakers (6–8 years old) and short utterances (1 or 2 words). For SID and SV (with adults' speech), an algorithm is developed for deriving subglottal features that are more informative (than SGRs) with regard to speaker discriminability. When combined with Mel-frequency cepstral coefficients (conventional speech features for SID and SV), subglottal features provide significant performance improvements, especially for short test utterances (5–10 seconds in duration).

The dissertation of Harish Arsikere is approved.

Jody Kreiman

Paulo Tabuada

Abeer Alwan, Committee Chair

University of California, Los Angeles

2014

To my family and teachers.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible but for the support of many people. I would first like to thank my academic adviser, Prof. Abeer Alwan, for being a wonderful mentor over the years and for teaching me the value of novel, ethical research. I am ever indebted to her for the faith she had in me—will always remember how she stood by me when I was unable to clear the PhD prelim exam. It has been a privilege to work under her supervision. I would also like to thank my committee members, Prof. Jody Kreiman, Prof. Kung Yao and Prof. Paulo Tabuada, for their interest in my research.

Special thanks to Prof. Steven Lulich (at Indiana University, Bloomington) for introducing me to the field of subglottal acoustics. His knowledge, enthusiasm and accessible nature meant that I always had a place to discuss my ideas and concerns in detail. I am also grateful to Prof. Mitchell Sommers and John Morton (at Washington University, St. Louis) for their timely suggestions and collaboration over the years. Many thanks to John for his mammoth effort in collecting and annotating our databases.

I feel very fortunate to have worked with Dr. Elizabeth Shriberg. My research internship at Microsoft was a great learning experience, thanks to her expertise and insightfulness. Her mentorship has immensely impacted my thought process both as a person and as a researcher.

My heartfelt thanks to the administrative staff in the department (especially Deeona, Mildri, Mandy and Michelle) for their patience and help during my stay at UCLA. Life would have been a lot tougher without them.

I was blessed with wonderful friends in and outside the lab—Gary, Lee Ngee, Zaid, Shankar, Abde, Jom, Gang, Julien, Juan, Hitesh and Anirudh. I am thankful

for their company and support, and for the many enjoyable discussions (technical and otherwise) I have had with them. Sincere thanks to my former lab mates—Shizhen, Wei, Yen, Jonas and Eleanor—for helping me find my bearings initially. Thanks also to my other friends—elsewhere in the US and back in India—who have been there for me over the years.

Lastly, but very importantly, a million thanks to my family members (especially my mother and brother) for their unconditional love and encouragement. I am ever indebted to them for the sacrifices they have made to ensure that I could focus on my career goals and research. Many thanks to Preethi for her love, understanding and moral support during the many stressful weeks I spent towards the end of my stay at UCLA.

This work was supported in part by NSF Grant Number 0905381.

Vita

2003–2007	Bachelor of Engineering (B. E.) Telecommunication Engineering Department R. V. College of Engineering, Bangalore, India
2007	Undergraduate Intern Texas Instruments, Bangalore, India
2007–2009	Master of Technology (M. Tech.) Electrical Engineering Department Indian Institute of Technology, Kanpur, India
2007–2009	Teaching Assistant Electrical Engineering Department Indian Institute of Technology, Kanpur, India
2009–present	Graduate Student Researcher Electrical Engineering Department University of California, Los Angeles, USA
2010	Teaching Assistant Electrical Engineering Department University of California, Los Angeles, USA
2013	Summer Research Intern Microsoft Corporation, Sunnyvale, California, USA

Publications

Harish Arsikere, Steven M. Lulich and Abeer Alwan, “Estimating speaker height and subglottal resonances using MFCCs and GMMs,” IEEE Signal Processing Letters, Vol. 21, pp. 159–162, 2014.

Harish Arsikere, Gary K. F. Leung, Steven M. Lulich and Abeer Alwan, “Automatic estimation of the first three subglottal resonances from adults’ speech signals with application to speaker height estimation,” Speech Communication, Vol. 55, pp. 51–70, 2013.

Steven M. Lulich, John R. Morton, *Harish Arsikere*, Mitchell S. Sommers, Gary K. F. Leung and Abeer Alwan, “Subglottal resonances of adult male and female native speakers of American English,” Journal of the Acoustical Society of America, Vol. 132, pp. 2592–2602, 2012.

Steven M. Lulich, Abeer Alwan, *Harish Arsikere*, John R. Morton and Mitchell S. Sommers, “Resonances and wave propagation velocity in the subglottal airways,” Journal of the Acoustical Society of America, Vol. 130, Issue 4, pp. 2108–2115, 2011.

Harish Arsikere, Steven M. Lulich and Abeer Alwan, “Automatic estimation of the first subglottal resonance,” Journal of the Acoustical Society of America, Vol. 129, Issue 5, pp. 197–203, May 2011.

Harish Arsikere and Abeer Alwan, “Frequency warping using subglottal resonances: complementarity with VTLN and robustness to additive noise,” ICASSP 2014 (accepted).

Harish Arsikere, Steven M. Lulich and Abeer Alwan, “Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency,” Proceedings of ICASSP, pp. 7922–7926, 2013.

Harish Arsikere, Gary K. F. Leung, Steven M. Lulich and Abeer Alwan, “Automatic estimation of the first two subglottal resonances in children’s speech with application to speaker normalization in limited-data conditions,” Proceedings of Interspeech, 2012.

Harish Arsikere, Gary K. F. Leung, Steven M. Lulich and Abeer Alwan, “Automatic height estimation using the second subglottal resonance,” Proceedings of ICASSP, pp. 3989–3992, 2012.

Steven M. Lulich, *Harish Arsikere*, John R. Morton, Gary K. F. Leung, Abeer Alwan and Mitchell S. Sommers, “Analysis and automatic estimation of children’s subglottal resonances,” Proceedings of Interspeech, pp. 2817–2820, 2011.

Harish Arsikere, Steven M. Lulich and Abeer Alwan, “Automatic estimation of the second subglottal resonance from natural speech,” Proceedings of ICASSP, pp. 4616–4619, 2011.

CHAPTER 1

Introduction

1.1 Overview and motivation

Speech production is a complex physiological process involving several organs and a series of acoustic events. The speech-production system can be viewed as being composed of three subsystems: (1) the subglottal system, (2) the larynx, and (3) the supraglottal system (also known as the vocal tract); see Figure 1.1. The subglottal system comprises the trachea, bronchi and lungs, and is responsible for generating and driving the airflow required for speech production. The larynx, a structure made of cartilage and muscle, is located at the top of the trachea, and is the place where the airflow driven upwards by the subglottal system is modulated. The vocal folds (or vocal cords), a pair of membranous structures situated within the larynx, are responsible for converting the airflow to a *source signal* that acts as the excitation input to the vocal tract. The source signal can be either a quasi-periodic train of pulses or a noise-like excitation depending on whether or not the vocal folds vibrate. The opening between the vocal folds is known as the glottis; it acts as an acoustic link between the subglottal and supraglottal systems and its area determines the degree of acoustic coupling. The vocal tract is composed of the oral, nasal and pharyngeal cavities, and its configuration is determined by the positions of the tongue, jaws, teeth, lips, and other articulators. The source signal is spectrally shaped by the vocal tract and converted to a sound pressure

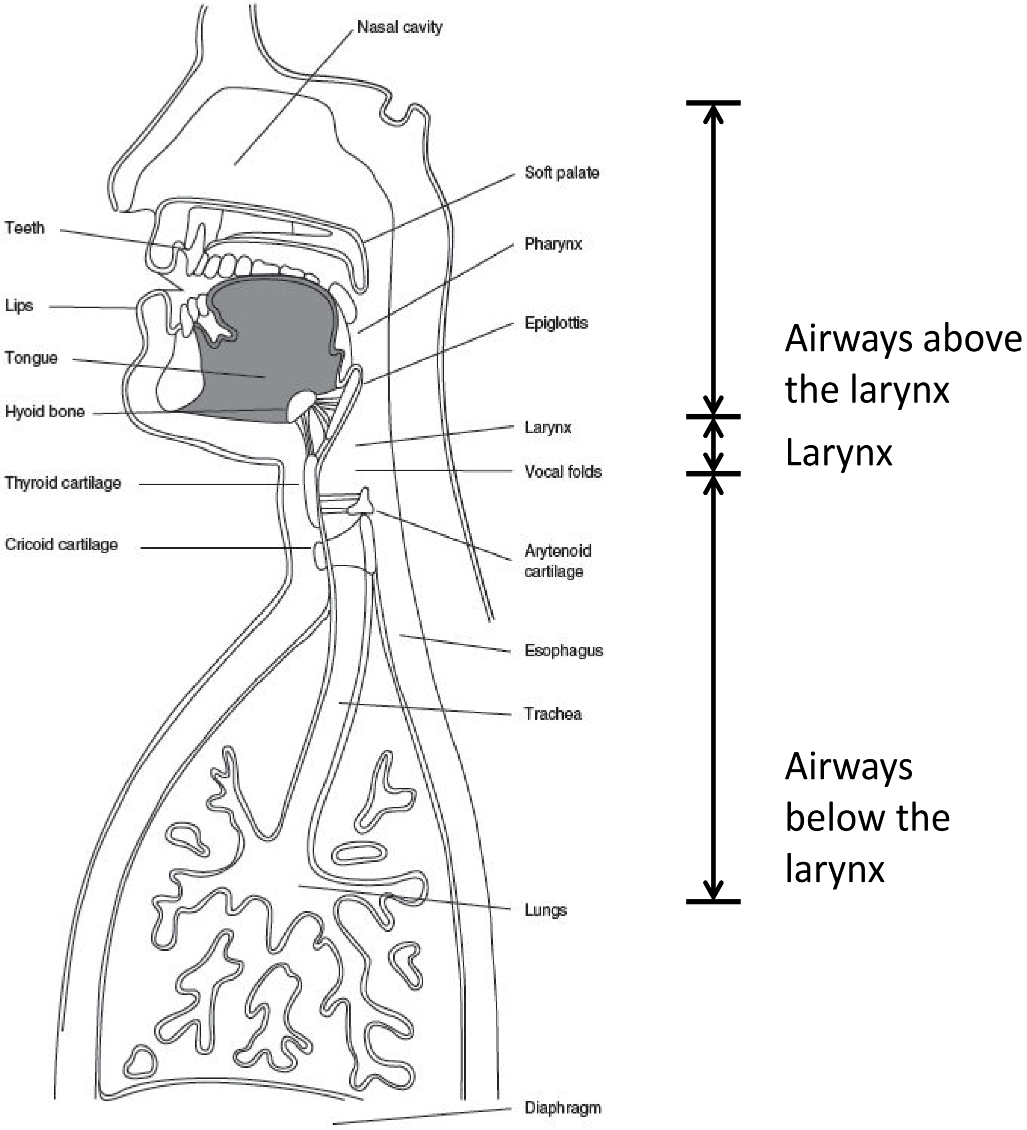


Figure 1.1: *A schematic representation of the three subsystems involved in speech production: the subglottal system (airways below the larynx), the larynx, and the supraglottal system (airways above the larynx); adapted from [KS11].*

waveform via radiation from the lips, thus resulting in the speech signal as we know it. The kind of sound produced depends on the nature of the source as well as the vocal-tract configuration.

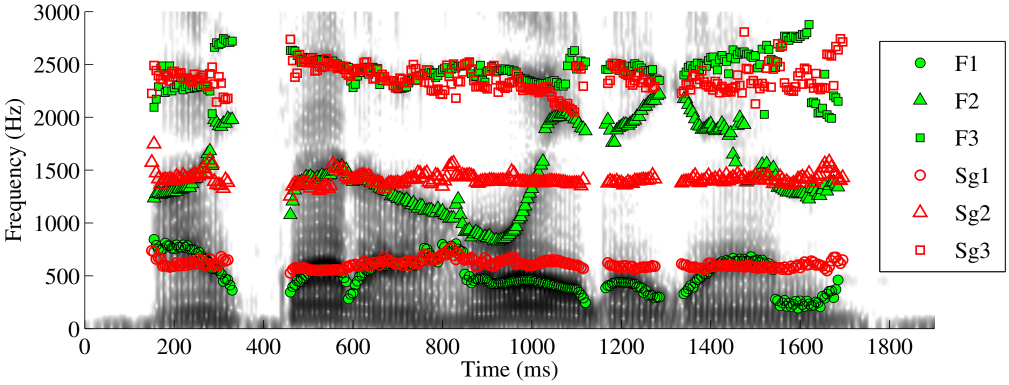


Figure 1.2: *Spectrogram of a speech signal (from a particular male speaker) superimposed with tracks of the first three vocal-tract formants, and tracks of the first three SGRs obtained from the corresponding time-synchronized accelerometer recording of subglottal acoustics. Clearly, formants vary much more than SGRs. $Sg3$ in this case was very weak compared to $Sg1$ and $Sg2$, and hence was tracked less accurately.*

Owing to absence of moving articulators and body parts, the acoustics of the subglottal system are much more stationary over time compared to the acoustics of the source and the vocal tract. Subglottal acoustics are therefore expected to characterize a speaker better, or at least provide information that is complementary to the characteristics of the source and the vocal tract. To substantiate the above statements, Figure 1.2 illustrates the difference between subglottal and supraglottal acoustics by comparing (for a particular male speaker) the first three subglottal resonances (SGRs), $Sg1$, $Sg2$ and $Sg3$, with the first three vocal-tract resonances (or formant frequencies), $F1$, $F2$ and $F3$. SGRs and their properties will be explained in detail later in Section 1.2.

The acoustic properties of the source and the vocal tract have been studied extensively in the past and have also found wide application is different areas of speech technology. Automatic methods have been developed to estimate parameters such as formant frequencies and the fundamental frequency ($F0$) of the source, and to extract other kinds of spectral, temporal and spectro-temporal features for tasks such as speech-activity detection, speech and speaker recognition, speech synthesis and compression, language identification, extraction of paralinguistic information, etc. In comparison, subglottal acoustics have found little application in main-stream speech technology, although there has been considerable effort to mathematically model the subglottal system and to understand its linguistic and phonological significance in the context of speech production and perception. The main goal of this dissertation, therefore, is to further our understanding of subglottal acoustics and to apply some of their properties to tasks that involve using or estimating speaker-specific characteristics.

The role of subglottal acoustics will be investigated in relevance to the following three problems: (1) body height estimation, (2) speaker normalization for automatic speech recognition (ASR), and (3) speaker identification (SID) and verification (SV). While these problems have been researched extensively in the past, the methods proposed thus far have been largely statistical in nature. Statistical methods provide good performance in general, but their efficacy tends to be lowered in limited-data conditions, and in scenarios where the test data are noisy and/or acoustically mismatched with the data used for training. Therefore, in this dissertation, the properties of subglottal acoustics are leveraged to develop hybrid knowledge-based and statistical solutions to the above three problems.

Given a speech signal (recorded using a microphone), information about the source and the vocal tract can be determined readily from it. However, to reliably determine the properties of subglottal acoustics, other modalities—invasive methods such as laryngectomy or noninvasive methods such as the use of an accelerometer—are necessary. Therefore, one of the challenges in using subglot-

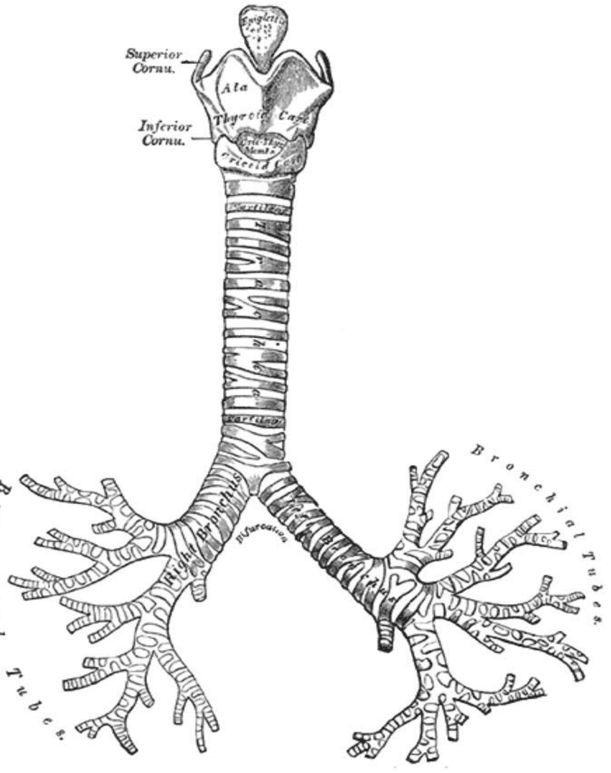


Figure 1.3: *The subglottal airways, including the trachea, main bronchi, and the bronchial tree down to about 6 generations (adapted from [Gra18]).*

tal acoustics for speech technology is to be able to extract subglottal information from the speech signal itself. Addressing this challenge is one of the major goals of this dissertation. Another goal is to collect a sizable corpus of time-synchronized speech and subglottal acoustics that would enable the development of techniques for meeting the above challenge.

1.2 The subglottal system and its resonances

Some of the material presented in this section is based on the studies by Lulich [Lul06, Lul10], and Chi and Sonderegger [CS07].

The subglottal airways comprise the trachea, the two main bronchi, and the rest of the bronchial tree. Each airway typically divides into two smaller airways, each of which divides further, and so on down to the level of the alveoli of the lungs, where gas exchange takes place during breathing. The trachea is usually referred to as the 0th generation of the tree, the main bronchi are referred to as the 1st generation, and so on (down to about 35 generations). Efficient gas exchange requires a large surface area, and in order to fit such a large area inside the limited volume of the chest cavity, there must be a large number of very narrow airways. After 6 or 7 generations of the bronchial tree, the airways become so narrow that acoustic fluctuations inside them turn out to be negligibly small, causing any more peripheral airways to be effectively cutoff from the subglottal acoustic system [Lul06]. Therefore, the terminal impedance below 6 or 7 generations has virtually no impact on the acoustics of the subglottal system (except at very low frequencies), and the airways can be modeled as being open at the periphery [Lul06]. Figure 1.3 shows the subglottal system down to about 6 generations.

A few mathematical models (based on electrical transmission-line theory) have been proposed to describe the acoustics of the subglottal airways [HKP01,HPK03, Lul06]. The models of Harper *et al.* [HKP01, HPK03] make the simplifying assumption that the bronchial tree branches symmetrically. The model proposed in [Lul06] does not make such an assumption and hence is more accurate, although it shares the typical assumption of binary branching. Figure 1.4 shows the subglottal input impedance obtained using the model proposed in [Lul06], for a certain combination of model parameters; SGRs are nothing but the natural frequencies (or poles) of the subglottal input impedance.

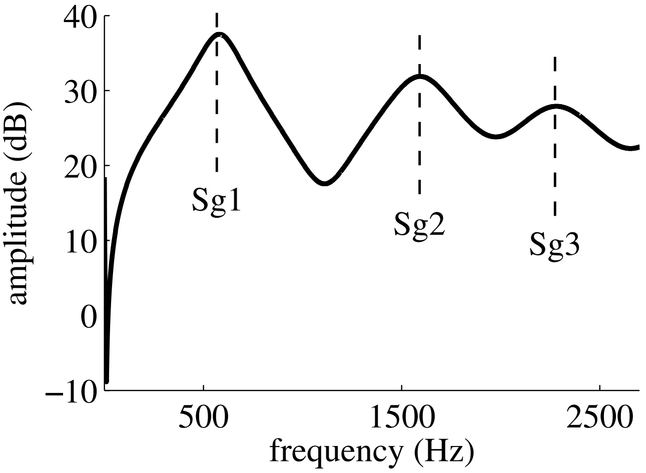


Figure 1.4: *Subglottal input impedance obtained using the mathematical model proposed in [Lul06], for a certain combination of model parameters. SGRs $Sg1$, $Sg2$ and $Sg3$ are the poles (natural frequencies) of the input impedance. Each pole is accompanied by a zero (or antiresonance).*

1.2.1 Coupling between the subglottal and supraglottal systems

The subglottal and supraglottal systems are acoustically coupled via the glottis when the vocal folds are open, and via the vocal-fold tissues themselves when the folds are closed. As a result of this coupling, each natural frequency of the uncoupled subglottal system contributes a pole-zero pair to the speech signal.

To understand the origin of zeros in the speech signal, consider the circuit model [CS07, Ste98] of the vocal tract-glottis-subglottal system, shown in Figure 1.5(a). Z_{sg} is the impedance of the subglottal system, Z_{vt} is the impedance of the vocal tract, Z_g is the glottal impedance, and $U_{s1} = U_{s2}$ are the twin (dipole) volume-velocity sources on either side of the vocal folds. If the glottal impedance is infinite (i.e., $Z_g = \infty$), Z_g is replaced by an open circuit and there are two equal and independent volume-velocity waves moving in opposite directions: one flowing into the vocal tract, and the other flowing into the subglottal system. On

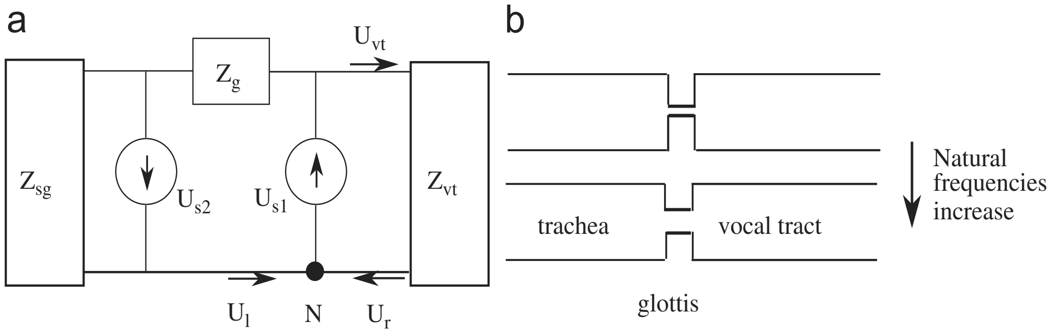


Figure 1.5: *(a) Circuit diagram of the speech production system, including the impedances of the vocal tract (Z_{vt}), glottis (Z_g), and subglottal airway (Z_{sg}), and the twin glottal volume velocity sources ($U_{s1} = U_{s2}$) due to vocal fold vibration. (b) Tube diagram of the speech production system, including the vocal tract, the glottis, and the subglottal airway (adapted from [Lul10]).*

the other hand, if the glottal impedance is finite, the volume velocity flowing into the vocal tract is affected by the subglottal impedance.

Consider the node labeled 'N'. The volume velocity flowing out of node N is U_{s1}. Volume velocity can flow into node N from the left, denoted by U_l, or from the right, denoted by U_r ($= U_{vt}$). From Kirchhoffs Current Law:

$$U_{s1} = U_l + U_r. \tag{1.1}$$

For frequencies at which the subglottal impedance is infinite (i.e. at the natural frequencies of the subglottal system), the left side of the circuit is open so that U_l $= U_{s2}$. Since $U_{s1} = U_{s2}$ and $U_r = U_{vt}$, Eq. (1.1) reduces to $U_s = U_s + U_{vt}$, or U_{vt} $= 0$. No volume velocity flows into the vocal tract, so that in the speech signal there is a zero at the same frequency as the frequency of a subglottal resonance.

To understand the origin of poles in the speech signal, consider two tubes coupled by means of a narrow constriction; see Figure 1.5(b). The constriction represents the glottis, while the tubes on the left and right represent the subglottal system and vocal tract, respectively. The poles of the speech signal are then at the natural frequencies of the combined system. If the glottal impedance is infinite so that there is no coupling between the two systems, the poles are simply the natural frequencies of the vocal tract. When there is some coupling due to a partially open glottis, the vocal-tract poles are shifted upward in frequency, and additional poles are introduced from the subglottal system. The subglottal poles are also shifted upward with respect to their uncoupled natural frequencies.

The subglottal system thus contributes zeros at its (uncoupled) natural frequencies and poles at slightly higher frequencies. For each natural frequency of the subglottal system, there is a pole-zero pair in the speech signal. This model applies to the subglottal-vocal tract system when acoustic coupling is achieved via the air column between the vocal folds. If the vocal folds are closed, it is possible that coupling could still be achieved across the vocal-fold tissues themselves. In this case, the zeros would still be at the (uncoupled) natural frequencies of the subglottal system, whereas the frequencies of the poles corresponding to the subglottal system and the vocal tract would not necessarily be the same as in the case of air-only coupling.

1.2.2 SGRs and phonological distinctive features

In the context of feature-based phonology (the science of speech sounds in regard to their distribution and pronunciation), understanding the connection between the acoustic properties of sounds and their phonological feature values has been of considerable importance. The quantal theory of speech production [Ste89] has been one of the most successful approaches to this problem, grounding the set of phonological distinctive features in the set of nonlinear articulation-to-acoustic

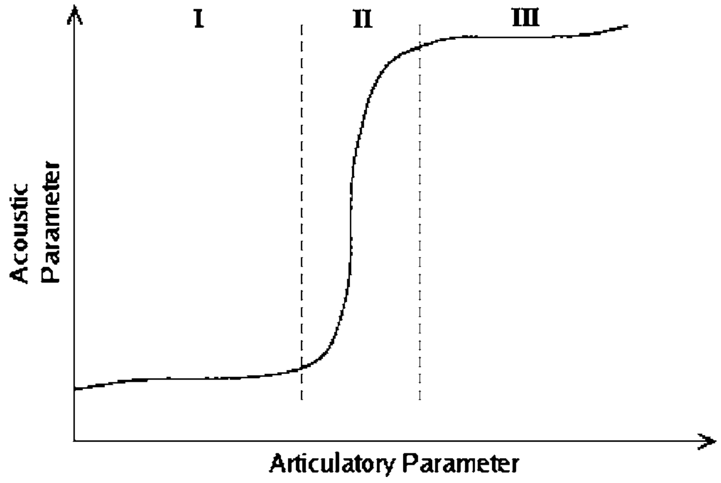


Figure 1.6: *Schematic of a quantal relation. Regions I and III are the 'states' corresponding to [+feature] and [-feature] values, and region II is the abrupt transition or 'boundary' between them (adapted from [CS07]).*

mappings. According to quantal theory, equal movements of speech articulators (the tongue body, for example) do not produce equal changes in the acoustic properties of speech sounds (formant frequencies, for example). In some regions of the articulator space, small movements lead to large acoustic changes, and these form what are known as landmarks [Ste02]; in other regions of the articulator space, large movements lead to small acoustic changes, and these define the stable regions which underlie distinctive features. Regions of the first type are called 'boundaries' and regions of the second type are called 'states'. A boundary and its two accompanying states define a single distinctive feature, where one state corresponds to the positive value of the feature and the other state corresponds to the negative value of the feature. The above ideas are summarized in Figure 1.6.

One set of boundaries and states arises from the acoustic coupling between the subglottal system and the vocal tract. As described in Section 1.2.1, the natural frequencies of the subglottal system introduce pole-zero pairs into the speech spectrum. Each pole-zero pair affects a relatively narrow band of frequencies in the spectrum, since the pole and zero largely cancel out (i.e., sum to zero, or close to zero) as the frequency increases or decreases away from the resonance. Thus, if a vocal-tract formant is far away from these narrow bands, it is not significantly affected by the pole-zero pairs; if a formant is near or within one of these narrow bands, however, its frequency and amplitude will be unstable.

It was first suggested in [Ste98] that such narrow, unstable regions might define acoustic boundaries between ± values of certain distinctive features. Later, several studies investigated the relationship between $Sg2$ (or more precisely, the unstable region due to $Sg2$'s pole-zero pair) and the distinctive feature [back]. In [Son04], $F2$ values of 53 languages (reported in the literature) were analyzed, and it was found that $Sg2$, on average, lies at the boundary between [-back] and [+back] vowels. It was also found that individual adult speakers of English tend to produce [-back] vowels with $F2$ higher than $Sg2$, and [+back] vowels with $F2$ lower than $Sg2$. This trend was subsequently observed in other languages such as High German and Swabian German [DLM11], Standard Korean [Jun09], and Standard Hungarian [CBG09,GLC11]. Analogously, a few studies [Jun09,GLC11] have shown that speakers tend to produce [+low] vowels with $F1$ higher than $Sg1$, and [-low] vowels with $F1$ lower than $Sg1$. Figure 1.7 illustrates these quantal relationships with an example. In the forthcoming chapters, the quantal nature of $Sg1$ and $Sg2$ will be used as the basis for automatic SGR estimation as well as speaker normalization for ASR.

1.2.3 Acoustic effects of SGRs on the speech signal

Coupling between the vocal tract and the subglottal system can usually be ignored while modeling the vocal-tract transfer function, but its effect becomes

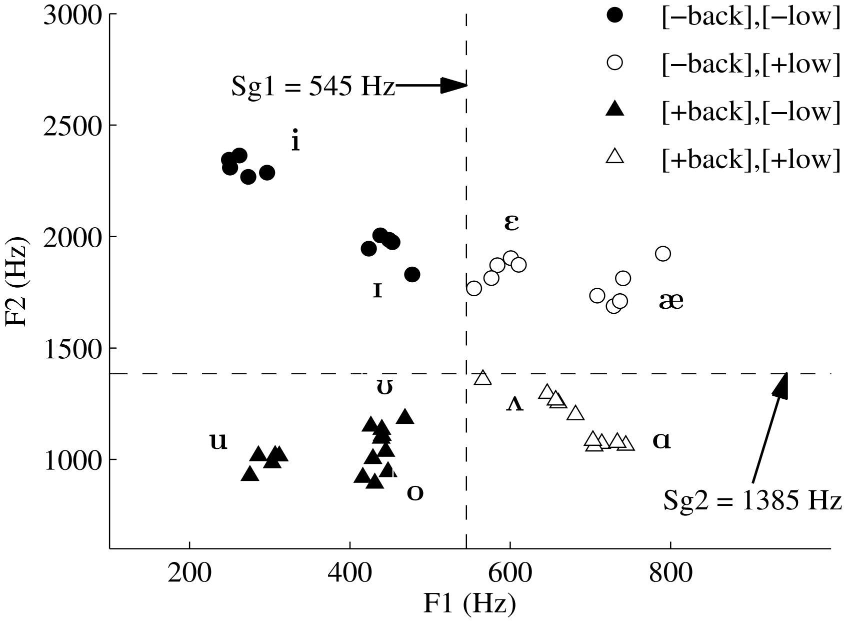


Figure 1.7: *Vowel space (of a male speaker) in the* $F1$*-*$F2$ *plane demonstrating the vowel-feature contrasts provided by* $Sg1$ *and* $Sg2$*.* $F1$ *is higher than* $Sg1$ *for [+low] vowels (empty symbols) and lower than* $Sg1$ *for [-low] vowels (filled symbols). Similarly,* $F2$ *is higher than* $Sg2$ *for [-back] vowels (circles) and lower than* $Sg2$ *for [+back] vowels (triangles).*

non-negligible when a vocal-tract formant approaches a subglottal resonance in frequency. Using a mathematical model of coupled resonators, Chi and Sonderegger [CS07] demonstrated that (1) formant amplitudes experience attenuation, and (2) formant frequencies appear to "jump", when vocal-tract formants lie in the vicinity of SGR-induced pole-zero pairs. Their model also suggested a positive correlation between the degree of subglottal coupling—which depends on the glottal area and impedance—and the magnitude of frequency-jump and amplitude-attenuation effects.

The frequency-jump effect (or discontinuity effect) can be understood by

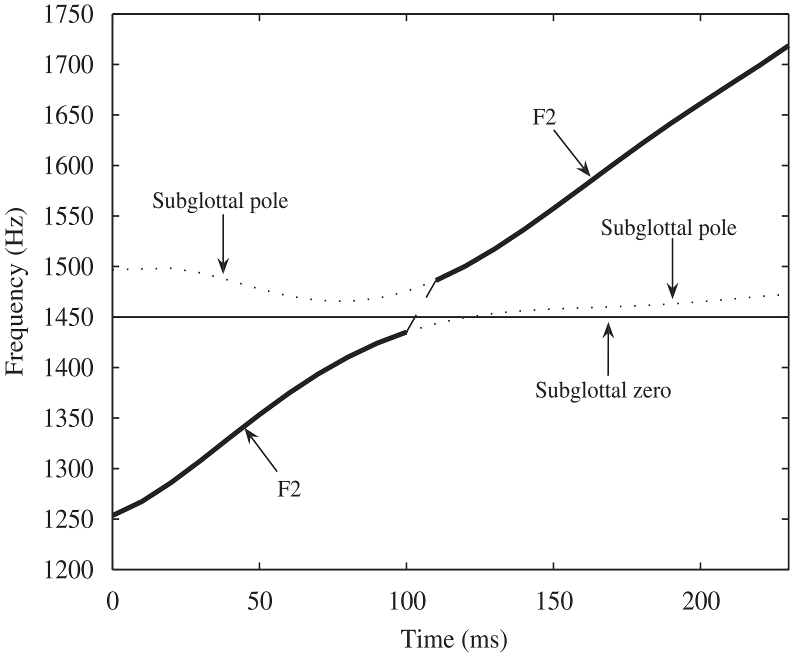


Figure 1.8: *Schematic of the $F2$ discontinuity due to $Sg2$. The horizontal line indicates the subglottal zero (from the pole-zero pair). The dotted line indicates the subglottal pole (from the pole-zero pair), and the thick solid line indicates $F2$. $F2$ and the subglottal pole swap affiliations at about 100 ms, giving rise to a discontinuity in the $F2$ trajectory (adapted from [Lul10]).*

studying, as an example, the interaction between $F2$ and $Sg2$. Figure 1.8 shows a schematic of the time-course of a rising $F2$ trajectory as it interacts with $Sg2$—there are two poles defined by the coupled vocal tract-subglottal system, and one zero defined by the subglottal system alone. Since the acoustics of the lower airway do not change much over time, the frequency of the zero is relatively constant. Early during the trajectory, the lower pole is defined by the vocal tract (it is 'affiliated' with the vocal tract), and this is the second formant. The higher pole is affiliated with the subglottal system, and this is the pole that is paired with the zero. Late in the trajectory, however, the higher pole is affiliated with the vocal tract and the lower pole is affiliated with the subglottal system. When the two poles are close together, they are affiliated more equally with both the vocal tract and the subglottal system. During this portion of the trajectory, the two poles are individually continuous, but their affiliations are not, thus giving rise to the perceived formant discontinuity.

Acoustic evidence of subglottal coupling has been found in back-to-front diphthongs such as [aɪ] and [ɔɪ], where $F2$ crosses $Sg2$, and in low-to-high diphthongs such as [aʊ], where $F1$ crosses $Sg1$ [CS07, Jun09, Lul10]. By careful selection of the analysis window length, shape and time spacing, it is often possible to discern frequency jumps and/or amplitude attenuations in trajectories of vocal-tract formants. An example of the coupling effects due to $Sg2$ is shown in Figure 1.9; note that the frequency of $Sg2$ as measured in the accelerometer signal (bottom-right panel) is identical to the frequency of the zero that is observed in the speech signal (top and bottom-left panels).

1.3 Automatic estimation of SGRs

SGRs can be measured noninvasively using accelerometer recordings of subglottal acoustics. When held against the skin of the neck at the location of the cricoid cartilage (which is inferior to the thyroid cartilage), an accelerometer captures the pressure fluctuations at the top of the trachea, thereby yielding a frequency spectrum whose peaks occur near the SGR frequencies. However, since the use of accelerometers in many real-life situations is unfeasible, it is important to be able to automatically *estimate* SGRs from speech signals.

Existing literature suggests two possible approaches to automatically estimating SGRs from speech signals: (1) *direct* estimation based on detecting the subtle

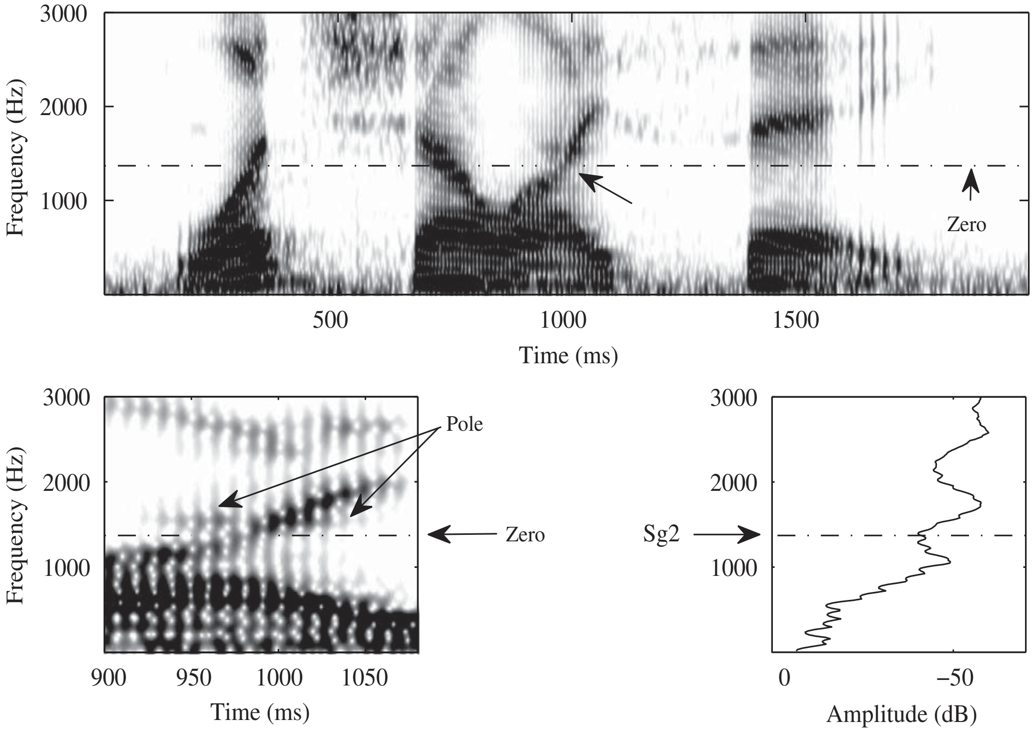


Figure 1.9: Top panel: *Spectrogram of the utterance "What shall I say?" The frequency of the subglottal zero is indicated by the horizontal dashed line. At about 1000 ms (as indicated by the arrow), a frequency discontinuity and amplitude attenuation can be observed in the second-formant trajectory; note that the segment around 1000 ms corresponds to the diphthong [aɪ].* Bottom-left panel: *An enlarged version of the spectrographic segment around 1000 ms. Note that the subglottal pole is clearly visible here.* Bottom-right panel: *Averaged spectrogram of the accelerometer signal. This figure has been adapted from [Lul10].*

effects of SGRs on vowel formants—frequency discontinuities and amplitude attenuations observed in the formant contours of back-to-front ([aɪ], [ɔɪ]) and low-to-high ([aʊ]) diphthongs [CS07, Jun09, Lul10]; and (2) *indirect* estimation based on the potential correlations between SGRs and formant frequencies (especially $F3$). Previous research efforts [WAL08, WLA08, WLA09a] have focused on $Sg2$

and $Sg3$ estimation, using a combination of both approaches.

In [WAL08], an automatic algorithm was proposed for estimating $Sg2$ and $Sg3$ in isolated American English (AE) vowels of adults as well as children. Estimation of $Sg2$ relied on detecting discontinuities (or jumps) in trajectories of $F2$. Such discontinuities are usually observed in back-to-front diphthongs ([aɪ] and [ɔɪ]) when $F2$ approaches and crosses $Sg2$ [CS07]. Given a vowel token, $F2$ was first tracked frame-by-frame using the SNACK toolkit [Sjo97]. Then, the track was inspected for frequency discontinuities by computing its smoothed first-order difference and comparing it with an empirically-set threshold (in Hertz). If a discontinuity was found, $Sg2$ was estimated as the average of the $F2$ values constituting the jump (see Figure 1.10). If no discontinuity was detected, $Sg2$ was estimated simply as the token's average $F2$. $Sg3$ was estimated with the help of Eq. (1.2), which was derived using a previously-proposed model of the subglottal airways [Lul06].

$$Sg3 = Sg2\{-0.3114[\log_{10}(Sg2) - 3.280]^2 + 1.436\} \tag{1.2}$$

The algorithm was evaluated indirectly by applying it to speaker normalization tasks and its performance was found to be vowel dependent. Specifically, the estimation accuracy was high for diphthongs but much poorer for other vowels.

The above $Sg2$ estimation algorithm was improved in [WLA08] and [WLA09a], but was customized to suit children's speech (unlike the above algorithm, which was applicable to adults as well as children). In [WLA08], a *rough* estimate of $Sg2$ was first obtained using the following empirical relation between $Sg2$ and $F3$ [Lul10]:

$$Sg2 = 0.636 \times F3 - 103.$$

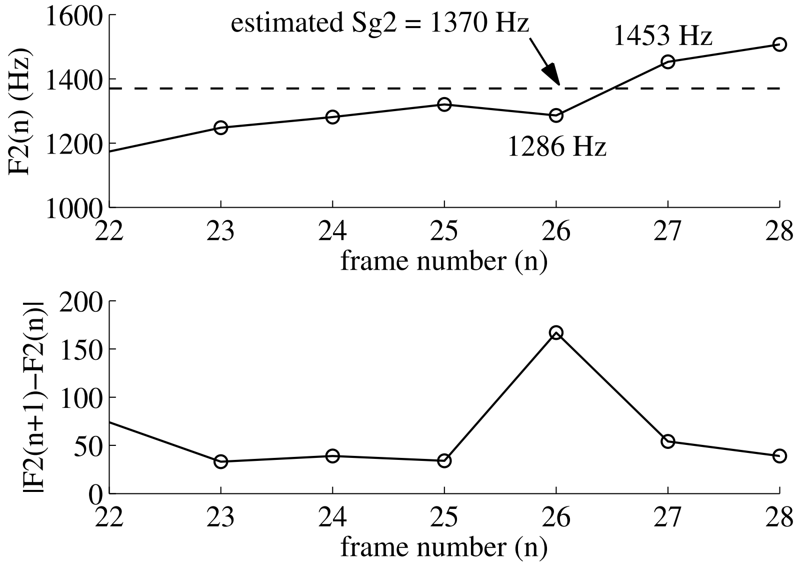


Figure 1.10: *Estimating $Sg2$ based on the $F2$ discontinuity observed in a token of [ɔɪ]—using the algorithm proposed in [WAL08]. The upper panel shows the frame-by-frame $F2$ track. The dashed line passes through the average of the low and high $F2$ values constituting the jump. The lower panel shows the absolute first difference of the $F2$ track.*

Then, a refined estimate was obtained by searching for an $F2$ jump within ±100 Hz of the initial estimate and computing a weighted average of the $F2$ values constituting the discontinuity, if a discontinuity was found. This procedure enabled reliable detection of $Sg2$-induced $F2$ jumps, especially in the presence of nearby, competing jumps that could be caused by other factors (e.g., interdental spaces) [HTT10]. In [WLA09a], the initial $Sg2$ estimate was obtained as in [WLA08], but its refinement relied on locating not only an $F2$ jump but also an accompanying attenuation in the second formant's energy prominence, a phenomenon which has been shown to be more robust than $F2$ discontinuities in indicating subglottal coupling effects [CS07]. The improved algorithms were more reliable than the algorithm in [WAL08], but their performance was still found to be vowel dependent.

The algorithms in [WAL08, WLA08, WLA09a] suffer from the following limitations. (1) Their approach is not well suited to estimating $Sg1$ because automatically detecting $Sg1$-induced coupling effects (in trajectories of $F1$) can be challenging [Jun09]. (2) Their practical applicability is rather limited because (a) their performance is data dependent, and (b) they can be applied only to isolated vowels (and not continuous or natural speech). (3) Detection of subglottal coupling effects requires very accurate formant tracking procedures. As the authors of [WLA09a] point out, "Manual verification and/or correction is applied through visually checking the tracking contours against spectrograms," implying that their algorithms are not completely automatic. The algorithms developed in this dissertation are fully automatic and can estimate the first three SGRs from continuous speech in a content- and language-independent manner.

1.4 Height estimation using speech signals

Automatic height estimation—estimating the height of an unknown speaker from his/her speech sample—could have potential applications in forensics, automatic analysis of telephone calls (e.g., 911 distress calls), and automatic speaker identification. In the past, researchers attempted to identify height-related features of speech based on the assumption that an anatomical correlation exists between speaker height and vocal-tract length (VTL). In fact, a study using magnetic resonance imaging techniques [FG99]—over a wide range of speaker ages and heights—provides some evidence in favor of this assumption. Motivated by a fundamental premise of speech-production theory that formant frequencies are inversely proportional to VTL, several studies have analyzed the correlation between speaker height and formant frequencies [DM95, Gon04, RKN05]; however, no strong correlations have been reported. A few studies have also investigated the relation between height and $F0$, but have found no significant correlation between the two [Gon04, Kun89]. More recently, [Dus05] has reported the correlations between speaker height and commonly-used vocal-tract features such as Mel-frequency cepstral coefficients (MFCCs) [DM80] and LPCs; the study shows that 57% of the variance in height can be explained using 31 vocal-tract features: the first 10 MFCCs, 16 LPCs and the first 5 formant frequencies.

A few studies have proposed automatic algorithms to estimate speaker height using speech signals. In [PH97], speech signals were parameterized using the first 19 MFCCs, and 11 height-dependent Gaussian mixture models (GMMs) were trained using data from all speakers in the TIMIT corpus [Gar88b]. The height of a given speaker was then estimated using the maximum *a posteriori* classification rule. With this approach, the height estimation error was found to be 5 cm or less for 72% of the speakers. However, it should be noted that the *same* set of speakers was used for both training and evaluation. In [GMF10a], support vector machine (SVM) regression was proposed for height estimation. The model was trained and evaluated using data from 462 and 168 speakers, respectively, in the TIMIT corpus. Training was accomplished by first extracting 6552 audio features from each utterance, and then subjecting them to a feature ranking procedure to choose the most relevant subset. The subset consisting of the top 50 features resulted in the best performance, yielding a mean absolute error (MAE) equal to 5.3 cm and a root mean squared error (RMSE) equal to 6.8 cm. The features consisted mostly of means, standard deviations, percentiles and quartiles of MFCCs, $F0$ and voicing probability. In [GMF10b], a similar algorithm using Gaussian-process regression was proposed for real-world indoor and outdoor scenarios, and results identical to those of [GMF10a] were achieved. Although the algorithms in [GMF10a] and [GMF10b] yield reasonably good results (MAE =

5.3 cm) using statistical measures of speech features, it is not clear as to how such features relate to speaker height.

Despite the correlation between VTL and speaker height, height estimation using vocal-tract information is difficult because the configuration of the vocal tract changes significantly during speech production. Specifically, as evident from [Dus05], [GMF10a] and [GMF10b], a large number of vocal-tract features are required to capture the correlation between height and VTL. In this dissertation, a novel approach to height estimation is proposed based on the observed relationship between speaker height and SGRs. Since the configuration of the subglottal system changes little over time, the proposed approach, in addition to having a physiological basis, is likely to be more efficient than existing techniques in terms of the number of features required for height estimation.

1.5 Speaker normalization for ASR

Automatic speech recognition (ASR) systems can be either speaker dependent (SD) or speaker independent (SI) depending on the source of the training data. An SD system can achieve high recognition accuracy, but requires a large amount of training data from the target speaker. It may also not generalize well to new speakers. On the other hand, SI systems are trained using data from a large speaker population, and their performance is, in general, worse compared to that of SD systems. However, SI systems are more commonly used in practice because they offer more flexibility and can be easily adapted to new speakers.

Inter-speaker variability (with regard to speech acoustics) poses a challenge to the design of SI-ASR systems. Inter-speaker acoustic variations are mostly caused by morphological differences in the vocal tract—especially vocal-tract length.

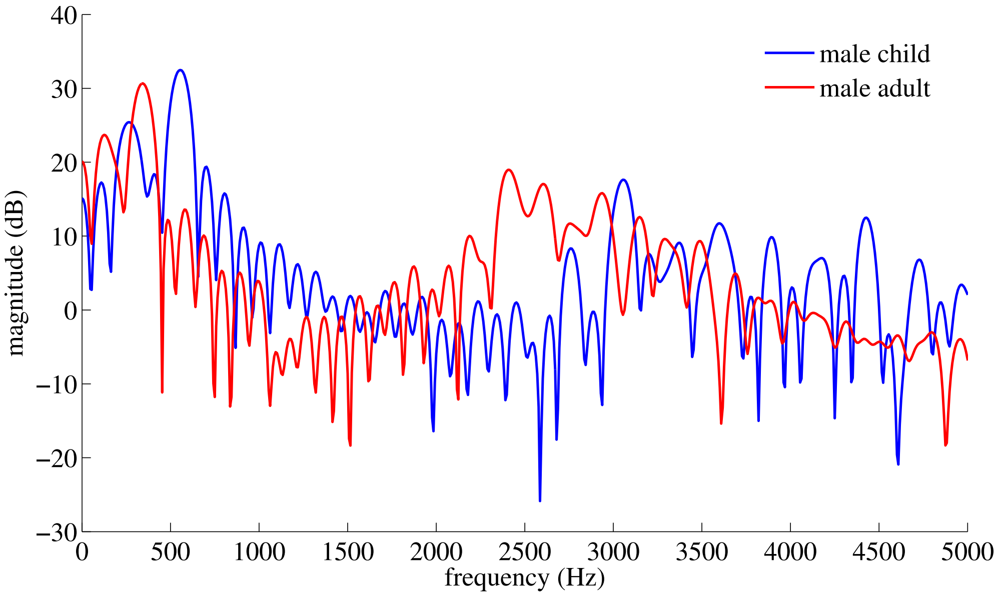


Figure 1.11: *Comparing the steady-state magnitude spectra of the vowel [i] (as in "heed") as enunciated by a male adult and a male child speaker. Vowel data were obtained from the database used in [HGC95].*

Typically, adult females have shorter vocal tracts compared to adult males, and children have shorter vocal tracts compared to adults [Wak77]. This implies, according to linear speech-production theory [Fan60], that children tend to have higher formant frequencies than adults, and adult females tend to have higher formant frequencies than adult males—Figure 1.11 provides an example using steady-state magnitude spectra of the vowel [i]. Consequently, the performance of SI-ASR systems varies significantly across speakers.

The effects of inter-speaker variability can be mitigated using speaker normalization and adaptation techniques. Speaker normalization employs frequency warping in the front-end feature domain. A widely-used approach to frequency warping involves a piece-wise linear function with a single parameter that controls the degree of spectral compression or expansion [LR98]. This approach is known as *conventional* vocal-tract length normalization (VTLN), and will henceforth be referred to as VTLN for simplicity. Nonlinear frequency warping has also been investigated in the past. Popular approaches include power-law transformation [EG96], all-pass transformation methods [McD00], and methods based on spectral shifting in the Mel or Bark scale [SU08, WLA09a]. In speaker adaptation, spectral variability is implicitly compensated for in the back-end acoustic-model domain by statistically tuning the SI model parameters to a given target speaker. Popular approaches include maximum-likelihood linear regression (MLLR) [LW95], constrained MLLR [Gal98], and maximum *a posteriori* (MAP) adaptation [GL94]. Typically, the number of parameters to be estimated for speaker adaptation is several orders of magnitude larger than the number of parameters to be estimated for speaker normalization. Therefore, speaker normalization techniques are generally preferred when the amount of data is limited. In this dissertation, the focus is only on speaker normalization.

Frequency-warping parameters are typically estimated using the maximum-likelihood (ML) criterion. When frequency warping is implemented directly in the power-spectrum domain, the optimal parameters are determined via an ML grid search over the parameter space [LR98,SU08]. On the other hand, when frequency warping is implemented as a linear transformation of cepstral features, the optimal parameters are determined via explicit maximization of the ML objective function (using the Expectation Maximization algorithm)—see [MSW04], for example. Another approach to estimating warping parameters is to define them as ratios (or differences) of formant frequencies (especially $F3$) and formant-like spectral peaks [EG96,GS97,ZW97,CA06]. More recently, $Sg2$ has also been used to compute warping factors (as ratios) and spectral shifts (as differences) for speaker normalization [WAL08, WLA08, WLA09a].

ML approaches perform better than ratio-based methods in general (see [ZW97] for comparison), but ratio-based methods tend to be more effective in limited-data conditions. One drawback of all the above methods is their sensitivity to noise. ML approaches are purely statistical in nature and hence likely to be less effective in noise, especially when the training data are relatively clean. On the other hand, ratio-based methods, including the $Sg2$-based methods of Wang *et al.* [WAL08, WLA08, WLA09a], are likely to be ineffective in noise owing to their stringent formant-tracking requirements (see Section 1.3 for a description of the $Sg2$-estimation algorithms developed by Wang *et al.*). This dissertation proposes a hybrid knowledge-based and statistical approach to normalization using the first three SGRs. The method is shown to be effective in clean as well as noisy environments, especially when data are limited.

1.6 Speaker identification and verification

The speech signal conveys several levels of information. Primarily, the speech signal conveys the message (i.e., the sequence of words) being spoken, but on a secondary level, the signal also conveys information about the identity of the talker. While ASR is concerned with decoding the word sequence in a given utterance, automatic *speaker* recognition is concerned with extracting the identity of the talker. Speaker recognition has found wide application in telephone-based financial transactions, information retrieval from speech databases, voice-based user authentication, etc.

The general area of speaker recognition involves two specific, closely-related tasks: speaker identification (SID) and speaker verification (SV). In SID, the goal is to determine which one of a group of known voices best matches the given voice sample. In SV, the goal is to verify, given a voice sample and an associated claim, if the talker is indeed the one he or she claims to be. In both tasks, the speech input can be either totally unconstrained (text independent) or constrained to be a known phrase (text dependent). This dissertation considers the text-independent case only. The success of SID and SV systems depends on extracting and modeling the speaker-dependent characteristics of the speech signal which can effectively distinguish one talker from another.

Mel-frequency cepstral coefficients (MFCCs), which capture the acoustics of the supraglottal vocal tract, have been widely used for both SID and SV. They have been shown to provide good performance with a number of modeling schemes such as simple Gaussian mixture models (GMMs) [RR95], GMMs adapted from universal background models (UBMs) [RQD00], support vector machine (SVM) supervectors [CSR06,YLL09], joint speaker and channel factors [KBO07,KOD08], and total-variability i-vectors [DKD11]. Other features that have been proposed and used in conjunction with MFCCs (via feature-level or score-level fusion) include those based on voice-source parameters [MY06], spectro-temporal modulation frequencies [Kin06], prosody [SFK05], word patterns and lexicon [Dod01], and articulatory parameters [LKG13]. This dissertation investigates the utility of *subglottal* features (capturing the acoustics of the tracheo-bronchial airways) for both SID and SV. The focus is specifically on cepstral coefficients extracted from subglottal acoustics (henceforth referred to as SGCCs) and their fusion with MFCCs for improved speaker-recognition performance.

1.7 Dissertation outline

The rest of this dissertation is organized as follows.

Chapter 2 describes the new databases (comprising simultaneous recordings of speech and subglottal acoustics) that were collected for the purposes of this dissertation. It also presents some important results of data analysis to motivate the techniques and algorithms developed in the following chapters.

Chapter 3 presents automatic algorithms to estimate SGRs in speech signals of adults and children. Algorithms are developed for natural speech, with emphasis on language independence and robustness to noise and data limitedness.

Chapters 4, 5 and 6 form the core of this dissertation; each of them focuses on a particular application of subglottal acoustics. Chapter 4 investigates body-height estimation using speech-based estimates of SGRs. The method is evaluated using a standard database of adults' speech, and the results are analyzed in light of physiological limits to height-estimation accuracy. Chapter 5 uses SGRs to develop a hybrid statistical and knowledge-based approach to speaker normalization. The emphasis is on speaker normalization for children, in clean as well as noisy environments. Empirical results are presented for a standard connected-digit ASR task. Chapter 6 proposes an automatic algorithm to estimate subglottal cepstral coefficients using speech signals. Estimated subglottal cepstra are used to provide complementary information to conventional MFCCs in SID and SV tasks. Experimental results are reported for two standard databases of adults' speech (one for SID and one for SV).

Chapter 7 summarizes the key results of this dissertation and provides directions for future work.

The material presented in this dissertation is based in part on the following published articles:

Harish Arsikere, Gary K. F. Leung, Steven M. Lulich and Abeer Alwan, "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation," Speech Communication, Vol. 55, pp. 51–70, 2013. (Chapters 2, 3 and 4)

Harish Arsikere, Steven M. Lulich and Abeer Alwan, “Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency,” Proceedings of ICASSP, pp. 7922–7926, 2013. (Chapter 5)

CHAPTER 2

Data collection and analysis

Given the diverse roles of the subglottal input impedance and its resonances in speech production, perception and, potentially, technology (see Chapter 1), it is important that their properties be well understood for a large number of speakers. The *exact* subglottal input impedance (and SGRs) can be measured only through invasive procedures such as laryngectomy [IMK76], placing miniature pressure transducers below the glottis [CB85], or using an endotracheal tube [HCS94]. Owing to the technically-challenging nature of these procedures, a popular, non-invasive alternative has been the use of an *accelerometer* placed against the skin of the neck [Che02, CS07, MLW08, WM09, CBG09, Lul10, GLC11]. In this case, the phonation volume velocity acts as the source that drives the subglottal input impedance, and the pressure at the top of the trachea relates to the motion of the neck tissues (and hence the accelerometer). This results in a frequency spectrum that is closely related to the input impedance, but one that is sampled by the source harmonics, and partially shaped by the source spectral envelope and the effects of acoustic coupling with the vocal tract. While the accelerometer-based method is less accurate than the invasive methods mentioned above, it enables easy and rapid data acquisition and is known to provide a reliable approximation of the exact input impedance [Che02].

The above studies were all based on small data sets—only [Lul10] and [WM09] analyzed data from more than 10 subjects (23 and 19, respectively). Also, none of the above studies, except [Lul10], involved children. Therefore, one of the goals of this dissertation was to obtain and analyze data from a large number of adult and child subjects. Speech and subglottal acoustics were recorded simultaneously, and the data were analyzed to understand the relationships between SGRs, formant frequencies, body height (or speaker height) and native language. The analyses were intended to aid the development of automatic algorithms for SGR estimation from speech signals, speech-based body height estimation, and speaker normalization for ASR. This chapter describes the corpora that were collected and presents some of the important analysis results.

2.1 The WashU-UCLA corpora

Data were collected in collaboration with the Washington University Psychology Department. Time-synchronized recordings of speech and subglottal acoustics were obtained from (1) 25 male and 25 female adult native speakers of American English (AE)—the WashU-UCLA Adults corpus, (2) 4 male and 2 female adult bilingual speakers of Mexican Spanish (MS), their first language, and AE—the WashU-UCLA Bilingual Adults corpus, and (3) 31 male and 12 female native AE-speaking children—the WashU-UCLA Kids corpus [LMA12, LAM11]. The native AE speakers were recorded at Washington University, while the bilingual speakers were recorded at UCLA. Adult speakers were aged between 18 and 24 years, while children were aged between 6 and 17 years.

2.1.1 Recording setup and material

Recordings of speech and subglottal acoustics were made with a SHURE PG27 microphone and a K&K Sound HotSpot accelerometer, respectively, while participants sat in a double-walled, sound attenuating booth. All signals were recorded at a sampling rate of 48 kHz and a resolution of 16 bits/sample. MATLAB was used to acquire and save the data via an M-Audio MobilePre USB pre-amplifier. The microphone was placed roughly 20 cm in front of the speaker and slightly to the right to avoid distortion due to airflow during high-airflow sounds. The speaker was instructed on how to hold the accelerometer against the skin of the neck at the cricoid cartilage below the level of the glottis.

The subjects were made to sit in front of a computer monitor that displayed sentences to be read aloud. Various consonant-vowel-consonant (CVC) words were embedded in the carrier phrase "*I said a* ____ *again*" (or "*Dije una* ____ *otra vez*" in the case of MS recordings) and displayed on the monitor in random order. The CVC words were divided into two lists, and each list was recorded in a separate session. For the native AE speakers, the first word list comprised 21 AE CVb words ('V' was one of 4 monophthongs or 3 diphthongs, and 'C' was one of [b], [d] or [g]), and the second word list comprised 14 AE hVd words ('V' was one of 9 monophthongs, 4 diphthongs, or the approximant [ɹ]). Rows 1 and 2 of Table 2.1 show the list of vowels recorded along with the corresponding values of the features [low] and [back]. For the bilingual speakers, the first word list was identical to the one used for the native AE speakers, while the second word list comprised 21 MS CVb words ('V' was one of 4 monophthongs or 3 diphthongs, and 'C' was one of [b], [d] or [g]). Row 3 of Table 2.1 shows the list of vowels recorded (the MS vowels are indicated in parentheses; note that each MS vowel is placed below an AE vowel that phonetically resembles it the most). Each target word was recorded 10 times for the adult native AE speakers, 7 times for the adult bilingual speakers, and up to 7 times for children. In addition to the above word lists, two accelerometer recordings of the sustained vowel [ɑː] were recorded for each speaker, in which there was special emphasis on obtaining a

Table 2.1: *List of vowels recorded in the WashU-UCLA corpora. The hVd words in American English (AE) were recorded for the native English speakers only, while the AE CVb words were recorded for all speakers. The CVb words in Mexican Spanish (MS) were recorded for the bilingual speakers only. Note that each MS vowel, indicated in parentheses, is placed below an AE vowel that phonetically resembles it the most. For monophthongs, values of the phonological features [low] and [back] are also indicated.*

	i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	o	ʊ	u	aɪ	aʊ	ɔɪ	ɹ
AE (hVd)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AE (CVb)	✓			✓		✓				✓	✓	✓	✓	
MS (CVb)	✓ (i)			✓ (e)		✓ (a)				✓ (u)	✓ (ai)	✓ (au)	✓ (oi)	
feature [low]	-	-		+	+	+	+	-	-	-				
feature [back]	-	-		-	-	+	+	+	+	+				

clear resonance structure up to $Sg3$. These high quality accelerometer recordings were obtained using WAVESURFER [SB00] by allowing the speaker and the experimenter to interactively adjust the placement of the accelerometer until the best quality signal was achieved. This was an important step because the quality of the accelerometer signals during the list recordings was expected to be somewhat variable. The age, height and gender of each speaker were also noted.

2.1.2 Manual annotation

To facilitate data analysis, vowel segments were manually annotated in all of the microphone recordings. Separate annotations were not required for the accelerometer recordings since they were time synchronized with the microphone signals.

Using PRAAT [BW09], the *beginning* and *end* of the target vowel in each of the microphone recordings were manually labeled by a single investigator (to ensure consistency). For monophthongs and the approximant [ɹ], the *middle* of the steady-state portion of the vowel was also labeled. For diphthongs, the *nucleus* was labeled either in the middle of the steady state, if a steady state existed, or just before the onset of rapid formant movements. All label files (one corresponding to each microphone recording) were saved in the PRAAT TextGrid format. The *beginning* of each target vowel in the CVb word lists (AE and MS) was labeled where the formants became visible immediately after the initial plosive consonant. The *beginning* of each target vowel in the hVd word list was labeled where the formants were visible and the waveform demonstrated a significant deviation from previous aspirated pulses. For both lists, the *end* of the vowel was labeled at the point of closure of the final stop consonant. In general, the placement of each label was guided by inspection of the spectrogram and waveform, and by listening.

2.2 Measurement methods

For the native AE speakers, only the hVd recordings were analyzed in this dissertation. For the bilingual speakers, both the AE and MS CVb recordings were analyzed (to investigate the effect of native language on SGRs). In all cases, only the monophthong vowels were considered (see Table 2.1 for a list of vowels). Several repetitions of each monophthong (microphone as well as accelerometer signals)—5 in the case of adults and 3 in the case of children—were chosen at random for measuring the first three formants, the fundamental frequency ($F0$), and the first three SGRs. All measurements were made in the steady-state regions (in accordance with the PRAAT TextGrid labels). While $F0$ and formants were mea-

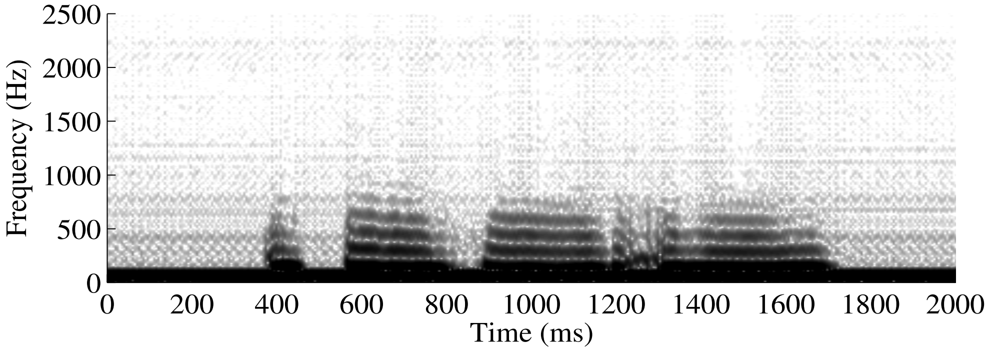


Figure 2.1: *Spectrogram of a sample accelerometer signal (taken from the WashU-UCLA Kids corpus) demonstrating the strong low-pass nature of subglottal acoustics. Note that the signal had a bandwidth of 24 kHz originally, but was down sampled to 5 kHz (2.5 kHz bandwidth) for display purposes.*

sured semi automatically (i.e., with manual tuning of parameters) using the pitch and formant tracks provided by WAVESURFER, SGRs were measured manually using a combination of FFT (fast Fourier transform) spectra, LPC (linear predictive coding) spectra, and autocorrelation-based smoothed spectral envelopes. Manual procedures were required for measuring SGRs because the accelerometer recordings were noisy and strongly low pass in nature owing to interference from the neck skin and tissues (see Figure 2.1).

2.2.1 Measuring $F0$ and formants

Using WAVESURFER, microphone signals were down sampled to 10 kHz and pre-emphasized using the frequency response: $H(\omega) = 1 - 0.97e^{-j\omega}$. $F0$ and the first three formants were obtained at 5 ms intervals after dividing the given signal into Hamming-windowed segments. For formant tracking, the window length was varied between 30 and 50 ms, and the LPC order was varied between 12 and 16 until the tracking contours aligned satisfactorily with the signal's spectrogram.

For $F0$ tracking, the ESPS (Entropic signal processing system) algorithm [Tal95] (built into WAVESURFER) was used. The window length was set to 7.5 ms and the minimum pitch parameter was set to 60 Hz. The maximum pitch parameter was 400 and 450 Hz for adults and children, respectively. The values of $F0$, $F1$, $F2$, and $F3$ near the labeled steady-state part of the vowel were averaged over five frames centered around the label location. $F0$ measurements were not used for the analysis presented in this chapter; their purpose will be made clear when SGR estimation algorithms are discussed later in Chapter 3.

2.2.2 Measuring SGRs

Given a vowel token of an accelerometer recording, a steady-state segment with length equal to 4 pitch periods was chosen for analysis. The segment was down sampled to between 6 and 10 kHz depending on how noisy the signal was at high frequencies, and then passed through a pre-emphasis filter as described in Section 2.2.1. The FFT spectrum was computed after applying a Hamming window. For LPC analysis, the LPC order was set to between 10 and 14 depending on whether the first two harmonics dominated the spectrum at low frequencies, thus requiring an increased number of poles to reveal $Sg1$. The smoothed spectral envelope was obtained by dividing the long analyis segment into smaller subframes, computing the autocorrelation function for each subframe after applying a Hamming window, and computing the FFT of the averaged autocorrelation function. The subframe size was set to between 0.9 and 1.1 pitch periods depending on the intended frequency resolution, and the overlap between successive subframes was 80% of the subframe size.

The FFT spectrum, the LPC spectrum, and the smoothed envelope were plotted on a single graph (as illustrated with an example in Figure 2.2), and with

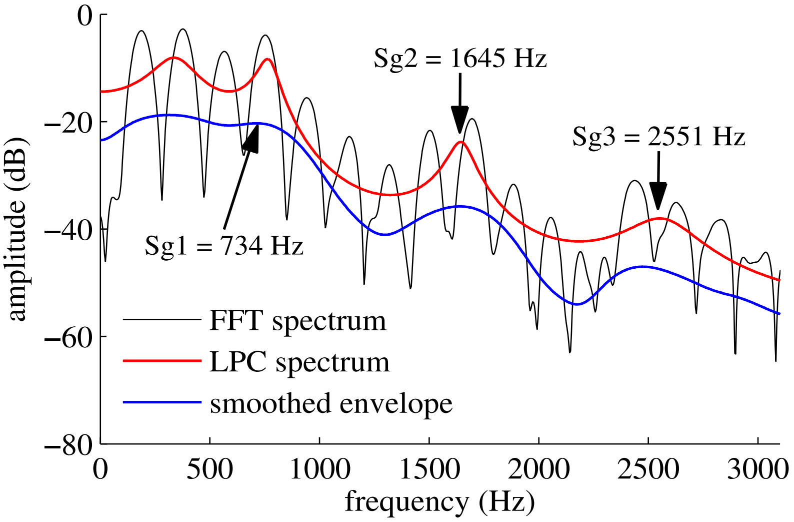


Figure 2.2: *FFT spectrum, LPC spectrum, and smoothed envelope for a sample accelerometer signal taken from the WashU-UCLA Kids corpus. Note that the first two harmonics are particularly high in energy, producing an additional peak in both the LPC spectrum and the smoothed envelope.*

the FFT spectrum as a guide, $Sg1$, $Sg2$, and $Sg3$ were measured from either the LPC spectrum or the smoothed envelope, depending on which was judged to give the more accurate result. If neither spectral representation was satisfactory for a particular SGR, the SGR frequency was not measured. Hence it is important to note that not all three SGRs were necessarily measured in every vowel token chosen for analysis. In general, it was more difficult to measure $Sg1$ and $Sg3$ than to measure $Sg2$. While measuring $Sg1$ was sometimes difficult (especially for high-pitched speakers) due to its proximity to strong low-frequency harmonics, measuring $Sg3$ was always difficult owing to the attenuation of high frequencies caused by the low-pass nature of the neck tissues and skin.

2.3 Analysis results

This section presents a few important results based on the measurements of formant and SGR frequencies. The discussion is focused towards motivating the algorithms to be presented in the coming chapters. For details regarding other analyses, the interested reader is referred to the study by Lulich *et al.* [LMA12].

2.3.1 Phonetic and language independence of SGRs

Owing to the stationary nature of subglottal acoustics (see Figure 2.1, for example), the SGRs of a given speaker are expected to be constant, regardless of the phonetic content or language spoken. This hypothesis was verified with the help of the manual SGR measurements that were obtained from the WashU-UCLA corpora (see Section 2.2.2). Several SGR measurements were available for each speaker, but they were not equally distributed across the monophthong vowels chosen for analysis. $Sg1$ and $Sg2$ measurements were available for all speakers, while no measurements of $Sg3$ could be made for 1 speaker in the WashU-UCLA Adults corpus and 12 speakers in the WashU-UCLA Kids corpus. For the bilingual speakers, roughly equal numbers of measurements were obtained from AE and MS vowels. The within-speaker coefficient of variation (COV)—defined as the ratio of standard deviation to mean—was computed for $Sg1$, $Sg2$ and $Sg3$ for each native AE speaker, and the within-speaker, cross-language COV (i.e., with measurements from AE and MS vowels combined) was computed for each bilingual speaker. For comparison, the within-speaker COVs of $F3$, which is known to be the least variable of the first three formants [PB52], were similarly computed.

Table 2.2 shows the average within-speaker percentage COVs for each database. The corresponding raw standard deviations are also shown. SGR COVs are on

Table 2.2: *Average within-speaker percentage coefficient of variation (COV) for SGRs and $F3$ in the WashU-UCLA corpora, along with the corresponding raw standard deviations (σ). For bilingual speakers, measurements from both AE and MS data were combined (cross-language COV and σ).*

	$Sg1$	$Sg2$	$Sg3$	$F3$
	%COV ($\sigma/\mu \times 100$)			
WashU-UCLA Adults corpus	5.0	2.2	2.5	8.3
WashU-UCLA Bilingual Adults corpus	4.5	2.3	2.7	7.5
WashU-UCLA Kids corpus	4.8	3.1	2.5	8.8
	σ (Hz)			
WashU-UCLA Adults corpus	30	32	57	231
WashU-UCLA Bilingual Adults corpus	25	32	61	203
WashU-UCLA Kids corpus	35	55	69	278

the order of 2–5%, indicating that the measurements vary very little about their mean values. In comparison, the average within-speaker COVs of $F3$ are about 2–3 times higher. These results confirm that SGRs can indeed be considered 'constant' for a given speaker and that average values are sufficient to characterize a speaker's SGRs. Henceforth in this dissertation, the term *ground truth SGRs* (or *actual SGRs*) will be used to refer to the mean SGR frequencies of a given speaker. To evaluate SGR estimation algorithms, the ground truth SGRs will be treated as reference values regardless of the phonetic content, language and duration of the test utterance.

Table 2.3: *Average ground truth SGRs for speakers in the WashU-UCLA corpora along with the corresponding minimum and maximum values (in square brackets).*

	*Sg*1 (*Hz*)	*Sg*2 (*Hz*)	*Sg*3 (*Hz*)
		WashU-UCLA Adults corpus	
Males	**542** [492, 622]	**1327** [1217, 1492]	**2198** [2039, 2449]
Females	**659** [580, 722]	**1511** [1382, 1610]	**2410** [2273, 2575]
		WashU-UCLA Bilingual Adults corpus	
Males	**533** [491, 556]	**1314** [1198, 1405]	**2160** [1931, 2343]
Females	**642** [626, 658]	**1503** [1493, 1513]	**2462** [2420, 2505]
		WashU-UCLA Kids corpus	
Males	**727** [532, 906]	**1720** [1261, 2160]	**2710** [2056, 3384]
Females	**752** [672, 831]	**1778** [1519, 2006]	**2720** [2417, 2980]

2.3.2 Ground truth SGRs of adults and children

Table 2.3 shows the average (across speakers) ground truth SGRs along with the corresponding minimum and maximum values (in square brackets) for the three WashU-UCLA corpora. These values serve to indicate the typical ranges of SGRs for young adults, and children between the age of 6 and 17 years. The ground truth SGRs of children are significantly higher than those of adults, on average. Within adults, the ground truth SGRs of females are significantly higher than those of males, on average. Note that gender does not have a significant effect in the case of children. These trends are largely the result of the dependence of SGRs on body height, which will be discussed further in Section 2.3.5.

2.3.3 Relationships between SGRs and formants

On the one hand, formant frequencies, especially $F1$ and $F2$, are known to vary considerably with phonetic content [PB52]. On the other hand, SGRs are practically independent of phonetic content (as Table 2.2 shows). Given these contrasting behaviors, this section tries to answer the following question: Can the two sets of resonances be combined in a way that would enable the automatic estimation of SGRs when only formant frequencies are known?

As discussed in Section 1.2.2, $Sg1$ forms a natural acoustic boundary between [+low] and [-low] vowels along the $F1$ dimension—$F1$ is typically below $Sg1$ for [-low] vowels and above $Sg1$ for [+low] vowels, and $Sg2$ forms a natural acoustic boundary between [+back] and [-back] vowels along the $F2$ dimension—$F2$ is typically below $Sg2$ for [+back] vowels and above $Sg2$ for [-back] vowels. Also, as discussed in Section 1.2.2, SGRs tend to play a role in the human perception of vowel phonological features. Based on these ideas, SGRs and formants were combined to define the following perceptually-motivated acoustic features of vowel height and backness:

$$\text{vowel height: } B_{f1,sg1} = F1(\text{Bark}) - Sg1(\text{Bark}), \tag{2.1}$$

$$\text{vowel backness: } B_{f2,sg2} = F2(\text{Bark}) - Sg2(\text{Bark}), \tag{2.2}$$

where $F1$(Bark), $F2$(Bark), $Sg1$(Bark) and $Sg2$(Bark) denote the first two formants and SGRs, respectively, on the Bark scale. The Bark scale is a psychoacoustical scale based on human hearing [Zwi61] and several definitions exist for it. In this dissertation, the following definition based on [Tra90] was used:

$$z = [(26.81f)/(1960 + f)] - 0.53, \tag{2.3}$$

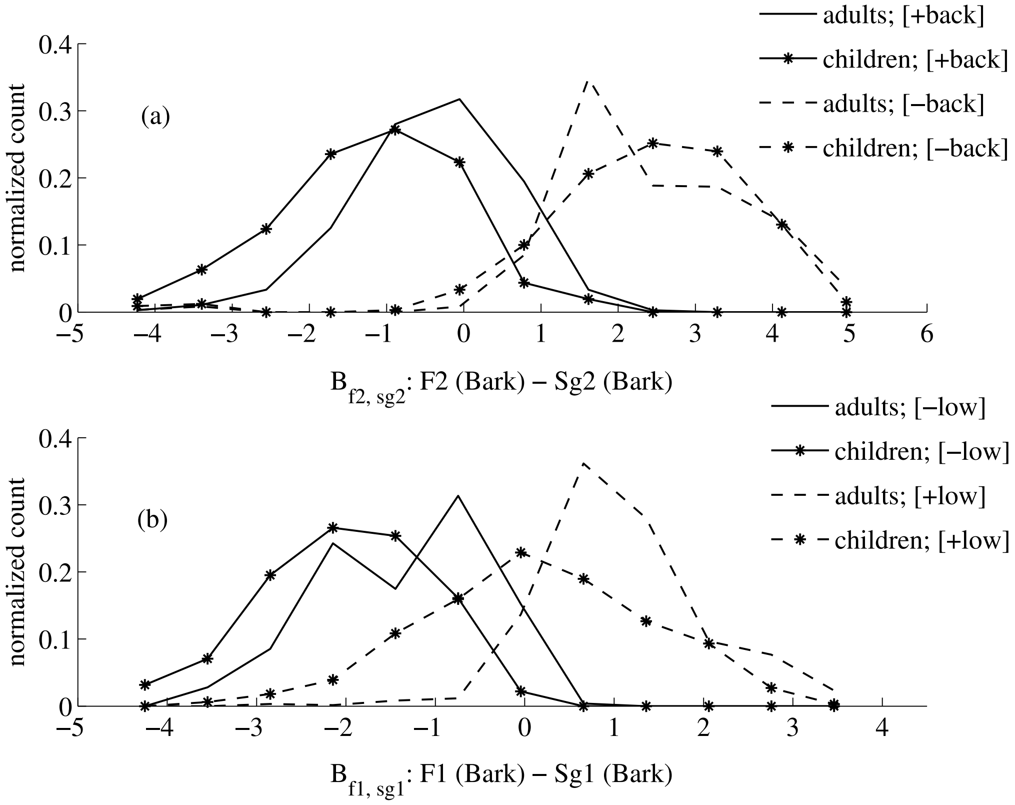


Figure 2.3: *Distributions of (a) $B_{f2,sg2}$, and (b) $B_{f1,sg1}$, based on vowel formant measurements and ground truth SGRs of native AE-speaking adults and children in the WashU-UCLA corpora.*

where z denotes the Bark value corresponding to a frequency f in Hertz. Equation (2.3) offers simplicity in terms of conversion between Hertz and Bark values, in addition to being as accurate as the other definitions.

Figure 2.3(a) shows the distribution of $B_{f2,sg2}$ for [+back] and [-back] vowels, and Figure 2.3(b) shows the distribution of $B_{f1,sg1}$ for [+low] and [-low] vowels. The distributions were obtained using vowel formant measurements (see Section 2.2.1) and ground truth SGRs of all the native AE speakers in the WashU-UCLA corpora. The clear separation between vowel classes ([+back] ver-

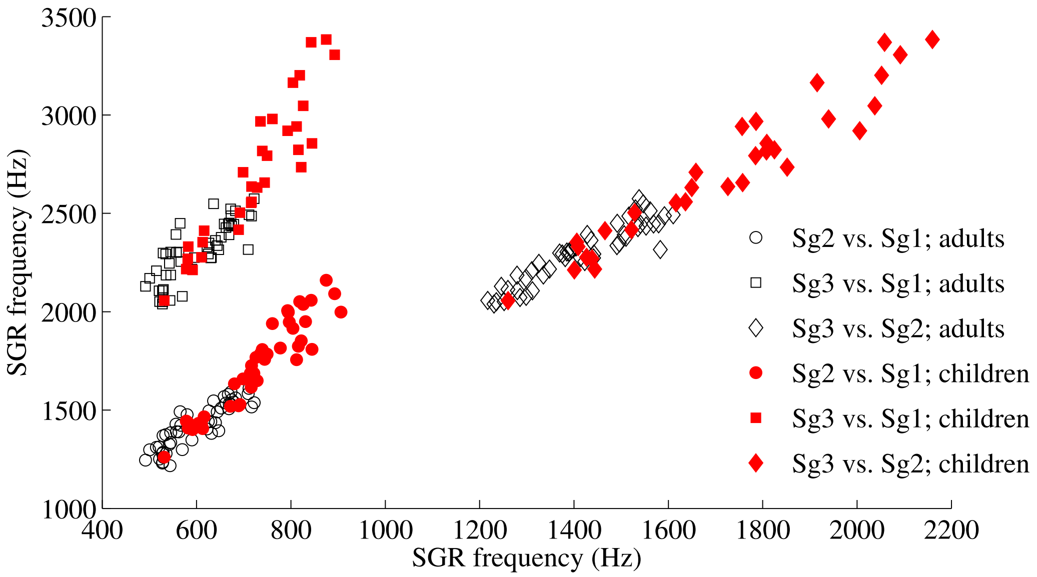


Figure 2.4: *Scatter plots of $Sg2$ versus $Sg1$, $Sg3$ versus $Sg1$, and $Sg3$ versus $Sg2$, for the native AE-speaking adults and children in the WashU-UCLA corpora.*

sus [-back], and [+low] versus [-low]) around 0 Bark indicates that $B_{f1,sg1}$ and $B_{f2,sg2}$ are reliable acoustic measures of vowel height and vowel backness, respectively. As will be shown in Chapter 3, these measures—involving both SGRs and formants—correlate strongly with well-known formant-only measures, thus enabling the automatic estimation of SGRs given only formant information.

2.3.4 Correlations among SGRs

It was shown in [LAA11] that SGRs conform fairly well to the uniform-tube model of Eq. (2.4):

$$SgN = \frac{(2N-1)c}{4l} \qquad N = 1, 2, 3, \tag{2.4}$$

Table 2.4: *Correlations among the ground truth SGRs of native AE-speaking adults and children in the WashU-UCLA corpora.*

	Adults			Children		
	males	females	overall	males	females	overall
$Sg1$ vs. $Sg2$	0.585	0.659	**0.885**	0.952	0.881	**0.938**
$Sg2$ vs. $Sg3$	0.936	0.727	**0.929**	0.976	0.937	**0.965**
$Sg3$ vs. $Sg1$	0.486	0.645	**0.813**	0.948	0.631	**0.926**

where c is the speed (wave propagation velocity) of sound in the subglottal airways and l is the length of the uniform tube (closed at the glottal end and open at the distal end) that approximates the subglottal system. If the uniform tube were an exact model of the subglottal system, SGRs would be rational multiples of one another and hence perfectly correlated ($Sg2 = 3 \times Sg1$, $Sg3 = 5 \times Sg1$, $Sg3 = \frac{5}{3} \times Sg2$). However, since Eq. (2.4) is only an approximation, the correlations among SGRs are expected to be less than 1.

Figure 2.4 shows the scatter plots of $Sg2$ versus $Sg1$, $Sg3$ versus $Sg1$, and $Sg3$ versus $Sg2$, for the native AE-speaking adults and children in the WashU-UCLA corpora. Each data point represents the ground truth SGRs of a particular speaker. Table 2.4 shows the corresponding correlation coefficients (overall and by gender). The correlations involving $Sg1$ are poorer than the correlation between $Sg2$ and $Sg3$ (especially for adults). This is presumably due to low-frequency subglottal *tissue resonances* (typically between 150 and 300 Hz), which have been shown to affect the frequency of $Sg1$ (without altering the $Sg2$ and $Sg3$ frequencies) in a speaker-specific manner depending on (1) their proximity to $Sg1$, and (2) the mechanical properties of the wall tissue that gives rise to them [LAA11]. The strong correlations between $Sg2$ and $Sg3$ (both overall and by gender) are at-

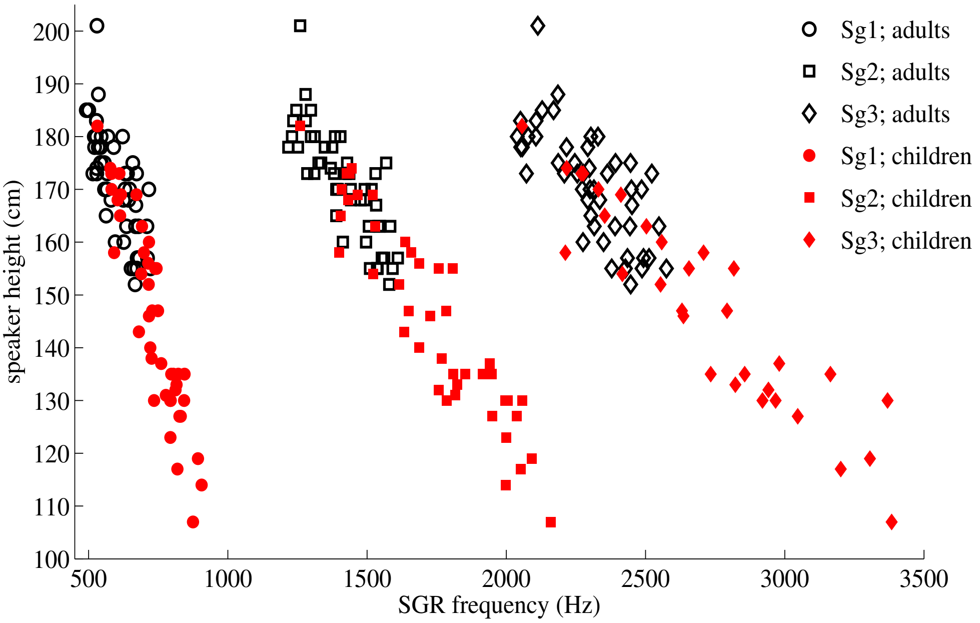


Figure 2.5: *Scatter plots of body height versus SGRs for the native AE-speaking adults and children in the WashU-UCLA corpora.*

tributable to the observation in [LAA11] that the uniform-tube model of Eq. (2.4) becomes a good approximation of the subglottal system at frequencies far from the tissue resonance. The correlation between $Sg2$ and $Sg3$ will be used in Chapter 3 for automatic $Sg3$ estimation. Note that the overall correlations are, on average, significantly higher than the within-gender correlations in the case of adults. This is because the SGRs of adult females differ significantly from those of adult males, on average (see Table 2.3).

2.3.5 SGRs versus body height

The trachea accounts for a significant portion of the overall subglottal-tract length [IMK76]. In [SP93], spectral features of tracheal sounds were shown to correlate strongly with body height (or speaker height). This, along with the fact that

Table 2.5: *Correlations between ground truth SGRs and height for the native AE-speaking adults and children in the WashU-UCLA corpora.*

	Adults			Children		
	males	females	overall	males	females	overall
$Sg1$ vs. height	-0.314	-0.297	**-0.742**	-0.929	-0.926	**-0.922**
$Sg2$ vs. height	-0.558	-0.527	**-0.810**	-0.943	-0.865	**-0.929**
$Sg3$ vs. height	-0.467	-0.338	**-0.719**	-0.950	-0.611	**-0.927**

Table 2.6: *Minimum, maximum and average height (in cm) of the native AE-speaking adults and children in the WashU-UCLA corpora.*

	Adults			Children		
	(18–24 years old)			(6–17 years old)		
	min.	max.	avg.	min.	max.	avg.
males	165	201	**178.4**	107	182	**145.7**
females	152	175	**163.6**	127	169	**143.8**

SGRs conform to a uniform-tube model (see Section 2.3.4), led to the hypothesis that SGRs might be correlated with body height.

Figure 2.5 shows the scatter plots of ground truth SGRs versus body height for the native AE-speaking adults and children in the WashU-UCLA corpora. Each data point corresponds to a particular speaker. Table 2.5 shows the corresponding correlation coefficients (overall and by gender). Children clearly show stronger correlations compared to adults, but this could be because the WashU-UCLA Kids corpus spans a larger height range compared to the WashU-UCLA Adults corpus (see Table 2.6). The overall correlations are significantly stronger than the within-gender correlations in the case of adults. This is because the average female speaker is significantly shorter than the average male speaker (see Table 2.6). In the case of children, on the other hand, the overall and within-gender correlations are comparable since the average speaker height is practically independent of gender. The correlations between SGRs and height will be used in Chapter 4 for the purpose of automatic height estimation from speech signals.

2.4 Conclusion

Time-synchronized recordings of speech and subglottal acoustics were collected from adult native AE speakers, adult bilingual AE and MS speakers, and native AE-speaking children. The first three formants and SGRs were measured in the steady-state regions of monophthong vowels. SGRs were found to be practically independent of phonetic content and native language; their average within-speaker COVs were on the order of 2–5%. The ground truth SGRs (averages of SGR measurements on a per-speaker basis) were significantly higher for children compared to adults, and for adult females compared to adult males. The correlations among SGRs and the correlations between SGRs and body height were found to be stronger (and less influenced by gender differences) for children ($0.92 \leq |r| \leq 0.97$) than for adults ($0.72 \leq |r| \leq 0.93$), in general. The Bark difference between $F1$ and $Sg1$ was found to be a reliable acoustic measure of vowel height, and the Bark difference between $F2$ and $Sg2$ was found to be a reliable acoustic measure of vowel backness. The above findings are important for automatic SGR estimation, height estimation and speaker normalization, as will be evident from the forthcoming chapters.

CHAPTER 3

Automatic estimation of SGRs

Motivated by the practical utility of SGRs—for applications such as height estimation (Chapter 4) and speaker normalization (Chapter 5)—and the need to estimate SGRs from speech signals in real time, this chapter develops automatic estimation algorithms for adults and children. The goal is to develop algorithms for natural speech, with an emphasis on language independence and on robustness to noise and data limitedness. Adults' speech (Sections 3.1 and 3.2) and children's speech (Section 3.3) will be dealt with separately.

3.1 SGR estimation for adults

The approach proposed for adults' speech is based on the vowel-feature contrasts provided by SGRs. Specifically, $Sg1$ estimation relies on the distinction it provides between [+low] and [-low] vowels, while $Sg2$ estimation relies on the distinction it provides between [+back] and [-back] vowels. Although there has been some research regarding the division of *tense* and *lax* [-back] vowels by $Sg3$ [Lul10, CBG09], no strong evidence exists either for or against it. Therefore, $Sg3$ estimation relies simply on its correlation with $Sg2$ (see Section 2.3.4 of Chapter 2). Data from 30 speakers (15 males, 15 females) in the WashU-UCLA Adults corpus were used to develop and train the SGR estimation algorithm. The training speakers were chosen such that their actual SGR frequencies were uniformly distributed in the range of ground truth values shown in Table 2.3. The key ideas behind estimating $Sg1$, $Sg2$ and $Sg3$ are described first followed by a description of the proposed automatic algorithm for natural speech.

3.1.1 Estimating $Sg1$

$Sg1$ estimation relied on three ideas: (1) defining a vocal tract-based measure of vowel height that can be computed using speech signals, (2) defining an $Sg1$-based measure of vowel height that can be computed using speech and subglottal (accelerometer) signals, and (3) developing a model to predict the $Sg1$-based measure from the vocal tract-based measure. In [SG86], the Bark difference between $F1$ and $F0$ was shown to be a reliable indicator of vowel height. However, an acoustic measure involving $F1$ and $F0$ may be problematic for two reasons: (1) $F1$ and $F0$ can be controlled fairly independently of each other, and (2) reliable estimation of $F1$ and $F0$ can be difficult when they are very close to each other (e.g., [-low] vowels produced by high-pitched speakers). Therefore, for the purposes of this dissertation, the Bark difference between $F3$ and $F1$, denoted henceforth as $B_{f3,f1}$, was used as the required vocal tract-based measure of vowel height. The choice of $B_{f3,f1}$ was motivated by a similar acoustic feature, namely the Bark difference between $F3$ and $F2$, denoted henceforth as $B_{f3,f2}$, which has been shown to be a reliable indicator of vowel backness [SG86, Chi85]. $B_{f1,sg1}$ (Bark difference between $F1$ and $Sg1$), which was shown in Chapter 2 to be a good indicator of vowel height, was used as the required $Sg1$-based measure.

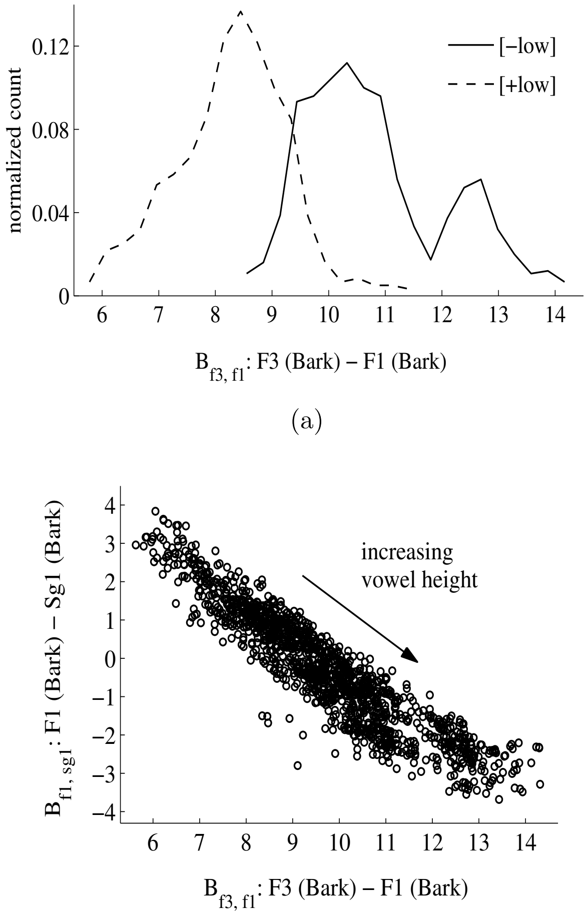
A total of 1350 tokens were used to train a model between $B_{f1,sg1}$ and $B_{f3,f1}$—5 repetitions of each of the 9 hVd monophthongs (see Table 2.1) from each of the 30 training speakers. The formant measurements and ground truth SGRs required for modeling were available through the data analysis procedures described in

Chapter 2. Note that the formant and SGR values were converted from Hertz to Bark (using Eq. (2.3)) before modeling. Figure 3.1 shows normalized histograms of $B_{f3,f1}$ for [-low] and [+low] vowels, and also a scatter plot of $B_{f1,sg1}$ versus $B_{f3,f1}$. $B_{f3,f1}$ separates the two vowel categories at approximately 9.5 Bark confirming that it is, like $B_{f1,sg1}$, a reliable measure of vowel height. More importantly, as evident from Figure 3.1(b), the two measures are strongly correlated ($r =$ -0.9241), suggesting that $B_{f3,f1}$ provides most of the information required for predicting $B_{f1,sg1}$.

A linear regression between $B_{f1,sg1}$ (dependent variable) and the first three powers of $B_{f3,f1}$ (independent variables) resulted in the following model:

$$B_{f1,sg1} = 0.011(B_{f3,f1})^3 - 0.269(B_{f3,f1})^2 + 1.322(B_{f3,f1}) + 2.455. \tag{3.1}$$

Although this regression model had a reasonably high r-squared (r^2) value (0.8702), a non-negligible portion of the variance in $B_{f1,sg1}$ (13%) was still not accounted for. The residual variance was observed to be due to individual speaker differences. Specifically, when the regression was performed separately for each speaker in the training set, the resulting model coefficients showed large spreads in their values: the coefficients related to the linear term ($B_{f3,f1}$) and the intercept term—terms with the two largest weights—were found to have COVs equal to 115 and 162%, respectively. To reduce the inter-speaker variability involved in predicting $B_{f1,sg1}$, two speaker-related features were used: $F3$ and $F0$ (both in Hertz). Note that $F0$ measurements were available through the data analysis procedures described in Chapter 2. When $F3$ and $F0$ (in that order) were added incrementally to the above third-order regression model, r^2 increased from 0.8702 to 0.9255 and from 0.9255 to 0.9724; the increase in each case was statistically significant ($p < 0.001$).



(a)

(b)

Figure 3.1: *(a) Normalized histograms of $B_{f3,f1}$ for [-low] and [+low] vowels; the boundary between the two classes is around 9.5 Bark. (b) Scatter plot (1350 data points) of $B_{f1,sg1}$ versus $B_{f3,f1}$ (r = -0.9241).*

The updated regression model is given by Eq. (3.2):

$$\begin{aligned} B_{f1,sg1} = 0.001(B_{f3,f1})^3 - 0.024(B_{f3,f1})^2 - 0.737(B_{f3,f1}) \\ + 0.002(F3) - 0.007(F0) + 3.903. \end{aligned} \tag{3.2}$$

With the updated model, the COVs of the coefficients related to the linear term and the intercept term reduced to 44 and 49%, respectively. It can thus be said that $F3$ and $F0$ were successful in reducing inter-speaker variability. Given $F0$, $F1$ and $F3$, $Sg1$ can be readily estimated using Eq. (3.2).

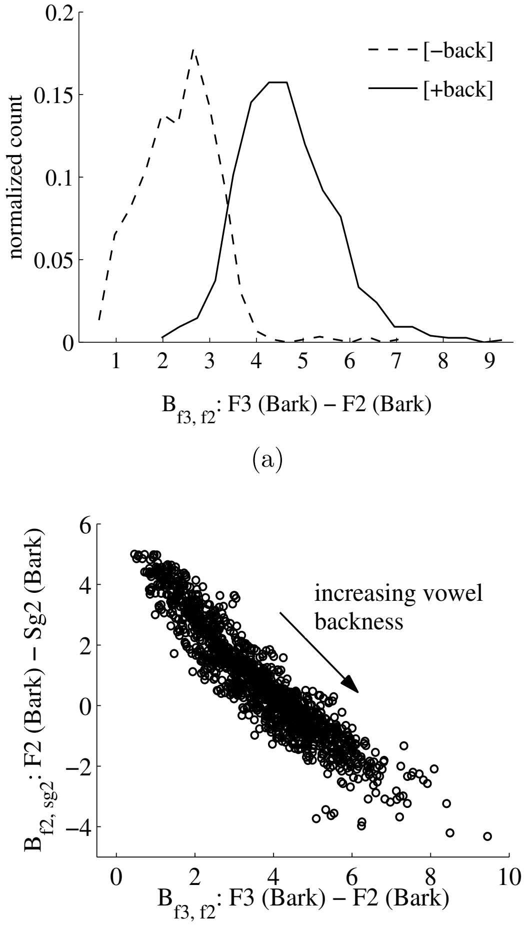
3.1.2 Estimating $Sg2$

$Sg2$ estimation was analogous to $Sg1$ estimation and relied on the following three ideas: (1) defining a vocal tract-based measure of vowel backness, (2) defining an $Sg2$-based measure of vowel backness, and (3) developing a model to predict the $Sg2$-based measure from the vocal tract-based measure. While $B_{f3,f2}$ was used as the required vocal tract-based measure (based on the findings in [SG86] and [Chi85]), $B_{f2,sg2}$ (Bark difference between $F2$ and $Sg2$), which was shown in Chapter 2 to be a good indicator of vowel backness, was used as the required $Sg2$-based measure.

A model between $B_{f2,sg2}$ and $B_{f3,f2}$ was trained using the same 1350 tokens that were used to develop the $Sg1$-estimation model. Figure 3.2 shows normalized histograms of $B_{f3,f2}$ for [-back] and [+back] vowels, and also a scatter plot of $B_{f2,sg2}$ versus $B_{f3,f2}$. $B_{f3,f2}$ separates the two vowel categories at approximately 3.5 Bark (which agrees well with the findings in [SG86] and [Chi85]) confirming that it is, like $B_{f2,sg2}$, a reliable measure of vowel backness. More importantly, as evident from Figure 3.2(b), the two measures are strongly correlated (r = -0.9352), suggesting that $B_{f3,f2}$ provides most of the information required for predicting $B_{f2,sg2}$.

A linear regression between $B_{f2,sg2}$ and the first three powers of $B_{f3,f2}$ resulted in the following model ($r^2 = 0.8905$):

$$B_{f2,sg2} = -0.004(B_{f3,f2})^3 + 0.134(B_{f3,f2})^2 - 1.958(B_{f3,f2}) + 6.182. \tag{3.3}$$



(a)

(b)

Figure 3.2: *(a) Normalized histograms of $B_{f3,f2}$ for [-back] and [+back] vowels; the boundary between the two classes is around 3.5 Bark. (b) Scatter plot (1350 data points) of $B_{f2,sg2}$ versus $B_{f3,f2}$ (r = -0.9352).*

As in the case of $Sg1$ estimation, the residual variance in the above model (11%) was minimized by using $F3$ and $F0$. When $F3$ and $F0$ (in that order) were added incrementally to the regression, r^2 increased from 0.8905 to 0.9429 and from 0.9429 to 0.9713; the increase in each case was statistically significant ($p < 0.001$).

The updated regression model is given by Eq. (3.4):

$$\begin{aligned} B_{f2,sg2} &= 0.001(B_{f3,f2})^3 + 0.009(B_{f3,f2})^2 - 1.089(B_{f3,f2}) \\ &\quad + 0.002(F3) - 0.007(F0) - 0.019. \end{aligned} \tag{3.4}$$

Given $F0$, $F2$ and $F3$, $Sg2$ can be readily estimated using Eq. (3.4).

3.1.3 Estimating $Sg3$

Although there has been some research indicating that $Sg3$ may lie at the boundary of *tense* and *lax* [-back] vowels [Lul10, CBG09], there is not enough evidence to suggest that the phenomenon occurs consistently in all speakers and languages. Therefore, $Sg3$ was simply estimated based on its strong correlation with $Sg2$ (as observed in Chapter 2). With the actual SGRs of the 30 speakers in the training set, a first-order linear regression between $Sg3$ and $Sg2$ resulted in Eq. (3.5) (r^2 = 0.8427):

$$Sg3 = 1.079 \times Sg2 + 763.676, \tag{3.5}$$

which provided an estimate of $Sg3$ once $Sg2$ was estimated using the procedure described in Section 3.1.2. For the training set used here, the RMS error between actual $Sg3$ and the $Sg3$ predicted using Eq. (3.5) (53 Hz) was much smaller than the corresponding RMS error incurred using Eq. (1.2) (275 Hz). Therefore, Eq. (3.5) is more reliable than Eq. (1.2) for estimating $Sg3$ from $Sg2$.

3.1.4 The automatic algorithm

One of the goals of this work was to estimate SGRs from continuous (and natural) speech. Following are the steps involved in going from a given speech signal to the estimates of the speaker's SGRs.

Table 3.1: *SNACK toolkit parameters for automatic formant and pitch tracking (as required by the SGR estimation algorithm for adults).*

parameter	value
window size	30 ms
window spacing	5 ms
window type	Hamming
LPC order	10
LPC method	autocovariance
$F0$ tracking algorithm	ESPS
minimum pitch	60 Hz
maximum pitch	400 Hz

1. Downsample the signal to 7 kHz and pre-emphasize the high frequencies by passing it through a filter with the following frequency response:

   $$H(\omega) = 1 - 0.97e^{-j\omega}.$$

   Since the first three formants of adult speakers usually lie below 3.5 kHz [PB52, HGC95], a sampling rate of 7 kHz suffices for formant tracking.

2. Track $F0$, $F1$, $F2$ and $F3$ automatically using the SNACK sound toolkit. The values of the formant tracking parameters and pitch tracking parameters are shown in Table 3.1. The chosen window size (30 ms) covers at least 2 to 3 pitch periods and the small window spacing (5 ms) ensures smooth formant tracks. The minimum (60 Hz) and maximum (400 Hz) pitch values accommodate the range of pitch frequencies observed in adults' speech.

3. Select all *voiced* frames using SNACK's binary voicing parameter: the prob-

ability of voicing (PV). SNACK sets PV to 1 or 0 depending on whether a given frame is voiced or unvoiced, respectively. Unvoiced frames need to be discarded because the fundamental and formant frequencies (required for SGR estimation) are not well defined for unvoiced speech.

4. Perform the following sequence of operations for each voiced frame in the given speech signal. The superscript k in all the following operations indicates the k^{th} voiced frame.

   - Obtain Bark values corresponding to $F1^k$, $F2^k$ and $F3^k$ using Eq. (2.3).
   - Compute $B^k_{f3,f1}$ and $B^k_{f3,f2}$.
   - Predict $B^k_{f1,sg1}$ from $\{B^k_{f3,f1},\ F3^k,\ F0^k\}$ using Eq. (3.2), and $B^k_{f2,sg2}$ from $\{B^k_{f3,f2},\ F3^k,\ F0^k\}$ using Eq. (3.4).
   - Recover $Sg1^k$ and $Sg2^k$ in Bark:

   $$\begin{aligned} Sg1^k(\text{Bark}) &= F1^k(\text{Bark}) - B^k_{f1,sg1}, \\ Sg2^k(\text{Bark}) &= F2^k(\text{Bark}) - B^k_{f2,sg2}. \end{aligned}$$

   - Convert $Sg1^k$ and $Sg2^k$ from Bark to Hertz by inverting Eq. (2.3).

5. At the end of Step 4, every voiced frame in the signal is associated with an estimate of $Sg1$ and $Sg2$. Then, estimate the speaker's $Sg1$ and $Sg2$ as the averages of the corresponding frame-level estimates:

   $$\begin{aligned} Sg1 &= \frac{1}{N_v}\sum_{k=1}^{N_v} Sg1^k \\ Sg2 &= \frac{1}{N_v}\sum_{k=1}^{N_v} Sg2^k, \end{aligned}$$

   where N_v denotes the total number of voiced frames.

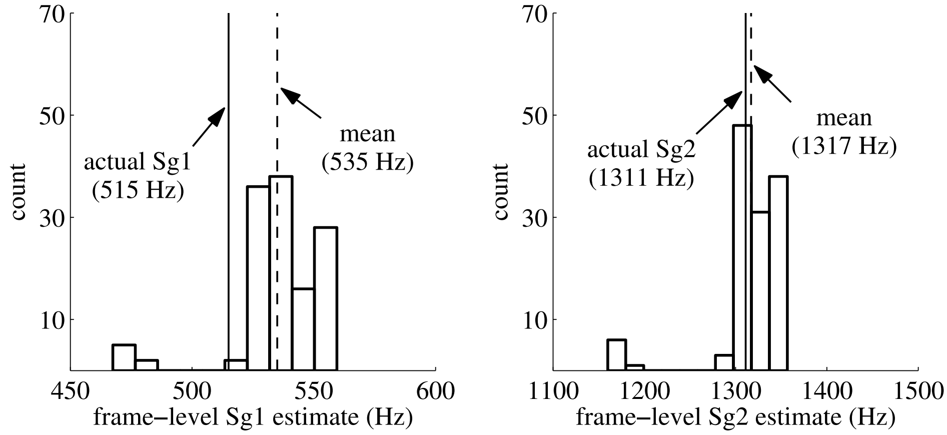


Figure 3.3: *Distributions of frame-level $Sg1$ estimates (left) and frame-level $Sg2$ estimates (right) obtained by applying the automatic SGR estimation algorithm to a microphone recording of "I said a heed again" (not in the training set) in the WashU-UCLA Adults corpus. In each case, the mean of the distribution is close to the actual SGR value.*

6. Estimate the speaker's $Sg3$ by plugging the above $Sg2$ estimate into Eq. (3.5).

It must be noted that while the regression models for SGR estimation were trained using formant frequencies and $F0$ measured in steady-state vowels, the actual algorithm was designed to use all voiced frames irrespective of their origin: vowels (steady-state or otherwise), voiced consonants or transition regions between voiced and unvoiced sounds. Although such an approach is expected to yield a few 'undesirable' frame-level estimates, natural speech contains enough vowel segments to skew the *averages* of frame-level estimates towards the actual ('desired') SGR values. Figure 3.3 illustrates with an example that the proposed frame-based approach is effective in estimating $Sg1$ and $Sg2$ from continuous speech. However, it is important to note that the proposed algorithm cannot estimate SGRs from purely unvoiced speech (e.g., whispered speech).

3.1.5 Analysis of noise robustness

The efficacy of the proposed estimation algorithm depends on the performance of SNACK (with regard to pitch and formant tracking). SNACK is a popular tool and is known to be accurate in clean conditions for both pitch tracking [TA13] and formant tracking [DCP06]. To assess its efficacy in noise (for the purpose of SGR estimation), noisy speech files were created using a subset of the WashU-UCLA Adults corpus, and the pitch and formant contours obtained from them were visually analyzed. Babble noise, which is more realistic than white noise, was added to the clean speech files using the Filtering and Noise-adding Tool (FaNT) [Hir05]. The noise file was obtained from the NOISEX-92 database [VS93], and the signal-to-noise ratio (SNR) was varied between 0 and 10 dB. The contours obtained from a given noisy utterance were considered reliable if they closely matched the contours obtained from the corresponding clean utterance. Note that pitch and formant tracking in noise are challenging problems if accurate results are required at every time instant. However, since the proposed SGR estimation algorithm uses only voiced segments and relies on averaging the frame-level estimates, the pitch and formant tracking requirements are not very stringent.

In general, SNACK's estimates of pitch and formants were found to be reliable in voiced regions with high local SNRs; an example is shown in Figure 3.4. These high-SNR voiced regions belonged largely to vowels (which are important for SGR estimation). Also, the unvoiced-to-voiced error (i.e., frames classified as unvoiced in clean but voiced in noise) was less than 5%, on average. This is beneficial to estimating SGRs because unvoiced segments must be discarded in the estimation process. Note that voiced-to-unvoiced errors are relatively harmless as long as at least a few voiced frames are correctly detected. In summary, therefore, SNACK

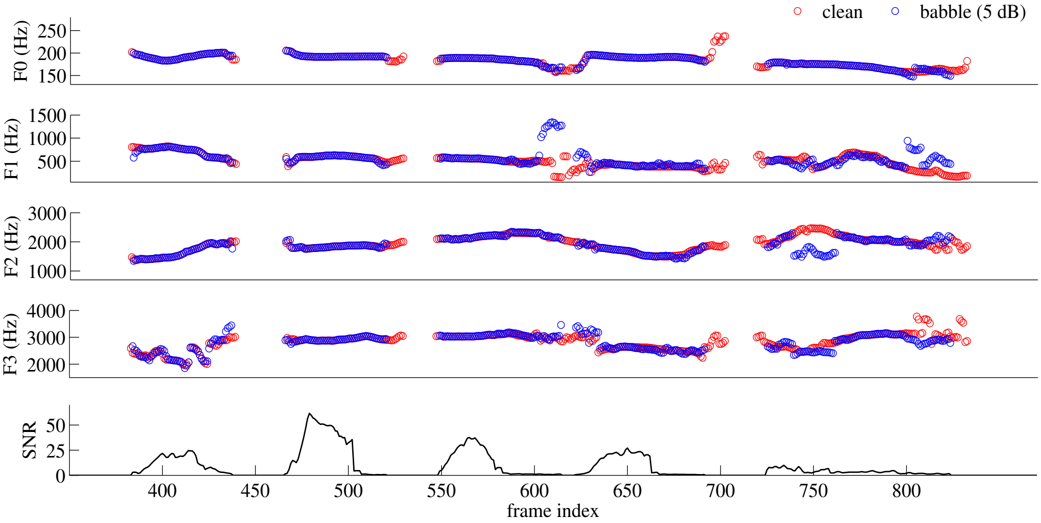


Figure 3.4: *Contours of $F0$ and formants obtained using SNACK for a sample utterance ("I said a hod again") in the WashU-UCLA Adults corpus. The contours in red correspond to the original clean signal and the contours in blue correspond to its noisy version (babble noise was added at an average SNR of 5 dB). The bottom panel shows the frame-by-frame SNRs for the noisy signal (not in dB, to better distinguish the high-SNR from the low-SNR regions). Note that the contours and SNRs are shown for only voiced segments (which are used for SGR estimation).*

is reasonably accurate for the purpose of SGR estimation in noisy adults' speech, even at an SNR as low as 0 dB.

3.2 Experiments with adults' speech

The proposed SGR estimation algorithm was evaluated using microphone recordings of 20 speakers (10 males, 10 females; different from the speakers in the training set) in the WashU-UCLA Adults corpus and all 6 speakers in the WashU-

UCLA Bilingual Adults corpus. The algorithm was evaluated additionally on the MIT tracheal resonance (TR) database [Son04].

The MIT TR database comprises time-synchronized recordings of speech and subglottal acoustics from 14 adult (7 male, 7 female) native AE speakers aged between 18 and 78 years. The recorded material comprises carrier phrases of the form *"____, say ____ again,"* where the blank is one of 16 dVd or 16 hVd words. There are a total of 160 utterances per speaker (5 repetitions per word). Further details regarding the database can be found in [Son04].

The SGR estimation algorithm was applied to both the carrier phrases in the WashU-UCLA corpora and the vowel tokens isolated from them. These experiments served to analyze the algorithm's performance with regard to spoken content, language (AE versus MS), and the amount of speech data used for estimation. Note that the algorithm was trained on AE vowels but applied to MS data without any modification.

Noisy speech files were created using the MIT TR database and the algorithm was applied to clean as well as noisy data. These experiments served to analyze the algorithm's robustness to noise and its efficacy under varying recording conditions (note that the WashU-UCLA corpora and the MIT TR database were collected with different equipment and in different recording conditions). The speech files were corrupted with four different noise types (babble, white, factory and pink—all from NOISEX-92) at three different SNRs (0, 5 and 10 dB). Ground truth SGRs were obtained from accelerometer recordings as described in Section 2.2.2.

3.2.1 Performance metrics

For ease of representation, let actual SGR values be denoted as $Sg1_a$, $Sg2_a$ and $Sg3_a$, and estimated SGR values be denoted as $Sg1_e$, $Sg2_e$ and $Sg3_e$. The SGR estimation algorithm was evaluated using two performance metrics: (1) average root mean squared error (RMSE), and (2) average mean-relative standard deviation (MSD). While RMSE quantifies estimation *accuracy*, MSD quantifies the *consistency* of estimation. Denoting the number of test speakers as N_s and the number of test utterances (isolated vowels or sentences) for the i^{th} speaker as M_i, the definitions of RMSE and MSD for the K^{th} SGR ($K = 1, 2, 3$) are as follows.

$$\text{RMSE} = \frac{1}{N_s}\sum_{i=1}^{N_s}\text{RMSE}^i, \qquad \text{RMSE}^i = \sqrt{\frac{1}{M_i}\sum_{j=1}^{M_i}(SgK_e^{ij} - SgK_a^i)^2} \tag{3.6}$$

$$\begin{aligned} \text{MSD} &= \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\frac{\sigma_e^i}{\mu_e^i}\times 100\right), \\ \mu_e^i &= \frac{1}{M_i}\sum_{j=1}^{M_i} SgK_e^{ij} \qquad \sigma_e^i = \sqrt{\frac{1}{M_i}\sum_{j=1}^{M_i}(SgK_e^{ij} - \mu_e^i)^2} \end{aligned} \tag{3.7}$$

In Eq. (3.6), SgK_a^i denotes the actual value of the K^{th} SGR of the i^{th} test speaker. In Eqs. (3.6) and (3.7), SgK_e^{ij} denotes the estimated value of the K^{th} SGR corresponding to the j^{th} utterance of the i^{th} test speaker. Note the resemblance between the RMSE definition of Eq. (3.6) and the average within-speaker standard deviation of SGR measurements reported in Table 2.2—they are both computed in reference to actual SGR values. Therefore, the numbers in Table 2.2 can be used as a rough guideline for interpreting the results of SGR estimation.

3.2.2 Results for the WashU-UCLA corpora

(a) Estimation using isolated vowels:

The algorithm was evaluated on (1) 13 AE vowels (10 tokens per vowel per speaker) in the hVd word list of the WashU-UCLA Adults corpus (the approximant [ɹ] was not used), and (2) all 7 vowels (21 tokens per vowel per speaker) in

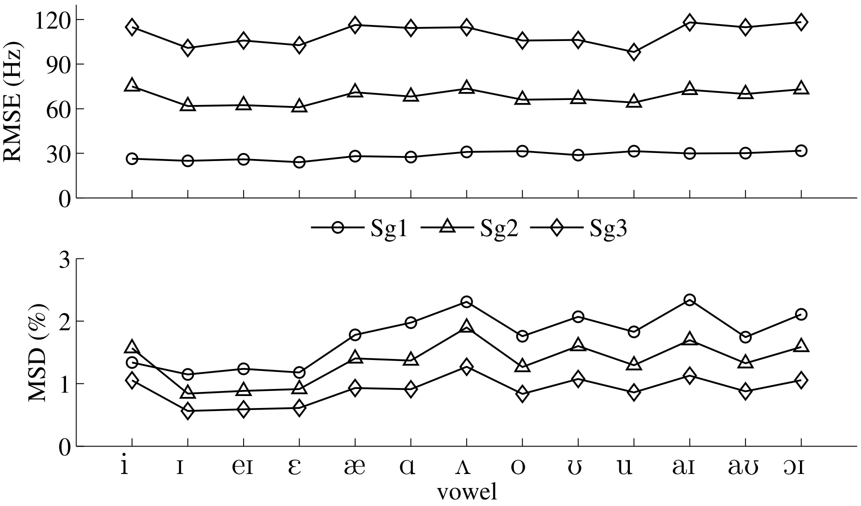


Figure 3.5: *SGR estimation using isolated vowels: overall RMSE and MSD corresponding to the monophthong and diphthong vowels recorded in the WashU-UCLA Adults corpus. For practical purposes, the performance can be considered to be vowel independent.*

the AE CVb word list and the MS CVb word list of the WashU-UCLA Bilingual Adults corpus.

Figure 3.5 shows the overall (males and females combined) RMSE and MSD corresponding to each monophthong and diphthong vowel in the WashU-UCLA Adults corpus. The following two observations can be made. (1) The algorithm's performance is slightly vowel dependent; this might be attributed, at least in part, to differences in the accuracy of automatic formant tracking. Specifically, it is easier to track formants when they are fairly 'steady' and well separated from one another (e.g., [ɛ], [æ] and [ɪ]) than when two or more of them are very closely spaced (e.g., [i] and [ɑ]) or rapidly changing over time (e.g., [aɪ] and [ɔɪ]). Nevertheless, the observed vowel dependence in performance is small enough to be ignored for practical purposes: RMSE ranges from 24 Hz ([ɛ]) to 32 Hz ([ɔɪ])

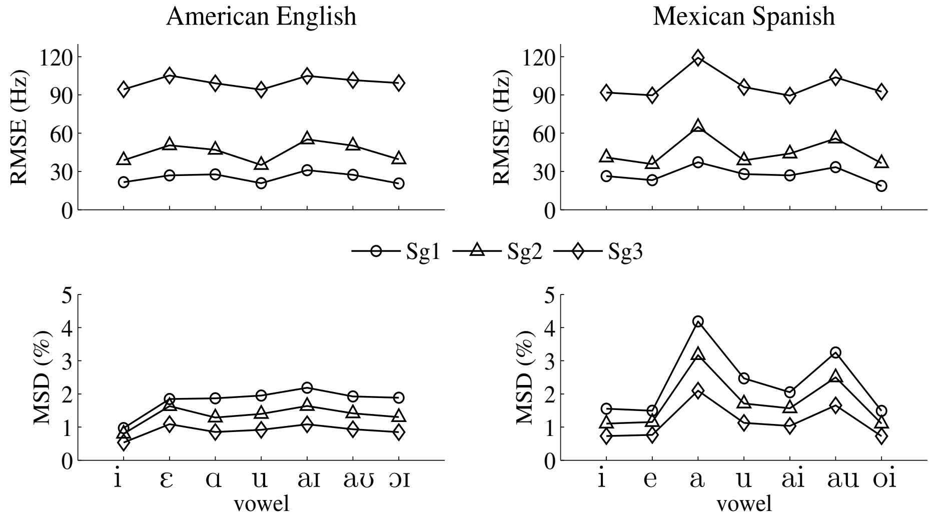


Figure 3.6: *SGR estimation using isolated vowels: overall RMSE and MSD corresponding to the AE (left) and MS (right) vowels recorded in the WashU-UCLA Bilingual Adults corpus. For practical purposes, the performance can be considered to be language independent.*

for $Sg1$, from 61 Hz ([ɛ]) to 75 Hz for $Sg2$ ([i]), and from 98 Hz ([u]) to 118 Hz ([ɔɪ]) for $Sg3$. (2) For all three SGRs, the RMSEs are on the order of (about 1 to 2 times) the average within-speaker standard deviations shown in Table 2.2, and the MSDs are less than 3%. Therefore, the algorithm's performance can be considered accurate within the measurement error.

Figure 3.6 shows the overall RMSE and MSD corresponding to all the AE and MS vowels in the WashU-UCLA Bilingual Adults corpus. As in the case of the native AE speakers (Figure 3.5), the algorithm's performance is only slightly vowel dependent. However, the more important observation here is that the algorithm is equally accurate and consistent for AE and MS vowels (with the exception of [au] and, particularly, [a]) despite being trained using AE data only. This *language independent* nature of the algorithm can be attributed to two factors. (1) Bark *differences* between vocal-tract formants ($B_{f3,f1}$ and $B_{f3,f2}$), which are es-

Table 3.2: *SGR estimation results using one utterance (less than 2 seconds) per estimate: RMSEs and MSDs for the WashU-UCLA Adults corpus. For practical purposes, the performance can be considered to be gender independent.*

	RMSE (Hz)			MSD (%)		
	$Sg1$	$Sg2$	$Sg3$	$Sg1$	$Sg2$	$Sg3$
Males	22	64	97	1.9	1.4	0.9
Females	32	65	125	1.4	1.2	0.8
Overall	**25**	**61**	**104**	**1.6**	**1.2**	**0.8**

Table 3.3: *SGR estimation results using one utterance (less than 2 seconds) per estimate: RMSEs and MSDs for the WashU-UCLA Bilingual Adults corpus.*

	$Sg1$		$Sg2$		$Sg3$	
	AE	MS	AE	MS	AE	MS
RMSE (Hz)	20	18	40	32	100	90
MSD (%)	1.2	1.2	0.9	0.9	0.6	0.6

sential to the SGR estimation algorithm, do not contain any language-specific information about vowels; they are simply acoustic measures of vowel height and backness. (2) Acoustic features such as $F3$ and $F0$, which provide auxiliary information for estimating SGRs, carry speaker-related information and hence do not vary significantly with the language spoken.

(b) Estimation using continuous speech (carrier phrases):

The test set consisted of carrier phrases from the WashU-UCLA Adults corpus (140 per speaker) and the WashU-UCLA Bilingual Adults corpus (147 in AE and 147 in MS, per speaker). All utterances were less than 2 seconds in duration.

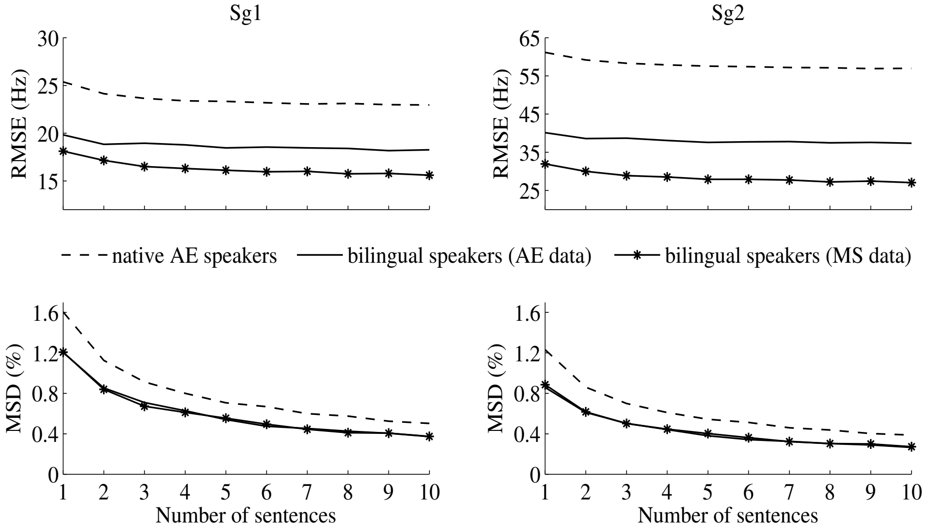


Figure 3.7: *SGR estimation results using up to 10 utterances per estimate: overall RMSE and MSD—corresponding to $Sg1$ (left) and $Sg2$ (right)—as a function of the number of utterances used. Note that in the bottom two panels, the curves for MS data overlap with the curves for AE data.*

The first three SGRs were estimated from each utterance in the test set. Hence, every SGR estimate was obtained using less than 2 seconds of speech. Table 3.2 shows the RMSE and MSD for the WashU-UCLA Adults corpus (separated by gender), and Table 3.3 shows the results for the WashU-UCLA Bilingual Adults corpus (separated by language). The following observations can be made from Table 3.2. (1) Compared to males, females have larger RMSEs but smaller MSDs (especially in the case of $Sg3$). The slightly larger values of females' RMSEs might be attributed, at least in part, to the fact that LPC-based formant estimation (used in the SNACK toolkit) is less accurate for speakers with high-pitched voices (usually females) as compared to speakers with low-pitched voices (usually males) [Mak75]. However, this gender dependence in performance is small enough to be ignored for practical purposes. (2) The overall RMSE for $Sg3$ estimation is

$\sim$100 Hz. In comparison, $Sg3$ estimation using Eq. (1.2) (which, like Eq. (3.5), requires an estimate of $Sg2$) incurs an overall RMSE in excess of 300 Hz. This reiterates that Eq. (3.5) is a more accurate model of the relation between $Sg2$ and $Sg3$. From Table 3.3, it can be observed once again that the algorithm is language independent. Also, as observed earlier, the RMSEs for both the native AE and bilingual speakers are comparable to the average within-speaker standard deviations reported in Table 2.2; the MSDs are less than 2% (implying that the variance in the estimates is very small).

The results in Tables 3.2 and 3.3 were obtained using one speech utterance (less than 2 seconds) per estimate. To see if the algorithm performed better with more data, the SGRs of the test speakers were estimated using up to 10 utterances per estimate. Figure 3.7 shows the overall RMSE and MSD—corresponding to $Sg1$ and $Sg2$—as a function of the number of utterances used for estimation ($Sg3$ shows the same trend as $Sg2$ since it is estimated from $Sg2$). As the number of sentences increases from 1 to 10, RMSE decreases slightly (by 11% for $Sg1$ and 10% for $Sg2$, on average), but MSD decreases considerably (by 67% for both $Sg1$ and $Sg2$, on average). Therefore, the algorithm's performance does improve as more speech data becomes available. The more attractive feature of the algorithm, however, is that it performs well even when data are limited.

3.2.3 Results for the MIT tracheal resonance database

The test set comprised 280 utterances (20 utterances chosen at random for each of the 14 speakers in the database) in each evaluation condition: clean speech, and speech corrputed with babble, white, factory and pink noise types (at SNRs of 0, 5 and 10 dB). The utterances were 2–3 seconds long, on average, and SGR estimates were obtained on a per-utterance basis. Babble noise was found to

Table 3.4: *SGR estimation results using one utterance (less than 3 seconds) per estimate: overall RMSEs (Hz) for the MIT tracheal resonance database in clean and babble noise (at different SNRs) conditions.*

	Clean	10 dB	5 dB	0 dB	Average
$Sg1$	28	30	31	34	**31**
$Sg2$	63	64	63	67	**64**
$Sg3$	113	116	114	119	**116**

be the most challenging condition. Therefore, for brevity, Table 3.4 shows the RMSEs for the case of babble noise only; results for the other noise types were either similar or slightly better.

Two important observations can be made from Table 3.4. (1) The SGR estimation algorithm is robust to noise, even at an SNR of 0 dB. This is attributable to the fact that the SNACK toolkit provides reasonably-accurate voicing detection, and pitch and formant estimation in voiced speech segments with high local SNR (see Figure 3.4). Note that the algorithm's performance is not severely affected despite the fact that babble noise has a large portion of its energy in the low frequencies (which might affect the estimation of $F0$ and the first two formants). (2) The RMSEs are comparable to those obtained for the WashU-UCLA corpora (see Tables 3.2 and 3.3), implying that the proposed algorithm is effective under varying recording conditions.

3.3 SGR estimation for children

For adults' speech, $Sg1$ was estimated using a regression model of the relationship between two vowel-height measures: $B_{f1,sg1}$ and $B_{f3,f1}$ (see Eq. (3.2)). Similarly,

$Sg2$ was estimated using a model of the relationship between two vowel-backness measures: $B_{f2,sg2}$ and $B_{f3,f2}$ (see Eq. (3.4)). $F3$ and $F0$ (in Hz) were used in the above models as auxiliary, speaker-related features. Since the relationship between Hertz and Bark frequencies is nonlinear—see Eq. (2.3), Eqs. (3.2) and (3.4) result implicitly in nonlinear relationships between $Sg1$ or $Sg2$, and the vocal tract parameters $F0$, $F1$, $F2$, and $F3$.

Similar regression models were trained for children (using a subset of the WashU-UCLA Kids corpus), but their r^2 values were found to be significantly lower than the r^2 values associated with Eqs. (3.2) and (3.4). This can be attributed, at least in part, to the higher degree of acoustic variability (with regard to the fundamental and formant frequencies) that children's speech is known to exhibit [LPN99]. One way of compensating for this large variability would be to train an explicit nonlinear model of the relationships between SGRs and vocal-tract parameters. In this dissertation, artificial neural networks (ANNs) were used as the required nonlinear models. The key idea was to train an ANN with $F0$, $F1$, $F2$ and $F3$ as inputs, and $Sg1$ and $Sg2$ as outputs. $Sg3$ was simply estimated based on its strong correlation with $Sg2$ (see Section 2.3.4).

Data from 25 speakers—ground truth SGRs and measurements of $F0$ and formant frequencies (see Chapter 2)—in the WashU-UCLA Kids corpus were used for training. An ANN with one hidden layer and 20 nodes (each having a sigmoid nonlinearity) was trained using MATLAB's Neural Network Toolbox. Raw measurements (not speaker-wise averages) of $F0$, $F1$, $F2$ and $F3$ formed the ANN's inputs, and the corresponding ground truth values of $Sg1$ and $Sg2$ formed the ANN's outputs. A first-order linear model was trained to estimate $Sg3$ from $Sg2$:

$$Sg3 = 1.47 \times Sg2 + 203. \tag{3.8}$$

Given a speech signal from a child speaker, the first three SGRs are estimated as follows. (1) Downsample the signal to 8 kHz and pre-emphasize it with the high-pass filter: $H(\omega) = 1 - 0.97e^{-j\omega}$. Since the first three formants of children typically lie below 4 kHz [LPN99], a sampling rate of 8 kHz suffices for formant tracking. (2) Obtain contours of $F0$, $F1$, $F2$ and $F3$ automatically using the SNACK toolkit. Use the same parameters as in Table 3.1, except for an LPC order of 12 (instead of 10) and a maximum pitch value of 450 Hz (instead of 400 Hz). A higher LPC order is necessary to account for the fact that children's voices, in general, are breathier than those of adults (resulting in a higher degree of spectral tilt) [Shu10]. (3) Select voiced frames using SNACK's binary voicing parameter. (4) Estimate $Sg1$ and $Sg2$ for each voiced frame by feeding its pitch and formant values as inputs to the pre-trained ANN. (5) Estimate $Sg1$ and $Sg2$ for the given utterance by averaging the corresponding frame-level estimates. (6) Estimate $Sg3$ using Eq. (3.8) and the estimated value of $Sg2$ obtained in step (5). Steps (1)–(6) are summarized in Figure 3.8. Note that the above algorithm is similar to the estimation algorithm for adults, except that $Sg1$ and $Sg2$ are estimated using an ANN instead of Eqs. (3.2) and (3.4).

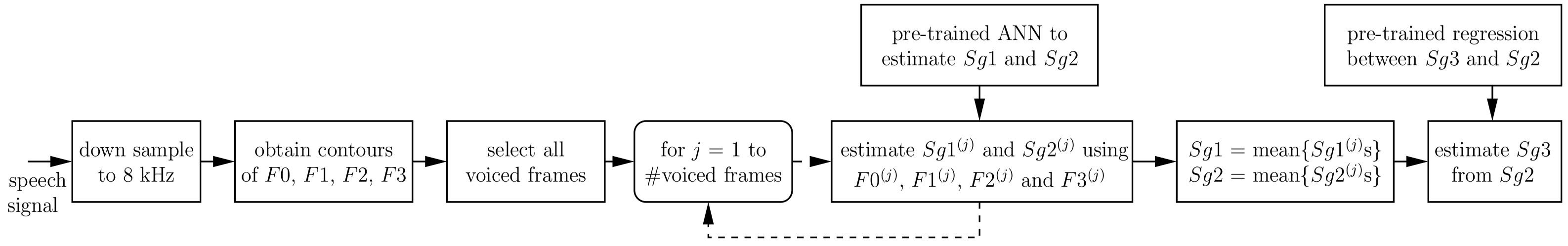

pre-trained ANN to estimate $Sg1$ and $Sg2$
pre-trained regression between $Sg3$ and $Sg2$
speech signal
down sample to 8 kHz
obtain contours of $F0$, $F1$, $F2$, $F3$
select all voiced frames
for $j = 1$ to #voiced frames
estimate $Sg1^{(j)}$ and $Sg2^{(j)}$ using $F0^{(j)}$, $F1^{(j)}$, $F2^{(j)}$ and $F3^{(j)}$
$Sg1 = \text{mean}\{Sg1^{(j)}\text{s}\}$
$Sg2 = \text{mean}\{Sg2^{(j)}\text{s}\}$
estimate $Sg3$ from $Sg2$

Figure 3.8: *Block diagram illustrating the steps involved in estimating the first three SGRs from speech signals of children. The SNACK toolkit was used for pitch and formant contour estimation (block 2) and voiced/unvoiced classification (block 3).*

Table 3.5: *RMSEs (in Hz) for SGR estimation from children's speech at 5 dB SNR. Data from 18 speakers (42 utterances per speaker) in the WashU-UCLA Kids corpus were used for evaluation. Results for $Sg3$ are based on data from 6 speakers only.*

	Clean	Babble	Car	Pink	White	Average
$Sg1$	64	74	63	64	64	**66**
$Sg2$	146	173	143	151	153	**153**
$Sg3$	111	196	111	150	146	**143**

The noise robustness of the above algorithm depends on the performance of SNACK (as in the case of adults). The efficacy of SNACK, in noise, was analyzed subjectively as per the procedure described in Section 3.1.5. In general, SNACK's $F0$ and formant estimates (in voiced speech) were found to be reliable at SNRs of 5 dB or more. In the case of adults, on the other hand, SNACK was found to be effective at SNRs as low as 0 dB. This suggests that improved $F0$ and formant tracking might be required for children's speech at low SNR levels, a topic beyond the scope of this dissertation.

3.3.1 Experiments with children's speech

To evaluate the SGR estimation algorithm for children, speech data from 18 speakers (not part of the training set) in the WashU-UCLA Kids corpus were used. For each speaker, 42 utterances (3 utterances corresponding to each of the 14 target words) were randomly chosen from the hVd recordings. The clean speech data were corrputed with 4 different noise types—babble, car, pink, and white—at an SNR of 5 dB (the lowest SNR level at which SNACK was found to be effective with children's speech, for the purpose of SGR estimation). Noise files were taken from the NOISEX-92 database, and FaNT was used to create the noisy data sets. Ground truth values of $Sg1$ and $Sg2$ were available for all 18 speakers, but ground truth $Sg3$ values were available for only 6 of them. RMSE, as defined by Eq. (3.6), was used as the performance metric. Note that the above noise types were chosen to mimic situations where children are likely to use speech-based applications; factory noise, which was used in the case of adults, was therefore not used.

The results are shown in Table 3.5. The RMSEs are about 2 to 3 times the average within-speaker standard deviations shown in Table 2.2—this is slightly worse compared to the performance achieved for adults (see Section 3.2.2). Note that the average RMSE for $Sg3$ is less than the average RMSE for $Sg2$. This is probably because of the small test sample used for $Sg3$ estimation (RMSEs are expected to be higher for $Sg3$ because its estimation depends on $Sg2$ through Eq. (3.8)). The algorithm is robust to different noise conditions, except babble. The larger performance drop (with respect to clean) in babble noise is attributable to the fact that speech babble has a large portion of its energy in the low frequencies (which affects the estimation of $F0$ and the first two formants). Note that the effect of babble noise was less severe in the case of adults (see Table 3.4), which again points to the fact that $F0$ and formant tracking are more challenging with children's speech (especially in noise).

3.4 Conclusion

In this chapter, automatic SGR-estimation algorithms for adults as well as children were developed and evaluated. The emphasis was on phonetic and langauge independence, noise robustness, and effectiveness with limited data.

For adults, $Sg1$ and $Sg2$ are estimated based on the fact that they form natural boundaries between [+low]/[-low] and [+back]/[-back] vowels, respectively. $Sg3$ is estimated based on its correlation with $Sg2$. The algorithm uses the SNACK toolkit for automatic pitch and formant tracking, and provides utterance-level SGR estimates by averaging the frame-level estimates obtained in voiced speech segments. Data from 30 native AE speakers in the WashU-UCLA Adults corpus are used for training, and the algorithm is evaluated using data from (1) 20 speakers in the WashU-UCLA Adults corpus, (2) 6 speakers in the WashU-UCLA Bilingual Adults corpus, and (3) 14 speakers in the MIT TR database. The algorithm's performance—in terms of RMSE (a measure of accuracy) and MSD (a measure of consistency)—is found to be practically independent of vowel content and language (AE versus MS). In addition, the algorithm is robust to different noise types (babble, factory, white and pink) at SNRs as low as 0 dB. Using just 2–3 seconds of speech, $Sg1$, $Sg2$ and $Sg3$ can be estimated, on average, to within 31, 64 and 116 Hz, respectively—these errors are 1 to 2 times the average within-speaker standard deviations in measured SGR frequencies.

In the case of children, $Sg1$ and $Sg2$ are estimated using an ANN-based nonlinear mapping between vocal-tract parameters ($F0$, $F1$, $F2$ and $F3$) and ground truth SGRs. $Sg3$ is estimated as in the case of adults. The algorithm is trained and evaluated using data from 25 and 18 speakers, respectively, in the WashU-UCLA Kids corpus. The algorithm (based on the SNACK toolkit as in the case of adults) is robust to different noise types (babble, car, white and pink) at SNRs of 5 dB or more. Robustness at lower SNR levels could possibly be achieved with more sophisticated pitch- and formant-tracking algorithms. The average RMSEs incurred in estimating $Sg1$, $Sg2$ and $Sg3$ are 66, 153 and 143 Hz, respectively—these errors are 2 to 3 times the average within-speaker standard deviations in measured SGR frequencies.

CHAPTER 4

Speaker height estimation using SGRs

This chapter presents a practical, physiologically-motivated approach to height estimation based on the use of SGRs. The approach is based on the assumption that the 'acoustic length' of the subglottal system is proportional to speaker height. 'Acoustic length' is the length of an equivalent uniform tube (closed at the glottal end and open at the distal end) whose input impedance approximates the actual input impedance of the subglottal system. The above assumption is supported by [LAA11], in which it was shown that the first three SGRs can be modeled as the resonances of a simple uniform tube whose length is equal to the height of the speaker divided by an empirically-determined scaling factor. The assumption finds further evidence in the correlations observed between height and SGR frequencies (see Chapter 2).

The proposed approach is essentially a combination of two key ideas: (1) modeling the correlations between speaker height and ground truth SGRs, and (2) estimating SGRs using speech signals (see Chapter 3). While the approach is applicable to adults and children alike, this chapter presents empirical results for adults only—speech databases with height information were unavailable in the case of children. It will be shown that the proposed approach is effective with limited data, requires very little training, and is robust to noise.

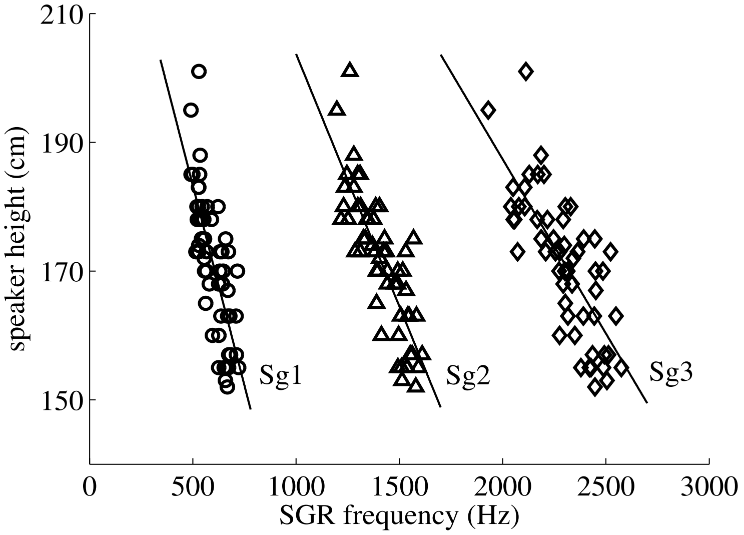


Figure 4.1: *Scatter plots of speaker height versus the first three SGRs (data correspond to speakers in the WashU-UCLA Adults corpus and the WashU-UCLA Bilingual Adults corpus). The solid lines represent first-order linear regression fits to the data. Speaker height correlates more strongly with* $Sg2$ *(r = -0.8256) than with* $Sg1$ *(r = -0.7586) or* $Sg3$ *(r = -0.7627).*

4.1 Method

The relationships between speaker height and SGRs were modeled using the ground truth SGRs and self-reported heights (in centimeters) of all speakers in the WashU-UCLA Adults corpus and the WashU-UCLA Bilingual Adults corpus (56 in total). Male speaker heights ranged from 165 to 201 cm while female speaker heights ranged from 152 to 175 cm. All three SGRs correlated negatively with height, but $Sg2$ accounted for more height variance ($r^2 = 0.68$) than $Sg1$ ($r^2 = 0.58$) or $Sg3$ ($r^2 = 0.58$). In contrast, [Dus05] reported that 31 vocal tract-based features (MFCCs, LPCs and formants) were necessary to explain 57% of the variance in height. This reinforces the hypothesis that SGRs are more suitable for height estimation than are vocal-tract features.

Using first-order linear regression, the following empirical relations were obtained between speaker height and SGR frequencies:

$$h = -0.124 \times Sg1 + 245.476 \tag{4.1}$$

$$h = -0.078 \times Sg2 + 282.107 \tag{4.2}$$

$$h = -0.054 \times Sg3 + 295.659, \tag{4.3}$$

where h denotes speaker height (in centimeters). Figure 4.1 shows scatter plots of speaker height versus ground truth SGRs, as well as the corresponding regression fits to the data. Given a speech signal, speaker height was estimated by first estimating $Sg1$, $Sg2$ and $Sg3$ using the algorithms developed in Chapter 3, and then using Eqs. (4.1), (4.2) or 4.3, respectively. The scatter plots suggest that the linear fits provided by Eqs. (4.1)–(4.3) are probably not the best models to predict the height of an unknown speaker. However, given that there were only 56 data points in the training set, first-order linear regression was considered the most appropriate solution. Regression relations were derived for the WashU-UCLA Kids corpus as well (although height-estimation experiments were not conducted); they were found to be very similar to Eqs. (4.1)–(4.3).

Although speaker height correlated most strongly with $Sg2$, all three linear models were considered for height estimation since the accuracy of the proposed method was affected not only by the correlations between height and SGRs, but also by the accuracy of SGR estimation. Note, however, that the interaction between height-estimation accuracy and SGR-estimation accuracy is nontrivial—since the correlations between height and SGRs are all less than 1, a small SGR-estimation error may not always result in a small height-estimation error. Figure 4.2 shows an example in support of this argument.

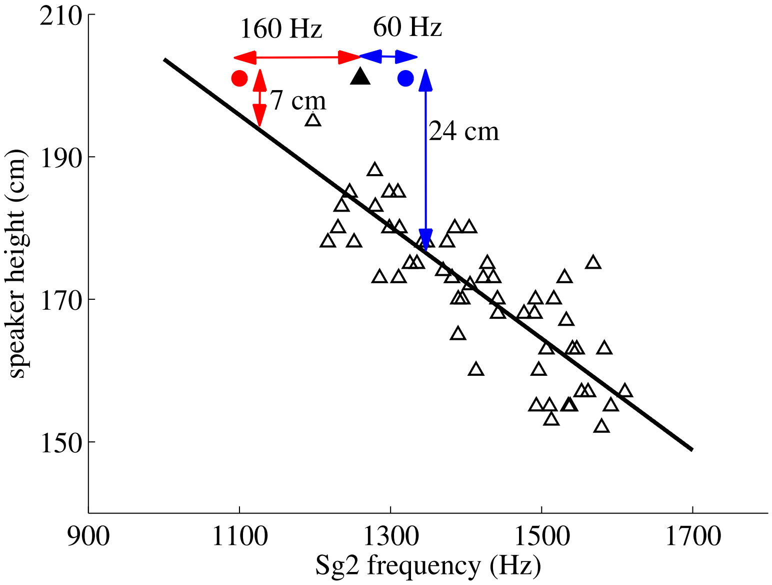


Figure 4.2: *An example to illustrate the fact that small SGR-estimation errors may not always result in small height-estimation errors. The solid black triangle corresponds to a training speaker with an actual $Sg2$ of 1260 Hz and height equal to 201 cm. If this speaker's $Sg2$ is overestimated by 60 Hz (blue circle), the error in height estimation is 24 cm. On the other hand, if $Sg2$ is underestimated by 160 Hz (red circle), the error in height estimation is only 7 cm.*

4.2 Experimental setup

The TIMIT database was used to evaluate the proposed height-estimation algorithm. The database contains a total of 6300 sentences in American English, 10 sentences spoken by each of 630 speakers (438 males, 192 females) from 8 major dialect regions of the United States. The average sentence duration is about 3 seconds. Data are sampled at 16 kHz and quantized at 16 bits/sample. The database also contains the height of each speaker in feet and inches. Further details regarding the database can be found in [Gar88b].

Data from 604 speakers (431 males, 173 females; a total of 6040 sentences) were used for evaluation. The remaining 26 speakers in the database were not part of the evaluation because their heights were outside the range spanned by the training data (used to derive Eqs. (4.1)–(4.3)). To see if the proposed method could estimate height using narrowband telephone speech (which can potentially benefit forensic applications—see [PH97]), a narrowband evaluation set was created by filtering TIMIT data with the ITU-T G.712 filter, which has a flat frequency response between 300 and 3400 Hz [ITU01]. Furthermore, to assess the noise robustness of the proposed method, babble, white, pink and factory noise (from the NOISEX-92 database) were added (using FaNT) to the narrowband evaluation set at an SNR of 0 dB. Note that the SGR estimation algorithm for adults, which forms the basis for height estimation, was found to be robust to different noise types at SNRs as low as 0 dB (see Chapter 3).

The height-estimation algorithms of [GMF10a] and [GMF10b] are the most accurate algorithms known to date—they yield an MAE (mean absolute error) of 5.3 cm and an RMSE of 6.8 cm over 168 speakers in the TIMIT database. For comparison purposes, MAE and RMSE were used as the performance metrics. The following equations were used to calculate MAE and RMSE:

$$\mathrm{MAE} = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{1}{M_i}\sum_{j=1}^{M_i}\left|h_a^i - h_e^{ij}\right| \tag{4.4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s}\frac{1}{M_i}\sum_{j=1}^{M_i}(h_a^i - h_e^{ij})^2}, \tag{4.5}$$

where N_s, M_i, h_a^i and h_e^{ij} denote the number of test speakers, the number of test utterances for the i^{th} speaker, the actual height of the i^{th} speaker, and the height estimate corresponding to the j^{th} utterance of the i^{th} speaker, respectively.

Depending on the amount of data used for height estimation, MAE and RMSE

Table 4.1: *Sentence- and speaker-level MAEs and RMSEs for automatic height estimation using $Sg1$, $Sg2$ and $Sg3$ (results are shown for clean, wideband TIMIT data). In comparison, the algorithms in [GMF10a] and [GMF10b] were reported to yield an overall MAE and RMSE of 5.3 cm and 6.8 cm, respectively.*

	Using $Sg1$	Using $Sg2$	Using $Sg3$
MAE_{st} (cm)	**5.4** (5.6, 5.0)	**5.5** (5.6, 5.4)	**5.7** (5.6, 5.9)
MAE_{sp} (cm)	**5.3** (5.5, 4.9)	**5.4** (5.5, 5.2)	**5.6** (5.5, 5.8)
RMSE_{st} (cm)	**6.8** (6.9, 6.4)	**6.9** (6.9, 6.9)	**7.1** (7.0, 7.4)
RMSE_{sp} (cm)	**6.6** (6.8, 6.2)	**6.7** (6.8, 6.7)	**7.0** (6.9, 7.3)

were calculated in two different ways. (1) When one sentence was used per height estimate, MAE and RMSE were calculated at the *sentence level*. In other words, M_i was equal to 10 (the number of sentences per speaker) in Eqs. (4.4) and (4.5). The sentence-level metrics will henceforth be denoted as MAE_{st} and RMSE_{st}. (2) When height was estimated using a single utterance formed by concatenating all 10 sentences of a given speaker, MAE and RMSE were calculated at the *speaker level* (M_i was equal to 1 in Eqs. (4.4) and (4.5)). The speaker-level metrics will henceforth be denoted as MAE_{sp} and RMSE_{sp}.

4.3 Results and discussion

This section presents the MAEs and RMSEs for height estimation in clean as well as noisy conditions. Additionally, the correlations achieved between actual height and estimated height are analyzed.

Table 4.1 lists the sentence- and speaker-level MAEs and RMSEs for the clean, wideband evaluation set. The following observations can be made from

Table 4.2: *Sentence-level MAEs and RMSEs for automatic height estimation using $Sg1$. Results are shown for narrowband TIMIT data in clean as well as noisy (0 dB SNR) conditions.*

	Clean	Babble	White	Pink	Factory
$\mathrm{MAE}_{\mathrm{st}}$ (cm)	5.4	5.5	5.7	5.6	5.4
$\mathrm{RMSE}_{\mathrm{st}}$ (cm)	6.8	6.9	7.2	7.0	6.8

Table 4.1. (1) Considering the overall (males and females combined) performance metrics, $Sg1$ and $Sg2$ are almost equally good for height estimation using speech signals. Despite a stronger correlation between speaker height and $Sg2$ (see Section 4.1), $Sg1$-based height estimation is superior for female speakers. This result is difficult to explain because (1) the actual SGRs of TIMIT speakers are unknown, and (2) the effect that SGR-estimation errors have on height-estimation accuracy is nontrivial. (2) $Sg3$ gives slightly poorer results than $Sg1$ and $Sg2$ (especially for female speakers). This is presumably because the estimation of $Sg3$ is indirect, requiring an intermediate estimate of $Sg2$. If estimated directly from speech data, $Sg3$ might be able to predict height as accurately as the other two SGRs. (3) The sentence-level metrics are slightly worse than the corresponding speaker-level metrics, but are still quite acceptable. This means that the proposed method is effective even when data are limited. (4) The overall $\mathrm{MAE}_{\mathrm{sp}}$ and $\mathrm{RMSE}_{\mathrm{sp}}$ for $Sg2$-based height estimation—5.4 cm and 6.7 cm—are comparable to the results in [GMF10a] and [GMF10b], while the overall $\mathrm{MAE}_{\mathrm{sp}}$ and $\mathrm{RMSE}_{\mathrm{sp}}$ for $Sg1$-based estimation—5.3 cm and 6.6 cm—are marginally better. Although the proposed method is not significantly better than the best existing algorithms, it is more efficient in two respects. (1) *Amount of training data and generalizability*: [GMF10a] and [GMF10b] trained and evaluated their algorithms on 462

and 168 TIMIT speakers, respectively (train-to-test ratio = 2.75), while the proposed method was trained on just 56 speakers in the WashU-UCLA corpora and evaluated on 604 speakers in the TIMIT corpus (train-to-test ratio < 0.1). (2) *Size of the feature set*: [GMF10a] and [GMF10b] used a 50-dimensional feature vector to estimate height, while the proposed method used just *one* feature ($Sg1$, $Sg2$ or $Sg3$).

Table 4.2 lists the sentence-level MAEs and RMSEs (corresponding to $Sg1$) for the narrowband evaluation set in clean as well as noisy conditions. Note that [GMF10a] and [GMF10b] did not evaluate their algorithms in noise. It is clear from the results for the clean condition that G.712 filtering has no effect on the performance of the proposed algorithm. There is a slight performance degradation in white and pink noise conditions, but the results are quite acceptable considering that the algorithm was evaluated at 0 dB SNR. These results enhance the utility of the proposed approach, especially given that the same models (Eqs. (4.1)–(4.3)) can be used regardless of the evaluation condition. In contrast, the algorithms proposed in [GMF10a], [GMF10b] and [PH97] are likely to suffer a larger performance degradation with filtered and/or noisy speech owing to their dependence on features derived from spectral envelopes (e.g., MFCCs).

In addition to MAE and RMSE, the correlation between actual height and estimated height (or equivalently, between actual height and estimated SGR frequencies) was considered important for the assessment of height-estimation performance. The correlation was fairly strong when male and female data were pooled together ($r = 0.71$ for all three SGRs), but not when they were treated separately ($r = 0.12$ for males, and 0.21 for females, for all three SGRs)—see Figure 4.3. In comparison, the within-gender correlations between ground truth SGRs and height (for the training speakers in the WashU-UCLA corpora) were

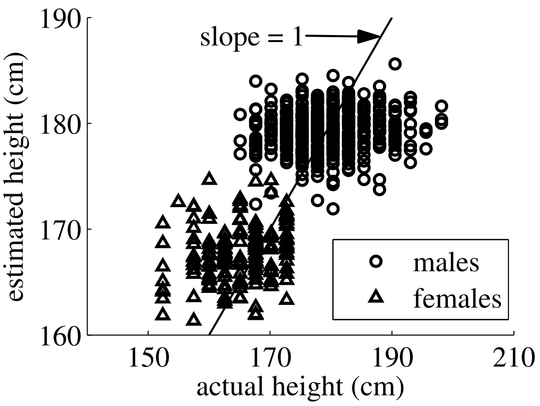


Figure 4.3: *Scatter plot (604 data points) of estimated height (using $Sg1$) versus actual height for wideband TIMIT data (clean speech). The correlation between the two quantities is poor within gender, suggesting that the proposed method requires further improvement.*

significantly better: $|r| = \{0.43\ (Sg1), 0.63\ (Sg2), 0.57\ (Sg3)\}$ for males, and $\{0.23\ (Sg1),\ 0.48\ (Sg2),\ 0.37\ (Sg3)\}$ for females. Therefore, the proposed approach needs improvement in terms of the within-gender correlations between estimated height and actual height (despite satisfactory MAEs and RMSEs within gender). Note that [GMF10a] and [GMF10b] did not report correlations for their algorithms. One possible approach to achieving improved performance would be to collect a larger database of subglottal acoustics (along with height information) that would allow the development of more sophisticated (probably nonlinear) models between SGRs and speaker height.

4.4 Existence of performance limits

This section tries to answer the question as to whether there exists a fundamental limit to the accuracy of speech-based height estimation. The limits defined by vocal tract-based and SGR-based approaches are assessed separately because these limits arise from different physiological constraints. From Section 4.3, it is clear that the within-gender correlation between actual and estimated height is an important indicator of height-estimation performance. The limit to height-estimation accuracy will therefore be assessed with respect to this metric.

SGR-based approaches require estimates of SGRs in order to estimate speaker height. Therefore, the maximum correlations that can be achieved between estimated and actual height are governed largely by the correlations between ground truth SGRs and actual height. The WashU-UCLA corpora suggest that these correlations (in magnitude) are roughly between 0.3 and 0.6, with an average value of 0.45 (statistically significant, $p < 0.05$). Therefore, it is probably correct to say that the correlations achievable using SGR-based approaches have a limiting value close to 0.5 (for the range of speaker heights encountered in this study). Since this limit probably arises from physiological constraints, it would also be interesting to see what those constraints are, and why the limit cannot possibly be higher than what it appears to be.

As mentioned in the beginning of this chapter, SGRs are determined primarily by the 'acoustic length' of the subglottal system. Physiologically, since the 'acoustic length' is expected to be correlated with the size of the lungs and the length of the trunk (or torso), SGRs are likely to be strongly correlated with trunk length. However, according to physiological data reported in [Hrd25], trunk length itself appears to be only moderately correlated with overall body height. Specifically, [Hrd25] reports that the ratio of trunk length and height is a function of height itself, and that short speakers (males as well as females) have larger trunk length-to-height ratios than tall speakers. Such a relationship between trunk length and height seems to be partly responsible for the weak correlations observed in Figure 4.3, with height being overestimated for short speakers and underestimated for tall speakers (for both genders). In essence, SGRs, when estimated well, may provide accurate estimates of trunk length but only moderately-accurate estimates of speaker height. In light of these observations, a value of 0.5 (as mentioned above) appears to be a reasonable estimate of the limiting correlation for SGR-based approaches.

In contrast to SGR-based methods, vocal tract-based approaches rely on the correlation between VTL and height. Figure 5 of [FG99] shows VTL as a function of height, and Table 5 of that paper reports the corresponding correlation coefficients separated by gender. Although the correlations are strong—roughly 0.8 for both males and females, they result from the fact that the data span a wide range of speaker heights within gender. To enable a comparison with the data used in this chapter, a subset of the data plotted in Figure 5 of [FG99] was analyzed (i.e., considering male speaker heights between 165 and 201 cm, and female speaker heights between 152 and 175 cm). The x- and y-coordinates of the data points were obtained with the help of the program TRACER, v.1.7 [Kar], and the within-gender correlation between VTL and height was found to be 0.3 for both males and females (not significantly different from 0.0, $p > 0.05$). Similarly, [RKN05] found the correlations between speaker height and the first four formants of schwa vowels—which have relatively open vocal-tract configurations—to be less than or equal to 0.3 for females (0.16 on average), and less than 0.59 for males (0.41 on average). Note that in the above analyses, the number of speakers was comparable to the size of the training data used in this chapter. For the range of speaker heights encountered in this study, a value of 0.3 (approximately) appears to be the limiting correlation for vocal tract-based approaches; this is considerably lower than the corresponding limit for SGR-based methods.

From the above arguments, it appears that the correlations between SGRs and speaker height determine the limit of height-estimation accuracy, although the limit itself can vary depending on the range of speaker heights under consideration. Furthermore, it is probably easier to achieve the SGR-based limit owing to the fact that the subglottal system of a given speaker, unlike his/her vocal tract, is effectively time invariant.

4.5 Conclusion

In this chapter, an SGR-based approach was proposed for estimating the height of an unknown speaker using his/her speech sample. The emphasis was on robustness to noise and data limitedness, and on achieving good performance with narrowband telephone speech.

The proposed method is motivated by the physiological correlation between overall body height and the effective length of the subglottal system, and is based on the correlation observed between SGRs and height. Using the ground truth SGRs and self-reported heights of speakers in the WashU-UCLA Adults corpus and the WashU-UCLA Bilingual Adults corpus, first-order linear models are trained to predict height given SGR frequencies. Given a speech signal, speaker height is estimated by first estimating SGRs (see Chapter 3) and then using the empirical relations between SGRs and height. The method is evaluated on 604 speakers in the TIMIT database. Three evaluation conditions are considered: (1) clean wideband speech (0–8 kHz bandwidth), (2) clean narrowband speech (300–3400 Hz bandwidth), and (3) noisy narrowband speech (babble, factory, white and pink noise added at 0 dB SNR).

Using about 3 seconds of clean speech data (wideband or narrowband), $Sg1$, $Sg2$ and $Sg3$ can estimate height, on average, to within 5.4, 5.5 and 5.7 cm, respectively. The degradation in height-estimation performance due to noise is minimal—less than 0.3 cm, on average. Actual and estimated height correlate well ($r \sim 0.7$) when the results for male and female speakers are pooled together, but not when they are considered separately ($r \leq 0.2$). One reason for this could be the simplistic nature of the first-order linear models used for prediction. It might be possible to achieve better within-gender correlations (close to 0.5) with the help of a larger training set that would allow the development of more sophisticated (probably nonlinear) models between SGRs and speaker height.

The proposed height-estimation method is an improvement over existing algorithms because (1) it achieves comparable performance while being more transparent (well-motivated features), efficient (small feature set and limited training data) and generalizable (test corpus much larger than the training corpus), (2) it can perform equally well in wideband and narrowband conditions with little degradation in the presence of noise, and (3) its optimal implementation is likely to perform better than the optimal vocal tract-based approach (which is expected to achieve within-gender correlations close to 0.3 or smaller).

CHAPTER 5

Speaker normalization using SGRs

Speaker normalization is an important component of SI ASR systems. It is particularly important in the context of ASR for children owing to the following reasons. (1) Despite a few efforts to collect databases of children's speech [MLU96, Esk96, SHC00, GG03, KYI05, BBD05], the majority of ASR systems (for real-world applications) are still trained using adults' speech. In other words, ASR for children happens typically with a *mismatched* recognition setup. Owing to the large acoustic differences between adults and children [LPN99, PN03], it becomes important to compensate for this mismatch using speaker normalization procedures. (2) Children's speech exhibits a high degree of inter-speaker acoustic variability (much higher compared to what is observed in adults' speech) [LPN99]. Therefore, speaker normalization becomes important even in a *matched* ASR setup (i.e., trained and tested on children).

Several speaker-normalization schemes have been proposed for children's ASR [PN03, SU08, WLA09b, WLA09a, GG03, EB05, GGB07]. However, the effect of noise, which is inevitable in most real-world environments such as classrooms and public kiosks, has not been accounted for. This chapter develops a novel, hybrid (partly knowledge-based and partly statistical) approach to noise-robust normalization by leveraging certain well-established properties of SGRs. Note that some studies report speaker-normalization results using a severely-mismatched setup that involves training only on adult males and testing on children [WLA09b, WLA09a, CA06]. Such a paradigm is not common in practice and hence is not considered in this dissertation.

An important practical consideration in speaker normalization is whether to apply it in *enrollment* mode, where a few words or utterances are used to estimate the frequency-warping parameters before actual recognition, or in *live* mode, where frequency warping and recognition happen on a per-utterance basis. The live mode offers better performance in general [EB05] and is also better suited to real-time recognition and recognition on multi-user systems (e.g., gaming consoles). For children's ASR, another important aspect to consider is the bandwidth of the speech signal. Since children's speech has useful spectral information above 4 kHz, better normalization can be achieved (especially in mismatched training conditions) when the bandwidth is 6 kHz or more [LR01, EB05]. However, for compatibility with existing systems and greater utility across application domains, most studies have proposed normalization schemes using a bandwidth of 4 kHz [WJ96, PN03, SU08, WLA09b, WLA09a]. In this dissertation, only the live mode and the narrowband paradigm are considered.

5.1 Motivation for an SGR-based approach

There are several reasons why it was hypothesized that SGRs might be effective for normalizing children's speech.

• Being speaker specific and content independent (see Chapter 2), SGRs are good candidates for *removing* speaker effects.

• Owing to their role as phonological vowel-feature boundaries, $Sg1$ and $Sg2$ can be used to implicitly normalize $F1$ and $F2$ (which are important carriers of phonemic information). Figure 5.1 shows two examples of how $Sg1$ and $Sg2$

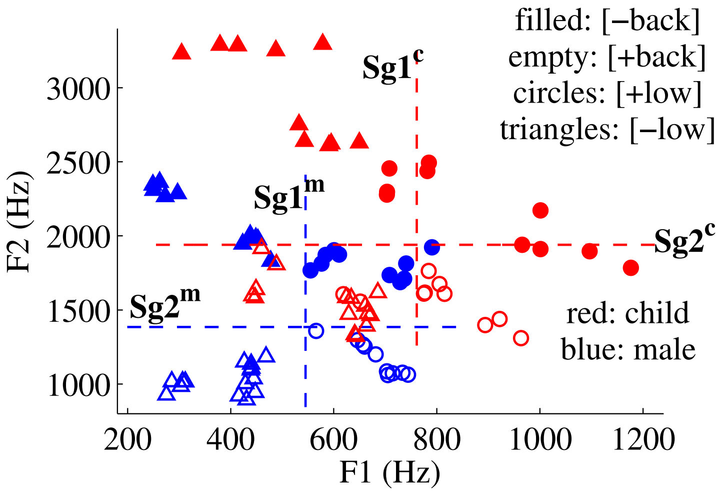


Figure 5.1: *$Sg1$ and $Sg2$ plotted in the $F1$-$F2$ plane for an adult male speaker (in blue; $Sg1^m$, $Sg2^m$) in the WashU-UCLA Adults corpus and a child speaker (in red; $Sg1^c$, $Sg2^c$; age = 11 years) in the WashU-UCLA Kids corpus. Note that $Sg1$ lies roughly between [+low] and [-low] vowels along the $F1$ dimension and $Sg2$ lies roughly between [+back] and [-back] vowels along the $F2$ dimension.*

divide the $F1$-$F2$ plane. The plot in blue is for an adult male speaker ($Sg1^m$, $Sg2^m$) in the WashU-UCLA Adults corpus and the plot in red is for a child speaker ($Sg1^c$, $Sg2^c$; age = 11 years) in the WashU-UCLA Kids corpus. It is evident from the figure that by mapping $Sg1^c$ to $Sg1^m$ and $Sg2^c$ to $Sg2^m$, the formant clusters of the child speaker can be aligned (roughly) with those of the male speaker. While $Sg3$ may not provide any vowel-feature contrasts, it can still be useful for normalizing $F3$ and higher formants.

- The performance of children's ASR is known to be correlated with speaker age and height—worse for the shorter, younger speakers compared to the taller, older ones [EB05]. Table 5.1 shows that the SGRs of children are also correlated with age and height, which implies that SGR-based normalization could potentially equalize ASR performance across different age and height groups. Also,

Table 5.1: *Correlations of ground-truth SGRs and* $F3$ *(per-speaker average) with speaker age and height based on data from 43 speakers, aged between 6 and 17 years, in the WashU-UCLA Kids corpus. Note that the correlations for* $Sg3$ *are based on data from 31 speakers only.*

	$Sg1$	$Sg2$	$Sg3$	$F3$
correlation with age	-0.89	-0.88	-0.91	-0.75
correlation with height	-0.94	-0.93	-0.94	-0.84

as Table 5.1 shows, the correlations for SGRs are significantly stronger than the correlations for $F3$. Since $F3$ is known to be closely related to vocal-tract length [Fan75], these results also suggest that SGR-based normalization might be more effective than VTLN.

• Statistical normalization schemes (like VTLN with an ML grid search) tend to be less effective in noise owing to environmental mismatch between test utterances and acoustic models (which are usually trained using relatively clean data). SGR-based normalization is expected to be more noise robust since SGRs can be estimated reliably even in the presence of noise (see Chapter 3).

5.2 The proposed approach

Based on the ideas described above, the proposed frequency-warping function was designed to map the first three SGRs of a given target utterance to the first three SGRs of a reference speaker. Figure 5.2 shows the proposed warping function in red. Denoting the reference and target SGRs with subscripts r and t, respectively,

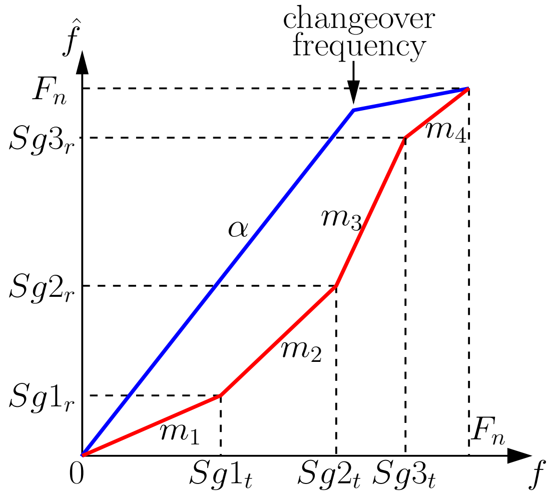


Figure 5.2: *The proposed warping function (shown in red) maps the SGRs of a given target utterance (subscript t) to those of a reference speaker (subscript r). The scalars m_1 to m_4 are the slopes of the lines constituting the warping function. The conventional piece-wise linear warping function (for VTLN) is shown in blue (α is the warping factor). F_n denotes Nyquist frequency.*

the function can be defined as:

$$\hat{f} = \begin{cases} m_1 f & 0 \leq f \leq Sg1_t \\ m_2(f - Sg1_t) + Sg1_r & Sg1_t < f \leq Sg2_t \\ m_3(f - Sg2_t) + Sg2_r & Sg2_t < f \leq Sg3_t \\ m_4(f - Sg3_t) + Sg3_r & Sg3_t < f \leq F_n, \end{cases} \tag{5.1}$$

where F_n denotes the Nyquist frequency, and f and $\hat{f}$ denote the frequency scales before and after warping, respectively. The scalars m_1 to m_4 are the slopes of the lines constituting the warping function, and can be easily computed given the reference and target SGRs. Figure 5.2 also shows the piece-wise linear warping

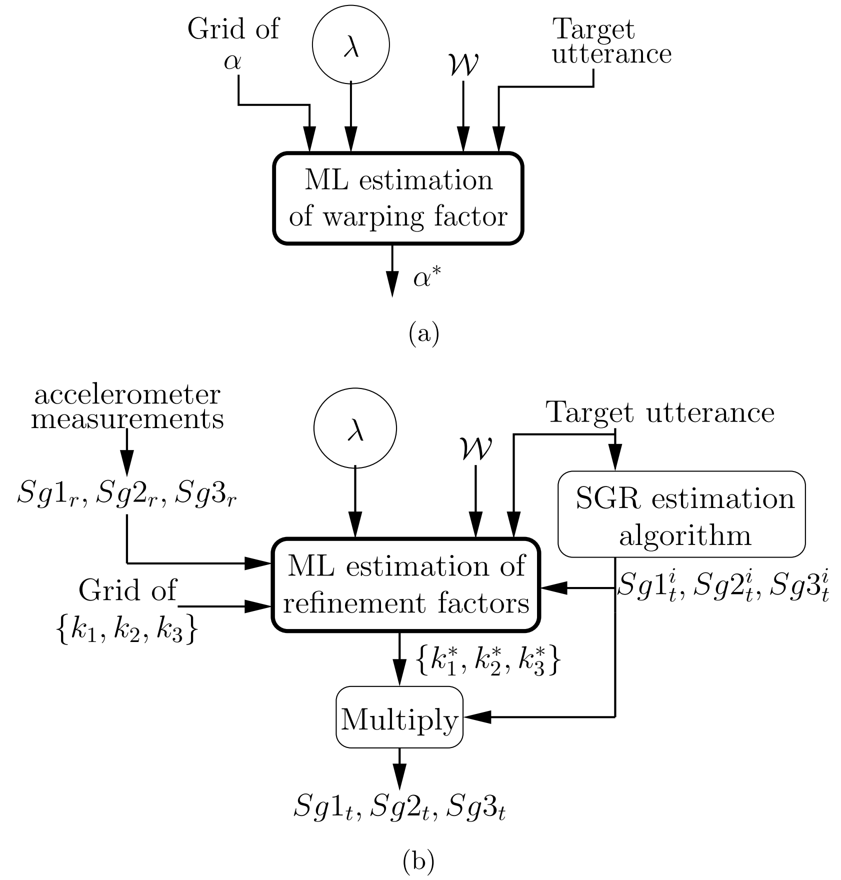


Figure 5.3: *(a) Estimating the optimal VTLN α. (b) Estimating target SGRs via optimal refinement of initial (signal-based) estimates.*

function used for VTLN—it is almost linear, with a slope of α, except for the changeover frequency near F_n (typically chosen to be between 0.8 and 0.9 times F_n) that ensures bandwidth preservation after warping. Both VTLN and SGR-based warping were implemented by inversely scaling the center frequencies and bandwidths of the feature-extraction filter bank. This implementation was first proposed in [LR98] and has been widely used for its computational efficiency.

5.2.1 Parameter estimation

Given an utterance, VTLN α is typically estimated using an ML grid search [LR98]:

$$\alpha^* = \underset{\alpha \in \mathcal{G}_\alpha}{\arg\max}\, P(\mathcal{X}^\alpha | \lambda, \mathcal{W}), \tag{5.2}$$

where $\mathcal{X}^\alpha$, α^*, $\mathcal{G}_\alpha$, λ, and $\mathcal{W}$ denote the sequence of α-warped feature vectors, the optimal α value, the search grid, the set of pre-trained hidden Markov models (HMMs), and the word-level transcription associated with the given utterance, respectively. Note that the estimation of VTLN α via Eq. (5.2) is purely statistical—it depends entirely on the parameters of λ.

SGR warping via Eq. (5.1) requires two sets of parameters—reference SGRs $Sg1_r$, $Sg2_r$, and $Sg3_r$, and target SGRs $Sg1_t$, $Sg2_t$, and $Sg3_t$. Reference SGRs were determined *a priori*, using manual SGR measurements obtained from accelerometer recordings of monophthong vowels. Accelerometer recordings were chosen based on the composition of the speech training data that were used for ASR—the WashU-UCLA Adults corpus was used when the training data comprised adults' speech, and the WashU-UCLA Kids corpus was used when the training data comprised children's speech. The reference value of each SGR was computed by averaging all the measurements available in the chosen corpus.

The target SGRs, for a given utterance, were determined in two steps. (1) Initial estimates ($Sg1_t^i$, $Sg2_t^i$, $Sg3_t^i$) were obtained using automatic SGR estimation algorithms—the algorithm proposed in Section 3.1 was used for adults (during acoustic model training), and the algorithm proposed in Section 3.3 was used for children (during training and testing). (2) To compensate for SGR estimation errors, and also to account for the small (yet non-negligible) within-speaker COVs of SGRs (see Chapter 2), the initial estimates were refined as per Eq. (5.3):

$$SgM_t = k_M^* \times SgM_t^i \qquad M \in \{1, 2, 3\}, \tag{5.3}$$

where $\{k_1^*, k_2^*, k_3^*\}$ denotes the set of optimal refinement factors, determined using an ML grid search:

$$\{k_1^*, k_2^*, k_3^*\} = \underset{\{k_1,k_2,k_3\} \in \mathcal{G}_k}{\arg\max} \; P(\mathcal{X}^{\{k_1,k_2,k_3\}} | \lambda, \mathcal{W}). \tag{5.4}$$

In Eq. (5.4), $\mathcal{G}_k$ denotes the 3-dimensional search grid for the refinement factors, and $\mathcal{X}^{\{k_1,k_2,k_3\}}$ denotes feature vectors corresponding to the parameters $\{k_1 \times Sg1_t^i, k_2 \times Sg2_t^i, k_3 \times Sg3_t^i\}$. Note that the refinement factors define a frequency range around the initial SGR estimates. For example, if the refinement factors are chosen to lie between 0.95 and 1.05, the target SGRs would be determined by searching within ±5% of the initial estimates. Figure 5.3 summarizes the procedure used to estimate warping parameters for VTLN and the proposed SGR-based approach. For convenience, $\mathcal{K}$ will henceforth be used to denote the triplet $\{k_1, k_2, k_3\}$.

While VTLN relies entirely on λ to estimate the best α, the ML grid search in Eq. (5.4) is preceded by an initialization step that is independent of λ. Since SGR estimation for children is fairly noise robust (down to an SNR of 5 dB), SGR-based warping is expected to be less sensitive to noise than VTLN, thus leading to better ASR performance in matched (training on children) as well as mismatched (training on adults) conditions. For convenience, the acronym SGRN will henceforth be used to refer to speaker normalization via SGR-based warping.

```
Extract unwarped features from the training utterances.
Train the set of unwarped models $\lambda^0$.
if <VTLN> then
    for $s = 1$ to # training utterances do
        Determine $\alpha^{(s)*}$ with respect to $\lambda^0$ and $\mathcal{W}^{(s)}$ (the true transcription of utterance $s$) as per Eq. (5.2).
        Extract warped features using $\alpha^{(s)*}$.
    end for
    Train the set of normalized models $\lambda^\alpha$ using the warped utterances.
else if <SGRN> then
    Choose reference SGR values that are appropriate for the training population (adults or children).
    Choose an SGR estimation algorithm—the algorithm in Section 3.1 for adults, and the algorithm in Section 3.3 for children.
    for $s = 1$ to # training utterances do
        Obtain initial SGR estimates.
        Determine $\mathcal{K}^{(s)*}$ with respect to $\lambda^0$ and $\mathcal{W}^{(s)}$ (the true transcription of utterance $s$) as per Eq. (5.4).
        Extract warped features using $\mathcal{K}^{(s)*}$.
    end for
    Train the set of normalized models $\lambda^{\mathcal{K}}$ using the warped utterances.
end if
```

Figure 5.4: *Pseudo-code of the training protocol for VTLN and SGRN.*

5.2.2 Training and testing protocols

Normalization of training data was supervised—true transcriptions were used to estimate the warping parameters. First, an initial set of models λ^0 was trained

```
for $s = 1$ to # testing utterances do
    Estimate $\mathcal{W}^{(s)}$ (the first-pass transcription of utterance $s$) using unwarped features and models $\lambda^0$.
    if <VTLN> then
        Determine $\alpha^{(s)*}$ with respect to $\lambda^\alpha$ and $\mathcal{W}^{(s)}$ as per Eq. (5.2).
        Extract warped features for $\alpha^{(s)*}$. Decode using $\lambda^\alpha$.
    else if <SGRN> then
        Use the same reference SGRs as during training. Use the algorithm in Section 3.3 for SGR estimation.
        Obtain initial SGR estimates. Determine $\mathcal{K}^{(s)*}$ with respect to $\lambda^\mathcal{K}$ and $\mathcal{W}^{(s)}$ as per Eq. (5.4).
        Extract warped features for $\mathcal{K}^{(s)*}$. Decode using $\lambda^\mathcal{K}$.
    end if
end for
```

Figure 5.5: *Pseudo-code of the testing protocol for VTLN and SGRN.*

using unwarped utterances. Then, for each training utterance, the optimal warping parameters (α^* for VTLN, and $\mathcal{K}^*$ for SGRN) were estimated with respect to λ^0 and the true word-level transcription. Finally, the warped utterances were used to train a set of normalized models (λ^α for VTLN, and $\lambda^\mathcal{K}$ for SGRN). The pseudo-code of the training protocol is summarized in Figure 5.4.

Normalization of testing data was unsupervised—first-pass transcriptions were used to estimate the warping parameters. Given a test utterance, a word-level transcription was first obtained using unwarped features and models λ^0. Then, the optimal warping parameters were estimated with respect to the normalized models and the first-pass transcription. Finally, the utterance was decoded using the warped features and the normalized models. The pseudo-code of the testing protocol is summarized in Figure 5.5. The ML estimation of warping paramters (for both training and testing) was implemented by *force aligning* the warped features with respect to word-level transcriptions.

5.3 Experimental setup

So far, no database has been collected for the purpose of children's ASR in noise. Therefore, noisy data sets were created using the TIDIGITS corpus (which has been used previously for children's ASR in quiet [SU08, CA06, WLA09b, WLA09a]). The TIDIGITS corpus contains training and testing sets for adults as well as children. The training set for adults has 112 speakers (55 male; 57 female), while the training and testing sets for children have 51 (25 male; 26 female) and 50 (25 male; 25 female) speakers, respectively. All children in the corpus are between 6 and 15 years of age. The recorded material consists of connected-digit sequences formed from a 11-word vocabulary—"one" to "nine", "zero", and "oh". The utterances have 1, 2, 3, 4, 5, or 7 digits each.

All signals were down sampled to 8 kHz (from the original sampling rate of 20 kHz). Two sets of acoustic models were trained—one using adults' speech (for mismatched ASR) and the other using children's speech (for matched ASR). No noise was added to the training data. Testing was done in both clean and noisy conditions. Using FaNT, clean speech files (in the testing set for children) were corrupted with babble, car, pink, and white noise types (from the NOISEX-92 corpus), at SNRs chosen randomly between 5 and 15 dB.

Experiments were conducted with two different front ends—MFCCs and power-normalized cepstral coefficients (PNCCs) [KS12]. PNCCs have been shown to be noise robust for adults' ASR. Both front ends used a 26-channel filter bank having triangular filters of constant bandwidth on the Mel scale. Speech signals were segmented into 25 ms frames spaced at 10 ms intervals. Each frame was parameterized by a 39-dimensional feature vector consisting of the first 13 cepstral coefficients, and their first- and second-order derivatives. Mean and variance normalization (MVN) was applied on a per-utterance basis. The recognizer was composed of monophone hidden Markov models (HMMs). The HMMs had 3 emitting states each, and each state had 6 diagonal-covariance Gaussian components.

For adults' speech (used only as training data), VTLN α was estimated by searching over [0.8,1.2] in steps of 0.02. For children's speech, the search grid depended on whether the ASR setup was matched or mismatched. For the matched setup, the search grid was the same as that used with adults' speech. For the mismatched setup, on the other hand, the search range was [0.7,1.1] to account for the fact that children's speech, with respect to adults' speech, is more likely to undergo spectral compression than expansion.

The reference SGRs for the mismatched setup were $\{601, 1419, 2304\}$ Hz (obtained from the WashU-UCLA Adults corpus), while the reference SGRs for the matched setup were $\{734, 1736, 2713\}$ Hz (obtained from the WashU-UCLA Kids corpus). In both cases, the SGR refinement factors were searched over [0.9,1.1] in steps of 0.05, thus allowing the initial SGR estimates to be refined by up to ±10%. Therefore, the search grid for $\mathcal{K}$ was 3-dimensional with 125 points.

5.4 Results and discussion

In all of the experiments discussed in this chapter, the performance of SGRN is compared with the performance of VTLN. Word error rate (WER) is used as the metric for comparing recognition performance. Based on the current literature, only the shift-based approach of [SU08] is known to be better than VTLN for children's speech—the authors of [SU08] showed (using a mismatched setup for digit recognition in clean conditions) that their approach reduced the WER by 7.5% relative to VTLN. However, instead of using the usual filter-bank front end for spectral smoothing, they used a periodogram averaging approach. When their shift-based method was implemented with the conventional filter-bank MFCC front end, no significant WER reduction was observed relative to VTLN. Therefore, only VTLN was chosen as the algorithm for comparison.

Normalization experiments were first performed with MFCCs, in matched as well as mismatched training conditions. The results are shown in Table 5.2, from which the following observations can be made.

- As expected, WERs for the matched case are much lower.

- Baseline (MVN) performance in clean is quite satisfactory for both matched and mismatched conditions. With the addition of noise, however, the WERs increase dramatically (except for the case of car noise, which is fairly stationary in nature). The increase in baseline WER, due to noise, is about 3.5–6 times in the mismatched condition, and about 11–17 times in the matched condition.

- For the mismatched case, VTLN provides large WER reductions relative to the baseline in both clean and noisy conditions ($\sim$33%, on average). This can be attributed to the large acoustic differences between adults and children, and shows the importance of speaker normalization in mismatched training, regardless of the environmental conditions. For the matched case, VTLN provides a 13% WER reduction (relative to the baseline) in clean, but only moderate performance improvements in noise. This indicates that in matched training conditions, compensating for the effects of noise is probably more important than normalizing the acoustic differences among speakers.

Table 5.2: *WERs for children's speech recognition using MFCCs (TIDIGITS database; children's testing set corrputed with different noise types; SNRs between 5 and 15 dB). Numbers in paranthesis are the WER reductions achieved by SGRN, relative to VTLN.*

	Clean	Babble	Car	Pink	White	Average
	Training on adults (mismatched condition)					
MVN	7.85	49.27	12.36	37.26	28.21	26.99
MVN + VTLN	2.64	33.97	4.00	27.92	21.87	18.08
MVN + SGRN	1.80	29.12	2.99	25.10	19.85	15.77
	(31.8%)	(14.3%)	(25.3%)	(10.1%)	(9.2%)	(12.8%)
	Training on children (matched condition)					
MVN	1.52	24.98	1.88	21.31	17.08	13.35
MVN + VTLN	1.33	23.66	1.70	21.32	15.99	12.80
MVN + SGRN	1.32	23.25	1.63	19.66	15.36	12.24
	(0.8%)	(1.7%)	(4.1%)	(7.8%)	(3.9%)	(4.4%)

Table 5.3: *WERs for children's speech recognition using PNCCs in a mismatched setup (same test data as for Table 5.2). Numbers in paranthesis are the WER reductions achieved by SGRN, relative to VTLN.*

	Clean	Babble	Car	Pink	White	Average
MVN	10.37	36.28	12.59	28.57	24.78	22.52
MVN + VTLN	3.33	19.38	3.62	17.08	16.74	12.03
MVN + SGRN	2.67	17.11	2.99	14.97	15.50	10.65
	(19.8%)	(11.7%)	(17.4%)	(12.4%)	(7.4%)	(11.5%)

• For the mismatched case, SGRN yields significant WER reductions relative to VTLN in clean as well as noisy conditions (especially car and babble noise).

The performance improvements provided by SGRN are attributable to (1) strong correlation between SGRs, and speaker age and height (which helps in mitigating the age- and height-induced acoustic mismatch between adults and children), and (2) noise robustness imparted by the semi-statistical estimation of warping parameters. For the matched case, SGRN is only marginally better than VTLN, which in turn is only marginally better than the baseline. This reiterates that in matched training conditions, noise compensation is probably more important than speaker normalization. However, as will be shown later in Section 5.4.1, SGRN, even in matched conditions, provides a significant performance improvement for the younger speakers.

Further experiments were performed in the mismatched case, using PNCCs. The results are shown in Table 5.3. PNCCs provide a much better baseline than MFCCs in noise, although there is some performance degradation in clean—this is in agreement with the results published previously on adults' ASR [KS12]. VTLN provides large WER reductions relative to the baseline (∼47%, on average), but SGRN is significantly better than VTLN in clean as well as noisy conditions. These results suggest that the performance gains offered by SGRN could possibly generalize to other filter bank-based features.

5.4.1 Results by speaker age

As previous studies have shown, recognition of children's speech becomes harder with decreasing speaker age, especially in mismatched training conditions [PN03, EB05]. Therefore, SGRN and VTLN were compared in terms of their normalization performance for different age groups. The utterances in the testing sets (both clean and noisy) were divided into 3 speaker groups: 6–8, 9–11, and 12–15 years, and WERs were computed separately for each of them. The results are shown in

Table 5.4 (averaged across clean and the four noise types) for both mismatched and matched training conditions.

For the youngest speaker group (6–8 years), SGRN yields large performance gains (relative to VTLN) in the mismatched condition—note that SGRN leads to better WER equalization across age groups. The performance gain is significant in the matched condition as well. Again, this is attributable to the strong correlation between SGRs, and speaker age and height. These results suggest that SGRN can be highly effective in ASR applications that are targeted towards young speakers (language learning, reading assessment, etc.).

Table 5.4: *WERs for children's speech recognition, by age group (same test data as for Table 5.2). Results are averaged across clean and the four noise types (babble, car, pink, and white). Numbers in paranthesis are the WER reductions achieved by SGRN, relative to VTLN.*

	Training on adults (MFCC)			Training on adults (PNCC)			Training on children (MFCC)		
	6–8 yrs	9–11 yrs	12–15 yrs	6–8 yrs	9–11 yrs	12–15 yrs	6–8 yrs	9–11 yrs	12–15 yrs
MVN	39.71	26.15	18.87	37.95	21.02	13.96	11.71	13.44	14.48
MVN + VTLN	23.16	17.91	14.41	16.50	11.72	9.22	11.71	12.94	13.29
MVN + SGRN	16.65	16.13	14.07	12.68	10.87	8.38	10.32	12.44	13.29
	(28.1%)	(9.9%)	(2.4%)	(23.2%)	(7.3%)	(9.1%)	(11.9%)	(3.9%)	(0.0%)

Table 5.5: *WERs for children's speech recognition, by utterance length (same test data as for Table 5.2). "Short" = 1 or 2 digits; "Medium" = 3–5 digits; "Long" = 7 digits (note: TIDIGITS does not have 6-digit utterances). Results are averaged across clean and the four noise types (babble, car, pink, and white). Numbers in paranthesis are the WER reductions achieved by SGRN, relative to VTLN.*

	Training on adults (MFCC)			Training on adults (PNCC)			Training on children (MFCC)		
	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
MVN	24.68	27.47	27.49	17.84	23.21	24.01	11.39	14.08	13.23
MVN + VTLN	18.22	18.42	17.42	10.50	12.42	12.23	11.44	13.48	12.41
MVN + SGRN	14.90	16.19	15.56	8.54	11.26	10.82	10.48	12.82	12.27
	(18.2%)	(12.1%)	(10.7%)	(18.7%)	(9.3%)	(11.5%)	(8.4%)	(4.9%)	(1.1%)

5.4.2 Results by utterance length

In ASR-driven applications such as pronunciation assessment and interactive gaming, incoming utterances (from child users) are expected to be short in duration—typically a few words or less. In other applications such as automatic assessment of reading and comprehension abilities, the utterances can be a little longer. Therefore, it is important that the chosen normalization scheme (1) is effective with limited data, and (2) can take advantage of longer utterances when available. SGRN and VTLN were compared in terms of their performance for different utterance lengths. The utterances in the testing sets (both clean and noisy) were divided into 3 groups: short (1 or 2 digits), medium (3–5 digits), and long (7 digits). The results are shown in Table 5.5 (averaged across clean and the four noise types) for both mismatched and matched training conditions.

In both matched and mismatched conditions, SGRN performs significantly better than VTLN for short utterances. For medium-duration and long utterances, SGRN is either better than (in the mismatched case) or comparable (in the matched case) to VTLN. SGRN is effective for both short and long utterances because (1) SGRs can be estimated with reasonable accuracy using small amounts of speech data, and (2) the proposed approach for estimating warping parameters is both knowledge based and statistical. Owing to its semi-statistical nature, SGRN, even for short utterances, is able to estimate three parameters with reasonable efficacy despite using a search grid of coarse resolution (spacing of 0.05 between successive points, versus 0.02 for VTLN).

5.4.3 Analysis of model compaction

Compact acoustic models are obtained by minimizing the inter-speaker variability in the training data via some form of speaker normalization and/or speaker-adaptive training [AMS96]. Model compaction not only leads to better normalization of test data, but also provides better *target-speaker* models when adaptation techniques such as MLLR and constrained MLLR are used (especially in conjunction with speaker normalization) [GGB07]. Although model adaptation was not investigated in this study, SGRN and VTLN were compared with regard to their ability to provide compact acoustic models for ASR.

Model compaction can be measured in terms of the separability (or discriminability) of acoustic units (monophone HMMs in the present case) in the feature space. The Bhattacharyya distance [Fuk90] was used as a measure of the separation between two given monophone HMMs (the Bhattacharyya distance has been used before to measure phone separability [GGB07]). Given two multidimensional Gaussian densities $\mathcal{N}(x; \mu_i, \Sigma_i)$ and $\mathcal{N}(x; \mu_j, \Sigma_j)$ (μ_i and μ_j denote the mean vectors; Σ_i and Σ_j denote the covariance matrices), the Bhattacharyya distance between them—denoted by $\mathcal{D}(i, j)$—is given by Eq. (5.5).

$$\begin{aligned}\mathcal{D}(i,j) &= \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) \\ &+ \frac{1}{2}\log\frac{\left|\frac{\Sigma_i+\Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}\end{aligned} \tag{5.5}$$

To compute Eq. (5.5) for a given pair of monophone HMMs, the Gaussian densities associated with their center states were used (note that the center state better reflects the acoustic characteristics of a phone, as compared to the initial and final states). As mentioned in Section 5.3, the recognizer used 6-component Gaussian mixtures to model the emitting densities of HMMs. For analyzing model compaction, however, all HMM densities were modeled using single-component Gaussians. The average Bhattacharyya distance (averaged over all monophone HMM pairs; denoted by $\overline{\mathcal{D}}$) was computed as per Eq. (5.6), where N denotes the

Table 5.6: *Average Bhattacharyya distance (computed by averaging over all monophone HMM pairs as per Eq.* (5.6)*) for different training conditions and front ends. The average Bhattacharyya distance is indicative of the degree of model compaction and separability.*

	MVN	MVN +VTLN	MVN +SGRN
Training on adults (MFCC)	7.21	7.50	7.64
Training on adults (PNCC)	6.46	6.81	7.05
Training on children (MFCC)	7.35	7.78	7.90

number of HMMs used in the recognizer.

$$\overline{\mathcal{D}} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathcal{D}(i,j) \tag{5.6}$$

Table 5.6 shows the value of $\overline{\mathcal{D}}$ for different training conditions and front ends. Clearly, speaker normalization (using VTLN or SGRN) provides more compact models compared to the baseline (MVN). More importantly, SGRN leads to better model compaction compared to VTLN, in both matched and mismatched training conditions. The efficacy of SGRN can therefore be attributed, in part, to its ability to provide compact acoustic models. Also, MLLR and constrained MLLR are expected to be more effective when applied in conjunction with SGRN than when applied in conjunction with VTLN.

5.4.4 Analysis of Robustness to Noise

It was shown earlier that SGRN performs significantly better than VTLN (in terms of WERs) in various noise conditions—see Tables 5.2 and 5.3. Here, the robustness of SGRN is analyzed by assessing as to how well the normalization parameters are estimated and, also, as to how close the observed performance is to the *optimal* performance. For this analysis, results corresponding to the mismatched training condition and the MFCC front end are used.

Table 5.7 shows the correlation coefficients between parameters estimated from clean utterances and parameters estimated from the corresponding noisy utterances. For VTLN, α is the only parameter that is estimated. For SGRN, on the other hand, the estimated parameters comprise the first three SGRs (obtained after the ML refinement step of Eq. (5.4)). It is clear from Table 5.7 that the correlations for SGRs ($Sg1$, in particular) are significantly higher than the correlations for α, especially in the more severe noise types (babble, pink and white). This can be attributed to the noise robustness of SGR estimation and to the semi-statistical nature of SGRN.

Table 5.8 shows the actual and oracle (or optimal) WERs for VTLN and SGRN. The oracle WERs were obtained by warping noisy utterances using parameters estimated from the corresponding clean utterances. The difference between the actual and oracle WERs is higher for VTLN (especially in the more severe noise types), indicating that SGRN is the more robust normalization scheme.

5.5 Relation to prior work

The present work is novel in several ways compared to previous studies on speaker normalization for children's ASR.

- This is the first known investigation of children's ASR in noise.

- Previous studies have used either purely statistical [LR98, SU08] or purely knowledge-based methods [EG96, WLA09b, WLA09a] to estimate the warping

Table 5.7: *Correlation coefficients between parameters estimated from clean utterances and parameters estimated from the corresponding noisy utterances (same test data as for Table 5.2; MFCC front end; models trained on adults).*

	Babble	Car	Pink	White	Average
α	0.28	0.64	0.34	0.39	0.41
$Sg1$	0.55	0.71	0.61	0.67	0.64
$Sg2$	0.47	0.74	0.44	0.43	0.52
$Sg3$	0.45	0.74	0.40	0.38	0.49

Table 5.8: *Actual and oracle WERs for children's speech recognition (same test data as for Table 5.2; MFCC front end; models trained on adults). The oracle WERs were obtained by warping noisy speech using parameters estimated from the corresponding clean data.*

	Clean	Babble	Car	Pink	White	Average
MVN + VTLN	2.64	33.97	4.00	27.92	21.87	18.08
MVN + VTLN (Oracle)	2.64	29.05	3.63	24.47	19.55	15.87
MVN + SGRN	1.80	29.12	2.99	25.10	19.85	15.77
MVN + SGRN (Oracle)	1.80	28.31	2.55	22.90	18.44	14.80

parameters. This study resulted in a hybrid approach that is not only effective with limited data (like knowledge-based methods) but can also take advantage of longer utterances (like statistical methods).

- In [WLA09b] and [WLA09a], $Sg2$ was simply used in place of a maximum-likelihood grid search to estimate the piece-wise linear warping factor or the Bark-scale shift factor. In contrast, this study employed an entirely new warping scheme based on the use of $Sg1$, $Sg2$ and $Sg3$.

- To estimate $Sg2$ from speech, [WLA09b] and [WLA09a] proposed algorithms that were based on detecting frequency discontinuities and amplitude attenuations of the second formant in diphthongs. Therefore, isolated vowels or words were used for estimating $Sg2$. Such an approach limits the practical utility of SGR-based normalization, especially in the presence of noise. The approach proposed here is based on SGR estimation algorithms that are applicable to *natural* speech, in clean as well as noisy conditions.

5.6 Conclusion

This study represents the first known effort to account for the effects of additive noise in the context of acoustic speaker normalization for children's ASR. Analysis of the WashU-UCLA Kids data revealed that SGRs (basis for SGRN) have stronger correlations with speaker age and height than does $F3$ (related to vocal-tract length). It was therefore hypothesized that SGRN might be more effective than VTLN, especially with regard to equalizing the effect of speaker age on ASR performance.

SGRN was based on a warping function that mapped the first three SGRs of a given target utterance to the first three SGRs of a reference speaker. The reference SGRs were chosen *a priori* based on the speaker population used for training ASR models. The target SGRs for a given utterance were first determined using the proposed SGR estimation algorithm and then refined by up to ±10% using a maximum-likelihood grid search. Normalization was applied to train as well as test data, and the optimal warping parameters (SGRs) were estimated using a two-pass approach involving forced alignment.

SGRN and VTLN were evaluated and compared via ASR experiments on the TIDIGITS database (noisy data sets were created by adding babble, car, pink and white noise types at SNRs ranging between 5 and 15 dB). Children's speech was recognized in both matched and mismatched training conditions. Two front ends were considered—standard MFCCs, and the recently-proposed, noise-robust PNCCs (which have been used successfully for adults' ASR in noise). On average, SGRN was found to be significantly better than VTLN in mismatched training (up to 32% relative reduction in WER) and slightly better than VTLN in matched training. Regardless of the training condition, SGRN offered significant WER reductions (relative to VTLN) for 6–8 year old speakers and 1 or 2 word utterances. These results suggest that SGRN can be highly effective in ASR applications involving young speakers and short responses (e.g., automatic pronunciation assessment and high-end gaming). It was also found that SGRN provided more compact models than VTLN, meaning that model adaptation (e.g., MLLR) can be expected to be more effective with SGRN-compacted models.

The efficacy of SGRN for children's ASR can be attributed to (1) speaker specificity and content independence of SGRs, (2) strong correlations between SGRs, and speaker age and height, (3) noise robustness of the proposed SGR estimation algorithm and its ability to provide reliable estimates with limited data, and (4) semi-statistical approach to estimating the optimal warping parameters. One limitation of SGRN in its current form is the use of a large 3-dimensional search grid—such an approach could be prohibitive for large vocabulary ASR. The grid search can be replaced by gradient search methods for higher efficiency as well as accuracy (see [PA06], for example).

CHAPTER 6

Speaker recognition using subglottal acoustics

6.1 Introduction

Previous chapters studied the properties of SGRs—by manually analyzing accelerometer recordings—and also developed automatic algorithms to estimate SGRs from speech signals. Although speech-based SGR estimates were found to be effective for speaker height estimation and adaptation (especially in limited-data conditions), pilot experiments on the TIMIT database showed that they were not discriminative enough for speaker recognition. Therefore, this chapter employs more informative spectral features in the form of subglottal cepstral coefficients (SGCCs)—they are computed just like MFCCs, except that they are based on subglottal acoustics instead of speech.

Subglottal features could benefit speaker recognition for two reasons. First, subglottal acoustics are speaker specific to some extent owing to their dependence on body height [SP93]. Second, the spectral characteristics of subglottal acoustics (for a given speaker) are much less variable than the spectral characteristics of speech. Figure 6.1 exemplifies this using vowel spectrograms of speech and their corresponding recordings of subglottal acoustics (data were obtained from the WashU-UCLA Adults corpus). The stationary nature of subglottal acoustics can be particularly beneficial when the amount of speech data (for enrollment and/or evaluation) is limited. One of the challenges, however, is to be able to

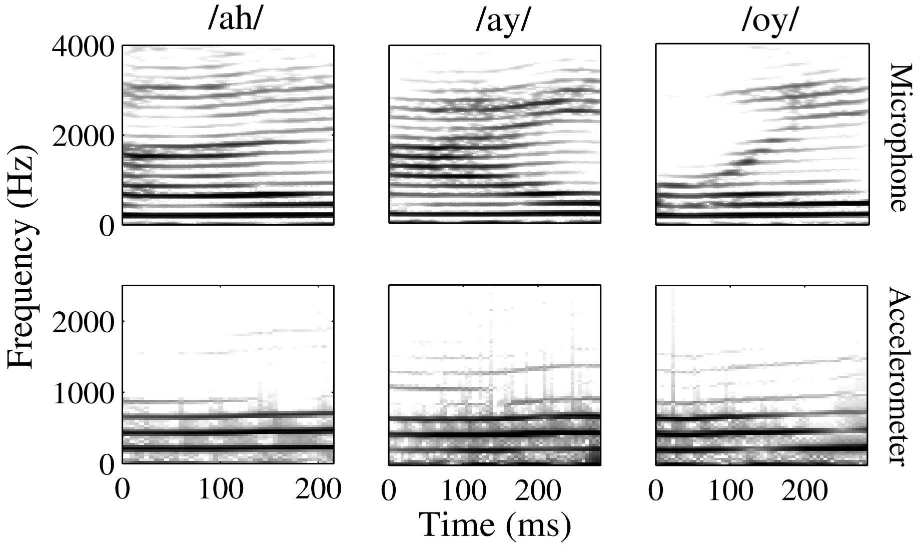


Figure 6.1: *Vowel spectrograms comparing the within-speaker variability of speech (top panel) and subglottal acoustics (bottom panel). Data are sampled from the recordings of a female speaker in the WashU-UCLA Adults corpus.*

estimate subglottal features (SGCCs) using speech, thus obviating the need for an accelerometer in real-world scenarios.

The proposed approach to estimating SGCCs from MFCCs is inspired by previous studies on speech-to-articulatory inversion [TBT08, GN11, LKG13]. The method proposed in these studies was to train joint statistical models from simultaneously-recorded speech and articulatory data, and then use those models to estimate articulatory trajectories from unseen utterances. The proposed approach to SGCC estimation is similar, except that simultaneous recordings of speech and subglottal acoustics are used.

In [LKG13], articulatory parameters (estimated from speech signals) were combined with MFCCs in an SV task. Using the classical UBM-GMM setup, it was shown that the combined system improved verification performance by 9–14% relative to the MFCC-only baseline. This chapter uses a feature-combination approach like [LKG13], but with two important differences. (1) The production-based features used here (SGCCs) are those of the subglottal, not supraglottal, system. (2) In [LKG13], a subset of the Wisconsin X-Ray Microbeam (XRMB) database (46 speakers) was used for SV experiments. In contrast, the proposed approach is evaluated on databases that are larger and also more commonly used for speaker recognition. Using the TIMIT database for SID and the NIST 2008 database for SV, SGCCs are shown to offer complementary information to the MFCC-only system.

6.2 Proposed framework

A score-level framework is proposed to fuse the information provided by MFCCs and SGCCs (it will be explained later why feature concatenation is difficult). An overview of the proposed framework is presented here (see Figure 6.2) and the implementation details are provided in Section 6.4.

Let the number of speakers to be enrolled for SID or SV be N. Enrollment data are used to train two sets of acoustic models: $\{\lambda_M^{(1)}, ..., \lambda_M^{(N)}\}$ for MFCCs, and $\{\lambda_S^{(1)}, ..., \lambda_S^{(N)}\}$ for estimated SGCCs (details about the SGCC estimator are provided in Section 6.3). Given an unseen test utterance, MFCC and SGCC scores ($\{\ell_M^{(1)}, ..., \ell_M^{(N)}\}$, $\{\ell_S^{(1)}, ..., \ell_S^{(N)}\}$) are computed with respect to the pre-trained models and then combined in a weighted fashion. The combined scores $\{\ell^{(1)}, ..., \ell^{(N)}\}$ are used to make a decision (binary for SV, and 1-of-N for SID).

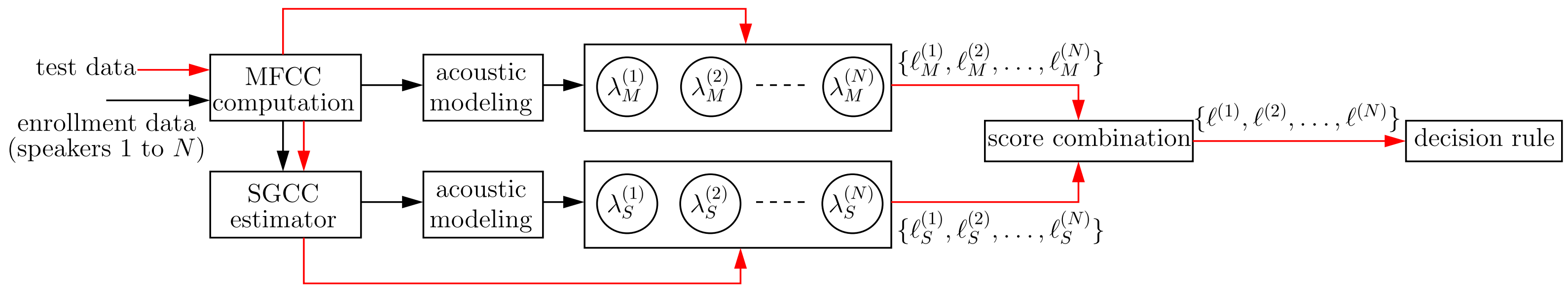


Figure 6.2: *Block diagram of the proposed SID/SV framework. The arrows in black correspond to training (enrollment) and the arrows in red correspond to evaluation. Subscripts M and S denote MFCCs and SGCCs, respectively. The λs denote speaker models and the ℓs denote acoustic model scores (for test data).*

6.3 Estimating SGCCs using MFCCs

In [TBT08], a Bayesian minimum mean squared error (MMSE) estimator was proposed for estimating articulatory parameters from speech acoustics. That approach is adopted here for SGCC estimation and is evaluated using the WashU-UCLA Adults corpus (which contains time-synchronized recordings of speech and subglottal acoustics). The basic mathematical framework for MMSE estimation is provided below (see [TBT08] for a detailed derivation) and the implementation details are deferred to Section 6.3.1.

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M]^\top$ and $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_S]^\top$ be M- and S-dimensional random vectors denoting MFCCs and the corresponding time-synchronized SGCCs, respectively. Let $\mathbf{Z} = [\mathbf{X}^\top \mathbf{Y}^\top]^\top$ denote the joint random vector. Since the distribution of $\mathbf{Z}$ is usually unknown, the simplest way to model it would be via a K-component GMM $\lambda^{(\mathbf{Z})}$:

$$p(\mathbf{z}|\lambda^{(\mathbf{Z})}) = \sum_{k=1}^{K} \nu_k^{(\mathbf{Z})} \mathcal{N}(\mathbf{z}; \mu_k^{(\mathbf{Z})}, \Sigma_k^{(\mathbf{Z})}), \tag{6.1}$$

where $\nu_k \mathcal{N}(\cdot; \mu_k, \Sigma_k)$ denotes the probability density function of the k^{th} mixture component, with mean μ_k, covariance Σ_k and weight ν_k. Once $\lambda^{(\mathbf{Z})}$ is available (from joint training data), the marginal and joint statistics of $\mathbf{X}$ and $\mathbf{Y}$ can be obtained using Eq. (6.2). Note that $\mathbf{Z}$ must be modeled using full covariances in order to extract the joint statistics of $\mathbf{X}$ and $\mathbf{Y}$.

$$\mu_k^{(\mathbf{Z})} = \begin{bmatrix} \mu_k^{(\mathbf{X})} \\ \\ \mu_k^{(\mathbf{Y})} \end{bmatrix}, \quad \Sigma_k^{(\mathbf{Z})} = \begin{bmatrix} \Sigma_k^{(\mathbf{XX})} & \Sigma_k^{(\mathbf{XY})} \\ \\ \Sigma_k^{(\mathbf{YX})} & \Sigma_k^{(\mathbf{YY})} \end{bmatrix} \tag{6.2}$$

Given an unseen test utterance, a sequence of MFCC vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ is first extracted from it. Then, for a given MFCC vector $\mathbf{x}_t$ $(1 \leq t \leq T)$, the SGCC vector is computed as the conditional mean (or MMSE estimate) of $\mathbf{Y}$:

$$\hat{\mathbf{y}}_t = E[\mathbf{Y}|\mathbf{x}_t] = \sum_{k=1}^{K} P(k|\mathbf{x}_t, \lambda^{(\mathbf{Z})})\zeta_{k,t}^{(\mathbf{Y})}, \tag{6.3}$$

where $E[\cdot]$ denotes the expectation operator, and $P(k|\mathbf{x}_t, \lambda^{(\mathbf{Z})})$ and $\zeta_{k,t}^{(\mathbf{Y})}$ are defined as in Eqs. (6.4) and (6.5), respectively.

$$P(k|\mathbf{x}_t, \lambda^{(\mathbf{Z})}) = \frac{\nu_k^{(\mathbf{Z})}\mathcal{N}(\mathbf{x}_t; \mu_k^{(\mathbf{X})}, \Sigma_k^{(\mathbf{XX})})}{\sum_{k'=1}^{K} \nu_{k'}^{(\mathbf{Z})}\mathcal{N}(\mathbf{x}_t; \mu_{k'}^{(\mathbf{X})}, \Sigma_{k'}^{(\mathbf{XX})})} \tag{6.4}$$

$$\zeta_{k,t}^{(\mathbf{Y})} = \mu_k^{(\mathbf{Y})} + \Sigma_k^{(\mathbf{YX})}\Sigma_k^{(\mathbf{XX})^{-1}}(\mathbf{x}_t - \mu_k^{(\mathbf{X})}) \tag{6.5}$$

In the present study, the MMSE estimator of Eq. (6.3) provides a mapping from the more-variable MFCC space to the less-variable SGCC space (can be viewed in some sense as a many-to-one mapping). On the other hand, in [TBT08], the same MMSE estimator provides a one-to-many mapping from speech acoustics to articulatory parameters.

6.3.1 Implementation details and evaluation setup

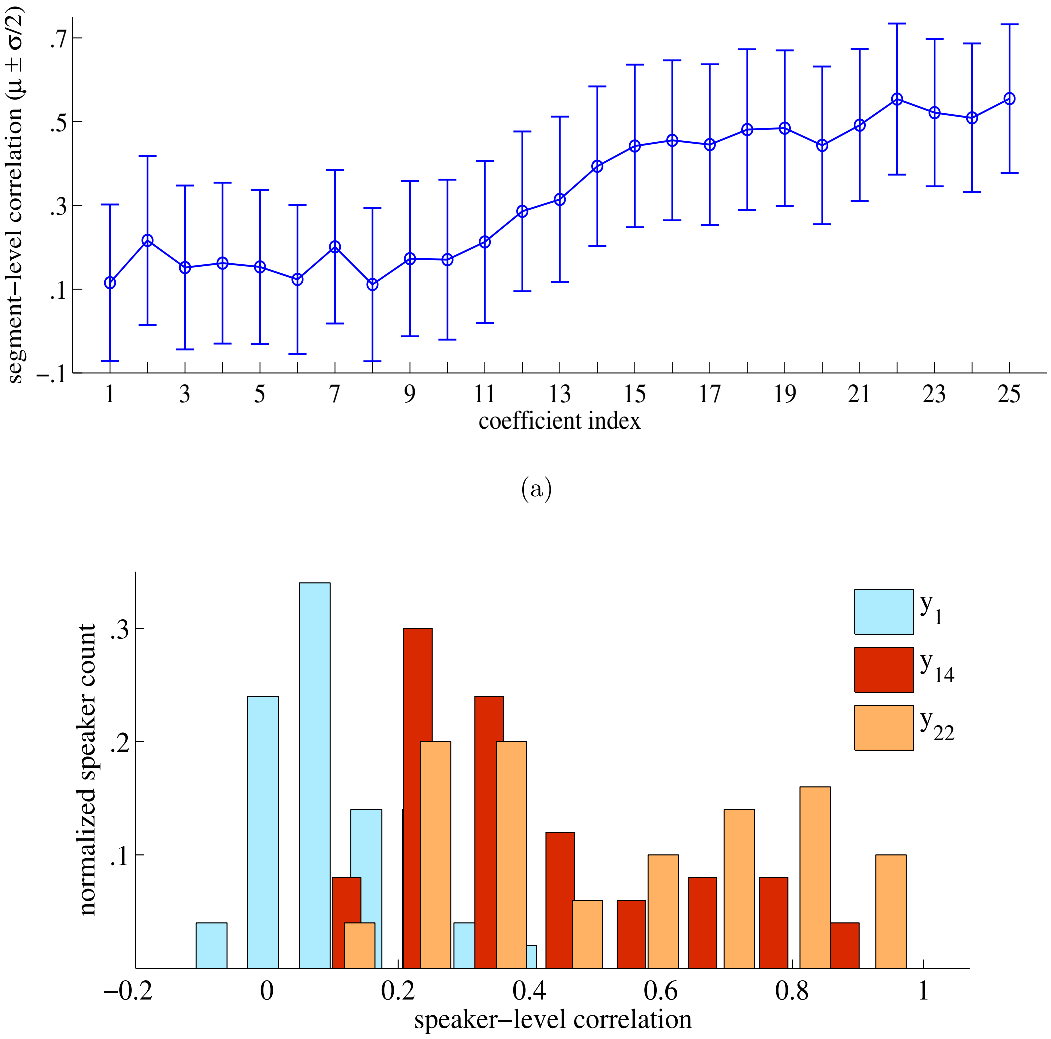
The databases used by studies on speech-to-articulatory inversion consist of read speech utterances (and time-synchronized articulatory trajectories) with good phonetic and lexical coverage. The WashU-UCLA Adults corpus, in contrast, consists only of short phrases of the form "I said a h[V]d again," where [V] is one of 9 monophthongs, 4 diphthongs, or the approximant [ɹ] (the corpus has 10 repetitions of each phrase from 50 adult speakers of American English—25 male and 25 female). To avoid redundancy in the training data that are used to estimate $\lambda^{(\mathbf{Z})}$ (note that all phrases have the same content except for the vowel [V]), only the vowel segments are isolated and used. However, since vowels form only a part of the speakers' phonetic space, there needs to be a way to deal with non-vowel segments while estimating SGCCs for speaker recognition. Section 6.4 explains how this is done.

The MMSE estimator (described above through Eqs. (6.1)–(6.5)) is evaluated using 5-fold cross validation. The available vowel samples (7000 in total: 50 speakers, 14 vowels, 10 repetitions) are split into 5 sets such that the data from any given speaker belong to exactly one set. All signals are down sampled to 8 kHz (from their original sampling rate of 48 kHz). MFCCs and SGCCs are extracted at 5 ms intervals using a 20 ms Hamming window and a 26-channel Mel filter bank. The zeroth cepstral coefficient is discarded; MFCCs x_1–x_{25} and SGCCs y_1–y_{25} are used to train $\lambda^{(\mathbf{Z})}$. The number of components K is set to 16—roughly one component per vowel (no significant improvements in performance were observed by increasing K beyond 16).

6.3.2 Results

SGCC estimates from all 5 test sets are pooled together for analysis. The utility of the estimates (for speaker recognition) is assessed in two ways: (1) by computing the correlation between actual and estimated SGCCs on a per-segment basis (i.e., correlations between actual and estimated time trajectories), and (2) by comparing actual and estimated SGCCs with regard to their ability to discriminate between speakers.

Figure 6.3(a) shows the average segment-level correlations (with error bars) for SGCCs y_1 to y_{25}—the values lie between 0.12 and 0.55, and are comparable to



(a)

(b)

Figure 6.3: *(a) Means (circles) and standard deviations (error bars) of the segment-level correlations (segment = vowel token) between actual and estimated SGCCs. Results from all 50 speakers in the WashU-UCLA Adults corpus are pooled together. (b) Distribution of speaker-level correlation (i.e., average segment-level correlation on a per-speaker basis) for three different cepstral coefficients (y_1, y_{14}, y_{22}).*

Feature set	J-Ratio
MFCCs (x_1–x_{25})	5.32
Actual SGCCs (y_1–y_{25})	5.89
Estimated SGCCs	5.79
MFCCs + actual SGCCs	9.04
MFCCs + estimated SGCCs	8.79

Table 6.1: *J-ratio, a measure of class separation (class = speaker), for different features (+ denotes concatenation). Features were extracted from isolated vowel recordings of speech and subglottal acoustics, for all 50 speakers in the WashU-UCLA Adults corpus.*

the correlations achieved for speech-to-articulatory inversion [GN11]. An important observation from Figure 6.3(a) is that the error bars are significantly large, suggesting a high degree of speaker variability in the estimator's performance. Figure 6.3(b) verifies this further via distributions of the speaker-level correlation (i.e., average segment-level correlation on a per-speaker basis). In essence, the discriminatory power of estimated SGCCs can be attributed, in part, to the speaker-dependent nature of the MMSE estimator.

The J-Ratio [Fuk90], a popular measure of class separation, is used to compare the actual and estimated SGCCs in terms of speaker discriminability. Given feature vectors for N speakers, the J-Ratio can be computed using Eqs. (6.6) and (6.7):

$$S_w = \frac{1}{N}\sum_{i=1}^{N} R_i \qquad S_b = \frac{1}{N}\sum_{i=1}^{N}(M_i - M_o)(M_i - M_o)^\top \tag{6.6}$$

$$J = \mathrm{trace}\{(S_b + S_w)^{-1} S_b\}, \tag{6.7}$$

where S_w is the within-class scatter matrix, S_b is the between-class scatter matrix, M_i is the mean vector for the i^{th} speaker, M_o is the mean of all M_is, and R_i is the covariance matrix for the i^{th} speaker (a higher J-Ratio means better separation). Table 6.1 shows the J-Ratio for different feature sets; it leads to two important observations. (1) SGCCs offer better separation than MFCCs. This is partly attributable to the stationarity of subglottal acoustics and the low within-class scatter that results from it. Despite the moderate correlations achieved by the MMSE estimator (Figure 6.3(a)), estimated SGCCs are comparable in performance to actual SGCCs. This suggests again that the discriminatory power of estimated SGCCs is partly due to the speaker-dependent nature of the estimator (Figure 6.3(b)). (2) SGCCs are complementary to MFCCs, as reflected by the significantly higher J-Ratios for the combined feature sets. Note that SGCCs are simply concatenated with MFCCs for this analysis; for speaker recognition experiments, the score-combination framework described in Section 6.2 will be followed (see Figure 6.2).

6.4 Speaker recognition experiments

The acoustic models for SID and SV are simple GMMs (as in [RR95]) and UBM-adapted GMMs (as in [RQD00]), respectively. Given enrollment data, speech segments are first detected using the algorithm proposed in [SKS99]. MFCCs x_0–x_{25} are extracted from the detected speech segments using a 20 ms Hamming window, a 10 ms frame shift, and a 26-channel Mel filter bank. Non-vowel speech frames must be discarded for SGCC estimation since the MMSE estimator is trained on isolated vowels only. Instead of using a vowel detector (which is difficult to implement and also computationally expensive), all strongly-voiced speech frames are retained. A normalized autocorrelation peak value of 0.6 is chosen as the threshold to select strongly-voiced frames (see [ALA14] for further details

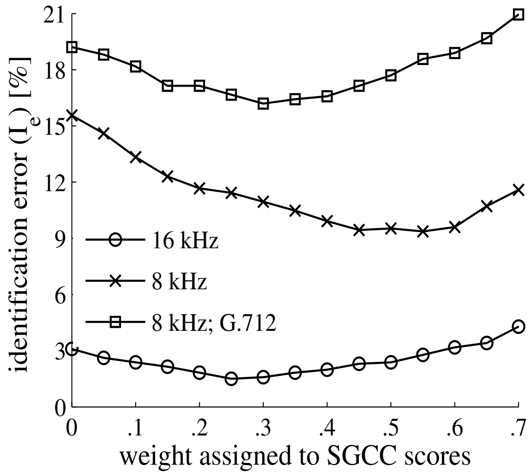


Figure 6.4: *Percent identification error (I_e) as a function of SGCC weight (0 weight = MFCCs only) for the TIMIT database.*

about the voicing detector). Using MFCCs x_1–x_{25}, SGCCs y_1–y_{25} are estimated and used to train the GMMs of the SGCC system. To train the MFCC GMMs, x_0–x_{12} and their first- and second-order derivatives are used. Note that feature concatenation is not possible here—MFCCs are computed for all speech frames whereas SGCCs are computed for strongly-voiced frames, only.

Given a test utterance, MFCCs and SGCCs are computed as described above. The features are scored with their respective models to obtain two sets of scores (see Figure 6.2). The scores are log likelihoods for SID and log likelihood ratios for SV. Each set of scores is normalized to the range [0,1]; this is essential before score combination since MFCC and SGCC scores are generally observed to have different dynamic ranges. The scores from the two systems are combined in a weighted fashion such that the weights (non-negative) sum to 1. The combined scores are used to make a decision. Note that the score-combination procedure is not rigorously optimized here; the focus is more on answering the question as to whether or not SGCCs are beneficial to SID and SV.

Data	Baseline system	Best combined system	Best SGCC weight
16 kHz	3.09	1.51 (51.1%)	0.25
8 kHz	15.56	9.37 (37.8%)	0.55
8 kHz; G.712	19.21	16.19 (15.7%)	0.30

Table 6.2: *Percent identification errors for the TIMIT database in three different conditions, for the baseline (MFCC-only) and the best combined systems (relative reductions in paranthesis).*

6.4.1 Speaker identification: TIMIT database

TIMIT consists of data (sampled at 16 kHz) from 630 speakers. Each speaker has 10 utterances: 2 "shibboleth" *sa* sentences, 5 phonetically-compact *sx* sentences, and 3 phonetically-diverse *si* sentences [Gar88a]. The average utterance length is around 3 seconds. The *sa* sentences are used individually as test trials and the remaining 8 sentences are used for acoustic modeling (as in [RR95]). MFCCs are modeled with 32-component GMMs and SGCCs are modeled with 16-component GMMs.

SID performance is evaluated in three different conditions: (1) wideband (16 kHz sampling rate), (2) narrowband (8 kHz sampling rate), and (3) filtered narrowband (8 kHz sampling rate; data are band-pass filtered using the ITU-T G.712 characteristic [ITU01], which has a flat frequency response from 300 to 3400 Hz). Note that for the filtered narrowband condition, the MMSE estimator is retrained after applying the G.712 characteristic to the vowel segments in the WashU-UCLA Adults corpus.

Figure 6.4 shows the percent identification error (I_e) as a function of the weight assigned to SGCCs, for the three evaluation conditions described above.

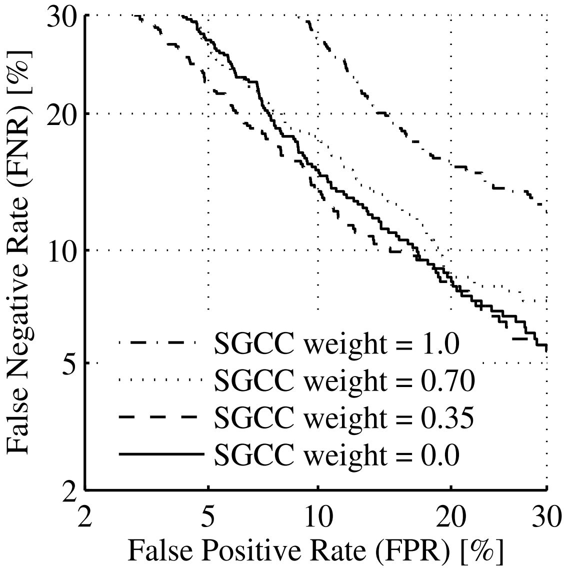


Figure 6.5: *Detection error tradeoff (DET) curves corresponding to different SGCC weights (0 weight = MFCCs only) for the 5 second test trials in the NIST 2008 database.*

Table 6.2 summarizes the results for the best combined systems along with the I_e reductions relative to their respective baselines. SGCCs are clearly effective and complementary (the optimal SGCC weight is less than 0.5, on average) to MFCCs, and one of the reasons for this is the short duration of the test utterances.

6.4.2 Speaker verification: NIST 2008 database

NIST 2008 data (used widely for evaluating SV algorithms) are similar to the filtered narrowband speech of TIMIT, but with significantly higher speaker and channel variability [MG09]. Segments from the “10-sec” condition (which has 10 second utterances from 1336 speakers) are used for this experiment. Data from 892 speakers (having just one utterance each) are used for UBM training. Data from the remaining 444 speakers (having at least two utterances each) are used for enrollment (one utterance) and evaluation (one utterance). The test trials are set up such that each test segment is claimed to belong to each of the 444 speakers, with only one of them being the target speaker. Hence, there are 197136 trials in total. A 128-component UBM is trained for both MFCC and SGCC systems, and speaker models are obtained via maximum *a posteriori* (MAP) adaptation of the UBMs. A relevance factor of 10 is chosen to adapt the means, covariances and component weights. The MSR Identity Toolbox is used for all experiments [SSH13].

Note that the above experimental setup is not a standard one. Typically, UBMs are trained on other corpora (Switchboard, Fisher, NIST 2006, etc.), and NIST 2008 data are used for enrollment and evaluation [KBO07,KOD08,DKD11]. Nevertheless, the above framework serves as a proof-of-concept to demonstrate the efficacy of SGCCs in the presence of speaker and channel variability.

Equal error rate (EER) is used as the performance metric. Evaluation on the 10 second test utterances results in a 4.3% EER reduction for the best combined system (SGCC weight = 0.35), relative to the MFCC-only baseline of 10.59%. The effect of SGCCs is stronger when the test utterances are truncated to 5 seconds each: the best combined system (SGCC weight = 0.35) shows a 10.5% reduction relative to the baseline EER of 12.84%. Detection error tradeoff (DET) curves for the 5 second test trials are shown in Figure 6.5.

6.4.3 Discussion

In both SID and SV tasks, the performance of the combined system drops below the baseline as the SGCC weight tends to 1. However, the J-Ratio analysis of Section 6.3.2 shows that estimated SGCCs, by themselves, can provide better speaker separation than MFCCs. This discrepancy could be arising due to (1) acoustic mismatch between the WashU-UCLA Adults corpus and the speaker recognition corpora, or (2) the simplistic approach used for selecting vowel-like frames for SGCC estimation. The above hypotheses could possibly be verified if the proposed fusion approach can be evaluated using a large, phonetically-balanced database (like TIMIT) of speech and subglottal acoustics.

6.5 Conclusion

Motivated by the speaker-specificity and stationarity of subglottal acoustics, this chapter investigated the utility of subglottal cepstral coefficients (SGCCs) for speaker identification (SID) and verification (SV). SGCCs can be computed using accelerometer recordings of subglottal acoustics, but such an approach is unfeasible in real-world scenarios. To estimate SGCCs from speech signals, the Bayesian minimum mean squared error (MMSE) estimator proposed in the speech-to-articulatory inversion literature was adopted. The joint distribution of SGCCs and speech MFCCs was modeled using the WashU-UCLA Adults corpus (containing simultaneous recordings of speech and subglottal acoustics), and the resulting model was used to obtain an MMSE estimate of SGCCs from unseen (test) MFCCs. Cross-validation experiments on the WashU-UCLA Adults corpus showed that the estimation efficacy, on average, was speaker dependent. A score-level fusion of MFCC and SGCC systems was found to outperform the MFCC-only baseline in both SID and SV tasks. On the TIMIT database (SID), the relative reduction in identification error was 16, 40 and 51% for G.712-filtered (300–3400 Hz), narrowband (0–4000 Hz) and wideband (0–8000 Hz) speech, respectively. On the NIST 2008 database (SV), the relative reduction in equal error rate was 4 and 11% for 10 and 5 second utterances, respectively. Results on the NIST 2008 database suggest that SGCCs could also be potentially effective with more sophisticated modeling schemes such as *i*-vectors.

CHAPTER 7

Summary and future work

This dissertation represents the first detailed investigation of subglottal acoustics from a speech-technology perspective. New databases of time-synchronized speech and subglottal acoustics were collected and analyzed to aid the development of automatic SGR estimation algorithms and novel, hybrid (partly knowledge-based and partly statistical) approaches to body-height estimation, speaker normalization for ASR, and speaker recognition. The emphasis, in general, was on language independence, noise robustness, and efficacy in limited-data conditions.

- **Data collection and analysis**

Three new databases were collected—the WashU-UCLA Adults corpus (50 native speakers of American English), the WashU-UCLA Bilingual Adults corpus (6 bilingual speakers of American English and Mexican Spanish), and the WashU-UCLA Kids corpus (43 native speakers of American English). The recorded material, in all three databases, consisted of CVC words embedded in neutral carrier phrases. Formants and SGRs were measured—using microphone and accelerometer signals, respectively—in the steady-state regions of monophthong vowels. SGRs were found to be practically independent of phonetic content and native language; their average within-speaker coefficients of variation were on the order of 2–5%. The ground truth SGRs (averages of SGR measurements on a per-speaker basis) were significantly higher for children compared to adults, and for adult females compared to adult males. In general, the correlations among SGRs and the correlations between SGRs and body height were found to be stronger (and less influenced by gender differences) for children (r values on the order of 0.9) than for adults (r values on the order of 0.7–0.9). The Bark difference between $F1$ and $Sg1$ was found to be a reliable acoustic measure of vowel height, and the Bark difference between $F2$ and $Sg2$ was found to be a reliable acoustic measure of vowel backness.

The databases collected in the present work as well as in previous studies have all been in nontonal languages (i.e., languages in which pitch is not used to convey lexical or grammatical information). To understand the potential interactions among SGRs, formants and tones, data must be collected in tonal languages such as Mandarin and Gujarati. Preliminary work in this direction had begun at the time this dissertation was written.

- **Automatic estimation of SGRs**

Automatic SGR-estimation algorithms were developed for both adults and children. For adults, $Sg1$ and $Sg2$ were estimated based on the fact that they form natural boundaries between [+low]/[-low] and [+back]/[-back] vowels, respectively. $Sg3$ was estimated based on its correlation with $Sg2$. The algorithm used the SNACK toolkit for automatic pitch and formant tracking, and provided utterance-level SGR estimates by averaging the frame-level estimates obtained in voiced speech segments. Based on experiments with the WashU-UCLA corpora and the MIT Tracheal Resonance database, the algorithm's performance, in terms of RMSE, was found to be practically independent of phonetic content and language. In addition, the algorithm was found to be robust to different noise types at SNRs as low as 0 dB. Using just 2–3 seconds of speech, $Sg1$, $Sg2$ and $Sg3$ were estimated, on average, to within 31, 64 and 116 Hz, respectively; these errors were 1 to 2 times the average within-speaker standard deviations in measured SGR frequencies. In the case of children, $Sg1$ and $Sg2$ were estimated using an ANN-based nonlinear mapping between vocal-tract parameters ($F0$, $F1$, $F2$ and $F3$) and ground truth SGRs. $Sg3$ was estimated as in the case of adults. The algorithm (based on the SNACK toolkit) was robust to different noise types at SNRs of 5 dB or more. The average RMSEs incurred in estimating $Sg1$, $Sg2$ and $Sg3$ were 66, 153 and 143 Hz, respectively; these errors were 2 to 3 times the average within-speaker standard deviations in measured SGR frequencies.

The proposed algorithms were based on empirically-derived relationships involving SGRs, formants and $F0$. Such an approach was adopted to make the algorithms simple, fast, noise robust and accurate enough for the applications considered in this dissertation. To achieve more accurate estimation of SGRs, one approach could be to jointly estimate the subglottal and supraglottal transfer functions such that, in conjunction, they best explain the given speech signal. Mathematical models of sound propagation in the subglottal tract (e.g., see [HKP01]) and vocal tract (e.g., see [SLT00]) could potentially be used to develop such an approach.

- **Body height estimation using SGRs**

The proposed method was motivated by the physiological correlation between overall body height and the effective length of the subglottal system, and was based on the correlation observed between SGRs and height. Using the ground truth SGRs and self-reported heights of speakers in the WashU-UCLA Adults corpus and the WashU-UCLA Bilingual Adults corpus, first-order linear models were trained to predict height given SGR frequencies. Given a speech signal, speaker height was estimated by first estimating SGRs and then using the empirical relations between SGRs and height. The method was evaluated on 604 speakers in the TIMIT database, in three evaluation conditions: (1) clean wideband speech, (2) clean narrowband speech, and (3) noisy narrowband speech. Using about 3 seconds of clean speech data (wideband or narrowband), speaker height could be estimated, on average, to within 5.4 cm. The degradation in height-estimation performance due to noise was minimal—less than 0.3 cm, on average. The proposed method was found to be comparable to state-of-the-art approaches in performance while being more transparent (well-motivated features), efficient (small feature set and limited training data) and generalizable (test corpus much larger than the training corpus).

Actual and estimated height correlated well ($r \sim 0.7$) when the results for male and female speakers were pooled together, but not when they were considered separately ($r \leq 0.2$). One reason for this could be the simplistic nature of the first-order linear models used for prediction. It might be possible to achieve better within-gender correlations (close to 0.5, as predicted using physiological data reported in this dissertation and in other studies) with the help of a larger training set that would allow the development of more sophisticated (probably nonlinear) models between SGRs and speaker height.

- **Speaker normalization using SGRs**

The proposed SGR-based normalization scheme (SGRN) was implemented using a warping function that mapped the first three SGRs of a given target utterance to the first three SGRs of a reference speaker. The reference SGRs were chosen *a priori* based on the speaker population used for training ASR models. The target SGRs for a given utterance were first determined using the proposed SGR-estimation algorithms and then refined by up to ±10% using a maximum-likelihood grid search. SGRN and conventional VTLN were evaluated on children's speech using the TIDIGITS database (noisy data sets were created by adding babble, car, pink and white noise types at SNRs ranging between 5 and 15 dB). Two front ends were considered—standard MFCCs, and the recently-proposed, noise-robust PNCCs (which have been used successfully for adults' ASR in noise). On average, SGRN was found to be significantly better than VTLN in mismatched training (up to 32% relative reduction in WER) and slightly better than VTLN in matched training. Regardless of the training condition, SGRN offered significant WER reductions (relative to VTLN) for 6–8 year old speakers and 1 or 2 word utterances. The efficacy of SGRN can be attributed to (1) speaker specificity and content independence of SGRs, (2) noise robustness of the proposed SGR-estimation algorithms and their ability to provide reliable estimates with limited data, and (3) semi-statistical approach to estimating the optimal warping parameters.

One limitation of SGRN in its current form is the use of a large 3-dimensional search grid—such an approach could be prohibitive for large vocabulary ASR. The grid search could potentially be replaced by gradient search methods for higher efficiency as well as accuracy, although deriving closed-form expressions of the cost function (for gradient search) would be a challenging task. Another direction for future work would be to evaluate SGRN in conjunction with standard speaker-adaptation schemes like MLLR and CMLLR.

- **Speaker recognition using subglottal cepstral coefficients**

Preliminary experiments on the TIMIT database showed that SGRs estimated from speech signals were not discriminative enough from a speaker-recognition perspective. Therefore, subglottal cepstral coefficients (SGCCs) were used for speaker identification (SID) and verification (SV). To estimate SGCCs from speech signals, the Bayesian minimum mean squared error (MMSE) estimator proposed in the speech-to-articulatory inversion literature was adopted. The joint distribution of SGCCs and speech MFCCs was modeled using the WashU-UCLA Adults corpus (containing simultaneous recordings of speech and subglottal acoustics), and the resulting model was used to obtain an MMSE estimate of SGCCs from unseen (test) MFCCs. Cross-validation experiments on the WashU-UCLA Adults corpus showed that the estimation efficacy, on average, was speaker dependent. A score-level fusion of MFCC and SGCC systems was found to outperform the MFCC-only baseline in both SID and SV tasks. On the TIMIT database (used for SID), the relative reduction in identification error was 16, 40 and 51% for G.712-filtered, narrowband and wideband speech, respectively. On the NIST 2008 database (used for SV), the relative reduction in equal error rate was 4 and 11% for 10 and 5 second utterances, respectively. SGCCs therefore were complementary to MFCCs, especially when speech data were limited.

Since the WashU-UCLA Adults corpus comprises carrier phrases of the form "I said a CVC again," the Bayesian MMSE estimator was trained using just the vowel segments extracted from the recordings. The estimator can possibly be trained better if a large, phonetically-balanced database (like TIMIT) of speech and subglottal acoustics were available; this is something that can be pursued in the future. Another direction for further investigation would be to evaluate the utility of SGCCs in SV systems based on state-of-the-art *i*-vectors.

References

[ALA14] H. Arsikere, S. M. Lulich, and A. Alwan. "Estimating speaker height and subglottal resonances using MFCCs and GMMs." *IEEE Signal Processing Letters*, **21**(2):159–162, 2014.

[AMS96] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. "A compact model for speaker-adaptive training." In *Proceedings of ICSLP*, pp. 1137–1140, 1996.

[BBD05] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. J. Russell, S. Steidl, and M. Wong. "The PF_STAR children's speech corpus." In *Proceedings of Interspeech*, pp. 2761–2764, 2005.

[BW09] P. Boersma and D. Weenink. "Praat: doing phonetics by computer (Version 5.1.05)[Computer program].", 2009.

[CA06] X. Cui and A. Alwan. "Adaptation of children's speech with limited data based on formant-like peak alignment." *Computer Speech and Language*, **20**(4):400–419, 2006.

[CB85] B. Cranen and L. Boves. "Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production." *Journal of the Acoustical Society of America*, **77**(4):1543–1551, 1985.

[CBG09] T. G. Csapó, Z. Bárkányi, T. E. Gráczi, T. Bőhm, and S. M. Lulich. "Relation of formants and subglottal resonances in Hungarian vowels." In *Proceedings of Interspeech*, pp. 484–487, 2009.

[Che02] H. A. Cheyne. *Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck*. PhD thesis, Massachusetts Institute of Technology, 2002.

[Chi85] L. A. Chistovich. "Central auditory processing of peripheral vowel spectra." *Journal of the Acoustical Society of America*, **77**:789–805, 1985.

[CS07] X. Chi and M. Sonderegger. "Subglottal coupling and its influence on vowel formants." *Journal of the Acoustical Society of America*, **122**:1735–1745, 2007.

[CSR06] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. "Support vector machines using GMM supervectors for speaker verification." *IEEE Signal Processing Letters*, **13**(5):308–311, 2006.

[DCP06] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan. "A database of vocal tract resonance trajectories for research in speech processing." In *Proceedings of ICASSP*, pp. 369–372, 2006.

[DKD11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech and Language Processing*, **19**(4):788–798, 2011.

[DLM11] G. Dogil, S. M. Lulich, A. Madsack, and W. Wokurek. "Crossing the quantal boundaries of features: subglottal resonances and Swabian diphthongs." *Tones and Features: Phonetic and Phonological Perspectives. De Gruyter Mouton*, pp. 137–148, 2011.

[DM80] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**:357–366, 1980.

[DM95] W. A. van Dommelen and B. H. Moxness. "Acoustic parameters in speaker height and weight identification: sex-specific behaviour." *Language and Speech*, **38**:267–287, 1995.

[Dod01] G. R. Doddington. "Speaker recognition based on idiolectal differences between speakers." In *Proceedings of Interspeech*, pp. 2521–2524, 2001.

[Dus05] S. Dusan. "Estimation of speaker's height and vocal tract length from speech signal." In *Proceedings of Interspeech*, pp. 1989–1992, 2005.

[EB05] D. Elenius and M. Blomberg. "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children." In *Proceedings of Interspeech*, pp. 2749–2752, 2005.

[EG96] E. Eide and H. Gish. "A parametric approach to vocal tract length normalization." In *Proceedings of ICASSP*, pp. 346–348, 1996.

[Esk96] M. S. Eskenazi. "KIDS: a database of children's speech." *Journal of the Acoustical Society of America (A)*, **100**:2759, 1996.

[Fan60] G. Fant. "Acoustic theory of speech production." *The Hague, The Netherlands*, 1960.

[Fan75] G. Fant. "Non-uniform vowel normalization." Technical report, Speech Transmission Laboratory, Royal Institute of Technology, Sweden, 1975.

[FG99] W. T. Fitch and J. Giedd. "Morphology and development of the human vocal tract: A study using magnetic resonance imaging." *Journal of the Acoustical Society of America*, **106**:1511–1522, 1999.

[Fuk90] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, New York, second edition, 1990.

[Gal98] M. J. F. Gales. "Maximum likelihood linear transformations for HMM-based speech recognition." *Computer Speech and Language*, **12**(2):75–98, 1998.

[Gar88a] J. S. Garofolo. "DARPA TIMIT acoustic-phonetic speech database." *National Institute of Standards and Technology*, **15**:29–50, 1988.

[Gar88b] J. S. Garofolo. "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database." *National Institute of Standards and Technology (NIST)*, 1988.

[GG03] D. Giuliani and M. Gerosa. "Investigating recognition of children's speech." In *Proceedings of ICASSP*, pp. 137–140, 2003.

[GGB07] M. Gerosa, D. Giuliani, and F. Brugnara. "Acoustic variability and automatic recognition of children's speech." *Speech Communication*, **49**:847–860, 2007.

[GL94] J.-L. Gauvain and C.-H. Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE Transactions on Speech and Audio Processing*, **2**(2):291–298, 1994.

[GLC11] T. E. Gráczi, S. M. Lulich, T. G. Csapó, and A. Beke. "Context and speaker dependency in the relation of vowel formants and subglottal resonances—evidence from Hungarian." In *Proceedings of Interspeech*, pp. 1901–1904, 2011.

[GMF10a] T. Ganchev, I. Mporas, and N. Fakotakis. "Audio features selection for automatic height estimation from speech." *Artificial Intelligence: Theories, Models and Applications*, pp. 81–90, 2010.

[GMF10b] T. Ganchev, I. Mporas, and N. Fakotakis. "Automatic height estimation from speech in real-world setup." In *Proceedings of the 18th European Signal Processing Conference*, pp. 800–804, 2010.

[GN11] P. K. Ghosh and S. S. Narayanan. "A subject-independent acoustic-to-articulatory inversion." In *Proceedings of ICASSP*, pp. 4624–4627, 2011.

[Gon04] J. González. "Formant frequencies and body size of speaker: a weak relationship in adult humans." *Journal of Phonetics*, **32**:277–287, 2004.

[Gra18] H. Gray. *Anatomy of the human body.* Lea & Febiger, 1918.

[GS97] E. B. Gouvêa and R. M. Stern. "Speaker normalization through formant-based warping of the frequency scale." In *Proceedings of Eurospeech*, pp. 1139–1142, 1997.

[HCS94] R. H. Habib, R. B. Chalker, B. Suki, and A. C. Jackson. "Airway geometry and wall mechanical properties estimated from subglottal input impedance in humans." *Journal of Applied Physiology*, **77**(1):441–451, 1994.

[HGC95] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler. "Acoustic characteristics of American English vowels." *Journal of the Acoustical Society of America*, **97**:3099–3111, 1995.

[Hir05] G. Hirsch. "FaNT–Filtering and Noise Adding Tool." Technical report, Niederrhein University of Applied Sciences (online: http://dnt.-kr.hsnr.de/download.html), 2005.

[HKP01] P. Harper, S. S. Kraman, H. Pasterkamp, and G. R. Wodicka. "An acoustic model of the respiratory tract." *IEEE Transactions on Biomedical Engineering*, **48**(5):543–550, 2001.

[HPK03] V. P. Harper, H. Pasterkamp, H. Kiyokawa, and G. R. Wodicka. "Modeling and measurement of flow effects on tracheal sounds." *IEEE Transactions on Biomedical Engineering*, **50**(1):1–10, 2003.

[Hrd25] A. Hrdlička. *The Old Americans.* The Williams and Wilkins Company, Baltimore, MD, 1925.

[HTT10] K. Honda, S. Takano, and H. Takemoto. "Effects of side cavities and tongue stabilization: Possible extensions of the quantal theory." *Journal of Phonetics*, **38**:33–43, 2010.

[IMK76] K. Ishizaka, M. Matsudaira, and T. Kaneko. "Input acoustic-impedance measurement of the subglottal system." *Journal of the Acoustical Society of America*, **60**(1):190–197, 1976.

[ITU01] ITU-T recommendation G.712. “Transmission performance characteristics of pulse code modulation channels.” 2001.

[Jun09] Y. Jung. *Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back]*. PhD thesis, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, 2009.

[Kar] M. Karolewski. “Tracer, v.1.7.” *Last accessed on 4/24/2012 (Online: http://sites.google.com/site/kalypsosimulation/Home/data-analysis-software-1).*

[KBO07] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. “Joint factor analysis versus eigenchannels in speaker recognition.” *IEEE Transactions on Audio, Speech and Language Processing*, **15**(4):1435–1447, 2007.

[Kin06] T. Kinnunen. “Joint acoustic-modulation frequency for speaker recognition.” In *Proceedings of ICASSP*, volume 1, pp. 665–668, 2006.

[KOD08] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. “A study of interspeaker variability in speaker verification.” *IEEE Transactions on Audio, Speech and Language Processing*, **16**(5):980–988, 2008.

[KS11] J. Kreiman and D. Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception.* John Wiley & Sons, 2011.

[KS12] C. Kim and R. M. Stern. “Power-normalized cepstral coefficients (PNCC) for robust speech recognition.” In *Proceedings of ICASSP*, pp. 4101–4104, 2012.

[Kun89] H. J. Künzel. “How well does average fundamental frequency correlate with speaker height and weight?” *Phonetica*, **46**:117–125, 1989.

[KYI05] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan. “TBALL data collection: the making of a young children’s speech corpus.” In *Proceedings of Interspeech*, pp. 1581–1584, 2005.

[LAA11] S. M. Lulich, A. Alwan, H. Arsikere, J. R. Morton, and M. S. Sommers. “Resonances and wave propagation velocity in the subglottal airways.” *Journal of the Acoustical Society of America*, **130**:2108–2115, 2011.

[LAM11] S. M. Lulich, H. Arsikere, J. R. Morton, G. Leung, M. S. Sommers, and A. Alwan. "Analysis and automatic estimation of children's subglottal resonances." In *Proceedings of Interspeech*, pp. 2817–2820, 2011.

[LKG13] M. Li, J. Kim, P. Ghosh, V. Ramanarayanan, and S. Narayanan. "Speaker verification based on fusion of acoustic and articulatory information." In *Proceedings of Interspeech*, pp. 1614–1618, 2013.

[LMA12] S. M. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. F. Leung, and A. Alwan. "Subglottal resonances of adult male and female native speakers of American English." *Journal of the Acoustical Society of America*, **132**:2592–2602, 2012.

[LPN99] S. Lee, A. Potamianos, and S. Narayanan. "Acoustics of children's speech: Developmental changes of temporal and spectral parameters." *Journal of the Acoustical Society of America*, **105**:1455–1468, 1999.

[LR98] L. Lee and R. Rose. "A frequency warping approach to speaker normalization." *IEEE Transactions on Speech and Audio Processing*, **6**:49–60, 1998.

[LR01] Q. Li and M. J. Russell. "Why is automatic recognition of children's speech difficult?" In *Proceedings of Eurospeech*, pp. 2671–2674, 2001.

[Lul06] S. M. Lulich. *The role of lower airway resonances in defining vowel feature contrasts.* PhD thesis, Massachusetts Institute of Technology, 2006.

[Lul10] S. M. Lulich. "Subglottal resonances and distinctive features." *Journal of Phonetics*, **38**:20–32, 2010.

[LW95] C. J. Leggetter and P. C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." *Computer Speech and Language*, **9**(2):171–185, 1995.

[Mak75] J. Makhoul. "Linear prediction: A tutorial review." *Proceedings of the IEEE*, **63**:561–580, 1975.

[McD00] J. W. McDonough. *Speaker compensation with all-pass transforms.* PhD thesis, Johns Hopkins University, 2000.

[MG09] A. F. Martin and C. S. Greenberg. "NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels." In *Proceedings of Interspeech*, pp. 2579–2582, 2009.

[MLU96] J. D. Miller, S. Lee, R. M. Uchanski, A. F. Heidbreder, B. B. Richman, and J. Tadlock. "Creation of two children's speech databases." In *Proceedings of ICASSP*, pp. 849–852, 1996.

[MLW08] A. Madsack, S. M. Lulich, W. Wokurek, and G. Dogil. "Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs." *Proceedings of LabPhon*, **11**:91–92, 2008.

[MSW04] J. McDonough, T. Schaaf, and A. Waibel. "Speaker adaptation with all-pass transforms." *Speech Communication*, **42**(1):75–91, 2004.

[MY06] K. S. R. Murty and B. Yegnanarayana. "Combining evidence from residual phase and MFCC features for speaker recognition." *IEEE Signal Processing Letters*, **13**(1):52–55, 2006.

[PA06] S. Panchapagesan and A. Alwan. "Multi-parameter frequency warping for VTLN by gradient search." In *Proceedings of ICASSP*, pp. 1178–1181, 2006.

[PB52] G. E. Peterson and H. L. Barney. "Control methods used in a study of the vowels." *Journal of the Acoustical Society of America*, **24**:369–381, 1952.

[PH97] B. L. Pellom and J. H. L. Hansen. "Voice analysis in adverse conditions: the centennial Olympic park bombing 911 call." In *40th Midwest Symposium on Circuits and Systems*, pp. 873–876, 1997.

[PN03] A. Potamianos and S. Narayanan. "Robust recognition of children's speech." *IEEE Transactions on Speech and Audio Processing*, **11**:603–616, 2003.

[RKN05] D. Rendall, S. Kollias, C. Ney, and P. Lloyd. "Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry." *Journal of the Acoustical Society of America*, **117**:944–955, 2005.

[RQD00] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital Signal Processing*, **10**(1):19–41, 2000.

[RR95] D. A. Reynolds and R. C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models." *IEEE Transactions on Speech and Audio Processing*, **3**(1):72–83, 1995.

[SB00] K. Sjölander and J. Beskow. “Wavesurfer—an open source speech tool.” In *Proceedings of Interspeech*, pp. 464–467, 2000.

[SFK05] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. “Modeling prosodic feature sequences for speaker recognition.” *Speech Communication*, **46**(3):455–472, 2005.

[SG86] A. K. Syrdal and H. S. Gopal. “A perceptual model of vowel recognition based on the auditory representation of American English vowels.” *Journal of the Acoustical Society of America*, **79**:1086–1100, 1986.

[SHC00] K. Shobaki, J.-P. Hosom, and R. Cole. “The OGI kids’ speech corpus and recognizers.” In *Proceedings of ICSLP*, 2000.

[Shu10] Y. Shue. *The voice source in speech production: Data, analysis and models.* PhD thesis, University of California Los Angeles, 2010.

[Sjo97] K. Sjölander. “The Snack sound toolkit.” *KTH, Stockholm, Sweden (Online: http://www.speech.kth.se/snack/)*, 1997.

[SKS99] J. Sohn, N. S. Kim, and W. Sung. “A statistical model-based voice activity detection.” *IEEE Signal Processing Letters*, **6**(1):1–3, 1999.

[SLT00] B. H. Story, A.-M. Laukkanen, and I. R. Titze. “Acoustic impedance of an artificially lengthened and constricted vocal tract.” *Journal of Voice*, **14**(4):455–469, 2000.

[Son04] M. Sonderegger. “Subglottal coupling and vowel space: An investigation in quantal theory.” *B. S. Thesis, Massachusetts Institute of Technology*, 2004.

[SP93] I. Sanchez and H. Pasterkamp. “Tracheal sound spectra depend on body height.” *American Review of Respiratory Disease*, **148**:1083–1083, 1993.

[SSH13] S. O. Sadjadi, M. Slaney, and L. Heck. “MSR Identity Toolbox v1.0: A MATLAB toolbox for speaker-recognition research.” *Speech and Language Processing Technical Committee Newsletter*, November 2013.

[Ste89] K. N. Stevens. “On the quantal nature of speech.” *Journal of Phonetics*, **17**:3–45, 1989.

[Ste98] K. N. Stevens. *Acoustic phonetics.* MIT Press, Cambridge, MA, 1998.

[Ste02] K. N. Stevens. "Toward a model for lexical access based on acoustic landmarks and distinctive features." *Journal of the Acoustical Society of America*, **111**(4):1872–1891, 2002.

[SU08] R. Sinha and S. Umesh. "A shift-based approach to speaker normalization using non-linear frequency-scaling model." *Speech Communication*, **50**:191–202, 2008.

[TA13] L. N. Tan and A. Alwan. "Multi-band summary correlogram-based pitch detection for noisy speech." *Speech Communication*, **55**:841–856, 2013.

[Tal95] D. Talkin. "A robust algorithm for pitch tracking (RAPT)." *Speech Coding and Synthesis*, pp. 495–518, 1995.

[TBT08] T. Toda, A. W. Black, and K. Tokuda. "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model." *Speech Communication*, **50**(3):215–227, 2008.

[Tra90] H. Traunmüller. "Analytical expressions for the tonotopic sensory scale." *Journal of the Acoustical Society of America*, **88**:97–100, 1990.

[VS93] A. Varga and H. J. M. Steeneken. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems." *Speech Communication*, **12**:247–251, 1993.

[Wak77] H. Wakita. "Normalization of vowels by vocal-tract length and its application to vowel identification." *IEEE Transactions on Acoustics, Speech and Signal Processing*, **25**(2):183–192, 1977.

[WAL08] S. Wang, A. Alwan, and S. M. Lulich. "Speaker normalization based on subglottal resonances." In *Proceedings of ICASSP*, pp. 4277–4280, 2008.

[WJ96] J. G. Wilpon and C. N. Jacobsen. "A study of speech recognition for children and the elderly." In *Proceedings of ICASSP*, pp. 349–352, 1996.

[WLA08] S. Wang, S. M. Lulich, and A. Alwan. "A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation." In *Proceedings of Interspeech*, pp. 1717–1720, 2008.

[WLA09a] S. Wang, Y.-H. Lee, and A. Alwan. "Bark-shift based nonlinear speaker normalization using the second subglottal resonance." In *Proceedings of Interspeech*, pp. 1619–1622, 2009.

[WLA09b] S. Wang, S. M. Lulich, and A. Alwan. "Automatic detection of the second subglottal resonance and its application to speaker normalization." *Journal of the Acoustical Society of America*, **126**:3268–3277, 2009.

[WM09] W. Wokurek and A. Madsack. "Comparison of manual and automated estimates of subglottal resonances." In *Proceedings of Interspeech*, pp. 1671–1674, 2009.

[YLL09] C. H. You, K. A. Lee, and H. Li. "An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition." *IEEE Signal Processing Letters*, **16**(1):49–52, 2009.

[ZW97] P. Zhan and M. Westphal. "Speaker normalization based on frequency warping." In *Proceedings of ICASSP*, pp. 1039–1042, 1997.

[Zwi61] E. Zwicker. "Subdivision of the audible frequency range into critical bands." *Journal of the Acoustical Society of America*, **33**(2):248, 1961.