Title
Subglottal Resonances: Coupling Effects and Application to Automatic Speaker Identification

Author
Leung, Gary Ka Fu

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Subglottal Resonances:

Coupling Effects and Application to Automatic Speaker Identification

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Electrical Engineering

by

Gary Ka Fu Leung

2012

ABSTRACT OF THE THESIS

Subglottal Resonances:

Coupling Effects and Application to Automatic Speaker Identification

by

Gary Ka Fu Leung

Master of Science in Electrical Engineering

University of California, Los Angeles, 2012

Professor Abeer Alwan, Chair

Subglottal resonances (SGRs) have been extensively studied in recent years due to their demonstrated advantages in different applications, such as speaker height estimation and speaker normalization in automatic speech recognition. In the interest of studying this area in speech processing, the current study does not only extend the previously explored oral-subglottal coupling effects, but also investigates the application of speaker identification with SGRs. By using newly-developed tools, a more generalized analysis of the coupling effect is conducted with a larger database compared to previous studies. In order to demonstrate the importance of SGRs, exploratory speaker identification experiments with SGR features from both "ground truth" measurements and statistical based estimation techniques are carried out. The results show the

effectiveness of SGR features with preliminary analysis, and several suggestions are made to motivate further study.

The thesis of Gary Ka Fu Leung is approved.

Jin Hyung Lee

Lieven Vandenberghe

Abeer Alwan, Committee Chair

University of California, Los Angeles

2012

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to express the deepest gratitude to my academic adviser, Professor Abeer Alwan for her guidance and support throughout my Master degree study at UCLA. I sincerely thank her for serving as my Master's Thesis Committee Chair and providing me the excellent opportunity to work as a graduate student researcher in the Speech Processing and Auditory Perception Laboratory (SPAPL) at UCLA. I have earned tremendous amount of valuable experience which lights up the path to the next stage in my life.

Secondly, I am indebted to all my colleagues in SPAPL for their insightful interactions. In particular, I would like to give special thanks to Harish Arsikere from SPAPL and Steven Lulich from Washington University, for sharing their knowledge and expertise in subglottal resonances.

Moreover, it is my honor to have the privilege to collaborate and co-author several publications [**1**],[**2**],[**3**],[**4**],[**5**] in the last two years with Professor Abeer Alwan, Professor Mitchell Sommers, Steven Lulich, Harish Arsikere and John Morton. I am grateful for their support and inspiration for the current study. Some portions of [**1**] and [**5**] are presented in Section 3.2 to aid the study presented in Chapter 3.

In addition, I would like to thank both Professor Jin Hyung Lee and Professor Lieven Vandenberghe for serving as my Master's Thesis Committee.

Last but not least, this thesis would not have been possible without the love and support from my parents and girlfriend.

This thesis is supported in part by the NSF grant number 0905381.

CHAPTER 1. INTRODUCTION

1.1 Motivation

The human speech production system has been studied for decades, but most of the research interest in the early days was focused on the supraglottal airway system, which consists of the vocal tract. In particular, acoustic properties of speech, such as fundamental frequency (F0) and formant frequencies (F1, F2, F3, etc.) have been well studied [**6**], [**7**]. In the recent past, several studies have focused on the subglottal airway system which includes the larynx, trachea, two main bronchi, and the bronchi tree as shown in Figure 1. Studies have shown that subglottal resonances (SGRs) are useful in several tasks, such as speaker normalization and adaptation for automatic speech recognition (ASR) [**1**], [**8**], [**9**], [**10**], [**11**], and speaker height estimation [**4**], [**5**]. Although SGRs can be measured non-invasively using an accelerometer [**12**] and [**13**], it is more practical if they can be estimated from speech signals.

Motivated by acoustical coupling between the subglottal and supraglottal systems [**6**], [**12**], [**13**], two major approaches have been developed to estimate SGRs automatically as in [**4**], [**5**], [**8**], [**9**], [**11**], [**14**], and [**15**]. One approach estimates SGRs indirectly from speech signals by the demonstrated boundary between [+low] and [-low] vowels by Sg1, and the boundary between [+back] and [-back] by Sg2. This approach has been proven to yield effective results for several applications [**1**] and [**4**]. The other approach focuses on direct estimation by using several properties of the oral-subglottal coupling effect and reinforces studies on the interactions

between supraglottal and subglottal systems. For example, [**10**] and [**11**] use the second formant frequency discontinuities and magnitude attenuations caused by the second subglottal resonance (Sg2) as part of the estimation method for different tasks. Although the previous studies such as [**12**] and [**16**] have analyzed the oral-subglottal coupling with limited-data, a more generalized study with a larger data set is desirable in order to further explore different properties of the coupling effects.

Studies such as [**3**] and [**12**] have shown that SGRs have relatively small intra-speaker variability compared to acoustic properties of the vocal tract. In other words, SGRs contain more speaker-specific information, which could be a desirable feature for automatic speaker identification. A similar argument is used in [**17**] for their proposed 8 acoustic parameters, which contain speaker-specific information and have been demonstrated to give comparable performance to the standard cepstral features for automatic speaker identification. In order to investigate this hypothesis, an exploratory study on automatic speaker identification is undertaken in this thesis.

1.2 The roles of subglottal and supraglottal systems in speech production

In order to understand the roles of subglottal and supraglottal systems in human speech production, it is helpful to start with the physical system and the corresponding mathematical model. Figure 2 shows a schematic diagram of a simple model proposed by [**18**] for simulating the speech production system. The subglottal system (modeled by the lungs, trachea and bronchi in the diagram) provides power for air to flow through the larynx and vocal tract [**19**]. It is the ‘battery’ of the speech production system, and it allows the generation of different sources for

different languages around the world [**6**]. The air flows through the vocal cords to produce voiced or unvoiced sounds. The generated sound propagates through the supraglottal system and is shaped by the vocal tract to produce speech. The vibration frequency of the vocal cords is called the fundamental frequency (F0) while the resonance frequencies of the vocal tract tube are called formant frequencies (e.g. F1, F2) [**19**]. Although the model described here is a simplified version which captures only a few important components of the complicated speech production system, it is sufficient to serve as a basis for understanding the overall system.

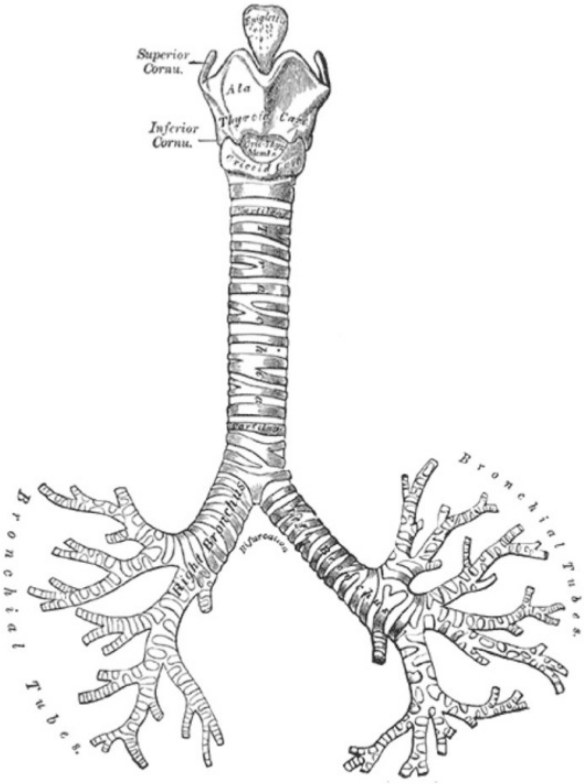

Figure 1: An anatomical sketch of the subglottal airway, including the larynx, trachea, two main bronchi, and the bronchiole tree (adapted from [**20**]).

In terms of mathematical models, the speech production system is commonly modeled by a linear time-invariant system as described in [**19**]. The excitation signal, $e(t)$, is modeled as a periodic signal or random signal for voiced or unvoiced speech, respectively. This source signal is then filtered by the vocal tract impulse response, $v(t)$, to produce the resulting speech waveform, $s(t)$. The filtering process is achieved by convolution in the time domain which is equivalent to multiplication in frequency domain. Figure 3 demonstrates this simplified model of the speech production system where $E(j\Omega)$, $V(j\Omega)$, and $S(j\Omega)$ are the Fourier transform representations for the excitation, vocal tract, and resulting signals, respectively.

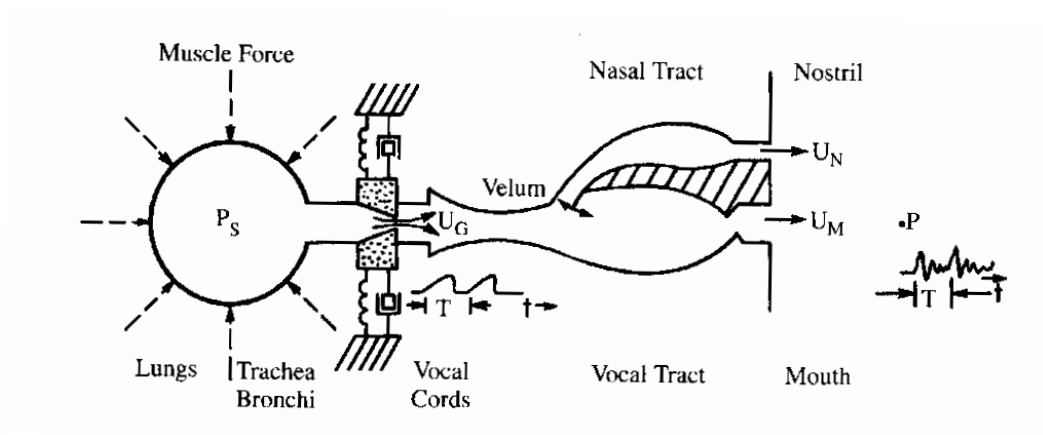

Figure 2: Schematized diagram of the speech production mechanism (adapted from [**18**])

According to the acoustic theory of speech production, most speech sounds, except consonants, can be characterized by using an all-pole model to represent the vocal tract transfer function. In some cases, the effect of zeros in nasals, for example, can also be approximated by additional poles [**21**]. Although the all-pole model captures most of the characteristics of sound

propagating through the vocal tract, there are still some uncertainties partly due to the behavior of the subglottal system.

Studies such as [**6**], [**12**], and [**13**] have shown that the lower airway introduces new zero-pole pairs, which correspond to SGRs, to the speech signal through coupling. This coupling is not only limited to air-only coupling when the glottis is open, but also can be achieved across the vocal fold tissues as described in [**13**].

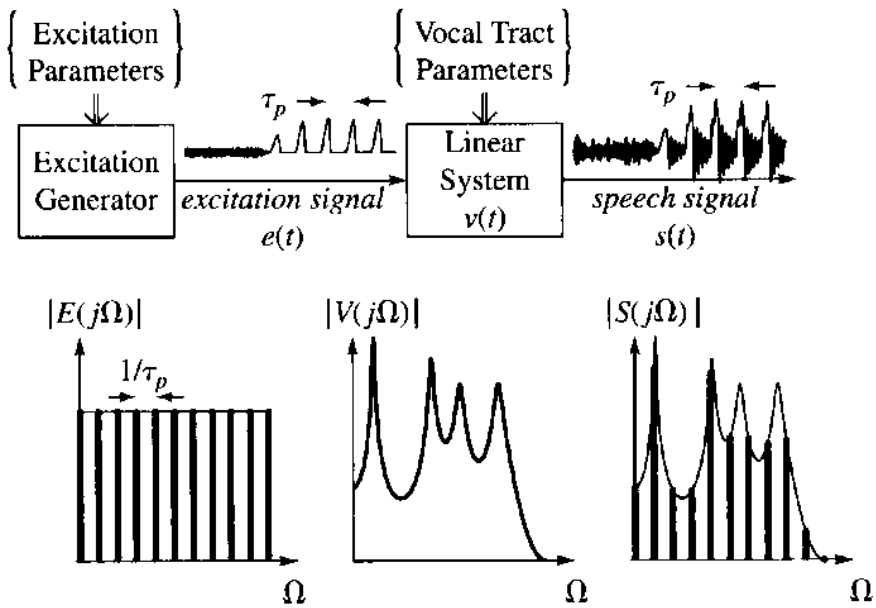

Figure 3: Linear speech production model representing the source, vocal tract, and resulting speech signal. The figure presents temporal and spectral representations on top and bottom, respectively (adapted from [**19**])

1.3 Existing Studies

Although there is some literature on the interaction between the subglottal and supraglottal systems, the coupling effect between the two cavities has been somewhat ignored due to the high complexity of the interaction. However, this effect becomes important when formant frequencies approach SGR frequencies, and more studies have started looking into this issue in the last decade [**10**], [**11**], [**12**], [**13**], [**16**].

In particular, the two pioneering studies on coupling effects, [**12**] and [**16**], model the interaction between supraglottal and subglottal cavities by using coupled resonators. As shown in Figure 4, a circuit model is introduced by [**22**] to model the coupled speech production system with both subglottal and supraglottal systems. By following the derivation in [**12**], the transfer function of the circuit, $T(\omega)$, can be calculated by inverse-filtering the volume velocity at the lips, U_m, by the output of the glottis, U_o. It is decomposed by introducing the volume velocity of the airflow , U_v, into the vocal tract as,

$$T(\omega) = \frac{U_m}{U_v}\frac{U_v}{U_o}. \tag{1}$$

The term U_m/U_v can be calculated by impedance matching, and it is commonly used in [**19**] to model the vocal tract transfer function when the subglottal coupling effect is ignored. The additional term, U_v/U_o, is determined by solving the circuit in Figure 4 which gives

$$\frac{U_v}{U_o} = \frac{Z_g}{Z_g+Z_v+Z_l}, \tag{2}$$

where Z_g, Z_v, and Z_l, are the impedances of the glottis, vocal tract and subglottal system, respectively. This term, U_v/U_o, characterizes the coupling effect by the subglottal system. When the coupling effect is ignored, the glottal impedance, Z_g, is assumed to be infinite and U_v/U_o will be equal to 1, as expected. Otherwise, a zero will be introduced at the local maximum of Z_l and a pole will be added at the local minimum of the sum of Z_g, Z_v, and Z_l corresponding to each SGR. A more detailed derivation is presented in [**12**].

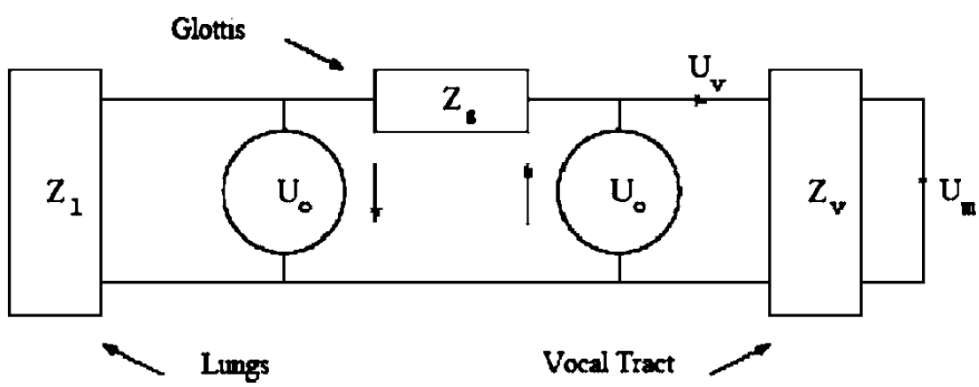

Figure 4: Equivalent circuit model of the subglottal and supraglottal systems (adapted from [**22**]). Z_l is the impedance of the subglottal system, Z_g is the impedance of the glottis, Z_v is the impedance of the vocal tract, U_v is air flow into the vocal tract, U_m is the volume velocity at the lips and the two U_o are the volume velocity sources.

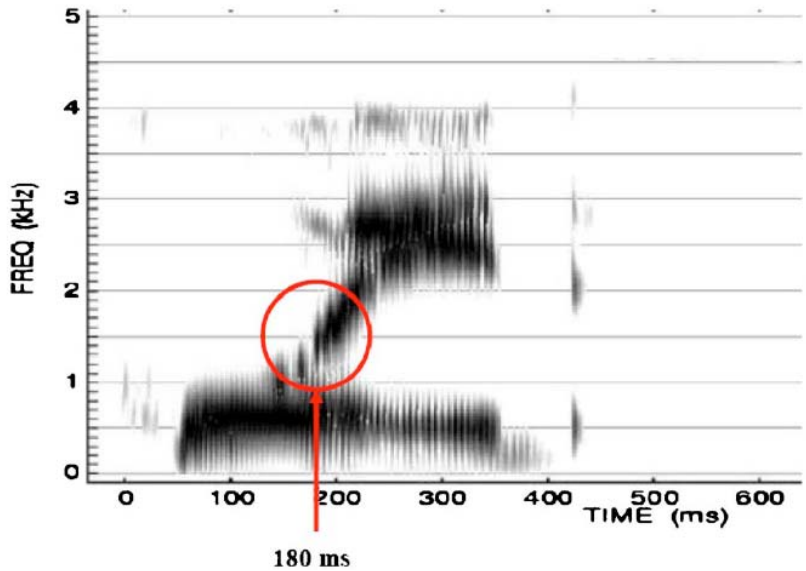

Figure 5: A spectrogram of a diphthong /ɔɪ/ spoken by a male speaker. Attenuation of the second formant (F2) and a discontinuity occur at 180 ms around the measured Sg2 at 1370Hz (adapted from [**12**]).

Based on this model, stimulation and analysis have been done to investigate an observed phenomenon as shown in Figure 5. The spectrogram shows a small frequency discontinuity and attenuation of the second formant (F2) around the measured Sg2. A simulator is built and tested with values in the range of a common male speaker as reported by other studies. The simulated model with coupling clearly outperforms the model without coupling in tracking the actual F2 path, especially by simulating the frequency jump around the second subglottal frequency region as shown in Figure 6. Although this simulated result demonstrates the power of the model to capture the coupling phenomenon, uncertainties still exist, including the vocal tract variability among speakers and the effect of formant amplitude attenuation [**16**].

In order to quantify and further investigate the coupling effect, both [**12**] and [**16**] have defined some preliminary procedures to measure and analyze some subsets of a relatively small

database. In [**12**], 6 speakers with up to 10 diphthong tokens of /aɪ/ and /ɔɪ/ (total of 60 or less tokens) were examined for formant frequency jumps and amplitude attenuation around the SGR frequency regions. On the other hand, 14 speakers with up to 4 diphthong tokens of /ɔɪ/ (total of 56 or less tokens) were used in [**16**] from the same database, which contains the ground truth SGR data recorded by an accelerometer simultaneously during speech recording with a microphone. Since the SGRs are relatively constant for a given speaker, averaged SGRs from both manual DFT measurement and automatic formant tracking are reported for every speaker.

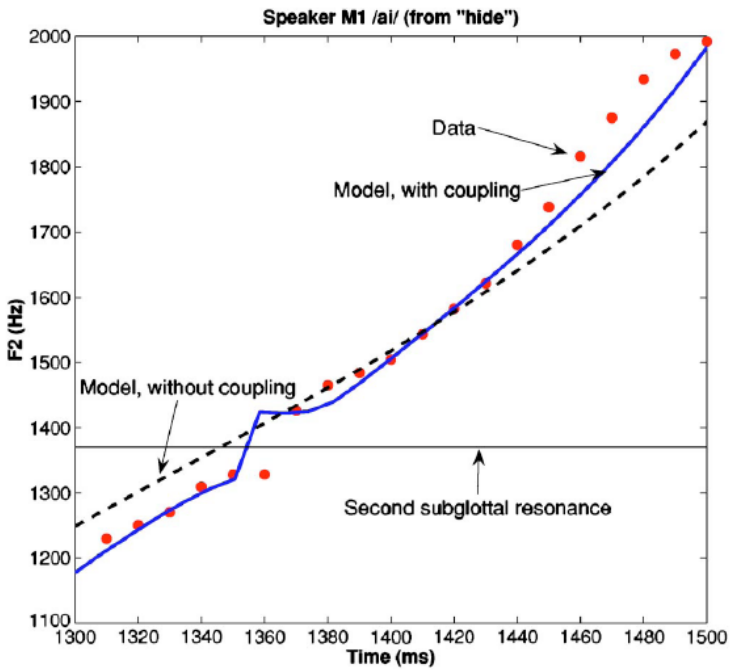

Figure 6: Measured data of a second formant (F2) track and the simulated model with and without coupling effect (adapted from [**12**]).

There are a few major and interesting findings from the two studies. First, the measurements of the frequency jump and amplitude attenuation are made manually by using a

shifting Hamming window for consecutive frames. However, the results are sensitive to both the size and position of the window. By comparing both open-phase and close-phase measurements, the study [**12**] provides evidence for their proposed model which predicts greater coupling effect with larger glottal areas. As a result, the study [**12**] demonstrates that it is experimentally better to use open-phase measurements by centering a half pitch period sized window over the lowest amplitude part of the pitch period until the largest peak in the next period [**12**]. To further investigate the relationship between the oral-subglottal coupling effect and size of glottal area, the correlation between the coupling effect and breathiness is studied in [**16**]. The study uses the difference between the first and second harmonics (H1-H2), which is highly correlated to breathiness [**16**], from 10 tokens of each speaker to show positive correlations with both sizes of frequency jump and amplitude attenuation for each gender. Although a frequency jump cannot always be found in every utterance, the amplitude attenuation has proven to be a better cue for the coupling as it always exists. In addition, a weak negative correlation between the size of a frequency jump and the corresponding amplitude attenuation is reported by [**12**]; however, the result is not consistent with results from [**16**].

Based on these two studies, estimation methods of the SGRs, Sg2 in particular, are developed in [**8**],[**9**], [**10**], and [**11**]. In order to apply SGRs to different applications, such as speaker normalization and adaption, these studies use the cues from frequency discontinuity and amplitude attenuation of diphthongs, such as /aɪ/, to estimate Sg2. When they fail to locate the coupling cues from an isolated vowel, especially in a monophthong, statistical approximations are used instead. Therefore, these estimation methods are self-reported to be unstable sometimes depending on the vowel content. In addition, the reliability of these estimations is tested on data with children's speech only, which highlights the need for further investigation.

In spite of the fact that several studies have looked into the coupling effect from both theoretical and practical points of view, there is room for improvement; hence, the current study is undertaken. For example, the data used by previous studies are relatively small, and it would be better to have a larger database for more reliable statistical analyses. Moreover, a more sophisticated procedure, as presented in [**3**], has been developed recently for measuring the SGRs from accelerometer signals. The new procedure may help improve the analysis by providing more precise and reliable ground truth measurements compared to the DFT-only method or Snack toolkit [**23**] used in previous studies. Moreover, more systematic methods of detecting glottal closure instant might be used to locate open-phase instead of searching the frequency spectrum manually. Finally, it would be helpful to extend the analysis of the coupling effect with more voice quality parameters other than H1-H2. With all these questions and possibilities, this study is formed and presented in the following chapters.

1.4 Organization of this thesis

The rest of the thesis is organized as follows. Chapter 2 contains an extended study of the properties of the coupling effects between subglottal and supraglottal systems using a large database. An exploratory study on speaker identification system with SGRs is presented in Chapter 3. Finally, Chapter 4 summarizes and concludes the thesis.

CHAPTER 2. ANALYSIS OF ORAL-SUBGLOTTAL COUPLING

2.1 Database Used

The WashU-UCLA corpus [**24**] is used for the study in this chapter. The corpus consists of simultaneous speech and subglottal acoustics recorded using a SHURE PG27 microphone and a K&K Sound HotSpot accelerometer, respectively. There are 25 male and 25 female native American English (AE) speakers from 18 to 25 years of age. Two separate sessions are recorded for each speaker. The first session consists of 21 AE 'CVb' words where 'V' has 4 monophthongs and 3 diphthongs in conjunction with three voiced stops, /b/, /d/, and /g/ in the 'C'. The other session has 14 AE 'hVd' words where the vowel set 'V', includes 9 monophthongs, 4 diphthongs and the approximant [ɹ]. The list of target vowels 'V', for both sessions is shown in Table 1. Each of the 35 words from both sessions is embedded in a phonetically neutral carrier phrase, 'I said a ___ again', and recorded 10 times for each speaker. Also, the target vowel in each recording is hand-labeled under careful inspection. There are 17500 microphone recordings, which are sampled at 48 kHz and quantized at 16 bits/sample, with their corresponding accelerometer waveforms. In addition, the corpus includes self-reported height, date of birth, and gender for each speaker as a reference. Some detailed analysis of the distributions of the database can be found in [**3**] and [**25**].

Table 1: The list of target monophthongs, diphthongs and approximant comprised in WashU-UCLA corpus for both sessions 1 and 2.

	Session 1	Session 2
Monophthongs	/i/, /ɛ/, /ɑ/, /u/	/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /o/, /ʊ/, /u/
Diphthongs	/aɪ/, /aʊ/, /ɔɪ/	/e/, /aɪ/, /aʊ/, /ɔɪ/
Approximant		/ɹ/

For the study presented in Chapter 2, a subset of 20 speakers is selected. Table 2 shows the gender balanced list of selected speakers for all the analysis in this chapter. For each selected speaker, 20 tokens of the diphthong /ɔɪ/ from session one of the corpus are used for analyzing frequency jumps and amplitude attenuations because F2 always rises from low to high by crossing Sg2. The total number of tokens, which is 400, is comparably larger than 60 tokens used in previous studies. Besides, multiple Sg2 values, between 10 to 30 tokens, from each speaker are measured from monophthong and approximant recordings to obtain the actual Sg2. For voice quality parameter estimations, such as H1-H2, monophthong recordings with four specific vowels - /ɛ/, /æ/, /ɑ/, and /ʌ/ from both sessions are used. The four vowels are commonly used for voice quality parameter studies, such as [**26**], because their first few harmonics and formants are relatively well separated. In order to obtain reliable statistical representations of speaker-wise voice quality parameters, 90 tokens from different combinations of the four vowels for each speaker are processed.

Table 2: The list of 20 speaker IDs selected from WashU-UCLA corpus by genders. The 10 male and 10 female speakers are selected by preliminary inspection of the quality of the spectrograms.

Male Speaker ID	11	12	15	22	38	41	43	52	53	64
Female Speaker ID	14	16	19	24	27	32	33	36	40	59

2.2 Measurement Methods

This section explains the two measuring methods used in this chapter. Since one of the motivations of this study is to investigate the coupling effect with more sophisticated and precise procedures in a large database, efforts have been made to research and develop software toolkits to analyze SGRs. Two new tools were developed and one existing tool was applied for measuring the data.

The first tool is implemented in MATLAB according to the procedure described in both [**3**] and [**5**] for measuring SGRs. Since SGR measurements are obtained based on visual inspection of the spectral characteristics of the accelerometer waveform by adjusting many parameters, it is important to have an integrated tool for measuring SGRs from a large database.

Table 3: List of parameters used in Snack toolkit for automatic pitch (F0) extraction.

Parameter	**Value**
Window Type	Hamming
Window Length	7.5 ms
Window Shift	5 ms
Method	ESPS
Maximum Pitch	400 Hz
Minimum Pitch	60 Hz

The Subglottal Resonance Measuring Tool (SRMT), as shown in Figure 7, is designed exclusively for the WashU-UCLA corpus to easily select any specific file from the pool of 17500 recordings and the corresponding accelerometer files. A user can choose a sampling frequency between 6 and 10 kHz. After applying the pre-emphasis filter with a coefficient of 0.97, a segment with the length of an adjustable multiple of the pitch period is extracted from the labeled steady state in order to achieve better frequency resolution. The pitch is automatically extracted by using the Snack toolkit [**23**] with the default parameters listed in Table 3. In order to visualize the signals, three spectral representations, including the discrete Fourier transform (DFT) spectrum, the linear predictive coding (LPC) spectrum, and the estimated wideband power spectral density (WPSD), are computed from the segment. Figure 8 shows the spectra of a sample accelerometer recording analyzed using SRMT. The estimated WPSD is described qualitatively as the envelope of the DFT spectrum and the procedures in [**27**] is implemented. The WPSD subdivides the segment into overlapping frames with adjustable overlapping percentage and frame size as multiples of the detected pitch period. The overlapping percentage is commonly set to 80% while the frame size ranges from 0.9 to 1.1 times the detected pitch period. A Hamming window is applied to each subdivided frame. The WPSD eventually outputs the DFT of the autocorrelation averaged over all frames. On the other hand, the LPC fitting quality is mainly controlled by tuning the LPC order which usually ranges from 10 to 18 depending on the sampling frequency. At the same time, the corresponding microphone signal is processed with the same procedures as the accelerometer signal and presented in another tab of the toolkit for reference purposes. Given all three spectral representations in accelerometer plots, SGRs are measured by choosing either the LPC spectral peak or WPSD spectral peak by visual inspection of their fitting to the spectral envelope, and the selected values can be saved for

further post processing. Although not all Sg2 are measureable for every recording, a fair number of tokens from monophthongs and the approximant /ɹ/, ranging from 10 to 30 per speaker, are obtained to calculate the speaker-wise Sg2. As shown in previous studies, the SGRs are fairly constant for a given speaker; hence, the speaker-wise Sg2 is calculated by averaging the measurements obtained from the specific speaker.

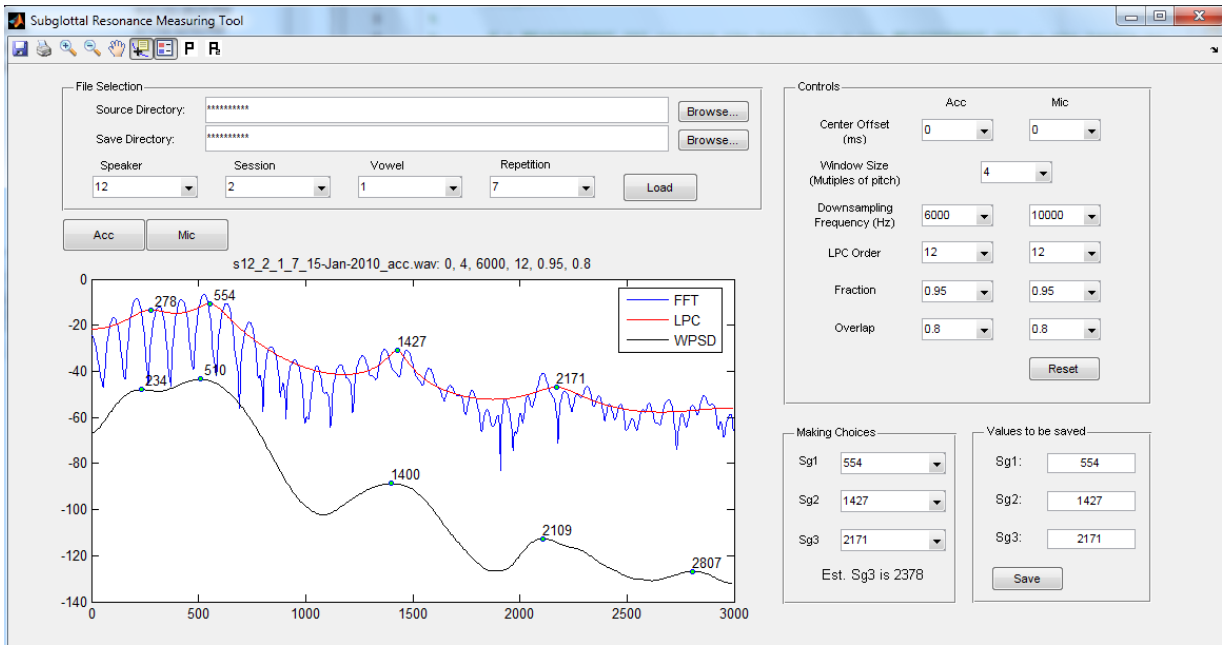

Figure 7: A screenshot of the Subglottal Resonance Measurement Tool.

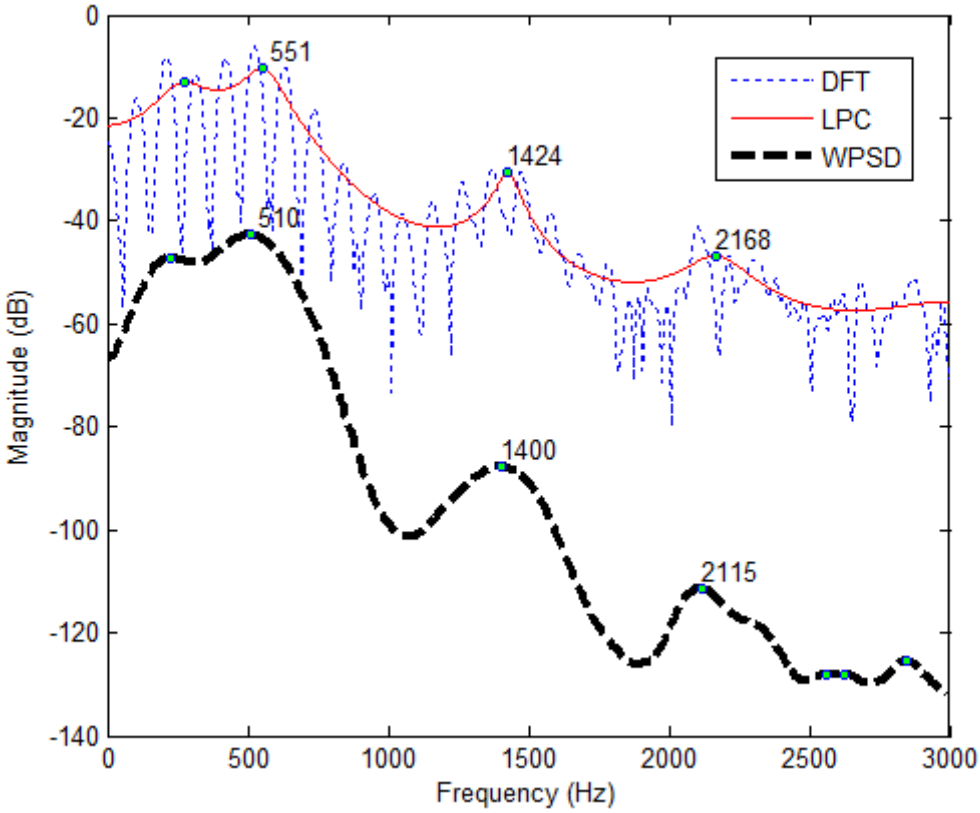

Figure 8: DFT, LPC, and WPSD spectra of the 7th accelerometer recording of the vowel /i/ from male speaker 12 in the WashU-UCLA corpus. The peaks (candidates of SGRs) are automatically marked on both LPC and WPSD spectra.

The second MATLAB implemented toolkit, the Discontinuity Tracking Tool, is used for visualization of formant discontinuities. As shown in Figure 9, it employs the same file selection features for the WashU-UCLA corpus as in the previously presented SGR measuring tool. Since the tool gives the formant track by collecting manual measurements for each frame, the target steady state portions of both microphone and accelerometer recordings are first extracted according to the corresponding label. These extracted signals are then processed in two separate stages.

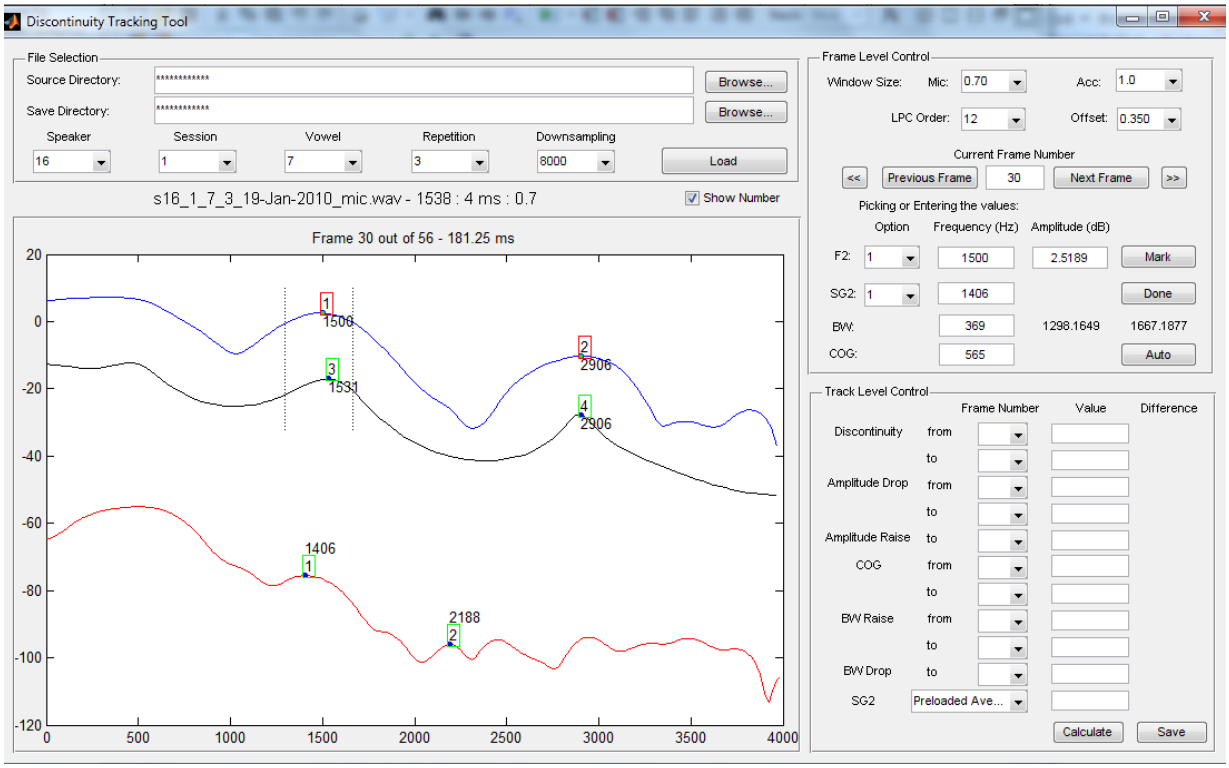

Figure 9: A screenshot of the Discontinuity Tracking Tool with the frame measuring stage.

During the anlysis, both microphone and accelerometer signals are first downsampled to either 8 or 10 kHz, depending on the high frequency noise level. In order to avoid influencing the amplitude of the microphone signal, the pre-emphasis filter with coefficient of 0.97 is only applied to the accelerometer signal. As demonstrated by previous studies, greater coupling can be measured when the glottis is more open in contrast with the close phase of the glottal cycle. A preliminary experiment with results similar to those in [**12**] shown that open-phase measurements have more prevalent frequency jumps and amplitude attenuation. Hence, open-phase measuring procedures are implemented in the toolkit and used. In order to position the analysis window more systematically, the open phase of a glottal cycle is estimated by detecting the glottal closure instant (GCI) with an epoch extraction algorithm from [**28**]. The algorithm

filters the speech signal with a zero frequency resonator in three major steps. Firstly, it removes low frequency bias by differencing the speech signal s[n] as,

$$x[n] = s[n] - s[n-1]\,. \tag{3}$$

The differenced signal from Eq. (3) is passed twice through a zero frequency resonator which is defined as,

$$y_i[n] = -\sum_{k=1}^{2} a_k y_i[n-k] + x[n]\,, \tag{4}$$

where a_1 = -2, a_2 = 1, and i = 1,2 for two stages of the cascaded filter. Finally, the unstable variation in $y_2[n]$ is removed by subtracting the averaged epoch,

$$y[n] = y_2[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_2[n+m], \tag{5}$$

where 2N + 1 is the number of samples in the averaging interval. The positive zero crossings of the zero frequency filtered signal, y[n], are the detected epochs. For cross verification, the DYPSA algorithm [**29**] with the implementation from [**30**] is used as a first pass reference track. After detecting GCIs, the size of the analyzing window is determined by an adjustable factor, ranging from 0.4 to 0.9 for microphone signal and 1 to 3 for accelerometer signal, of the averaged length between two GCIs. At the same time, both microphone and accelerometer signals are divided into frames with a Hamming window centered half of the window size before each GCI. In other words, the window approximately covers only the open phase of a glottal cycle. However, the proportion of the open and close phases varies from person to person, depending partly on the voice breathiness of the speaker; therefore, another option of manually

adjusting the center of the window by visual inspection of the spectra is implemented, and the measurements for both options are presented later in this chapter. Both the LPC spectrum of the microphone signal and the DFT spectra of both signals are displayed. After inspecting and selecting the best fitting F2 and Sg2 peaks from each frame, all parameters, including the Sg2, F2, A2, second formant bandwidth (B2), and $\mathbb{C}$, are saved for further processing.

By collecting data from consecutive frames (usually 10 to 15 frames), both F2 and A2 tracks are displayed in the discontinuity tracking stage. Since the frame-wise Sg2 measurements from the first stage are always unstable and speaker-wise Sg2 values are sometimes used as a stable alternative estimate. By inspecting both F2 and A2 tracks visually, parameters, including frequency jump, amplitude drop, amplitude rise, and Sg2 value are measured for analysis. As shown in Figure 10, an F2 jump of 219 Hz is observed with an A2 drop of 16.15 dB and a rise of 13.47 dB for this particular utterance. This example also demonstrates two commonly observed phenomena: A2 attenuation extends a few frames, and F2 shifts up after crossing Sg2. Although most of the F2 discontinuities are observable after manually adjusting some parameters, some measurements are subjective and some special cases will be discussed in the next section.

The last toolkit used in this chapter, VoiceSauce [**31**], is well-known toolkit for automatic voice quality parameter measurement developed at UCLA SPAPL. It is implemented in MATLAB and measures several voice quality parameters from a given speech signal. These parameters include (but not limited to) H1-H2, H1-A1 (the difference between the first harmonic and the first formant prominence), and H1-A3 (the difference between the first harmonic and the third formant prominence). In addition, the toolkit has a desirable feature which applies the magnitude correction formula from [**32**] to remove the influence of vocal tract resonances for improving the correlation between the voice quality parameters and the actual voice quality of a

given speaker. As the toolkit is designed for voiced and steady vowels, only the voice quality parameters from the steady states of the target vowels are measured here.

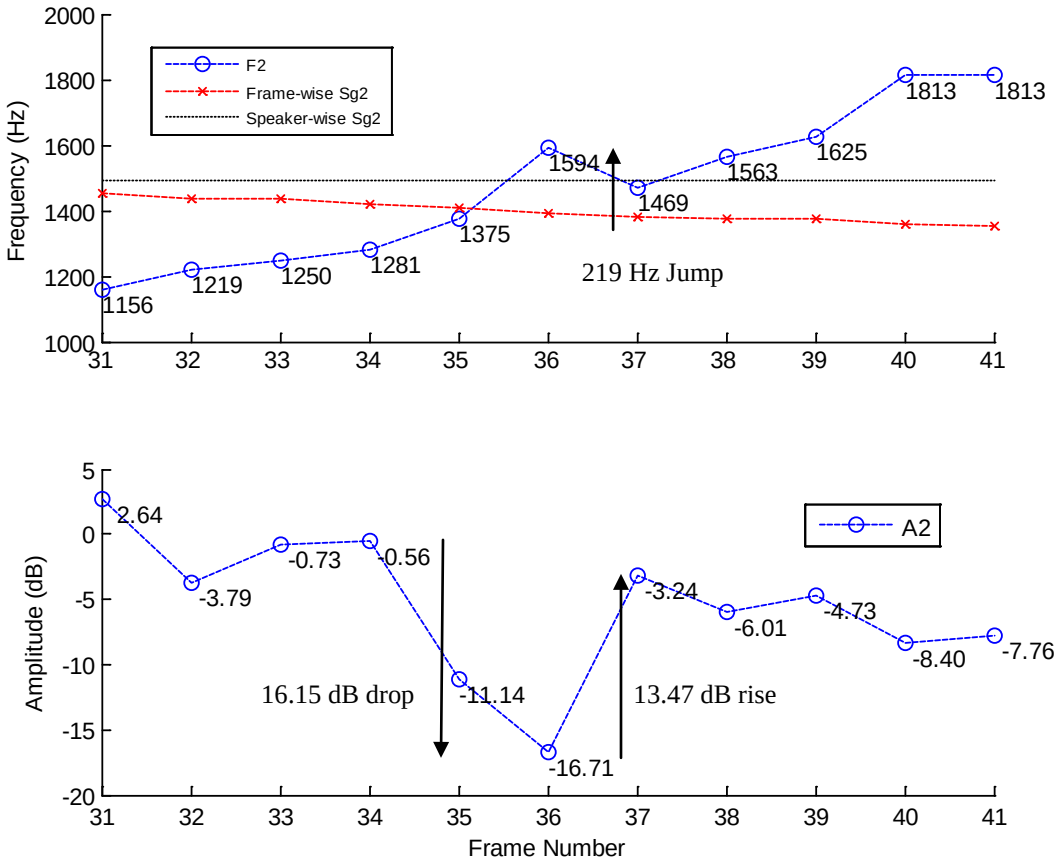

Figure 10: F2 and A2 tracks of the vowel /ɔɪ/ from female speaker 24. The measured frame-wise Sg2 is very close to the speaker-wise Sg2 of 1491 Hz. A F2 jump of 219 Hz is observed between frame number 35 and 36 with the corresponding A2 drops 16.15 dB from frame 35 to 36 and rises 13.47 dB from frame 36 to 37.

2.3 Analysis and Discussion

Although the statistics of SGRs for all 50 speakers in the WashU-UCLA corpus have been reported in both[**3**] and [**5**], the study in this chapter uses only a subset of the corpus, and the statistics of measured Sg2 from the 20 chosen speakers are presented in Appendix A. By comparing with the Sg2 statistics of all 50 speakers in the corpus, intra-speaker means and standard deviations for the chosen speakers are evenly distributed, and their inter-speaker averages are very close to the full corpus values. Consequently, this subset can be claimed as having good Sg2 representativeness of the whole corpus while being relatively larger than the data used in previous studies.

As mentioned in the previous section, the measurements of F2 jump and A2 attenuation are subjective and sometimes vary with the choice of window size and window position. In some tokens, one or both of the jump and attenuation are non-measureable with any combination of controlling parameters. In order to cross verify the sensitivity of window size and position to the measurements, two setups, denoted by S1 and S2, have been used and the numbers of measurable tokens with different measuring parameters are listed in Table 4. The measurements in S1 are determined by using both a fixed window size controlling factor of 0.7 and a fixed window positioning factor of 0.35 to process every token. On the other hand, S2 applies both varying window size and window centering position in the interest of getting the best fitting spectral peaks. Although the percentages of getting measurable F2 jump is higher in S1 than that in S2, it can be explained by the order of the procedures that the subjective measurements on S2 are determined with more conservative considerations after processing S1. By taking out the tokens without A2 attenuation from the measureable F2 jump list, the percentages of getting both the F2 jump and A2 attenuation in a given token for both setups are roughly the same. Although

A2 attenuation is claimed as a more robust cue for the coupling effect by previous studies, uncertainties have been found during the measuring process in the current study. There is a common observable phenomenon that A2 attenuation can occur across a couple of frames without aligning the center with the F2 jump, and there is lack of evidence to prove the association between the two. The possibility of such a wide spread and shifted A2 attenuation due to other articulatory factors, such as interdental spaces, cannot be eliminated. In order to improve the credibility of the measurements, A2 attenuations with such uncertainty are ignored, and this reduces the robustness of A2 attenuation as a cue for detecting the coupling effect in the current study.

Table 4: The settings and numbers of tokens for the two measuring setups used in Chapter 2 are presented. The table shows the number of speakers (Speakers), the window size controlling factor (Size), the window positioning factor (Pos.), the number of target words (Words), the total number of tokens used for measurement (Total Tokens), the number of measureable F2 jump tokens (F2 Tokens), and the number of measureable F2 jump and A2 attenuation tokens (F2-A2 Tokens) for each setup. The percentages in both F2 Tokens and F2-A2 Tokens represent the ratio of the number of measureable tokens over the corresponding total number of tokens.

Setup	Speakers	Size	Pos.	Words	Total Tokens	F2 Tokens; %	F2-A2 Tokens; %
S1	20	0.7	0.35	2	400	385; 96%	321; 80%
S2	20	Vary	Vary	1	200	162; 81%	155; 78%

There are two major issues suggested by previous studies that are probably corresponding to some non-measureable tokens in the current study. As suggested by [**9**], there might be other factors, such as the pole-zero pairs from the interdental spaces, interacting with formants and SGRs in the coupling effect. The discontinuities of F2 caused by the interdental spaces occur in a

wide range of frequencies, and the coupling effect from SGR is predicted to be stronger than from interdental spaces during the open phase of the glottal cycle [**33**]. However, the possibility of other factors influencing the measurement of F2 discontinuity around Sg2 frequency cannot be eliminated, and further investigation of this issue is necessary. Another issue is mentioned in [**12**] that the small analyzing window, which enables capturing the discontinuity, results in poor frequency resolution. From Rayleigh's criterion, an analysis window can only resolve two peaks that are apart from each other by at least half of main lobe width of the window. For instance, a 5 millisecond Hamming window is used with a window size controlling factor of 0.7 to measure male speech with F0 of 140 Hz. Hence, the corresponding main lobe width is around 800 Hz with sampling frequency of 10 kHz, and this window cannot resolve peaks that are less than 400 Hz apart. When F2 crosses Sg2, their poles can easily get closer than a few hundred hertz where a small window fails to resolve their peaks by giving a single smoothed peak between the two poles in the spectrum. Given such trade-off between time and frequency resolutions, the discontinuity measurements in this study can only be claimed best possible estimates.

Due to the uncertainties from the measuring process, only tokens with measurable both F2 jump and A2 attenuation, which account for 78% to 80% of the total number of tokens in the two setups as shown in Table 4, are used for analysis. The averaged F2 jumps and A2 attenuations for all speakers with different setups are listed in Table 5 toTable 8. The four tables show the statistics of the discontinuity measurements separated by gender (from the 20 speakers in S1 and S2). Each table presents statistics of the measurements, including the minimum F2 jump (Min.), the maximum F2 jump (Max.), the averaged F2 jump (F2 Avg.), the averaged A2 drop (Drop), the averaged A2 rise (Rise), and the averaged mean between drop and rise (A2 Avg.), from all tokens with both measureable F2 jump and A2 attenuation. Moreover, the last

two columns of each table give the correlation coefficient (Corr.) and the corresponding correlation significance (p-value) between all pairs of F2 jumps and A2 attenuations for each speaker. Furthermore, the last row of each table shows the correlation coefficient and the correlation significance of all the paired measurements from a given gender in the corresponding setup. Although a previous study, [**12**], showed a weak negative correlation between the size of F2 jump and the corresponding A2 attenuation for four of the six speakers in the study, a discrepancy is found in the current study. A positive or sometimes weak positive correlation between the size of F2 jump and A2 attenuation is found in majority of the speakers, including 16 out of 20 speakers in S1 and 13 out of 20 speakers in S2. Nevertheless, the speaker-level correlations in both the previous and current studies are not always statistically significant. When all the measureable tokens are combined together, weak positive correlations of 0.34 and 0.23 for S1 and S2, respectively, are found with p-value less than 0.005. Although a trend of positive correlation is observed with the collected data, a prominent conclusion cannot be made due to inconsistent results in some speakers. Given the dependency between A2 attenuation and the time F2 passes the Sg2 introduced zero-pole pair, one might suspect that the time resolution for the current setup is not sufficient in certain tokens to capture the exact amplitude changes. Nevertheless, the trade-off between time and frequency resolution creates obstacles for the current study to proceed further in this regard.

Table 5: The statistics of the discontinuity measurements for the 10 male speakers in S1. There are 160 tokens with both measureable F2 jump and A2 attenuation.

Males	F2 Jump (Hz)			A2 Attenuation Average (dB)			Statistics	
ID	Min.	Max.	F2 Avg.	Drop	Rise	A2 Avg.	Corr.	P-value
11	94	266	135	4.9	7.6	6.2	**0.5672**	0.01410
12	94	344	165	10.6	2.5	6.6	**0.1546**	0.64987
15	125	312	183	7.3	4.7	6.0	-0.0320	0.90295
22	125	235	169	6.8	4.5	5.6	**0.6195**	0.01377
38	93	250	151	7.2	6.1	6.6	**0.5317**	0.03401
41	109	218	138	4.7	2.8	3.7	**0.0844**	0.77417
43	110	344	184	8.8	5.7	7.2	**0.2951**	0.20655
52	140	390	220	9.2	4.7	6.9	-0.5119	0.03566
53	132	297	172	6.9	4.3	5.6	**0.5931**	0.00948
64	141	563	303	12.4	7.7	10.1	**0.4074**	0.14826
Over **160** tokens from 10 male speakers							0.3791	0.00001

Table 6: The statistics of the discontinuity measurements for the 10 female speakers in S1. There are 161 tokens with both measureable F2 jump and A2 attenuation.

Females	F2 Jump (Hz)			A2 Attenuation Average (dB)			Statistics	
ID	Min.	Max.	F2 Avg.	Drop	Rise	A2 Avg.	Corr.	P-value
14	63	375	157	7.3	4.0	5.6	**0.5372**	0.01771
16	62	250	131	6.2	3.9	5.1	**0.6819**	0.00362
19	187	437	226	10.9	4.9	7.9	**0.1075**	0.67105
24	94	344	205	11.4	8.4	9.9	-0.0365	0.90147
27	125	234	181	4.0	4.1	4.0	-0.0784	0.75702
32	125	312	243	13.1	4.8	9.0	**0.1775**	0.52674
33	94	281	149	5.4	4.8	5.1	**0.2860**	0.24992
36	110	187	142	12.4	6.5	9.5	**0.6666**	0.01791
40	125	313	209	8.3	3.4	5.8	**0.6783**	0.00388
59	125	219	184	8.0	5.5	6.7	**0.1447**	0.60690
Over **161** tokens from 10 female speakers							0.3135	0.00005

Table 7: The statistics of the discontinuity measurements for the 10 male speakers in S2. There are 80 tokens with both measureable F2 jump and A2 attenuation.

Males	F2 Jump (Hz)			A2 Attenuation Average (dB)			Statistics	
ID	Min.	Max.	F2 Avg.	Drop	Rise	A2 Avg.	Corr.	P-value
11	125	438	257	11.5	9.1	10.3	**0.7882**	0.03525
12	125	250	198	17.7	6.5	12.1	**0.4388**	0.38400
15	78	422	272	12.5	7.9	10.2	**0.0586**	0.90067
22	140	453	297	7.3	3.1	5.2	**0.6803**	0.09261
38	141	266	200	10.2	7.1	8.6	-0.2170	0.54705
41	141	297	206	8.3	6.8	7.5	**0.3876**	0.44774
43	187	328	233	12.3	7.1	9.7	**0.0281**	0.94726
52	187	391	267	7.8	4.5	6.2	**0.2126**	0.58292
53	171	297	207	10.4	4.8	7.6	-0.2624	0.56968
64	219	500	373	12.0	6.2	9.1	-0.1473	0.72776
Over **80** tokens from 10 male speakers							0.1791	0.12420

Table 8: The statistics of the discontinuity measurements for the 10 female speakers in S2. There are 75 tokens with both measureable F2 jump and A2 attenuation.

Females	F2 Jump (Hz)			A2 Attenuation Average (dB)			Statistics	
ID	Min.	Max.	F2 Avg.	Drop	Rise	A2 Avg.	Corr.	P-value
14	125	281	192	7.5	3.1	5.3	-0.0241	0.95913
16	125	250	179	9.4	4.2	6.8	-0.2213	0.63351
19	218	469	313	12.3	7.8	10.1	**0.0634**	0.86182
24	156	375	263	12.1	9.0	10.5	**0.3831**	0.39632
27	156	297	222	5.1	3.8	4.5	**0.1567**	0.66557
32	219	344	281	12.2	5.1	8.7	-0.0002	0.99957
33	156	344	228	5.6	5.0	5.3	-0.1418	0.76167
36	187	375	228	12.7	4.5	8.6	**0.2110**	0.64971
40	187	281	250	8.3	3.3	5.8	**0.1877**	0.72179
59	187	281	226	8.5	7.2	7.8	**0.2598**	0.49956
Over **75** tokens from 10 female speakers							0.30646	0.00570

2.3.1 Correlation with H1-H2

Another focus of the study in this chapter is the investigation of the correlation between voice quality parameters and the Sg2 coupling effect. It is natural to study such correlation due to

the dependency between oral-subglottal coupling and glottal area. From the physiological point of view [**34**], a speaker with breathier voice has a larger open quotient (OQ), which is the proportion of a glottal cycle during which the glottis is open. One would expect to observe greater coupling from a breathier speaker, and the OQ is highly correlated to the amplitude of the first harmonic relative to that of the second (H1-H2) [**26**]. In addition to the investigation of the correlation between H1-H2 and subglottal coupling, two more voice quality parameters are explored, including the amplitude of the first harmonic relative to that of the first-formant prominence (H1-A1) and the amplitude of the first harmonic relative to that of the third-formant spectral peak (H1-A3). Both parameters are commonly used for voice quality measurement where H1-A1 is correlated with the presence of a posterior glottal chink, and H1-A3 reflects the source spectral tilt. However, the results in both H1-A1 and H1-A3 cases are inconclusive, and their analysis can be found in Appendix B.

All voice quality parameters are measured from a large number of tokens for each speaker by using VoiceSause [**31**]. Each token generates 3 voice quality parameters by averaging the multi-point measurements over time. After eliminating the outlier tokens by inspecting their F0 tracking contours, the three speaker-wise voice quality parameters are calculated by averaging over all tokens for a given speaker. In order to validate the credibility of the voice quality parameters, three subsets of data, V1, V2, and V3 from a given speaker are used separately to generate the speaker-wise parameters. In V1, the speaker-wise parameters are generated by averaging multi-point measurements from 90 tokens of the vowels /æ/, /ɑ/, and /ʌ/ for a given speaker while V2 uses the same number of tokens from the vowels /ɛ/, /æ/, and /ɑ/. Although the voice quality parameter measuring toolkit is designed to give meaningful results with steady monophthong vowels only, the last subset of data, V3, which contains the extracted

steady-state portion from the same diphthong tokens, /ɔɪ/, as in discontinuity measurements, is also processed by the toolkit for exploratory experiments. The correlation coefficients between the size of the F2 jump and different voice quality parameters with the three subsets of data are presented in Table 9 and Table 10. The tables show the inter-speaker correlation coefficients between the size of F2 jump and H1-H2. The (*) in Table 9 indicates the use of harmonic magnitude correction formula to reduce the influence of vocal tract resonances [**32**]. The correlation coefficients with different setups and voice quality parameters are relatively consistent across V1 and V2 regardless of the use of harmonic magnitude correction formula. However, small disagreements shown in V3 as expected due to the unreliable tracking algorithm from the voice quality measuring tool on diphthongs. Nevertheless, the three subsets of data provide a cross validating framework to explore the relationships between the size of jump and H1-H2.

Table 9: Correlation coefficients between the size of F2 jump and H1-H2 with correction formula to reduce the influence of vocal tract resonances.

		S1 - F2 Jump		S2 - F2 Jump	
		Male	Female	Male	Female
V1	H1*-H2*	0.52	0.47	0.43	0.47
V2	H1*-H2*	0.51	0.42	0.45	0.37
V3	H1*-H2*	0.40	0.57	0.41	0.32

Table 10: Correlation coefficients between the size of F2 jump and H1-H2 without using any correction formula.

		S1 - F2 Jump		S2 - F2 Jump	
		Male	Female	Male	Female
V1	H1-H2	0.54	0.41	0.50	0.40
V2	H1-H2	0.49	0.38	0.40	0.34
V3	H1-H2	0.34	0.18	0.30	0.10

Based on the observation that the measured F2 jumps for female have higher mean and larger variation than those for male speakers as shown in Figure 11, results in this chapter are reported with respect to gender. This observation confirms the hypothesis that the size of the frequency jump is correlated with the breathiness of a given speaker because studies, such as [**26**] and [**35**], have shown that females are generally breathier than males with a higher mean and variance of voice quality parameters. Moreover, both V1 and V2 confirm the hypothesis and previous study [**16**] that positive correlations, ranged from 0.40 to 0.54 for males and 0.37 to 0.47 for females, can be found between H1-H2 and the size of the jump across S1 and S2. In spite of the slightly weaker correlation results and unstable parameter measurements from V3, results consistently show positive correlations. The numerical values of these correlation coefficients may change with more precise measurement methods for both voice parameters and F2 jump, but the consistently positive correlation trend provides a possible evidence to support the hypothesis that the frequency discontinuity is mainly contributed by the oral-subglottal coupling through the glottis. To further study the interactions between oral-subglottal coupling and articulatory coupling, such as the interdental space, a corpus with both accelerometer recording and magnetic resonance imaging (MRI), for subglottal resonances and articulatory features, respectively, with synchronized microphone speech signals is necessary.

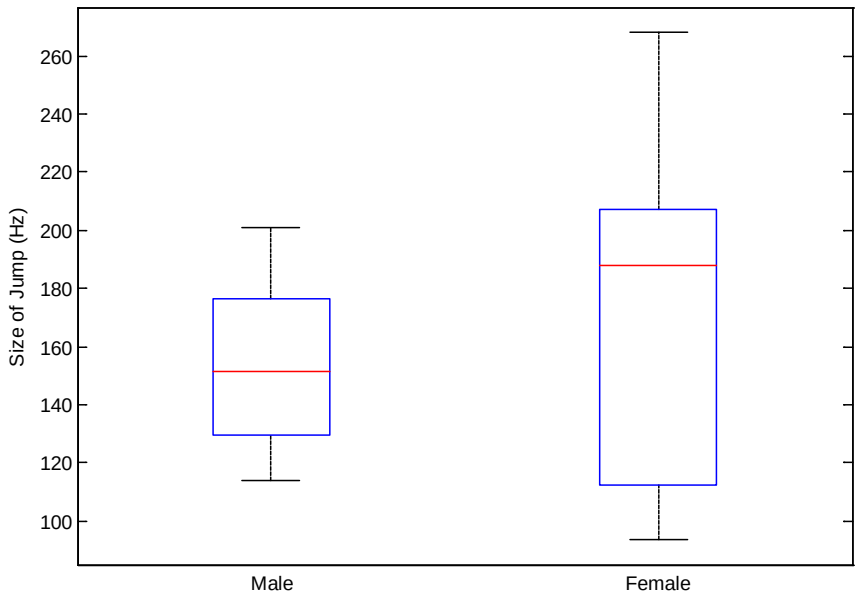

Figure 11: The boxplot of the size of F2 jump from tokens with both frequency jump and amplitude attenuation in S1 by gender. The bottom and top of the box are the 25th and 75th percentile, respectively, and the band near the middle of the box is the median.

2.4 Summary

In this chapter, we revisited and extended an investigation on the SGR coupling effect. The current study does not only use a larger corpus than previous studies, but also defines more accurate procedures for measuring the SGR coupling effect. Two new tools are developed for data measurement. All discontinuity measurements are determined by using open-phase procedures due to the prevalent F2 jump and A2 attenuation observed in a preliminary experiment and previous studies. The results can be grouped in two parts.

The first part is the correlation between the size of the F2 jump and A2 attenuation. Although the current study demonstrates a trend of positive correlation between the two which is

in contrast with previous studies, there is lack of evidence to support a strong conclusion. Additional synchronized MRI data and better time-frequency analysis methods could be helpful to solve this problem. Preliminary experiments with wavelet and Wigner distribution analysis methods are explored but without success, and further investigation is necessary. The second part of the study analyzes the relationships between H1-H2 and the size of F2 jump with validation from different measurement setups. All results in this part confirm the positive correlation between the two measures, which is also consistent with a previous study. This correlation provides evidence to support the claim of contribution to formant discontinuity from subglottal cavity. However, a more concrete conclusion cannot be made unless, again, a speech corpus with both MRI and accelerometer synchronized signals is available.

All in all, the study in this chapter reinforces the previously proposed model of oral-subglottal coupling effect and develops specific tools for future studies.

CHAPTER 3. SPEAKER IDENTIFICATION USING SUBGLOTTAL RESONANCES

3.1 Introduction to Speaker Identification Systems

Since ancient time, humans have been looking for ways to confirm the identities of each other for different purposes. In modern times, almost everyone uses some form of authentication methods such as passwords and signatures, but these methods are relatively easy to be stolen, forged, or forgotten. As a result, biometric recognition systems by using cues, including fingerprint, voice, face, retina, DNA and so on, come into play with various advantages as discussed in [**36**]. Among all these measures, voice is the easiest one to capture by machines because only a microphone is required. Moreover, it contains both physical and behavioral characteristics of each individual speaker. The physical characteristics come from the inter-speaker variations in sizes and constructions of speech production organs, such as vocal tract, larynx, and nasal cavity. On the other hand, speaking rhythm, intonation, and accent are considered as behavioral characteristics. With these properties, researchers have been continuously working on both speaker identification and speaker verification systems. Some classical and recent techniques for these systems have been presented in several tutorials and survey publications, including but not limited to [**37**], [**38**], [**39**], and [**40**]. Although there are different techniques and applications associated with identification and verification systems, both of them require discriminative features from speech signals. As mentioned in previous chapters,

SGRs have relatively small within-speaker variability. These criteria make SGRs ideal candidates for discriminating speakers using speech, but an easy and reliable measurement or estimation method, which will be presented in next section, is necessary. The goal of this chapter is to explore the possible role of SGRs in speaker identification.

In a speaker identification system, as shown in Figure 12, an individual is identified from a known pool of people. Generally speaking, there are two stages, training and identification. In the training stage, selected features are extracted from enrolling utterances for all speakers in the identifying pool to build corresponding models for each speaker individually. This stage is usually performed under the supervision of a professional operator and can be carried out offline before deploying the system. With all trained models for each speaker, the system progresses to the identifying stage. In the second stage, the same selected features are extracted from speech of an unknown incoming speaker, and the speaker is identified by comparing the features against all trained models individually. If the unknown speaker is restricted to be one of the trained speakers, the system is considered as close-set identification; otherwise, it is an open-set system, which is not the focus of this study. In addition, speaker identification systems can be divided into the text-dependent and text-independent types. Although text-dependent systems usually perform better, the current study is restricted to text-independent type for exploratory purposes. In order to evaluate the performance of any speaker identification system, identification rate, which is the percentage of correctly identified speakers over the total number of speakers to be identified, is used in the current study; however, the complementary evaluation metric, namely identification error rate, is used in some other studies [**17**].

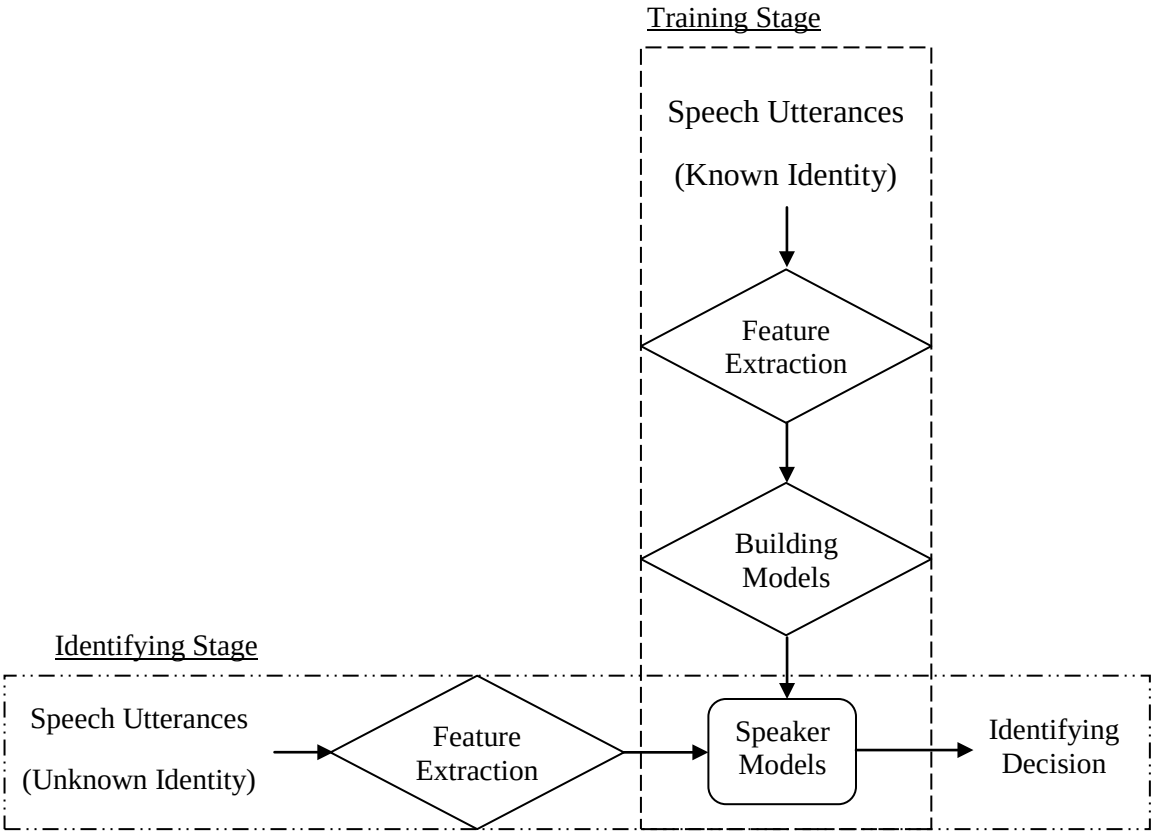

Figure 12: Block diagram of a speaker identification system.

Over the last few decades, many features, such as short-term spectral features and prosodic features, have been studied for speaker identification. Among all these features, Mel-frequency cepstral coefficients (MFCCs) [**41**] are the most popular due to their close approximation to human auditory system and robust performance on different speech processing tasks, including speech recognition and speaker recognition. Although MFCCs achieve good performance on speaker identification tasks [**42**] , [**43**], speaker information is implicitly embodied in the feature. Therefore, the current study explores the effect of adding additional features with explicit speaker information embodied, such as SGRs, in addition to the MFCCs for speaker identification.

In order to compute MFCCs, the Fast Fourier Transform (FFT) magnitude spectrum, which is composed using a 512-point FFT, is first calculated for each frame of the input utterance. The computed spectra are processed by a Mel-frequency filterbank. The following formula relates linear frequency f (Hz) and Mel frequency ν:

$$\nu = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right), \tag{7}$$

The last step in computing MFCCs is applying discrete cosine transform to the log magnitude of the Mel-frequency filtered signal as:

$$\mathrm{mfcc_m}[n] = \frac{1}{R}\sum_{r=1}^{R} \log S_m[r] \times \cos\left[\frac{2\pi}{R}\left(r + \frac{1}{2}\right) n\right]. \tag{8}$$

In Eq. 8, $S_m[r]$ is the magnitude of the mth frame and r^{th} mel-filterbank output. The MFCCs are computed for n = 1, 2 … N where N is the number of cepstral coefficients to be retained and R is the number of filters, and they are set to 13 and 26, respectively, in the current study. The zeroth coefficient is normally excluded in the 13-dimensional feature vector because it represents the averaged logarithmic power which has minimal speaker information [**40**]. Finally, the first and second order temporal derivatives of the feature vector, which are commonly called the delta and delta-delta features, are usually attached to the computed MFCCs in order to capture dynamic information. This makes the final dimension of the feature vector 39 for each frame (13 MFCCs, deltas, and delta-deltas).

After extracting the features from available training utterances of a given speaker, a model representing the speaker, regardless of the speech content, has to be constructed. This study uses the most commonly used paradigm for text-independent speaker identification,

namely the Gaussian Mixture Model (GMM). It was first demonstrated by [**42**] and [**43**] that GMMs can be used for speaker identification, and many studies have investigated the extensions of the paradigm since then. Using the same notation as in [**42**], the Gaussian mixture density, denoted by λ, is a weighted sum of M component densities for a given speaker as described in Figure 13. It is formulated by,

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) , \tag{9}$$

where $\vec{x}$ is the feature matrix with each column representing the D-dimensional feature vector extracted from each frame. The p_i and b_i for i = 1, 2 … M are the mixture weights and component densities, respectively. The ith component density is defined by a D-variate Gaussian function as,

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \sum_i^{-1}(\vec{x} - \vec{\mu}_i)\right\}, \tag{10}$$

while $\vec{\mu}_i$ and $\sum_i$ are mean vector and covariance matrix, respectively. Moreover, the mixture weights of all M component densities should sum up to unity. A common notation to denote the mean vectors, covariance matrices and mixture weights from all component densities is,

$$\lambda = \left\{p_i, \vec{\mu}_i, \sum_i\right\}, i = 1, 2, \dots, M . \tag{11}$$

There is an important reason for the popularity of GMMs in speaker identification, and that is the expectation-maximization (EM) algorithm [**44**]. The algorithm is basically an iterative procedure to estimate an optimal set of parameters, but it guarantees monotonic convergence to the parameters in a small number of iterations. In addition, there are a few more reasons behind the

success of GMMs in speaker identification, but they are not the main focus of this study, and the interested reader is referred to studies such as [**42**] and [**43**] for more details.

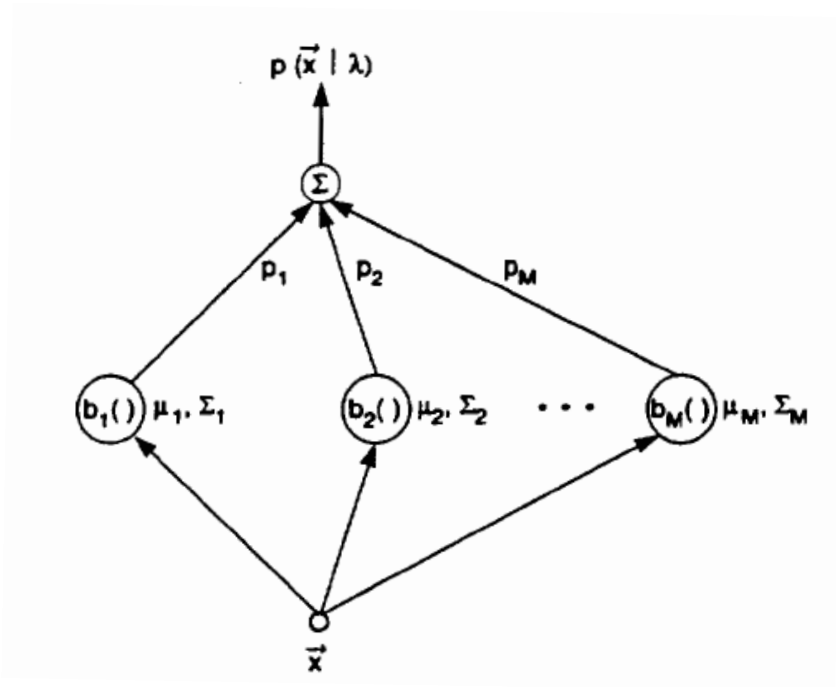

Figure 13: A Gaussian mixture density from M weighted sum of Gaussian densities. The p_i and b_i for i = 1,2 … M are the mixture weights and component densities, respectively (adapted from [**42**]).

After collecting all trained GMM models, $\lambda_1, \lambda_2, \ldots, \lambda_S$, for the pool of S speakers, an unknown speaker can be identified by searching for the model which gives the maximum a-posteriori probability. As presented in [**42**], this can be reduced to a Maximum Log Likelihood problem as,

$$\hat{S} = \arg\max_{1 \leq k \leq S} \sum_{t=1}^{T} \log\, p(\vec{x}_t | \lambda_k)\,, \tag{12}$$

where $\hat{S}$ is the identified speaker and $p(\vec{x}_t | \lambda_k)$ is calculated using Eq. 9.

The basic building blocks of the most popular speaker identification system have been presented. Although there are more advanced versions of this system, this basic but robust backbone system is adequate for the current exploratory study to investigate the role of SGRs in speaker identification. The standard GMM implementation from the Statistics Toolbox of MATLAB is used for constructing the speaker models in the training stage and calculating the negative log-likelihood in the identification stage. In order to proceed further, a fundamental and important issue, which is an easy and reliable estimating algorithm of SGRs from microphone signal, has to be explored. The next section considers this issue and presents a solution.

3.2 Automatic Estimation of the First and Second Subglottal Resonances

As discussed in previous chapters, there are two major approaches for estimating SGRs from microphone signals. The first approach directly estimates SGRs by detecting their interactions with formant frequencies. In particular, existing techniques, including [**10**] and [**11**], estimate Sg2 and third subglottal resonance (Sg3) by detecting frequency jumps and amplitude attenuations in isolated vowels. However, the previous chapter has shown several uncertainties of the coupling effect, such as influence from inter-dental cavity, that make the estimation unstable. Besides, this approach only works for some specific isolated vowels, which is not a good property for text-independent speaker identification or any other speech content independent task. In addition, this direct estimation approach has very limited extendibility

because it is not designed for estimating the first subglottal resonance (Sg1) and it heavily relies on the accuracy of the formant tracking algorithm. As a result, the second approach, which indirectly estimates SGRs from speech signals by using correlations with vowel boundaries, is used in this chapter to extract SGRs for speaker identification. The methods presented in this section contain some portions of co-authored works in both [**1**] and [**5**].

The indirect estimation approach is motivated by the findings from [**6**] which defines vowel boundaries from SGRs. In particular, Sg1 creates a boundary between [+low] and [-low] vowels, and Sg2 acts similarly for [+back] and [-back] vowels. Although [**13**] has proposed a possible boundary between tense and lax [-back] vowels by Sg3, further evaluation is necessary and Sg3 estimation method based on such relationship has not yet been developed. However, a strong correlation is observed between Sg2 and Sg3 for adults so that study in [**5**] estimates Sg3 from Sg2 by a first-order linear regression model with r-squared (r^2) value of 0.8427 as,

$$Sg3 = 1.079 \times Sg2 + 763.676\,. \tag{13}$$

Despite the proven robustness of this estimation model, such estimated Sg3 is not evaluated in the current study of speaker identification because of its strong dependence on the estimation of Sg2.

By following the well-studied boundary definition for [+back] and [-back] vowels, the estimation method of Sg2 is first developed by using the correlation between the vocal tract-based and Sg2-based measures of vowel backness. In [**45**], the Bark difference between F3 (third formant) and F2 (second formant), denoted by B_{32} , is used to separate [+back] and [-back] vowels at a boundary value of 3 Bark. This boundary is regarded as vocal tract based measure

since both F2 and F3 can be measured from microphone signals. The Bark value can be obtained from Hertz by using the formula from [**46**] as:

$$z = \frac{26.81 \times f}{1960 + f} - 0.53, \tag{14}$$

where f is frequency in Hertz and z is the converted Bark value. Based on the motivation of dividing the vowel space into Bark sclae, [-back] and [+back] vowels are shown to be separable by the Bark difference between F2 and Sg2, denoted by $B_{2,s2}$, at roughly 1 Bark. Therefore, $B_{2,s2}$ is proven to be a reliable Sg2-based measure of vowel backness and can be computed from both the microphone and accelerometer signals. Although a regression model with r^2 value of 0.8905 can be built to correlate B_{32} with $B_{2,s2}$, two speaker-related features, F0 and F3, are added to the model to reduce inter-speaker variability. The final regression model presented in [**5**] for estimating Sg2 from adults has r^2 value of 0.9713 is:

$$\begin{aligned} B_{2,s2} = {} & 0.001(B_{32})^3 + 0.009(B_{32})^2 - 1.083(B_{32}) \\ & + 0.002(F3) - 0.007(F0) - 0.019\,. \end{aligned} \tag{15}$$

This is a trained linear regression model, denoted by $M2_a$, with $B_{2,s2}$ as the dependent variable, and B_{32}^3, B_{32}^2, B_{32}, F3, and F0 as the independent variables.

With a similar strategy and the observation of the boundary between [+low] and [-low] vowels formed by Sg1, another regression model is built. Instead of B_{32} and $B_{2,s2}$, the Bark difference between F3 and F1, denoted by B_{31}, and the Bark difference between F1 and Sg1, denoted by $B_{1,s1}$, are used. By the same argument as in the previous model, F0 and F3 are used to compensate for inter-speaker variability. The resulting regression model for estimating Sg1 from adults speech has r^2 value of 0.9724:

$$B_{1,s1} = 0.001(B_{31})^3 - 0.024(B_{31})^2 - 0.737(B_{31}) + 0.002(F3) - 0.007(F0) + 3.903\,. \tag{16}$$

Sg1 from adults speech is estimated by this trained linear regression model, denoted by $M1_a$, with $B_{1,s1}$ as the dependent variable, and B_{31}^3, B_{31}^2, B_{31}, F3, and F0 as the independent variables.

In order to explore the possibility of applying children's SGRs to speaker identification, a similar indirect estimation approach for children from [**1**] is used. For simplicity, the linear regression models for predicting Sg1 and Sg2 for children are denoted by $M1_c$ and $M2_c$, respectively in the current study. Model $M1_c$ is trained with $B_{1,s1}$ as the dependent variable, and B_{10} (the Bark different between F1 and F0) and F3 as independent variables. Similarly, $M2_c$ is trained with $B_{2,s2}$ as the dependent variable, and B_{32}^2 , B_{32} , F3 and F0 as the independent variables. However, children have a large acoustic variability mainly because of their growing vocal tract length. As a result, each of the models $M1_c$ and $M2_c$ is split into two models based on the averaged F3, which is highly correlated with vocal-tract length. The two split models trained with recordings from children who have an average F3 less than 3300 Hz are denoted by $M1_c^l$ and $M2_c^l$, while the other pair of models is denoted by $M1_c^g$ and $M2_c^g$ for children with averaged F3 greater than or equal to 3300 Hz. The boundary value of 3300 Hz is empirically chosen according to the distribution of averaged F3 values from 25 children, and the interested reader is referred to [**1**] for more details.

By using these trained regression models, frame-level Sg1 and Sg2 can be estimated, but speaker-level SGR values are instead used as the final estimation results. It is meaningful to use speaker-level SGR values for different applications because SGRs are roughly constant within a given speaker as discussed before. Moreover, it is practical to estimate SGRs from continuous

speech, but existing automatic formant frequencies and F0 tracking algorithms do not always yield reliable results, especially around the transition regions between voiced and unvoiced sounds. In other words, these frame-level SGR estimates with continuous speech are very sensitive to measurement errors, but this problem can be minimized by averaging frame-level estimates to get more accurate speaker-level SGR values. Although all regression models are trained by steady-state vowels, the speaker-level SGR estimates are evaluated on both isolated vowels and continuous speech to have reasonable high accuracy as presented in both [**1**] and [**5**]. In addition, the resulting estimates have been demonstrated to be helpful for both speaker height estimation in adults and speaker normalization on children's speech. The goal of the remaining sections of this chapter is to explore the role of SGRs in speaker identification by using the presented estimation method for both adults and children.

3.3 Databases Used

In order to have a multidimensional investigation of SGRs in speaker identification, three speech corpuses are used in this chapter, including the WashU-UCLA corpus, the TIMIT database [**47**], and the CID database [**48**].

The WashU-UCLA corpus, as presented in Section 2.1, is used for preliminary experiments because "ground truth" SGR values can be measured from the synchronized accelerometer recordings while SGR estimation method has to be used in other databases. With the "ground truth" SGR values, one can expect to estimate a theoretical upper bound of performance of speaker identification with SGRs as features. However, this corpus is not designed to support speaker identification research. For instance, it has only 50 speakers, which is relatively small compared to other standard speaker identification databases, but the balanced gender distribution enables a fair exploratory investigation on the role of SGRs to speaker identification across gender. Moreover, the corpus is recorded under ideal conditions with the same microphone, and all the 35 recorded words from two sessions of the corpus are embedded in a phonetically neutral carrier phrase. In other words, the corpus contains limited variability in several aspects, such as microphone variability, intersession variability, and phonetic variability, but it is sufficient and helpful for an exploratory study in the current chapter.

In order to generalize the study with a more commonly used database for speaker identification, the TIMIT database is used. The database contains 10 recorded utterances from each of the 630 speakers (438 males and 192 females) in 8 major dialect regions of the United States. The 10 sentences from each speaker include 2 dialect-specific (SA) sentences, 3 phonetically-diverse (SI) sentences and 5 phonetically-compact (SX) sentences. The utterances

are recorded in a clean studio environment with a fixed wideband headset in one session. The recordings are sampled at 16 kHz and quantized at 16 bits/sample. This database has rich phonetic variability with relatively large number of speakers. Although this database is not recommended for evaluating speaker recognition systems primarily due to the ideal recording conditions [**49**], it is widely available and remains one of the popular databases used for exploratory investigation of speaker identification systems [**40**]. The speaker identification performance with this database is expected to be higher than extemporaneous speech because of the limited variability [**50**].

To further extend the current study to children's speech, the MIC recordings, which are acquired by using high-fidelity recording equipment, from the CID database are used. Since the SGR estimation method for children's speech presented in the last section is developed for children from 8 to 17 years of age, only the subset of children with the same age range from the database is used. There are 323 children in this subset including 179 males and 144 females. Each speaker has recorded 49 utterances on average (15946 utterances in total) with minimum imitation, and all the recordings are sampled at 16 kHz. All speakers are asked to read a list of sentences (details can be found in [**48**]). As expected, acoustic variations in children's speech is higher than adults. The interested reader is referred to [**51**] for detailed analyses of the duration, formant, and pitch in the database. Although the database is not designed for speaker identification, it is commonly used for investigation in different speech applications with children, such as in [**52**] and [**53**], and the number of children is sufficiently large for the exploratory study on speaker identification in the current chapter.

3.4 Methods

The identification rate (IR) is adopted as the performance measure for all experiments in the current chapter and can be calculated as,

$$\mathrm{IR} = \frac{\text{Total number of correctly identified tokens}}{\text{Total number of tested tokens}} \times 100\,. \tag{17}$$

Since the "ground truth" SGR values can only be obtained from the WashU-UCLA corpus, SGR estimates from presented methods are used for all experiments with both TIMIT and CID databases.

3.4.1 Pilot Experiment

As a pilot experiment, the SGR features are first evaluated using the WashU-UCLA corpus with "ground truth" SGR values measured from accelerometer signals. All the measurements are acquired by using the Subglottal Resonance Measuring Tool presented in Section 2.3, and speaker-level SGR values are calculated by averaging over multiple utterances. Since the one of the goals of this study is to explore the effect of appending SGRs to MFCCs as discussed in Section 3.1, different ways of combining *speaker-level* SGR values and *frame-level* MFCCs are evaluated. Although the ideal method is to measure the corresponding frame-level SGR values and append them to MFCCs, this is not quite possible for all recordings due to both the unstable quality and unknown formant-SGR interactions in some utterances. On the other hand, adding speaker-level SGR values solely to MFCCs is violating the structure of GMM and loses the ability to account for within-speaker SGR variability. In order to address these challenges,

random variations from different ranges are attached to speaker-level SGR values to create simulated frame-level SGR values. The ranges are set to different multiples (from 1 to 3 in the current study) of the within speaker standard deviations (SDs) for different SGRs as reported in [**5**]. Although speaker information can be embodied in any frame of a given utterance, such as vowel information from voiced frames and speech pausing information from non-speech frames, the presented SGR estimation method only relies on information from voiced frames. In order to have a fair comparison, only voiced frames, which are extracted using the Snack Toolkit with default settings, are used in the pilot experiment. Each utterance from the WashU-UCLA corpus is around 4 to 5 seconds long, with roughly 25% being voiced frames. Since the corpus does not have well-defined train and test sets, utterances are partitioned manually. In the training set, 1 sentence from each of 4 specific target vowel sets (/i/, /æ/, /ɑ/, /u/) in session 2 of the corpus are selected for every speaker, and 3 sets of data are evaluated for cross-validation purpose . For example, the set with sentence identification 3 (SID 3) is constructed by selecting the third sentence from each of the 4 specific vowel sets. In other words, each speaker model is trained with 18 seconds long utterance or, equivalently, 4.5 seconds of voiced speech on average. The widely used 39-diemensional MFCCs (with delta and delta-delta) are used as baseline, and the simulated frame-level SGRs are evaluated by appending them to MFCCs. After testing with different model orders, 32-mixture GMM model is empirically chosen for all experiments in the current study for a consistent comparison among different settings. In the testing stage, 10 sentences are randomly selected from session 1 of the corpus for each speaker (total of 500 sentences for 50 speakers) in order to get better inter-session variability. The selected sentences are tested individually with all the trained models, and the identification decision is made based on maximum a-posteriori probability as discussed in Section 3.1.

3.4.2 Experiments with TIMIT and CID databases

To the author's knowledge, the most common baseline on the TIMIT database is achieved by a slightly different MFCC feature set compared to the commonly used 39 dimensional MFCCs as in [**17**] and [**40**]. The new MFCC feature set, which is first introduced in [**54**] and evaluated on the TIMIT database by [**50**] and [**55**], uses all coefficients from a 24 channel Mel frequency-spaced filter except the zeroth cepstral coefficient to get a final 23 dimensional feature vector. Although temporal derivatives, such as delta and delta-delta, are commonly used to help identifying speaking styles and durations, they are not included in the 23 dimensional MFCCs. One of the possible explanations is that recordings from the TIMIT database have limited variability in both speaking styles and durations due to the designed speech content recorded under studio condition. Furthermore, environmental compensation for removing time-invariant channel effects, such as cepstral mean normalization, is not used because the database does not contain acoustic noise and microphone variability. The commonly used partitioning method selects 8 sentences, including 2 SA, 3 SI and 3 SX sentences (approximately 24 seconds), from each speaker to train the models while the remaining 2 SX sentences (approximately 3 seconds each) are tested individually. In order to compare situations with limited training data, experiments with the same testing configuration but different number of training data are evaluated. In the case of 5 training sentences, 3 SI and 2 SA sentences are selected while only 3 SI sentences are used in the case of 3 training sentences.

To extend the study in children's speech, the 23 dimensional MFCCs are evaluated on the CID database. Since children have both large inter-speaker and intra-speaker variation in pronunciation duration, the lengths of recorded utterances in the CID database have high variations. In order to maintain both phonetic variability and duration of training data, utterances

are randomly chosen with a minimum total duration of selected utterances to train a model for a given speaker. For example, in order to get a minimum of 4 seconds of speech for training, variable numbers of utterances ranging between 1 and 4 are selected while 3 of the remaining sentences are randomly selected to test the trained models individually. The number of testing sentences is fixed to 3 for all experiments on the CID database in order to reduce the experimental variability.

3.5 Results and Discussion

This section will present and discuss the results of experiments with different databases. The appended SGR features contribute differently in each of the cases, but the general advantage of using the features can be observed throughout the discussion.

3.5.1 Pilot Experiment: WashU-UCLA corpus

The results of all pilot experiments from the WashU-UCLA corpus are presented in Table 11 to Table 13 for different SGRs and ranges of variation.

Since Sg1 is influenced by the interferences from the lower harmonics and Sg3 is attenuated by the low-pass nature of the skin when acquiring accelerometer signals, Sg2 is relatively easier to measure and performs the best as a single SGR feature when appended to MFCCs in the pilot experiment. The IR improves from 69.9% to 86.8% by adding simulated frame-level Sg2 with 1 SD-ranged random variations as shown in Table 11. As expected, the performance decreases as the range of random variation increases because the simulated frame-

level Sg2 values are more deviated from the “ground truth” speaker-level value, and the wide spread variation increases the probability of getting Sg2 values overlapped between speakers. Meanwhile, the within speaker Sg2 standard deviation is 32 Hz and the corresponding average root mean squared error (RMSE) of the presented estimation method is 61 Hz. Hence, the IRs with 3 SD ranged random variations approximately account for variations from both within speaker standard deviation and estimation error. Along the line of simulating the estimation error from “ground truth” Sg2, one might expect to get smaller SGR estimation error in the training stage with 4 sentences than in the testing stage with only 1 sentence. In order to simulate this scenario, an experiment with mismatched ranges of random variations during training and testing stages is evaluated, and the corresponding IRs are presented in the last column of Table 11. In all cases, the additional Sg2 feature improves IR by different amounts and suggests further exploration.

Motivated by the performance improvements achieved by attaching Sg2 to MFCCs, Sg1 and Sg3 are appended in two stages on top of Sg2. The first three columns of Table 12 show the resulting IRs by appending both Sg1 and Sg2 with different ranges of random variations to MFCCs. The additional Sg1 gives about 2% to 3% improvements on top of the achievement by Sg2 while Sg3 adds another 2% to 3% jump as shown in the last three columns of Table 12. The final averaged IR by using all 3 SGRs with 1 SD random variation is 92.1%. Although there is a 22.2% absolute improvement over the baseline with MFCCs, this is the ideal scenario without considering the SGR estimation error in practical applications.

Table 11: The IR on the WashU-UCLA corpus with 39 dimensional MFCCs and different variations of additional "ground truth" Sg2 as features. The SG2_1SD , for example, represents random variation within 1 SD of Sg2 is used to create the simulated frame-level Sg2 values. The last column presents the IR with mismatched ranges of variation in train and test stage.

SID	MFCC_39	MFCC_39 + SG2_1SD	MFCC_39 + SG2_2SD	MFCC_39 + SG2_3SD	MFCC_39 + SG2Train_1SD + SG2Test_3SD
3	73.4%	88.6%	84.2%	81.2%	85.6%
6	68.8%	86.2%	80.1%	79.6%	83.4%
8	67.6%	85.6%	80.1%	77.8%	81.8%
Average	**69.9**%	**86.8**%	**81.4**%	**79.5**%	**83.6**%

Table 12: The IR on the WashU-UCLA corpus with 39 dimensional MFCCs and different variations of additional "ground truth" SGRs as features. The SG123_1SD, for example, represents random variations within 1 SD of Sg1, Sg2 and Sg3 are used to create the simulated frame-level SGR values.

SID	MFCC_39 + SG12_1SD	MFCC_39 + SG12_2SD	MFCC_39 + SG12_3SD	MFCC_39 + SG123_1SD	MFCC_39 + SG123_2SD	MFCC_39 + SG123_3SD
3	90.0%	84.0%	83.8%	93.8%	86.4%	83.6%
6	87.6%	85.0%	79.4%	92.4%	84.8%	84.6%
8	89.6%	84.2%	81.2%	90.0%	84.0%	84.8%
Average	**89.1**%	**84.4**%	**81.5**%	**92.1**%	**85.1**%	**84.3**%

In order to verify the claim that the IRs with 3 SD-ranged random variation simulates the IRs with estimation errors in practical situations, experiments are carried out by replacing the “ground truth” Sg1 and Sg2 values by estimated Sg1 and Sg2 values. In addition, the previously determined within-speaker SD values, which are used to guide the range of frame-level random variations, are replaced by standard deviation from the actual frame-level estimates. Following

the discussion in Section 3.2, only Sg1 and Sg2 estimations are evaluated because the best available Sg3 estimation method solely relies on the estimation of Sg2 without additional speaker specific information. The results are presented in Table 13 with two different approaches for calculating the estimated speaker-level SGR values. Although arithmetic mean of all frame-level estimates is used to obtain speaker-level SGR values in previous studies, another approach using the median instead of mean is evaluated in the current study because median is sometimes less sensitive to outliers. The results of both approaches are very close to each other that imply the frame-level SGR estimates are more likely to be Gaussian distributed. By comparing the results from estimated SGR and "ground truth" SGR with 3 SD random variations, they are off by about 6% which suggests the variations from estimation errors are larger than expected. Nevertheless, the averaged IR with estimated Sg1 and Sg2 is still 4.2% (from 69.9% to 75.7%) better than the baseline; however, this result can be biased because WashU-UCLA corpus is used to evaluate the pilot experiments and train the presented estimation methods. Therefore, an extended study on a commonly evaluated database for speaker identification is undertaken.

Table 13: The IR on the WashU-UCLA corpus with 39 dimensional MFCCs and additional estimated Sg1 and Sg2 with random variations within 1 SD from the estimates attached as features. The IR, with speaker-level estimate calculated by either averaging or taking median from the frame-wise estimations, is shown in column 2 and 3, respectively.

SID	MFCC_39 + SG12_Est_Mean_1SD	MFCC_39 + SG12_Est_Med_1SD
3	79.6%	79.6%
6	75.8%	73.6%
8	70.4%	74.0%
Average	**75.3**%	**75.7**%

3.5.2 Adults: TIMIT database

The IR by using the popular TIMIT database is expected to be higher than experiments with extemporaneous speech due to the near ideal recording condition. By evaluating the optimized 23 dimensional MFCC feature with the TIMIT database, an IR of 99.37%, which matches the results from [**50**] and [**55**], is achieved by training models with 8 sentences. Although the number of identifying speakers has been shown by [**50**] to have negligible influence on speaker identification performance, the amount of training data always remains a concern to statistical classifiers. In order to compare situations with limited training data, baseline experiments with the same testing configuration and different amount of training data are evaluated. The IR with models trained by 5 sentences is 97.78%, and the more interesting case with only 3 sentences has an IR of 84.68% as shown in Table 14. The performance with 3 training sentences is not good enough because commercial applications of biometric identification require at least 98% accuracy [**36**].

Table 14: IRs on the TIMIT database with different number of training sentences from adults. The number of testing sentences are fixed to 2 for all cases and the sentences are tested individually.

Num. of Training Sentences	Num. of Testing Sentences (Individually)	IR
8	2	99.37 %
5	2	97.78 %
3	2	84.68 %

Since the identification performance for both experiments with 8 and 5 training sentences are sufficient for commercial applications, SGR features with standard deviations from speaker-

level estimates are evaluated with baseline MFCCs for 3 training sentences. The same estimation procedure for simulating frame-level SGR values as in the pilot experiment is used, and the speaker-level SGR values are calculated by the mean of frame-level SGR estimates. The resulting IRs of attached Sg1 and Sg2 with 1 SD random variations are 84.60% and 83.02% , respectively. These results are just comparable to the baseline performance without advantages as in the pilot experiments, and the features with Sg1 give better performance than that with Sg2. One of the possible explanations of these results is that the estimation method is biased in the pilot experiment as discussed previously, and the SGR estimation accuracy is not sufficient to discriminate a large number of speakers. The reported Sg1 and Sg2 inter-speaker ranges for American English speakers are 230 Hz and 393 Hz across genders, respectively. By taking the number of enrolled speakers and estimation errors into account, their SGR values are heavily overlapped which makes the features less discriminative. Moreover, the RMSE and the mean-relative standard deviation (MSD), which is used to quantify the consistency of estimation, for Sg1 is 25 Hz and 1.6 % compared to 61 Hz and 1.2 % for Sg2. The ratio between the RMSE and range of SGRs for Sg1 and Sg2 are about 0.11 and 0.16, respectively. Therefore, the estimation of Sg1 clearly has smaller error with higher consistency than that of Sg2, and these advantages probably account for the higher identification performance. In addition, SGRs theoretically do not exist for the non-speech portions of an utterance, and silence frames usually degrade the training of speaker models as discussed by [**40**]. However, the well-known MFCC baseline for the TIMIT database trains models with all frames from an utterance as reported by [**55**]. In order to explore the problem with silence, speech frames are extracted from the TIMIT database by using the provided word boundaries from a dynamic string alignment program. The baseline result with only speech frames drops 4.97 % (from 84.68% to 79.71%) as shown in Table 15, but

the additional Sg1 features with different ranges of random variations consistently give small improvements instead of degradations as in the experiments with both speech and non-speech frames. These experiments with only speech frames are theoretically more meaningful than the experiments with all frames because SGRs are naturally present only in speech portions of an utterance, but the degradation between baseline results remains an issue. One of the possible explanations is that the TIMIT database contains only reading sentences that are recorded consecutively in a single session for each speaker. Hence, the durations of silences from each utterance may contain similar patterns for a given speaker and help speaker identification as a discriminative feature. Nevertheless, Sg1 shows advantages in a more meaningful experimental setup, but further exploration is necessary.

Table 15: IRs on the TIMIT database with only speech frames extracted from given word boundaries. The numbers of adult training and testing sentences are fixed to 3 and 2, respectively. The baseline feature is the 23 dimensional MFCCs and additional Sg1 with different ranges of random variations are evaluated.

Num. of Training Sentences	Num. of Testing Sentences (Individually)	Features	IR
3	2	Speech Frames + MFCC_23	79.71 %
3	2	Speech Frames + MFCC_23 + Sg1_1SD	80.03 %
3	2	Speech Frames + MFCC_23 + Sg1_2SD	79.78 %
3	2	Speech Frames + MFCC_23 + Sg1_3SD	80.32 %

3.5.2 Children's Speech: CID database

In order to further investigate the argument that estimated SGRs are close together in the inter-speaker SGR ranges in the TIMIT database, experiments are undertaken with children's speech from the CID database because children have higher SGR inter-speaker variations than adults. By combining the reported SGRs from both [**1**] and [**2**], the Sg1 and Sg2 inter-speaker ranges for children between 8 to 17 years old are 351 Hz and 847 Hz, respectively. With these higher than adult ranges and less crowded speaker space (only 323 children in the CID database are evaluated), the SGRs are expected to help speaker identification to a greater extent. With minimum 4 seconds of training speech, the baseline result with 23 dimensional MFCCs and exploratory results with appended SGR features are presented in Table 16. There is a 1.86% improvement achieved by the additional Sg1 feature with 1 SD ranged random variation as expected, but degradation is observed in the case with Sg2 feature. The results of appending Sg1,

Sg2 and Sg3 together are ignored due further degradations. These observations can be possibly justified by the high RMSE, which is 144 Hz and more than double to that of adults, for estimating Sg2 with children's speech. In the interest of verifying the advantage of using Sg1 feature, another set of experiments with minimum 6 seconds of training speech is tested. The new baseline IR increases by 5.37% to 93.81% with the additional training data, and the Sg1 feature further improves the accuracy by an extra 0.82% to 94.63%. These consistently improving performances with children by appending Sg1 feature give evidence that the possible limiting factor of speaker identification is the overlapping of speakers in SGR inter-speaker ranges.

Table 16: IRs on the CID database with minimum 4 seconds of training speech. The baseline result is shown in the first row, and IRs with additional SGR features are presented for comparison.

Minimum total duration of training sentences	Num. of Testing Sentences (Individually)	Features	IR
> 4 seconds	3	MFCC_23	88.44 %
> 4 seconds	3	MFCC_23 + Sg1_1SD	90.30 %
> 4 seconds	3	MFCC_23 + Sg2_1SD	84.52 %

To further investigate the effects of estimation errors to speaker identification, another set of experiments by training models with only voiced frames from minimum 8 seconds of speech is evaluated, and the results are shown in Table 17. The voiced frames, which account for less than half of the total number of frames, are selected by using the Snack Toolkit [**23**]. The baseline IR with equivalently less than 4 seconds of frames drops about 2.78%, but the additional

Sg1 feature improves the accuracy by 2.37% which is higher than the improvement from the set of experiments with minimum 4 seconds of utterances. This encouraging improvement is meaningful because the speaker-level SGR estimates are determined by using F0, F1 and F3 from only voiced frames as discussed previously. In order to show the advantage of using SGR instead of acoustic measures, the IR with models trained by appending extracted frame-level F0, F1 and F3 directly to the baseline MFCCs instead of Sg1 is presented in Table 17. Although Sg1 is estimated from the same acoustic measurements, the appended F0, F1 and F3 degrade the performance by 0.83% from the baseline instead of improving as with the SGR feature.

Table 17: IRs on the CID database with voiced frames from minimum 8 seconds of training speech. The baseline result is shown in the first row, and the IR with additional Sg1 features follows. The last row shows result with additional acoustic features, including F0, F1 and F3.

Minimum total duration of training sentences	Num. of Testing Sentences (Individually)	Features	IR
> 8 seconds	3	Voiced Frames + MFCC_23	85.66 %
> 8 seconds	3	Voiced Frames + MFCC_23 + Sg1_1SD	88.03 %
> 8 seconds	3	Voiced Frames + MFCC_23 + F013	84.83 %

3.5 Summary

The exploratory study presented in this chapter is the first of its kind to examine the possible role of SGRs in speaker identification. Although this study uses the basic GMM classifier which was

proposed for speaker identification over a decade ago, it remains the most popular approach in the literature today due to its reliable performance. By using the “ground truth” SGRs in the pilot experiments, strong improvements can be observed which motivate the rest of the study. However, additional variations from estimation errors reduce the improvement for the WashU-UCLA database. In order to extend and verify the study, the features are evaluated with 630 speakers from the well-known TIMIT database; however, the SGR features unexpectedly give small degradation to the performance. This result might be explained by the heavily overlapped SGRs with estimation errors in the limited inter-speaker ranges. For a more meaningful setup, evaluations for SGR features with only speech frames, excluding silence frames, show better performance, but the overlapping problem remains a challenge. To further explore this limitation, children’s speech, which have higher inter-speaker SGR ranges, from the CID database are used for evaluation. Although the estimation error is higher for children’s speech, the speaker identification performance is improved by the additional Sg1 feature, and this result supports the claim that the limiting factor in speaker identification with SGR features is the crowding of SGRs in the inter-speaker ranges. Nevertheless, more accurate SGR estimations can reduce the overlapping of these features and help the performance approach the theoretical upper limit as presented in the pilot experiments.

Since this is an exploratory study, a simple and straight forward method is used to append SGR features to the well-known baseline MFCC features. The current appending method is sensitive to estimation errors, and the best available estimation methods are not designed for determining frame-level SGRs. Therefore, some further investigations on both estimating and appending SGR features are necessary. A possible suggestion for future work will be combining short-term cepstral features and the SGRs as long-term features by incorporating different kinds

of classifiers. This idea is inspired by studies on speaker age classification which utilize the advantages from both long-term and short-term features, and similar feature characteristics have been observed in the current study. For instance, MFCCs are categorized as short-term feature which contain more phonetic information of an utterance while long-term features, such as pitch, jitter and shimmer, carry more paralinguistic information as discussed in [**56**]. Since SGRs do not vary too much with the content of speech, modeling them as long-term features might be more meaningful and support-vector-machine (SVM) would be a better classifier for these features. In addition, SGRs can be possibly combined with other features with speaker-specific information, such as the set of acoustic parameters proposed by [**17**], because they do not directly contain overlapped information. With all these possibilities and the demonstrated advantages throughout the exploratory study, further investigations on speaker identification with SGR features are highly motivated.

CHAPTER 4. SUMMARY

The current study presents two interesting topics on subglottal resonances in speech processing. The first topic presented in Chapter 2 extends previous studies on the oral-subglottal coupling effect, and Chapter 3 investigates an unexplored topic of speaker identification with SGRs.

In order to explore the coupling effect in detail, more sophisticated tools are developed. The new tools do not only enable proper measurements of F2 discontinuities for the presented study, but also contain useful features for future explorations. In the interest of examining and extending the previous studies, measurements are acquired from a recently collected and relatively larger corpus than has been used. Part of the analysis on the collected data matches the results from previous studies, but conclusive results cannot be obtained due to several observed uncertainties. This skepticism arises from both time-frequency resolution trade-off and uncertain influences from articulatory coupling. Nevertheless, the study provides cues for correlations between voice quality parameters and F2 discontinuities, but further investigation is required.

In order to motivate studies on SGRs by introducing possible applications, an exploratory study on speaker identification with SGR features is presented. Since existing SGR direct estimation methods rely on unstable cues from oral-subglottal coupling, a more reliable indirect estimation method by utilizing statistical relationships between SGRs and vowel boundaries is used. This estimation method has been successfully applied to different applications, such as speaker height estimation and speaker normalization on ASR, but it is designed for speaker-level estimates while the most popular speaker identification approach takes advantages from frame-

level features. Moreover, a modified version of this estimation method for children's speech is also presented in order to increase the extendibility of the current study. For exploratory purposes, a straight forward method by appending statistically bounded random variations to the speaker-level estimates is developed to simulate frame-level SGR values. The same appending method is applied to the "ground truth" speaker-wise SGR measurements to simulate frame-level SGRs, and the pilot experiments with these simulated features attached to baseline cepstral features demonstrate significant improvements. However, results with estimated speaker-wise SGRs show some drawbacks which are possibly caused by the overlapping of SGR values in the inter-speaker ranges. This explanation is supported by the better identification performances on children's speech because children have relatively wider inter-speaker SGR ranges. Although this simple appending method may not be the best for the SGR features, it is good enough for this exploratory study to demonstrate the speaker discriminative ability of SGRs. The identification performance is expected to be higher as in the pilot experiments with more accurate SGRs estimation methods in the future. In addition, some possible and sophisticated classification models, such as a hybrid model combining GMM and SVM for cepstral and SGR features, respectively, are suggested at the end of Chapter 3 to motivate further explorations.

This is the end of the current study, but it is also the beginning of all the possibilities discussed throughout the chapters.

APPENDIX

A. The Sg2 measurements for the 20 speakers in Chapter 2

Speaker-wise statistics of Sg2 measurements collected for analysis in Chapter 2 are presented below. The 10 male speakers are listed on the left table while the 10 female speakers are on the right hand-side table. The intra-speaker Sg2 measurement statistics including minimum (Min), maximum (Max), mean ($\overline{Sg2}$), and standard deviation (SD) are all in Hertz (Hz). Inter-speaker averages are reported at the bottom of the table.

Table 18: The Sg2 measurements for the 20 speakers in Chapter 2 separated by gender.

Males	Sg2 (Hz)			
ID	Min	Max	$\overline{Sg2}$	SD
11	1203	1391	1280	44
12	1357	1416	1385	13
15	1201	1326	1252	32
22	1270	1346	1311	27
38	1270	1355	1326	28
41	1201	1230	1217	11
43	1230	1354	1298	37
52	1234	1313	1286	25
53	1178	1301	1235	42
64	1258	1434	1335	39
Avg.	**1240**	**1347**	**1293**	**30**

Females	Sg2 (Hz)			
ID	Min	Max	$\overline{Sg2}$	SD
14	1492	1586	1530	27
16	1484	1586	1538	30
19	1508	1594	1547	29
24	1453	1516	1491	17
27	1328	1539	1413	51
32	1551	1613	1583	21
33	1320	1445	1382	29
36	1484	1574	1541	25
40	1465	1563	1510	33
59	1514	1660	1592	39
Avg.	**1460**	**1568**	**1513**	**30**

B. Correlation between discontinuities and voice quality parameters

The correlation coefficients between the size of the F2 jump and different voice quality parameters with the three subsets of data are presented in Table 19 and Table 20. The tables show the inter-speaker correlation coefficients between the size of F2 jump and H1-H2. The (*) in Table 19 indicates the use of harmonic magnitude correction formula to reduce the influence of vocal tract resonances [**32**].

Table 19: Correlation coefficients between the size of F2 jump and two voice quality parameters with correction formula to reduce the influence of vocal tract resonances.

		S1 - F2 Jump		S2 - F2 Jump	
		Male	Female	Male	Female
V1	H1*-A1*	0.04	0.01	-0.06	0.01
	H1*-A3*	-0.07	0.63	0.07	0.57
V2	H1*-A1*	-0.13	0.21	-0.18	0.07
	H1*-A3*	-0.02	0.67	0.05	0.63
V3	H1*-A1*	0.39	0.61	0.03	0.29
	H1*-A3*	0.25	0.53	0.31	0.60

Table 20: Correlation coefficients between the size of F2 jump and two voice quality parameters without using any correction formula.

		S1 - F2 Jump		S2 - F2 Jump	
		Male	Female	Male	Female
V1	H1-A1	0.14	0.43	0.03	0.24
	H1-A3	-0.15	0.44	-0.04	0.39
V2	H1-H2	0.49	0.38	0.40	0.34
	H1-A3	-0.24	0.45	-0.16	0.40
V3	H1-A1	0.44	0.72	0.16	0.38
	H1-A3	0.01	0.79	-0.06	0.45

Both H1-A1 and H1-A3 can be viewed as measures of the spectral slope. Besides, studies such as [**57**], have demonstrated the relationship between spectral slope and glottis air leakage. Although positive correlations are observed between the H1-A3 with the size of the F2 jump for females across different setups, the results are inconclusive due to the inconsistent findings from male speech. Such correlation differences between the two genders may be caused by gender-related breathiness as described in Chapter 2. For the correlations with H1-A1, no consistent results are found for a conclusion. By the same argument of the dependency of oral-subglottal coupling effect on the glottal area as presented in Chapter 2, stronger coupling should be observed during the close phase from a speaker with posterior glottal chink. In other words, speakers with a posterior glottal chink should have smaller F2 jump size differences between open and close phase measurements. Since H1-A1 is correlated with the presence of a posterior glottal chink, its correlation with the frequency jump size difference between open and closed phase measurements can be interesting.

Although the correlation between the size of F2 jump and H1-A3 is positive for female speech, the result is not convincing due to both the reliability concern for high pitch speakers and the inconsistent results with male speakers. Moreover, the H1-A1 study also gives inconclusive results, but a hypothesis with the comparison between open and close phase measurements is suggested for future work. The analysis on A2 attenuation is not presented in detail due to the uncertainties of the A2 measurements with the developed procedure. However, improved ways for quantifying the amplitude attenuation, such as an amplitude only study by increasing time resolution around the frequency jump region, can be investigated to get a better overall picture of the oral-subglottal coupling effect.

BIBLIOGRAPHY

[1] H. Arsikere, G. Leung, S. Lulich, and A. Alwan, "Automatic estimation of the first two subglottal resonances in children's speech with application to speaker normalization in limited-data conditions," in *Interspeech*, 2012, submitted for publication.

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

[3] S. Lulich, J. Morton, H. Arsikere, M. Sommers, G. Leung, and A. Alwan, "Subglottal resonances of adult male and female native speakers of American English," *Journal of the Acoustical Society of America*, submitted for publication.

[4] H. Arsikere, G. Leung, S. Lulich, and A. Alwan, "Automatic height estimation using the second subglottal resonance," in *ICASSP*, 2012, pp. 3989-3992.

[5] H. Arsikere, G. Leung, S. Lulich, and A.Alwan, "Automatic estimation of the first three subglottal resonances in adults' speech with application to speaker height estimation using speech signals," *Speech Communication Journal*, submitted for publication.

[6] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: The MIT Press, 1998.

[7] A. Liberman, *Speech: A special code.*: The MIT Press, 1996.

[8] S. Wang, Y. Lee, A. Alwan, "Bark-shift based nonlinear speaker normalization using the second subglottal resonance," in *Interspeech*, 2009, pp. 1619-1622.

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

[10] S. Wang, A. Alwan and S. Lulich, "Speaker normalization based on subglottal resonances," in *ICASSP*, 2008, pp. 4277-4280.

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

[12] X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1735-1745, 2007.

[13] S. Lulich, "Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, no. 1, pp. 20-32, Jan. 2010.

[14] H. Arsikere, S. Lulich, and A. Alwan, "Automatic estimation of the first subglottal resonance," *Journal of Acoustic Soceity of America (Express Letters)*, vol. 129, pp. 197-203, 2011.

[15] H. Arsikere, S. Lulich and A. Alwan, "Automatic estimation of the second subglottal resonance from natural speech," in *ICASSP*, 2011, pp. 4616-4619.

[16] M. Sonderegger, "Subglottal coupling and vowel space: An investigation in quantal theory," Physics B.S. thesis, Cambridge, MA, 2004.

[17] S. Manocha, and S. Vishnubhotla C. Espy-Wilson, "A new set of features for text-independent speaker identification," in *Interspeech*, 2006, pp. 1475-1478.

[18] J. Flanagan, C. Coker, L. Rabiner, R. Schafer and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum*, vol. 7, no. 10, pp. 22-45, October 1970.

[19] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ: Pearson Higher Education, Inc., 2011.

[20] H. Gray, *Anatomy of the Human Body*. Philadelphia, PA: Lea & Febiger, 1918.

[21] B. Atal and S. Hanauer, "Speech analysis and aynthesis by linear prediction of the speech wave," *Journal of Acoustical Society of America*, vol. 50, no. 2, pp. 637-655, August 1971.

[22] H. Hanson and K.Stevens, "Sub-glottal resonances in female speakers and their effect on vowel spectra," in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 3, Stockholm, 1995, pp. 182-185.

[23] K. Sjölander. (1997) The Snack sound toolkit. [Online]. http://www.speech.kth.se/snack/

[24] S. Lulich, J. Morton, M. Sommers, H. Arsikere, Y. Lee and A.Alwan, "A new speech corpus for studying subglottal acoustics in speech production, perception, and technology.," *Journal of the Acoustical Society of America*, vol. 128, no. 4, p. 2288, 2010, (Abstract).

[25] H. Arsikere, Y. Lee, S. Lulich, J. Morton, M. Sommers, and A. Alwan, "Relations among subglottal resonances, vowel formants, and speaker height, gender, and native language.," *Journal of the Acoustical Society of America*, vol. 128, no. 4, p. 2288, 2010, (Abstract).

[26] H. Hanson and E. Chuang, "Glottal characteristics of male speakers: acoustic correlates and comparison with female data.," *Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 1064-1077, 1999.

[27] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 40-45, Jan. 1999.

[28] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[29] P. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm.," *IEEE Transaction on Speech and Audio Processing*, vol. 15, pp. 34-43, Jan. 2007.

[30] M. Brookes. (2006) Voicebox: a speech processing toolbox for MATLAB. [Online]. http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[31] Y. Shue. (2010) VoiceSauce: a program for voice analysis. [Online]. http://www.ee.ucla.edu/~spapl/voicesauce/

[32] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation.," in *ICASSP*, Montreal, Canada, 2004, pp. 669-672.

[33] K. Honda, S. Takano, H. Takemoto, "Effects of side cavities and tongue stabilization: Possible extensions of the quantal theory," *Journal of Phonetics*, vol. 38, no. 1, pp. 33-43, January 2010.

[34] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of Acoustical Society of America*, vol. 87, no. 2, pp. 820-857, 1990.

[35] M. Södersten and P. Lindestad, "Glottal closure and perceived breathiness during phonation in normally speaking subjects," *Journal of speech and hearing research*, vol. 33, pp. 601-611, Sep. 1990.

[36] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4-20, Jan. 2004.

[37] J. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437-1462, Sep. 1997.

[38] D. Reynolds, "An overview of automatic speaker recognition technology," in *ICASSP*, 2002, pp. 4072-4075.

[39] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectos," *Speech Communication*, vol. 52, no. 1, pp. 12-40, Jan. 2010.

[40] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23-61, 2nd Quarter 2011.

[41] S. Davis and P. Mermelstein, "Comparison of parametric representations for mnosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[42] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transaction on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[43] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91-108, 1995.

[44] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models," University of Berkeley, Technical Report ICSI-TR-97-021, 1997.

[45] A. Syrdal and H. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *Journal of Acoustic Society of America*, vol. 79, no. 4, pp. 1086-1100, 1986.

[46] H. Traunmuller, "Analytical expressions for the tonotopic sensory scale," *Journal of Acoustic Society of America*, vol. 88, no. 1, pp. 97-100, 1990.

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

[48] J. Miller, S. Lee, R. Uchanski, A. Heidbreder, B. Richman, and J. Tadlock, "Creation of two children's speech databases," in *ICASSP*, 1996, pp. 849-852.

[49] J. Campbell and D. Reynolds, "Corpora for evaluation of speaker recognition systems," in *ICASSP*, 1999, pp. 829-832.

[50] D. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," *The Lincoln Laboratory Journal*, pp. 173-192, 1995.

[51] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: duration, pitch and formants," in *EUROSPEECH*, 1997, pp. 473-476.

[52] M. Iseli, Y. Shue and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *Journal of Acoustic Society of America*, vol. 121, no. 4, pp. 2283-2295, 2007.

[53] G. Chen, X. Feng, Y. Shue and A. Alwan, "On using voice source measures in automatic gender classification of children's speech," in *Interspeech*, 2010, pp. 673-676.

[54] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639-643, Oct. 1994.

[55] D. Reynolds, M. Zissman, T. Quatieri, G. O'Leary and B. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *ICASSP*, 1995, pp. 329-332.

[56] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *Interspeech*, 2007, pp. 2277-2280.

[57] B. Cranen and J. Schroeter, "Modeling a leaky glottis," *Journal of Phonetics*, vol. 23, no. 1-2, pp. 165-177, Apr. 1995.

[58] J. Grey and J. Gordon, "Perceptual effects of spectral modifications on musical timbres.," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493-1500, 1978.