Title

The analysis and applications of Subglottal Resonances in height estimation and speaker identification and normalization

Author

Guo, Jinxi

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The analysis and applications of Subglottal Resonances in height estimation and speaker identification and normalization

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Electrical Engineering

by

Jinxi Guo

2015

ABSTRACT OF THE THESIS

The analysis and applications of Subglottal Resonances in height estimation and speaker identification and normalization

by

Jinxi Guo

Master of Science in Electrical Engineering

University of California, Los Angeles, 2015

Professor Abeer A. H. Alwan, Chair

The subglottal acoustic system refers to the acoustic system below the glottis, which consists of the trachea, bronchi and lungs. Compared to the supraglottal system, the configuration of the subglottal system is relatively fixed and more speaker dependent. Previous research showed that the natural frequencies of the subglottal system, which are referred to as subglottal resonances (SGRs), form the boundaries of vowel classes for several languages. Results in previous studies also indicate that SGRs correlate well with the standing height for adult speakers. Motivated by these properties, SGRs have been used in different applications including adults' height estimation and speaker normalization for automatic speech recognition (ASR). In this thesis, our knowledge of SGRs is leveraged to extend the utility of SGRs to more language and applications, including: (1) finding the relationship between SGRs and vowel class, speaker height and F0 variation for Mandarin speakers; (2) finding the relationship between SGRs, height estimation

and speaker normalizations for children's speech; (3) the investigation of SGRs for speaker identification (SID) under noisy conditions.

The results indicate that SGRs, similar to English speakers, divide the vowel space for Mandarin speakers and there exist strong inverse relationships between SGRs and speaker height, and between SGRs and trunk length. Moreover as a tonal language, while F0 varies over time within a vowel, there is no statistically-significant variation of SGRs in Mandarin speech. For the study of children speech, an age-dependent SGRs estimation algorithm is designed. The experiments show that the algorithm is effective for children height estimation and speaker normalization. For SID, SGRs are used as noise robust features to provide complementary information to state-of-the-art noise robust features, such as power normalized cepstral coefficients. A two-stage framework is developed and the results show that SGRs provide significant performance improvements.

The thesis of Jinxi Guo is approved.

Alan Laub

Yingnian Wu

Abeer A. H. Alwan, Committee Chair

University of California, Los Angeles

2015

To my family, teachers and friends

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere gratitude to my advisor Prof. Abeer Alwan for all her support and guidance throughout my graduate study here in UCLA. It has been a privilege to work under her supervision. I would like also express my appreciation to Prof. Alan Laub and Prof. Yingnian Wu for their time in serving as members of my thesis committee. I would like to thank all my lab-mates at UCLA, for their help. Last but not least, I would like to thank my family and friends for all their support.

Chapter 1

Introduction

1.1 Overview and motivation

The speech-production system can be viewed as being composed of 3 subsystems: (1) the subglottal system, (2) the larynx, and (3) the supraglottal system (known as the vocal tract). The subglottal system comprises the trachea, bronchi and lungs, and is responsible for generating and driving the airflow required for speech production [1]. Figure 1.1 shows the subglottal system. The subglottal resonances (SGRs) are the natural frequency of the subglottal system, and correspond to the complex conjugate pairs of poles in the subglottal input impedance [2].

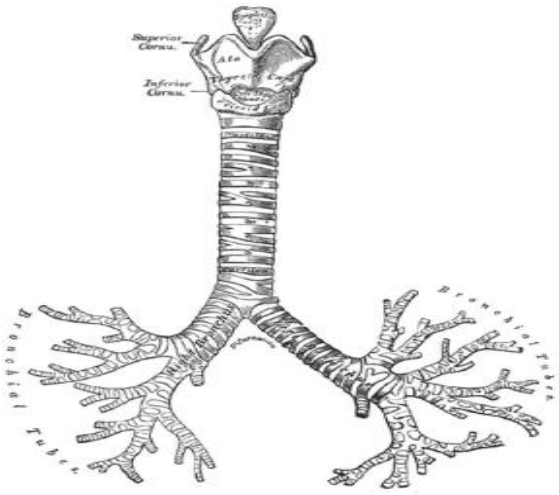

Figure 1.1: The subglottal airways, including the trachea, main bronchi, and the bronchial tree down to about 6 generations.

Owing to the absence of moving articulators and body parts, the acoustics of the subglottal system are much more stationary over time compared to the acoustics of the source and the vocal tract. Therefore, SGRs are expected to remain relativly constant for a given speaker. SGRs have been extensively studied as speaker-specific acoustic features in dividing the vowel space into distinctive regions for many languages [3-6]. Several automatic estimation algorithms for SGRs are also developed [3, 7, 8]. Different applications of SGRs have been used in speech technology, including: body height estimation for adult speakers; speaker normalization for adult speakers; speaker verification under matched clean condition. The main goal of this thesis is to extend the applications of subglottal resonances to solve the following three problems: (1) SGRs for Mandarin speakers, (2) SGRs for children speech and height estimation and speaker normalization for children speech, and (3) speaker identification under noisy and mismatched condition.

1.2 The relationship between the supraglottal and subglottal system

In speech production, a linear time invariant model is composed of the following: the voiced input, which is the excitation from the glottis; the transfer function of the vocal tract; and the voice output, which is the speech signal. An illustration of the system is displayed in Figure1.2.

When acoustic coupling occurs between the vocal tract and subglottal system, additional zero-pole pairs occur. The zero-pole pairs derive from the natural frequencies of the lower airway right below the glottis, which refer to the subglottal acoustics. These frequencies are known as subglottal resonances.

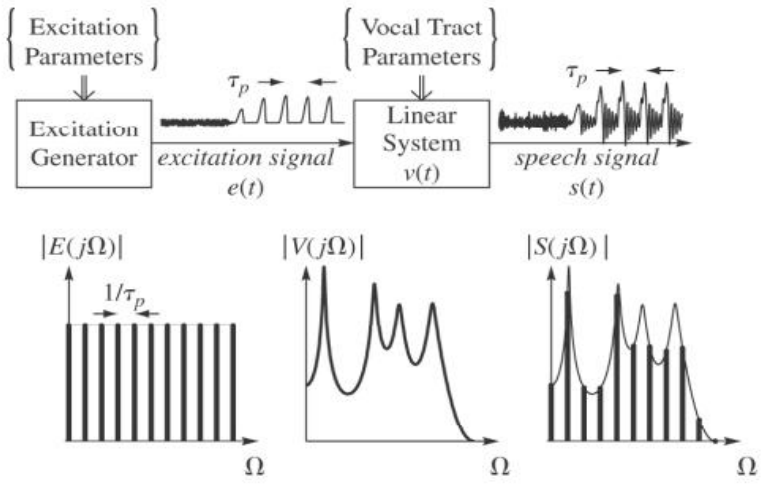

Figure 1.2: A linear time-invariant model of speech production

The subglottal resonances have been used in classifying the vowel space in the frequency domain into distinctive regions. The dimensions of the vowel space correspond to the first and second formant frequencies (F1 and F2) of the different vowels. The first subglottal resonances (Sg1) form the natural boundary of the high and low vowels; while the second subglottal resonance (Sg2) separates the front and back vowels, which are characterized by F2. The relationship between the first two subglottal resonances and vowel space are illustrated in Figure 1.3. For the third subglottal resonances (Sg3) there is no clear conclusion, even though some research showed that it forms the division between tense and lax vowels. These relationships have been found in several languages, including American English [3], German [4], Korean [5], and Hungarian [6]. All of these are non-tonal languages, and no research has been done on tonal languages, such as Mandarin. The influence of F0 on SGRs is interesting to study.

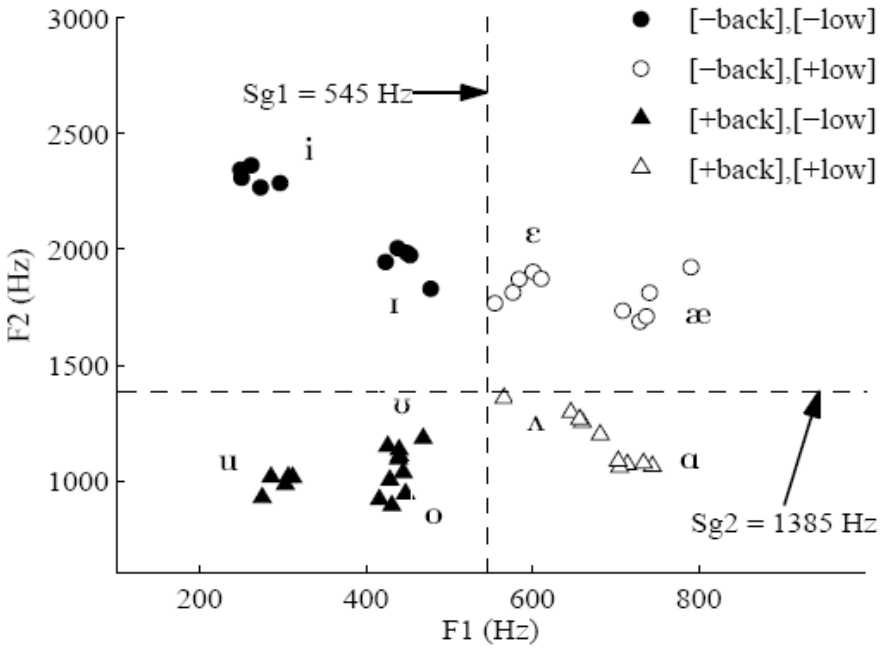

Figure 1.3: Vowel space in the F1-F2 plane demonstrating the vowel-feature contrasts provided by Sg1 and Sg2.

1.3 Automatic estimation of subglottal resonances

SGRs can be measured noninvasively using accelerometer recordings of subglottal acoustics. When held against the skin of the neck at the location of the place which is inferior to the thyroid cartilage, an accelerometer captures the signal which can represent subglottal acoustics, and thereby yielding a spectrum whose peaks occur near the SGR frequencies. However, since using accelerometers in real-life situations is unfeasible, it is important to find ways to automatically estimate SGRs from speech signals.

Previous research suggests several approaches to automatically estimating SGRs from speech signals: (1) direct estimation based on detecting the subtle effects of SGRs on vowel

formant frequency discontinuities and amplitude attenuations observed in the formant contours of back-to-front and low-to-high diphthongs [3, 9, 10]; (2) indirect estimation based on the potential correlation between SGRs and formant frequencies [7]. Among the previous methods, the algorithm in [10] gives the-state-of-the-art estimation results.

In [10], the algorithm was based on the following central idea: Sg1 acts as a boundary between high and low vowels so that two acoustic features characterizing vowel frontness – the Bark difference between the third and first formants (denoted as B_31) and the Bark difference between F1 and Sg1 (denoted as B_(1,s1)) – are correlated. Similarly, for Sg2 estimation, the Bark difference between F3 and F2 (denoted B_32) was found to be related to the bark difference between F2 and Sg2 (denoted as B_(2,s2)) since both measures characterize vowel backness. An empirical equation was derived to predict B_(1,s1) from a linear combination of the first three powers of B_31 and a constant term. The same approach was also applied to B_(2,s2) and B_32 to predict B_(2,s2). Sg3 is estimated based on its correlation with Sg2 using a first-order linear regression. These empirical relations allowed the first three SGRs to be estimated from a speech signal once the first three formants are tracked automatically.

The previous research doesn't develop a good SGR estimation algorithm for children speech. In this thesis, an age-dependent SGR estimation algorithm is proposed for children speech and used in different applications.

1.4 Height estimation using speech

Automatic height estimation, estimating the speaker's height from speech samples, could have potential applications in forensics, automatic analysis of telephone calls (e.g., 911 calls), and automatic speaker identification. A few studies have proposed automatic algorithms to estimate

adult speakers' height using speech signals. In [11], the first 19 MFCCs are extracted as features from speech signals and Gaussian mixture models (GMMs) were trained using data from all speakers in the TIMIT corpus [12]. With this approach, the height estimation error was found to be 5cm or less for 72% of the speakers, when training and testing used the same set of speakers. In [13], support vector machine (SVM) regression was proposed for height estimation. 6552 audio features from each utterance were extracted and then used in a feature ranking procedure to choose the top 50 features. The results yielded a mean absolute error (MAE) equal to 5.3cm. Although the algorithms yield reasonably good results using statistical modeling of speech features, but use a large number of features

In [10], an approach to height estimation is proposed based on the negative correlation between speaker height and SGRs. The algorithm only uses one of the first three SGRs as a 1 dimension feature to estimate height for adult speakers, and yields a MAE equal to 5.3cm. This method is likely to be more efficient than existing techniques in terms of the number of features required for height estimation.

In this thesis, automatic height estimation algorithm using SGRs is applied to children speech, since no research has been done in this area, an age-dependent height estimation algorithm is proposed for different age groups.

1.5 Speaker normalization for ASR

Most of the Automatic speech recognition (ASR) systems are speaker independent (SI). Inter-speaker variability poses a challenge to the design of SI-ASR systems. Inter-speaker acoustic variations are mostly caused by differences in the vocal tract, especially vocal-tract length. Typically, adult females have shorter vocal tracts compared to adult males, and children have

shorter vocal tracts compared to adults [14]. Hence, children tend to have higher formant frequencies than adults, and adult females tend to have higher formant frequencies than adult males. Figure 1.4 shows an example of the steady-state magnitude spectra of the vowel [i] of a male adult and a male child. Consequently, the ASR performance decreases significantly without using a speaker normalization technique.

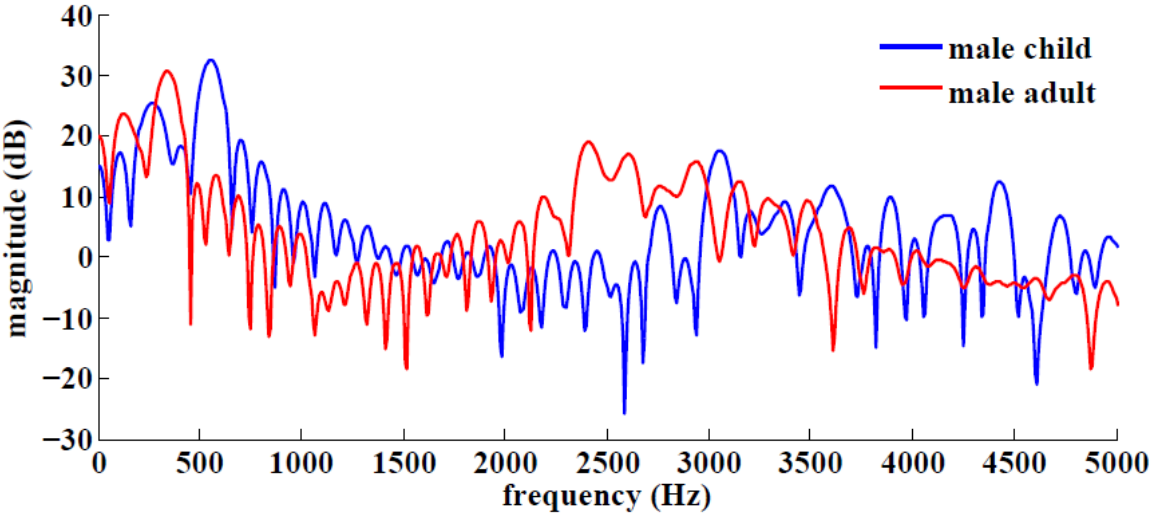

Figure 1.4: Comparing the steady-state spectra of the vowel [i] from a male adult and a male child speaker.

The effect of inter-speaker variability can be mitigated using speaker normalization. Speaker normalization is focused on frequency warping in the front-end feature domain. A widely-used approach is known as vocal-tract length normalization (VTLN), which involves a piece-wise linear function with a single parameter which controls the degree of spectral compression or expansion [15].

Frequency-warping parameters are typically estimated using the maximum-likelihood (ML) criterion. In order getter better performance under limited-data and noise conditions, SGRs have been used to compute warping factors (as ratios) for speaker normalization. The basic idea is to

map the spectrum of a target speaker to a reference speaker using SGRs. In [16], Sg2 has been used to compute warping factors. In [17], a nonlinear warping framework using the first three SGRs and F3 was proposed. The SGR-based speaker normalization method also shows robustness under noisy conditions [18]. The Figure 1.5 shows the frequency warping using SGRs.

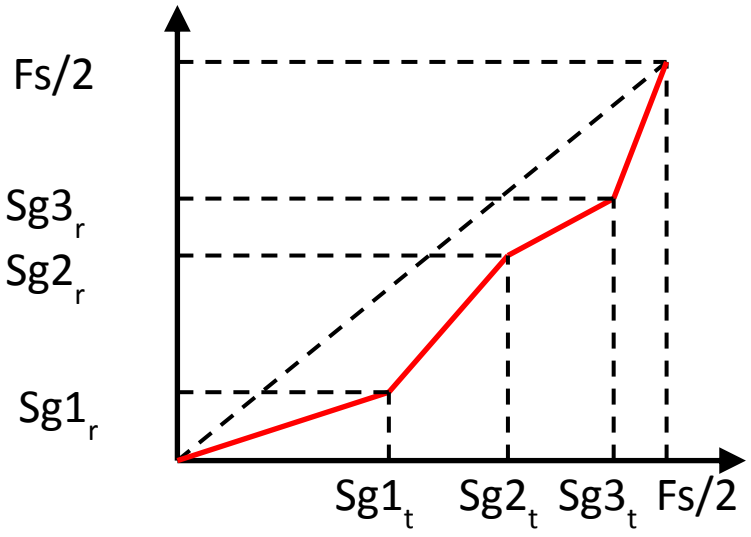

Figure 1.5: The warping functions maps the SGRs of a given target utterance to those of a reference speaker.

The previous SGR-based normalization techniques are more focused on adult speakers. Even though some research has been done for children speech normalization, there is no research that has been done on a large dataset of children speech. In this thesis, an age-dependent speaker normalization algorithm is investigated for children speech.

1.6 Speaker recognition

Speaker recognition refers to machine identification of the talker, which has found wide application in telephone-based financial transactions, voice-based user authentication, etc.

Speaker recognition involves two tasks: speaker identification (SID) and speaker verification (SV). In both tasks, the speech input can be either text independent or dependent. Most research is focused on text-independent tasks. For speaker recognition, extracting and modeling the speaker-dependent characteristics of the speech signal is very important.

Mel-frequency cepstral coefficients (MFCCs), which capture the acoustics of the supraglottal vocal tract, have been widely used for speaker recognition from clean speech [19]. MFCCs have been shown to be an effective feature set with a number of backend modeling schemes such as GMMs adapted from universal background models (GMM-UBM) [20], i-vectors frameworks [21]. Under noise condition, some noise robust features have been proposed, such as high order linear prediction cepstral coefficients (LPCCs) [22] and power-normalized cepstral coefficients (PNCCs) [23].

Previous research investigated the utility of subglottal acoustic features for speaker recognition. In [19], the cepstral features of subglottal signal are estimated from speech and are used in SV under clean conditions. In this thesis, the noise robustness characteristics of SGRs are used as front-end features for speaker identification under noisy conditions.

1.7 Thesis outline

The rest of this thesis is organized as follows.

Chapter 2 describes the new database for the study of subglottal acoustics for Mandarin speakers. It also presents small important results of data analysis to motivate the algorithms developed in the following chapters.

Chapter 3 presents the study of SGRs for children speakers. An automatic algorithm is developed to estimate SGRs for different age groups. The new estimation algorithm is applied to height estimation and speaker normalization for children speech.

Chapter 4 proposes a noise robust speaker identification algorithm, which investigates using SGRs as noise robust features for SID. SGRs are used to provide complementary information to LPCCs and PNCCs in SID tasks. Experimental results are reported for two standard databases of adults' speech.

Chapter 5 summarizes the key results and provides directions for future work.

Chapter 2

The analysis of subglottal resonances for Mandarin speakers

This chapter starts with a description of the Mandarin corpus. Then some important analysis of SGRs for Mandarin speakers is studied and the corresponding results are presented.

2.1 The UCLA Mandarin corpora

The Mandarin vowel system contains 6 standard vowels, which are [a], [o], [e], [i:], [u], [ü]. A short, retroflex [i] is also studied and distinguished from the long [i:]. Therefore there are 7 vowels in total in our corpus, as illustrated in Table 2.1. The place-of-articulation feature [+/− back], specifying the tongue position, is also shown in Table 2.1. [+] indicates that the tongue dorsum bunches and retracts slightly to the back of the mouth, while [−] indicates that the tongue extends slightly forward.

Table 2.1: Phonological classification of Mandarin vowels.

Vowel	i:	i	ü	e	a	u	o
[+/-Back]	-	-	-	central	central	+	+

As a tonal language, there are four tones in Mandarin, which are flat, rising, falling-rising and falling, as illustrated in Table 2.2. Table 2.2 also shows the symbols of the four tones, which indicates the corresponding F0 trajectory. The vowels in our database are in a ‘pV’, ‘shV’ or ‘xV’ context. For each context and each tone, we used a corresponding Pinyin (the official phonetic system for transcribing the Mandarin pronunciations of Chinese characters into the Latin alphabet) and a Chinese character, as illustrated in Table 2.3. Each character was embedded in the carrier phrase, “我把__的说一遍” (“I said __de once”(English version), “wǒ bǎ__de shuō yi biàn” (Pinyin version)).

Table 2.2: Four tones in Mandarin.

Tones	First	Second	Third	Fourth
Pitch	Flat	Rising	Falling-	Falling
Symbol	ˉ	ˊ	ˇ	ˋ

Table 2.3: The Mandarin corpus.

Vowel	i:	i	ü	e	a	u	o
Pinyin	(xī)	(shī)	(xū)	(shē)	(shā)	(shū)	(pō)
	(xí)	(shí)	(xú)	(shé)	(shá)	(shú)	(pó)
	(xǐ)	(shǐ)	(xǔ)	(shě)	(shǎ)	(shǔ)	(pǒ)
	(xì)	(shì)	(xù)	(shè)	(shà)	(shù)	(pò)
Character	吸	诗	虚	奢	沙	书	坡
	席	石	徐	蛇	啥	熟	婆
	喜	史	栩	舍	傻	鼠	叵
	细	事	旭	设	厦	竖	破

Acoustic data were collected with simultaneous speech and subglottal recordings for 20 native speakers of Mandarin (10 males and 10 females). Speech data were recorded using a Shure PG27 condenser microphone and subglottal data were obtained using a K&K Sound 'Hot Spot' accelerometer attached to the skin of the neck below the thyroid cartilage. All recordings were sampled at 48 kHz and digitized at 16bits/sample. All speakers, aged between 18 and 24 years, were recorded in one session, and every word (character) was recorded at least 4 times. Speaker height (standing height) and sitting height (trunk length) were measured before recording, with the sitting height being measured from the speakers' hip bone to top of the head. In this corpus, speaker height ranged from 165-195cm for males, and from 150-170cm for females. Speaker sitting height ranged from 73-81cm for males, and from 63-77cm for females.

2.2 Analysis methods

2.2.1 Measurements

The first two formants of each vowel were measured from the microphone signals in their steady-state regions, and Sg2 was measured from the corresponding accelerometer signals, using

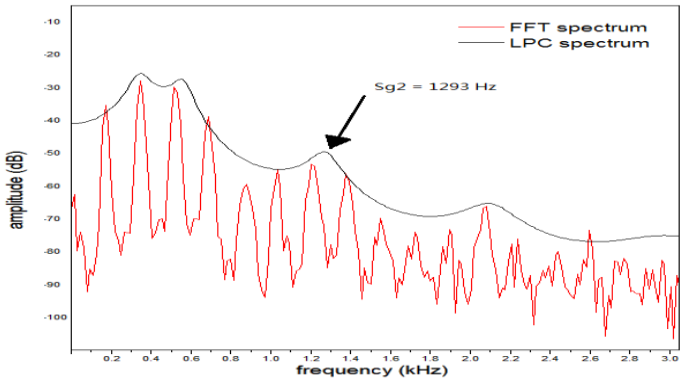

Figure 2.1: FFT spectrum and LPC spectrum for a sample accelerometer signal taken from the Mandarin corpus

Wavesurfer [24]. Both signals were down-sampled to 8000 Hz before analysis. Before measuring Sg2 from accelerometer signals, each vowel segment was divided into three parts: beginning, middle and end. The reason is that the three different parts representing different F0 regions help us to study the correlation between Sg2 and F0. Sg2 measurements of each part were acquired by visual inspection of the resonance peaks in LPC spectra at the middle point of the corresponding part, as illustrated in Figure 2.1.

2.2.2 The relationship between Sg2 and vowel class

To investigate whether Sg2 divides the vowel space in Mandarin, means (Mean) and standard deviations (Std) of Sg2 measurements were calculated for each speaker. A frequency range from Mean-Std to Mean+Std was chosen to represent the Sg2 interval for each speaker as in [25], since we hypothesized that Sg2 measurements might exhibit increased variance during the large pitch excursions typical of Mandarin tones. For each vowel and each tone, the average values of the first two formants measured from the four respective repetitions were used. The Sg2 interval was then compared with the second formant (F2) for all seven vowels and four tones in order to test whether Sg2 defines the boundary of front and back vowels.

2.2.3 The relationship between Sg2 and standing height and sitting height (trunk length)

Mean Sg2 was used to investigate correlations between Sg2 and both standing and sitting height measurements. Mean Sg2 values along with standing and sitting height and gender are illustrated in Table 2.4. IDM is the abbreviation for speaker ID (male), IDF for speaker ID (female), STH (cm) for standing height, SIH (cm) for sitting height and SG2 (Hz) for Sg2. The data in Table 2.4 are in line with what we found before for English [2].

Table 2.4: Mean Sg2 values along with standing and sitting height and gender for each speakers. IDM (speaker ID (male)), IDF (Speaker ID (female)), SG2 (Sg2 in Hz), STH (standing height in cm) and SIH (sitting height in cm)

IDM	SG2	STH	SIH	IDF	SG2	STH	SIH
1	1292	177	77	2	1458	162	70
11	1240	171	77	3	1348	170	67
9	1193	195	81	18	1595	150	65
7	1342	165	75	12	1489	164	65
17	1355	170	73	16	1515	153	68
4	1332	178	78	6	1494	153	63
13	1354	174	74	8	1371	157	77
19	1268	175	76	10	1470	159	69
5	1268	175	76	14	1473	151	68
15	1336	171	76	20	1460	163	73

2.2.4 The relationship between Sg2 and F0

In the study of possible F0-Sg2 interaction, we chose the 3 corner vowels [u], [i:] and [a]. For every vowel, Sg2 measurements from 10 speakers (5 males and 5 females which are randomly chosen but cover the height range in the database), 4 tones and 4 utterances were selected. For each utterance, Sg2 measurement of 3 parts (beginning, middle and end of the vowel) were used for study.

Hypothesis testing [26] was applied to detect possible F0 influence on Sg2 using paired t-tests. The null hypothesis was that the mean difference between Sg2 in two parts of a vowel (beginning and middle, middle and end, beginning and end) is zero.

Since the paired t-test requires the difference between pairs to be normally distributed, it was necessary to verify this property of our data [27]. Figure 2.2 shows the 8 class histogram of Sg2 difference value between the middle and end of the vowel [u] with the first tone. The

vertical line denotes the mean. Similar to the case in Figure 2, most of the data fit a normal distribution.

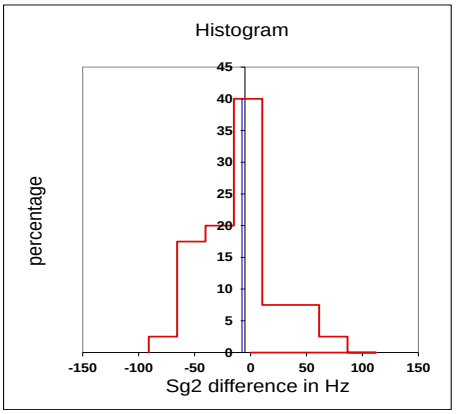

Figure 2.2: Histogram of Sg2 difference for all 10 selected speakers between beginning and middle of vowel[u], first tone.

2.3 Analysis results

2.3.1 Sg2 and distinctive feature [+back] and [-back]

Two speakers (denoted as Female3 and Male11) were selected to show the division of vowel space by Sg2, as representative examples. Vowel plots for each speaker are shown in Figure 2.3. The plots are the sample for a given vowel regardless of the tones. The mean Sg2 is indicated by the solid horizontal lines and the upper and lower dashed lines indicate the Sg2 interval bounded by the values Mean+Std and Mean-Std.

In general, the vowel space is divided by Sg2, with front vowels on one side of Sg2 and back vowels on the other side. This result is consistent with previous studies [3-6].

Based on our observation of average Sg2 of 10 male and 10 female Mandarin speakers, we should note that the location of the central vowels which are close to Sg2 in vowel space such as

[a] and [e], varies considerably for different speakers. For some speakers the central vowels show up above the upper boundary of the Sg2, while in some other cases they are located below the lower boundary of Sg2.

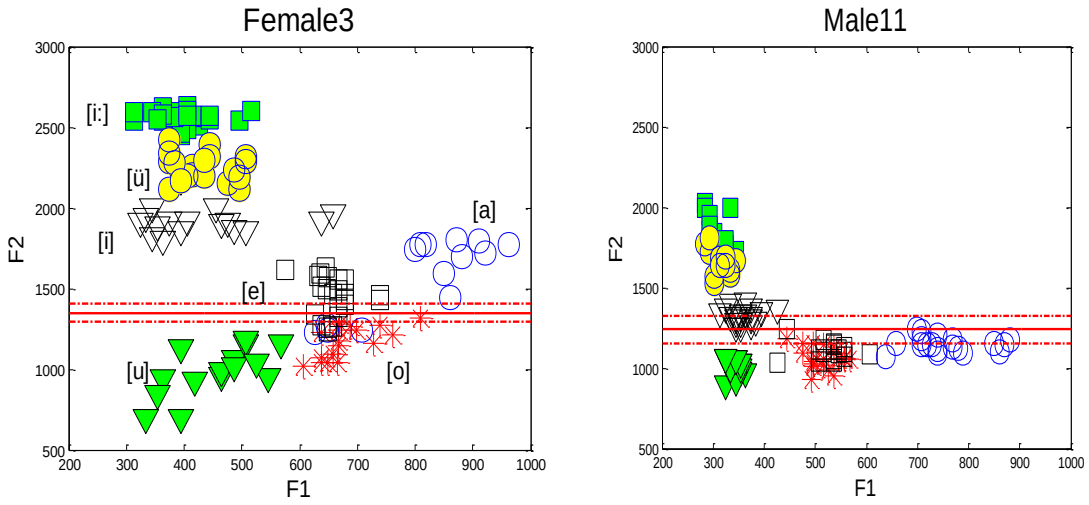

Figure 2.3: Vowel plots of the two speakers (Female3 and Male11). Horizontal dashed lines indicate Sg2 interval (Mean ±Std). Different symbols represent different vowels. The vowel identities are labeled in the vowel plot for speaker Female3.

2.3.2 Correlation between Sg2 and speaker's standing height and sitting height

The scatter plot of height and sitting height versus Sg2 (mean Sg2 for given speaker) is shown in Figure 2.4 (all 20 speakers).

The correlation coefficients for Sg2 (mean Sg2 for a given speaker) with standing height and sitting height were also calculated for 10 male speakers, 10 female speakers and all 20 speakers respectively, as illustrated in Table 2.5.

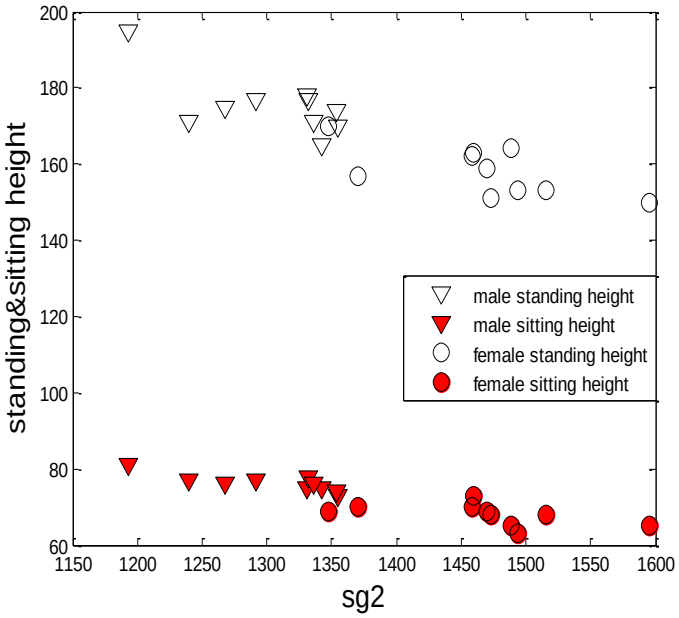

Figure 2.4: Plots of the Sg2 versus standing and sitting height for male and female speakers

Table 2.5: Correlation between Sg2 and standing and sitting height for male, female and all speakers. (All correlations are significant)

Gender	Males	Female	Combined
Correlation	-0.6776	-0.6602	-0.8699
Correlation(sitt	-0.8063	-0.5488	-0.8733

For male speakers, the inverse correlation of Sg2 with sitting height is stronger than the correlation between Sg2 and standing height. A possible reason for this is that SGR frequencies are actually determined primarily by the 'acoustic length' (effective length of the subglottal system) [28]. Physiologically, the 'acoustic length' may be expected to be correlated with the size of the lungs and the length of the trunk, Sg2 is likely to be more strongly correlated with sitting height (trunk length). According to physiological data reported in [29], trunk length itself

appears to be moderately correlated with overall body height. Such a relationship between trunk length and standing height seems to be partly responsible for the relatively weaker correlations with standing height [8]. In essence, the result that Sg2 has stronger correlation with sitting height but relatively weaker correlation with standing height for male speakers is reasonable.

As for the female speakers, the opposite trend is observed: correlation with sitting height is weaker compared to correlation with standing height. This can be explained by two interesting cases. Speaker 3 with a 1348Hz average Sg2 has a 170cm standing height but a 67cm sitting height, while Speaker 8 with a 1371Hz Sg2 has a 157cm standing height and a 77cm sitting height, which indicates that one is tall but has a short trunk length, while the other speaker is short but has a relatively long trunk length.

Overall, the correlation coefficients calculated by using all data from the 20 speakers are high for both standing height and sitting height cases. The correlation in the sitting height case is somewhat higher than standing height.

2.3.3 Relationships between F0 and Sg2

Table 2.6 shows the t value of Sg2 difference and their upper boundary. If the t value is less than the upper boundary, we accept the null hypothesis, i.e. the difference of Sg2 is zero. As shown in the table, under a significance level of 0.01, there is no statistically significant Sg2 difference between different parts of the vowel for all three vowels and all four tones.

Under the assumption that three measurements sufficiently capture change of Sg2 over time, we find no statistically significant Sg2 variation in any tone of any of vowel. Even if we loosen the significance level to 0.05 and upper bound of t reduces to 2.022, there are only 2 outliers, while in most cases we still accept the null hypothesis. A higher t value denotes a larger

difference between mean Sg2 of different parts. And we notice that the fourth tone has the widest F0 range, whereas its t values are not significantly higher than the other tones. Therefore it can be inferred that despite the large range of F0 variation within a vowel, Sg2 tends to stay at about the same value.

Table 2.6: The results of t-test of Sg2 comparisons between different vowel parts. (significance level of 0.01)

Tone	Parts of	t(uppe	t(/u:/	t(/i:/)	t(/a:/
First	Begin-	2.708	1.28	1.90	1.07
	Mid-End		0.26	0.88	0.10
	Begin-		1.21	0.90	0.10
Second	Begin-		0.59	0.53	1.43
	Mid-End		2.05	1.04	0.95
	Begin-		1.67	1.33	1.73
Third	Begin-		1.25	0.60	1.19
	Mid-End		1.84	0.74	1.52
	Begin-		1.30	1.28	2.06
Fourth	Begin-		0.84	0.67	0.28
	Mid-End		0.23	0.88	0.64
	Begin-		1.10	1.28	0.15

2.4 Conclusions

Our results illustrate that, as in other languages studied before, in Mandarin, Sg2 is the boundary between front and back vowels. It is also important to note that the neutral vowels in Mandarin are not always located in the Sg2 interval and they could be located outside the Sg2's upper and lower boundary.

We also showed that, Sg2 is more inversely correlated with sitting height (trunk length) than standing height for our database. This result is more concrete for male speakers, since there were 2 female speakers who have an unexpected ratio of trunk length-to-standing height.

The result of the correlation between Sg2 and F0 is also illustrated. Paired t-test was used to test the Sg2 of different regions of the vowel, which represent different F0 values. We illustrated that there is no statistically-significant difference between Sg2 with different F0 values within a vowel.

For future work, further interaction and correlation between SGRs, formants and F0 will be studied in Mandarin, such as the coupling effect between subglottal acoustic system, source model and vocal tract. The results in this chapter will also have application in automatic speech recognition and height estimation for Mandarin speakers.

Chapter 3

The analysis and applications of subglottal resonances for children speech

This chapter starts with the analysis of SGRs for children speech. Then an age-dependent estimation algorithm for SGRs is proposed. Two applications for children speech are investigated, which are height estimation and speaker normalization.

3.1 Analysis and automatic estimation of the first three SGRs

3.1.1 WashU-UCLA Children Dataset

The WashU-UCLA child corpus comprises simultaneous recordings of microphone and subglottal accelerometer signals from 46 child speakers (33 males, 13 females) of American English. The speakers are aged between 6 and 17 years: 24 speakers were between the age of 6 and 11 years (18 males, 6 females), 22 were between the age of 11 and 17 (15 males, 7 females). Every speaker was recorded in two sessions: one with 14 hVd words (10 monophthongs – in which we include the approximant [ɹ] – and 4 diphthongs) and the other with 21 CVb words (4 monophthongs and 3 diphthongs, in three different consonant contexts). Every word, embedded in the carrier phrase, “I said a ____ again”, was recorded repeatedly until each child successfully

said the sentence at least 3 times. Only the monophthong hVd words and the corresponding carrier phrases were used in this study. Moreover, speaker height was recorded in the corpus and ranged from 105cm to 182cm.

3.1.2 Analysis of SGRs for children speech

SGR analysis was conducted on all the recordings of the 10 monophthongs: 2760 microphone recordings and 2852 subglottal recordings. For each speaker, the first three formants were measured from the microphone signals in the steady-state region using Snack [30]. The first three SGRs were measured from the corresponding accelerometer signals by visual inspection of the resonance peaks in LPC spectra using Wavesurfer [20]. Both microphone and accelerometer signals were down-sampled to 8kHz before analysis.

Previous research [2] showed that Sg1 acts as a boundary between high and low vowels and Sg2 forms a boundary between front and back vowels for adult speech, and this chapter investigated whether Sg1 and Sg2 divided the vowel space for children's speech as well. Table 3.1 shows the percentage of speakers in which Sg1 and Sg2 successfully divided the vowel space. The percentages are high indicating that SGRs divide the vowel space of children as well as adults.

Table 3.1: Percentage of speakers, separated by age group, whose SGRs successfully divided the vowel space.

Age Group	Below 11	Above 11	All speakers
Sg1	87.5%	95.5%	91.3%
Sg2	91.6%	95.5%	93.5%

To investigate the relationship between Sg1, Sg2 and Sg3 for children's speech, scatter plots of Sg3 versus Sg1 and Sg2 are shown in Figure 3.1. The results indicate that Sg3 is correlated with Sg1 (r=0.88) but more strongly correlated with Sg2 (r=0.92). Therefore, a first-order linear regression was trained using Sg2 and Sg3 and the result is Eq. 3.1.

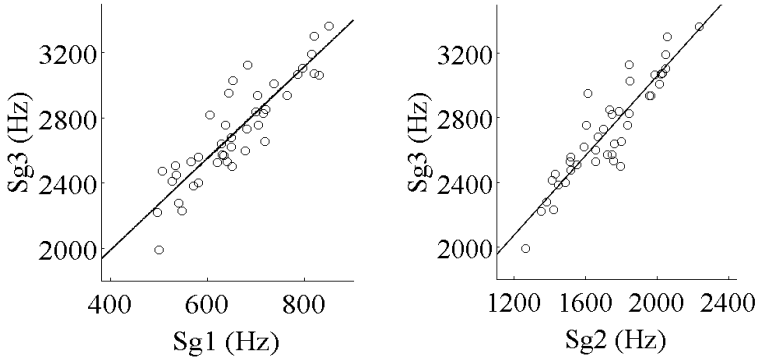

Figure 3.1: Scatter plots of Sg3 vs. Sg1 (left) and Sg3 vs. Sg2 (right). Also shown are first-order linear regression. Sg1 and Sg3 are correlated (r=0.88) while Sg2 and Sg3 are more strongly correlated (r=0.92).

$$\mathrm{Sg3} = 1.233(\mathrm{Sg2}) + 593.424 \qquad (3.1)$$

3.1.3 Automatic estimation of SGRs for children speech

Estimation algorithms for the first three subglottal resonances were proposed for adults in [8]. The algorithm was based on the following central idea: Sg1 acts as a boundary between high and low vowels so that two acoustic features characterizing vowel frontness – the Bark difference between the third and first formants (denoted as B_{31}) and the Bark difference between F1 and Sg1 (denoted as $B_{1,s1}$) – are correlated. Similarly, for Sg2 estimation, the Bark difference between F3 and F2 (denoted B_{32}) was found to be related to the bark difference between F2 and Sg2 (denoted as $B_{2,s2}$) since both measures characterize vowel backness. An empirical equation was derived to predict $B_{1,s1}$ from a linear combination of the first three powers of B_{31} and a

constant term. The same approach also applied to $B_{2,s2}$ and B_{32} to predict $B_{2,s2}$. Sg3 is estimated based on its correlation with Sg2 using a first-order linear regression, as in Eq. 3.1. These empirical relations allowed the first three SGRs to be estimated from a speech signal once the first three formants are tracked automatically.

A previous study [31] derived empirical relations to estimate Sg1 and Sg2 for children's speech, but the dataset was relatively small, and the estimation algorithm for Sg3 was not investigated. In this study, all the empirical relations to estimate the first three SGRs were derived using a larger dataset of 46 speakers in the WashU-UCLA child corpus.

When training the regression model using the data from all the speakers together, the results showed a relatively low r-squared (r^2) value. However, when we separated the speakers into two different age groups, below 11 and above 11, both of the regression models trained on each group resulted in larger values of r^2, as illustrated in Table 3.2.

Table 3.2: R-squared values for the SGR estimation models of Sg1 and Sg2 when trained on speakers separated by age group, as well as when trained on all speakers.

Age Group	Below 11	Above 11	All speakers
r^2 for Sg1	0.91	0.92	0.85
r^2 for Sg2	0.91	0.93	0.85

Therefore, we train and test the SGR estimation algorithms separately for the two different age groups using a cross-validation method. Within each age group, each time we randomly chose around 60% of the speakers to train the regression model and the rest to test the estimation algorithm. Given a test speech signal, the detailed steps involved in estimating SGRs are the same as in [8].

3.1.4 Performance analysis of the algorithm

The SGR estimation algorithm was evaluated using two performance metrics: the mean and standard deviation of the root mean squared errors (across speakers and 5 cross-validation tests), denoted as μ_{rms} and σ_{rms}, respectively, both in units of Hz. Table 3.3 shows the performance of the automatic estimation algorithm in each age group and the whole dataset.

Noted that in applications, when age information is not available, broad age group estimation algorithms should be used before estimating SGRs. In section 3.3 of this chapter, average F3 is used as a threshold to predict the age group of each speaker when estimating the SGRs for speaker normalization.

The best regression models to estimate Sg1 and Sg2 for younger children during cross-validation are presented by Eq. 3.2 and Eq. 3.3. Eq. 3.4 and Eq. 3.5 present the best models for older children. The regression equations of Sg3 for both age groups are dependent on Sg2 in a similar way as in Eq. 3.1. Therefore, Sg3 for both age groups can be estimated from the Sg2 estimates using Eq. 3.1.

Table 3.3: Mean and standard deviation of RMS error, in Hz, of SGR estimation for the set of 'younger' children (Y), 'older' children (O) and both sets 'combined' (C).

	Sg1			Sg2			Sg3		
	Y	O	C	Y	O	C	Y	O	C
μ_{rms}	48	53	51	131	126	128	170	166	168
σ_{rms}	31	35	34	70	66	69	81	77	79

$$B_{1,s1} = 0.0002(B_{31})^3 + 0.003(B_{31})^2 - 0.907(B_{31}) + 8.310 \qquad (3.2)$$

$$B_{2,s2} = -0.011(B_{32})^3 + 0.184(B_{32})^2 - 1.870(B_{32}) + 5.290 \quad (3.3)$$

$$B_{1,s1} = -0.001(B_{31})^3 + 0.011(B_{31})^2 - 0.776(B_{31}) + 6.601 \quad (3.4)$$

$$B_{2,s2} = -0.003(B_{32})^3 + 0.062(B_{32})^2 - 1.534(B_{32}) + 4.477 \quad (3.5)$$

3.2 Height estimation for children speakers

3.2.1 Methods

Previous work on adult height estimation using speech signals has shown a strong negative correlation between the three SGRs and height [8]. This section tests a similar hypothesis for children. Using measurements for Sg1, Sg2 and Sg3 for each speaker, as well as information about the speakers' heights, a scatter plot of height versus the SGRs for all children is presented in Figure 3.2.

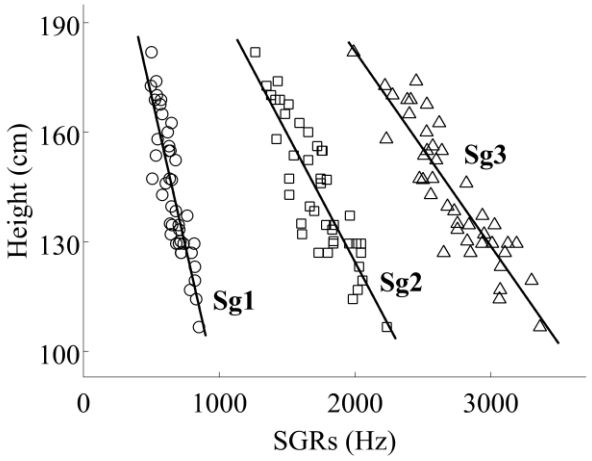

Figure 3.2: Scatter plots of all child speaker height vs. each of the first three SGRs. Also shown are first-order linear regression fits. Speaker height correlates strongest with Sg3 (r=-0.90), but is also correlated with Sg1 (r=-0.88) and Sg2 (r=-0.88).

The inverse correlation of each SGR with height is strong, and therefore a first-order linear regression is used to model the relationship between each SGR and height. The empirical relations were obtained between speaker height and SGR frequencies, as illustrated in Eqs. 3.6, 3.7 and 3.8. These equations are different from the height estimation equations for adults in [8].

$$h = -0.166(Sg1) + 254.497 \text{ (r=-0.88)} \quad (3.6)$$

$$h = -0.070(Sg2) + 264.793 \text{ (r=-0.88)} \quad (3.7)$$

$$h = -0.053(Sg3) + 288.821 \text{ (r=-0.90)} \quad (3.8)$$

3.2.2 Experiments and results

Motivated by the results of 3.1, the height estimation algorithm was tested used a cross-validation method in which the child speakers were grouped into two different categories: age under 11 years and age above 11 years. In each category, 60% of the speakers were chosen to train the first order linear regression model between height and each empirically measured SGR (ground truth), and the rest were used to test the model. In each age group, after the models were trained, using the method proposed in Section 3.1, Sg1, Sg2 and Sg3 were estimated for each testing speech signal. Finally, the trained linear regressions between SGRs and height, along with the three computed SGRs from the test data, were used to estimate the speakers' heights, and the results were compared with the actual height measurements. The height was calculated for each voiced frame, and the estimated height for each test speaker was the average number across all frames.

This procedure was repeated 5 times for each age group, and each time, the root mean squared errors (RMSE) and mean average errors (MAE) were recorded. Table 3.4 displays the average RMSE and MAE of this experiment for both age groups.

Additionally, to verify the necessity of the age-dependent SGR estimation model, the experiment was repeated again but with all child speakers grouped into a single category. Estimation of Sg1, Sg2 and Sg3 for the speakers used a model obtained in a similar way as Eqs. 1-5 in 3.1 but trained assuming age-independence of SGRs. The average RMSE (cm) and MAE (cm) of this experiment are also shown in Table 3.4.

Table 3.4: Mean average error and root mean squared error of the height estimation algorithms when trained and tested on the set of 'younger' children (Y), 'older' children (O) and 'all' children (A).

	Using Sg1			Using Sg2			Using Sg3		
	Y	O	A	Y	O	A	Y	O	A
MAE	3.8	5.0	9.4	4.3	4.9	10	4.3	4.9	11
RMSE	4.8	6.2	10	5.9	6.5	11	6.0	6.6	12

The resulting regression equations of height versus each SGR during cross-validation training were similar to Eqs. 3.6, 3.7 and 3.8 for both the younger and older groups, and therefore, Eqs. 3.6, 3.7 and 3.8 can be used to estimate height using SGRs regardless of age. However, RMSE and MAE were smaller when using different SGR estimation models for different age groups, suggesting the necessity for age-dependent SGR regression models. Thus, the height estimation algorithm can simplify to age-dependent SGR estimation models in combination with age-independent linear regressions of height versus SGRs. Observing the values in Table 3.4 reveals that Sg1 returns the most accurate height estimation (MAE of 3.8cm)

for children under the age of 11 years, while Sg2 and Sg3 return the most accurate height estimation (MAE of 4.9cm) for children above 11. Note that the height estimation error for older children is similar to that of adult speech [8].

3.3 Speaker normalization for automatic speech recognition

3.3.1 Methods and algorithm for comparison

Motivated by the success of the age-dependent SGRs estimation algorithm and the results on height estimation, we investigated speaker normalization using the new age-dependent framework. The SGR warping scheme is the same as in [17]: the test speakers' SGRs are warped onto a reference speaker's SGRs, and in case of errors in SGR estimation, scaling factors were used to fine-tune the SGRs in a maximum likelihood approach similar to that used in VTLN techniques. We have shown in the previous sections that SGRs are estimated differently for the age groups below and above 11. So, for normalization we estimated the age group of the speaker by thresholding the average F3 for each speaker. Since the effect of Sg3 for normalization has not been clearly studied before, we include Sg3 in the experiments. The various experiments performed using the estimated SGRs are: (1) age-independent Sg2 warping (2) age-independent Sg3 warping (3) age-independent {Sg1, Sg2, Sg3} warping, (4) age-dependent Sg3 warping using oracle age information and (5) F3 based age-dependent Sg3 warping (using F3 as a threshold to predict different age groups). We focus primarily on Sg3 because initial experiments showed that Sg3 yields best results. We have also compared the results of these experiments with the CVTLN and age-independent {Sg1, Sg2, F3} warping in a previous paper [17].

3.3.2 Normalization experiment and results

The automatic speech recognition (ASR) system used for our experiments was trained using adult speech and tested on SGR-warped children's speech. The TIDIGITS database was used for both adult speech (training) and children's speech (testing). The features used are the first thirteen Mel-frequency cepstral coefficients (MFCCs c0–c12) and their first and second-order derivatives computed using 25ms frames spaced at 10ms intervals. All signals are down sampled to 8kHz. The training and testing sets comprise data from 112 adults (55 males, 57 females) and 50 children (25 boys, 25 girls; 6–15 years old), respectively. Monophone hidden Markov models (HMMs) are used for recognition. The HMMs have 3 emitting states each, and each state has 6 Gaussian components.

Normalization is applied only to the testing data and not to the training data. The reference SGRs used in our experiment were obtained by taking the average of all the estimated SGRs of the adult speakers in the training set, which were Sg1ref =604.9Hz, Sg2ref=1357.4Hz and Sg3ref =2228.3Hz. The F3 used for separating the speakers into the 2 age groups was 3kHz.

The hidden Markov model toolkit (HTK) was used for all experiments, and word error rate (WER) was used as the performance metric. Results for all our experiments are shown in Table 3.5.

The results show that the experiments using only Sg3 produce the lowest WERs followed by the Sg2 and Sg1 warping schemes. One possible reason is that, as it has been shown in Section 3.2 that Sg3 has the strongest correlation with height, Sg3 may also have strong correlation with the vocal tract length (VTL). The combination of Sg1, Sg2 and Sg3 gives a WER that lies between that obtained by using only Sg3, only Sg2 and only Sg1. Among the

warping schemes involving only Sg3, the highest WER is obtained using the age-independent Sg3 estimation. The lowest WER occurred when using oracle age information to estimate the SGRs (~26% WER reduction relative to CVTLN and {Sg1, Sg2, F3} warp). Automatic

Table 3.5: Word error rates (%) for ASR experiments.

Experiment Type	WER (%)
Baseline	9.9
CVTLN	2.7
{Sg1,Sg2,F3} warp	2.7
Age-independent Sg1	3.4
Age-independent Sg2	2.8
{Sg1,Sg2,Sg3} warp	2.8
Age-independent Sg3	2.47
Age-dependent Sg3 using F3-based age estimation	2.09
Age-dependent Sg3 using oracle age information	1.96

estimation of age group using F3 also produced WER lower than the CVTLN, {Sg1, Sg2, F3} with results comparable to the oracle age-dependent Sg3 warping scheme. Though F3 alone is not a perfect measure to estimate age, it has been observed that it is good enough to roughly separate the speakers into 2 age groups to estimate SGRs.

3.4 Conclusions

In this chapter, an age-dependent scheme for automatic height estimation and speaker normalization is proposed for children's speech. Analysis indicates that children below and above 11 years old show different acoustic properties, and therefore, an automatic age-dependent SGR estimation algorithm is applied to each age group. The first three SGRs were estimated using a similar method adapted from adult speech but with age dependency considerations. Good results were achieved for each SGR. Using the algorithm for estimating SGRs and the inverse

relation between SGRs and height, speaker height can be automatically estimated. The proposed height estimation algorithm performs well in each age group. Using a cross-validation method, speaker height can be estimated to within 3.8cm for younger children and 4.9cm for older children, on average. Motivated by the differences between each age group, a linear frequency warping method using age-dependent Sg3 was applied to the TIDIGITS speech recognition task. The results show that the proposed method outperforms CVLTN and other SGR-based warping schemes.

For future work, we will evaluate the effectiveness of the algorithm on a larger database. Moreover, age estimation for children's speech using SGRs will also be studied.

Chapter 4

Noise robust speaker identification using subglottal resonances

4.1 Proposed framework

We propose a two-stage frame work to fuse the information provided by SGRs and the cepstral features.

Traditional score level combination of two separate feature systems doesn't work for several reasons. Firstly, since SGRs are relatively constant features of low dimension, generative models like GMMs are probably not a good choice for speaker models based on SGRs. Instead, a discriminate classifier, such as multilayer perception (MLP), may give better performance. Moreover, since SGRs have negative correlation with speakers' height, speakers similar in height might have similar SGRs. Using SGRs to perform identification tasks among a large amount of speakers is not discriminative enough. However, SGRs may work much better within a small set of speakers.

Therefore, during the first stage of the proposed system, we use the cepstral features (PNCC&LPCC) as the front-end feature to find the top N most likely speaker models for a test utterance. Within these N speakers, the SGRs are used as the new features in the second stage. MLP is used as the classifier to retrain the speaker models and generate new scores for these N

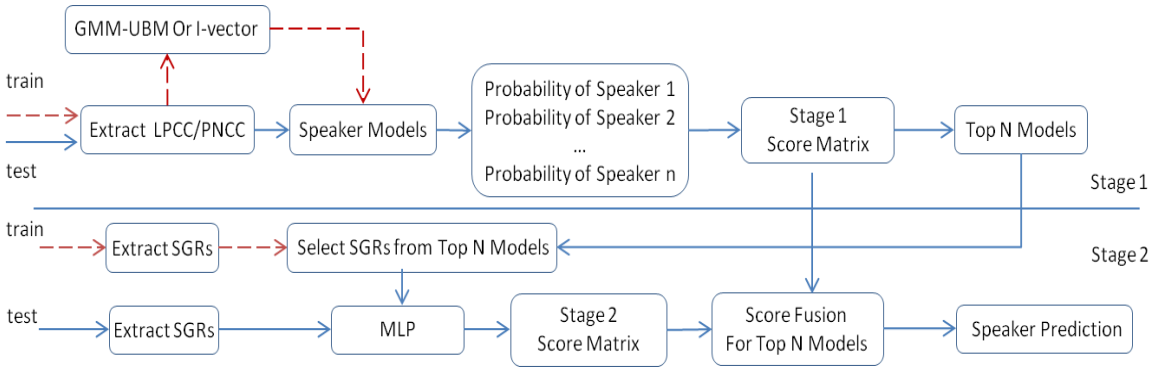

Figure 4.1: System flow chart of the proposed speaker identification algorithm.

speaker models with respect to the corresponding test utterance. The cepstral and SGRs scores of the N speakers are then combined in a weighted fashion and the combined scores are used to make the final decision. An overview of the proposed framework is presented in Figure 4.1 and the implementation details are provided in Section 4.3.

4.2 SGRs estimation

This section provides the details about the SGRs estimation algorithm and the estimation results on the large databases.

4.2.1 Estimation algorithm

The SGR estimation algorithm is based on the algorithm proposed in [8]. The algorithm is derived on certain well-established phonological relations between SGRs and formant frequencies of our previous research [2, 8, 17, 18], and is known to be reasonably accurate in estimating the first three SGRs on WashU-UCLA datasets [32]. One of our goals here is to test whether the algorithm gives consistent estimates under different noise conditions on large SID datasets.

4.2.2 Results

Four additive noises (i.e., babble, factory, pink and white) collected from the NOISEX-92 database were used for representing different noise conditions. The speech segment was degraded by adding a specific type of noise at SNRs of 5, 10, 15, 20 dB, respectively, using FaNT [33]. SGRs for all 630 Speakers in TIMIT are estimated based on one clean utterance per speaker, as well as the same utterance with additive noise of different SNRs.

To quantify the SGR estimation accuracy for all speakers, RMSEavg and ith speaker's RMSE for the Kth SGR (K = 1, 2, 3) are defined as follows. Denoting the number of speakers as N, and the number of utterances for the ith speaker as Mi,

$$\mathrm{RMSE}_{\mathrm{avg}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{RMSE}^{\mathrm{i}} \tag{4.1}$$

$$\mathrm{RMSE}^{\mathrm{i}} = \sqrt{\frac{1}{\mathrm{M_i}} \sum_{\mathrm{j}=1}^{\mathrm{M_i}} \left(\mathrm{SgK}_{\mathrm{noise,avg}}^{\mathrm{ij}} - \mathrm{SgK}_{\mathrm{Clean,avg}}^{\mathrm{i}} \right)^2} \tag{4.2}$$

Table 4.1 shows the RMSEavg (overall Root Mean Square Error) of the averaged SGR estimates under all noise types compared to clean for a given SNR. As expected, the RMSE of SGRs is fairly small over all SNRs. Table 4.2 also demonstrates that the overall RMSEavg across all SNRs is small for a given noise type. The results confirm on the robustness of the SGR estimation algorithm on a much larger data set. Similarly, the experiment on NIST SRE08 draws the same conclusion. Thus, it is beneficial to incorporate the SGRs for noise robust speaker identification tasks.

Table 4.1: Overall RMSE of SGRs under several SNRs (TIMIT)

RMSE(Hz)	5dB	10dB	15dB	20dB
SGR1	45.0753	33.0358	23.3100	15.6578
SGR2	86.8863	67.1728	48.9838	34.5018
SGR3	93.7503	72.4795	52.8535	37.2275

Table 4.2: Overall RMSE of SGRs under several noise types (TIMIT)

RMSE(Hz)	Babble	Factory1	Pink	White
SGR1	19.3118	25.4771	31.9324	42.7997
SGR2	37.5271	49.8199	63.0521	86.7192
SGR3	40.4917	53.7557	68.0332	93.5700

4.3 SID experiments and results

All experiments were conducted under mismatched conditions with clean training utterances evaluated against noisy test utterances. Details about the noisy speech are stated in Section 4.2.2.

The TIMIT SID acoustic models are UBM GMMs (as in [20]), since TIMIT only has files with short utterances, the UBM GMM framework is adequate to use here. On the other hand, the state of the art i-vector/PLDA model is used on the NIST08 dataset. Given the enrollment data, speech segments are first detected using a statistic-based voiced activity detection (VAD) algorithm [34] to discard non-speech frames. In total four front-end features are extracted: LPCCs, PNCCs, SGRs, and MFCCs. For MFCCs and PNCCs, we use x1-x20 and their first- and second- order derivatives, so in total 60 dimension features. For LPCCs, x1-x24 is used for our experiment. Note that all the cepstral features are computed for all speech frames whereas SGRs are computed for voiced frames, only.

Given a test utterance, the cepstral features and SGRs are computed as described above. The cepstral features are scored with their respective models to obtain several sets of scores. The scores are log likelihoods for SID and are normalized to the range [0, 1]. The top 3 speakers are selected for the test utterance with their respective normalized scores. For these 3 speakers, SGRs are used as the new features and MLP as the new classifier, and new normalized scores are generated. The scores from the two stages are combined in a weighted fashion, and the weight ranges from 0 to 1. The combined scores are used to make a decision.

4.3.1 SID on TIMIT database

TIMIT consists of data (sampled at 16kHz) from 630 speakers, and each speaker said 10 utterances [12]. The average utterance length is around 3 seconds. One of the 10 utterances is used as the test trial for each speaker and the remaining 9 sentences are used for acoustic modeling. Cepstral features are modeled with 128-component GMMs.

SID performance is evaluated in 16 different conditions (four SNRs with four noise types), and Table 4.3 shows the results for SID on TIMIT. SGRs are substantially effective and complementary to LPCC and PNCC, since it gives significant improvement over the baseline. The combined features system performs the best across all the noise conditions. Since the MFCC baseline was low, we didn't evaluate it with SGRs. The improvement in the matched clean conditions is small, since the baseline is already above 99.7%.

Table 4.3: SID accuracies under different noise and SNR combinations for TIMIT (boldface numbers indicate best results)

	MFCC	*PNCC*	*LPCC*	*PNCC+SGRs*	*LPCC+SGRs*
Babble					
5dB	46.7	53.5	64.7	56.4	**70.8**
10dB	85.2	80.6	93	82.4	**94.1**
15dB	95.7	89.8	98	92.0	**98.4**
20dB	97.7	91.7	99.2	92.4	**99.3**
Factory					
5dB	14.3	24	23.1	**30**	27.8
10dB	41.9	59.3	56.6	**63.4**	60.5
15dB	73.8	83.6	81.9	84.7	**86.9**
20dB	92.5	91.4	95.8	92.4	**97.3**
Pink					
5dB	4.6	18.9	7.4	**22.1**	10.9
10dB	17	34.7	26.1	**39.9**	31.6
15dB	42.4	57.9	52.8	**61.3**	59.7
20dB	71.9	80.6	78.8	83.2	**85.2**
White					
5dB	3	8	1.7	**11.8**	1.9
10dB	7.3	19.2	9	**22**	11.7
15dB	16.4	35.7	22.8	**41.5**	28.7
20dB	38.7	56.5	45.3	**59**	52.2

4.3.2 SID on NIST 2008 database

NIST 2008 data is widely used for evaluating speaker verification (SV) algorithms [35]. Compared with TIMIT, it has higher speaker and channel variability. Note unlike the standard SV task, in this chapter we only focus on the closed-set SID task and demonstrate the efficacy of the SGRs in the presence of larger speaker and channel variability. Therefore, we randomly drew 947 speakers from the evaluation dataset (3conv part of the training set). Each speaker has three telephone conversations. From the three telephone conversations, a 10sec piece is extracted as the test file and the rest is used for training the system.

Since, our experiment is only concerned with the closed-set SID task, the training data can be used as the development data to set up the i-vector/PLDA system. A gender independent UBM of 1024 GMM components was built. A total variability matrix T of 400 factors was used and the dimension of the resultant i-vector was further reduced via PLDA modeling with 200 latent components.

The best weights of the feature scores from the TIMIT database are used for the NIST 2008 database to evaluate the performance. Table 4.4 summarizes the results for the performance of the proposed combined features system and the baseline. Since the baselines for 20 dB and matched clean condition are already very high and the improvement is small, we don't show the results here. Similarly, the combined features system outperforms the other baselines in all noise conditions, which show the significant complementary effect for SGRs to the baseline cepstral features.

Table 4.4: SID accuracies under different noise and SNR combinations for NIST SRE 08

	MFCC	*PNCC*	*LPCC*	*PNCC+SGRs*	*LPCC+SGRs*
Babble					
5dB	16.6	46.2	37.9	**50.6**	43.6
10dB	45.7	76.1	70.3	**80.5**	76.5
15dB	75.0	89.8	90.5	92.0	**92.8**
Factory					
5dB	20.5	44.5	40.5	**49.6**	46.1
10dB	54.9	75.2	74.3	**79.4**	78.2
15dB	84.3	89.6	93.9	91.4	**95.2**
Pink					
5dB	17.6	47.7	24.0	**53.2**	30
10dB	53.7	77.8	63.3	**81.4**	6.2
15dB	85.2	90.7	89.5	**92.8**	91.1
White					
5dB	8.7	37.3	4.6	**39.8**	6.6
10dB	40.2	52.1	22.1	**55.7**	27.6
15dB	59	75.1	54.6	**77.0**	60.1

4.3.3 Discussion

In the first stage of the SID experiment, the top 3 speakers' models are selected with respect to the test utterance. (We tried using the top 5 speaker models to conduct the experiment, but it did not give better results.) Note that PNCC+SGRs achieves the highest accuracy in most cases.

4.4 Conclusions

In this chapter, a two-stage noise robust speaker identification system is proposed to demonstrate the efficacy of the SGRs as noise-robust features. SID experiments on TIMIT and NIST 2008 database demonstrates that SGRs can provide complimentary speaker information to noise robust features, such as PNCCs.

Chapter 5

Summary and future work

5.1 Conclusions

This thesis extends the previous research on subglottal acoustics to a wider range of analysis and applications, and especially on Mandarin and children speakers, and noisy conditions.

For Mandarin speakers, a new database was collected (20 native Mandarin speakers). The analysis result shows that Sg2 forms the natural boundary of the back and front vowels. Moreover, there exist inverse relationships between standing height and Sg2, and between trunk length and Sg2, with a stronger correlation between trunk length and Sg2. For the relationship between F0 and Sg2, while F0 varies over time within a vowel, there is no statistically-significant variation of Sg2.

For children speech, due to the different acoustics properties of the children above the age of 11 and those under 11, an age-dependent SGR estimation algorithm is investigated. The height estimation employs a negative correlation between SGRs and height, and the mean absolute height estimation error was found to be less than 3.8cm for younger children and 4.9cm for the older children. In addition, using TIDIGITS, a linear frequency warping scheme using age-dependent Sg3 gives statistically-significant word error rate reductions (up to 26%) relative to conventional VTLN.

The utility of SGRs are also investigated for noise-robust speaker identification (SID). A two-stage framework is proposed which combines the SGRs with different cepstral features. Experiments with the TIMIT and NIST 2008 databases show that SGRs, when used in conjunction with PNCCs and LPCCs, can improve the performance significantly (2-6% absolute accuracy improvement) across all noise conditions and in mismatched situations.

5.2 Future work

Previous research is more focused on knowledge-based subglottal feature analysis and extraction, such as using formant information to estimate SGRs. More sophisticated and meaningful subglottal features will be studied and estimated from the speech signal, by nonlinear feature mapping from speech signals using Deep Neural Networks (DNNs). DNNs have been shown to produce reliable nonlinear mapping between features. The new SGRs features, such as the spectrum or log spectrum of the subglottal signal, may improve the previous SGR-based applications. The detailed subglottal features may help a lot under the i-vector framework, when testing or training utterances are short (less than 30s). Moreover, i-vector extraction from the detailed subglottal features may also be helpful for DNN-based speaker adaptation for ASR.

Reference

[1] Shizhen Wang, "Rapid Speaker Normalization and Adaptation with Applications to Automatic Evaluation of Children's Language Learning Skills", Ph.D. dissertation, University of California, Los Angeles, 2010.

[2] Lulich, S. M., Morton, J. R., Arsikere, H., Sommers, M. S., Leung, G. K., and Alwan, A., "Subglottal resonances of adult male and female native speakers of American English", The Journal of the Acoustical Society of America, 132(4): 2592-2602, 2012.

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

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

[5] Jung, Y., "Subglottal effects on the vowels across language: Preliminary study on Korean", JASA 125:2638, 2009.

[6] Csapó, T. G., Bárkányi, Z., Gráczi, T. E., Bohm, T., and Lulich, S. M., "Relation of formants and subglottal resonances in Hungarian vowels", Interspeech, 2009.

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

[8] Arsikere, H., Leung, G. K. F., Lulich, S. M., and Alwan, A. "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation", Speech Commun, 2013, 55(1): 51-70..

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

[10] Arsikere, H., Leung, G. K. F., Lulich, S. M., and Alwan, A. "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation", Speech Commun, 2013, 55(1): 51-70

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

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

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

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

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

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

[17] H. Arsikere, S. M. Lulich, and A. Alwan. "Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency." ICASSP. 2013.

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

[19] Harish Arsikere, H.A. Gupta and Abeer Alwan, "Speaker recognition via fusion of subglottal features and MFCCs," Interspeech 2014, pp. 1106-1110.

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

[21] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. “Frontend factor analysis for speaker verification.” IEEE Transactions on Audio, Speech and Language Processing, 19(4):788–798, 2011.

[22] D. A. Reynolds, “Experimental Evaluation of Features for Robust Speaker Identification,” IEEE Trans. on Speech and Audio Processing (TSAP), vol. 2, no. 4, pp. 639-643, Oct. 1994.

[23] C. Kim and R. M. Stern, “Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition,” Proc. ICASSP, 2012, pp. 4101-4104.

[24] Sjölander, K., Wavesurfer. KTH, Stockholm, Sweden. Online: https://www.speech.kth.se/wavesurfer/, 2012.

[25] Gráczi, T. E., Lulich, S. M., Csap´o, T. G., and Beke, A., "Context and speaker dependency in the relation of vowel formants and subglottal resonances - evidence from Hungarian", Proceedings of Interspeech 1901–1904, 2010.

[26] Zhang, T.D. and Wang, W.K., "Probability theory and Mathematical Statistics Fourth Edition", Zhejiang University. Higher Education press, Beijing, China, 2008.

[27] McDonald, J.H., "Handbook of biological statistics", Department of Biological Sciences, University of Delaware, Delaware, DE, United States, Online: http://udel.edu/~mcdonald/statpaired.html, 2014.

[28] S. M. Lulich, A. Alwan, H. Arsikere, J. R. Morton, and M. S. Sommers, "Resonances and wave propagation velocity in the subglottal airways", The Journal of the Acoustical Society of America, 130(4): 2108-2115, 2011.

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

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

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

[32] Alwan, Abeer, Steven Lulich, and Mitchell Sommers. The Subglottal Resonances Database LDC2015S03. Hard Drive. Philadelphia: Linguistic Data Consortium, 2015.

[33] H. G. Hirsch, "F a NT-Filtering and Noise Adding Tool," 2005.

[34] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1–3, 1999.

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