UNIVERSITY OF CALIFORNIA

Los Angeles

An Exploration Study of the Effect of Voice Quality on Subglottal Resonances

A thesis submitted in partial satisfaction

of requirements for the degree Master of Science

in Electrical Engineering

by

Yi-Hui Lee

2010

The thesis of Yi-Hui Lee is approved.

Nhan Levan

Aydogan Ozcan

Abeer Alwan, Committee Chair

University of California, Los Angeles

2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Professor Abeer Alwan for her guidance, teaching, and support throughout my graduate study at UCLA Speech Processing and Auditory Perception Laboratory (SPAPL). I greatly appreciate her time and effort in reviewing my thesis. I also want to thank her for providing me the opportunity to present at Interspeech 2009 and Speech and Language Technology in Education 2009. I have tremendously benefitted from these experiences.

I would also like to thank SPAPL alumni Shizhen Wang, and Steven Lulich at Washington University for their knowledge and expertise in speaker normalization techniques and subglottal resonances, respectively. I want to especially thank Shizhen for co-authoring the conference paper presented in Chapter 2 (Wang *et al.*, 2009)

I would like to express my appreciation to Professor Nhan Levan and Professor Aydogan Ozcan for their time in serving as members of my Master Thesis' committee.

I would like to thank the following people for supporting my education. First, I would like to thank Eva Borgstrom and the Borgstrom family for their generous donation through the Borgstrom Fellowship. Second, I would like to thank the National Science Foundation and Professor Abeer Alwan again for funding my current research on subglottal resonances. I would not have been able to focus on my coursework and research full time without them.

Lastly, I must give my thanks to my parents for their love, encouragement, and spiritual support.

ABSTRACT OF THE THESIS

An Exploration Study
of the Effect of Voice Quality
on Subglottal Resonances

by

Yi-Hui Lee

Master of Science in Electrical Engineering
University of California, Los Angeles, 2010
Professor Abeer Alwan, Chair

Recently, researchers in speech production modeling have gained more knowledge about the coupling between the vocal tract and the lower airway system, giving rise to acoustic features known as subglottal resonances. Subglottal resonances have been extensively researched for classifying the vowel space in frequency into distinctive regions for many languages. With known correlations between subglottal resonances and formant frequencies, these features have also been used in speech technology, such as speaker normalization techniques for automatic speech recognition system. The use of estimated second subglottal resonance in frequency warping functions for speaker normalization has significantly improved the performance of the recognition system in comparison to conventional methods. However, current subglottal resonances information used in these

applications are based on normal speech production. To determine if subglottal resonances are affected by variations in phonation, an exploration study is undertaken. The study focuses on the effects of first and second subglottal resonances of the breathy voice, lax voice, and tense voice relative to modal voice. A small corpus is created with the speech signals and subglottal resonances data from five English speakers using a bi-directional microphone and an accelerometer, respectively. Results of the study suggest that further experimentation is required to formulate generalizable conclusions.

CHAPTER 1

INTRODUCTION

1.1. Motivation

Recently, researchers in speech production modeling have gained more knowledge about the coupling between the vocal tract and the lower airway system, giving rise to acoustic features known as subglottal resonances. Subglottal resonances have been extensively researched as speaker-specific acoustic features in dividing the vowel space into distinctive regions (Lulich and Chen, 2009). With known correlations between subglottal resonances and formant frequencies, these features have also been used in speech technology, such as estimating the warping factor in speaker normalization techniques for automatic speech recognition systems (Wang et al., 2009). However, current subglottal resonances information used in these applications are based on normal phonation. To determine if subglottal resonances are affected by variations in phonation, an exploration study is undertaken.

1.2. Speech Production Model with Coupling to the Subglottal System

In speech production, a linear time-invariant source model is composed of the following: the voiced input is the glottis (source function), the transfer function is the vocal tract

(above the glottis to the opening of the mouth), and the voiced output is the speech signal. An illustration of this system is displayed in Figure 1.1.

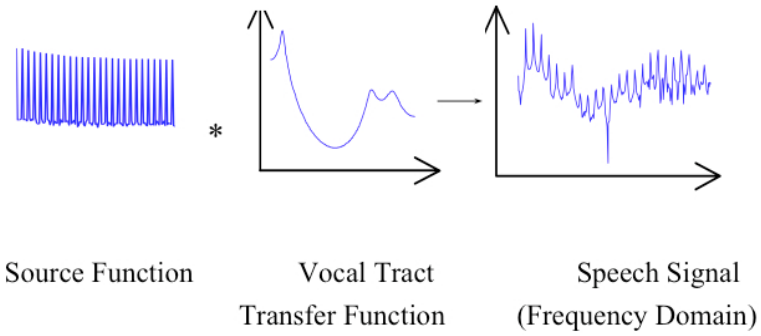

Figure 1.1 A linear time-invariant model of speech production

However, when acoustic coupling occurs between the vocal tract and the subglottal system, additional zero-pole pairs occur in the speech spectra. The zero-pole pairs derive from the natural frequencies of the lower airway right below the glottis, which contains the trachea, the two main bronchi, the bronchioles, and finally the alveoli (see Figure 1.2). These frequencies are known as subglottal resonances.

With the acoustic coupling between the subglottal system and the vocal tract, Chi and Sonderegger have proposed a modified speech production model based on circuit theory (Chi and Sonderegger, 2007; Lulich, 2009). Figure 1.3 is the circuit diagram of the proposed system. Z_{VT} is the impedance of the vocal tract, Z_g is the impedance of the glottis, and Z_{sg} is the impedance of the subglottal system. U_{s1} and U_{s2} are the twin glottal volume velocity sources produced by vocal fold vibration.

The resulting transfer function in the model effectively incorporates an additional zero-pole pair per subglottal resonance. With this model, researchers are able to visualize

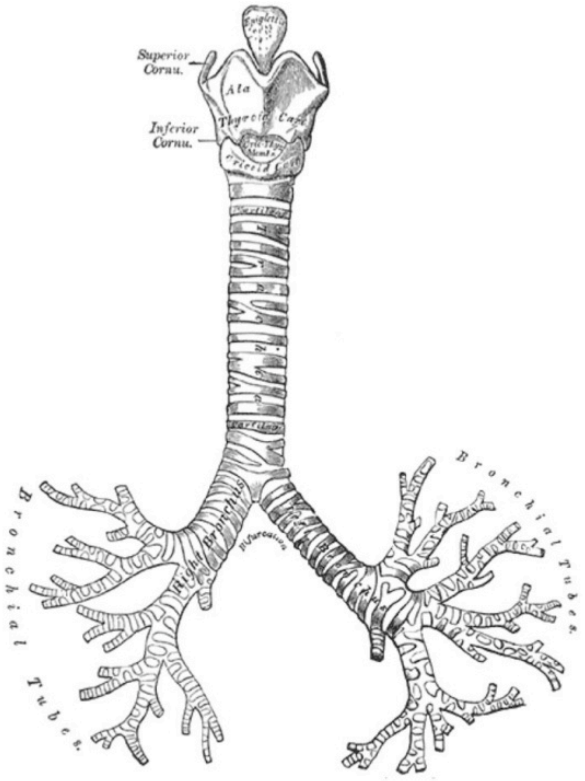

Figure 1.2 An anatomical sketch of the lower airway, including the trachea, the two main bronchi, and bronchiole tree (adapted from Gray, 1918).

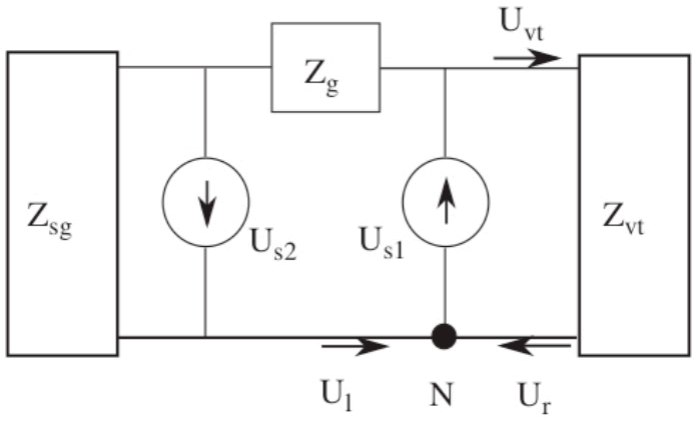

Figure 1.3 Speech production model coupled with subglottal system as a circuit diagram. Z_{VT} is the impedance of the vocal tract, Z_g is the impedance of the glottis, Z_{sg} is the impedance of the subglottal system, U_{s2} and U_{s1} are the twin glottal volume velocity sources due to vocal fold vibration.

and understand the relationship between subglottal resonances and formants of speech sounds.

1.3. Subglottal Resonances and Formant Frequencies

The first three subglottal resonances have been used in classifying the vowel space in frequency into distinctive regions. The dimensions of the vowel space correspond to the first and second formant frequencies (F1 and F2) of the vowel sound. The first subglottal resonance (Sg1) forms the division between the low and non-low vowels, which are grouped by F1. The second subglottal resonance (Sg2) separates the front vowels and the back vowels, which are characterized by F2. Finally, the third subglottal resonance (Sg3) forms the division between the tense and lax vowels, which are classified by the duration of the vowels (Stevens, 1997). These relationships have been supported in multiple languages, including American English (Lulich 2009), Korean (Jung, 2009), German (Madsack et al., 2008), and Hungarian (Csapó et al., 2009). In addition, these studies have shown that the subglottal resonances are speaker-specific and content-independent due to limited articulation factors during speech production.

Of the three subglottal resonances, the Sg2 has been extensively studied because of its robust interactions with formants. In Chi and Sondergger's study on back-front diphthongs (sound formed by the combination of two vowels) of adult speakers, attenuation of F2 prominence and discontinuity in the F2 track occur near the second subglottal resonance (Chi and Sondergger, 2007). An example of these effects is shown in Figure 1.4. In another study of child vowel spaces, empirical analysis has presented a linear relationship between Sg2 and the third formant (F3). By combining these two

findings, estimation of subglottal resonances from speech data becomes possible in automatic speech recognition systems.

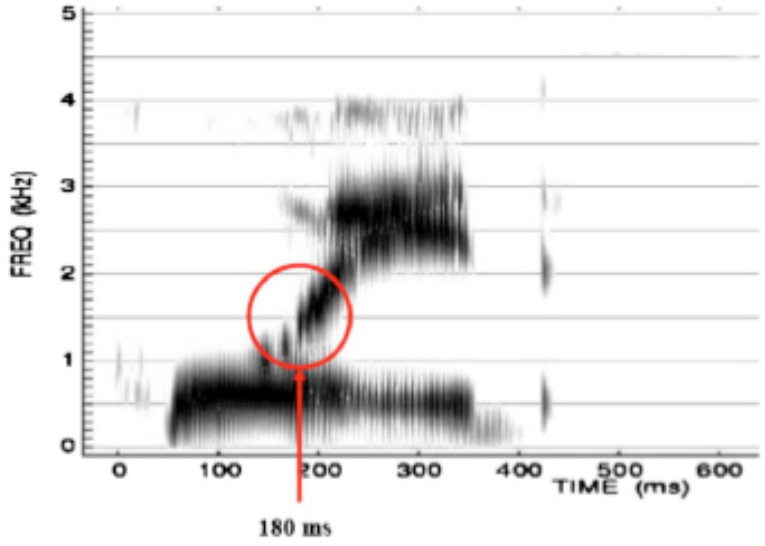

Figure 1.4 Spectrogram for a diphthong spoken by a male speaker. As shown in the circled region at 180ms, the attenuation of F2 prominence and discontinuity in the F2 track occur near the subbglottal resonance, at 1370Hz (adapted from Chi and Sondergger, 2007).

1.4. Organization of Thesis

This thesis is organized into four chapters. Chapter 2 discusses the usage of subglottal resonances in speaker normalization techniques. Chapter 3 presents the exploration study of the effect of voice quality on subglottal resonances. Finally, Chapter 4 presents a summary and concluding remarks.

CHAPTER 2

USING SUBGLOTTAL RESONANCE INFORMATION IN SPEAKER NORMALIZATION

Because the second subglottal resonance (Sg2) has been shown to be context-independent, this speaker-specific feature has been incorporated in a speaker normalization scheme for an automatic speech recognition (ASR) system.

2.1. Speaker Normalization Techniques

In ASR, the robustness of the system often degrades due to speaker differences between the training and testing data. Differences such as age, gender, and even language generally affect the speech spectra, creating spectral mismatches. To reduce this effect, speaker normalization is applied in ASR using a technique known as frequency warping.

Equation 1 provides a general frequency warping scheme, where f is the frequency scale in Hz, $W_\alpha(f)$ is the warping function, and α is the optimal warping factor.

$$S'(f) = S(W_\alpha(f)) \tag{1}$$

One of the most popular normalization techniques is vocal tract length normalization (VTLN), which is denoted as

$$W_\alpha(f) = \alpha \cdot f \tag{2}$$

and assumes that differences in the speakers' vocal tract lengths result in scaled spectra of each other.

Even though this technique is commonly used, VTLN only provides a crude approximation to reduce spectral mismatch due to vocal tract variations. In Umesh's study, it has been observed that the warping factor between speakers is frequency dependent (Umesh et al., 2002).

Several nonlinear frequency warping functions have also been proposed. Based on psycho-acoustical observations in auditory perception studies on speaker normalization (Bladon et al., 1984), approaches such as shift-based Mel-scale and Bark-scale frequency warping functions have been introduced:

$$W_{\alpha}(z) = z + \alpha \tag{3}$$

where z represents the warped spectra of either the Mel domain (Umesh et al., 2002) or the Bark domain (Sinha and Umesh, 2008). The Mel domain and its frequency shift (Mel shift) representation are described in Equations 4 and 5, while the Bark domain and its frequency shift (Bark shift) representation are described in Equations 6 and 7.

The Mel Domain:

$$z = Mel(f) = 1127 \log(1 + \frac{f}{700}) \tag{4}$$

$$f' = e^{\frac{\alpha}{1127}} \cdot f + 700(e^{\frac{\alpha}{1127}} - 1) \tag{5}$$

The Bark Domain:

$$z = Bark(f) = 6 \log(\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}) \tag{6}$$

$$f' = 300e^{\frac{\alpha}{6}}[\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}] - \frac{300e^{-\frac{\alpha}{6}}}{\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}} \tag{7}$$

Equation 7 can further be simplified with the following criteria

$$\begin{cases} f' = e^{\frac{\alpha}{6}} \cdot f & \text{for } f >> 600 \text{ Hz} \\ f' = e^{\frac{\alpha}{6}} \cdot f + 600(e^{\frac{\alpha}{6}} - 1) & \text{for } f >> 600 \text{ Hz} \end{cases} \tag{8}$$

For high frequencies $f >>$ 600 Hz, the Bark shift corresponds to a linear scaling, while for low frequencies f << 600 Hz, the Bark shift results in a nonlinear scaling, which closely resembles the Mel shift. In general, the Bark shift warping function stretches or compresses lower frequencies more than higher frequencies.

In order to preserve the bandwidth after Bark shift warping, a piece-wise nonlinear warping function is applied as shown in Figure 2.1 with a different warping factor α (Wang et al., 2009). Equation 9 provides a mathematical expression of this function.

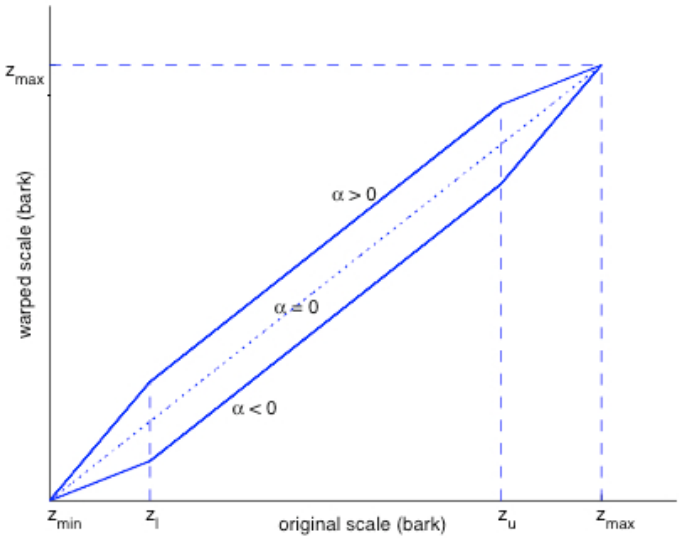

Figure 2.1 Piecewise bark shift warping function, where $\alpha > 0$ shifts the Bark scale upward, $\alpha < 0$ shifts the Bark scale downward, and $\alpha = 0$ means no warping. z_l and z_u represent the lower and upper discontinuity in bark scale, respectively

$$W_\alpha(z)=\begin{cases}\dfrac{z_l+\alpha-z_{\min}}{z_l-z_{\min}}\cdot(z-z_{\min})+z_{\min} & \text{if } z\le z_l\\ z+\alpha & \text{if } z_l\le z\le z_u\\ \dfrac{z_{\max}-z_u-\alpha}{z_{\max}-z_l}\cdot(z-z_u)+z_u+\alpha & \text{if } z>z_u\end{cases} \tag{9}$$

In this case, $z_{\min}$ and $z_{\max}$ remain constant after applying the warping factor. This property essentially increases the computational efficiency. Furthermore, the effect of bandwidth mismatch after frequency warping is compensated.

2.2. Evaluation of Warping Factor using Second Subglottal Resonance

Generally for both linear and nonlinear speaker normalization techniques, the optimal warping factor, α, is normally evaluated using a maximum likelihood (ML) based grid search approach based on the observations given an acoustic model λ:

$$\alpha=\arg\max_{\alpha\in G}\sum_{r=1}^{R}\log p(O_r(W_\alpha(f))\,|\,\lambda,s_r) \tag{10}$$

where s_r is the transcription of the r-th speech file O_r, and G is the search grid. Although this approach is speaker specific, it is not context-independent. Therefore, Sg2 will be used to evaluate the warping factor for all three techniques mentioned earlier to demonstrate comparable or better performance than traditional approaches.

However, prior to calculating the warping factor, an automatic estimation of Sg2 from speech files is required. The foundation of the algorithm is mainly based on F2 discontinuity, energy prominence in F2 (E2), and average F3. (Note: as mentioned in the pervious chapter, the average F3 and Sg2 relation is only applicable to children's speech.) From empirical analysis, it has been shown that as F2 approaches Sg2, an attenuation of

5-12dB in E2 always occurs, while an F2 discontinuity in the range of 50-300Hz often occurs. By combining the two properties, the detection algorithm, Sg2DJ is implemented as follows:

A. Track F2 and E2 frame by frame using LPC analysis and dynamic programming. The F2 tracking algorithm is similar to that used in Snack (Wavesurfer), with parameters specifically tuned to provide reliable F2 tracking results on children's speech. Manual verification and/or correction is applied through visually checking the tracking contours against the spectrogram.

B. Search automatically within ±100 Hz around the initial estimated Sg2 ($\tilde{Sg}2$) for F2 discontinuities (F2d) and E2 attenuation (E2a).

$$\tilde{Sg}2 = 0.363 \times F_3 - 103 \tag{11}$$

C. Apply the following decision rules for final Sg2 Estimation:

   a. If F2d and E2a correspond to the same frame, the estimated Sg2 is evaluated as

$$\begin{aligned} \hat{Sg}2 &= \beta \cdot F2_{high} + (1-\beta) \cdot F2_{low} \\ \hat{\beta} &= \arg\min_{\beta} E\{(\hat{Sg}2 - Sg2)^2\} \end{aligned} \tag{12}$$

   where $F2_{high}$ and $F2_{low}$ are the F2 values on the high and low frequency side of the discontinuity, respectively; β is a weight in the range (0,1) that controls the closeness of the detected Sg2 value to $F2_{high.}$ The optimal value of β is estimated using minimum mean square error criterion on the training data (Figure 2.2).

b. If F2d and E2a correspond to different frames but E2a is present, the estimated Sg2 is evaluated as an average of F2 values. The F2 values are from three consecutive frames with the centered frame corresponding to the E2a frame (Figure 2.3).

c. If E2a is not present, the estimated Sg2 is calculated using Equation 11.

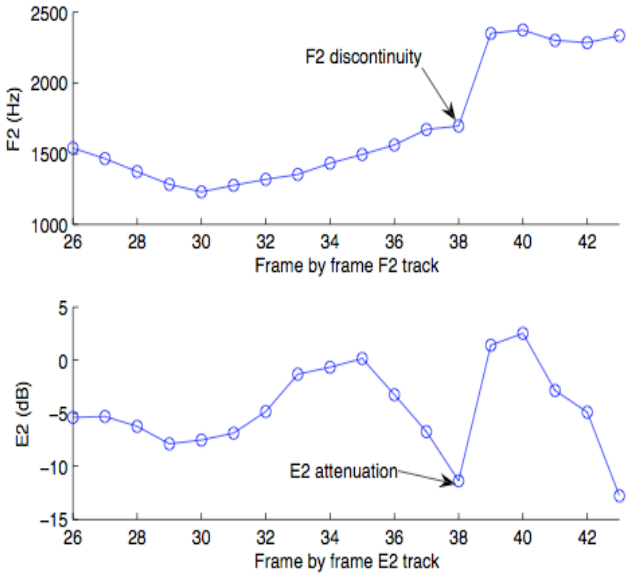

Figure 2.2 Example of Decision a: the joint estimation method where F2 discontinuity and E2 attenuation correspond to the same location (frame 38).

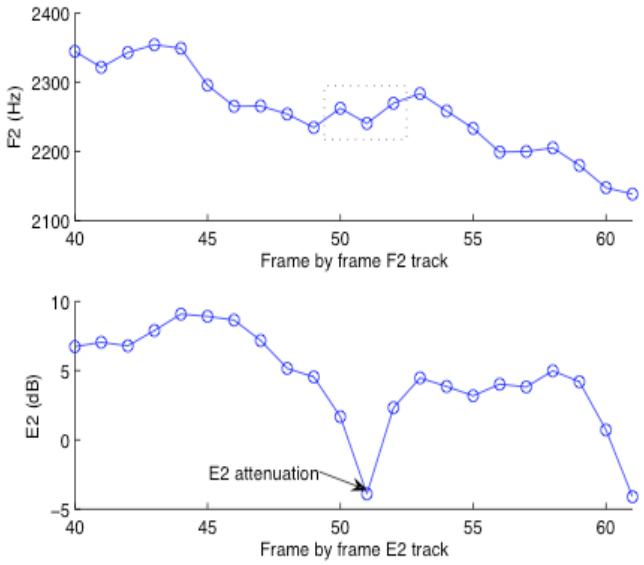

Figure 2.3 Example of Decision b: the joint estimation method where F2 discontinuity (not detectable) and E2 attenuation (frame 51) are mismatched. The average F2 value within the dotted box is used to estimate Sg2.

After automatically estimating Sg2, the optimal warping factor, α, is then evaluated given the test speaker's Sg2 ($Sg2_{test}$) and reference speaker's Sg2 ($Sg2_{ref}$):

$$\alpha = \begin{cases} Sg2_{ref} / Sg2_{test} & \text{for linear scaling} \\ Mel(Sg2_{ref}) - Mel(Sg2_{test}) & \text{for Mel shift} \\ Bark(Sg2_{ref}) - Bark(Sg2_{test}) & \text{for Bark shift} \end{cases} \tag{13}$$

The ML-based speaker normalization in Equation 10 involves an exhaustive grid search to find an optimal warping factor, which is time consuming and requires a certain amount of data to be effective. On the contrary, the main computational cost for Sg2-based speaker normalization methods come from F2 and E2 tracking using linear prediction coding (LPC) analysis, which can be done efficiently. Since Sg2 has been shown to be context-independent and relatively constant for a given speaker, the estimation can be

performed well with limited data. In this scenario, Sg2 clearly has an advantage over ML-based grid search.

2.3. Performance of Sg2-based Speaker Normalization

For computational efficiency, all normalization methods are implemented by modifying the Mel or Bark filter bank analysis instead of warping the power spectrum. The Mel frequency cepstrum coefficient (MFCC) features are used for Mel shift, and the perceptual linear prediction cepstrum coefficient (PLPCC) features are used for Bark shift. In this experiment, combinations of linear and nonlinear normalization with ML-based and Sg2-based speaker normalization methods are evaluated.

In order to use a consistent framework to ensure fair comparison between ML-based linear and nonlinear normalization, search grid parameters need to be adjusted such that linear and nonlinear warped spectra covers roughly the same frequency range. For linear scaling (LS), a grid of 21 search points is used with a step size of 0.01. For nonlinear warping, the step size is approximately 0.07 for Bark-shift (BS) and 10 for Mel shift (MS) in their respective warped spectral domain.

Two databases are used to evaluate the performance of these different normalization techniques on children's automatic speech recognition: TIDIGITS and TBall (Wang et al., 2009). For both databases, the speech signals are segmented into 25ms frames with a 10ms shift. Each frame is parameterized by a 39-dimensional feature vector consisting of 12 static MFCC/PLPCC plus log energy, and their first- and second-order derivatives. In addition, cepstral mean subtraction is applied in all cases. Word error rate (WER) is used for performance evaluation.

For the TIDIGITS study, monophone-based acoustic models were used with 3 states and 6 Gaussian mixtures in each state. The acoustic models are trained on 55 adult male speakers and tested on 50 children, with data of 1, 4, 7, 10, or 15 connected digits. The optimal warping factor is evaluated from randomly selected utterances of the test subset. The ML search grid is [0.8, 1.0] for LS, [-1.4, 0.0] for BS, and [-200, 0.0] for MS. The baseline WER is 37.63% using MFCC features and 37.47% using PLPCC features.

For the TBall study, 55 HFW and 55 BPST words are collected from 189 children in grades of 1 or 2. Approximately two-thirds of the data (12) is used for training, while the remaining third for testing. Three randomly chosen words, including a diphthong, are used for estimating the warping factor, since the most reliable Sg2 estimation requires F2 transition from back to front vowel. The ML search grid is [0.9,1.1] for LS, [-0.7, 0.7] for BS, and [-100. 100] for MS. The baseline WER is 7.75% using MFCC features and 8.35% using PLPCC features.

Warping	**1**	**4**	**7**	**10**	**15**
LS-ML	7.48	6.34	5.42	4.99	4.91
MS-ML	6.33	5.47	4.48	4.11	4.08
LS-Sg2	6.11	5.57	5.05	5.07	5.03
MS-Sg2	5.29	4.81	4.05	4.13	3.99

Table 2.1 WER on TIDIGITS using MFCC features with varying normalization data from 1 to 15 digits

Warping	**1**	**4**	**7**	**10**	**15**
LS-ML	7.62	6.90	5.78	5.64	5.25
BS-ML	6.21	5.63	4.56	4.30	4.13
LS-Sg2	6.15	5.17	5.51	5.47	5.39
BS-Sg2	5.17	4.76	4.09	4.11	4.05

Table 2.2 WER on TIDIGITS using PLPCC features with varying normalization data from 1 to 15 digits

Tables 2.1 and 2.2 show results on TIDIGITS with various amounts of normalization data for MFCC and PLPCC features, respectively. The warping abbreviation is denoted as AA-BB, where AA stands for the type of frequency warping and BB stands for type of the warping factor estimation.

When comparing LS with MS in Table 2.1 and LS with BS in Table 2.2, the nonlinear frequency warping functions show better performances than linear warping function for all conditions, which is in agreement with the literature. However, comparing the WER of Sg2 and ML for the linear warping function in Table 2.1, Sg2 performs better for digits of 1, 4, or 7, while ML performs slightly better for digits of 10 or 15. This result suggests for ASR system with limited data, Sg2 may be a better technique in estimating the warping factor than ML.

MFCC		**PLPCC**	
Warping	**WER**	**Warping**	**WER**
LS-ML	6.86	LS-ML	6.99
MS-ML	5.91	BS-ML	5.82
LS-Sg2	6.10	LS-Sg2	6.33
MS-Sg2	4.89	BS-Sg2	4.71

Table 2.3 WER on TBall children's data using MFCC and PLPCC features with 3 normalization words

Table 2.3 shows results of the TBall database with direct comparison between MFCC and PLPCC. All combinations of normalization techniques have shown a reduction in WER in comparison to the baseline. In this study, all linear and nonlinear warping functions using the Sg2-based warping factor estimation have shown better results than their counterparts using the ML-based grid search approach. In addition, the combination

of bark-shift and second subglottal resonance has the best performance among the discussed methods.

2.4. Summary

The second subglottal resonance has shown great potential in frequency warping based speaker normalization. After automatically estimating the second subglottal resonance using both frequency discontinuity and energy attenuation, the frequency warping factor is evaluated and applied to both linear and nonlinear frequency warping functions: linear shifting, Mel shifting, and Bark shifting. All three speaker normalization schemes using this feature have shown better performance than using traditional maximum likelihood-based grid search approach, especially with limited data in children's ASR system.

CHAPTER 3

AN EXPLORATION STUDY OF THE EFFECT OF VOICE QUALITY ON SUBGLOTTAL RESONANCES

As mentioned in Chapters 1 and 2, subglottal resonances have been found to improve the performance of ASR systems because of their context-independent property and relationship to other acoustic properties. However, it is important to note that these studies assumed normal speech production. To investigate whether variations in speech production can affect the subglottal resonances, specifically Sg1 and Sg2, different types of voice quality are examined in this exploration study.

3.1. Voice Quality Characterizations

Voice quality is the mannerism of speaking. It is often characterized by physiological terms such as: vocal fold vibrations, vocal tract tension, glottal airflow, etc. There are many different types of voice quality; this study will focus on only four fundamental voice qualities (Laver, 1980):

A. Modal Voice

   This is the neutral mode of phonation, or normal speaking voice. The vibration of the vocal folds is periodic within the range of the fundamental frequency, efficient, and without audible friction.

B. Breathy Voice

This mode of voice quality is considered inefficient and with slight audible friction due to aspiration. The acoustic energy of this voice quality is inversely proportional to the magnitude of breathiness present. The vibration of the vocal folds is still periodic.

C. Lax Voice

This phonation is an extreme case of breathy voice in which the vocal folds completely relax. In comparison to modal voice, lax voice has less acoustic energy.

D. Tense

This voice quality is the direct counterpart of lax voice. The muscles in the vocal tract and vocal fold are tense, creating higher acoustic energy than modal voice.

3.2. Relationship between Voice Quality and Subglottal Resonances

Several studies have proposed potential effects of voice quality on subglottal resonances based on either mathematical or pseudo modeling of voiced speech production. Cranen and Boves proposed a two-mass system with glottal leakage (vocal folds do not close completely at end of glottal cycle) to determine the relationship between breathy phonation and subglottal resonances (pressure). Their model suggested that an increase in glottal leakage area with constant vertical phase difference would decrease subglottal resonances (Cranen and Boves, 1987). In another study, Austin and Titze used excised larynges of dogs and pseudotrachea to investigate the effect of subglottal resonances

based upon vocal fold vibration (Austin and Titze, 1997). Acoustic pressure was measured using an electroglottalgraph and a subglottal pressure transducer. Their study suggested that when the vocal folds are maximally apart, a reduction of subglottal pressure occurs.

Based on the definitions of the four different voice qualities and the proposed results from previous studies, the following are some hypotheses on the relationship between voice quality and subglottal resonances with the modal voice quality as control (baseline):

A. Breathy Voice

   Because breathy voice incorporates simultaneous aspiration during speech production, this increase in leaking airflow at the glottis could potentially decrease the subglottal resonances based on Cranen and Boves' results.

B. Lax Voice

   When comparing the subglottal resonances of this phonation with modal voice, the acoustic energy should be less. In addition, because lax voice provides complete relaxation in the vocal tract and vocal fold, it is likely to decrease the subglottal resonances due to an increase in glottal area.

C. Tense Voice

   When comparing the subglottal resonances of this phonation with modal voice, the acoustic energy should be much higher. In addition, because tense voice provides complete opposite properties of lax voice, it is likely to slightly increase the subglottal resonances due to a decrease in glottal area.

These hypotheses have a higher chance in affecting Sg1 than Sg2 because Sg1 can be influenced by low formants and the fundamental frequency (pitch) information.

To evaluate the hypothesized effect of four different voice qualities on subglottal resonances, data collection is required.

3.3. Data Collection

This section is divided into two parts. The first part focuses on the word list, while the second part focuses on the experimental setup.

3.3.1. Target Word List

Three different vowels are used in the target word list: /**eɪ**/, /**aɪ**/, and /**ɔɪ**/. These vowels are also known as diphthongs that combine two isolated vowels (i.e. /**e**/ and /**ɪ**/) together. There are six words with the structure consonant-vowel (CV) and consonant-vowel-consonant (CVC):

CV	**CVC**
Bay	Hay'ed
Buy	Hide
Boy	Hoid

Table 3.1 Target word list containing CV and CVC sets

The CV set contains the voice plosive /**b**/ followed by the diphthong, while the CVC set contains an aspiration consonant /**h**/, followed by the diphthong, and ending with a voiced plosive /**d**/. These two sets are selected to provide two different contexts since acoustic features of vowels may be influenced by their surrounding consonants.

3.3.2. Equipment Setup

To minimize environmental noise corruption of the data, the recordings are conducted in a sound booth. The equipment used for this experiment include: a bi-directional microphone, an accelerometer, a pre-amplifier, a monitor, a mouse, a keyboard, and a laptop. The pre-amplifier and the laptop are situated outside of the sound booth to reduce the inherent machine noise during the recording.

Inside the sound booth, a bi-directional microphone is used to capture the speech signal while an accelerometer is used to capture the subglottal resonances. The microphone is bi-directional in order to provide noise cancellation from the sides and is situated slightly off center from the speaker to minimize aspiration effect. The accelerometer is placed at the cricoid cartilage of the speaker in order to sense and capture the vibration of the cartilage during speech production. This location is selected because it is connected to the first cartilage ring of the trachea via the cricotracheal ligament. An anatomical sketch of the cricoid cartilage is shown in Figure 3.3. To record both signals simultaneously, the pre-amplifier is used to both amplify the signal, and to support dual channel recording at a sampling rate of 44.1kHz. The wirings among the devices are done through an insulated panel in the sound booth. Finally, the mouse and keyboard are only to enable and disable recording, and to label the files, respectively.

In order to visualize the recordings in real time, the recording software, Adobe Audition, is used. This software supports stereo recording, and monitors clipping. In addition, the software enables both waveforms (time domain) and spectrograms (time-

frequency domain) of the signals with adjustable windowing and sampling properties, providing convenient access to verify the quality of the recordings.

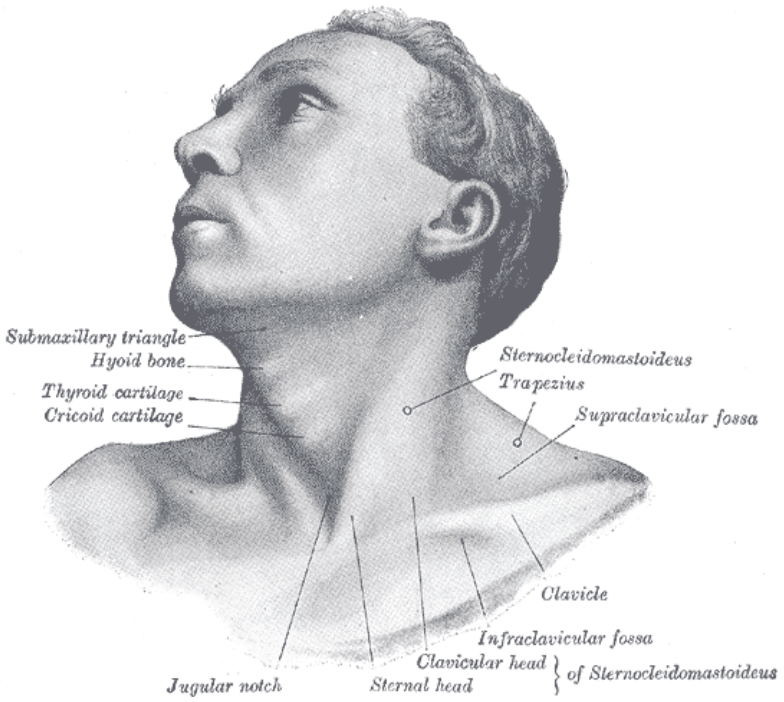

Figure 3.1 An anatomical sketch with the labeled location of the cricoid cartilage (adapted from Gray, 1918).

3.3.3. Recording Session

For this exploration study, a total of five English speakers (3 males and 2 females) are recorded. To control this study, the subjects are asked to speak in four different vocal qualities (modal, breathy, tense, and lax) based on Laver's recordings and definitions (1980) prior to the recording sessions.

In the beginning of the session, the speaker is instructed to pronounce each target word with five repetitions under the four different vocal qualities. To provide consistent vocal quality across the words, the speaker is asked to only change the voice quality after

completing the target word list with repetitions. After the speaker understands the recording process, the speaker will place and hold on to the accelerometer beneath the thyroid and say the target word into the microphone for calibration. Calibration is performed to ensure the quality of the recording, maximize the strength of the signal, and prevent clipping. Once the aforementioned recording parameters are adjusted, the data collection proceeds.

After the recording, the collected stereo wavefiles are parsed and labeled for the vowel in each repetition. Then, the files are imported to MATLAB to split the two channels in order to focus on the subglottal resonances information from the accelerometer. Finally, Adobe Audition is used to manually estimate the Sg1 and Sg2 in the spectrogram viewing of the accelerometer data.

During the estimation, the waveform is first re-sampled to 8kHz to increase resolution of the data. Then, the spectrogram is computed using a Blackmann-Harris window of 25ms with 512-point fast Fourier transform. The gradient used for energy evaluation is based on the default setting. The parameters may be adjusted based on the quality of the data.

3.4. Results for First Subglottal Resonance

For the following results, the five speakers involved in the study will be represented as Speaker G#, where G is the gender of the speaker and # is the speaker ID (i.e. male speaker 1 is denoted as Speaker M1 and female speaker 2 is denoted as Speaker F2). The averaged Sg1 data from the five repetitions for each target word are listed in Tables 3.2 – 3.6.

	Modal	Breathy	Lax	Tense
Hay'ed	553.00	553.00	575.70	568.10
Hide	537.80	560.60	530.30	613.60
Hoid	590.90	568.10	539.70	573.80
Bay	530.30	606.00	575.70	568.10
Buy	583.30	560.60	553.00	606.00
Boy	545.40	606.00	530.30	556.50
Average	556.78	575.72	550.78	581.02
Stdev	24.79	23.94	21.02	23.12

Table 3.2 Average Sg1 of each target word (in Hz) of M1

	Modal	Breathy	Lax	Tense
Hay'ed	492.40	488.60	522.70	477.20
Hide	492.40	454.50	439.30	666.60
Hoid	500.00	439.30	590.90	522.70
Bay	437.50	471.50	460.10	454.50
Buy	517.00	613.60	568.10	676.10
Boy	482.90	448.80	437.50	465.90
Average	487.03	486.05	503.10	543.83
Stdev	26.81	64.89	67.14	101.50

Table 3.3 Average Sg1 of each target word (in Hz) of M2

	Modal	Breathy	Lax	Tense
Hay'ed	560.14	494.28	515.13	595.34
Hide	587.82	481.76	472.68	601.46
Hoid	554.56	583.58	536.32	462.70
Bay	545.44	484.65	545.42	569.68
Buy	531.80	481.78	539.34	552.90
Boy	545.40	592.38	604.52	558.03
Average	554.19	519.74	535.57	556.69
Stdev	19.09	53.13	42.99	50.03

Table 3.4 Average Sg1 of each target word (in Hz) of M3

	Modal	Breathy	Lax	Tense
Hay'ed	519.66	536.32	510.56	556.02
Hide	518.16	551.16	509.04	549.96
Hoid	530.28	534.8	527.24	563.02
Bay	543.94	569.16	527.22	543.88
Buy	531.84	546.91	502.98	515.10
Boy	533.28	533.28	506.86	528.76
Average	529.53	545.27	513.98	542.79
Stdev	9.53	13.72	10.57	17.89

Table 3.5 Average Sg1 of each target word (in Hz) of F1

	Modal	Breathy	Lax	Tense
Hay'ed	636.60	594.65	659.05	704.30
Hide	644.05	602.20	651.70	708.30
Hoid	651.50	772.70	575.70	666.60
Bay	628.70	613.60	586.95	712.50
Buy	621.20	606.00	583.30	655.25
Boy	628.70	704.50	666.60	689.35
Average	635.13	648.94	620.55	689.38
Stdev	11.19	72.98	42.66	23.66

Table 3.6 Average Sg1 of each target word (in Hz) of F2

For all five speakers, the average Sg1 results for the target words under the same voice quality display low variations based on the standard deviations. This finding supports the context-independent property of Sg1. To compare the four different phonations closely, the difference in average Sg1 between modal and each of the remaining vocal qualities are computed in Tables 3.7 – 3.11.

When comparing the average Sg1 differences in Tables 3.7 – 3.11, the relationships among breathy, lax, and tense with respect to modal voice quality do not display consistent increment or decrement in the subglottal resonance. For example, Speaker M1 has average relative Sg1 differences of 18.93Hz for breathy voice, -6.00Hz for lax voice, and 24.23Hz for tense voice. On the other hand, Speaker F1 has average relative Sg1

	Breathy	Lax	Tense
Hay'ed	0.00	22.70	15.10
Hide	22.80	-7.50	75.80
Hoid	-22.80	-51.20	-17.10
Bay	75.70	45.40	37.80
Buy	-22.70	-30.30	22.70
Boy	60.60	-15.10	11.10
Average	18.93	-6.00	24.23
Stdev	41.97	35.17	31.02

Table 3.7 Average Sg1 difference (in Hz) for each voice quality relative to modal voice of M1

	Breathy	Lax	Tense
Hay'ed	-3.80	30.30	-15.20
Hide	-37.90	-53.10	174.20
Hoid	-60.70	90.90	22.70
Bay	34.00	22.60	17.00
Buy	96.60	51.10	159.10
Boy	-34.10	-45.40	-17.00
Average	-0.98	16.07	56.80
Stdev	57.93	55.92	86.75

Table 3.8 Average Sg1 difference (in Hz) for each voice quality relative to modal voice of M2

	Breathy	Lax	Tense
Hay'ed	-65.87	-45.02	35.20
Hide	-106.06	-115.14	13.64
Hoid	29.02	-18.24	-91.86
Bay	-60.79	-0.02	24.24
Buy	-50.02	7.54	21.10
Boy	46.98	59.12	12.63
Average	-34.46	-18.63	2.49
Stdev	59.52	58.46	46.94

Table 3.9 Average Sg1 difference (in Hz) for each voice quality relative to modal voice of M3

	Breathy	Lax	Tense
Hay'ed	16.66	-9.10	36.36
Hide	33.00	-9.12	31.80
Hoid	4.52	-3.04	32.74
Bay	25.22	-16.72	-0.06
Buy	15.07	-28.86	-16.74
Boy	0.00	-26.42	-4.52
Average	15.75	-15.54	13.26
Stdev	12.35	10.35	23.02

Table 3.10 Average Sg1 difference (in Hz) for each voice quality relative to modal voice of F1

	Breathy	Lax	Tense
Hay'ed	-41.95	22.45	67.70
Hide	-41.85	7.65	64.25
Hoid	121.20	-75.80	15.10
Bay	-15.10	-41.75	83.80
Buy	-15.20	-37.90	34.05
Boy	75.80	37.90	60.65
Average	13.82	-14.58	54.26
Stdev	68.21	43.93	25.04

Table 3.11 Average Sg1 difference (in Hz) for each voice quality relative to modal voice of F2

differences of 15.75Hz for breathy voice, -15.54Hz for lax voice, and 13.26Hz for tense voice. Even though both speakers have lowest Sg1 for lax voices, the highest Sg1 for Speaker M1 is the tense voice while the highest Sg1 for Speaker F1 is the breathy voice. Based on this result, using absolute relative difference in Sg1 to evaluate the effect of voice quality on subglottal resonance is not sufficient. Instead, this effect should be determined by whether Sg1 of each voice quality is greater than or less than the modal voice. Figures 3.2 – 3.5 show the results of breathy voice, lax voice, and tense voice, respectively. The unit of measurement is frequency (freq) of occurrence, which represents the averaged Sg1 for each target word.

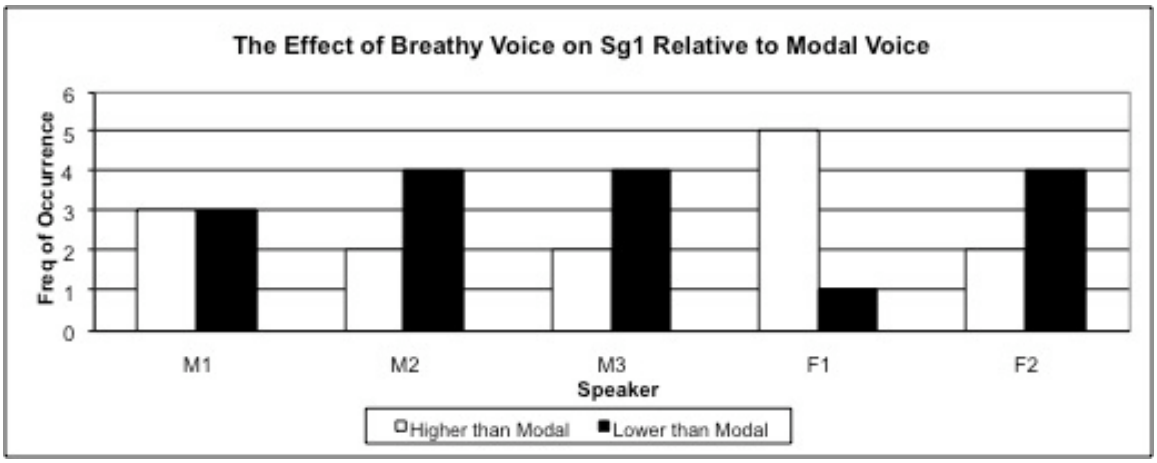

Figure 3.2 The effect of breathy voice on Sg1 relative to modal voice

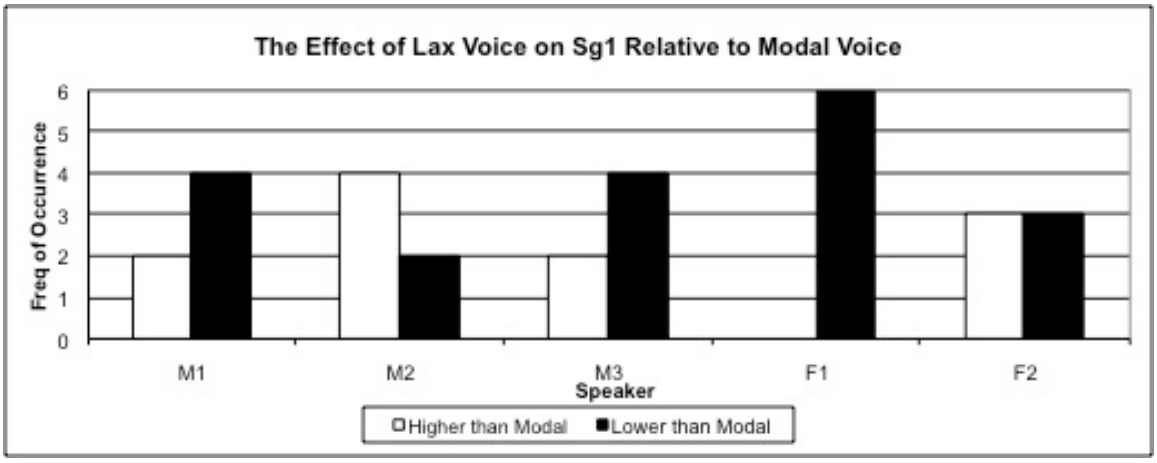

Figure 3.3 The effect of lax voice on Sg1 relative to modal voice

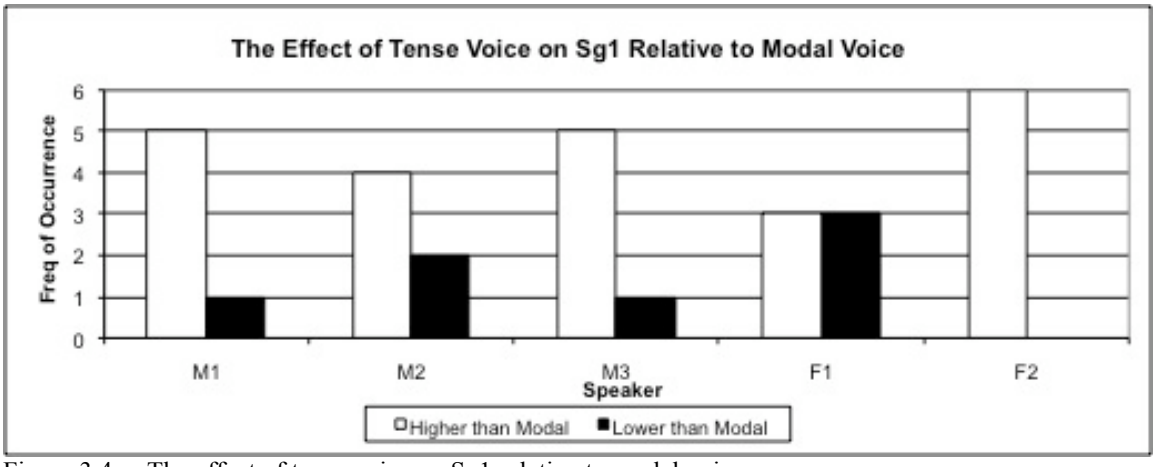

Figure 3.4 The effect of tense voice on Sg1 relative to modal voice

For the effect of breathy voice on Sg1 relative to modal voice, four out of five speakers have shown have equal or more frequency of occurrences to have a decrease in Sg1. This pattern also occurs for lax voice effect. However, for tense voice, all speakers have equal or higher Sg1 than that of modal voice. After averaging the results in Figures 3.2 – 3.4, Figure 3.5 shows the relationships among the different voice qualities, supporting the proposed hypotheses.

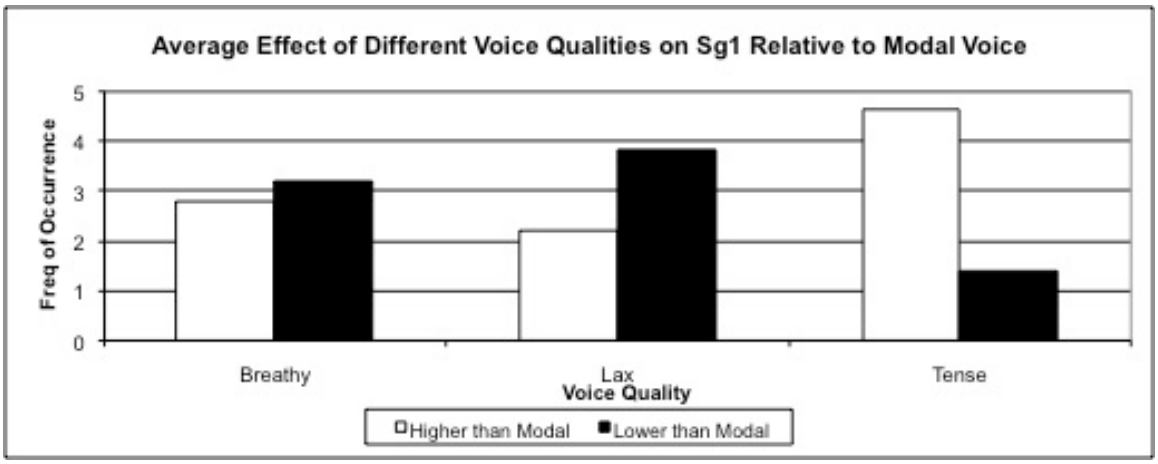

Figure 3.5 Averaged results on the effect of different voice qualities on Sg1 relative to modal phonation

3.5. Results for Second Subglottal Resonance

The averaged Sg2 data from the five repetitions for each target word are listed in Tables 3.12 – 3.16. The standard deviations across the target words pronounced in the same phonation are generally low with the exception of speaker M3, with value as high as 106.96Hz for breathy voiced words. However, in general the context-independent property still holds for all speakers, especially in modal phonation.

	Modal	Breathy	Lax	Tense
Hay'ed	1204.50	1219.60	1257.50	1242.40
Hide	1212.10	1212.10	1227.20	1221.60
Hoid	1244.30	1238.60	1204.90	1198.80
Bay	1204.50	1208.97	1255.03	1212.10
Buy	1234.80	1287.80	1265.10	1212.10
Boy	1212.10	1227.20	1257.50	1204.50
Average	1218.72	1232.38	1244.54	1215.25
Stdev	16.76	29.19	23.41	15.38

Table 3.12 Average Sg2 of each target word (in Hz) of M1

	Modal	Breathy	Lax	Tense
Hay'ed	1090.90	1019.90	1060.2	1098.40
Hide	1090.90	1022.70	1000.00	1060.60
Hoid	1090.90	1083.30	1060.60	1037.80
Bay	1145.42	1073.82	1098.82	1052.24
Buy	1085.20	1096.50	1125.00	1079.50
Boy	1064.74	1067.00	1060.20	1113.62
Average	1094.68	1060.54	1067.47	1073.69
Stdev	26.85	31.98	42.42	28.84

Table 3.13 Average Sg2 of each target word (in Hz) of M2

	Modal	Breathy	Lax	Tense
Hay'ed	1321.16	1565.50	1348.93	1314.35
Hide	1318.14	1350.35	1288.88	1307.36
Hoid	1292.10	1401.50	1318.15	1325.13
Bay	1315.10	1606.08	1521.10	1538.20
Buy	1356.26	1382.53	1387.82	1284.74
Boy	1319.64	1524.05	1356.02	1409.07
Average	1320.40	1471.67	1370.15	1363.14
Stdev	20.60	106.96	81.31	95.75

Table 3.14 Average Sg2 of each target word (in Hz) of M3

	Modal	Breathy	Lax	Tense
Hay'ed	1474.18	1463.56	1475.70	1454.44
Hide	1462.04	1465.88	1472.60	1522.68
Hoid	1454.08	1448.42	1475.70	1536.30
Bay	1478.66	1490.50	1465.86	1499.90
Buy	1484.80	1486.32	1487.82	1443.86
Boy	1468.12	1474.18	1459.04	1464.60
Average	1470.31	1471.48	1472.79	1486.96
Stdev	11.24	15.59	9.80	38.19

Table 3.15 Average Sg2 of each target word (in Hz) of F1

	Modal	Breathy	Lax	Tense
Hay'ed	1473.25	1503.75	1443.15	1412.85
Hide	1477.25	1575.70	1488.40	1534.05
Hoid	1518.90	1568.10	1553.00	1472.65
Bay	1446.90	1492.40	1424.15	1408.80
Buy	1458.30	1545.00	1503.75	1545.45
Boy	1462.20	1392.40	1446.90	1389.80
Average	1472.80	1512.89	1476.56	1460.60
Stdev	25.05	67.89	47.89	67.40

Table 3.16 Average Sg2 of each target word (in Hz) of F2

To compare the four different phonations, the difference in average Sg2 between modal and each of the remaining vocal qualities are computed in Tables 3.17 – 3.21. To provide better evaluation of the data (as discussed for Sg1), Figures 3.6 – 3.8 display graphical representation of the aforementioned table on the whether average Sg2 difference is greater than or less than 0Hz.

	Breathy	Lax	Tense
Hay'ed	15.10	53.00	37.90
Hide	0.00	15.10	9.50
Hoid	-5.70	-39.40	-45.50
Bay	4.47	50.53	7.60
Buy	53.00	30.30	-22.70
Boy	15.10	45.40	-7.60
Average	13.66	25.82	-3.47
Stdev	20.96	34.99	28.83

Table 3.17 Average Sg2 difference (in Hz) for each voice quality relative to modal voice of M1

	Breathy	Lax	Tense
Hay'ed	-71.00	-30.70	7.50
Hide	-68.20	-90.90	-30.30
Hoid	-7.60	-30.30	-53.10
Bay	-71.60	-46.60	-93.18
Buy	11.30	39.80	-5.70
Boy	2.26	-4.54	48.88
Average	-34.14	-27.21	-20.98
Stdev	40.04	43.49	49.51

Table 3.18 Average Sg2 difference (in Hz) for each voice quality relative to modal voice of M2

	Breathy	Lax	Tense
Hay'ed	244.34	27.76	-6.81
Hide	32.21	-29.26	-10.78
Hoid	109.40	26.05	33.03
Bay	290.98	206.00	223.10
Buy	26.27	31.56	-71.52
Boy	204.41	36.38	89.43
Average	151.27	49.75	42.74
Stdev	111.86	80.26	103.13

Table 3.19 Average Sg2 difference (in Hz) for each voice quality relative to modal voice of M3

	Breathy	Lax	Tense
Hay'ed	-10.62	1.52	-19.74
Hide	3.84	10.56	60.64
Hoid	-5.66	21.62	82.22
Bay	11.84	-12.80	21.24
Buy	1.52	3.02	-40.94
Boy	6.06	-9.08	-3.52
Average	1.16	2.47	16.65
Stdev	8.13	12.65	47.55

Table 3.20 Average Sg2 difference (in Hz) for each voice quality relative to modal voice of F1

	Breathy	Lax	Tense
Hay'ed	30.50	-30.10	-60.40
Hide	98.45	11.15	56.80
Hoid	49.20	34.10	-46.25
Bay	45.50	-22.75	-38.10
Buy	86.70	45.45	87.15
Boy	-69.80	-15.30	-72.40
Average	40.09	3.76	-12.20
Stdev	59.76	31.39	66.95

Table 3.21 Average Sg2 difference (in Hz) for each voice quality relative to modal voice of F2

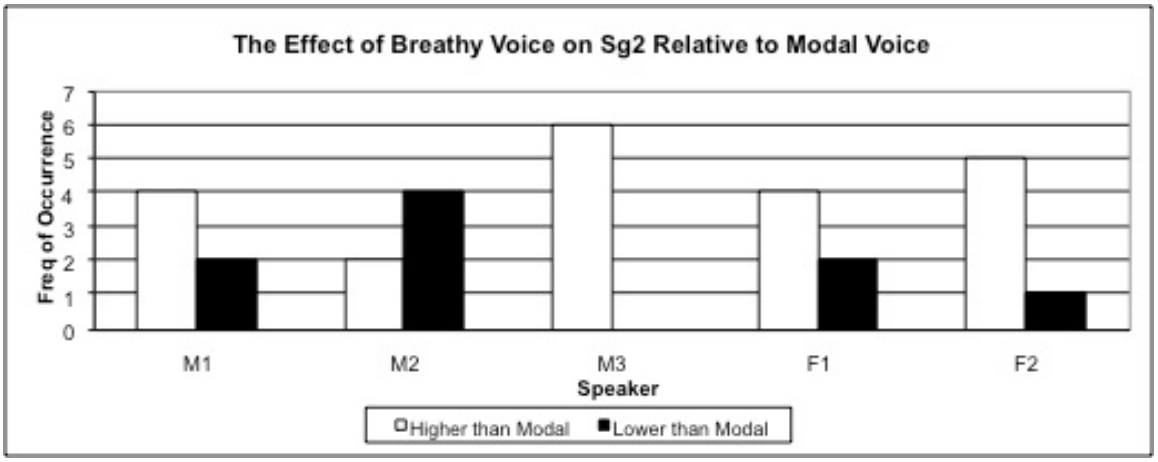

Figure 3.6 The effect of breathy voice on Sg2 relative to modal voice

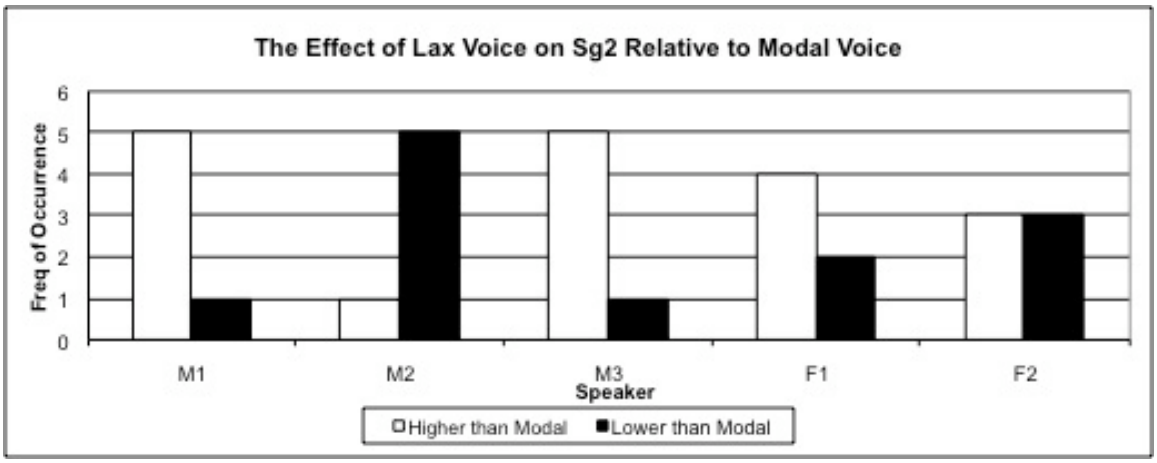

Figure 3.7 The effect of lax voice on Sg2 relative to modal voice

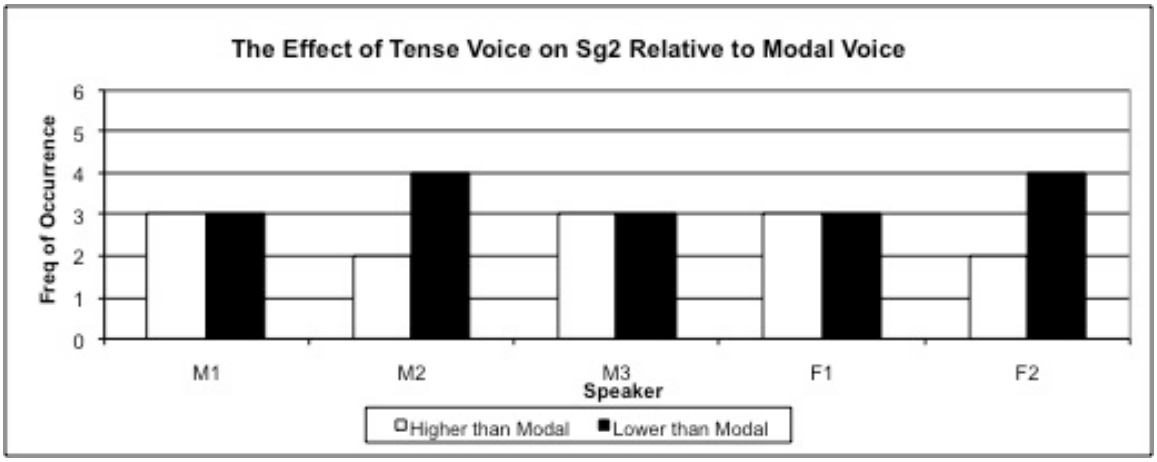

Figure 3.8 The effect of tense voice on Sg2 relative to modal voice

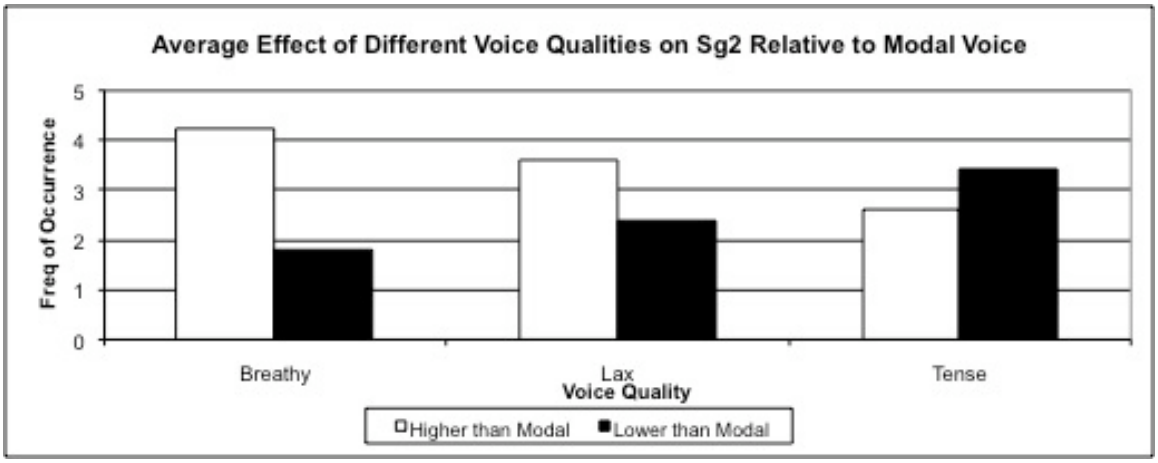

Figure 3.9 Averaged results on the effect of different voice qualities on Sg2 relative to modal voice

Interestingly, the results of the effect of voice qualities on Sg2 are actually the opposite of Sg1. Four out of five speakers have shown equal or more occurrences to have an increase in Sg2 from modal voice to breathy voice. This pattern also occurs for the lax voices. For tense voice, all speakers have equal or lower Sg2 than that of modal voice. The summary of this phenomenon is shown in Figure 3.9.

3.6. Results for Acoustic Energy for Subglottal Resonances

The acoustic energy information of the subglottal resonances are evaluated from the spectrograms of the accelerometer signals. Figures 3.10 are the spectrograms for the target word 'Hoid' spoken by Speaker F1 in four different phonations: modal, breathy, lax, and tense. With the spectrogram of modal voice as the baseline, the acoustic energies for both breathy and lax are lower and attenuate faster; on the contrary, the acoustic energy for tense voice is significantly higher and does not attenuate at all. These results have been shown consistently for all speakers in the study and henceforth supporting the hypothesis stated in Section 3.3.

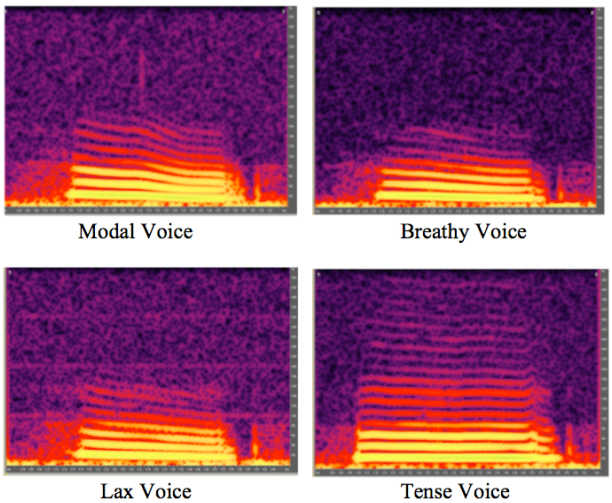

Figure 3.10 Spectrograms of the target word 'Hoid' spoken by Speaker F1

3.7. Summary

In this chapter, four different voice qualities: modal, breathy, lax, and tense, are introduced and examined in an exploration study of their impact on subglottal resonances. The data are collected using an accelerometer from three male and two female speakers with a total of 120 utterances for each speaker. After manually processing the waveforms in Adobe Audition, the hypothesized effects of breathy, lax, and tense voice qualities relative to modal voice quality have been supported in acoustic energy and Sg1. The acoustic energy of lax voices is lower relative to modal voice while the acoustic energy of tense voices is higher relative to modal voice. For Sg1 results, four out of five speakers have shown that Sg1 of both breathy voices and lax voices to be lower than modal voices, and Sg1 of tense voices to be higher than modal voices due to the area of glottal flow. On the contrary, these results are not present in Sg2. Further experimentation is still required to formulate a generalizable conclusion due to the following issues:

First, to understand the relationships between different modes of phonation and subglottal resonances fully, a larger database is required. Ideally, the participants of the study should be phoneticians that have great control over their speech productions.

Second, the analysis in this study is performed manually due to inaccuracy of automatic formant-tracking algorithms. However, this potentially introduces inherent human error in the estimated subglottal resonances.

CHAPTER 4

CONCLUSION

The correlations between subglottal resonances and formant frequencies have enabled the usage of subglottal resonances information in speech technology. In Chapter 2, an automatic Sg2 estimation algorithm is implemented to evaluate the warping factors of linear and nonlinear frequency warping functions used in speaker normalization. The combination of Bark-shift based nonlinear warping function with Sg2-based warping factor has provided the best performance among the conventional speaker normalization techniques (linear and nonlinear warping function using ML-based grid searches) in children's ASR. In addition, the proposed speaker normalization technique has low computational cost and performed well for limited data sets.

In Chapter 3, to determine whether variations in speech production can affect the subglottal resonances (Sg1 and Sg2), an exploration study is conducted. Four different phonations are examined in this study: modal voice, breathy voice, lax voice, and tense voice. Based on prior works on the relationships between voice qualities and subglottal resonances using mathematical and pseudo modeling, hypotheses are formulated and validated with accelerometer data of five English speakers. The results for Sg1 have supported the hypotheses of breathy, lax, and tense voice qualities, in which subglottal resonances of breathy and lax phonations would be lower than that of modal phonation

while subglottal resonances of tense phonations would be higher than that of modal phonation due to glottal air flow. On the contrary, the results for Sg2 implied the opposite effect to occur. Finally, the results for acoustic energy of subglottal resonances have supported the hypotheses, in which lax voices have less acoustic energy than modal voices and tense voices have more acoustic energy than modal voices.

Since the subglottal resonances are manually estimated from limited data, further experimentation is still required to formulate a solid understanding between voice qualities and subglottal resonances. The ideal exploration study would contain the following: a larger database with speech and accelerometer data of phoneticians to provide better control of phonations; alternative manual analysis methods such as evaluating the formants frame by frame in spectra domain can be applied with spectrogram readings to detect subglottal resonances of breathy and tense phonations due to acoustic energy issues; and an effective and accurate automatic analysis of the accelerometer data should be used to reduce inherent human error.

REFERENCES

Austin, S. F., and Titze, I, R., "The effect of subglottal resonance upon vocal fold vibrations," *Journal of Voice*, Vol. 11, 1997, pp. 391-402.

Bladon R., Henton, C., and Pickering, J., "Towards an auditory theory of speaker normalization," *Language & Communication*, Vol. 4, No. 1, 1984, pp. 59-69.

Chi, X. and Sonderegger, M., "Subglottal coupling and its influence on vowel formants," *Journal of Acoustical Society of America*, Vol. 102, 2007, pp. 2380-2389.

Cranen, B. and Boves, L., "On subglottal formant analysis," *Journal of Acoustical Society of America*, Vol. 81, 1987, pp. 734-746.

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

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

Jung, Y., "Subglottal effects on the vowels across language: Preliminary study on Korean," *Journal of Acoustical Society of America*, Vol. 125, 2009, pp. 2320-2327.

Laver, J., *The Phonetic Description of Voice Quality*, Cambridge University Press: New York, NY, 1980.

Lulich, S. M., "Subglottal resonances and distinctive features," *J. Phonetics*, 2009, doi:10.1016/j.wocn.2008.10.006.

Lulich, S. M. and Chen, N., "Automatic classification of consonant-vowel transitions based on subglottal resonances and second formant frequencies," *Proceedings of Meetings on Acoustics*, Vol. 6, 2009.

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 *Proceedings of LabPhon*, Vol. 11, 2008, pp. 91-92.

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

Stevens, K. N., *Acoustic Phonetics*. MIT Press: Cambridge, MA, 1998.

Umesh, S., Cohen, L., and Nelson, D., "Frequency waqrping and the Mel scale," *IEEE Signal Processing Letters,* Vol. 9, No. 3, March 2001, pp. 104-107.

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