Automatic detection of the second subglottal resonance and its application to speaker normalization[a]

Shizhen Wang[b]
Department of Electrical Engineering, University of California, Los Angeles, California 90095

Steven M. Lulich
Speech Communication Group, MIT, Cambridge, Massachusetts 02139

Abeer Alwan
Department of Electrical Engineering, University of California, Los Angeles, California 90095

(Received 24 December 2008; revised 24 September 2009; accepted 25 September 2009)

Speaker normalization typically focuses on inter-speaker variabilities of the supraglottal (vocal tract) resonances, which constitute a major cause of spectral mismatch. Recent studies have shown that the subglottal airways also affect spectral properties of speech sounds, and promising results were reported using the subglottal resonances for speaker normalization. This paper proposes a reliable algorithm to automatically estimate the second subglottal resonance (Sg2) from speech signals. The algorithm is calibrated on children's speech data with simultaneous accelerometer recordings from which Sg2 frequencies can be directly measured. A cross-language study with bilingual Spanish-English children is performed to investigate whether Sg2 frequencies are independent of speech content and language. The study verifies that Sg2 is approximately constant for a given speaker and thus can be a good candidate for limited data speaker normalization and cross-language adaptation. A speaker normalization method using Sg2 is then presented. This method is computationally more efficient than maximum-likelihood based vocal tract length normalization (VTLN), with performance better than VTLN for limited adaptation data and cross-language adaptation. Experimental results confirm that this method performs well in a variety of testing conditions and tasks. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3257185]

PACS number(s): 43.72.Ne, 43.72.Fx [DOS] Pages: 3268–3277

I. INTRODUCTION

A major cause of performance degradation in automatic speech recognition (ASR) systems is inter-speaker variations in acoustic characteristics (fundamental and formant frequencies, etc.), which are mostly caused by differences in the supraglottal speech production system (vocal tract and vocal fold apparatus). Accordingly, speaker normalization, which aims to reduce these acoustic variabilities, typically focuses on supraglottal variations. Vocal tract length normalization (VTLN) is one of the most popular methods for reducing the effects of speaker-dependent vocal tract variability through a speaker-specific frequency warping function (linear, bilinear, or piece-wise linear) (Wegman *et al.*, 1996; Lee and Rose, 1998; Pitz and Ney, 2003; Umesh *et al.*, 2005; McDonough *et al.*, 2004; McDonough, 2000). Warping factors are typically estimated based on the maximum likelihood (ML) criterion over the adaptation data through an exhaustive grid search or warping-factor specific models (Wegman *et al.*, 1996; Lee and Rose, 1998). Linear frequency warping can be implemented directly in the power spectrum domain or in the cepstral domain through the linearization of VTLN (Pitz and Ney, 2003; Umesh *et al.*, 2005; McDonough *et al.*, 2004). Along with the linearization of VTLN, the warping factor can be estimated using the expectation maximization algorithm with an auxiliary function (McDonough *et al.*, 2004). Other frequency warping functions have also been studied. A class of transforms, known as all-pass transforms (APTs), was proposed to perform VTLN and studied in detail in McDonough, 2000 for two classes of conformal maps, namely, rational all-pass transforms and sine-log all-pass transforms. It was demonstrated that using multiple-parameter warping functions is more effective than single-parameter ones.

Another way to reduce spectral variability is to explicitly align spectral formant positions or formant-like spectral peaks, especially the third formant (F_3), and to define the warping factors as formant frequency ratios (Eide and Gish, 1996; Gouvea and Stern, 1997; Claes *et al.*, 1998; Cui and Alwan, 2006; Zhan and Westphal, 1997; Wang *et al.*, 2007). In formant-based frequency warping methods, formant positions of different speakers are transformed into a normalized frequency space. Eide and Gish (1996) proposed a nonlinear warping function based on a parameter estimated using F_3 frequency. Zhan and Westphal (1997) extended this formant-based algorithm and compared the performance with ML-based methods. Gouvea and Stern (1997) explored the performance of frequency warping using the first three formant frequencies. Claes *et al.* (1998) proposed a linear approximation of VTLN for reasonably small warping factors estimated based on average F_3 values. Cui and Alwan (2006) proposed

[a] Portions of this article were presented at ICASSP 2008 and Interspeech 2008.

[b] Author to whom correspondence should be addressed. Electronic mail: szwang@ee.ucla.edu

a novel spectral formant-like peak alignment method, with a focus on F_3, to reduce spectral mismatch between adults and children's speech (Cui and Alwan, 2006; Wang *et al.*, 2007). Based on the idea of frequency transformation for digital filters, Wang *et al.* (2004) treated formant structures as filters and developed a bilinear transform with parameters estimated using average F_3 frequency and bandwidth values. Due to coarticulation, clarity, speed, and other factors, formant frequencies vary considerably within an utterance and thus make the performance of formant normalization content-dependent.

Besides the effects of the supraglottal system, recent studies show that the subglottal airways also affect spectral properties of speech sounds (Hanson and Stevens, 1995; Stevens, 1998; Sonderegger, 2004; Chi and Sonderegger, 2004; Lulich, 2006; Chi and Sonderegger, 2007; Lulich *et al.*, 2007). The coupling between the supraglottal and subglottal systems has been shown to be non-negligible when a vocal tract formant approaches a subglottal resonance in frequency. At such a point, the formant prominence amplitude will be attenuated and the prominent spectral peak will jump in frequency to skip the subglottal resonance. Many studies have been done to model and analyze the subglottal resonances for adults' speech, particularly focusing on American English, although a few studies in Korean and German have shown similar results (Jung, 2008; Madsack *et al.*, 2008).

Children's speech analysis and recognition have drawn increasing attention for educational purposes, and more efforts have been devoted to ASR's applications on children's speech. Due to developmental changes in vocal tract and vocal ford apparati, children's speech demonstrates high acoustic variabilities, which makes children's ASR more challenging compared to adults' ASR. The performance of an ASR system developed for adult speech decreases drastically when employed to recognize children's speech. Furthermore, recognition performance for children is usually lower than that achieved for adults even when using a recognition system trained on children's speech. Such challenges require further studies on children's speech.

In this paper, we focus on children's speech and explore the hypothesis that subglottal resonances can be used for speaker normalization, much like formant alignment techniques. We first describe the subglottal system and outline the theory of its effects on speech. We then use this theory to implement an automatic detector of the second subglottal resonance. Third, we address the question whether subglottal resonances are constant for a given speaker regardless of the language being spoken. Finally, we implement a speaker normalization scheme based on the second subglottal resonance (Sg2) and evaluate its performance on several tasks. We compare the performance with VTLN and demonstrate the effectiveness of this method.

II. EFFECTS OF COUPLING TO THE SUBGLOTTAL ACOUSTIC SYSTEM

The configuration of the acoustic system below the glottis consists of the trachea, bronchi, and lungs. Similar to the vocal tract, the acoustic input impedance of the subglottal system is characterized by a series of poles (or resonances) and zeros. Unlike the supraglottal system, however, the configuration of the subglottal system is essentially fixed and thus the poles and zeros are expected to remain constant for any given speaker. Like formant frequencies, subglottal resonances are generally higher for female speakers than for male speakers, and there are substantial individual differences from speaker to speaker. It has been shown that the lowest three subglottal resonances are around 600, 1450, and 2200 Hz for adult males, and 700, 1600, and 2350 Hz for adult females (Stevens, 1998).

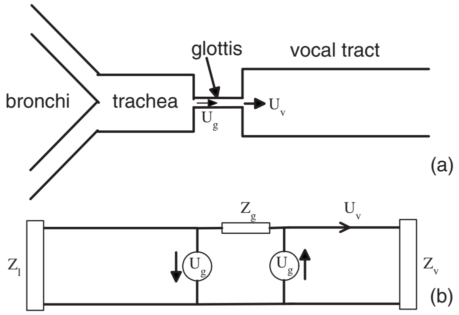

FIG. 1. Schematic model of vocal tract with acoustic coupling to the trachea through the glottis (a) and the equivalent circuit model (b) (Adapted from Stevens, 1998).

Figure 1 shows a schematic model of vocal tract coupling to the trachea through the glottis and its equivalent circuit model, where Z_l is the impedance of the subglottal system, Z_g is the glottal impedance, Z_v is the impedance looking into the vocal tract from the glottis, U_g is the volume velocity through the glottis, and U_v is the airflow into the vocal tract. Coupling between the subglottal and supraglottal airways is thought to occur primarily when the glottis is open, such as during a voiceless consonant or the open phase of glottal vibration in a voiced sound, although Lulich (2009) and Lulich *et al.* (2009) suggested that coupling may also occur when the vocal folds are closed, either by means of a posterior glottal opening or the vocal fold tissue itself. During coupling, each subglottal resonance contributes a pole-zero pair to the speech spectrum. The frequency of the zero is the same as that of the subglottal resonance, while the pole is shifted upward in frequency away from the resonance and depends somewhat on the vocal tract configuration. This is because the zero is a function only of the part of the entire system behind the source (that is, the subglottal airways), while the pole is a function of the entire system, including the subglottal and supraglottal airways (Lulich, 2009; Chi and Sonderegger, 2007).

The pole-zero pair introduced in the speech spectrum around Sg2 generally falls within the range of 1300–1500 Hz for adult males, and between 1400 and 1700 Hz for adult females (Stevens, 1998). It is somewhat higher in frequency for children (Jung *et al.*, 2008). When F2 crosses the Sg2 pole-zero pair, F2 can jump in frequency or diminish in amplitude, or both, resulting in a discontinuity in the F2 trajectory (Chi and Sonderegger, 2007). This is illustrated in Fig. 2 for an 8-year-old girl speaking the word *boy*, and it is schematically represented in Fig. 3. In both figures, F2 rises from a low frequency to a high frequency, crossing the Sg2 pole-zero pair along the way. The F2 discontinuity in

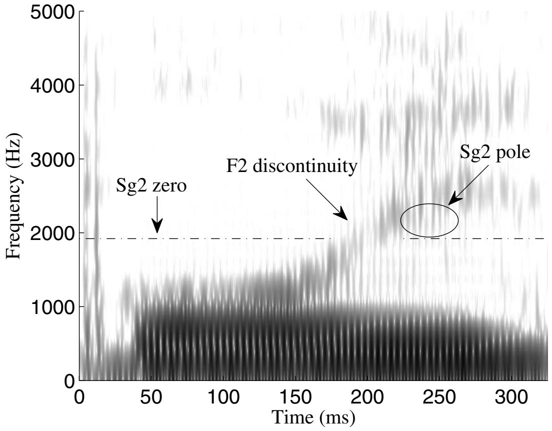

FIG. 2. Spectrogram for the word *boy* from an 8-year-old girl. The second subglottal resonance Sg2 for this speaker is 1920 Hz.

Fig. 2 is marked by a diminished amplitude in the vicinity of the zero. The Sg2 pole has a very low amplitude except during the time when F2 is nearby. In Fig. 2 the diffuse energy between F2 and the zero at 250 ms is likely due to the Sg2 pole, its amplitude decreasing as F2 continues to rise.

Recent studies (Lulich, 2006; Lulich *et al.*, 2007; Sonderegger, 2004) have shown that the acoustic contrasts for some phonological distinctive features are dependent on the subglottal resonances. As illustrated in Fig. 4, for example, the vowel feature [back][1] is dependent on the frequency of Sg2, such that a vowel with F2 > Sg2 is [−back] and a vowel with F2 < Sg2 is [+back]. The ability of Sg2 to underlie the distinctive feature [back] is likely derived from the fact that Sg2 is roughly constant for a given speaker. Subglottal resonances could potentially be affected by lung volume, larynx height, and glottal configuration. Lung volume has been shown not to significantly affect the subglottal resonances in one study (Cheyne, 2001), and reported accelerometer measurements of subglottal resonances across utterances (in which phonetic content was varied and voice quality was uncontrolled—both of which may affect laryngeal height and glottal configuration) have had standard deviations on the order of 30 Hz or less (Chi and Sonderegger, 2007). Thus, although the influence of lung volume, larynx height, and glottal configuration on subglottal resonances invites further research, the available evidence appears to indicate that subglottal resonances are roughly constant under normal speaking conditions.

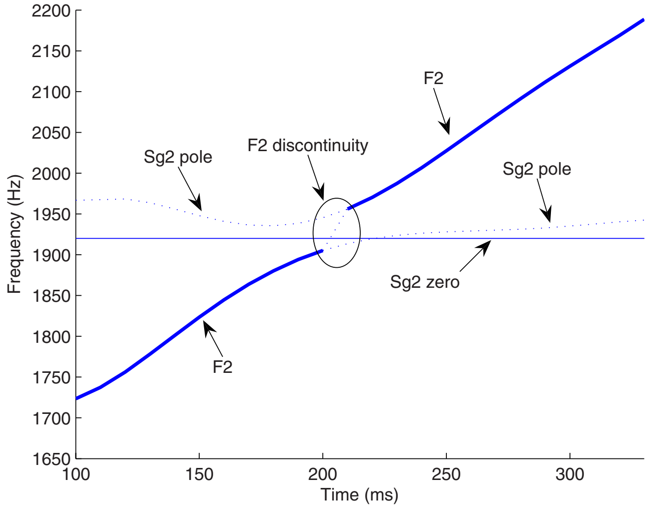

FIG. 3. (Color online) Illustration of the F2 discontinuity caused by Sg2. The bold solid line corresponds to the most prominent spectral peak (F2), which has a jump in frequency and a decrease in amplitude when F2 is crossing the subglottal resonance Sg2. The dotted line represents the Sg2 pole, which varies somewhat in frequency and amplitude when F2 is nearby. The horizontal thin solid line represents the Sg2 zero, which is roughly constant (adapted from Stevens, 1998).

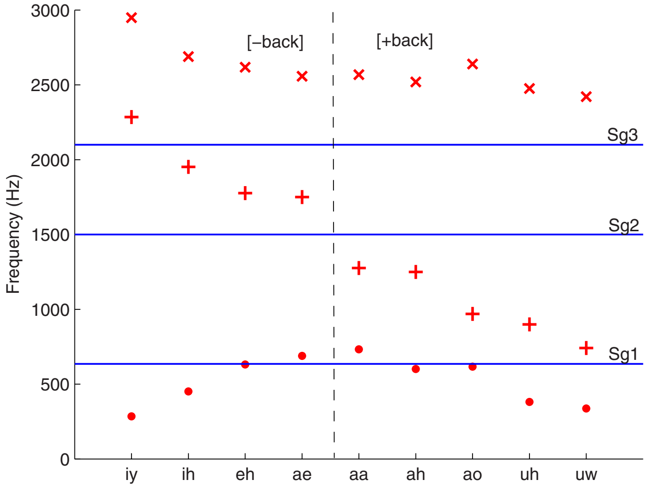

FIG. 4. (Color online) Illustration of the relative positions of vowel formants F1 (·), F2 (+), and F3 (×) and the subglottal resonances (Sg1, Sg2, and Sg3) for an adult male speaker. For the vowels /i, ɪ, ɛ, æ/ F2 > Sg2, and they are therefore [−back]. For the vowels /a, ʌ, ɔ, ʊ, u/ F2 < Sg2, and they are therefore [+back] (adapted from Lulich, 2006).

For this reason, Sg2 might be useful in speaker normalization, since it is context independent but speaker dependent. Sg1 and Sg3 have also been claimed to play a role in distinguishing different classes of speech sounds, but Sg2 has been more thoroughly studied. In this paper, therefore, we focus on Sg2 estimation and its application to speaker normalization.

III. ESTIMATING THE SECOND SUBGLOTTAL RESONANCE

A. Automatic estimation of Sg2 frequency

As noted above, when F2 crosses Sg2, there is a discontinuity in the F2 trajectory. Based on this discontinuity, an automatic Sg2 detector (Sg2D1) was developed in Wang *et al.* (2008a). The Snack sound toolkit (Snack, 2005) was used to generate the F2 trajectory. All experiments were done in clean conditions. The tracking parameters were specifically tuned to provide reliable F2 tracking results on children's speech. Manual verification and/or correction were applied through visually checking the tracking contours against the spectrogram. (Note that this method is limited by the accuracy of the formant tracker, which is known to encounter difficulties in high-pitched speech such as that produced by children.) The F2 discontinuity was detected based on the smoothed first order difference of the F2 trajectory, as shown

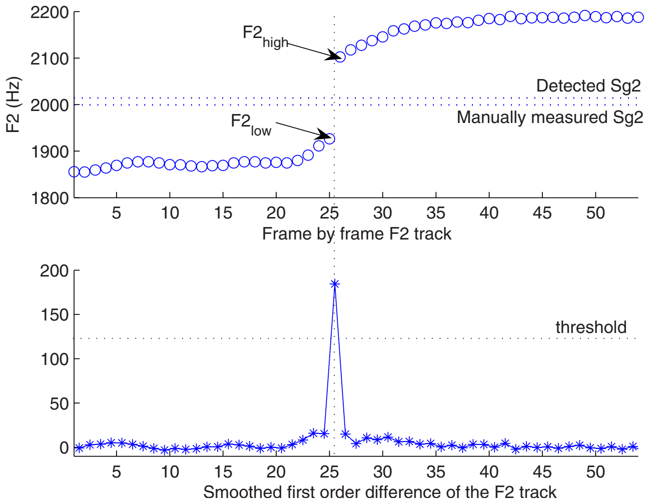

FIG. 5. (Color online) An example of the detection algorithm applied in Wang *et al.*, 2008.

in Fig. 5. If the F2 values on the high and low frequency sides of the discontinuity are $F2_{high}$ and $F2_{low}$, respectively, then Sg2D1 is estimated as

$$\widehat{Sg2} = \frac{F2_{high} + F2_{low}}{2}. \quad (1)$$

If no such discontinuity was detected, Sg2D1 used the mean F2 over the utterance. In many such cases, such as during a monophthong, F2 is consistently above or below Sg2, and the mean F2 value is either too high or too low. Thus, the estimated Sg2 values are dependent on the speech sound analyzed. Furthermore, discontinuities in F2 may arise from other factors beside the subglottal resonances, including pole-zero pairs from the interdental spaces (Honda *et al.*, 2009). These discontinuities occur a few hundred hertz higher than Sg2 discontinuities, but are sometimes more prominent than Sg2 discontinuities and can therefore be mistakenly detected as Sg2.

To address both issues, we developed an improved Sg2 estimation algorithm (Sg2D2) (Wang *et al.*, 2008b). We first detected F3 and obtained an estimate of Sg2 using a formula derived in Lulich, 2009:

$$\widetilde{Sg2} = 0.636 \times F3 - 103. \quad (2)$$

Note that the derivation of this formula was based on a linear regression on children's speech data which have available simultaneous accelerometer recordings, and its extension to adults' speech may still need further refinements.

We then searched for a discontinuity within ±100 Hz of this estimate using the original algorithm. The range ±100 Hz was chosen based on calculated standard deviations of Sg2 on the calibration data. If no discontinuity in this range was found, $\widetilde{Sg2}$ was used. If a discontinuity was found, we estimated Sg2 using the following equation:

$$\widehat{Sg2} = \beta \times F2_{high} + (1-\beta) \times F2_{low}, \quad (3)$$

where 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 was calibrated over the data described below using the minimum mean square error criterion:

TABLE I. Comparison of Sg2 estimates for two algorithms over various vowel contents, where Sg2D1 refers to the algorithm used in Wang *et al.* (2008a), Sg2D2 is the new algorithm, Sg2M is the manual measurement from speech spectrum, and Sg2Acc is the "ground truth" measurement from the accelerometer signal. For each algorithm the average Sg2 estimates (Hz) are shown (with standard deviations in parentheses). The two speakers with an " *" symbol are those used for calibration.

Speaker	Sg2D1	Sg2D2	Sg2M	Sg2Acc
1	2135 (531)	2194 (95)	2193 (97)	2176
2	2115 (334)	1766 (137)	1719 (112)	1646
3	2586 (467)	2718 (143)	2634 (135)	2679
4	2098 (358)	1823 (151)	1781 (129)	1614
5*	2065 (267)	2021 (79)	2013 (76)	1970
6*	1612 (251)	1689 (72)	1681 (65)	1648

$$\beta = \arg\min E\{(\widehat{Sg2} - Sg2)^2\} \quad (4)$$

and was found to be 0.65 in our experiments.

B. Calibration of the Sg2 estimation algorithm

To verify and calibrate our Sg2 estimation algorithm, acoustic data were collected from six female children aged 2–17 years old (speakers G1–G6 in Lulich, 2009). The children were native speakers of American English and all of them except the youngest were recorded repeating the phrase "hVd, say hVd again" three times for each of the vowels [i], [ɪ], [ɛ], [æ], [a], [ʌ], [o], [ʊ], and [u]. The subjects also recited the alphabet, counted to 10, and recited a few short sentences. The recording list was presented in random order and verbally prompted by the experimenter. The youngest speaker (G1) was recorded counting to 10, reciting the alphabet, and answering questions of the sort "What is this?," in which the experimenter pointed to his hand or head, for instance. All utterances were recorded in a sound-isolated chamber using a SHURE BG4.1 uni-directional condenser microphone and an accelerometer. Both the speech and accelerometer signals were digitized at 16 kHz. Microphone signals of each speaker were used to measure average F3 and the discontinuity in the F2 track. An independent direct measure of the average Sg2 for each speaker was obtained from an accelerometer signal. The accelerometer was attached to the skin of the neck below the larynx so that the measured vibration of the neck skin is directly related to the acoustic pressure variations in the air column at the top of the trachea (Cheyne, 2001; Chi and Sonderegger, 2007). The accelerometer signal can therefore act as a stand-in for the subglottal input impedance, in which the subglottal resonances appear as formants in the accelerometer spectrum.

The detection algorithms Sg2D1 and Sg2D2 were calibrated (to estimate discontinuity thresholds for both Sg2D1 and Sg2D2, and β for Sg2D2) on data from two of the recorded children and tested on the remaining four. The values measured from the accelerometer data were used as the "ground truth" Sg2 frequencies (henceforth denoted by "Sg2Acc"). The average Sg2 estimates (with standard deviations) over various vowel contents are shown in Table I. Compared to Sg2D1, the updated algorithm Sg2D2 estimates

TABLE II. Detailed comparison of Sg2 estimates for the two algorithms on two speakers. For vowels above the double line, there are no discontinuities in the F2 trajectory, and Sg2D1 uses the mean F2 as Sg2, while Sg2D2 uses Eq. (2) ($\widetilde{Sg2}$) to make an estimate; for vowels below the double line, the F2 discontinuity is detectable, and Sg2D1 uses Eq. (1), while Sg2D2 uses Eq. (3). The row "Avg.(std)" shows the mean (and standard deviation) for each algorithm.

	Speaker 1 (age 6) Sg2Acc: 2176 Hz		Speaker 2 (age 13) Sg2Acc: 1646 Hz	
Vowel	Sg2D1	Sg2D2	Sg2D1	Sg2D2
[i]	2987	2312	2563	1971
[ɪ]	2515	2306	2439	1909
[e]	2894	2115	2629	1998
[ɛ]	2799	2291	2378	1867
[æ]	2382	2289	2350	1863
[a]	1599	2020	1796	1700
[ʌ]	1687	2243	1948	1704
[o]	1512	2185	1497	1613
[ʊ]	1578	2228	1964	1717
[u]	1739	2071	1825	1631
[au]	1841	2114	1974	1617
[aɪ]	2103	2170	2072	1709
[ɔɪ]	2115	2183	2063	1659
Avg.(std)	2135 (531)	2194 (95)	2115 (334)	1766 (137)

Sg2 much better with less variance across vowels. The observed standard deviation values of Sg2D2 are similar to those from manually measured Sg2's (Sg2M)[2] in this study and those found in other studies (Jung *et al.*, 2008).

The performance of these two algorithms was investigated in more detail for each vowel for two speakers and the results are shown in Table II and Fig. 6. As stated earlier, if no discontinuity in the F2 track is detected (as for the vowels above the double line, Table II), Sg2D1 uses the mean F2 as Sg2 and thus is highly dependent on vowel content. Sg2D2, on the other hand, uses a formula to estimate Sg2 from F3 which is less content-dependent than F2. In such cases, it can be seen that the formula in Sg2D2 gives much closer estimates to the ground truth, especially for mid and back vowels. For the case when there is a discontinuity in the F2 trajectory (as for the diphthongs below the double line), both algorithms work well when the F2 discontinuity is from Sg2, as for speaker 1. In this case, Sg2D1 gave an estimate within about 70 Hz of the true Sg2 value, while the Sg2D2 estimate was within less than 10 Hz. For speaker 2, where the most prominent F2 discontinuity was probably from the interdental space, Sg2D1 gave an estimate hundreds of hertz above the Sg2 value, while Sg2D2 roughly located the correct Sg2 value using Eq. (2). Thus, Sg2D2 is less prone to mistakenly detecting discontinuities not caused by Sg2. In addition to diphthongs, discontinuities in F2 should also be detectable in certain consonant-vowel transitions (Lulich, 2009). Since Sg2D2 performs consistently better than Sg2D1, we will focus only on Sg2D2 in the following experiments. As shown in Tables I and II, and Fig. 6, the proposed detector produces Sg2 estimates close to the ground truth. Also, as will be shown in (Sec. V), the estimated Sg2 helps to improve ASR's performance on children's speech, which is of primary interest to us.

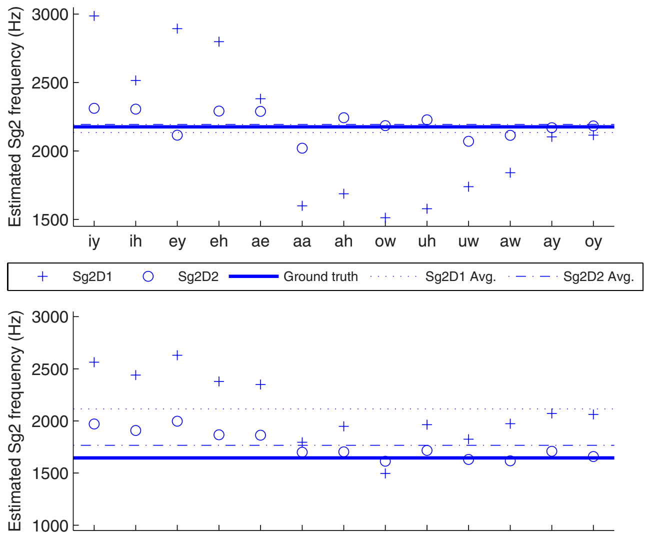

FIG. 6. (Color online) Comparison of Sg2 estimates for the two speakers in Table II: left panel for speaker 1 and right panel for speaker 2.

IV. VARIABILITY OF SUBGLOTTAL RESONANCE Sg2

The acoustic characteristics of children's speech have been shown to be highly different from those of adults' speech, in terms of pitch and formant frequencies, segmental durations, and temporal and spectral intra- and inter-speaker variabilities (Lee *et al.*, 1999; Huber *et al.*, 1999). Studies of subglottal resonances (Hanson and Stevens, 1995; Sonderegger, 2004; Chi and Sonderegger, 2004; Lulich, 2006; Chi and Sonderegger, 2007; Lulich *et al.*, 2007), however, have mainly focused on adults' speech in English with little effort devoted to children's speech or to other languages (but see Jung, 2008; Jung *et al.*, 2008; Madsack *et al.*, 2008). This section analyzes children's speech in English and Spanish, investigating the variabilities of Sg2 under different contents and across languages.

To examine the cross-language variability of Sg2 frequencies, we recorded a database (ChildSE) of 20 bilingual Spanish-English children (ten boys and ten girls) in the first or second grade (around 6 and 7 years old, respectively) from a bilingual elementary school in Los Angeles. The recorded speech consisted of words containing front, mid-, back, and diphthong vowels. There were four English words (*beat, bet, boot*, and *bite*) and five Spanish words (*calle* "street," *casa* "house," *quitar* "to take out," *taquito* "taco," and *cuchillo* "knife"), all of which were familiar to the children. Prior to the recording, children were instructed to practice as many times as they wanted. Both text and audio samples for each target word were available for prompt, and children decided what prompt they needed during recording and what language they wanted to record first. There were three repetitions for each word, and children spoke all the words in one language in a row with 3 s pause between words, and then repeated them. After they finished the recordings in one language, there was about a 1 min pause before they began the recordings in the other language. Recordings were made with 16 kHz sampling rate and 16-bit

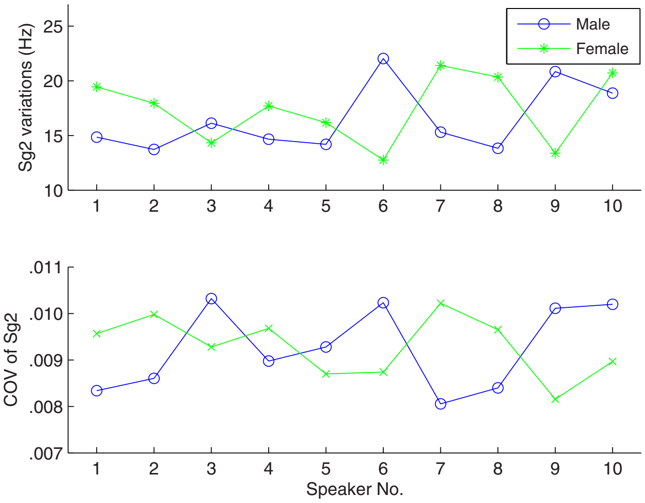

FIG. 7. (Color online) Average within-speaker Sg2 standard deviations and the COVs against contents and repetitions.

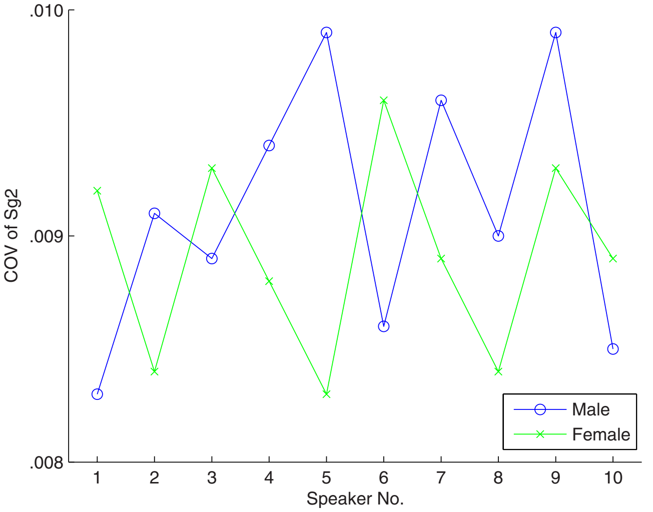

FIG. 8. (Color online) Cross-language within-speaker COV of Sg2 for ten boys and ten girls.

resolution. Like the English word *bite* [baɪt], the Spanish words *calle* [kaje] and *cuchillo* [kutʃijo] had obvious F2 discontinuities. We used these words with diphthongs to estimate Sg2 frequencies. Therefore, for each speaker, there were three English tokens and six Spanish tokens for the Sg2 estimation.

The within-speaker standard deviations were calculated on the Sg2 values estimated from the six Spanish tokens for each speaker. We also calculated the within-speaker coefficients of variation (COVs), a measure of dispersion of a probability distribution, which was computed as the ratio of the standard deviation to the mean Sg2 value for each speaker. As shown in Fig. 7, the within-speaker Sg2 standard deviations are around 20 Hz and the COV is less than 0.01. No significant difference in the COVs is observed between genders. A similar trend is observed for the within-speaker Sg2 standard deviations calculated from the English tokens. Compared to the COV of formant frequencies (Lee *et al.*, 1999), which are usually around 0.10, the COV of Sg2 is about one order of magnitude smaller. Therefore, the within-speaker Sg2 variability is negligible since they are sufficiently small compared to formant variabilities. This means that for a given speaker Sg2 is relatively constant against contents and repetitions.

Since Sg2 frequency for a given speaker does not depend on the contexts, we calculated the Sg2 COV for each speaker over the three English tokens and six Spanish tokens and viewed this as the Sg2 cross-language variabilities, as plotted in Fig. 8. The cross-language Sg2 COVs are less than 0.01, and there is no significant difference between genders. The cross-language COVs are similar to the within-speaker COVs, indicating that the cross-language effects are not significant for Sg2 frequencies and the Sg2 frequency for a given speaker is independent of languages.

Because of its invariability across speech content and language, Sg2 was judged to be applicable to speaker normalization. Since Sg2 is content-independent, we hypothesized that the performance of speaker normalization using Sg2 should be robust and independent of the amount of adaptation data available. This would make the Sg2 normalization method greatly suitable for limited data adaptation, which is often the case in ASR applications.

On the other hand, the language-independent property of Sg2 makes cross-language adaptation possible based on Sg2 normalization. Theoretically, with Sg2 normalization acoustic models trained in one language could be adapted with data in any other language,

V. SPEAKER NORMALIZATION RESULTS

Similar to formant normalization, the warping ratio for Sg2 normalization is defined as

$$\alpha = \mathrm{Sg2}_r/\mathrm{Sg2}_t, \qquad (5)$$

where $\mathrm{Sg2}_r$ is the reference Sg2 and $\mathrm{Sg2}_t$ is the Sg2 of the test speaker. The reference Sg2 is defined as the mean value of all the training speakers' Sg2's. The Sg2 values are detected using the Sg2D2 algorithm. In this section, we evaluate the content dependency of Sg2 normalization and also its use for cross-language normalization.

A. Comparison of VTLN and Sg2 frequency warping

Figure 9 shows F1, F2, and F3 values from a 9-year-old girl before and after warping using VTLN (Lee and Rose, 1998) and the Sg2 ratio. The line Sg2 is the reference second subglottal resonance for an adult male speaker (as in Fig. 4). Compared to Fig. 4, unwarped data demonstrate an obviously different pattern as to the relative positions of the formants with respect to the reference Sg2. For instance, the back vowels [ʊ] and [u] have higher F2 values than the reference Sg2, while in Fig. 4 F2's of all the back vowels lie below the Sg2 line. It is necessary to apply frequency warping to achieve the reference formant position pattern. Both VTLN (in circles) and Sg2 (in squares) warp the formants close to the reference pattern, although Sg2 warping yields a formant pattern more similar to the reference speaker's.

To examine the effects of warping in more detail, we plotted in Fig. 10 the reference F1, F2, and F3 values versus

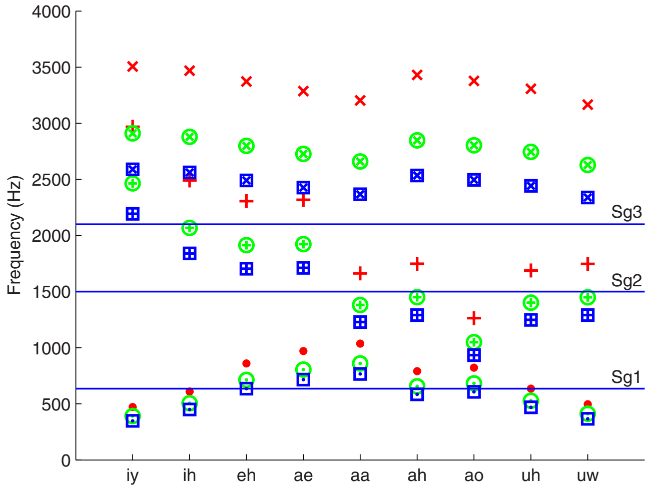

FIG. 9. (Color online) Vowel formants F1 (·), F2 (+), and F3 (×) before and after VTLN (in circles) and Sg2-based (in squares) warping for a 9-year-old girl's vowels. The lines "Sg1," "Sg2," and "Sg3" are the reference subglottal resonances from the same speaker as in Fig. 4.

the normalized values. Sg2 warping aligns the test speaker's formants more closely to the reference speaker's formants (Fig. 4), as indicated by the proximity of the data points to the diagonal line (with slope 1). In ASR such warping results in greatly reduced spectral mismatch between test and reference speakers, and thus can lead to better ASR performance.

B. Effectiveness of Sg2 normalization

Since VTLN has been shown to provide significant performance improvement on children's speech recognition, we first evaluate the subglottal normalization method on a connected digits recognition task of children's speech using the TIDIGITS database. Speech signals were segmented into 25 ms frames, with a 10 ms shift. Each frame was parametrized by a 39-dimensional feature vector consisting of 12 static Mel Frequency Cepstral Coefficients (MFCCs) plus log energy, and their first- and second-order derivatives. Acoustic Hidden Markov Models (HMMs) were monophone-based with three states and six Gaussian mixtures in each state. VTLN was implemented based on a grid search over [0.7, 1.2] with a stepsize of 0.01. The scaling factor producing maximal average likelihood was used to warp the frequency axis (Lee and Rose, 1998).

In this setup, acoustic models were trained on 55 adult male speakers and tested on 50 children. The baseline word accuracy is 55.76%. For each child, the adaptation data, which consisted of 1, 4, 7, 10, 13, or 16 digits, were randomly chosen from the test subset to estimate the Sg2 and VTLN warping factors. For comparison, we also evaluated the performance of manually measured Sg2, which in some sense can be viewed as the upper bound of this Sg2 normalization method. For each speaker, the manual Sg2 was measured from only diphthong words containing obvious F2 discontinuities in the spectrum and, independent of adaptation data, the same Sg2 value was applied for normalization. Figure 11 shows the recognition accuracy for VTLN, F3, and Sg2 warpings with various amounts of adaptation data, where Sg2M represents results using the manually measured subglottal resonance.

When the amount of adaptation data is small, Sg2 normalization offers better performance than VTLN. For instance, with only one digit for normalization, Sg2 normalization outperforms VTLN by more than 2%. VTLN outperforms Sg2D2 when more data are available, while the Sg2M provides slightly better performance to VTLN even with 16 adaptation digits. The improvements of Sg2 normalization over VTLN for up to ten adaptation digits are statistically significant for $p < 0.05$. Although automatic detection of Sg2 was fairly accurate, it was not exact and there is thus a gap between the performances of the automatic detection method and that of Sg2M. With more accurate Sg2 detection algorithms, we may expect closer performance to that of the manual Sg2.

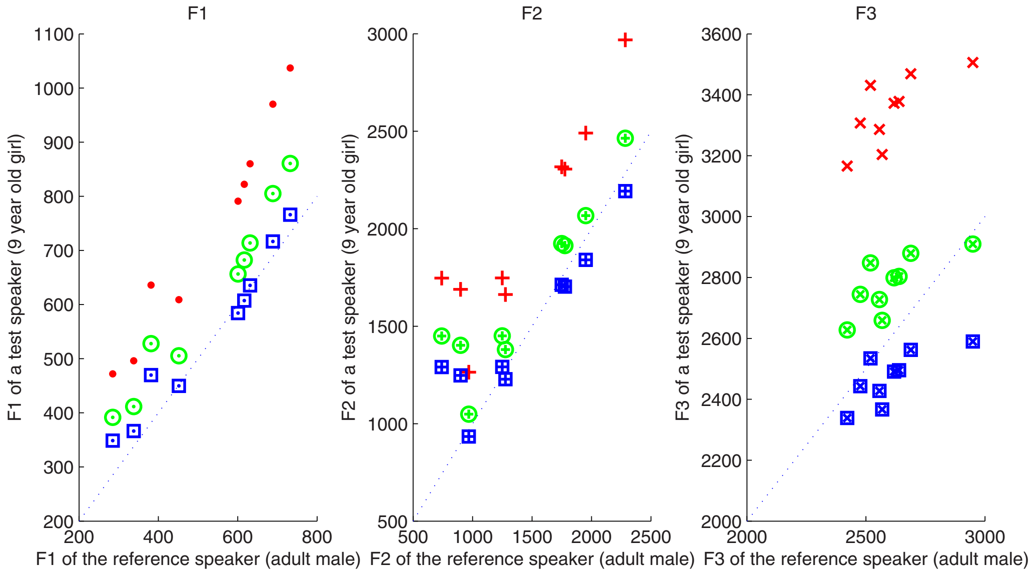

FIG. 10. (Color online) Vowel formants F1 (·), F2 (+), and F3 (×) from the reference speaker (Fig. 4) versus those from the test speaker (Fig. 9) before and after warping (VTLN in circles, Sg2 in squares). The dotted line is $y=x$ which means perfect match between reference and test speakers.

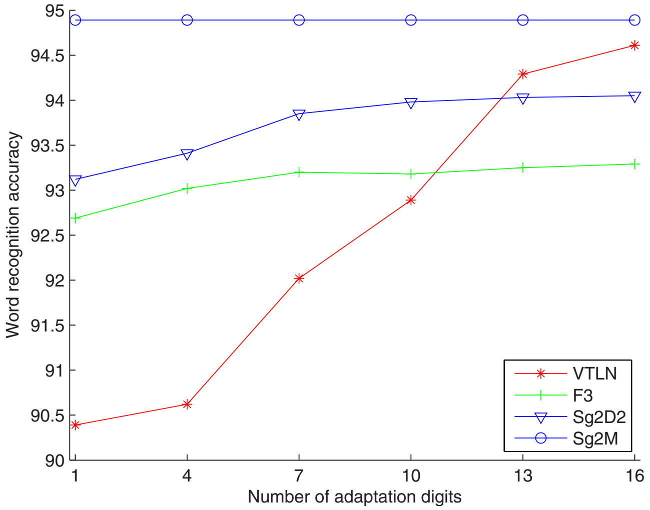

FIG. 11. (Color online) Speaker normalization performance on TIDIGITS with various amounts of adaptation data.

C. Comparison of vowel content dependency

As discussed in Sec. III B, Sg2 is not always detectable from acoustic signals, and thus the Sg2 detectability in adaptation data are important to the normalization performance. It is shown in Wang (2008a) that the normalization performance using Sg2D1 algorithm is highly content-dependent. To investigate the content dependency of the proposed algorithm Sg2D2, we evaluated its normalization performance on TIDIGITS database with one adaptation digit. For each child, the adaptation data were limited to only one digit but with varying vowels from front vowel (e.g., [ɪ] in six), central vowel (e.g., [ʌ] in one), back vowel (e.g., [u] in two) to diphthong (e.g., [aɪ] in five). The adaptation digits were chosen such that F2 discontinuities, if any, come only from vowel contents without any possible interferences from consonant-vowel transitions (Lulich, 2009).

The performance comparison for VTLN, F3, and Sg2 normalizations is shown in Fig. 12. It can be seen that the choice of adaptation data can potentially have an effect on the normalization performance for all three methods. Among them, VTLN is least affected by the choice of adaptation data (the performance standard deviation is 0.55), while F3 normalization is highly data dependent. The performance of Sg2 normalization is less content sensitive compared to F3 normalization, but more content-dependent than VTLN. We expect that the content dependency of Sg2 normalization will decrease with improved Sg2 detection algorithms. In spite of its greater content dependency, on average Sg2 normalization provides better performance than VTLN.

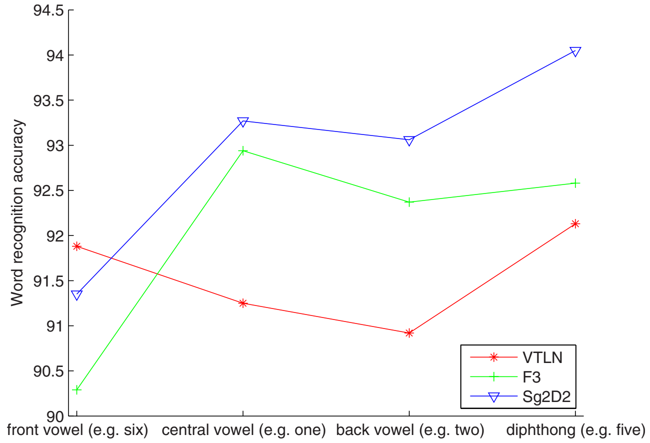

FIG. 12. (Color online) Performance comparison of VTLN, F3 and Sg2D2 using one adaptation digit with various vowel contents.

TABLE III. Performance comparison (word recognition accuracy) on RM1 with one adaptation utterance.

Accuracy	Mismatched	Matched
Baseline	59.10	92.47
F3	79.01	92.58
VTLN	86.65	93.91
Sg2	88.37	94.05

D. Performance on RM1 database

Since the TIDIGITS setup is a highly mismatched case, the experiments demonstrate the effectiveness of subglottal resonance-based speaker normalization. To further verify the effectiveness of this method, we also test the performance on a medium vocabulary recognition task using the DARPA Resource Management RM1 continuous speech database. As a next step, we tested the method on the RM1 database for both a medium-mismatched case and a matched case. Triphone acoustic models were applied with three states and four Gaussian mixtures per state using the same features as in the TIDIGITS experiments. For the mismatched case, HMM models were trained on 49 adult male speakers from the speaker independent (SI) portion of the database, and tested on 23 adult female speakers in the SI portion. The baseline word recognition accuracy was 59.10%. For the regular test on RM1, the HMM models were trained on the SI training portion of the database with 72 adult speakers, and tested on the SI testing set. The baseline performance was 92.47% word recognition accuracy. In both cases, the same utterance was used to estimate the Sg2 and VTLN warping factor for all speakers. Table III shows the results.

For the mismatched case, Sg2 normalization provides better performance than VTLN with about 1.5% absolute improvement. This improvement is statistically significant for $p < 0.01$. For the matched case, Sg2 normalization provides comparable performance to that of VTLN. From the computation point of view, Sg2 normalization is more efficient than VTLN, since VTLN relies on an exhaustive grid search over the warping factors to maximize the likelihood of the adaptation data, while for Sg2 normalization the main computational cost comes from formant tracking which can be estimated efficiently.

E. Cross-language speaker normalization

The language-independent property of Sg2 makes cross-language adaptation possible based on Sg2 normalization. In our experiments, training and test data were in English, while the adaptation data were in either English or Spanish. The warping factors were estimated from the adaptation data us-

TABLE IV. Performance comparison (word recognition accuracy) of VTLN and Sg2 normalization using English (four words) and Spanish (five words) adaptation data. The acoustic models were trained and tested using English data.

	Language of adaptation data	
Method	English	Spanish
VTLN	86.61	82.35
Sg2	86.59	85.97

ing Sg2D2 and applied to the test data to warp the spectrum. English adaptation data were collected for comparison.

The performance was evaluated on the Technology Based Assessment of Language and Literacy (TBall) project database (Kazemzadeh *et al.*, 2005), and the English high frequency words for first and second grade students were used in the test. Monophone acoustic models were trained on speech data from native English speakers. The test data were from the same 20 speakers as in the ChildSE. The ChildSE utterances (only one repetition) were used as adaptation data, and for each speaker there were four English words and five Spanish words for adaptation.

The typical text-dependent VTLN method using HMM recognizers for warping factor searching is not quite suitable in this scenario, because decoding Spanish speech with English phoneme models could itself introduce a systematic error due to different phonetic characteristics between these two languages. Instead, for a fair and reasonable comparison, text-independent VTLN is applied, which uses Gaussian mixture models (GMMs) for warping factor searching. A GMM with 512 mixtures was trained on English training set, and then applied to calculate the likelihood for each warping factor in the range [0.8, 1.2] with a step size of 0.01. The warping factor with the highest likelihood was chosen as the VTLN warping factor. Compared to the text-dependent VTLN used in Wang (2008), this text-independent method provides similar performance with English adaptation data, but much better for Spanish adaptation data. The subglottal resonance was estimated using Sg2D2 for each word, and the average was used as the speaker's Sg2 frequency. The Sg2 warping factor was calculated using Eq. (5).

The normalization performance is shown in Table IV for VTLN and Sg2 using English and Spanish adaptation data. When adaptation data are in English, which is the same language as for the acoustic models, Sg2 normalization and VTLN give comparably good results. For Spanish adaptation data, however, the performance of VTLN degrades, while the performance of Sg2 normalization remains similar as for English adaptation data. Sg2 normalization, therefore, produces more robust results than VTLN when performing cross-language adaptation. The performance difference between using Sg2D2 and using VTLN is statistically significant with Spanish adaptation data for $p < 0.01$.

VI. SUMMARY AND DISCUSSION

This paper presents a reliable algorithm for estimating the second subglottal resonance (Sg2) from acoustic signals. The algorithm provides Sg2 estimates very close to actual Sg2 values as determined from direct measurements using accelerometer data. With the proposed algorithm, Sg2 standard deviation over contents and languages was investigated with children's data for English and Spanish. Analysis shows that for a given speaker, the second subglottal resonance does not appear to vary with speech sounds, repetitions, and even across languages. Based on such observations, a speaker normalization method is proposed using the second subglottal resonance. This normalization method defines the warping factor as the ratio of the reference subglottal resonance over that of the test speaker.

A variety of evaluations show that the second subglottal resonance normalization performs better than or comparable to VTLN, especially for limited adaptation data. An obvious advantage of this method is that the subglottal resonances remain roughly constant for a specific speaker. This method is potentially independent of the amount of available adaptation data, which makes it suitable for limited data adaptation.

Cross-language experimental results shows that Sg2 normalization is more robust across languages than VTLN, and no significant performance variations are observed for Sg2 when the adaptation data are changed from English to Spanish. The fact that Sg2 is independent of language should make it possible to adapt acoustic models with available data from any language. The method is also computationally more efficient than VTLN.

The Sg2 variations found in this paper are similar to what has been reported elsewhere. However, given the small number of subglottal resonance studies, more data may need to be collected and analyzed in order to refine the characterization of subglottal resonance variability. For future work, we will further improve the accuracy of the Sg2 detector, evaluate the effectiveness of this method on a large vocabulary database, and test the performance in noisy conditions.

ACKNOWLEDGMENT

This work was supported in part by NSF Grant No. 0326214

[1] The place of articulation feature [+/−back] specifies the tongue positions during speech production: [+back] segments are produced with the tongue dorsum bunched and retracted slightly to the back of the mouth, while [−back] segments are bunched and extended slightly forward.

[2] The manual Sg2's were estimated through visually examining the speech spectrogram, and then applying Eq. (2) or Eq. (3) depending on the existence of F2 discontinuities.

Cheyne, H. A. (**2001**). “Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck,” Ph.D. thesis, MIT, Cambridge, MA.

Chi, X., and Sonderegger, M. (**2004**). “Subglottal coupling and vowel space,” J. Acoust. Soc. Am. **115**, 2540.

Chi, X., and Sonderegger, M. (**2007**). “Subglottal coupling and its influence on vowel formants,” J. Acoust. Soc. Am. **122**, 1735–1745.

Claes, T., Dologlou, I., Bosch, L., and Compernolle, D. V. (**1998**). “A novel feature transformation for vocal tract length normalization in automatic speech recognition,” IEEE Trans. Speech Audio Process. **11**, 549–557.

Cui, X., and Alwan, A. (**2006**). “Adaptation of children's speech with limited data based on formant-like peak alignment,” Comput. Speech Lang. **20**, 400–419.

Eide, E., and Gish, H. (**1996**). “A parametric approach to vocal tract length normalization,” in *Proceedings of ICASSP*, pp. 346–349.

Gouvea, E., and Stern, R. (**1997**). “Speaker normalization through formant-based warping of the frequency scale," in *Proceedings of Eurospeech*, pp. 1139–1142.

Hanson, H., and Stevens, K. N. (**1995**). "Subglottal resonances in female speakers and their effect on vowel spectra," in Proceedings of 13th International Congress of Phonetic Sciences, Stockholm, Vol. **3**, pp. 182–185.

Honda, K., Takano, S., and Takemoto, H. (**2009**). "Effects of side cavities and tongue stabilization: Possible extensions of quantal theory," J. Phonetics In Press. Doi: 10.1016/j.wocn.2008.11.002.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., and Johnson, K. (**1999**). "Formants of children women and men: The effect of vocal intensity variation," J. Acoust. Soc. Am. **106**, 1532–1542.

Jung, Y. (**2008**). "Acoustic articulatory evidence for quantal vowel categories across languages," poster presented at the Harvard-MIT HST Forum.

Jung, Y., Lulich, S. M., and Stevens, K. (**2008**). "Development of subglottal quantal effects in young children," J. Acoust. Soc. Am. **124**(4), 2519.

Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., and Alwan, A. (**2005**). "TBall data collection: The making of a young children's speech corpus," in *Proceedings of Eurospeech*, pp. 1581–1584.

Lee, L., and Rose, R. (**1998**). "A frequency warping approach to speaker normalization," IEEE Trans. Speech Audio Process. **6**, 49–60.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.

Lulich, S. M. (**2006**). "The role of lower airway resonances in defining vowel feature contrasts," Ph.D. thesis, MIT, Cambridge, MA.

Lulich, S. M. (**2009**). "Subglottal resonances and distinctive features," J. Phonetics In press. Doi: 10.1016/j.wocn.2008.10.006.

Lulich, S. M., Bachrach, A., and Malyska, N. (**2007**). "A role for the second subglottal resonance in lexical access," J. Acoust. Soc. Am. **122**, 2320–2327.

Lulich, S. M., Zañartu, M., Mehta, D. D., and Hillman, R. E. (**2009**). "Source-filter interaction in the opposite direction: Subglottal coupling and the influence of vocal fold mechanics on vowel spectra during the closed phase," J. Acoust. Soc. Am. **125**(4), 2638.

Madsack, A., Lulich, S. M., Wokurek, W., and Dogil, G. (**2008**). "Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs," Lab. Phon. **11**, 91–92.

McDonough, J. (**2000**). "Speaker compensation with all-pass transforms," Ph.D. thesis, Johns Hopkins University, Baltimore, MD.

McDonough, J., Shaaf, T., and Waibel, A. (**2004**). "Speaker adaptation with all-pass transforms," Sov. Phys. Crystallogr. **42**, pp. 75–91.

Pitz, M., and Ney, H. (**2003**). "Vocal Tract Normalization as Linear Transformation of MFCC," in Proceedings of Eurospeech, pp. 1445–1448.

Snack Sound Toolkit (**2005**). http://www.speech.kth.se/snack/ (Last viewed August, 2008).

Sonderegger, M. (**2004**). "Subglottal coupling and vowel space: An investigation in quantal theory," thesis, MIT, Cambridge, MA.

Stevens, K. N. (**1998**). *Acoustic Phonetics*, MIT, Cambridge, MA.

Umesh, S., Zolnay, A., and Ney, H. (**2005**). "Implementing frequency-warping and VTLN through linear transformation of conventional MFCC," in Proceedings of Interspeech, pp. 269–272.

Wang, X., Wang, B., and Qi, D. (**2004**). "A bilinear transform approach for vocal tract length normalization," in Proceedings of ICARCV, pp. 547–551.

Wang, S., Cui, X., and Alwan, A. (**2007**). "Speaker adaptation with limited data using regression-tree based spectral peak alignment," IEEE Trans. Audio, Speech, Lang. Process. **15**, pp. 2454–2464.

Wang, S., Alwan, A., and Lulich, S. M. (**2008a**). "Speaker normalization based on subglottal resonances," in Proceedings of ICASSP, pp. 4277–4280.

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

Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B. (**1996**). "Speaker normalization on conversational telephone speech," in Proceedings of ICASSP, Vol. **1**, pp. 339–341.

Zhan, P., and Westphal, M. (**1997**). "Speaker normalization based on frequency warping," in Proceedings of ICASSP, pp. 1039–1041.