Automatic estimation of the first two subglottal resonances in children's speech with application to speaker normalization in limited-data conditions

Harish Arsikere[1], *Gary K. F. Leung*[1], *Steven M. Lulich*[2] *and Abeer Alwan*[1]

[1]Dept. of Electrical Engineering, University of California, Los Angeles, CA 90095, USA
[2]Dept. of Psychology, Washington University, Saint Louis, MO 63130, USA

hari.arsikere@gmail.com, garyleung@ucla.edu, slulich@wustl.edu, alwan@ee.ucla.edu

Abstract

This paper proposes an automatic algorithm for estimating the first two subglottal resonances (SGRs)—$Sg1$ and $Sg2$—from continuous speech of children, and applies it to automatic speaker normalization in mismatched, limited-data conditions. The proposed algorithm is based on the observation that $Sg1$ and $Sg2$ form phonological vowel feature boundaries, and is motivated by our recent SGR estimation algorithm for adults. The algorithm is trained and evaluated, respectively, on 25 and 9 children, aged between 7 and 18 years. The average RMS errors incurred in estimating $Sg1$ and $Sg2$ are 55 and 144 *Hz*, respectively. By applying the proposed algorithm to a connected digits speech recognition task, it is shown that: 1) a linear frequency warping using $Sg1$ or $Sg2$ is comparable to or better than maximum likelihood-based vocal tract length normalization (ML-VTLN), 2) the performance of SGR-based frequency warping is less content dependent than that of ML-VTLN, and 3) SGR-based frequency warping can be integrated into ML-VTLN to yield a statistically-significant improvement in performance.

Index Terms: subglottal resonances, children's speech, automatic estimation, limited data, speaker normalization

1. Introduction

Recent research on speaker normalization for automatic speech recognition (ASR) has shown that the second subglottal resonance ($Sg2$) can be effective in normalizing children's speech to acoustic models trained on adults, especially when enrollment data is limited or is in a language different from that of the training data [1]. Motivated by the fact that $Sg2$ remains fairly constant for a given speaker,—phonetic content has little effect on the frequency of $Sg2$ [2]—speaker normalization was achieved in [1] by linearly warping the frequency axis of a given test speaker using the estimated $Sg2$ frequency. $Sg2$ was estimated based on: (1) an empirical relation between $Sg2$ and the third formant frequency ($F3$) [3], and (2) the observation that $Sg2$ induces discontinuities in the second formant frequency ($F2$) of vowels [2, 3]. Experiments on speaker normalization revealed that in limited-data conditions, $Sg2$-based frequency warping performed better than maximum likelihood-based vocal tract length normalization (ML-VTLN) [4]. In addition, $Sg2$ was found to be effective in performing cross-language speaker normalization. In [5], the $Sg2$ estimation algorithm was improved by incorporating the observation that $Sg2$, apart from inducing frequency discontinuities, can cause attenuations in the amplitude of the second formant [2, 3]. The improved $Sg2$ estimation algorithm was used to perform Bark shift-based nonlinear speaker normalization.

Although [1] and [5] apply $Sg2$ successfully to speaker normalization, their methods have two major limitations: (1) the $Sg2$ estimation algorithms require isolated vowels for achieving satisfactory performance since they are based on detecting subtle acoustic events in speech that are caused by subglottal coupling, and (2) the accuracy of $Sg2$ estimation depends on the phonetic content of the vowel used (see Fig. 6 in [1]), and consequently, the performance of $Sg2$-based normalization is more content dependent than that of ML-VTLN (see Fig. 12 in [1]). In order to overcome these limitations and develop more practical approaches to speaker normalization using subglottal resonances (SGRs), this study proposes an automatic algorithm for estimating the first two SGRs—$Sg1$ and $Sg2$—from continuous children's speech. The proposed algorithm is similar to our recent SGR estimation algorithms for adults [6, 7, 8], but is slightly modified to account for the larger acoustic variability observed in children's speech. By applying the proposed algorithm to a connected digits ASR task, it is shown that $Sg1$, like $Sg2$, can be effective in speaker normalization, and that in limited-data conditions, SGR-based normalization is comparable to ML-VTLN while being much less sensitive to the content spoken. It is useful to be able to perform speaker normalization using $Sg1$ since $Sg2$ estimation using the proposed algorithm requires a good estimate of $F3$, which may be difficult to obtain in narrow band children's speech (e.g., telephone speech). While the third subglottal resonance ($Sg3$) appears to complement $Sg2$ in improving speaker normalization performance [9], such a possibility is not investigated here.

Vocal tract length differences among speakers has been known to be a major source of inter-speaker variability (see [10], for example), and several parametric [11, 12] as well as maximum likelihood-based (ML-based) [13, 4] approaches have been proposed to alleviate the degradation in ASR performance that can result from it. Although ML-based approaches outperform parametric approaches in general (see [14], for example), parametric methods like $Sg2$-based warping can be more effective in limited-data conditions [1, 5]. It is shown in this study that a simple integration of SGR-based frequency warping into ML-VTLN [4] can yield a statistically-significant improvement in performance when data is limited.

Section 2 describes the proposed SGR estimation algorithm for children and presents the results of its evaluation. In Section 3, SGR-based frequency warping and its integration into ML-VTLN are discussed. Section 4 presents the results of speaker normalization experiments, and Section 5 concludes the paper.

2. Automatic estimation of $Sg1$ and $Sg2$

Automatic algorithms for estimating $Sg1$ and $Sg2$ of adult speakers were proposed recently in [6] and [7]. In [6], $Sg1$

was estimated using a model relating two correlated measures of vowel height: the Bark difference between the first formant frequency ($F1$) and $Sg1$—denoted as $B_{1,s1}$—was predicted from the Bark difference between $F1$ and the fundamental frequency ($F0$)—denoted as B_{10}. Similarly, in [7], $Sg2$ was estimated using a model relating two correlated measures of vowel backness: the Bark difference between $F2$ and $Sg2$—denoted as $B_{2,s2}$—was predicted from the Bark difference between $F3$ and $F2$—denoted as B_{32}. In [15], it was shown that the algorithms developed in [6] and [7] gave satisfactory results for children taller than 150 *cm*. In a more recent study [8], the algorithm in [7] was improved by including $F3$ and $F0$ in the model for predicting $B_{2,s2}$. Since the algorithms in [6] and [8] were effective with a limited amount of speech data and also independent of spoken content, modified versions of these algorithms were developed in this study for children's speech.

2.1. Methods

For simplicity, let the models for predicting $B_{1,s1}$ and $B_{2,s2}$ be denoted as M1 and M2, respectively. For training M1 and M2, we used data from 25 speakers—17 boys, 8 girls—in a recently-collected children's corpus, which, at present, contains simultaneous recordings of speech and subglottal acoustics from 29 children—20 boys, 9 girls—aged between 7 and 18 years. In this study, microphone as well as accelerometer (subglottal) signals of 9 different vowels—[i], [ɪ], [ɛ], [æ], [ɑ], [ʌ], [o], [ʊ] and [u]—were extracted from the corpus, and used for training M1 and M2. Further details of the recording procedures and the recorded material can be found in [15].

As in our previous studies [6, 8], measurements of $F0$ and formant frequencies $F1$–$F3$ were required for training M1 and M2. For each speaker, microphone signals of 3–5 tokens per vowel were chosen for analysis. Using Wavesurfer [16], $F0$ and $F1$–$F3$ were measured semi-automatically from the steady-state portions of the chosen tokens, i.e., the parameters of Wavesurfer were adjusted manually until the $F0$ and formant contours were found to be satisfactory. 'Ground truth' values of $Sg1$ and $Sg2$ were also required for model training. The 'ground truth' SGR frequencies of each speaker were obtained by making several measurements of $Sg1$ and $Sg2$ in the accelerometer signals of vowel tokens, and finding their averages. On average, $Sg1$ and $Sg2$ were measured in 11 and 13 tokens per speaker, respectively. The procedure for measuring $Sg1$ and $Sg2$ was identical to the one described in [15]: information from three different spectral representations—Fourier transforms, linear prediction spectra and power spectral density estimates—was combined to obtain reliable measurements. 'Ground truth' $Sg1$ values ranged between 532 and 893 *Hz* (mean = 727 *Hz*), while 'ground truth' $Sg2$ values ranged between 1261 and 2160 *Hz* (mean = 1734 *Hz*).

For training M1 and M2, features B_{10}, $B_{1,s1}$, B_{32} and $B_{2,s2}$ were computed using the formant and $F0$ frequencies as well as the 'ground truth' SGRs of all 25 speakers in the training set. As in our previous studies [6, 8], B_{10} was found to correlate strongly with $B_{1,s1}$ ($r^2 = 0.92$), while B_{32} was found to correlate strongly with $B_{2,s2}$ ($r^2 = 0.83$). Model M1 was trained by performing a linear regression between $B_{1,s1}$ and $\{B_{10}, F3\}$ ($r^2 = 0.93$); the increase in r^2 (from 0.92 to 0.93) caused by the inclusion of $F3$, was statistically significant ($p < 0.001$). Model M2 was similarly trained by performing a linear regression between $B_{2,s2}$ and $\{B_{32}^2, B_{32}, F3, F0\}$ ($r^2 = 0.92$); the increase in r^2 (from 0.83 to 0.92) caused by the inclusion of B_{32}^2, $F3$ and $F0$, was statistically significant ($p < 0.001$). It must be pointed out that a further addition of B_{10}^2 (analogous to B_{32}^2) and $F0$ to model M1 did not result in any significant improvement in modeling accuracy.

In order to account for the large acoustic variability in children's speech, we also investigated the idea of splitting M1 and M2 into 2 models each. Since differences in vocal tract lengths is a major source of acoustic variability, models M1 and M2 were split based on the speakers' average $F3$ values ($F3$ is known to be a good indicator of vocal tract length). Figure 1(a) shows the distribution of average $F3$ values for the 25 speakers in the training set. The median of this distribution—approximately 3300 *Hz*—was used for model splitting: model M1^l was trained using data from speakers whose average $F3$ values were less than 3300 *Hz* (15 speakers), and model M1^g was trained using data from the remaining 10 speakers. Models M2^l and M2^g were trained similarly. In our training set, the 'threshold value' of 3300 *Hz* corresponded roughly to an age of 10 years (see Fig. 1(b)). Since most children reach puberty at the age of 12–13 years, a 'threshold value' of 3300 *Hz* was considered reasonable.

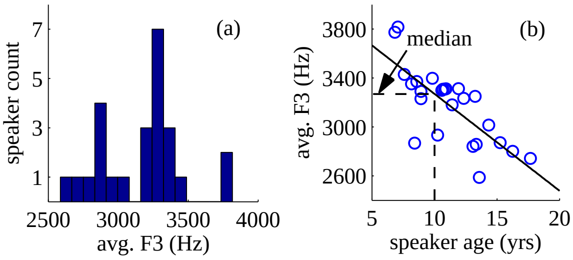

Figure 1: (a) Distribution of average $F3$ values in the training set of the SGR estimation algorithm (median ≈ 3300 *Hz*). (b) Scatter plot of average $F3$ versus speaker age. An average $F3$ of 3300 *Hz* corresponds roughly to 10 years (close to puberty).

2.2. The algorithm

In this study, we investigated the efficacy of two approaches—one based on using M1 and M2, and the other based on using (M1^l;M2^l) *or* (M1^g;M2^g). Let the algorithms based on these two approaches be denoted as A and A', respectively. Given a speech signal, the steps in estimating $Sg1$ and $Sg2$ are:

1. Down sample the signal to 8 *kHz*.
2. Perform pre-emphasis with the filter: $H(z) = 1 - 0.97z^{-1}$.
3. Track $F0$ and $F1$–$F3$ automatically using Snack [17]. Set the window length to 30 *ms* and the window spacing to 5 *ms*.
4. Select voiced frames using Snack's binary voicing parameter.
5. Compute the average $F3$ over all voiced frames.
6. For algorithm A, use M1 and M2. For algorithm A', use (M1^l;M2^l) or (M1^g;M2^g) depending on whether the average $F3$ is less than or greater than 3300 *Hz*, respectively.
7. Estimate $Sg1$ and $Sg2$ for each voiced frame. To estimate $Sg1$, predict $B_{1,s1}$ using the model selected in Step 6, subtract it from $F1$ (Bark), and convert the resulting value from Bark to Hertz. Estimate $Sg2$ in a similar fashion.
8. Estimate $Sg1$ and $Sg2$ for the given utterance by averaging the frame-level estimates obtained in Step 7.

2.3. Evaluation and results

Algorithms A and A' were evaluated on a small test set consisting of 9 speakers—4 boys, 5 girls—aged between 8 and 16 years. Out of the 9 test speakers, 4 were chosen from the recently-collected children's corpus (see Sec. 2.1) while the re-

Table 1: Mean and standard deviation of RMS errors (across speakers)—μ_{rms} and σ_{rms}—incurred in estimating $Sg1$ and $Sg2$ using algorithms A and A'.

	$Sg1$		$Sg2$	
	A	A'	A	A'
μ_{rms} (*Hz*)	55	55	155	144
σ_{rms} (*Hz*)	37	38	81	72

maining 5 were chosen from a previously-collected corpus [3] containing microphone and accelerometer recordings of 9 children (only speakers who were 7 years or older were chosen for evaluation). The 'ground truth' SGR frequencies of the test speakers were obtained as described in Sec. 2.1. 'Ground truth' $Sg1$ values ranged between 610 and 845 *Hz* (mean = 711 *Hz*), while 'ground truth' $Sg2$ values ranged between 1427 and 1965 *Hz* (mean = 1645 *Hz*). Algorithms A and A' were applied to every utterance in the test set. For comparing the performance of A and A', the mean and the standard deviation of root mean squared errors (across speakers)—μ_{rms} and σ_{rms}—were used.

Table 1 compares the performance of A and A'. While A' did not perform any better than A in estimating $Sg1$, it offered a small performance improvement—7% reduction in μ_{rms} and 11% reduction in σ_{rms}—in estimating $Sg2$. In essence, A' was not only more accurate in estimating $Sg2$, but also more consistent across speakers. Therefore, in all our speaker normalization experiments, model M1 (algorithm A) was used for estimating $Sg1$, while models M2^l and M2^g (algorithm A') were used for estimating $Sg2$ (in children's speech). Since formant and $F0$ tracking are more error prone in children's speech than in adults' speech, it is important to note that a non-negligible portion of SGR estimation errors could be the result of errors in $F0$ and/or formant tracking. Nevertheless, the errors incurred by A and A' are acceptable from a speaker normalization point of view (see Sec. 3 for an explanation).

3. Speaker normalization algorithms

The proposed SGR estimation algorithm was applied to speaker normalization in a mismatched, connected digits ASR task using the TIDIGITS database [18]. Acoustic hidden Markov models (HMMs) were trained on 55 adult males and tested on 25 boys and 25 girls (train and test speakers were chosen as per the documentation contained in the TIDIGITS database). Although we could have simulated a mismatched scenario using both male and female data for training, we chose the above setup so that the improvements due to normalization could be more easily observed. HMMs were monophone based, had a left-to-right topology, and contained 3 emitting states each. The output state distributions were six-component Gaussian mixtures with diagonal covariance matrices. Twelve Mel-frequency cepstral coefficients (MFCCs) plus log energy, and their first- and second-order derivatives, were used for parameterizing speech signals; features were extracted using a frame length of 25 *ms* and a frame spacing of 10 *ms*. Word error rate (WER) was used as the performance metric in all experiments. The baseline (without normalization) WER was 37.6%. In order to reduce the mismatch between adult males and children, we applied the following normalization algorithms to children's speech.

I. ML-VTLN: VTLN is a normalization scheme which is based on the assumption that vocal tract lengths of speakers are related linearly to one another [4]. Given the data of a particular test speaker, we implemented ML-VTLN as follows. (1) Extract features $\mathcal{X}$ from one enrollment utterance. (2) Using the set of pre-trained HMMs λ, find the transcription W corresponding to the enrollment utterance. (3) Extract warped features $\mathcal{X}^{\alpha}$ for $\alpha \in [0.7, 1.3]$ by varying α in steps of 0.02 (similar to the implementation in [1]). $\mathcal{X}^{\alpha}$ is the same as $\mathcal{X}$, except that the frequency axis is warped linearly by a factor α. (4) Find the 'best' α: $\hat{\alpha} = \arg\max_{\alpha} P(\mathcal{X}^{\alpha}|\lambda, W)$. (5) Recognize the remaining utterances after frequency warping by a factor $\hat{\alpha}$.

II. SGR-based normalization (SGRN): SGR-based normalization is a parametric approach which is inspired by the phonetic invariance of SGRs and the relationship between SGRs and formant frequencies—$Sg1$ forms a boundary between [+low] and [-low] vowels, while $Sg2$ forms a boundary between [+back] and [-back] vowels [19, 3]. Before implementing SGRN on the test data, $Sg1$ and $Sg2$ were estimated for all the training speakers (using our previous algorithms for adults), and 'reference' values ($Sg1_r$ and $Sg2_r$) were calculated by averaging the estimated SGRs. Then, given the data of a particular test speaker, SGRN was implemented using the following steps. (1) Estimate $Sg1$ and $Sg2$ from one enrollment utterance. (2) Find the warp factor: $\alpha_s = Sg1_r/Sg1$ for normalization with $Sg1$, and $\alpha_s = Sg2_r/Sg2$ for normalization with $Sg2$. (3) Perform Step 5 of ML-VTLN. It must be noted that α_s will differ from its 'true value' due to SGR estimation errors. A simple calculation using the nominal 'ground truth' values of $Sg1$ and $Sg2$ and the RMS estimation errors shown in Table 1 reveals that α_s tends to have an error of just 7–8%, on average.

III. SGR-based warping for ML-VTLN (ML-SVTLN): The advantage of ML-VTLN is its ability to estimate the warp factor with respect to statistical models (HMMs), while the advantage of SGRN is its reduced complexity due to the absence of a grid search. ML-SVTLN combines the benefits of both techniques by obtaining a better first-pass transcription than the one obtained in ML-VTLN (with little additional complexity). Using the same notation as above, the steps in ML-SVTLN are as follows. (1) Pick an enrollment utterance; estimate $Sg1$ and $Sg2$. (2) Find α_s, where $\alpha_s = Sg1_r/Sg1$ or $Sg2_r/Sg2$. (3) Extract features $\mathcal{X}^{\alpha_s}$ from the enrollment utterance, and perform recognition with λ to obtain the transcription W. (4) Perform Steps 3–5 of ML-VTLN. An algorithm in parallel with ML-SVTLN (and bypassing the use of SGRs) is one that estimates the optimal warp factor in two iterations of ML-VTLN. However, such a procedure is inefficient since it requires two recognition passes.

4. Normalization experiments and results

The TIDIGITS database contains utterances of 1, 2, 3, 4, 5 or 7 digits. All normalization experiments were performed in limited-data conditions since only one enrollment utterance was used in every case. First, in order to compare the content dependence of VTLN with that of SGRN—using $Sg1$ (SGRN1) and $Sg2$ (SGRN2)—, we performed normalization experiments with single-digit enrollment utterances ("one" to "nine", "zero", and "oh"). The resulting WERs are shown in Fig. 2(a). Clearly, SGRN1 and SGRN2 not only outperformed ML-VTLN on average (the difference in WERs was statistically significant at $p < 0.05$), but they also proved to be much more consistent across spoken content. This is an important improvement over the results in [1], where SGR-based normalization was found to be highly content dependent. The consistency of SGRN is further exemplified by Fig. 2(b), which compares the standard deviations of warp factors obtained from ML-VTLN and SGRN2. SGRN1 and SGRN2 were comparable in performance (see Fig. 2(a)), and the warp factors obtained using the two approaches were very similar (r = 0.84, mean absolute difference

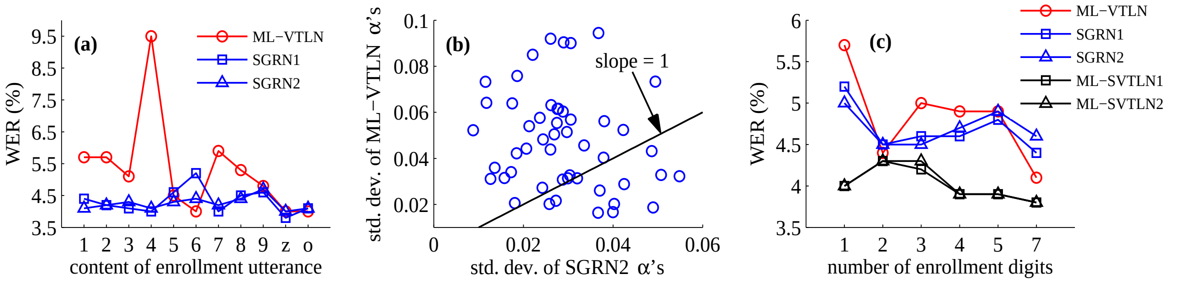

Figure 2: (a) Comparison of ML-VTLN and SGRN for different single-digit enrollment utterances ("one" to "nine", "zero", and "oh"). (b) Standard deviation of ML-VTLN warp factors versus standard deviation of SGRN2 warp factors (one data point per test speaker). (c) Comparison of ML-VTLN, SGRN and ML-SVTLN for varying amounts of enrollment data.

= 0.024). Therefore, a piece-wise linear warping involving $Sg1$ and $Sg2$ did not yield any significant improvement over using $Sg1$ or $Sg2$ alone. It must be noted that ML-VTLN, simply by virtue of being an ML-based method, should theoretically outperform SGRN. This does not always happen in practice, particularly in mismatched conditions, because the ML-VTLN warp factor is affected by first-pass transcription errors.

Next, we compared the performance of ML-VTLN, SGRN and ML-SVTLN—using $Sg1$ (ML-SVTLN1) and $Sg2$ (ML-SVTLN2)—by varying the length of enrollment utterances from 1 to 7 digits. The resulting WERs are shown in Fig. 2(c). Although ML-SVTLN1 and ML-SVTLN2 performed better than ML-VTLN in all cases (the difference in WERs was statistically significant at $p < 0.01$), the improvement was larger when ML-VTLN performed poorly (compare the WERs for 4 and 7 digits, for example). SGRN1 and SGRN2 were, on average, comparable to ML-VTLN in performance, although they were outperformed as the amount of enrollment data increased. For completeness, we also performed linear warping using $F3$ (as described in [11]). It resulted in a WER of 5.4% with 7 enrollment digits, clearly suggesting that SGR-based warping is the better approach in limited-data conditions.

5. Conclusions

In this paper, an automatic algorithm for estimating the first two SGRs in continuous children's speech was proposed. $Sg1$ was estimated by modeling the relation between two measures of vowel height, while $Sg2$ was estimated by modeling the relation between two measures of vowel backness. The idea of choosing automatically between two $F3$-dependent models—based on a 'threshold value' of 3300 Hz—was investigated; such an approach provided some improvement in $Sg2$ estimation, but not in $Sg1$ estimation. The proposed SGR estimation algorithm was used in two speaker normalization schemes (SGRN and ML-SVTLN) in mismatched, limited-data conditions. SGRN (with $Sg1$ or $Sg2$) was not only comparable to ML-VTLN in performance, but it also had the following advantages: (1) no grid search was required for estimating the warp factor, and (2) warp factor estimation was much less susceptible to the content of enrollment data. Moreover, when information about $Sg1$ or $Sg2$ was provided to ML-VTLN (ML-SVTLN), a significant improvement in performance was observed.

6. Acknowledgments

The authors thank John R. Morton, Karl Wassmann and Prof. Mitchell S. Sommers for their role in recording child subjects. This work was supported in part by NSF Grant No. 0905381.

7. References

[1] S. Wang, S. M. Lulich, and A. Alwan, "Automatic detection of the second subglottal resonance and its application to speaker normalization," *JASA*, vol. 126, pp. 3268–3277, 2009.

[2] X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *JASA*, vol. 122, pp. 1735–1745, 2007.

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

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

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

[6] H. Arsikere, S. M. Lulich, and A. Alwan, "Automatic estimation of the first subglottal resonance," *J. Acoust. Soc. Am. (Express Letters)*, vol. 129, pp. 197–203, 2011.

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

[8] H. Arsikere, G. Leung, S. M. Lulich, and A. Alwan, "Automatic height estimation using the second subglottal resonance," in *Proceedings of ICASSP (to appear)*, 2012.

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

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

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

[12] E. B. Gouvêa and R. M. Stern, "Speaker normalization through formant-based warping of the frequency scale," in *Fifth European Conference on Speech Communication and Technology*, 1997.

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

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

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

[16] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proceedings of ICSLP*, 2000, pp. 464–467.

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

[18] R. Leonard, "A database for speaker-independent digit recognition," in *Proceedings of ICASSP*, 1984, pp. 328–331.

[19] Y. Jung, "Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back]," *Ph.D. Thesis, Harvard-MIT Division of Health Sciences and Technology, MIT*, 2009.