Age-dependent height estimation and speaker normalization for children's speech using the first three subglottal resonances

Jinxi Guo[1], *Rohit Paturi*[1], *Gary Yeung*[1], *Steven M. Lulich*[2], *Harish Arsikere*[3] *and Abeer Alwan*[1]

[1] Dept. of Electrical Engineering, University of California, Los Angeles, CA 90095, USA
[2]Dept. of Speech and Hearing Sciences, Indiana University, Bloomington
[3]Data Analytics Lab, Xerox Research Center-India (XRCI), Bangalore, Karnataka, India

lennyguo@g.ucla.edu, rohitpaturi@ucla.edu, garyyeung@g.ucla.edu, alwan@ee.ucla.edu, slulich@indiana.edu, harish.asikere@xerox.com

Abstract

This paper proposes an age-dependent scheme for automatic height estimation and speaker normalization of children's speech, using the first three subglottal resonances (SGRs). Similar to previous work, our analysis indicates that children above the age of 11 years show different acoustic properties from those under 11. Therefore, an age-dependent model is investigated. The estimation algorithms for the first three SGRs are motivated by our previous research for adults. The algorithms for the first two SGRs have been applied to children's speech before. This paper proposes a similar approach to estimate Sg3 for children. The algorithm is trained and evaluated on 46 children, aged between 6-17 years, using cross-validation. Average RMS errors in estimating Sg1, Sg2 and Sg3 using the age-dependent model are 51, 128 and 168 Hz, respectively. The height estimation algorithm employs a negative correlation between SGRs and height, and the mean absolute height estimation error was found to be less than 3.8cm for the younger children and 4.9cm for the older children. In addition, using TIDIGITS, a linear frequency warping scheme using age-dependent Sg3 gives statistically-significant word error rate reductions (up to 26%) relative to conventional VTLN.

Index Terms: children's speech, subglottal resonances, height estimation, speaker normalization.

1. Introduction

Previous research on the applications of subgottal resonance (SGRs) has focused primarily on height estimation and speaker normalization. Automatic estimation of an unknown speaker's height from speech can benefit forensics and provide useful speaker information. In [1], an automatic height estimation algorithm was proposed for adults, based on the strong negative correlations between SGRs and speaker height. One important feature of this algorithm is that the amount of training data and the number of features used for height estimation are very small in comparison with other studies [2, 3]. Although the algorithm works well on adults' speech, the height estimation algorithm for children's speech has not been investigated.

Previous research also shows that SGRs can be effective in speaker normalization for automatic speech recognition especially for children's speech under mismatched and limited-data conditions. In [4], a linear frequency warping algorithm using either Sg1 or Sg2 was proposed. In [5], a non-linear frequency warping scheme was used, which is based on mapping the SGRs and the third formant frequency of a given utterance to reference values. Both the results of [4] and [5] indicate that SGR-based normalization is comparable to or better than the conventional form of vocal tract length normalization (CVTLN) [6] and is not sensitive to content. Although frequency warping using Sg1, Sg2 and formants gives good performance for children's speech, the effect of Sg3 for children speaker normalization has not been investigated.

In this paper, we analyzed the first three SGRs (Sg1, Sg2, Sg3) and their relationship with formants using the WashU-UCLA child corpus, a subset of which was used in [7]. Based on previous research [8] and our analysis, children with different age groups show different acoustic properties. The age of 11 years was chosen as a cutoff since this is the approximate age at which a child reaches puberty. Using this cutoff, new age-dependent regression models were trained to estimate Sg1, Sg2 and Sg3 for children's speech. With these models, an automatic height estimation algorithm is proposed for children's speech of different age groups, and a new method using age-dependent Sg3 is also applied to speaker normalization.

In Section 2, we present the relationship between SGRs and formants for children's speech, propose an age-dependent SGR estimation algorithm and show the results of its evaluation. Section 3 describes the height estimation algorithm and corresponding results. Section 4 explains the linear frequency warping scheme using age-dependent Sg3 and the results of the speaker normalization experiment. Section 5 concludes the paper.

2. Analysis and automatic estimation of the first three SGRs

2.1. Dataset

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

Work supported in part by NSF Grant No. 0905381.

height was recorded in the corpus and ranged from 105cm to 182cm.

2.2. Analysis

SGR analysis was conducted on all the recordings of the 10 monophthongs: 2760 microphone recordings and 2852 subglottal recordings. For each speaker, the first three formants were measured from the microphone signals in the steady-state region using Snack [9]. The first three SGRs were measured from the corresponding accelerometer signals by visual inspection of the resonance peaks in LPC spectra using Wavesurfer [10]. Both microphone and accelerometer signals were down-sampled to 8kHz before analysis.

2.2.1. The relationship between Sg1, Sg2 and vowel class

Previous research [11] showed that Sg1 acts as a boundary between high and low vowels and Sg2 forms a boundary between front and back vowels for adult speech, and this paper investigated whether Sg1 and Sg2 divided the vowel space for children's speech as well. Table 1 shows the percentage of speakers in which Sg1 and Sg2 successfully divided the vowel space. The percentages are high indicating that SGRs divide the vowel space of children as well as adults.

Table 1. *Percentage of speakers, separated by age group, whose SGRs successfully divided the vowel space.*

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

2.2.2. The relationship between Sg1, Sg2 and Sg3

To investigate the relationship between Sg1, Sg2 and Sg3 for children's speech, scatter plots of Sg3 versus Sg1 and Sg2 are shown in Figure 1. The results indicate that Sg3 is correlated with Sg1 (r=0.88) but more strongly correlated with Sg2 (r=0.92). Therefore, a first-order linear regression was trained using Sg2 and Sg3 and the result is Eq. 1.

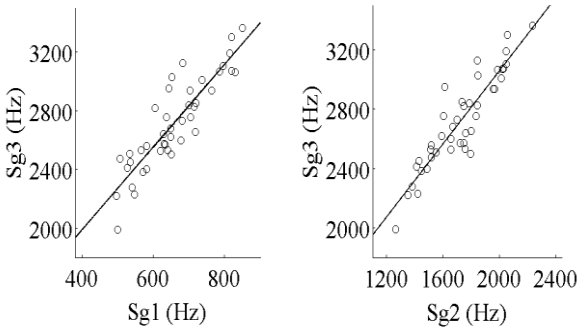

Figure 1: *Scatter plots of Sg3 vs. Sg1 (left) and Sg3 vs. Sg2 (right). Also shown are first-order linear regression. Sg1 and Sg3 are correlated (r=0.88) while Sg2 and Sg3 are more strongly correlated (r=0.92).*

$$Sg3 = 1.233(Sg2) + 593.424 \quad (1)$$

2.3. Automatic estimation

Estimation algorithms for the first three subglottal resonances were proposed for adults in [1]. The algorithm was based on the following central idea: Sg1 acts as a boundary between high and low vowels so that two acoustic features characterizing vowel frontness – the Bark difference between the third and first formants (denoted as B_{31}) and the Bark difference between F1 and Sg1 (denoted as $B_{1,s1}$) – are correlated. Similarly, for Sg2 estimation, the Bark difference between F3 and F2 (denoted B_{32}) was found to be related to the bark difference between F2 and Sg2 (denoted as $B_{2,s2}$) since both measures characterize vowel backness. An empirical equation was derived to predict $B_{1,s1}$ from a linear combination of the first three powers of B_{31} and a constant term. The same approach also applied to $B_{2,s2}$ and B_{32} to predict $B_{2,s2}$. Sg3 is estimated based on its correlation with Sg2 using a first-order linear regression, as in Eq. 1. These empirical relations allowed the first three SGRs to be estimated from a speech signal once the first three formants are tracked automatically.

A previous study [7] derived empirical relations to estimate Sg1 and Sg2 for children's speech, but the dataset was relatively small, and the estimation algorithm for Sg3 was not investigated. In this study, all the empirical relations to estimate the first three SGRs were derived using a larger dataset of 46 speakers in the WashU-UCLA child corpus.

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

Table 2. *R-squared values for the SGR estimation models of Sg1 and Sg2 when trained on speakers separated by age group, as well as when trained on all speakers.*

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

Therefore, we train and test the SGR estimation algorithms separately for the two different age groups using a cross-validation method. Within each age group, each time we randomly chose around 60% of the speakers to train the regression model and the rest to test the estimation algorithm. Given a test speech signal, the detailed steps involved in estimating SGRs are the same as in [1].

2.4. Performance analysis of the algorithm

The SGR estimation algorithm was evaluated using two performance metrics: the mean and standard deviation of the root mean squared errors (across speakers and 5 cross-validation tests), denoted as μ_{rms} and σ_{rms}, respectively, both in units of Hz. Table 3 shows the performance of the automatic estimation algorithm in each age group and the whole dataset.

Noted that in applications, when age information is not available, broad age group estimation algorithms should be used before estimating SGRs. In Section 4 of this paper, average F3 is used as a threshold to predict the age group of

each speaker when estimating the SGRs for speaker normalization.

The best regression models to estimate Sg1 and Sg2 for younger children during cross-validation are presented by Eq. 2 and Eq. 3. Eq. 4 and Eq. 5 present the best models for older children. The regression equations of Sg3 for both age groups are dependent on Sg2 in a similar way as in Eq. 1. Therefore, Sg3 for both age groups can be estimated from the Sg2 estimates using Eq. 1.

Table 3. *Mean and standard deviation of RMS error, in Hz, of SGR estimation for the set of 'younger' children (Y), 'older' children (O) and both sets 'combined' (C).*

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

$$B_{1,s1} = 0.0002(B_{31})^3 + 0.003(B_{31})^2 - 0.907(B_{31}) + 8.310 \quad (2)$$
$$B_{2,s2} = -0.011(B_{32})^3 + 0.184(B_{32})^2 - 1.870(B_{32}) + 5.290 \quad (3)$$

$$B_{1,s1} = -0.001(B_{31})^3 + 0.011(B_{31})^2 - 0.776(B_{31}) + 6.601 \quad (4)$$
$$B_{2,s2} = -0.003(B_{32})^3 + 0.062(B_{32})^2 - 1.534(B_{32}) + 4.477 \quad (5)$$

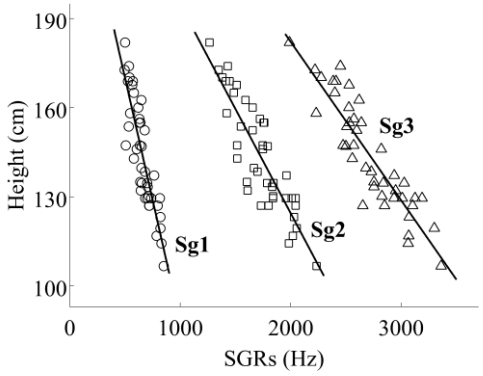

Figure 2: *Scatter plots of all child speaker height vs. each of the first three SGRs. Also shown are first-order linear regression fits. Speaker height correlates strongest with Sg3 (r=-0.90), but is also correlated with Sg1 (r=-0.88) and Sg2 (r=-0.88).*

3. Height estimation

3.1. Methods

Previous work on adult height estimation using speech signals has shown a strong negative correlation between the three SGRs and height [1]. This section tests a similar hypothesis for children. Using measurements for Sg1, Sg2 and Sg3 for each speaker, as well as information about the speakers' heights, a scatter plot of height versus the SGRs for all children is presented in Figure 2.

The inverse correlation of each SGR with height is strong, and therefore a first-order linear regression is used to model the relationship between each SGR and height. The empirical relations were obtained between speaker height and SGR frequencies, as illustrated in Eqs. 6, 7 and 8. These equations are different from the height estimation equations for adults in [1].

$$h = -0.166(Sg1) + 254.497 \text{ (r=-0.88)} \quad (6)$$
$$h = -0.070(Sg2) + 264.793 \text{ (r=-0.88)} \quad (7)$$
$$h = -0.053(Sg3) + 288.821 \text{ (r=-0.90)} \quad (8)$$

3.2. Experiments and results

Motivated by the results of Section 2, the height estimation algorithm was tested used a cross-validation method in which the child speakers were grouped into two different categories: age under 11 years and age above 11 years. In each category, 60% of the speakers were chosen to train the first order linear regression model between height and each empirically measured SGR (ground truth), and the rest were used to test the model. In each age group, after the models were trained, using the method proposed in Section 2, Sg1, Sg2 and Sg3 were estimated for each testing speech signal. Finally, the trained linear regressions between SGRs and height, along with the three computed SGRs from the test data, were used to estimate the speakers' heights, and the results were compared with the actual height measurements. The height was calculated for each voiced frame, and the estimated height for each test speaker was the average number across all frames.

This procedure was repeated 5 times for each age group, and each time, the root mean squared errors (RMSE) and mean average errors (MAE) were recorded. Table 4 displays the average RMSE and MAE of this experiment for both age groups.

Additionally, to verify the necessity of the age-dependent SGR estimation model, the experiment was repeated again but with all child speakers grouped into a single category. Estimation of Sg1, Sg2 and Sg3 for the speakers used a model obtained in a similar way as Eqs. 1-5 in Section 2 but trained assuming age-independence of SGRs. The average RMSE (cm) and MAE (cm) of this experiment are also shown in Table 4.

Table 4. *Mean average error and root mean squared error of the height estimation algorithms when trained and tested on the set of 'younger' children (Y), 'older' children (O) and 'all' children (A).*

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

The resulting regression equations of height versus each SGR during cross-validation training were similar to Eqs. 6, 7 and 8 for both the younger and older groups, and therefore, Eqs. 6, 7 and 8 can be used to estimate height using SGRs regardless of age. However, RMSE and MAE were smaller when using different SGR estimation models for different age groups, suggesting the necessity for age-dependent SGR regression models. Thus, the height estimation algorithm can simplify to age-dependent SGR estimation models in combination with age-independent linear regressions of height versus SGRs. Observing the values in Table 4 reveals that Sg1 returns the most accurate height estimation (MAE of 3.8cm) for children under the age of 11 years, while Sg2 and Sg3 return the most accurate height estimation (MAE of 4.9cm) for children above 11. Note that the height estimation error for older children is similar to that of adult speech [1].

4. Speaker normalization

4.1. Methods and algorithm for comparison

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

4.2. Normalization experiment and results

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

Normalization is applied only to the testing data and not to the training data. The reference SGRs used in our experiment were obtained by taking the average of all the estimated SGRs of the adult speakers in the training set, which were $Sg1_{ref}$ =604.9Hz, $Sg2_{ref}$=1357.4Hz and $Sg3_{ref}$ =2228.3Hz. The F3 used for separating the speakers into the 2 age groups was 3kHz.

The hidden Markov model toolkit (HTK) was used for all experiments, and word error rate (WER) was used as the performance metric. Results for all our experiments are shown in Table 1.

The results show that the experiments using only Sg3 produce the lowest WERs followed by the Sg2 and Sg1 warping schemes. One possible reason is that, as it has been shown in Section 3 that Sg3 has the strongest correlation with height, Sg3 may also have strong correlation with the vocal tract length (VTL). The combination of Sg1, Sg2 and Sg3 gives a WER that lies between that obtained by using only Sg3, only Sg2 and only Sg1. Among the warping schemes involving only Sg3, the highest WER is obtained using the age-independent Sg3 estimation. The lowest WER occurred when using oracle age information to estimate the SGRs (~26% WER reduction relative to CVTLN and {Sg1, Sg2, F3} warp). Automatic estimation of age group using F3 also produced WER lower than the CVTLN, {Sg1, Sg2, F3} with results comparable to the oracle age-dependent Sg3 warping scheme. Though F3 alone is not a perfect measure to estimate age, it has been observed that it is good enough to roughly separate the speakers into 2 age groups to estimate SGRs.

Table 5. *Word error rates (%) for ASR experiments.*

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

5. Conclusions

In this paper, an age-dependent scheme for automatic height estimation and speaker normalization is proposed for children's speech. Analysis indicates that children below and above 11 years old show different acoustic properties, and therefore, an automatic age-dependent SGR estimation algorithm is applied to each age group. The first three SGRs were estimated using a similar method adapted from adult speech but with age dependency considerations. Good results were achieved for each SGR. Using the algorithm for estimating SGRs and the inverse relation between SGRs and height, speaker height can be automatically estimated. The proposed height estimation algorithm performs well in each age group. Using a cross-validation method, speaker height can be estimated to within 3.8cm for younger children and 4.9cm for older children, on average. Motivated by the differences between each age group, a linear frequency warping method using age-dependent Sg3 was applied to the TIDIGITS speech recognition task. The results show that the proposed method outperforms CVLTN and other SGR-based warping schemes.

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

6. References

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

[2] T. Ganchev, I. Mporas, N. Fakotakis. "Audio features selection for automatic height estimation from speech," *Artificial Intelligence: Theories, Models and Applications* 2010: 81-90

[3] T. Ganchev, I. Mporas, and N. Fakotakis. "Automatic height estimation from speech in real-world setup," in *Proc. of the 18th European Signal Processing Conf.* 2010: 800-804.

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

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

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

[7] 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.

[8] S Lee, A. Potamianos, and S. Narayanan. "Acoustics of children’s speech: Developmental changes of temporal and spectral parameters." *The Journal of the Acoustical Society of America* 105.3: 1455-1468, 1999

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

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

[11] S. M. Lulich, “Subglottal resonances and distinctive features,” *Journal of Phonetics*, vol. 38, pp. 20-32, 2010.