Analysis and automatic estimation of children's subglottal resonances

Steven M. Lulich[1]*, Harish Arsikere*[2]*, John R. Morton*[1]
Gary K. F. Leung[2]*, Abeer Alwan*[2]*, Mitchell S. Sommers*[1]

[1]Dept. of Psychology, Washington University, St. Louis, Missouri
[2]Dept. of Electrical Engineering, UCLA, Los Angeles, California
slulich@wustl.edu, harishan@ucla.edu, jrmhvc333@yahoo.com
garyleung@ucla.edu, alwan@ee.ucla.edu, msommers@wustl.edu

Abstract

Models and measurements of subglottal resonances are generally made from adult data, but there are several applications in which it would be useful to know about subglottal resonances in children. We therefore conducted an analysis of both new and old recordings of children's subglottal acoustics in order 1) to produce a fuller picture of the variability of children's subglottal resonances, and 2) to confirm that existing models of subglottal acoustics can be reasonably applied to children. We also tested the effectiveness of recent algorithms for estimating children's subglottal resonances from speech formants and the fundamental frequency, which were originally formulated based on adult data. It was found that these algorithms are effective for children at least $150cm$ tall.

Index Terms: subglottal resonances, child speech, speech production, speaker normalization

1. Introduction

Previous research on subglottal resonances (SGRs) has focused primarily on adult speakers [1, 2, 3, 4, 5, 6, 7]. To the best of our knowledge, direct recordings and analysis of children's subglottal acoustics has been carried out in only one previous study [1]. In that study, only the second subglottal resonance (Sg2) was reported, for seven children ranging in age from 2 years, 2 months (2;2) to 16 years, 9 months (16;9). Six of these children were female, and only one was male. A few additional studies have estimated the first and second subglottal resonances of children from their putative effects on formant trajectories [8, 9], but such measurements are not always straight-forward and are always less conclusive than direct measurements on recordings of subglottal acoustics.

There are two particularly important applications for the study of SGRs in children. On the one hand, it has been claimed that SGRs play a role in the development of phonological inventories [1, 8, 10]. Understanding children's SGRs and their relation to formants (phonetically) and sound contrasts (phonologically) may therefore yield insights into the development of language skills in children, and perhaps also mechanisms of diachronic language change. On the other hand, automatic speech recognition (ASR) is increasingly being used in tools aimed at evaluating and improving children's academic achievement, and ASR systems for children's speech may benefit from incorporating information about children's SGRs [9].

In this paper, we reanalyzed data from the seven children reported in [1] to include measurements of the first three SGRs (Sg1, Sg2, Sg3). We also made new recordings and measured the SGRs from six male and two female children ranging in age from infancy (0;1) to 17 years, 6 months (17;6). The data were used to test the accuracy of 1) recent physical models of subglottal acoustics, and 2) automatic SGR estimation algorithms which were developed on the basis of adult data. In Section 2 we describe our recording, analysis, and automatic estimation methods. Results are presented and discussed in Section 3, and Section 4 concludes the study.

2. Methods

2.1. Recordings

The seven children recorded and analyzed in [1] are referred to as B1 (male) and G1-6 (female). Except for children B1 and G1, these children were recorded saying sentences of the form 'Say __ again', where hVd words were inserted in the blank. Vowels used in the hVd words are given in Table 1. Child G1 was too young to read these sentences, so instead the experimenter pointed to objects (e.g. his head, hand, etc.) and she named them. Further details can be found in [1]. Child B1 was recorded using a similar procedure to Child G1 ([1] mistakenly omitted this detail). Each of the children was recorded by a microphone and, simultaneously, an EMkay BU-1771 accelerometer glued to the skin of the neck below the thyroid cartilage. The accelerometer signal can be considered a good approximation to the subglottal input impedance, because the motion of the tissues and skin of the neck are related to the pressures at the top of the trachea, and the phonation source is principally a volume velocity (dipole) source [10]. All of these recordings were made with a sampling rate of $16kHz$ and $16bit$ resolution. The children's heights were also measured.

New recordings of subglottal acoustics were obtained from six male children, referred to as B4-B9,and two female children, referred to as G7-8. Children B4-8 were recorded in a sound-attenuated booth at Washington University, reading sentences of the form 'I said a __ again', where the blank was filled with 'target' hVd words as well as CVb words, where the C was one of [b], [d], or [g]. The sentence was constructed so as to provide a neutral phonetic context for the target words. The complete set of words is given in Table 1. Microphone recordings were made with simultaneous accelerometer recordings. The accelerometer was a K&K Sound *HotSpot*, and was pressed against the skin of the neck below the thyroid cartilage by the speakers themselves. Before making these recordings, two additional recordings of just the accelerometer signal were made while the children produced a sustained [a] vowel, with feedback from the experimenter in order to obtain the highest possible quality of accelerometer signal. All of these recordings were made with a sampling rate of $48kHz$ and $16bit$ resolution.

Table 1: *List of vowels recorded in hVd and CVb words.*

hVd	i	ɪ	ɛ	æ	a	ʌ, o, ʊ	u	aɪ, aʊ, ɔɪ	r
bVb	i	-	ɛ	-	a	-	-	aɪ, aʊ, ɔɪ	-
dVb	i	-	ɛ	-	a	-	u	aɪ, aʊ, ɔɪ	-
gVb	i	-	ɛ	-	a	-	u	aɪ, aʊ, ɔɪ	-

The children's heights were also measured.

Child B9 was recorded using a separate protocol, due to his age. Recordings were made somewhat periodically over several months (up to age 9 months). For each recording session, both microphone and accelerometer signals were obtained. The K&K Sound *HotSpot* accelerometer was placed (usually) on his back near the spine in the vicinity of the T2 vertebral process, or on the chest just below the neck on the sternum. The accelerometer was held in place with light pressure by the child's father until the child produced a suitable vocalization. These recordings usually lasted only a few seconds, and no more than about one minute. The recordings were made with a sampling rate of $16kHz$ and $16bit$ resolution. This child's length (i.e. height) was also measured the same week as the recordings.

In addition to the child recordings, we used the WashU-UCLA Corpus (male speakers 12, 13, 15, 17, 21, 22, 41, 44, and female speakers 14, 16, 18, 19, 20, 24, 25) of adult speech and accelerometer recordings [11] for training automatic SGR estimation algorithms [12, 13]. The recording procedure was identical to the one used to record child speakers B4-8.

2.2. Analysis

The first three subglottal resonances (Sg1, Sg2, Sg3) for children B1, B4-B8, and G1-6 were measured by a combination of LPC (linear predictive coding), autocorrelation-based smoothed spectra [14], and visual inspection of FFT spectra and spectrograms computed from the accelerometer signals. In all three methods the signals were down-sampled to $8kHz$. For the FFT analysis, a Hamming window was used with length equal to 4 pitch periods. For the LPC analysis, the LPC order was between 10 and 14, depending on whether the first two harmonics (H1 and H2) dominated the spectrum at low frequencies, thus requiring an increased number of poles to reveal Sg1.

The autocorrelation-based smoothed spectra were computed using the method described in [14]. We chose a segment of speech 4 pitch periods long in the steady-state portion of the vowel. This segment (or 'frame') was then divided into a number of subframes, and a Hamming window was applied to each subframe. The autocorrelation function was found for each subframe and averaged over all the available subframes. An FFT spectrum of the average was then taken. and this was the smoothed spectrum. The frame size was 4 pitch periods long, and the subframe size was between 0.9 and 1.1 pitch periods long, depending on whether H1 and H2 dominated the spectrum at low frequencies, thus requiring increased frequency resolution. The overlap between the subframes was 80% of the subframe size.

The FFT spectrum, the LPC spectrum, and the smoothed spectrum were plotted on a single graph (as shown in Figure 1), and with the FFT spectrum as a guide, the frequencies of Sg1, Sg2, and Sg3 were measured from either the LPC spectrum or the smoothed spectrum, depending on which was judged to give the more accurate result. Note that this procedure is more accurate than the one employed in [1] and [9] for Speakers B1 and G1-6, and the values of Sg2 reported here therefore supersede the values reported previously.

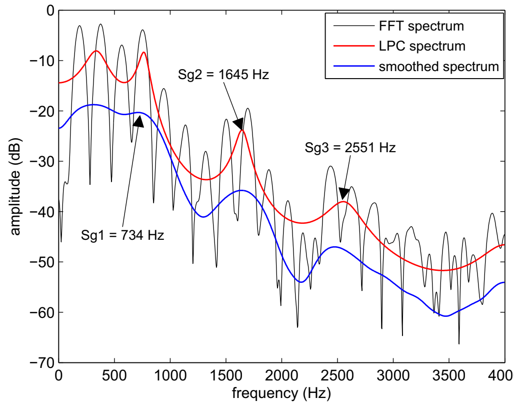

Figure 1: *FFT, LPC, and smoothed spectra of a sample accelerometer signal produced by Speaker G4. Note that the lowest two harmonics are particularly high in energy, producing an additional peak in both the LPC and smoothed spectra.*

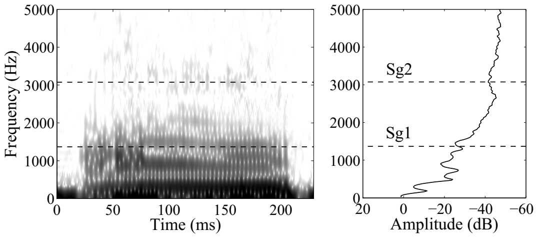

Figure 2: *Spectrogram (left) and averaged FFT spectrum (right) of an accelerometer recording from child B9 at approximately 4 months of age, showing the location of Sg1 and Sg2.*

For Speakers B4-6 and G2-6, the analysis was performed for one token of each of the 9 monophthongs and [r] in the hVd context, resulting in a total of 10 measurements per speaker. In all 10 cases, Sg1 and Sg2 were measured. Sg3 could not be measured in all cases due to its proximity to the noise floor in accelerometer recordings. For Speakers B1 and G1, Sg1 and Sg2 measurements were made in 10 different utterances, but without consideration of the vowel identity due to the different recording procedure and generally lower quality of the accelerometer signal. Sg3 was also measured in as many of these cases as possible. For each speaker, the mean values of Sg1, Sg2, and Sg3 were recorded as the 'ground truth'. For Speakers B7-8, measurements were made based on the high quality accelerometer signals using the spectrogram and the FFT and LPC spectra.

For Speaker B9, Sg1 and Sg2 were measured by visual inspection of FFT spectra and spectrograms. Sg3 was not visible in any recording, and Sg2 was present in only a few recordings. When possible, measurements taken over multiple vocalizations from a single recording session were averaged together. An example spectrogram with both Sg1 and Sg2 present is given in Figure 2. Note that these measurements provide the first longitudinal data on subglottal acoustics ever reported, to the best of our knowledge.

2.3. Automatic Estimation

We tested three algorithms for estimating Sg1 and Sg2 on Speakers B1, B4-6 and G2-6. The first algorithm, denoted $A1$, was developed and evaluated in [9] using the recordings of Speakers G2-6 (it was called 'Sg2D2' in [9] to distinguish

Table 2: *Summary of algorithms for estimating children's SGRs.*

	SGRs	Sample	Speakers
$A1$	Sg2	Isolated vowels	G4-6
$A2_a$	Sg1, Sg2	Isolated vowels	G4-6
$A2_b$	Sg1, Sg2	Continuous speech	B1, B4-6, G2-6

it from an earlier, less accurate algorithm). This algorithm estimated Sg2 in a two-step process: first, an initial estimate of Sg2 was obtained from the third formant (F3) using an empirical relation between Sg2 and F3 [1]; second, a frequency jump in the second formant (F2) trajectory within $\pm 100Hz$ of this initial estimate was searched for. If a frequency jump occurred, the F2 values at the beginning and end of the jump were averaged to yield a refined Sg2 estimate. If a frequency jump did not occur, the initial estimate of Sg2 based on F3 alone was retained. Since the linear relation between Sg2 and F3 is not valid beyond puberty [1], algorithm $A1$ was not successful when applied to adult data. Moreover, this algorithm requires vowels to be isolated before use.

The second and third algorithms, denoted $A2_a$ and $A2_b$, were recently developed and evaluated using adult data [12, 13]. These two algorithms estimate Sg2 by taking measurements of F2 and F3, and they estimate Sg1 by taking measurements of F0 and F1. For Sg2 estimation the bark difference between F3 and F2, denoted ${}_{F3}D_{F2}$, was found to be related to the bark difference between F2 and Sg2, ${}_{F2}D_{Sg2}$, by means of a cubic polynomial (cf. Eq. 1). These parameters were chosen because they both can be used as a measure of vowel 'backness' [1, 15]. Similarly, for Sg1 estimation the bark difference between the first formant (F1) and the fundamental frequency (F0), ${}_{F1}D_{F0}$, was related to the bark difference between F1 and Sg1, ${}_{F1}D_{Sg1}$, since both of these parameters can be used as a measure of vowel 'height' [10, 15]. In [13], a cubic polynomial was used to fit the data, but in this paper we use a linear fit, since it is computationally simpler and fits the data almost equally well (cf. Eq. 2). Eleven adult speakers were used in [12, 13] to discover these relations. For this study, the number of speakers was expanded to 15. Equation 1 accordingly differs slightly from the original equation in [12]. From Equations 1 and 2 it is straight-forward to obtain the estimated values of Sg1 and Sg2 in linear (Hertz) frequency. Algorithm $A2_a$ was applied to isolated vowels which had previously been identified and labeled, and its results are therefore directly comparable with those of [9]. Algorithm $A2_b$ was applied to running speech without any prior labeling. See [12, 13] for further details.

$$\begin{aligned}{}_{F2}D_{Sg2} &= 0.0061 \cdot ({}_{F3}D_{F2})^3 + 0.0205 \cdot ({}_{F3}D_{F2})^2 \\ &\quad - 1.5926 \cdot ({}_{F3}D_{F2}) + 5.8099, (r^2 = 0.8908)\end{aligned} \tag{1}$$

$${}_{F1}D_{Sg1} = 0.9190 \cdot ({}_{F1}D_{F0}) - 3.8218, (r^2 = 0.9546) \tag{2}$$

$A2_a$ was applied to Speakers G4-6, since $A1$ was also evaluated on these speakers (but [9] reported vowel-by-vowel Sg2 estimates only for Speaker G5). $A2_b$ was applied to Speakers B1, B4-6, and G2-6. In this paper, we report two metrics of algorithm performance: the average error of the estimates, and the standard deviation of the estimates. Table 2 summarizes the three algorithms.

3. Results and Discussion

3.1. Analysis

Mean values of Sg1, Sg2, and Sg3 for each of the child speakers except B9 are given in Table 3, along with their heights and ages. The heights and SGRs for the various recordings of speaker B9 are given in Table 4.

Table 3: *Age, height, and SGRs for each of the child speakers except B9.*

ID	Age (yr;mo)	Height (cm)	Sg1 (Hz)	Sg2 (Hz)	Sg3 (Hz)
B1	9;0	134.6	781	1965	3152
B4	11;5	142.9	704	1712	2682
B5	16;3	174.0	506	1438	2421
B6	17;8	181.9	494	1249	2019
B7	7;1	106.7	789	2187	3463
B8	8;2	130.2	688	1903	2944
G1	2;2	85.4	1056	2668	–
G2	6;10	121.8	882	2142	3389
G3	9;0	135.6	760	1891	3198
G4	12;6	154.9	691	1635	2709
G5	13;11	164.2	650	1513	2505
G6	16;9	162.1	655	1547	2565
G7	10;8	129.5	668	2096	3179
G8	10;10	137.2	627	1913	2815

Table 4: *Recording session number (#), height (Ht.), and SGRs for child speaker B9. A degree of measurement error in the longitudinal height and SGR measurements from one recording to the next is apparent, but the overall trend is toward growth and lower frequency SGRs.*

#	Ht. (cm)	Sg1 (Hz)	Sg2 (Hz)	#	Ht. (cm)	Sg1 (Hz)	Sg2 (Hz)
1	52.1	1417	–	13	59.7	1405	–
2	53.3	1659	–	14	59.7	1000	–
3	55.9	1734	–	15	62.2	1088	–
4	57.2	1316	–	16	62.9	1367	3075
5	57.2	1240	–	17	62.9	1278	–
6	55.9	1265	3468	18	69.9	1260	3417
7	55.9	1468	–	19	66.0	1189	–
8	58.4	1375	3107	20	66.0	1088	–
9	58.4	1341	–	21	71.1	1054	3362
10	58.4	1550	–	22	71.8	1265	–
11	59.7	1493	–	23	71.8	850	–
12	59.7	1645	–				

Panel A in Figure 3 shows the relation between speaker height and SGRs. Sg1 is indicated by circles, Sg2 by squares, and Sg3 by diamonds. The green symbols are for child B9; blue symbols are for children G1-8; red symbols are for children B1 and B4-8; and black symbols are for adults (previously reported in [16]). The solid lines are the first three quarter-wavelength resonances of a uniform tube open at the distal end, the length of which is varied as the height divided by 8.561 [16]. The child data extend the trend of the adult data, showing that SGRs increase as height decreases, and that SGRs can be approximately predicted by height alone assuming that they are quarter-wavelength resonances of an equivalent uniform tube. The SGRs for children with heights in the range of $120cm$ to $150cm$ are a bit higher in frequency than expected, and this could be due to a faster rate of overall growth relative to the rate of growth of the subglottal airways during this age range (6 to 12 yrs). Interestingly, the longitudinal data for child B9 follow the quarter-wavelength curves exceptionally well.

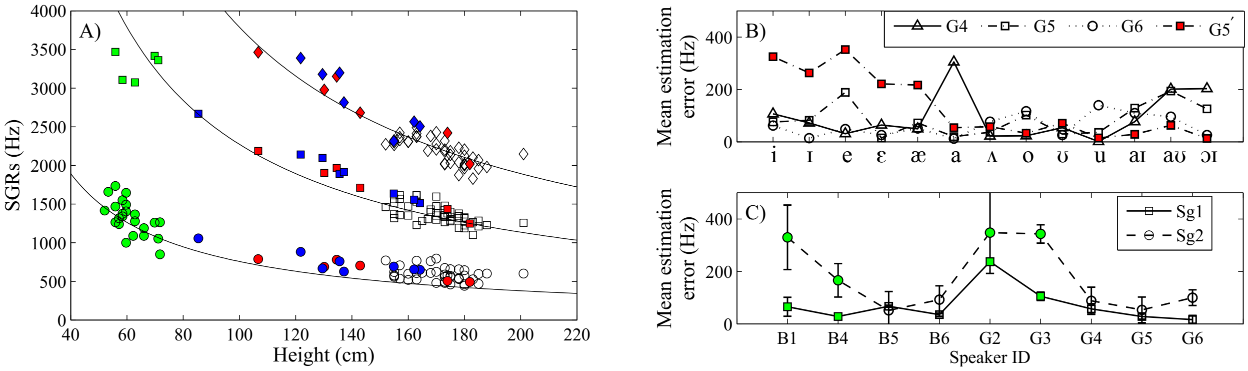

Figure 3: Panel A: *SGRs versus height (h) for adults [16] (black symbols), female children G1-G8 (blue symbols), male children B1, B4-B8 (red symbols), and male child B9 (green symbols). Circles: Sg1; squares: Sg2; diamonds: Sg3. Solid lines: quarter-wavelength resonances of an equivalent uniform tube with length equal to $h/8.561$ [16].* Panel B: *Vowel-by-vowel mean Sg2 estimate errors (Hz) for speakers G4-6 using algorithm $A2_a$, and using algorithm $A1$ for speaker G5 (red; denoted G5′).* Panel C: *Mean estimation errors and standard deviations of estimated values for Sg1 and Sg2 using algorithm $A2_b$. Green filled symbols represent speakers less than $150cm$ tall.*

3.2. Automatic Estimation

Panel B in Figure 3 shows the vowel-by-vowel estimation errors for Speakers G4-G6 using algorithms $A1$ and $A2_a$. Panel C shows the speaker-by-speaker results of $A2_b$. This algorithm results in mean errors similar to those of $A2_a$ for speakers G4-6. For speakers taller than approximately $150cm$, the mean errors are less than $100Hz$ for Sg2, and less than $70Hz$ for Sg1. The accuracy of $A2_b$ decreases for speakers shorter than approximately $150cm$ (B1, B4, G2-3, green filled symbols). For the taller speakers, the standard deviation of the estimated Sg1 and Sg2 values are less than $20Hz$ and $75Hz$, respectively, indicating a high degree of reliability of the algorithm as it applies across multiple utterances. For estimation of Sg2 for shorter speakers, $A1$ is more accurate than either $A2_a$ or $A2_b$.

4. Conclusion

This study was intended to produce a fuller picture of children's SGRs and their relationship to adult values, and to test the effectiveness of recent algorithms for estimating children's SGRs from speech formants and the fundamental frequency, which were originally formulated based on adult data [12, 13]. Only one previous study has presented children's SGRs measured from accelerometer signals directly [1]. Those signals were reanalyzed here and combined with new recordings, including a number of longitudinal recordings representing the first 9 months of infancy (Speaker B9). The SGRs of children are related to height in much the same way as those of adults, although it appears that children between 6 and 13 years of age grow taller at a faster rate than their subglottal airways, thus resulting in SGR frequencies slightly higher than predicted.

For children shorter than approximately $150cm$, the new algorithms developed in [12, 13] to estimate Sg2 in adult speech are less accurate than the algorithm presented in [9]. For taller children the new algorithms are superior. Moreover, the new algorithms are able to estimate Sg1, which the algorithm in [9] could not do. One of the new algorithms ($A2_b$) can be applied to unlabeled, running speech without any appreciable decrease in accuracy.

5. Acknowledgements

This work was supported in part by NSF Grant No. 0905250.

6. References

[1] Lulich, S. M., "Subglottal resonances and distinctive features", J. Phonetics, 38(1):20–32, 2010.

[2] Chi, X., M. Sonderegger, "Subglottal coupling and its influence on vowel formants", J. Acoust. Soc. Am. (JASA), 122:1735-1745, 2007.

[3] Habib, R. H., R. B. Chalker, B. Suki, A. C. Jackson, "Airway geometry and wall mechanical properties estimated from subglottal input impedance in humans", J. Appl. Physiol., 77(1):441-451, 1994.

[4] Ishizaka, K., M. Matsudaira, T. Kaneko, "Input acoustic-impedance measurement of the subglottal system", JASA, 60(1):190-197, 1976.

[5] Cranen, B., L. Boves, "On subglottal formant analysis", JASA, 81(3):734-746, 1987.

[6] Klatt, D. H., L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", JASA, 87(2):820-857, 1990.

[7] Hanson, H. M., K. N. Stevens, "Sub-glottal resonances in female speakers and their effect on vowel spectra", Proc. XIIIth ICPhS, Stockholm, 3:182-185, 1995.

[8] Jung, Y., "Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back]", PhD dissertation, MIT, 2009.

[9] Wang, S., A. Alwan, S. M. Lulich, "Automatic detection of the second subglottal resonance and its application to speaker normalization", JASA, 126:3268-3277, 2009.

[10] Stevens, K. N., "Acoustic Phonetics", MIT Press: Cambridge, 1998.

[11] Lulich, S. M., J. R. Morton, M. S. Sommers, H. Arsikere, Y.-H. Lee, A. Alwan, "A new speech corpus for studying subglottal acoustics in speech production, percpetion, and technology", JASA, 128(4):2288(A), 2010.

[12] Arsikere, H., S. M. Lulich, A. Alwan, "Automatic estimation of the second subglottal resonance from natural speech", ICASSP, 4616-4619, 2011.

[13] Arsikere, H., S. M. Lulich, A. Alwan, "Automatic estimation of the first subglottal resonance", JASA, 129(5):EL197-EL203, 2011.

[14] Umesh, S., L. Cohen, N. Marinovic, D. J. Nelson, "Scale transform in speech analysis", IEEE Trans. Speech Audio Proc., 7(1):40-45, 1999.

[15] Syrdal, A. K., H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels", JASA, 79(4):1086-1100, 1986.

[16] Arsikere, H., Y.-H. Lee, S. M. Lulich, J. R. Morton, M. S. Sommers, A. Alwan, "Relations among subglottal resonances, vowel formants, and speaker height, gender, and native language", JASA, 128(4):2288(A), 2010.