

AUTOMATIC ESTIMATION OF THE SECOND SUBGLOTTAL RESONANCE FROM NATURAL SPEECH

Harish Arsikere* Steven M. Lulich^{†,‡} Abeer Alwan*

* Department of Electrical Engineering, University of California, Los Angeles

† Department of Psychology, Washington University, Saint Louis

‡ Department of Speech and Hearing Sciences, Indiana University, Bloomington
harishan@ucla.edu, slulich@wustl.edu, alwan@ee.ucla.edu

ABSTRACT

This paper deals with the automatic estimation of the second subglottal resonance (Sg2) from natural speech spoken by adults, since our previous work focused only on estimating Sg2 from isolated diphthongs. A new database comprising speech and subglottal data of native American English (AE) speakers and bilingual Spanish/English speakers was used for the analysis. Data from 11 speakers (6 females and 5 males) were used to derive an empirical relation among the second and third formant frequencies (F2 and F3) and Sg2. Using the derived relation, Sg2 was automatically estimated from voiced sounds in English and Spanish sentences spoken by 20 different speakers (10 males and 10 females). On average, the error in estimating Sg2 was less than 100 Hz in at least 9 isolated AE vowels and less than 40 Hz in continuous speech consisting of English or Spanish sentences.

Index Terms— subglottal resonances, automatic estimation, bilingual speech, speaker normalization

1. INTRODUCTION

Subglottal resonances (SGRs) have received increasing attention in the last few years, in both the speech science and engineering communities. It has been hypothesized that Sg2 defines the boundary between front and back vowels, and that the first subglottal resonance (Sg1) defines the boundary between high and low vowels [1, 2]. Strong evidence has recently been presented in support of the [±back] division by Sg2 [3]. It has also been demonstrated that Sg2 affects the way humans perceive the distinctive feature [back] when F2 crosses Sg2 in transitioning from a high to a low value [4]. The fact that Sg2 lies at the boundary between front and back vowels has also been demonstrated in other languages [5, 6, 7].

Acoustic coupling between the subglottal system and the vocal tract has an influence on the frequencies and amplitudes of vowel formants. Chi and Sonderegger studied the influence of Sg2 on the second formant in detail [8]. They found that a discontinuity in F2 and a dip in the amplitude of the second formant (A2) can often be observed in the vicinity of Sg2 in back-to-front diphthongs of American English. Wang et al. [9] developed an algorithm for automatically estimating Sg2 from children’s speech based on the observations made in [8] and on a relation between F3 and Sg2 derived in [3]. In fact, the estimation algorithm was successfully applied to perform normalization and cross-language adaptation of children’s speech for use in Automatic Speech Recognition (ASR) tasks.

In this paper, algorithms are proposed to automatically estimate Sg2 in adults’ speech, based on the relation between two measures

of vowel *backness*. Section 2 describes the database used. Section 3 describes novel methods for measuring Sg2, the procedure used for deriving an empirical relation among F2, F3 and Sg2, and the algorithms for automatically estimating Sg2 using the derived relation. The results of automatic estimation are presented in Section 4. Section 5 summarizes the paper.

2. DATABASE

A database comprising simultaneous speech and subglottal recordings was recently collected [10] with the intention of studying the properties of SGRs and their effects on speech. Speech data were recorded using a Shure PG27 condenser microphone and subglottal data were obtained using a K&K Sound ‘Hot Spot’ accelerometer. All recordings were sampled at 48 kHz and digitized at 16 bits/sample. The database consists of two sets. Set 1 comprises data from 25 female and 25 male adult native speakers of American English (AE) aged between 18 and 25 years. Set 2 comprises data from 4 female and 6 male adult bilingual speakers of Mexican Spanish and AE aged between 18 and 25 years. Every speaker was recorded in two sessions. The first session, which was common to all the speakers in Sets 1 and 2, involved recording 21 nonsense CVb words embedded in the phrase “*I said a ____ again*”, where ‘C’ was one of the voiced stops [b], [d] and [g] and ‘V’ was one of the vowels in column 1 of Table 1. For native AE speakers, the second session involved recording 14 nonsense hVd words embedded in the same carrier phrase, where ‘V’ was one of the vowels in column 2 of Table 1. The second session for bilingual speakers involved recording 21 nonsense CVb words embedded in the Spanish phrase “*Dije una ____ otra vez*”, where ‘C’ was one of the voiced stops [b], [d] and [g] and ‘V’ was one of the vowels in column 3 of Table 1. Each word was repeated 10 times by the native AE speakers and 7 times by bilingual speakers. The start, steady state and end times of the target vowel were labeled manually in each microphone recording. Data from only 31 subjects were used for the present study - 11 for training (6 females and 5 males in Set 1) and 20 for testing (8 fe-

Sets 1&2, Session 1	Set 1, Session 2	Set 2, Session 2
[i], [ε], [a], [u]	[i], [i], [ε], [æ], [a], [A], [o], [ō], [u], [r]	[i], [e], [o], [u]
[ai], [aō], [oi]	[e], [ai], [aō], [oi]	[ai], [au], [oi]

Table 1. List of vowels recorded in sessions 1 and 2 for speakers in Set 1 (native AE) and Set 2 (bilingual). Monophthongs and diphthongs are listed above and below the double line, respectively.

males and 8 males in Set 1, 2 females and 2 males in Set 2). It must be noted that both isolated vowels and complete sentences (carrier phrases with target words) were used in our experiments.

3. METHODS

3.1. A Bark scale relation between F2, F3 and Sg2

An algorithm based on a linear relation between F3 and Sg2 was previously developed to estimate Sg2 in children’s speech [9]. A similar approach could not be used in the case of adults because F3 and Sg2 were found to be weakly correlated. We hypothesized, however, that the *Bark* difference between F3 and F2 (denoted f_3D_{f2}) would be correlated with the *Bark* difference between F2 and Sg2 (denoted f_2D_{s2}), since both measures can be used to represent vowel *backness* [11, 3]. The relation between a frequency f in Hz and its corresponding Bark value z is given by [12]

$$z = [(26.81f)/(1960 + f)] - 0.53. \quad (1)$$

Front vowels have high F2, for which f_3D_{f2} is usually less than 3 Bark. The converse is true for back vowels. Since f_3D_{f2} can be computed readily from speech, our goal of automatic Sg2 estimation required finding a relation between f_3D_{f2} and f_2D_{s2} .

Data from 6 female speakers (14, 16, 18, 19, 20 and 24) and 5 male speakers (12, 13, 15, 17 and 21) belonging to Set 1 were used to derive a relation between f_3D_{f2} and f_2D_{s2} . First, the *ground truth* Sg2 of each speaker was obtained using 30 accelerometer and 6 microphone signals, as follows. Sg2 was *directly* measured in 3 accelerometer signals of each of the monophthongs recorded in Session 2, in a *semi-automatic* manner using Snack [13]. Signals were down sampled to 6 kHz since the first 3 SGRs are expected to lie below 3 kHz, and the formant tracker’s LPC order was set to 12. A 49 ms Hamming window spaced at 5 ms intervals was used. In general, the above parameters resulted in the best alignment of the formant contours with the spectrograms, although small changes had to be made in some cases. For each token, the resonance of the accelerometer signal in the range of 1100-1700 Hz was recorded as the measured value. The correctness of the accelerometer measurements was ascertained by measuring Sg2 *indirectly* in 3 microphone signals of each of the diphthongs [aɪ] and [ɔɪ]. For each token, F2 was tracked semi-automatically using Snack. A window length between 1 and 3 pitch periods was chosen in order to clearly discern the Sg2-induced jump in F2. Figure 1(a) shows one such example. As shown in the figure, the average of the high and low F2 values constituting the jump was recorded as the measured value. In roughly 90% of the diphthongs analyzed, the jump in F2 was clearly observable, and the indirect and direct measurements agreed to within 40 Hz of each other. Finally, the mean of all the Sg2 measurements was recorded as the *ground truth*. Figure 1(b) shows the mean and standard deviation of Sg2 measurements for all training speakers. Standard deviations across vowels range between 25 Hz and 78 Hz and their corresponding Coefficients of Variation (COVs) (ratio of standard deviation to mean) range between 1.8% and 5.9%. Therefore, an estimate of Sg2 which lies within 5%-10% or within 100 Hz of the ground truth can be considered to be reasonably good.

Once the ground truths were obtained for all speakers in the training set, 5 measurements of F2 and F3 were made in the steady-state portion of each of the monophthongs (except [r]) recorded in Session 2. In all, 495 tokens were analyzed. The vowel [r] was not used because its third formant drops significantly below the nominal value. As with Sg2, F2 and F3 were obtained using Snack as described above, but the microphone signals were down sampled to 10

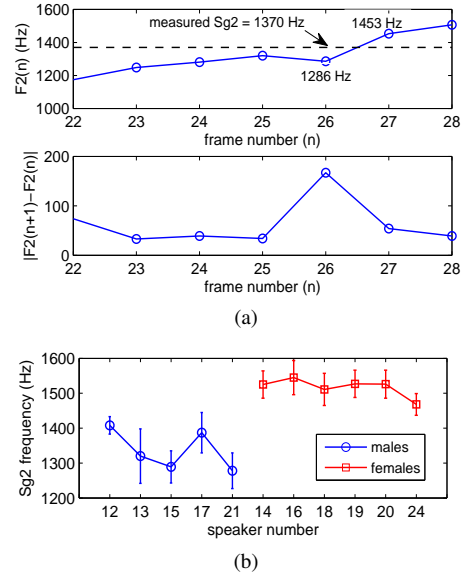


Fig. 1. (a) An example of the F2 jump observed in the diphthong [ɔɪ]. The upper panel shows the frame by frame F2 track. The dashed line passes through the average of the low and high F2 values constituting the F2 jump. The lower panel shows the absolute first difference of the F2 track. (b) Mean and standard deviation of Sg2 measurements for all training speakers.

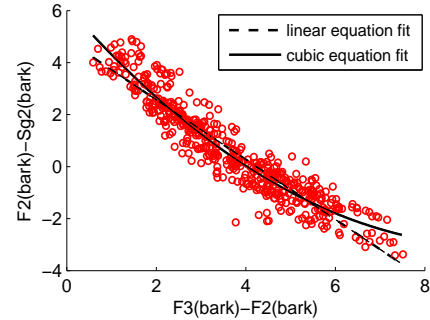


Fig. 2. A scatter plot of the Bark difference between F2 and Sg2 versus the Bark difference between F3 and F2 obtained using 5 measurements each from 9 vowels spoken by 11 speakers (495 tokens).

kHz for formant tracking. All ground truth Sg2 values and the formant measurements were converted to corresponding Bark values using Eq. (1). Then, 495 f_3D_{f2} values and their corresponding f_2D_{s2} values were computed. Figure 2 shows a scatter plot of f_2D_{s2} versus f_3D_{f2} . Clearly, the two quantities have a high degree of correlation ($\rho = -0.9396$). Since F3 is always higher than F2, f_3D_{f2} is always positive. However, f_2D_{s2} can be positive or negative depending on whether F2 is higher or lower than Sg2, respectively. As f_3D_{f2} increases, the degree of vowel *backness* increases, and when f_3D_{f2} is around 4 Bark, f_2D_{s2} starts assuming negative values. This is reasonable because vowels with the feature [+back] have f_3D_{f2} values higher than 3 Bark on average [11]. The figure also shows a linear fit ($r^2 = 0.8758$) and a cubic polynomial fit ($r^2 = 0.8908$) to the data. For the automatic estimation of Sg2, we decided to use the following equation describing the cubic polynomial since it forms a better fit to the data than the linear relation.

$$f_2D_{s2} = -0.0004(f_3D_{f2})^3 + 0.1075(f_3D_{f2})^2 - 1.9540(f_3D_{f2}) + 6.1555 \quad (2)$$

3.2. Automatic estimation of Sg2 from vowels

Ten tokens of each of the vowels (except [ɪ]) recorded in Session 2 were excised from data belonging to 8 female speakers (25, 26, 27, 28, 35, 36, 37, 40) and 8 male speakers (22, 23, 29, 31, 38, 41, 43, 44) in Set 1. Given a particular vowel token, Sg2 was estimated using a *frame-by-frame* approach. F3 and F2 were tracked *automatically* (default settings without manual adjustments) using Snack and converted to Bark values using Eq. (1). For each frame i , a Sg2 estimate was obtained as follows. First, $f_3 D_{f_2}^i$ was computed. Then, $f_2 D_{s_2}^i$ was computed using Eq. (2). Finally, $Sg2^i(\text{Bark})$ was calculated by subtracting $f_2 D_{s_2}^i$ from $F2^i(\text{Bark})$. All Bark Sg2 estimates were converted to Hz by inverting Eq. (1), and Sg2 for the given vowel token was evaluated by averaging all the frame-by-frame estimates. Data from the bilingual speakers were not used for this experiment.

3.3. Automatic estimation of Sg2 from continuous speech

Estimating Sg2 from continuous speech is important because one might not have access to excised vowels in real world scenarios. For this experiment, up to 3 sentences of continuous speech were used for each speaker in the testing set. In addition to speakers mentioned in Section 3.2, data belonging to 2 female speakers (1, 6) and 2 male speakers (3, 4) in Set 2 were used. Every sentence, either in English or in Spanish, consisted of one of the carrier phrases mentioned in Section 2 with one of the CVb or hVd words embedded in it. The technique adopted to estimate Sg2 is as follows. First, F2 and F3 were extracted automatically frame-by-frame from the entire length of continuous speech presented. Then, all *voiced* frames were selected with the help of a parameter called *Probability of Voicing (PV)* returned by Snack. Snack sets PV to 1 for voiced frames and to 0 for unvoiced frames. A Sg2 estimate was computed for each voiced frame by following the procedure outlined in the Section 3.2. Finally, a Gaussian distribution was estimated from the pool of Sg2 values obtained for voiced frames, and its mean was recorded as the final Sg2 estimate. In case of bilingual speakers, two separate estimates were obtained for English and Spanish sentences.

4. RESULTS AND DISCUSSION

Figure 3(a) shows results of automatic estimation in excised vowels for a particular male speaker (41) in Set 1 who is representative of the test set. For this particular speaker, the algorithm yields smaller estimation errors for front vowels ([i], [ɪ], [e], [æ]) and diphthongs ([eɪ], [aʊ], [ɔɪ]) as compared to mid ([ɑ], [ʌ]) and back ([o], [ʊ], [u]) vowels. Results of estimation from vowels for all native AE speakers in the test set are summarized in Table 2. Column 2 shows the number of estimates that lie within 10% of the ground truth. Column 3 shows the number and categories of vowels for which the average estimation error is less than 100 Hz. These vowels, according to our reasoning in Section 3.1, can be considered ‘good’ for automatic estimation. Clearly, these ‘good’ categories are speaker dependent. This is due to the fact that Eq. (2) captures the average characteristics of the set of training speakers, and hence might not represent all the test speakers equally well. Except for speakers 22 and 38, the number of ‘good’ vowels is at least 9. The algorithm’s poor performance for these two speakers was the result of erroneous tracking of F3, which was either because F3 was very weak even after pre-emphasis or because F4 was assumed to be F3 owing to its larger energy. After manual reconfiguration of the formant tracker for these two speakers, the average estimation error for diphthongs fell below 100 Hz, and consequently, the number of ‘good’ vowels increased to 9.

spkr	est. error < 10% (no. of tokens)	average est. error < 100 Hz (no. of vowels)
22	69/130	7/13 (m, b)
23	108/130	10/13 (f, m, d)
29	112/130	11/13 (f, m, d)
31	120/130	13/13 (f, m, b, d)
38	74/130	6/13 (f, m)
41	111/130	11/13 (f, b, d)
43	85/130	9/13 (f, m, d)
44	126/130	13/13 (f, m, b, d)
<hr/>		
25	127/130	11/13 (f, m, d)
26	112/130	10/13 (f, b, d)
27	94/130	10/13 (f, d)
28	113/130	13/13 (f, m, b, d)
35	116/130	12/13 (f, m, b, d)
36	98/130	9/13 (f, d)
37	118/130	10/13 (m, b, d)
40	111/130	10/13 (f, m, b)

Table 2. Results of Sg2 estimation from excised vowels. Results for males/females are shown above/below the double line. Columns 2 and 3 indicate the number of estimates within 10% of the ground truth, and the number of vowels for which average estimation error (averaged over all tokens of a given vowel) is less than 100 Hz, respectively. Corresponding vowel categories are shown in parentheses, where ‘f’, ‘m’ and ‘b’ denote *front*, *mid* and *back* vowels and ‘d’ *diphthongs*, respectively.

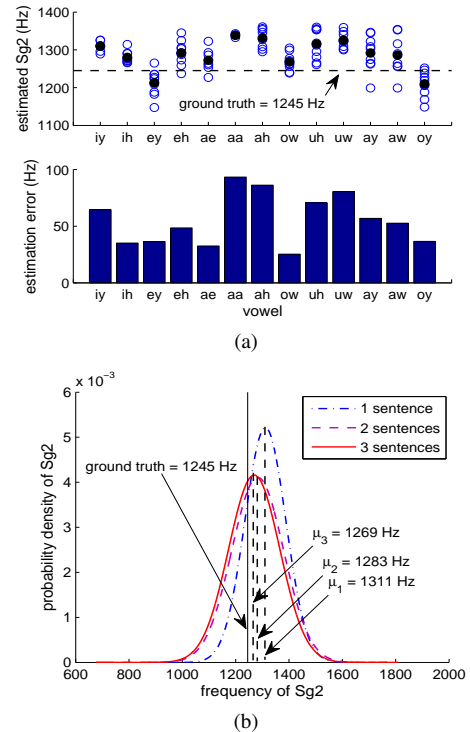


Fig. 3. Automatic estimation of Sg2 for speaker 41: (a) The upper panel shows Sg2 estimates from several tokens of each vowel. Empty and filled circles denote individual and average estimates, respectively. The lower panel shows average estimation errors. (b) Sg2 estimation from continuous speech. Each density function was estimated from frame-by-frame Sg2 estimates.

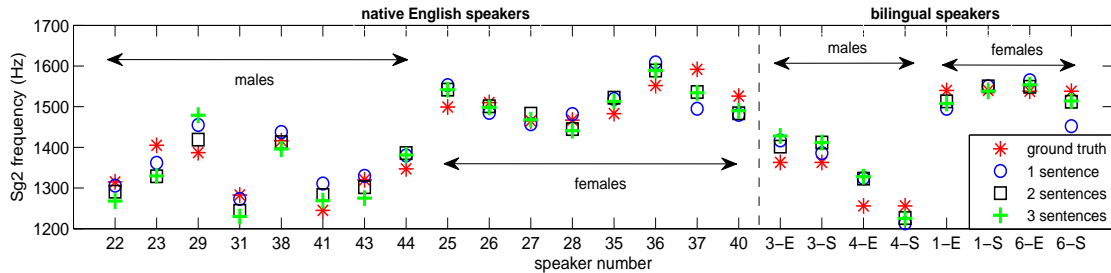


Fig. 4. Sg2 estimation for all test speakers using up to 3 sentences of speech. Results for native AE speakers and bilingual speakers are shown to the left and right of the dashed line, respectively. For bilingual speakers, ‘E’ denotes English data and ‘S’ denotes Spanish data.

Figure 3(b) shows results of automatic estimation from continuous speech for speaker 41. Each Gaussian density function was estimated from a different amount of data, which varied from 1 to 3 sentences. Figure 4 shows results of estimation from continuous speech for all 20 speakers in the test set. Both Figure 3(b) and Figure 4 show that for most speakers, the estimation may improve only slightly as the amount of data increases from 1 to 3 sentences. Hence for practical purposes, one short sentence of continuous speech with some voiced segments can be considered to be sufficient for estimating Sg2. With one sentence of data, the estimation error ranges from 8 Hz (speaker 27) to 97 Hz (speaker 37) and its average over all the test speakers is 40 Hz. This is well within the range of observed standard deviations for Sg2. It is important to note that 1-sentence estimation errors for speakers 22 and 38 are very small - 9 Hz and 22 Hz respectively - although the estimation from vowels is poor in their case. In general, Sg2 estimation is better in continuous speech than in isolated vowels owing to the availability of a larger number of voiced frames. Another important thing to note is that in case of bilingual speakers, the algorithm performs equally well with English and Spanish data. The reason is two fold. Firstly, our approach was based on relating two acoustic measures of vowel backness without incorporating any explicit information about language-specific characteristics of vowels. Secondly, we used only Bark *differences* and not absolute values. Therefore, the algorithm’s performance does not suffer despite the fact that English and Spanish differ significantly in their phonetic content. It must also be noted that the ground truth was assumed to be the same for English and Spanish data because SGRs have been shown to be almost independent of language [9].

5. CONCLUSION

In this paper, algorithms were proposed to estimate Sg2 from adults’ speech. To the best of our knowledge, this is the first attempt at estimating Sg2 from continuous speech and from isolated vowels other than back-to-front diphthongs [9]. In order to make the algorithms independent of spoken content, a novel technique based on relating two acoustic measures of vowel backness was developed. An empirical relation was derived between two perceptually motivated quantities - the Bark difference between F3 and F2, and the Bark difference between F2 and Sg2 - and was used to automatically estimate Sg2 from isolated vowels and continuous speech. It was shown that, on average, the error in estimating Sg2 was less than 100 Hz in at least 9 isolated AE vowels, and less than 40 Hz in continuous speech consisting of English or Spanish sentences. In future, we plan to use the proposed algorithms, in conjunction with algorithms for the automatic estimation of Sg1, for speaker normalization in ASR systems. We also plan to extend our methods to children’s speech and compare them with existing algorithms.

6. ACKNOWLEDGMENTS

We are thankful to John R. Morton for recording and labeling the database, and to Dr. Mitchell S. Sommers for providing valuable suggestions. Thanks to Gary Leung and Juan Cortes for help with measurements. Work supported in part by the NSF.

7. REFERENCES

- [1] K. N. Stevens, “On the quantal nature of speech,” *Journal of Phonetics*, vol. 17, pp. 3–45, 1989.
- [2] H. M. Hanson and K. N. Stevens, “Subglottal resonances in female speakers and their effect on vowel spectra,” in *Proceedings of ICPhS*, 1995, vol. 95.
- [3] S. M. Lulich, “Subglottal resonances and distinctive features,” *Journal of Phonetics*, vol. 38, pp. 20–32, 2010.
- [4] S. M. Lulich, A. Bachrach, and N. Malyska, “A role for the second subglottal resonance in lexical access,” *Journal of the Acoustical Society of America (JASA)*, vol. 122, 2007.
- [5] A. Madsack, S. M. Lulich, W. Wokurek, and G. Dogil, “Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs,” *Proc. LabPhon*, vol. 11, pp. 91–92, 2008.
- [6] T. G. Csapó, Z. Bárkányi, T. E. Grácz, T. Bóhm, and S. M. Lulich, “Relation of formants and subglottal resonances in Hungarian vowels,” in *Proceedings of Interspeech*, 2009, pp. 484–487.
- [7] Y. Jung, “Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back],” *PhD Thesis, Harvard-MIT Division of Health Sciences and Technology, MIT*, 2009.
- [8] X. Chi and M. Sonderegger, “Subglottal coupling and its influence on vowel formants,” *JASA*, vol. 122, 2007.
- [9] 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.
- [10] S. M. Lulich, J. R. Morton, M. S. Sommers, H. Arsikere, Y.-H. Lee, and A. Alwan, “A new speech corpus for studying subglottal acoustics in speech production, perception, and technology(A),” *(A) JASA*, vol. 128, pp. 2288, 2010.
- [11] A. K. Syrdal and H. S. Gopal, “A perceptual model of vowel recognition based on the auditory representation of American English vowels,” *JASA*, vol. 79, pp. 1086–1100, 1986.
- [12] H. Trau Müller, “Analytical expressions for the tonotopic sensory scale,” *JASA*, vol. 88, pp. 97–100, 1990.
- [13] K. Sjölander, “The Snack sound toolkit,” *KTH, Stockholm, Sweden (Online: <http://www.speech.kth.se/snack/>)*, 1997.