

Automatic estimation of the first subglottal resonance

Harish Arsikere^{a)}

Department of Electrical Engineering, University of California, Los Angeles, California 90095
harishan@ucla.edu

Steven M. Lulich^{b)}

Department of Psychology, Washington University, Saint Louis, Missouri 63130
slulich@wustl.edu

Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles, California 90095
alwan@ee.ucla.edu

Abstract: This letter focuses on the automatic estimation of the first subglottal resonance (Sg1). A database comprising speech and subglottal data of native American English speakers and bilingual Spanish/English speakers was used for the analysis. Data from 11 speakers (five males and six females) were used to derive an empirical relation among the first formant frequency, fundamental frequency, and Sg1. Using the derived relation, Sg1 was automatically estimated from voiced sounds in English and Spanish sentences spoken by 22 different speakers (11 males and 11 females). The error in estimating Sg1 was less than 50 Hz, on average.

© 2011 Acoustical Society of America

PACS numbers: 43.72.Ar, 43.70.Kv, 43.72.Ne [DOI]

Date Received: January 1, 2011 Date Accepted: February 18, 2011

1. Introduction

Subglottal resonances (SGRs) have recently been used as a basis for speaker normalization in automatic speech recognition (ASR). Wang *et al.*¹ estimated the second subglottal resonance (Sg2) using measurements of the third formant frequency (F3) and the effects of Sg2 on trajectories of the second formant frequency (F2).² The ratio of estimated Sg2 frequencies of the test and reference speakers was used as the *frequency warping factor* for speaker normalization. A similar study also estimated the third subglottal resonance (Sg3) from a model of subglottal acoustics and performed *piece-wise linear* frequency warping using both Sg2 and Sg3.³ Speaker normalization using Sg1 has not been attempted yet due to the lack of reliable algorithms for estimating Sg1. However, we predict that incorporating Sg1 into speaker normalization will yield additional benefits over using just Sg2 and Sg3, since Sg1 lies at the boundary of [+low] and [−low] vowels just like Sg2 lies at the boundary of [+back] and [−back] vowels.⁴

We propose two algorithms to automatically estimate Sg1 in excised vowels and continuous speech based on the relation between two measures of the vowel feature [+low]. Section 2 describes the database used. Section 3 describes novel methods for measuring Sg1, the procedure used for deriving an empirical relation among the fundamental frequency (F0), F1, and Sg1 and the algorithms for automatically estimating Sg1 using the derived relation. The results of automatic estimation are presented and discussed in Sec. 4. Section 5 summarizes the paper.

^{a)} Author to whom correspondence should be addressed.

^{b)} Also at: Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN 47405.

2. Database

A database comprising simultaneous speech and subglottal recordings was recently collected⁵ with the intention of studying the properties of SGRs and their effects on speech. Speech data were recorded using a Shure PG27 condenser microphone (Shure Inc., Niles, IL) and subglottal data were obtained using an *accelerometer*. All recordings were sampled at 48 kHz and digitized at a resolution of 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 yr. Set 2 comprises data from four female and six male adult bilingual speakers of Mexican Spanish and AE aged between 18 and 25 yr. Every native AE speaker was recorded in two sessions. The first session 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 [i], [ɛ], [a], [u], [ai], [au], and [ɔi]. In the second session, recordings were made of 14 nonsense hVd words embedded in the same carrier phrase, where “V” was one of the vowels [i], [ɪ], [e], [ɛ], [æ], [ɑ], [ʌ], [o], [ɔ], [u], [ai], [au], [ɔi], and [r]. Every bilingual speaker was also recorded in two sessions. The first session was the same as that of the native AE speakers. The second session 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 [i], [e], [o], [u], [ai], [au], and [oi]. Each utterance was repeated ten times by the native AE speakers and seven 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 33 subjects were used for the present study; 11 were used for training (male, AE—12, 13, 15, 17, and 21; female, AE—14, 16, 18, 19, 20, and 24) and 22 for testing (male, AE—22, 23, 29, 31, 38, 41, 43, 44, and 49; male, bilingual—3 and 4; female, AE—25, 26, 27, 28, 33, 35, 36, 37, and 40; female, bilingual—1 and 6), thus ensuring gender balance in both training and test sets. It must be noted that the training set was deliberately kept smaller than the test set in order to assess the generalizability of the proposed estimation algorithms to unseen data.

3. Methods

Previous studies on SGRs^{4,6} have shown that Sg1 usually lies in the range of 500–800 Hz and that females, on average, have higher values of Sg1 than males. Due to acoustic coupling between the subglottal and supraglottal systems, Sg1 has an effect on the frequency (F1) and prominence (A1) of the first formant. For example, in the diphthong [aʊ], F1 often shows a discontinuity and A1 experiences an attenuation as the first formant approaches and crosses Sg1.⁴ Based on this *a priori* knowledge of Sg1, the following methods were devised for our analysis.

3.1. A Bark scale relation between F0, F1, and Sg1

The influence of Sg1 on F1 and A1 can be used to automatically estimate Sg1 in low-to-high diphthongs like [aʊ].⁴ However, our aim was to develop a generic algorithm for estimating Sg1 in any given vowel. We hypothesized that the *Bark* difference between F1 and F0 (denoted ${}_1D_{f0}$) would be correlated with the *Bark* difference between F1 and Sg1 (denoted ${}_1D_{s1}$), since both can be considered as acoustic measures of the vowel feature [+low].^{4,7} The relation between a frequency f in hertz and its corresponding Bark value z is given by⁸

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

High vowels have low F1, for which ${}_1D_{f0}$ is usually less than 3 Bark. The reverse is true for low vowels. Since ${}_1D_{f0}$ can be computed readily from speech, our goal of automatic Sg1 estimation required finding a relation between ${}_1D_{f0}$ and ${}_1D_{s1}$.

Data from six female speakers (14, 16, 18, 19, 20, and 24) and five male speakers (12, 13, 15, 17, and 21) in set 1 were used to obtain a relation between ${}_1D_{f0}$ and ${}_1D_{s1}$.

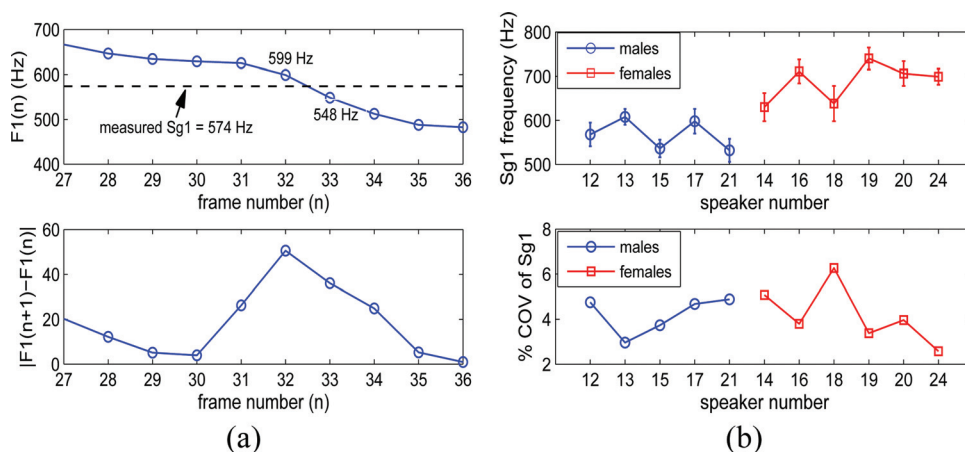


Fig. 1. (Color online) Measurement of Sg1: (a) The upper panel shows the frame-by-frame F1 track in a token of the vowel [au] spoken by speaker 13 (male). The dashed line passes through the average of the F1 values constituting the largest absolute difference (frames 32 and 33). The lower panel shows the absolute first difference of the F1 track. (b) The upper panel shows the mean and standard deviation of Sg1 measurements for all training speakers. The lower panel shows the corresponding percentage COVs.

First, the *actual value* of Sg1 was obtained using 27 accelerometer and 5 microphone signals for each speaker. Sg1 was *directly* measured from three accelerometer signals of each of the vowels [i], [ɪ], [ɛ], [æ], [ɑ], [ʌ], [o], [ʊ], and [u], in a *semi-automatic* manner using Snack.⁹ Signals were down-sampled to 6 kHz since the first three SGRs are expected to lie below 3 kHz and the formant tracker’s LPC (linear predictive coding) order was set to 12. A 49 ms Hamming window spaced at 5 ms intervals was used. In some cases, the above settings were slightly adjusted after visual inspection of the formant tracks and spectrograms. For each token, the resonance of the accelerometer signal in the range of 500–800 Hz was recorded as the measured value. However, it must be pointed out that the measurement of Sg1 in accelerometer signals was not always easy because of its proximity to high-energy harmonics of the fundamental frequency and its interaction with the first formant. In order to verify the correctness of the measurements in accelerometer data, Sg1 was also measured *indirectly* in five microphone signals of the diphthong [au]. For each token, F1 was tracked semi-automatically using Snack. A window length between 1 and 3 pitch periods was chosen in order to clearly discern the Sg1-induced discontinuity in F1. Figure 1(a) shows one such example. As shown in the figure, the two F1 values that yield the largest difference (frames 32 and 33 in this case) are on either side of the observed discontinuity. Therefore, for each token of [au], Sg1 was measured as the average of the F1 values that comprise the largest absolute difference. In roughly 80% of the diphthong tokens analyzed, the discontinuity in F1 was clearly observable and the indirect and direct measurements agreed to within 30 Hz of each other. In the remaining tokens, the indirect measurement was slightly biased, since *two* discontinuities could be observed and the one closer to the direct measurements was chosen for averaging. Finally, the mean of all the Sg1 measurements was recorded as the *actual value*. Figure 1(b) shows the mean, standard deviation, and the percentage coefficient of variation (COV—ratio of standard deviation to mean) of Sg1 measurements for all training speakers. As expected, females have slightly higher Sg1 values than males. Standard deviations range between 18 and 40 Hz and the percentage COVs range between 2.6 and 6.3%. Therefore, an estimate of Sg1 that lies within 5%–10% or within 50 Hz of the actual value can be considered to be reasonably good.

Once the actual values were obtained, five measurements of F1 and F0 were made in the steady-state portion of each of the vowels [i], [ɪ], [ɛ], [æ], [ɑ], [ʌ], [o], [ʊ], and [u], for all speakers in the training set. In all, 495 tokens were analyzed. As before, F1 and F0 values were obtained semi-automatically using Snack, except that the

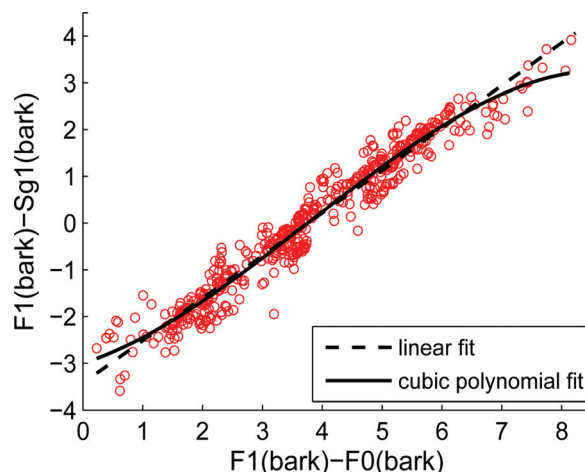


Fig. 2. (Color online) A scatter plot of the Bark difference between F1 and Sg1 versus the Bark difference between F1 and F0 obtained using five measurements from nine vowels spoken by 11 speakers (495 tokens).

microphone signals were down-sampled to 10 kHz (by default) for formant tracking. All actual Sg1 values and the F1 and F0 measurements were converted to corresponding Bark values using Eq. (1). Then, 495 $f_1D_{f_0}$ values and their corresponding f_1D_{s1} values were computed. Figure 2 shows a scatter plot of f_1D_{s1} versus $f_1D_{f_0}$. Clearly, the two quantities have a high degree of correlation ($\rho = 0.974$). Since F1 is always higher than F0, $f_1D_{f_0}$ is always positive. f_1D_{s1} can be positive or negative depending on whether F1 is higher or lower than Sg1, respectively. As $f_1D_{f_0}$ increases, the measure of the feature [+low] increases, and when it is around 4 Bark, f_1D_{s1} starts assuming positive values. This is reasonable because vowels with the feature [+low] have $f_1D_{f_0}$ values higher than 3 Bark on average.⁷ The figure also shows a linear fit ($r^2 = 0.9492$) and a cubic polynomial fit ($r^2 = 0.9533$) to the data. For the automatic estimation of Sg1, we decided to use the following equation describing the cubic polynomial since it forms a slightly better fit to the data than the linear relation

$$f_1D_{s1} = -0.0135(f_1D_{f_0})^3 + 0.1523(f_1D_{f_0})^2 + 0.4168(f_1D_{f_0}) - 3.5046 \quad (2)$$

3.2. Automatic estimation of Sg1 in vowels

Ten tokens of each of the vowels [i], [ɪ], [e], [ɛ], [æ], [a], [ʌ], [o], [ʊ], [u], [ai], [au], and [ɔɪ] were excised from data belonging to nine female speakers (25, 26, 27, 28, 33, 35, 36, 37, and 40) and nine male speakers (22, 23, 29, 31, 38, 41, 43, 44, and 49) in set 1. Given a particular vowel token, Sg1 was estimated using a *frame-by-frame* approach. F1 and F0 were tracked *automatically* (default settings without manual adjustments) using Snack and converted to Bark values using Eq. (1). For each frame i , a Sg1 estimate was obtained as follows: First, $f_1D_{f_0}^i$ was computed. Then, $f_1D_{s1}^i$ was computed using Eq. (2). Finally, $Sg1^i(\text{Bark})$ was calculated by subtracting $f_1D_{s1}^i$ from $F1^i(\text{Bark})$. All the frame-by-frame Bark Sg1 estimates were converted to hertz by inverting Eq. (1) and Sg1 for the given vowel token was evaluated by averaging them. Data from the bilingual speakers were not used for this experiment because each bilingual speaker was recorded saying just seven vowels (in three different contexts).

3.3. Automatic estimation of Sg1 in continuous speech

Estimating Sg1 in continuous speech is important because one might not have access to excised vowels in real world scenarios. For this experiment, up to three sentences of continuous speech were used for each speaker in the testing set. In addition to speakers

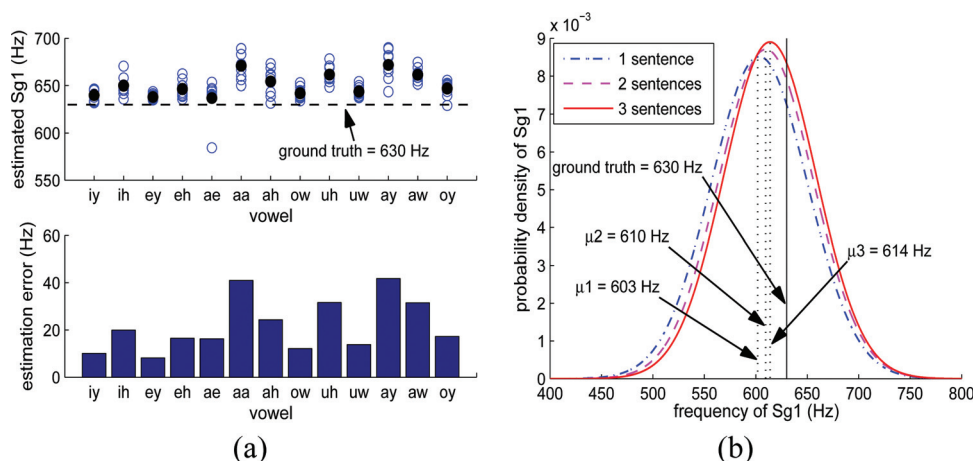


Fig. 3. (Color online) Automatic estimation of Sg1 for speaker 27 (female). (a) The upper panel shows Sg1 estimates in several tokens of each vowel. Empty and filled circles denote individual and average estimates, respectively. The lower panel shows average estimation errors. (b) Sg1 estimation from continuous speech: each density function was estimated from accumulated frame-by-frame Sg1 estimates.

mentioned in Sec. 3.2, data belonging to two female speakers (1 and 6) and two male speakers (3 and 4) in set 2 were used. Every sentence, either in English or in Spanish, consisted of one of the carrier phrases mentioned in Sec. 2 with one of the CVb or hVd words embedded in it. The technique adopted to estimate Sg1 is as follows: First, F1 and F0 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 Sg1 estimate was computed for each voiced frame by following the procedure outlined in Sec. 3.2. Finally, a Gaussian distribution was estimated from the pool of Sg1 values obtained for voiced frames and its mean was recorded as the final Sg1 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 female speaker (27) in set 1 who is a representative of the test set. For this speaker, the highest estimation error is in the case of the vowel [aɪ] (42 Hz). The estimation error averaged over all vowels is found to be 22 Hz, which is in the observed range of standard deviations of Sg1. The average percentage estimation error over all vowels is 3.5%, which is in the observed range of percentage COVs of Sg1. It must be noted that the estimation errors are mostly uniform across all vowels. Figure 4(a) shows results of estimation in vowels for all speakers in the test set. For each speaker, Sg1 was estimated in 130 vowel tokens. The percentage number of estimates that lie within 10% of the actual value ranges between 94 and 100% and is above 95% for 17 out of 18 speakers in the test set. The percentage number of estimates that lie within 50 Hz of the actual value ranges between 90 and 100% and is above 95% for 13 out of 18 speakers. The estimation error averaged over all 130 tokens ranges between 9 and 33 Hz and is below 25 Hz for 14 out of 18 speakers.

Figure 3(b) shows results of automatic estimation in continuous speech for speaker 27. Each Gaussian density function was estimated from a different amount of data, which varied from 1 to 3 sentences. Figure 4(b) shows results of automatic estimation in continuous speech for all 22 speakers in the test set. Both Figs. 3(b) and 4(b) show that for most speakers, the estimation may improve only slightly as the

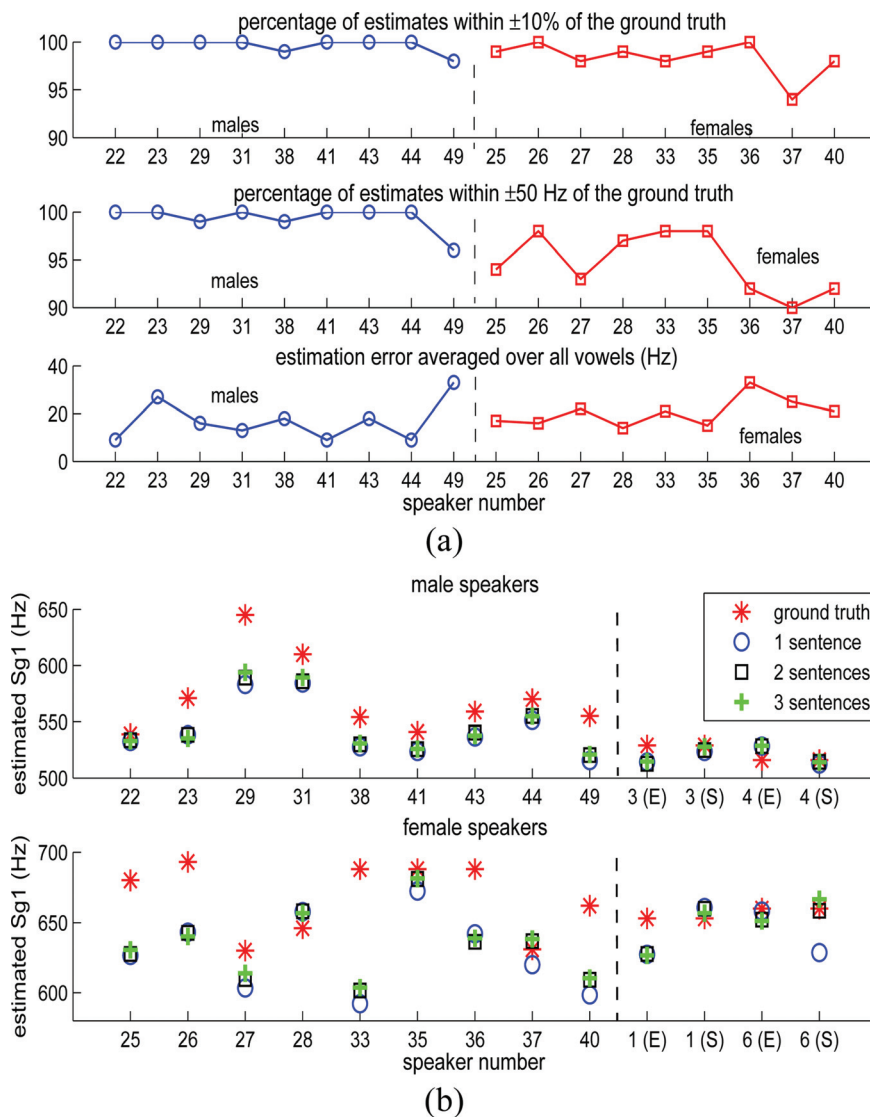


Fig. 4. (Color online) (a) Sg1 estimation in excised vowels spoken by native English speakers. In each panel, results for males/females are shown to the left/right of the dashed line. For every test speaker, Sg1 is estimated in 130 vowel tokens. The top and middle panels show the percentage of estimates which fall within 10% and within 50 Hz of the actual value, respectively. The bottom panel shows the estimation error averaged over all the 130 vowel tokens. (b) Sg1 estimation in continuous speech. Results for native English/bilingual speakers are shown to the left/right of the dashed line. The top and bottom panels show 1-sentence, 2-sentence, and 3-sentence estimation results for male and female speakers, respectively. For bilingual speakers, “E” and “S” denote estimation in English and Spanish sentences, respectively.

amount of data increases from one to three sentences. Hence for practical purposes, one short sentence of continuous speech with some voiced segments can be considered to be sufficient for estimating Sg1. The estimation error ranges from 2 Hz (speaker 6 in set 2) to 96 Hz (speaker 33 in set 1) and its average over all test speakers is 28 Hz. This is well within the range of observed standard deviations for Sg1. An important thing to note is that in case of bilingual speakers, the algorithm performs equally well with English and Spanish data. The reason is twofold. First, our approach was based on relating two acoustic measures of the feature [+low], without incorporating any

explicit information about language-specific characteristics of vowels. Second, 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 actual value was assumed to be the same for English and Spanish data because SGRs have been shown to be almost independent of the language spoken.¹

5. Conclusion

In this paper, algorithms were proposed to estimate Sg1 in adults' speech. To the best of our knowledge, this is the first attempt to estimate Sg1 using purely supraglottal acoustics. In order to develop content-independent algorithms, a novel approach based on relating two acoustic measures of the vowel feature [+low] was proposed. An empirical relation was derived between two perceptually motivated quantities—the Bark difference between F1 and F0 and the Bark difference between F1 and Sg1. The derived relation was used to develop algorithms for the automatic estimation of Sg1 in vowels and in continuous speech. It was shown that, on average, the proposed algorithms can estimate Sg1 to within 50 Hz of the actual value from voiced sounds in English and Spanish sentences. As part of our future work, we plan to use the proposed algorithms in automatic speaker normalization tasks.

Acknowledgments

The authors are thankful to John R. Morton for recording and labeling the database and to Dr. Mitchell S. Sommers for providing valuable suggestions. The authors wish to thank Gary Leung and Juan Cortes for help with the measurements. This work was supported in part by the NSF.

References and links

- ¹S. Wang, S. M. Lulich, and A. Alwan, "Automatic detection of the second subglottal resonance and its application to speaker normalization," *J. Acoust. Soc. Am.* **126**, 3268–3277 (2009).
- ²X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *J. Acoust. Soc. Am.* **122**, 1735–1745 (2007).
- ³S. Wang, A. Alwan, and S. M. Lulich, "Speaker normalization based on subglottal resonances," in *Proceedings of ICASSP* (2008), pp. 4277–4280.
- ⁴Y. 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, Cambridge, MA, 2009.
- ⁵S. 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)," *J. Acoust. Soc. Am.* **128**, 2288 (2010).
- ⁶T. G. Csapó, Z. Bărkányi, 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.
- ⁷A.K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.* **79**, 1086–1100 (1986).
- ⁸H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *J. Acoust. Soc. Am.* **88**, 97–100 (1990).
- ⁹K. Sjölander, "The Snack sound toolkit," Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, <http://www.speech.kth.se/snack/> (1997). (last accessed 9 March 2011).