Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 51–70

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation ☆

Harish Arsikere [a,*], Gary K.F. Leung [a], Steven M. Lulich [b], Abeer Alwan [a]

[a] Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA
[b] Department of Psychology, Washington University, Saint Louis, MO 63130, USA

Received 11 February 2012; received in revised form 24 May 2012; accepted 8 June 2012
Available online 4 July 2012

Abstract

Recent research has demonstrated the usefulness of subglottal resonances (SGRs) in speaker normalization. However, existing algorithms for estimating SGRs from speech signals have limited applicability—they are effective with *isolated vowels* only. This paper proposes a novel algorithm for estimating the first three SGRs ($Sg1, Sg2$ and $Sg3$) from *continuous* adults' speech. While $Sg1$ and $Sg2$ are estimated based on the phonological distinction they provide between vowel categories, $Sg3$ is estimated based on its correlation with $Sg2$. The RMS estimation errors (approximately 30, 60 and 100 Hz for $Sg1, Sg2$ and $Sg3$, respectively) are not only comparable to the standard deviations in the measurements, but also are independent of vowel content and language (English and Spanish). Since SGRs correlate with speaker height while remaining roughly constant for a given speaker (unlike vocal tract parameters), the proposed algorithm is applied to the task of height estimation using speech signals. The proposed height estimation method matches state-of-the-art algorithms in performance (mean absolute error = 5.3 cm), but uses much less training data and a much smaller feature set. Our results, with additional analysis of physiological data, suggest the existence of a limit to the accuracy of speech-based height estimation.

Keywords: Subglottal resonances; Automatic estimation; Bilingual speakers; Speaker height

1. Introduction

Recent research has shown that the subglottal resonances (SGRs) can be used effectively in automatic speaker normalization algorithms, especially when the training and testing conditions are acoustically mismatched (models trained for adults but tested on children, for example), or when the amount of speaker enrollment data is limited (Wang et al., 2008a, 2009a). It has also been shown that the second subglottal resonance ($Sg2$) can be used to adapt acoustic models trained in a particular language, say English, to a speaker whose enrollment data is in a different language, say Spanish (cross-language adaptation) (Wang et al., 2008b). By definition, SGRs are the resonance frequencies of the subglottal (below the glottis) input impedance measured from the top of the trachea. For measuring SGRs *noninvasively*, an accelerometer is commonly used (Cheyne, 2002; Chi and Sonderegger, 2007; Lulich, 2010). When held against the skin of the neck at the location of the cricoid cartilage (which is inferior to the thyroid cartilage), an accelerometer captures the pressure fluctuations at the top of the trachea, thereby yielding a frequency spectrum whose peaks occur near the SGR frequencies. However, since the use of accelerometers in many real-life situations is unfeasible, it is necessary to *estimate* SGRs from speech signals in order to use them for tasks such as automatic speaker normalization and adaptation.

☆ Parts of this article appeared in the proceedings of ICASSP 2011 and will appear in the proceedings of ICASSP 2012.

* Corresponding author. Address: Electrical Engineering Department, University of California, Los Angeles, 56-125B Engineering IV Building, Box 951594, Los Angeles, CA 90095, USA. Tel.: +1 310 729 1135.
E-mail addresses: hari.arsikere@gmail.com, harishan@ucla.edu (H. Arsikere), garyleung@ucla.edu (G.K.F. Leung), slulich@wustl.edu (S.M. Lulich), alwan@ee.ucla.edu (A. Alwan).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.06.004

Hence, motivated by the practical utility of SGRs and the need to estimate them from speech signals in real time, the present study focuses on developing an automatic algorithm that can estimate the first three SGRs ($Sg1, Sg2$ and $Sg3$) of adult speakers using as little speech data as possible. In addition, this study applies the proposed algorithm to the task of automatic speaker height estimation using speech signals. Arriving at an unknown speaker's height from speech data can be beneficial to forensics, automatic analysis of telephone calls and, possibly, automatic speaker identification.

Before trying to understand the existing SGR estimation algorithms, their limitations and the ways in which the proposed algorithm overcomes them, it is important and useful to take note of the following well-established findings. (1) The subglottal system and the supraglottal vocal tract are *acoustically* coupled through the glottis, causing each SGR to contribute a pole-zero pair to the speech signal (Stevens, 1998; Lulich, 2010). These pole-zero pairs interrupt formant trajectories of vowels (diphthongs in particular), causing *frequency discontinuities* and *amplitude attenuations* (Stevens, 1998; Chi and Sonderegger, 2007; Jung, 2009; Lulich, 2010). (2) SGRs play a role in defining vowel feature contrasts in several languages, including American English (Stevens, 1998; Lulich, 2010), High German and Swabian German (Dogil et al., 2011), Standard Korean (Jung, 2009), and Standard Hungarian (Csapó et al., 2009; Grászi et al., 2011). In particular, the first subglottal resonance acts as a boundary between [+low] and [−low] vowels, while the second subglottal resonance plays the same role with respect to [+back] and [−back] vowels. (3) SGRs are roughly constant for a given speaker, regardless of the content and the language spoken; however, SGRs do differ from speaker to speaker (Chi and Sonderegger, 2007; Madsack et al., 2008; Wang et al., 2009b; Jung, 2009; Csapó et al., 2009; Arsikere et al., 2010).

Existing literature on the relations and interactions between SGRs and formants (Stevens, 1998; Chi and Sonderegger, 2007; Lulich, 2010; Jung, 2009) suggests two possible approaches for automatically estimating SGRs from speech signals: (1) *direct* estimation based on detecting the subtle effects of SGRs on vowel formants (frequency discontinuities and amplitude attenuations), and (2) *indirect* estimation based on the potential correlations between SGRs and formant frequencies. Previous research efforts (Wang et al., 2008a,b, 2009a) have focused on the automatic estimation of $Sg2$ and $Sg3$, using a combination of both approaches.

In Wang et al. (2008a), an automatic algorithm was proposed for estimating $Sg2$ and $Sg3$ in isolated American English (AE) vowels of adults as well as children. Estimation of $Sg2$ relied on detecting a discontinuity (or jump) in the second formant frequency ($F2$) track; such a discontinuity can usually be observed in back-to-front diphthongs—[aɪ] and [ɔɪ]—when $F2$ approaches and crosses $Sg2$ (Chi and Sonderegger, 2007). Given a vowel token, $F2$ was first tracked frame-by-frame using Snack (Sjölander, 1997). Then, the track was inspected for a frequency discontinuity by computing its smoothed first-order difference and comparing it with an empirically-set threshold (in Hertz). If a discontinuity was found, $Sg2$ was estimated as the average of the $F2$ values constituting the jump. If no discontinuity was detected, $Sg2$ was estimated simply as the token's average $F2$. $Sg3$ was estimated with the help of the following empirical relation between $Sg2$ and $Sg3$, which was derived using a previously-proposed model of the subglottal airways (Lulich, 2006).

$$Sg3 = Sg2\{-0.3114[\log_{10}(Sg2) - 3.280]^2 + 1.436\} \quad (1)$$

The algorithm was evaluated indirectly by applying it to automatic speaker normalization tasks, and its performance was found to be dependent on the vowel used. Specifically, the estimation accuracy was high for diphthongs [aɪ] and [ɔɪ] but much poorer for other vowels.

The above $Sg2$ estimation algorithm was improved in Wang et al. (2008b, 2009a), but was customized to suit children's speech (unlike the above algorithm, which was applicable to adults as well as children). In Wang et al. (2008b), a *rough* $Sg2$ estimate was first obtained using the following empirical relation between $Sg2$ and the third formant frequency ($F3$) (Lulich, 2010).

$$Sg2 = 0.636 \times F3 - 103$$

Then, the estimate was refined by searching for an $F2$ jump within ± 100 Hz of the initial estimate, and computing a weighted average of the $F2$ values constituting the discontinuity, if a discontinuity was found. This procedure enabled reliable detection of the $Sg2$-induced $F2$ jump, especially in the presence of nearby, competing jumps that could be caused by other factors (e.g. inter-dental spaces) (Honda et al., 2010). In Wang et al. (2009a), the initial $Sg2$ estimate was obtained as in Wang et al. (2008b), but its refinement relied on locating not only an $F2$ jump but also an accompanying attenuation in the second formant's energy prominence, a phenomenon which has been shown to be more robust than $F2$ discontinuities in indicating subglottal coupling effects (Chi and Sonderegger, 2007). Although the improved algorithms were more reliable than the algorithm in Wang et al. (2008a), their performance was still dependent on the vowel used.

In this study, we develop a completely different approach to the SGR estimation problem because the previously-proposed algorithms suffer from the following limitations. (1) The algorithms' approaches are not suitable for estimating $Sg1$ because it can be very difficult to detect $Sg1$-induced coupling effects in trajectories of the first formant. It is important to be able to estimate $Sg1$ because, like $Sg2$ and $Sg3$, $Sg1$ could play a role in automatic speaker normalization. (2) The algorithms are not well-suited to practical applications because: (a) their performance is data dependent, and (b) they can be applied only to isolated vowels (and not continuous or natural speech). (3) Detection of subglottal coupling effects requires very accurate formant tracking procedures. As the authors of Wang

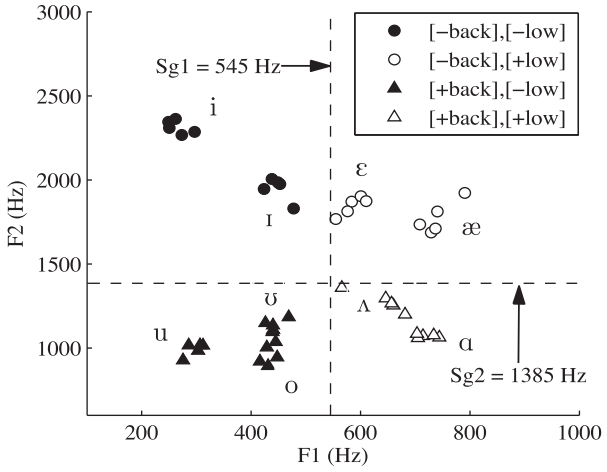

Fig. 1. Vowel space of a particular male speaker (part of this study) in the *F*1–*F*2 plane, demonstrating the vowel feature contrasts provided by *Sg*1 and *Sg*2. *F*1 (first formant) is greater than *Sg*1 for [+low] vowels (empty symbols) and less than *Sg*1 for [−low] vowels (filled symbols). Similarly, *F*2 is greater than *Sg*2 for [−back] vowels (circles) and less than *Sg*2 for [+back] vowels (triangles).

et al. (2009a) point out, "Manual verification and/or correction is applied through visually checking the tracking contours against spectrograms", which implies that their algorithms are not completely automatic.

To address these issues, we propose a novel and fully automatic algorithm that can estimate the first three SGRs from continuous speech in a content-independent and language-independent manner. However, the focus in this paper will be on adults' speech only. While existing algorithms rely on detecting the acoustic events (in speech) induced by SGRs, the proposed algorithm is based on the vowel feature contrasts provided by SGRs. Specifically, *Sg*1 estimation relies on the distinction it provides between [+low] and [−low] vowels, while *Sg*2 estimation relies on the distinction it provides between [+back] and [−back] vowels. Fig. 1 shows an example of the feature contrasts provided by *Sg*1 and *Sg*2 *F*1 (first formant frequency) is greater than *Sg*1 for [+low] vowels and less than *Sg*1 for [−low] vowels, while *F*2 is greater than *Sg*2 for [−back] vowels and less than *Sg*2 for [+back] vowels. Although there has been some research regarding the division of *tense* and *lax* [−back] vowels by *Sg*3 (Lulich, 2010; Csapó et al., 2009), no strong evidence exists either for or against it. Therefore, *Sg*3 is estimated simply by exploiting its correlation with *Sg*2.

1.1. Speaker height estimation

Previous efforts aimed at identifying height-related features of speech have been based largely on the assumption that an anatomical correlation exists between speaker height and vocal tract length (VTL). In fact, a study using magnetic resonance imaging techniques (Fitch and Giedd, 1999)—over a wide range of speaker ages and heights—provides some evidence in favor of this assumption. Motivated by a fundamental premise of speech production theory that formant frequencies are inversely proportional to VTL, several studies have analyzed the correlation between speaker height and formant frequencies (van Dommelen and Moxness, 1995; González, 2004; Rendall et al., 2005); however, no strong correlations have been reported. A few studies have also investigated the relation between height and fundamental frequency (*F*0), but have found no significant correlation between the two (González, 2004; Künzel, 1989). More recently, Dusan (2005) has reported the correlations between speaker height and commonly-used vocal tract features such as Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and linear prediction coefficients (LPCs); the study shows that 57% of the variance in height can be explained using 31 vocal tract features: the first 10 MFCCs, 16 LPCs and the first 5 formant frequencies. Unlike Fitch and Giedd (1999), the other studies mentioned above restrict their data and analyses to adult speakers; thus, they are perhaps more representative of the height estimation problem that we investigate in this study.

A few studies have proposed algorithms for automatically estimating speaker height from speech. In Pellom and Hansen (1997), speech signals were parameterized using the first 19 MFCCs, and 11 height-dependent Gaussian mixture models (GMMs) were trained using data from all speakers in the TIMIT corpus (Garofolo, 1988). The height of a given test speaker was then estimated using the maximum *a posteriori* classification rule. With this approach, the height estimation error was found to be 5 cm or less for 72% of the speakers. However, it should be noted that the *same* set of speakers was used for both training and evaluation. In Ganchev et al. (2010a), an SVM-based regression model was proposed for height estimation. The model was trained and evaluated using data from 462 and 168 speakers, respectively, in the TIMIT corpus. Training was accomplished by first extracting 6552 audio features from each utterance, and then subjecting them to a feature ranking procedure to choose the most relevant subset. The subset consisting of the top 50 features resulted in the best performance, yielding a mean absolute error (MAE) equal to 5.3 cm and a root mean squared error (RMSE) equal to 6.8 cm. The features consisted mostly of means, standard deviations, percentiles and quartiles of MFCCs, *F*0 and voicing probability. In Ganchev et al. (2010b), a similar algorithm using a Gaussian process based regression scheme was proposed to estimate speaker height in real-world indoor and outdoor scenarios, and results identical to those of Ganchev et al. (2010a) were achieved. Although the algorithms in Ganchev et al. (2010a,b) yield reasonably good results (MAE = 5.3 cm) using statistical measures of speech features, it is not clear how such features relate to speaker height.

Despite the correlation between VTL and speaker height, height estimation using vocal tract information is difficult because the configuration of the vocal tract changes significantly during speech production. Specifically, as evident

from Dusan (2005) and Ganchev et al. (2010a,b), a large number of vocal-tract features are required to capture the correlation between height and VTL. In comparison with the vocal tract, the configuration of the subglottal system of a given speaker changes little over time. This is readily exemplified by the observation that a speaker's formants vary much more than his/her SGRs (see Fig. 2). Therefore, in this paper, we propose a novel approach to height estimation based on the assumption that the 'acoustic length' of the subglottal system is proportional to speaker height. 'Acoustic length' can be defined as the length of an equivalent uniform tube (closed at one end) whose input impedance closely matches the actual input impedance of the subglottal system. The above assumption is supported by a recent study (Lulich et al., 2011) which showed that the first three SGRs can be modeled as the resonances of a simple uniform tube whose 'acoustic length' is approximately equal to the height of the speaker divided by an empirically-determined scaling factor. Hence, we attempt to capture the assumed relationship between height and 'acoustic length' by modeling the observed correlation between height and SGR frequencies (Arsikere et al., 2010; Lulich et al., 2011). In essence, the proposed approach not only has a physiological basis, but also is more efficient compared to existing techniques in terms of the number of features required for estimating height.

The rest of this paper is organized as follows. Section 2 describes the databases used in this study. Section 3 explains the procedure for analyzing accelerometer signals manually, the algorithm for automatically estimating SGRs from speech signals, and the method for estimating speaker height using SGRs estimated from speech. Experiments on SGR estimation and height estimation, followed by their results, are presented in Section 4. Section 5 sheds light on the limit as to how well speaker height might be estimated from speech signals, and also compares the methods proposed in this study with some of our previous techniques. Section 6 summarizes this paper and presents the major conclusions.

2. Databases used

The present study comprises four tasks: (1) training the SGR estimation algorithm, (2) deriving empirical relations between speaker height and SGRs (for height estimation), (3) evaluating the SGR estimation algorithm, and (4) evaluating the height estimation procedure. To accomplish the above tasks, data from the following corpora were used: the WashU-UCLA corpus (Lulich et al., 2010), the WashU-UCLA bilingual corpus, the MIT tracheal resonance database (Sonderegger, 2004; Chi and Sonderegger, 2004) and the TIMIT database (Garofolo, 1988). Following is a brief description of the relevant portions of each corpus:

- The *WashU-UCLA corpus*—used for tasks 1 to 3—consists of simultaneous microphone and subglottal accelerometer recordings from 50 adult native AE speakers (25 males, 25 females) aged between 18 and 25 years. Speakers are identified as s9, s10 and so on up to s68 (some numbers between 9 and 68 were not used). For each speaker, two word lists were recorded (in separate sessions): one with 14 AE *hVd* words ('*V*': 9 monophthongs, 4 diphthongs and the approximant [ɹ]) and the other with 21 AE *CVb* words ('*V*': 4 monophthongs and 3 diphthongs, '*C*': [b], [d] and [g]). Table 1 shows the list of vowels recorded along with the corresponding values of the features [low] and [back]. Although the vowel [ɛ] is [−low] phonologically (Stevens, 1998), we consider it to be [+low] acoustically because its average *F*1 is greater than *Sg*1 for most speakers in this study (for example, see Fig. 1). Every word, embedded in the carrier phrase, "*I said a* ___ *again*", was recorded 10 times. The *start*, *steady-state* and *end* times of the 'target' vowel were hand-labeled in each recording using Praat (Boersma and Weenink). All recordings were sampled at 48 kHz and quantized at 16 *bits/sample*. The corpus also contains self-reported speaker heights in

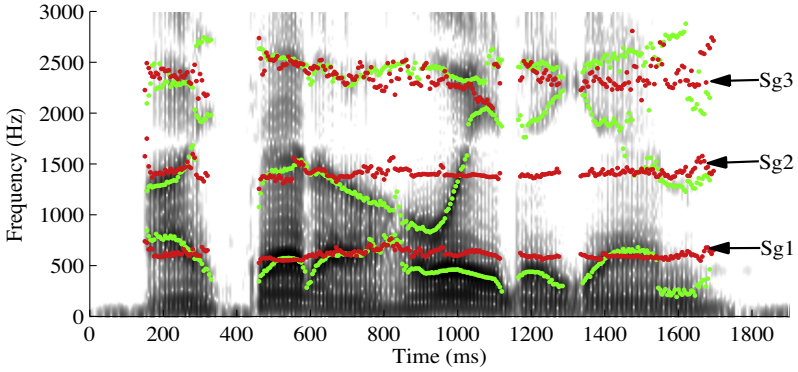

Fig. 2. Spectrogram of a speech signal (from the same speaker as in Fig. 1) superimposed with tracks of the first three vocal tract formants (green), and tracks of the first three SGRs (indicated by arrows) obtained from the corresponding accelerometer signal (red). Clearly, formants vary much more than SGRs. *Sg*3 in this case was very weak compared to *Sg*1 and *Sg*2, and hence it was tracked less accurately. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
List of vowels recorded in the WashU-UCLA corpus (native English speakers) and the WashU-UCLA bilingual corpus (native Spanish speakers). The *hVd* words in American English (AE) were recorded from native English speakers only; the AE *CVb* words were recorded from all speakers; and the *CVb* words in Mexican Spanish (MS) were recorded from bilinguals only. Note that each Spanish vowel is placed below a phonetically-similar English vowel. For monophthongs, the values of the features [low] and [back] are also indicated.

AE, *hVd* context	i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	o	ʊ	u	aɪ	aʊ	ɔɪ	ɹ
AE, *CVb* context	i			ɛ		ɑ				u	aɪ	aʊ	ɔɪ	
MS, *CVb* context	i			e		a				u	ai	au	oi	
Feature [low]	−	−		+	+	+	+	-	-	-				
Feature [back]	−	−		−	−	+	+	+	+	+				

centimeters. In this study, only recordings containing the *hVd* words were used. Further details of the corpus can be found in Lulich et al. (2011).

- The *WashU-UCLA bilingual corpus*—used for tasks 2 and 3—consists of simultaneous microphone and accelerometer recordings from 6 adult bilingual speakers (4 males, 2 females) of Mexican Spanish (MS)—their native language—and AE. All speakers were aged between 20 and 25 years. Speaker numbers range from s1 to s6. Each speaker was recorded in two sessions; one involved recording 21 AE *CVb* words (identical to the WashU-UCLA corpus) while the other involved recording 21 MS *CVb* words ('*V*': 4 monophthongs and 3 diphthongs, '*C*': [b], [d] and [g]). The list of vowels can be found in Table 1. Spanish words were embedded in the carrier phrase, "*Dije una ___ otra vez*" (meaning "*I said a ___ again*"). Each word (AE and MS) was recorded 7 times. All other procedures (labeling, height recording) and settings (sampling rate, bit resolution) were identical to those of the WashU-UCLA corpus. In this study, both AE and MS recordings were used.
- The *MIT tracheal resonance database*—used for task 3—consists of microphone and accelerometer recordings from 14 adults (7 males, 7 females) aged between 18 and 78 years. Males and females are numbered from M1 to M7 and F1 to F7, respectively. Out of the 14 subjects, 11 were native AE speakers, while M1, M7 and F7 were native speakers of Canadian English, British English and Mandarin, respectively. For each speaker, up to 16 *dVd* and 16 *hVd* words were recorded 5 times each by embedding them in the carrier phrase, " ___, *say* ___ *again*". Details of the recorded material can be found in Sonderegger (2004). All signals were low-pass filtered to 4.8 kHz, sampled at 10 kHz and quantized at 16 *bits/sample*.
- The *TIMIT database*—used for task 4—contains a total of 6300 AE sentences, 10 sentences spoken by each of 630 speakers (438 males, 192 females) from 8 major dialect regions of the United States. All signals were sampled at 16 kHz and quantized at 16 *bits/sample*. The database also contains speaker heights in feet and inches. In this study, data from only 604 speakers (431 males, 173 females) were used (as will be explained below). Further details of the corpus can be found in Garofolo (1988).

3. Methods

This section explains our methods for (1) manually analyzing accelerometer signals to obtain SGR measurements, (2) developing and training the automatic SGR estimation algorithm and (3) developing the automatic height estimation procedure based on the SGR estimation algorithm.

3.1. Manual analysis of accelerometer signals

For each speaker in the WashU-UCLA corpus, the WashU-UCLA bilingual corpus and the MIT tracheal resonance database, several measurements of the first three SGRs were made by manually analyzing their accelerometer signals. This was necessary to obtain the *actual* SGR frequencies (or *'ground truth'* values) of each speaker. The terms *actual SGRs* and *'ground truth' SGRs* will henceforth be used interchangeably. We first describe our procedure for measuring SGRs in a given accelerometer signal and then explain how the 'ground truth' SGR values were calculated for each speaker.

All SGR measurements were made from accelerometer signals of vowel tokens. Given a token of sufficient quality, the first three SGRs were measured as follows. (1) The token was down sampled to between 6 and 10 kHz depending on how noisy the signal was at high frequencies, and then passed through a standard pre-emphasis filter with a pre-emphasis coefficient of 0.97. (2) A segment, approximately four pitch periods in length, was chosen from the steady-state portion of the token. In case of the two WashU-UCLA corpora, the choice of the steady-state segment was guided by the Praat labels. However, in case of the MIT tracheal resonance database, the segment was chosen just by visually inspecting the spectrogram. (3) The discrete Fourier transform (DFT) spectrum, the LPC spectrum and the estimated wideband power spectral density (WPSD) (Nelson, 1997) of the chosen segment were computed, and the prominent peaks in the LPC spectrum and WPSD were identified using a simple peak-picking algorithm. The LPC order was varied between 12 and 16 until the envelope underlying the DFT spectrum was represented satisfactorily. The WPSD can be treated qualitatively as a smoothed envelope of the DFT spectrum. It was obtained, according to the approach outlined in Umesh et al. (1999), by subdividing the vowel segment into

several overlapping frames, calculating an autocorrelation function for each frame after applying a Hamming window, and computing the DFT of the averaged autocorrelation function. The overlap between successive frames was fixed at 80% of the frame size, and the frame size itself was varied between 0.9 and 1.1 times the pitch period such that the resulting envelope was visually of the best possible quality. Fig. 3 shows the three spectral representations of a sample accelerometer segment from speaker s12 in the WashU-UCLA corpus. (4) Each SGR was measured by choosing either the LPC peak or the WPSD peak depending on which of the two provided a more accurate representation of the envelope. If neither spectral representation was satisfactory for a particular SGR, the SGR frequency was not measured. Hence, it is important to note that not all three SGRs were necessarily measured in every vowel token chosen for analysis. In general, it was more difficult to measure $Sg1$ and $Sg3$ than to measure $Sg2$. While the measurement of $Sg1$ was sometimes difficult (especially for high-pitched speakers) due to its proximity to strong low-frequency harmonics (for example, see Fig. 3), the measurement of $Sg3$ was always difficult owing to the attenuation of high frequencies caused by the low-pass nature of the neck tissues and skin.

To calculate the 'ground truth' SGR values, we first obtained several SGR measurements for each speaker using accelerometer signals of vowel tokens. For speakers in the WashU-UCLA corpus and the MIT tracheal resonance database, monophthong vowels as well as the approximant [ɹ] in the *hVd* word list were analyzed. For speakers in the WashU-UCLA bilingual corpus, monophthongs in both the AE *CVb* word list and the MS *CVb* word list were analyzed. It must be noted that the tokens chosen for analysis were not distributed equally across vowels. Table 2 shows the minimum, maximum and average number of $Sg1, Sg2$ and $Sg3$ measurements obtained per speaker (for each database analyzed). $Sg3$ could not be measured for one speaker (speaker s18 in the WashU-UCLA corpus). For the bilingual speakers, roughly equal numbers of measurements were obtained from AE and MS vowels. Once the measurements were obtained, we verified if SGRs of a given speaker are invariant to spoken content and language (shown previously for children, in Wang et al. (2009b)) by calculating: (1) the within-speaker coefficient of variation (COV)—defined as the ratio of standard deviation to mean—of $Sg1, Sg2$ and $Sg3$ for each speaker in the WashU-UCLA corpus and the MIT tracheal resonance database and (2) the within-speaker, cross-language COV (measurements from AE and MS vowels combined) of $Sg1, Sg2$ and $Sg3$ for each speaker in the WashU-UCLA bilingual corpus. Table 3 shows the average percentage COVs and the corresponding average standard deviations for each database. COVs of the order of 2–5% indicated that the measurements varied very little about their mean values, thus confirming that SGRs are indeed 'constant' for a given speaker. Hence, the 'ground truth' $Sg1, Sg2$ and $Sg3$ of a given speaker were taken to be the averages of the corresponding measurements. Table 4 shows the minimum, maximum and average values of the 'ground truth' SGRs for each database analyzed. In the rest of this paper, only the 'ground truth' SGRs will be used (and not the measurements in individual tokens).

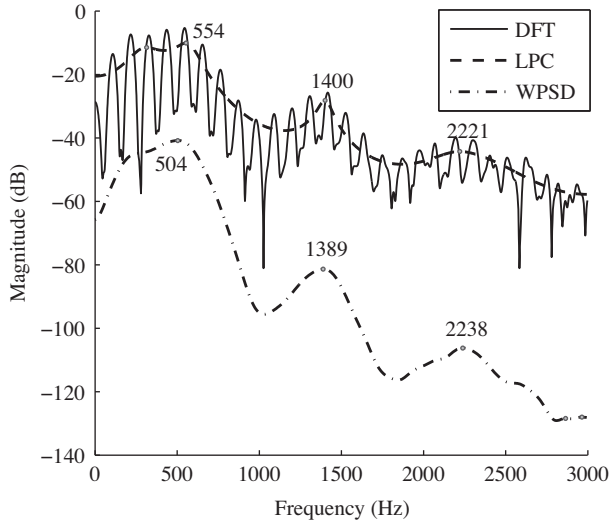

Fig. 3. DFT, LPC and WPSD spectra obtained using the steady-state portion of an accelerometer signal (vowel [i]) recorded from speaker s12 in the WashU-UCLA corpus. The numbers near the LPC and WPSD spectral peaks indicate candidates for the measured values of $Sg1, Sg2$ and $Sg3$.

3.2. Automatic estimation of SGRs from speech signals

Here, we focus on the first major goal of this study: automatic estimation of the first three SGRs from speech signals. Data from 30 speakers (15 males, 15 females) in the WashU-UCLA corpus were used to develop and train the SGR estimation algorithm. The training speakers were chosen such that their actual SGR frequencies were uniformly distributed in the ranges of 'ground truth' values shown in Table 4. We first describe our approaches to the estimation of $Sg1, Sg2$ and $Sg3$, and then explain the algorithm that estimates them automatically from continuous speech.

3.2.1. Estimation of Sg1

Previous studies have shown that $Sg1$ lies at the boundary of [+low] and [−low] vowels along the $F1$ dimension (see Section 1). Motivated by this finding, $Sg1$ estimation relied on three main ideas: (1) defining a vocal tract-based measure of vowel *height* that can be computed using speech signals, (2) defining an $Sg1$-based measure of vowel height that can be computed using speech and subglottal (accelerometer) signals, and (3) developing a model to predict the $Sg1$-based measure from the vocal tract-based measure. In Syrdal and Gopal (1986), the Bark difference between $F1$ and $F0$ was shown to be a reliable indicator of vowel height. Based on this finding, an $Sg1$ estimation algorithm was proposed in Arsikere et al. (2011a), in which the Bark

Table 2
Minimum, maximum and average number of $Sg1, Sg2$ and $Sg3$ measurements across speakers. For bilingual speakers, roughly equal numbers of measurements were obtained from AE and MS vowels (combined numbers are shown here).

	$Sg1$			$Sg2$			$Sg3$		
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
WashU-UCLA corpus	3	30	**15**	8	30	**15**	0	19	**7**
WashU-UCLA bilingual corpus	15	22	**18**	16	27	**21**	3	22	**13**
MIT tracheal resonance database	6	20	**12**	9	19	**12**	1	8	**6**

Table 3
Average within-speaker percentage coefficient of variation (COV) and standard deviation (σ) of $Sg1, Sg2$ and $Sg3$. For bilingual speakers, measurements from both AE and MS data were combined (cross-language COVs and standard deviations).

	$Sg1$		$Sg2$		$Sg3$	
	%COV	σ (Hz)	%COV	σ (Hz)	%COV	σ (Hz)
WashU-UCLA corpus	5.0	30	2.2	32	2.5	57
WashU-UCLA bilingual corpus	4.5	25	2.3	32	2.7	61
MIT tracheal resonance database	3.7	22	2.1	30	2.2	49

Table 4
Minimum, maximum and average values of the 'ground truth' SGRs of speakers in the WashU-UCLA corpus, the WashU-UCLA bilingual corpus and the MIT tracheal resonance database. The values for males and females are significantly different.

	$Sg1$ (Hz)			$Sg2$ (Hz)			$Sg3$ (Hz)		
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
WashU-UCLA corpus									
Males	492	622	**542**	1217	1492	**1327**	2039	2449	**2198**
Females	580	722	**659**	1382	1610	**1511**	2273	2575	**2410**
WashU-UCLA bilingual corpus									
Males	491	556	**533**	1198	1405	**1314**	1931	2343	**2160**
Females	626	658	**642**	1493	1513	**1503**	2420	2505	**2462**
MIT tracheal resonance database									
Males	515	567	**534**	1289	1382	**1347**	2166	2344	**2230**
Females	575	681	**642**	1373	1587	**1507**	2141	2550	**2424**

difference between $F1$ and $F0$ was used to predict an $Sg1$-based measure of vowel height: the Bark difference between $F1$ and $Sg1$, denoted henceforth as $B_{1,s1}$. However, a vocal tract-based measure involving $F1$ and $F0$ may be problematic because of two reasons: 1) $F1$ and $F0$ can be controlled fairly independently of each other and 2) reliable estimation of $F1$ and $F0$ can be difficult when they are very close to each other (e.g., [-low] vowels produced by high-pitched speakers). Therefore, in this study, the Bark difference between $F3$ and $F1$, denoted henceforth as B_{31}, was used as the required vocal tract-based measure of vowel height. B_{31} was chosen because a similar acoustic feature, namely the Bark difference between $F3$ and $F2$, denoted henceforth as B_{32}, has been shown to be a reliable indicator of vowel *backness* (Syrdal and Gopal, 1986; Chistovich, 1985). As in Arsikere et al. (2011a), $B_{1,s1}$ was used as the required $Sg1$-based measure of vowel height.

In order to develop a model for predicting $B_{1,s1}$ from B_{31}, we first obtained formant frequency measurements from microphone recordings of all 30 speakers in the training set. For each speaker, the first three formants were measured in the steady-state regions of 5 tokens each of 9 monophthongs (in the hVd word list); hence, a total of 1350 vowel tokens were analyzed. Wavesurfer (Sjölander and Beskow, 2000), a speech analysis tool designed using Snack, was used to obtain formant measurements semi-automatically; formant tracking parameters were manually adjusted until the tracking contours aligned satisfactorily with the spectrograms. Once all the measurements were obtained, the formant frequencies as well as the actual SGR frequencies of the training speakers were converted to Bark values using Eq. (2) (Traunmüller, 1990):

$$z = [(26.81f)/(1960 + f)] - 0.53, \quad (2)$$

where z is the Bark value corresponding to a frequency f in Hertz. Several definitions of the Bark scale exist (see references in Traunmüller (1990)), but the one given by Eq. (2) offers simplicity in terms of conversion between Hertz and Bark values, in addition to being as accurate as the other definitions. Finally, for each of the 1350 tokens analyzed, the values of B_{31} and $B_{1,s1}$ were computed. Fig. 4 shows normalized histograms of B_{31} and $B_{1,s1}$ for [−low] and [+low] vowels, and also a scatter plot of $B_{1,s1}$ versus B_{31}. B_{31} separates the two vowel categories at approximately

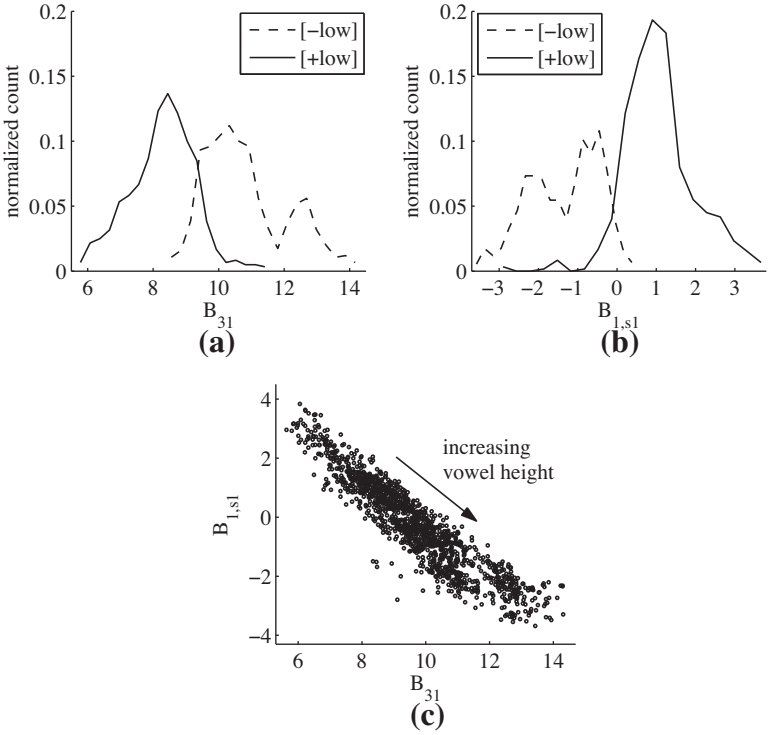

Fig. 4. Two measures of vowel height—B_{31} (vocal tract based) and $B_{1,s1}$ ($Sg1$ based). (a) Normalized histograms of B_{31} for [−low] and [+low] vowels; the boundary between the two classes is roughly at 9.5 Bark. (b) Normalized histograms of $B_{1,s1}$ for [−low] and [+low] vowels; the boundary between the two classes is roughly at 0 Bark. (c) Scatter plot (1350 data points) of $B_{1,s1}$ versus B_{31} showing that they are strongly correlated ($r = -0.9241$).

9.5 Bark (Fig. 4(a)) and $B_{1,s1}$ provides a boundary at approximately 0 Bark (Fig. 4(b)), confirming that B_{31} and $B_{1,s1}$ are indeed reliable measures of vowel height. More importantly, as evident from Fig. 4(c), the two measures are strongly correlated ($r = -0.9241$), suggesting that B_{31} provides most of the information required for predicting $B_{1,s1}$.

A linear regression (using data from all 1350 tokens) between $B_{1,s1}$ (dependent variable) and the first three powers of B_{31} (independent variables) resulted in the following model:

$$B_{1,s1} = 0.011(B_{31})^3 - 0.269(B_{31})^2 + 1.322(B_{31}) + 2.455.$$

Although this regression model had a reasonably high r-squared (r^2) value (0.8702), a non-negligible portion of the variance in $B_{1,s1}$ (13%) was still not accounted for. The residual variance was observed to be due to individual speaker differences. Specifically, when the regression was performed separately for each speaker in the training set, the resulting model coefficients showed large spreads in their values: the coefficients related to the linear term (B_{31}) and the intercept term—terms with the two largest weights—were found to have COVs equal to 115% and 162%, respectively. To reduce the inter-speaker variability involved in predicting $B_{1,s1}$, two speaker-related features were used: $F3$ and $F0$ (both in Hertz). Steady-state $F0$ values of all 1350 vowel tokens used for modeling were obtained automatically using Snack; the ESPS pitch tracking algorithm was used with a frame length of 30 ms and a frame spacing of 5 ms. When $F3$ and $F0$ (in that order) were added incrementally to the above third-order regression model, r^2 increased from 0.8702 to 0.9255 and from 0.9255 to 0.9724; the increase in each case was statistically significant ($p < 0.001$). The updated regression model, which predicts $B_{1,s1}$ using B_{31}, $F3$ and $F0$, is:

$$B_{1,s1} = 0.001(B_{31})^3 - 0.024(B_{31})^2 - 0.737(B_{31}) + 0.002(F3) - 0.007(F0) + 3.903. \quad (3)$$

With the updated model, the COVs of the coefficients related to the linear term and the intercept term were equal to 44% and 49%, respectively. Thus, it can be said that $F3$ and $F0$ were successful in reducing inter-speaker variability. Given $F0, F1$ and $F3, Sg1$ can be readily estimated using Eq. (3).

3.2.2. Estimation of Sg2

Since $Sg2$ acts as a boundary between [+back] and [−back] vowels along the $F2$ dimension (see Section 1), its estimation, like the estimation of $Sg1$, relied on three main ideas: (1) defining a vocal tract-based measure of vowel backness, (2) defining an $Sg2$-based measure of vowel backness, and (3) developing a model to predict the $Sg2$-based measure from the vocal tract-based measure. While B_{32} was used as the required vocal tract-based measure (based on the findings in Syrdal and Gopal (1986) and Chistovich (1985)), the Bark difference between $F2$ and

$Sg2$, denoted henceforth as $B_{2,s2}$, was used as the required $Sg2$-based measure.

In order to develop a model for predicting $B_{2,s2}$, we computed B_{32} and $B_{2,s2}$ for the 1350 vowel tokens that were used previously for developing the $Sg1$-estimation model. Fig. 5 shows normalized histograms of B_{32} and $B_{2,s2}$ for [−back] and [+back] vowels, and also a scatter plot of $B_{2,s2}$ versus B_{32}. B_{32} separates the two vowel categories at approximately 3.5 Bark (Fig. 5(a))—which agrees well with the findings in Syrdal and Gopal (1986) and Chistovich (1985)—and $B_{2,s2}$ provides a boundary at approximately 1 Bark (Fig. 5(b)), confirming that B_{32} and $B_{2,s2}$ are reliable measures of vowel backness. More importantly, as evident from Fig. 5(c), the two measures are strongly correlated ($r = -0.9352$), suggesting that B_{32} provides most of the information required for predicting $B_{2,s2}$.

A linear regression between $B_{2,s2}$ and the first three powers of B_{32} resulted in the following model ($r^2 = 0.8905$):

$$B_{2,s2} = -0.004(B_{32})^3 + 0.134(B_{32})^2 - 1.958(B_{32}) + 6.182.$$

In Arsikere et al. (2011b), $Sg2$ was estimated using a similar regression model derived from a smaller training set consisting of 11 speakers in the WashU-UCLA corpus. As in the case of $Sg1$ estimation, the residual variance in the above model (11%) was minimized by using $F3$ and $F0$. When $F3$ and $F0$ (in that order) were added incrementally to the regression, r^2 increased from 0.8905 to 0.9429 and from 0.9429 to 0.9713; the increase in each case was statistically significant ($p < 0.001$). The updated regression model, which predicts $B_{2,s2}$ using $B_{32}, F3$ and $F0$, is:

$$B_{2,s2} = 0.001(B_{32})^3 + 0.009(B_{32})^2 - 1.089(B_{32}) + 0.002(F3) - 0.007(F0) - 0.019. \quad (4)$$

Given $F0, F2$ and $F3, Sg2$ can be readily estimated using Eq. (4).

3.2.3. Estimation of Sg3

Although there has been some research indicating that $Sg3$ may lie at the boundary of *tense* and *lax* [−back] vowels (Lulich, 2010; Csapó et al., 2009), there is not enough evidence to suggest that the phenomenon occurs consistently in all speakers and languages. Therefore, $Sg3$ is estimated based on its potential correlation with the other two SGRs. Using the actual SGR frequencies of the 30 speakers in the training set, $Sg3$ was found to be moderately correlated with $Sg1$ ($r = 0.7712$) but strongly correlated with $Sg2$ ($r = 0.9180$). This is evident from the scatter plots shown in Fig. 6. A first-order linear regression between $Sg3$ and $Sg2$ resulted in Eq. (5) ($r^2 = 0.8427$):

$$Sg3 = 1.079 \times Sg2 + 763.676, \quad (5)$$

which was used to estimate $Sg3$ once an estimate of $Sg2$ was obtained using the approach outlined in Section 3.2.2. For our training set (30 data points), the RMS error between actual $Sg3$ and the $Sg3$ predicted using Eq. (5) (53 Hz), was much smaller than the corresponding RMS error

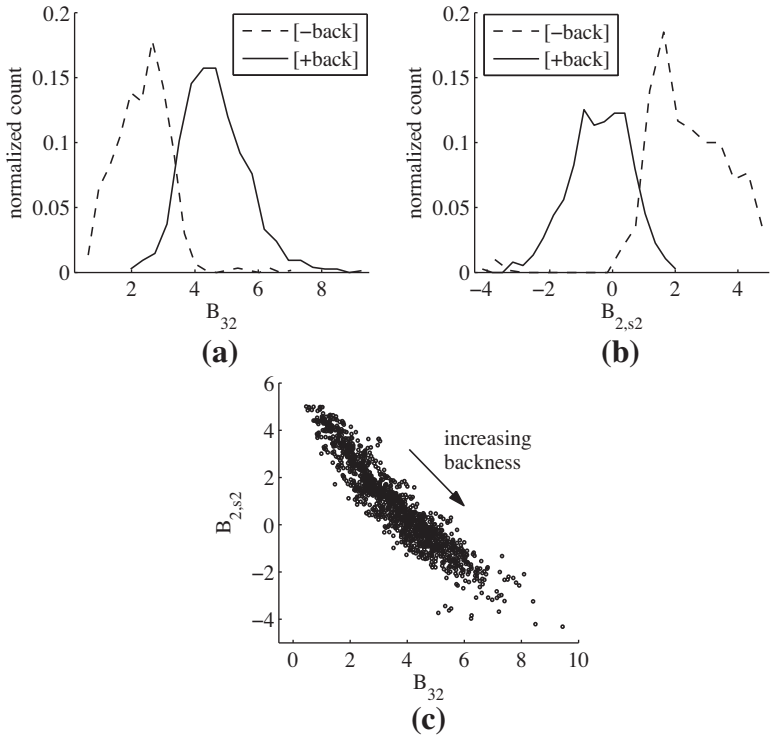

Fig. 5. Two measures of vowel backness—B_{32} (vocal tract based) and $B_{2,s2}$ ($Sg2$ based). (a) Normalized histograms of B_{32} for [−back] and [+back] vowels; the boundary between the two classes is roughly at 3.5 Bark. (b) Normalized histograms of $B_{2,s2}$ for [−back] and [+back] vowels; the boundary between the two classes is roughly at 1 Bark. (c) Scatter plot (1350 data points) of $B_{2,s2}$ versus B_{32} showing that they are strongly correlated ($r = -0.9352$).

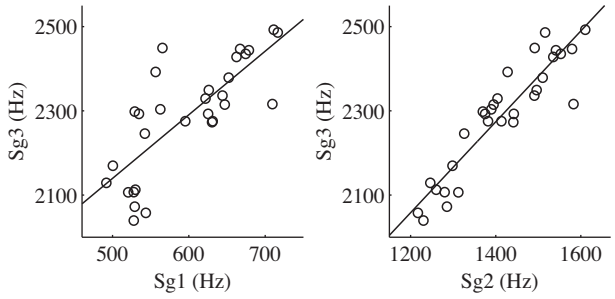

Fig. 6. Scatter plots (30 data points each) of $Sg3$ versus $Sg1$ (left) and $Sg3$ versus $Sg2$ (right). The solid lines represent first-order linear regression fits to the data. $Sg3$ was correlated moderately with $Sg1$ ($r = 0.7712$) but strongly with $Sg2$ ($r = 0.9180$).

Table 5
Snack parameters for automatic formant tracking and pitch tracking (required by the SGR estimation algorithm).

Parameter	Value
Window size	30 ms
Window spacing	5 ms
Window type	Hamming
LPC order	10
LPC method	Autocovariance
$F0$ tracking algorithm	ESPS
minimum pitch	60 Hz
maximum pitch	400 Hz

incurred using Eq. (1) (275 *Hz*). Therefore, Eq. (5) is more reliable than Eq. (1) for estimating $Sg3$ from $Sg2$. Next, we present the automatic SGR estimation algorithm that incorporates the ideas described so far.

3.2.4. The automatic algorithm

In this study, our goal was to estimate SGRs from continuous speech. Since formant frequencies and $F0$ vary considerably over time, the automatic SGR estimation algorithm warranted a frame-based approach. We now explain the steps involved in going from a given speech signal to the estimates of the speaker's SGRs.

1. Downsample the signal to 7 kHz and pre-emphasize the high frequencies by passing it through a filter with the following frequency response.

   $$H(\omega) = 1 - 0.96e^{-j\omega}$$

   Since the first three formants of adult speakers usually lie below 3.5 kHz (Peterson and Barney, 1952; Hillenbrand et al., 1995), a sampling rate of 7 kHz suffices for formant tracking.
2. Track $F0, F1, F2$ and $F3$ automatically using the Snack sound toolkit. The values of the formant tracking parameters and pitch tracking parameters are shown in Table 5. The chosen window size (30 ms) covers at least 2 to 3 pitch periods and the small window spacing (5 ms) ensures smooth formant tracks. The minimum (60 Hz) and maximum (400 Hz) pitch values accommodate the range of pitch frequencies observed in adults' speech.
3. Select all *voiced* frames using Snack's binary voicing parameter: the probability of voicing (PV). Snack sets PV to 1 or 0 depending on whether a given frame is voiced or unvoiced, respectively. Unvoiced frames need to be discarded because the fundamental frequency and formant frequencies (required for SGR estimation) are not well defined for unvoiced speech.
4. Perform the following sequence of operations for each voiced frame in the given speech signal. The superscript k in all the following operations indicates the kth voiced frame.
   - Obtain Bark values corresponding to $F1^k, F2^k$ and $F3^k$ using Eq. (2).
   - Compute B^k_{31} and B^k_{32}.
   - Predict $B^k_{1,s1}$ from $\{B^k_{31}, F3^k, F0^k\}$ using Eq. (3), and $B^k_{2,s2}$ from $\{B^k_{32}, F3^k, F0^k\}$ using Eq. (4).
   - Recover $Sg1^k$ and $Sg2^k$ in Bark:

     $$Sg1^k(\text{Bark}) = F1^k - B^k_{1,s1},$$
     $$Sg2^k(\text{Bark}) = F2^k - B^k_{2,s2}.$$

   - Convert $Sg1^k$ and $Sg2^k$ from Bark to Hertz by inverting Eq. (2).
5. At the end of Step 4, every voiced frame in the signal is associated with an estimate each of $Sg1$ and $Sg2$. Then, estimate the speaker's $Sg1$ and $Sg2$ as the averages of the corresponding frame-level estimates:

   $$Sg1 = \frac{1}{N_v}\sum_{k=1}^{N_v} Sg1^k,$$
   $$Sg2 = \frac{1}{N_v}\sum_{k=1}^{N_v} Sg2^k,$$

   where N_v denotes the total number of voiced frames.
6. Estimate the speaker's $Sg3$ by plugging the above $Sg2$ estimate into Eq. (5).

Steps 1 to 6 are summarized in Fig. 7. It must be noted that while the regression models for SGR estimation were trained using formant frequencies and $F0$ measured in steady-state vowels, the actual algorithm was designed to use all voiced frames irrespective of their origin: vowels (steady-state or otherwise), voiced consonants or transition regions between voiced and unvoiced sounds. Although such an approach is bound to yield a few 'undesirable' frame-level estimates, natural speech contains enough vowel segments to skew the *averages* of frame-level estimates towards the actual ('desired') SGR values. Fig. 8 illustrates with an example that the proposed frame-based approach is effective in estimating $Sg1$ and $Sg2$ from continuous speech. However, it is important to observe that

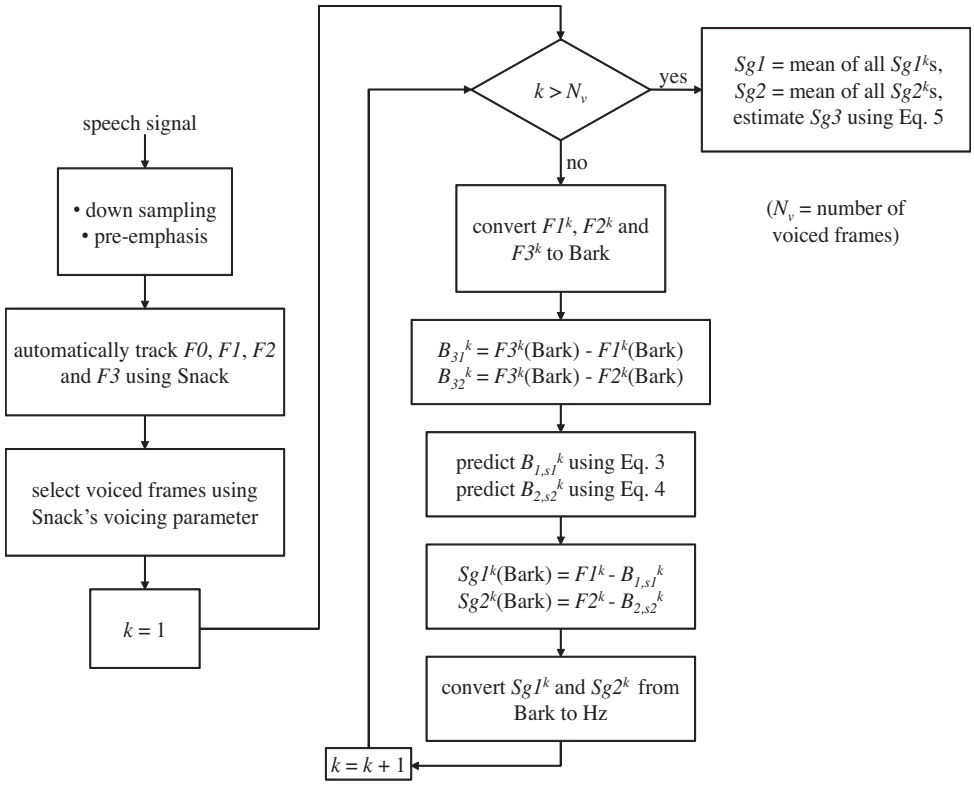

Fig. 7. Flowchart illustrating the steps involved in estimating $Sg1, Sg2$ and $Sg3$ from a given speech signal.

the proposed algorithm cannot estimate SGRs from purely unvoiced speech (e.g., whispered speech).

3.3. Automatic speaker height estimation

Our method for estimating speaker height from speech signals was motivated by the previously-observed correlation between SGRs and height (Arsikere et al., 2010; Lulich et al., 2011). As mentioned in Section 1.1, our approach is physiologically motivated because the correlation between height and SGRs is believed to be the result of an underlying correlation between height and 'acoustic length' (length of an equivalent uniform tube whose input impedance matches that of the subglottal system).

For developing models that predict height from SGR frequencies, we used the 'ground truth' SGRs and self-reported heights (in centimeters) of all speakers in the WashU-UCLA corpus and the WashU-UCLA bilingual corpus (56 in total). Male speaker heights ranged from 165 to 201 cm while female speaker heights ranged from 152 to 175 cm. All three SGRs correlated negatively with height (see Fig. 9), but $Sg2$ accounted for more height variance ($r^2 = 0.68$) than $Sg1$ ($r^2 = 0.58$) or $Sg3$ ($r^2 = 0.58$). In contrast, Dusan (2005) reported that 31 vocal tract-based

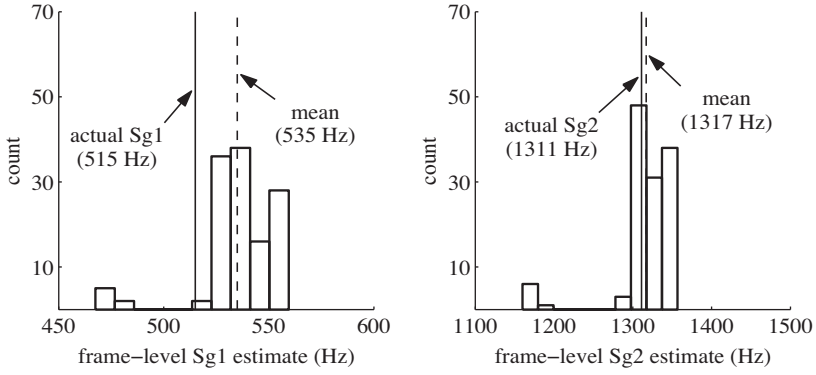

Fig. 8. Distributions of frame-level $Sg1$ estimates (left) and frame-level $Sg2$ estimates (right) obtained by applying the automatic SGR estimation algorithm to a microphone recording of "I said a heed again" from speaker s22 (not in the training set) in the WashU-UCLA corpus. In each case, the mean of the distribution is close to the actual SGR value.

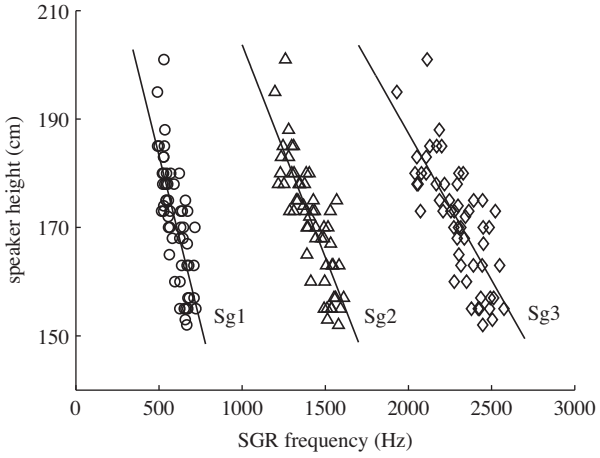

Fig. 9. Scatter plots of speaker height versus the first three SGRs (56 data points each for $Sg1$ and $Sg2$, 55 data points for $Sg3$). The solid lines represent first-order linear regression fits to the data. Speaker height correlates more strongly with $Sg2$ ($r = -0.8256$) than with $Sg1$ ($r = -0.7586$) or $Sg3$ ($r = -0.7627$).

features (MFCCs, LPCs and formants) were necessary to explain 57% of the variance in height. SGRs are therefore more suitable for height estimation than vocal tract features.

Using first-order linear regression, the following empirical relations were obtained between speaker height and SGR frequencies:

$$h = -0.124 \times Sg1 + 245.476, \tag{6}$$

$$h = -0.078 \times Sg2 + 282.107, \tag{7}$$

$$h = -0.054 \times Sg3 + 295.659, \tag{8}$$

where h denotes speaker height (in centimeters). Given a speech signal, speaker height was estimated by first estimating $Sg1, Sg2$ and $Sg3$ using the proposed SGR estimation algorithm, and then using Eqs. (6)–(8), respectively. Although speaker height correlated most strongly with $Sg2$, all the above equations were considered for height estimation because our method was affected not only by the correlations between height and SGRs, but also by the accuracy of SGR estimation.

4. Experiments and results

4.1. Automatic estimation of SGRs

The proposed automatic SGR estimation algorithm was evaluated using microphone recordings from 20 speakers (10 males, 10 females) in the WashU-UCLA corpus (different from the training set), all 6 speakers in the WashU-UCLA bilingual corpus and all 14 speakers in the MIT tracheal resonance database. Two sets of experiments were performed. (1) In order to assess its performance with regard to the content spoken, the algorithm was applied to isolated vowel tokens obtained (using Praat labels) from the two WashU-UCLA corpora. (2) In order to assess its ability to estimate SGRs from continuous or natural speech, the algorithm was applied to complete sentence recordings (carrier phrases) of all three corpora. Both sets of experiments were useful in analyzing the algorithm's performance with regard to language, since the same algorithm was applied to AE as well as MS data. In addition, since the MIT tracheal resonance database and the two WashU-UCLA corpora were recorded using different equipment, our experiments were helpful in analyzing the algorithm's reliability under varying recording conditions.

For ease of representation, let us denote actual SGR values as $Sg1_a, Sg2_a$ and $Sg3_a$, and estimated SGR values as $Sg1_e, Sg2_e$ and $Sg3_e$. The SGR estimation algorithm was evaluated using two performance metrics: (1) average root mean squared error (RMSE_{avg}) and (2) average mean-relative standard deviation (MSD_{avg}). While RMSE_{avg} was used to quantify estimation *accuracy*, MSD_{avg} was used to quantify the *consistency* of estimation. Denoting the number of test speakers as N_s and the number of test utterances (isolated vowels or sentences) for the ith speaker as M_i, the definitions of RMSE_{avg} and MSD_{avg} for the Kth SGR ($K = 1, 2, 3$) are as follows.

$$\text{RMSE}_{avg} = \frac{1}{N_s}\sum_{i=1}^{N_s}\text{RMSE}^i, \qquad \text{RMSE}^i = \sqrt{\frac{1}{M_i}\sum_{j=1}^{M_i}(SgK_e^{ij} - SgK_a^i)^2} \tag{9}$$

$$\text{MSD}_{avg} = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\frac{\sigma_e^i}{\mu_e^i} \times 100\right), \qquad \mu_e^i = \frac{1}{M_i}\sum_{j=1}^{M_i}SgK_e^{ij}, \qquad \sigma_e^i = \sqrt{\frac{1}{M_i}\sum_{j=1}^{M_i}(SgK_e^{ij} - \mu_e^i)^2}. \tag{10}$$

In Eq. (9), SgK_a^i denotes the actual value of the Kth SGR of the ith test speaker. In Eqs. (9) and (10), SgK_e^{ij} denotes the estimated value of the Kth SGR corresponding to the jth utterance of the ith test speaker. It must be noted that the definition of RMSE_{avg} is similar to that of the average within-speaker standard deviations of SGR measurements (reported in Table 3). Likewise, the definition of MSD_{avg} resembles that of the average within-speaker percentage COVs of SGR measurements (reported in Table 3). Therefore, the values in Table 3 will be used as a rough guideline for interpreting the results of the SGR estimation algorithm.

4.1.1. Estimation using isolated vowels

The algorithm was evaluated on: (1) 13 AE vowels in the *hVd* word list of the WashU-UCLA corpus (the approximant [ɹ] was not used) and (2) all 7 vowels in the AE *CVb* word list and the MS *CVb* word list of the WashU-UCLA bilingual corpus. Each speaker in the

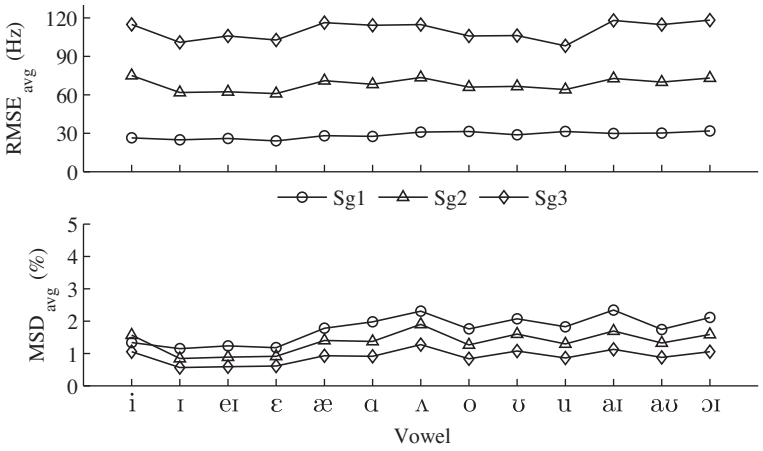

Fig. 10. SGR estimation using isolated vowels: overall $RMSE_{avg}$ and MSD_{avg} corresponding to the monophthongs and diphthongs recorded in the WashU-UCLA corpus. For practical purposes, the performance can be considered to be vowel independent.

WashU-UCLA corpus had 10 tokens per vowel, while each speaker in the WashU-UCLA bilingual corpus had 21 tokens per vowel.

Fig. 10 shows the overall (males and females combined) $RMSE_{avg}$ and MSD_{avg} corresponding to all the monophthongs and diphthongs in the WashU-UCLA corpus. The following two observations can be made. (1) The algorithm's performance is slightly vowel dependent; this might be attributed, at least in part, to differences in the accuracy of automatic formant tracking. Specifically, it is easier to track formants when they are fairly 'steady' and well separated from one another (e.g., [ɛ], [æ] and [I]) than when two or more of them are very closely spaced (e.g., [i] and [a]) or rapidly changing over time (e.g., [I] and [ɔI]). Nevertheless, the observed vowel dependence in performance is small enough to be ignored for practical purposes: $RMSE_{avg}$ ranges from 24 Hz ([ɛ]) to 32 Hz ([ɔI]) for *Sg*1, from 61 Hz ([ɛ]) to 75 Hz for *Sg*2 ([i]), and from 98 Hz ([u]) to 118 Hz ([ɔI]) for *Sg*3. (2) For all three SGRs, $RMSE_{avg}$ across vowels is of the same order as the average within-speaker standard deviations shown in Table 3, and MSD_{avg} across vowels is comparable to the average within-speaker percentage COVs shown in Table 3. Therefore, the algorithm's performance can be considered accurate to within measurement error.

Fig. 11 shows the overall $RMSE_{avg}$ and MSD_{avg} corresponding to all the AE and MS vowels in the WashU-UCLA bilingual corpus. As in the case of the native English speakers (Fig. 10), the algorithm's performance in the case of the bilingual speakers is only slightly vowel dependent. However, the more important observation here is that the algorithm is equally accurate and consistent for AE and MS vowels (with the exception of [au] and, particularly, [a]) despite being trained using AE data only. This *language independent* nature of the algorithm can be attributed to two factors. (1) Bark *differences* between vocal tract formants (B_{31} and B_{32}), which are essential to the SGR estimation algorithm, do not contain any language-specific information about vowels; they are simply acoustic measures of vowel height (B_{31}) and backness (B_{32}). (2) Acoustic features such as *F*3 and *F*0, which provide auxiliary information for estimating SGRs, do not vary significantly with the language spoken (since they carry speaker-related

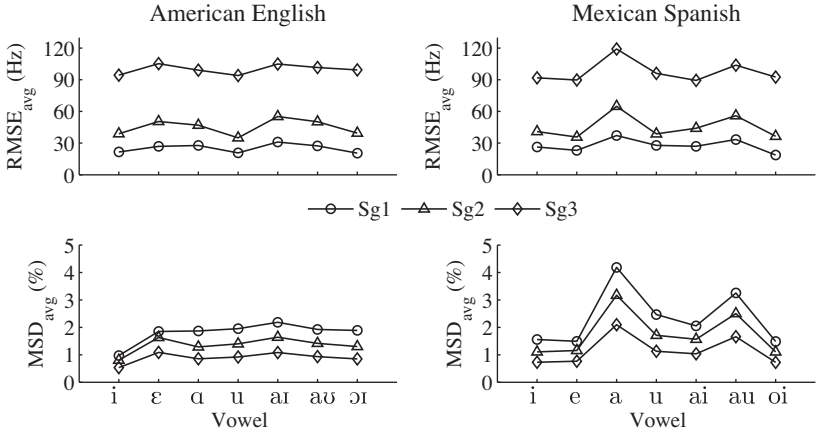

Fig. 11. SGR estimation using isolated vowels: overall $RMSE_{avg}$ and MSD_{avg} corresponding to the AE (left) and MS (right) vowels recorded in the WashU-UCLA bilingual corpus. For practical purposes, the performance can be considered to be language independent.

information). Next, we show that the algorithm is effective in estimating SGRs from continuous speech.

4.1.2. Estimation using continuous speech

The test set consisted of complete sentences from the WashU-UCLA corpus (140 per speaker), the WashU-UCLA bilingual corpus (147 in AE and 147 in MS, per speaker) and the MIT tracheal resonance database (between 85 and 170 per speaker). All sentences were less than 2 s in duration.

The first three SGRs were estimated from each sentence in the test set. Hence, every SGR estimate was obtained using less than 2 seconds of speech. Table 6 shows the results—$RMSE_{avg}$ and MSD_{avg}—for the WashU-UCLA corpus and the MIT tracheal resonance database (separated by gender), and Table 7 shows the results for the WashU-UCLA bilingual corpus (separated by language). The following observations can be made from Table 6. (1) Compared to males, females have larger values of $RMSE_{avg}$ but smaller values of MSD_{avg} (especially in the case of *Sg*3). The slightly larger values of females' $RMSE_{avg}$ might be attributed, at least in part, to the fact that LPC-based formant estimation is less accurate for speakers with high-pitched voices (usually females) as compared to speakers with low-pitched voices (usually males) (Makhoul, 1975). However, this gender dependence in performance is small enough to be ignored for practical purposes. (2) The overall $RMSE_{avg}$ for *Sg*3 estimation is approximately 100 Hz. In comparison, *Sg*3 estimation using Eq. (1) (which, like Eq. (5), requires an estimate of *Sg*2) incurs an overall $RMSE_{avg}$ in excess of 300 Hz. Therefore, Eq. (5) provides a more accurate model of the relation between *Sg*2 and *Sg*3. From Table 7, it can be observed once again that the algorithm is language independent. Also, as observed earlier in Section 4.1.1, the values of $RMSE_{avg}$ and MSD_{avg} corresponding to all three databases (Tables 6 and 7) are comparable respectively to the average within-speaker standard deviations and percentage COVs reported in Table 3.

The results in Tables 6 and 7 were obtained by providing the algorithm with one sentence of data (less than 2 seconds) per estimate. To see if the algorithm performed better with more data, we estimated the SGRs of every test speaker by providing the algorithm with up to 10 sentences per estimate. Fig. 12 shows the overall $RMSE_{avg}$ and MSD_{avg}—corresponding to *Sg*1 and *Sg*2—as a function of the number of sentences used for estimation (*Sg*3 shows the same trend as *Sg*2 since it is estimated from *Sg*2). As the number of sentences increases from 1 to 10, $RMSE_{avg}$ decreases slightly (by 11% for *Sg*1 and 10% for *Sg*2, on average), but MSD_{avg} decreases considerably (by 67% for both *Sg*1 and *Sg*2, on average). Therefore, the algorithm's performance does improve as the amount of available data increases. The more attractive feature of the algorithm, however, is that it performs well even when data is limited (which can be useful for automatic speaker normalization and adaptation).

Table 7
SGR estimation using one sentence of continuous speech: overall $RMSE_{avg}$ and MSD_{avg} for the WashU-UCLA bilingual corpus.

	*Sg*1		*Sg*2		*Sg*3	
	AE	MS	AE	MS	AE	MS
$RMSE_{avg}$ (Hz)	20	18	40	32	100	90
MSD_{avg} (%)	1.2	1.2	0.9	0.9	0.6	0.6

4.2. Automatic estimation of speaker height

To evaluate the proposed height estimation procedure, we used data from 604 speakers (431 males, 173 females) in the TIMIT corpus—10 sentences (each between 1 and 4 s in duration) per speaker for a total of 6040 sentences. The remaining 26 speakers in the corpus were not part of the evaluation because their heights were outside the range spanned by the training data (used for deriving Eqs. (6)–(8)). Given a speech utterance, speaker height was estimated by first estimating *Sg*1, *Sg*2 and *Sg*3, and then using Eqs. (6)–(8), respectively.

To the best of our knowledge, the height estimation algorithms proposed in Ganchev et al. (2010a,b) are the most accurate of all the existing algorithms—they yield an MAE (mean absolute error) of 5.3 cm and an RMSE of 6.8 cm over 168 speakers in the TIMIT corpus. To compare the proposed method with the algorithms in Ganchev et al. (2010a,b), MAE and RMSE were used as the perfor-

Table 6
SGR estimation using one sentence of continuous speech: $RMSE_{avg}$ and MSD_{avg} for the WashU-UCLA corpus and the MIT tracheal resonance database. For practical purposes, the performance can be considered to be gender independent.

	*Sg*1			*Sg*2			*Sg*3		
	Males	Females	Overall	Males	Females	Overall	Males	Females	Overall
WashU-UCLA corpus									
$RMSE_{avg}$ (Hz)	22	32	**25**	64	65	**61**	97	125	**104**
MSD_{avg} (%)	1.9	1.4	**1.6**	1.4	1.2	**1.2**	0.9	0.8	**0.8**
MIT tracheal resonance database									
$RMSE_{avg}$ (Hz)	25	30	**28**	52	61	**57**	74	129	**101**
MSD_{avg} (%)	2.8	1.3	**2.1**	2.0	1.1	**1.6**	1.3	0.7	**1.0**

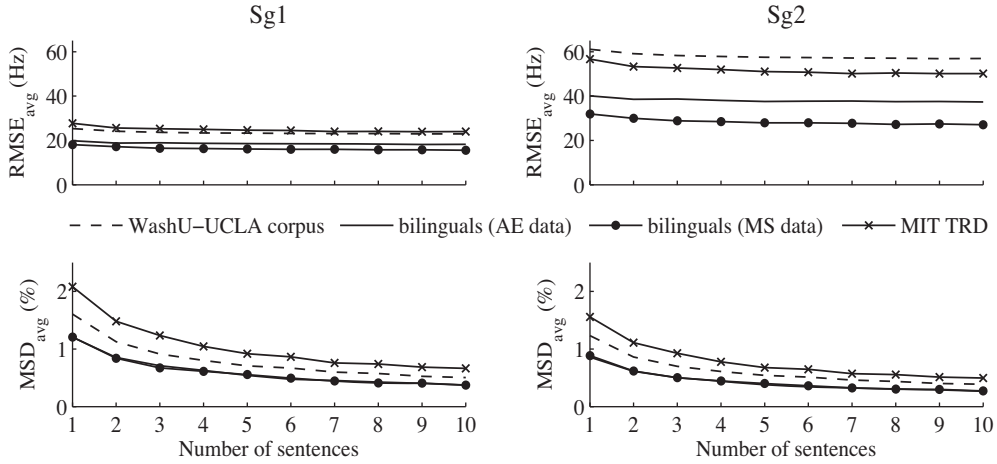

Fig. 12. SGR estimation using continuous speech: overall RMSE_{avg} and MSD_{avg}—corresponding to $Sg1$ (left) and $Sg2$ (right)—as a function of the number of sentences used for estimation. As the amount of data increases, RMSE_{avg} decreases only slightly, but MSD_{avg} decreases significantly (it must be noted that in the bottom two panels, the curves for MS data overlap the curves for AE data).

mance metrics. The following equations were used to calculate MAE and RMSE:

$$\text{MAE} = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{1}{M_i}\sum_{j=1}^{M_i}\left|h_a^i - h_e^{ij}\right| \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s}\frac{1}{M_i}\sum_{j=1}^{M_i}(h_a^i - h_e^{ij})^2}, \tag{12}$$

where N_s, M_i, h_a^i and h_e^{ij} denote the number of test speakers, the number of test utterances for the ith speaker, the actual height of the ith speaker, and the height estimate corresponding to the jth utterance of the ith speaker, respectively.

Depending on the amount of data used for estimating height, MAE and RMSE were calculated in two different ways. (1) When one sentence of speech data was used per height estimate, MAE and RMSE were calculated at the *'sentence level'*. In other words, M_i was equal to 10 (the number of sentences per speaker) in Eqs. (11) and (12). The sentence-level metrics will henceforth be denoted as MAE_{st} and RMSE_{st}. (2) When height was estimated using a single utterance formed by concatenating all 10 sentences of a given speaker, MAE and RMSE were calculated at the *'speaker level'*. In other words, M_i was equal to 1 in Eqs. (11) and (12). The speaker-level metrics will henceforth be denoted as MAE_{sp} and RMSE_{sp}. MAE_{sp} and RMSE_{sp} were expected to be smaller than MAE_{st} and RMSE_{st} because the SGR estimation algorithm (the basis for height estimation) performed better when data was not limited (see Fig. 12).

Table 8 lists the sentence-level and speaker-level MAEs and RMSEs corresponding to automatic height estimation using $Sg1, Sg2$ and $Sg3$. The following observations can be made from Table 8. (1) Considering the overall (males and females combined) performance metrics, $Sg1$ and $Sg2$ are almost equally good for estimating speaker height from speech signals. Despite a stronger correlation between speaker height and $Sg2$ (see Section 3.3), $Sg1$-based height estimation is superior for female speakers. This could possibly be because the SGR estimation algorithm is more accurate in estimating $Sg1$ than $Sg2$ (see Tables 6 and 7); verifying if that is actually the case is difficult because the actual SGR values of TIMIT speakers are unknown. (2) $Sg3$ gives slightly poorer results than $Sg1$ and $Sg2$ (especially for female speakers). This is presumably because the estimation of $Sg3$ is indirect, requiring an intermediate estimate of $Sg2$. If estimated directly from speech data, $Sg3$ might be able to estimate height as accurately as the other two SGRs. (3) The sentence-level metrics are slightly worse than the corresponding speaker-level metrics, but are still quite acceptable. This means that the proposed method is effective even when data is limited. (4) The overall MAE_{sp} and RMSE_{sp} for $Sg2$-based height estimation—5.4 cm and 6.7 cm—are comparable to the results in Ganchev et al. (2010a,b), while the overall MAE_{sp} and RMSE_{sp} for $Sg1$-based estimation—5.3 cm and 6.6 cm—are marginally better. Although the proposed method is not significantly better than the best existing algorithms, it is much more efficient in two respects: (1) *amount of training data and generalizability*: Ganchev et al. (2010a,b) trained and evaluated their algorithms on 462 and 168 TIMIT speakers, respectively (train-to-test ratio = 2.75), while the proposed method was trained on just 56 speakers in the WashU-UCLA corpora and evaluated on 604 speakers in the TIMIT corpus (train-to-test ratio < 0.1); (2) *size of the feature set*: Ganchev et al. (2010a,b) used a 50-dimensional feature vector to estimate height, while the proposed method used just *one* feature ($Sg1, Sg2$ or $Sg3$).

In addition to MAE and RMSE, the correlation between actual height and estimated height (or equivalently, between actual height and estimated SGR frequencies) was considered important to the assessment of height estimation performance. The correlation was fairly

Table 8
Sentence-level and speaker-level MAEs and RMSEs for automatic height estimation using $Sg1$, $Sg2$ and $Sg3$. In comparison, the algorithms in Ganchev et al. (2010a,b) were reported to yield an overall MAE and RMSE of 5.3 cm and 6.8 cm, respectively.

	Using $Sg1$			Using $Sg2$			Using $Sg3$		
	Males	Females	Overall	Males	Females	Overall	Males	Females	Overall
MAE_{st} (cm)	5.6	5.0	**5.4**	5.6	5.4	**5.5**	5.6	5.9	**5.7**
MAE_{sp} (cm)	5.5	4.9	**5.3**	5.5	5.2	**5.4**	5.5	5.8	**5.6**
$RMSE_{st}$ (cm)	6.9	6.4	**6.8**	6.9	6.9	**6.9**	7.0	7.4	**7.1**
$RMSE_{sp}$ (cm)	6.8	6.2	**6.6**	6.8	6.7	**6.7**	6.9	7.3	**7.0**

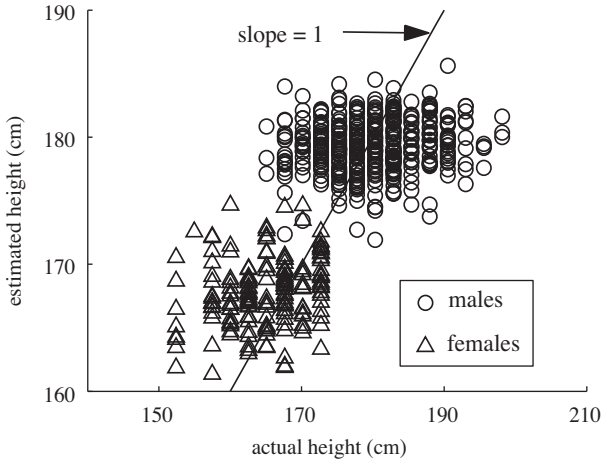

Fig. 13. Scatter plot of estimated height (using $Sg1$) versus actual height (604 data points). The correlation between the two quantities is poor within gender, suggesting that the proposed method requires further improvement.

strong when male and female data were pooled together ($r = 0.71$ for all three SGRs), but not when they were treated separately ($r = 0.12$ for males, and 0.21 for females, for all three SGRs), as shown in Fig. 13. In comparison, the within-gender correlations between 'ground truth' SGRs and height (for data in the WashU-UCLA corpora) were significantly better: $|r| = \{0.43\ (Sg1),\ 0.63\ (Sg2),\ 0.57\ (Sg3)\}$ for males, and $\{0.23\ (Sg1),\ 0.48\ (Sg2),\ 0.37\ (Sg3)\}$ for females.

4.2.1. Analysis of height estimation performance

Although the proposed height estimation method yields satisfactory MAEs and RMSEs within gender (see Table 8), it needs improvement in terms of the correlation coefficients between estimated height and actual height. However, considering that the proposed method is as accurate as the best existing algorithms while being much more simple and efficient, it is important to understand why the correlations between estimated and actual height are as poor as they are, and to find ways in which they can be improved.

The proposed method estimates speaker height as a linear function of estimated SGR frequencies. This means, as mentioned earlier, that the correlation between estimated and actual height is identical to the correlation between estimated SGRs and actual height (except for a change in sign). Naturally, therefore, the within-gender correlations between estimated and actual height tend to be weakened by SGR estimation errors. While the proposed SGR estimation algorithm is fairly accurate—errors are comparable to the standard deviations in measurements (see Section 4.1.2)—, it probably needs to be improved further from the height estimation point of view. It must also be noted that the proposed algorithm estimates SGRs somewhat indirectly (owing to its dependence on $F0$ and formant measures), and is constrained to be suitable for limited-data applications such as rapid speaker normalization. More direct approaches (relying explicitly on the interactions between SGRs and speech signals), or approaches without data-related constraints, are likely to be more accurate than the proposed algorithm.

Apart from minimizing SGR estimation errors, a possible solution for improving the within-gender correlations would be to develop alternative models between speaker height and SGRs. Based on the findings in Lulich et al. (2011), we previously tried using a uniform-tube model between SGRs and height (Arsikere et al., 2012). Although the linear models proposed in this paper (Eqs. (6)–(8)) yield better results than the uniform-tube model (see the discussion in Section 5.2 for details), a first-order linear regression is probably not the optimal solution. However, since the amount of data available at present is limited (56 speakers with known heights and SGRs), it is not yet clear as to what kind of model would be the most appropriate for height estimation.

While there exist well-defined solutions (at least two) to improve our present results, algorithms such as those presented by Ganchev et al. (2010a,b) seem to offer little room for improvement (mainly because the many features used by them are not physiologically motivated).

4.2.2. Height estimation from telephone speech

Estimating an unknown speaker's height from narrowband telephone speech can be of importance to forensic applications (see Pellom and Hansen, 1997). To see if the proposed method could estimate height using telephone speech, it was applied to a narrowband evaluation set generated by filtering TIMIT data (from the 604 speakers used earlier) with the ITU-T G.712 filter, which has a flat frequency response between 300 and 3400 Hz (ITU-T recommendation G.712, 2001). The resulting MAEs, RMSEs and correlation coefficients were identical to those obtained for

wideband (unfiltered) data, confirming that the proposed method did not suffer any degradation in narrowband conditions. In contrast, the algorithms proposed in Ganchev et al. (2010a,b) and Pellom and Hansen (1997) are likely to suffer a performance degradation with filtered speech owing to their dependence on features derived from spectral envelopes (e.g., MFCCs).

5. Discussion

In Section 5.1, we address the question as to what the limit might be for estimating speaker height from speech signals, and in Section 5.2, we compare the methods proposed in this paper with some of our previously-proposed techniques.

5.1. Height estimation using speech signals: existence of performance limits

From Section 4.2.1, it is clear that the proposed height estimation method can be improved further. It is important, however, to have an idea of the limit to the accuracy of speech-based height estimation. To this end, it is necessary to assess the limits defined by vocal tract-based and SGR-based approaches separately, because these limits arise from different physiological constraints. From Section 4.2, the most important metric for evaluating height estimation performance appears to be the within-gender correlations between estimated and actual height. Therefore, the limit of height estimation accuracy will be assessed with respect to this metric.

SGR-based approaches require estimates of SGRs in order to estimate speaker height. Therefore, the maximum correlations that can be achieved between estimated and actual height are governed largely by the correlations between 'ground truth' SGRs and actual height. The WashU-UCLA corpora suggest that these correlations (in magnitude) are roughly between 0.3 and 0.6, with an average value of 0.45 (statistically significant, $p < 0.05$; see Section 4.2 for details). Therefore, it is probably correct to say that the correlations achievable using SGR-based approaches have a limiting value close to 0.5 (for the range of speaker heights encountered in this study). Since this limit probably arises from physiological constraints, it would also be interesting to find out what those constraints are, and why the limit cannot possibly be higher than what it appears to be.

As mentioned in Sections 1.1 and 3.3, SGR frequencies are determined primarily by the 'acoustic length' of the subglottal system. Physiologically, since the 'acoustic length' is expected to be correlated with the size of the lungs and the length of the trunk (or torso), SGRs are likely to be strongly correlated with trunk length. However, according to physiological data reported in Hrdlička (1925), trunk length itself appears to be only moderately correlated with overall body height. Specifically, Hrdlička (1925) reports that the ratio of trunk length and height is a function of height itself, and that short speakers (males as well as females) have larger trunk length-to-height ratios than tall ones. Such a relationship between trunk length and height seems to be partly responsible for the weak correlations observed in Fig. 13, with height being overestimated for short speakers and underestimated for tall ones (for both genders). In essence, SGRs, when estimated well, may provide accurate estimates of trunk length but only moderately-accurate estimates of speaker height. In light of these observations, a value of 0.5 (as mentioned above) appears to be a reasonable estimate of the limiting correlation for SGR-based approaches.

In contrast to SGR-based methods, vocal tract-based approaches rely on the correlations between VTL and height. Fig. 5 of Fitch and Giedd (1999) shows VTL as a function of height, and Table 5 of the paper reports the corresponding correlation coefficients separated by gender. Although the correlations are strong—roughly 0.8 for both males and females—they result from the fact that the data spans a wide range of speaker heights within gender. To enable a comparison with the data used in this study, we analyzed a subset of the data plotted in Fig. 5 of Fitch and Giedd (1999) (male speaker heights between 165 and 201 cm, and female speaker heights between 152 and 175 cm). The x- and y-coordinates of the data points were obtained with the help of the program *Tracer, v.1.7* (Karolweski), and the within-gender correlation between VTL and height was found to be 0.3 for both males and females (not significantly different from 0.0, $p > 0.05$). Similarly, Rendall et al. (2005) found the correlations between speaker height and the first four formants of schwa vowels—which have relatively open vocal tract configurations—to be less than or equal to 0.3 for females (0.16 on average), and less than 0.59 for males (0.41 on average; the correlations were higher among males probably because the range of male speaker heights in their study was about 20% smaller than the range of heights in our data). Note that in the above analyses, the number of speakers is comparable to the size of our own training data. For the range of speaker heights encountered in this study, a value of 0.3 (approximately) appears to be the limiting correlation for vocal tract-based approaches; this is considerably lower than the corresponding limit for SGR-based methods.

It therefore appears that the correlations between SGRs and speaker height determine the limit of height estimation accuracy, although the limit itself can vary depending on the range of speaker heights under consideration. Furthermore, it is probably easier to achieve the SGR-based limit owing to the fact that the subglottal system of a given speaker, unlike his/her vocal tract, is effectively time invariant.

5.2. Comparison with previous techniques

Before proceeding to the major conclusions of this study, it is important to understand how the methods

proposed in this paper differ from our previously-proposed techniques.

- *Estimation of SGRs* – As mentioned in Section 3.2.1, the $Sg1$ estimation algorithm in Arsikere et al. (2011a) differs from the proposed approach in that it uses a different vocal tract-based measure of vowel height—the Bark difference between $F1$ and $F0$ (instead of the Bark difference between $F3$ and $F1$). When applied to the 20 test speakers in the WashU-UCLA corpus, the algorithm in Arsikere et al. (2011a) yields an overall RMSE_{avg} of 27 Hz. In comparison, the result given by the proposed algorithm—25 Hz—is better by 7%. The $Sg2$ estimation algorithm proposed in Arsikere et al. (2011b) differs from the proposed approach in that it does not use $F3$ and $F0$ for predicting $B_{2,s2}$. When applied to the 20 test speakers in the WashU-UCLA corpus, the algorithm in Arsikere et al. (2011b) yields an overall RMSE_{avg} of 121 Hz, which, in comparison with the result given by the proposed algorithm—61 Hz—is worse by almost 50%. This confirms that $F3$ and $F0$ are indeed important to the prediction of $B_{2,s2}$ and $B_{1,s1}$. Finally, as mentioned in Section 3.2.3, the proposed model for estimating $Sg3$ from $Sg2$ (Eq. (5)) is significantly better than the one suggested by Wang et al. (2008a) (Eq. (1)).
- *Height estimation using SGRs* – While the proposed height estimation method uses empirical relations of the form $h = \alpha x + \beta$ (h denotes speaker height, x denotes SGR frequency, and α and β are constants), our previous approach to height estimation used a uniform-tube model (Arsikere et al., 2012). The uniform-tube model comprises empirical relations of the form $h = C/x$, where C is a constant incorporating the speed of sound and an empirical scaling factor relating speaker height and the 'acoustic length' of the subglottal system (Lulich et al., 2011). Despite being better motivated on physiological grounds, the uniform-tube model of Arsikere et al. (2012) performed worse than the proposed method: across 563 speakers in the TIMIT corpus, an MAE_{sp} of 5.6 cm was achieved using $Sg2$. The uniform-tube model ($h = C/x$) does not perform as well as expected, probably because it results in height estimation errors that depend both on SGR estimation errors and the SGR frequencies themselves: $\Delta h = (-C/x^2)\cdot \Delta x$. In contrast, height estimation errors due to the proposed linear model are not a function of SGR frequencies: $\Delta h = \alpha \cdot \Delta x$.

6. Summary and conclusions

In this study, a novel algorithm for automatically estimating the first three subglottal resonances from speech signals of adults was developed. In addition, the algorithm was applied to the task of automatic speaker height estimation using speech signals.

Two recently-collected databases comprising simultaneous recordings of speech and subglottal acoustics—the WashU-UCLA corpus and the WashU-UCLA bilingual corpus—were used to train the SGR estimation algorithm and the height estimation procedure. Along with the two WashU-UCLA corpora, the MIT tracheal resonance database and the TIMIT speech corpus were used for evaluation purposes.

SGR frequencies of subjects with accelerometer recordings were measured with the help of LPC and WPSD spectra. The SGR measurements of a given speaker showed very small spreads about their mean values: the average within-speaker standard deviations of SGR measurements ranged between 22–30 Hz for $Sg1$, 30–32 Hz for $Sg2$ and 49–61 Hz for $Sg3$. The 'ground truth' values (averages of SGR measurements) of $Sg1, Sg2$ and $Sg3$ ranged approximately between 530–660 Hz, 1310–1510 Hz and 2160–2460 Hz, respectively. Female speakers had higher SGR frequencies than male speakers, on average.

Data from 30 speakers in the WashU-UCLA corpus were used to train the SGR estimation algorithm. $Sg1$ was estimated with the help of a model trained to predict an $Sg1$-based measure of vowel height ($B_{1,s1}$) from a vocal tract-based measure (B_{31}) and two speaker-related features ($F3$ and $F0$). Similarly, $Sg2$ was estimated with the help of a model trained to predict an $Sg2$-based measure of vowel backness ($B_{2,s2}$) from a vocal tract-based measure (B_{32}), $F3$ and $F0$. $Sg3$ was estimated using an empirically-derived first-order linear equation relating $Sg3$ and $Sg2$. Given a continuous speech signal, SGR estimates were obtained for every voiced frame in the signal and the averages of the frame-level estimates were calculated. The algorithm was evaluated on 20 speakers in the WashU-UCLA corpus, and all speakers in the WashU-UCLA bilingual corpus and the MIT tracheal resonance database. The algorithm's performance (in terms of RMSE_{avg} and MSD_{avg}) was found to be practically independent of vowel content as well as language (AE or MS). With less than 2 s of continuous speech per estimate, the average RMS errors incurred in estimating $Sg1, Sg2$ and $Sg3$ were less than 28, 61 and 104 Hz, respectively. The algorithm's performance, particularly its consistency, improved with the amount of speech data used for estimation. The proposed algorithm is therefore effective in estimating the first three SGRs from a limited amount of continuous speech, in a content-independent and language-independent manner.

It is important to note that the proposed SGR estimation algorithm is designed to suit real-time limited-data applications such as rapid speaker normalization and adaptation. If there is no restriction on the amount of speech data that can be used, it might be possible to develop more sophisticated and accurate algorithms for estimating SGRs.

Speaker height was estimated with the help of empirical first-order linear relations between height and SGR frequen-

cies, which were derived using data from speakers in the two WashU-UCLA corpora. Given a speech signal, speaker height was estimated by first estimating the SGR frequencies and then using the empirical relations between height and SGRs. This procedure was evaluated on 604 speakers in the TIMIT corpus. With up to 4 seconds of speech, mean absolute errors of 5.4, 5.5 and 5.7 cm were incurred in estimating speaker height using $Sg1$, $Sg2$ and $Sg3$, respectively; the errors reduced by 0.1 cm when 30–40 s of speech data was used. Actual height correlated well with estimated height when male and female data were pooled together, but not when they were considered separately (although the within-gender MAE and RMSE were satisfactory). This appears to be primarily due to the errors incurred in estimating SGR frequencies. The within-gender correlations between estimated height and actual height can probably be improved up to a certain limit (roughly 0.5) by developing more accurate SGR estimation algorithms and more appropriate models between speaker height and SGR frequencies. Despite its limitations, the proposed height estimation method is an improvement over existing algorithms because: (1) it achieves comparable performance while being much more transparent (well motivated features), efficient (small feature set and limited training data) and generalizable (test corpus much larger than the training corpus); (2) there exist well-defined solutions to improve it in the future; (3) it can perform equally well in wideband and narrowband (e.g., telephone speech) conditions; and (4) its optimal implementation is likely to perform better than the optimal vocal tract-based approach (which is expected to achieve within-gender correlations close to 0.3 or smaller).

Acknowledgments

We would like to thank John R. Morton for his role in recording and labeling the WashU-UCLA corpora. We are also thankful to Dr. Mitchell S. Sommers for his valuable suggestions, and to Melissa Erickson for help with manual measurements. The work was supported in part by NSF Grant no. 0905381.

References

Arsikere, H., Lee, Y.H., Lulich, S.M., Morton, J.R., Sommers, M.S., Alwan, A., 2010. Relations among subglottal resonances, vowel formants, and speaker height, gender, and native language (A). J Acoust Soc Amer 128, 2288.

Arsikere, H., Lulich, S.M., Alwan, A., 2011a. Automatic estimation of the first subglottal resonance. J. Acoust. Soc. Am. (Express Lett.) 129, 197–203.

Arsikere, H., Lulich, S.M., Alwan, A., 2011b. Automatic estimation of the second subglottal resonance from natural speech. In: Proc. of ICASSP, pp. 4616–4619.

Arsikere, H., Leung, G., Lulich, S.M., Alwan, A., 2012. Automatic height estimation using the second subglottal resonance. In: Proc. of ICASSP, pp. 3989–3992.

Boersma, P., Weenink, D., Praat Speech Processing Software. Institute of Phonetics Sciences of the University of Amsterdam. <http://www.praat.org>.

Cheyne, H.A., 2002. Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck. Ph.D. Thesis, MIT.

Chi, X., Sonderegger, M., 2004. Subglottal coupling and vowel space (A). J. Acoust. Soc. Amer. 115, 2540.

Chi, X., Sonderegger, M., 2007. Subglottal coupling and its influence on vowel formants. J. Acoust. Soc. Amer. 122, 1735–1745.

Chistovich, L.A., 1985. Central auditory processing of peripheral vowel spectra. J. Acoust. Soc. Amer. 77, 789–805.

Csapó, T.G., Bárkányi, Z., Gráczi, T.E., Bőhm, T., Lulich, S.M., 2009. Relation of formants and subglottal resonances in Hungarian vowels. In: Proc. of Interspeech, pp. 484–487.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366.

Dogil, G., Lulich, S.M., Madsack, A., Wokurek, W., 2011. Crossing the quantal boundaries of features: subglottal resonances and Swabian diphthongs. In: Goldsmith, J.A., Hume, E., Wetzels, W.L. (Eds.), Tones and Features: Phonetic and Phonological Perspectives. De Gruyter Mouton, pp. 137–148.

Dusan, S., 2005. Estimation of speaker's height and vocal tract length from speech signal. In: Proc. of Interspeech, pp. 1989–1992.

Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Amer. 106, 1511–1522.

Ganchev, T., Mporas, I., Fakotakis, N., 2010a. Audio features selection for automatic height estimation from speech. Artif Intell Theor Models Appl, 81–90.

Ganchev, T., Mporas, I., Fakotakis, N., 2010b. Automatic height estimation from speech in real-world setup. In: Proc. of the 18th European Signal Processing Conf., pp. 800–804.

Garofolo, J.S., 1988. Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. National Institute of Standards and Technology (NIST).

González, J., 2004. Formant frequencies and body size of speaker: a weak relationship in adult humans. J. Phonetics 32, 277–287.

Gráczi, T.E., Lulich, S.M., Csapó, T.G., Beke, A., 2011. Context and speaker dependency in the relation of vowel formants and subglottal resonances-evidence from Hungarian. In: Proc. of Interspeech.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Amer. 97, 3099–3111.

Honda, K., Takano, S., Takemoto, H., 2010. Effects of side cavities and tongue stabilization: Possible extensions of the quantal theory. J. Phonetics 38, 33–43.

Hrdlička, A., 1925. The Old Americans. The Williams and Wilkins Company, Baltimore, MD.

ITU-T recommendation G.712, 2001. Transmission performance characteristics of pulse code modulation channels.

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

Karolewski, M. 1.7. Last accessed on 4/24/2012. <http://sites.google.com/site/kalypsosimulation/Home/data-analysis-software-1>.

Künzel, H.J., 1989. How well does average fundamental frequency correlate with speaker height and weight? Phonetica 46, 117–125.

Lulich, S.M., 2006. The role of lower airway resonances in defining vowel feature contrasts. Ph.D. Thesis, MIT.

Lulich, S.M., 2010. Subglottal resonances and distinctive features. J. Phonetics 38, 20–32.

Lulich, S.M., Morton, J.R., Sommers, M.S., Arsikere, H., Lee, Y.H., Alwan, A., 2010. A new speech corpus for studying subglottal acoustics in speech production, perception, and technology (A). Journal of the Acoustical Society of America 128, 2288.

Lulich, S.M., Alwan, A., Arsikere, H., Morton, J.R., Sommers, M.S., 2011. Resonances and wave propagation velocity in the subglottal airways. J. Acoust. Soc. Amer. 130, 2108–2115.

Madsack, A., Lulich, S.M., Wokurek, W., Dogil, G., 2008. Subglottal resonances and vowel formant variability: a case study of high German monophthongs and Swabian diphthongs. Proc. LabPhon 11, 91–92.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63, 561–580.

Nelson, D., 1997. Correlation based speech formant recovery. In: Proc. of ICASSP, pp. 1643–1646.

Pellom, B.L., Hansen, J.H.L., 1997. Voice analysis in adverse conditions: the centennial Olympic park bombing 911 call. In: 40th Midwest Symposium on Circuits and Systems, pp. 873–876.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175–184.

Rendall, D., Kollias, S., Ney, C., Lloyd, P., 2005. Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice-acoustic allometry. J. Acoust. Soc. Amer. 117, 944–955.

Sjölander, K., 1997. The Snack sound toolkit. KTH, Stockholm, Sweden. <http://www.speech.kth.se/snack/>.

Sjölander, K., Beskow, J., 2000. Wavesurfer-an open source speech tool. In: Proc. of ICSLP, pp. 464–467.

Sonderegger, M., 2004. Subglottal coupling and vowel space: an investigation in quantal theory. B.S. Thesis, MIT.

Stevens, K.N., 1998. Acoustic Phonetics. MIT Press, Cambridge, MA.

Syrdal, A.K., Gopal, H.S., 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. JASA 79, 1086–1100.

Traunmüller, H., 1990. Analytical expressions for the tonotopic sensory scale. JASA 88, 97–100.

Umesh, S., Cohen, L., Marinovic, N., Nelson, D.J., 1999. Scale transform in speech analysis. IEEE Trans. Speech Audio Process. 7, 40–45.

van Dommelen, W.A., Moxness, B.H., 1995. Acoustic parameters in speaker height and weight identification: sex-specific behaviour. Lang Speech 38, 267–287.

Wang, S., Alwan, A., Lulich, S.M., 2008a. Speaker normalization based on subglottal resonances. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280.

Wang, S., Lulich, S.M., Alwan, A., 2008b. A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation. In: Proc. of Interspeech, pp. 1717–1720.

Wang, S., Lee, Y.H., Alwan, A., 2009a. Bark-shift based nonlinear speaker normalization using the second subglottal resonance. In: Proceedings of Interspeech, pp. 1619–1622.

Wang, S., Lulich, S.M., Alwan, A., 2009b. Automatic detection of the second subglottal resonance and its application to speaker normalization. J. Acoust. Soc. Amer. 126, 3268–3277.