Age, sex, and vowel dependencies of acoustic measures related to the voice source[a)]

Markus Iseli,[b)] Yen-Liang Shue,[c)] and Abeer Alwan[d)]
Department of Electrical Engineering, University of California Los Angeles, 405 Hilgard Avenue, Los Angeles, California 90095

(Received 22 February 2006; revised 23 January 2007; accepted 24 January 2007)

The effects of age, sex, and vocal tract configuration on the glottal excitation signal in speech are only partially understood, yet understanding these effects is important for both recognition and synthesis of speech as well as for medical purposes. In this paper, three acoustic measures related to the voice source are analyzed for five vowels from 3145 CVC utterances spoken by 335 talkers (8–39 years old) from the CID database [Miller *et al.*, Proceedings of ICASSP, 1996, Vol. **2**, pp. 849–852]. The measures are: the fundamental frequency (F_0), the difference between the "corrected" (denoted by an asterisk) first two spectral harmonic magnitudes, $H_1^*-H_2^*$ (related to the open quotient), and the difference between the "corrected" magnitudes of the first spectral harmonic and that of the third formant peak, $H_1^*-A_3^*$ (related to source spectral tilt). The correction refers to compensating for the influence of formant frequencies on spectral magnitude estimation. Experimental results show that the three acoustic measures are dependent to varying degrees on age and vowel. Age dependencies are more prominent for male talkers, while vowel dependencies are more prominent for female talkers suggesting a greater vocal tract-source interaction. All talkers show a dependency of F_0 on sex and on F_3, and of $H_1^*-A_3^*$ on vowel type. For low-pitched talkers ($F_0 \leqslant 175$ Hz), $H_1^*-H_2^*$ is positively correlated with F_0 while for high-pitched talkers, $H_1^*-H_2^*$ is dependent on F_1 or vowel height. For high-pitched talkers there were no significant sex dependencies of $H_1^*-H_2^*$ and $H_1^*-A_3^*$. The statistical significance of these results is shown.
 [DOI: 10.1121/1.2697522]

PACS number(s): 43.70.Gr [BHS] Pages: 2283–2295

I. INTRODUCTION

For almost half a century, research has been conducted on the nature of the glottal voice source signal, and glottal source parameters have been estimated using various procedures and algorithms. In the past, the study of the voice source signal has mainly centered on voice synthesis and speech coding applications. However, recent studies (Fant *et al.*, 2000; Sluijter and Van Heuven, 1996; Sluijter *et al.*, 1997) have shown that a relationship exists between the characteristics and/or parameters of the glottal voice source signal, and voice quality. A better knowledge of the relationship of acoustic measures that characterize the voice source with speaker properties such as sex and age, and with context or sound type such as vowel, would benefit the understanding of the human voice production mechanism and help improve voice analysis for a variety of speech processing and medical applications.

The human voice production mechanism can be roughly divided into three parts: lungs, vocal folds, and vocal tract. Air pressure from the lungs causes air to flow through the glottis, which is the airspace between the vocal folds. In voiced speech the vocal folds open and close quasiperiodically and thus convert the glottal air flow (air volume velocity) into a train of flow pulses, called the voice source excitation signal. This signal then passes through the vocal tract, which functions as an acoustic filter that shapes the spectrum of the sound, and at the end of the vocal tract the volume velocity signal is modified by the lip impedance. The speech pressure waveform measured in front of the lips can be approximated by the time derivative of the volume velocity signal (Rabiner and Schafer, 1978). This radiation effect is typically included in the source function, i.e., the source signal is modeled as the derivative of the glottal flow volume velocity. Sounds produced with nonvibrating vocal folds, such as in the fricative /f/, are called unvoiced sounds and are not studied in this paper.

Here, we use the linear source-filter model of speech production (Fant, 1960), in which the derivative of the glottal flow volume velocity acts as the source, sometimes also referred to as the excitation, and the vocal tract acts as the linear filter. The fact that source and filter are assumed to be independent of each other is one reason that this is the simplest and commonly used model for speech production. Early models of the source signal used a simple impulse train for modeling voiced signals. More recent studies model the shape of the glottal airflow or its derivative in the time-domain (Ananthapadmanabha, 1984; Fant *et al.*, 1985; Hedelin, 1984; Holmes, 1973; Klatt and Klatt, 1990; Rosenberg, 1971). Frequency-domain representations for some of those models were presented in Fant (1995) and Doval and d'Alessandro (1999). In this paper, the Liljencrants-Fant

[a)]Portions of this paper were presented at ICASSP04 and ICASSP06.
[b)]Electronic mail: iseli@ee.ucla.edu
[c)]Electronic mail: yshue@ee.ucla.edu
[d)]Electronic mail: alwan@ee.ucla.edu

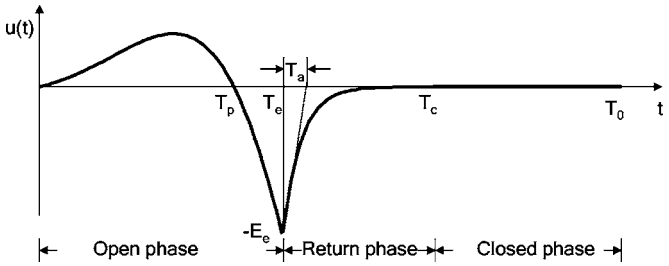

FIG. 1. The LF model and its parameters: instant of maximum airflow (T_p), instant of maximum airflow derivative (T_e), effective duration of return phase (T_a), beginning of closed phase (T_c), fundamental period (T_o), and amplitude of maximum excitation of glottal flow derivative (E_e).

(LF) model by Fant *et al.* (1985) is used to generate synthetic stimuli. It models the glottal volume velocity derivative, hence incorporating the effect of lip radiation, and is illustrated in Fig. 1. Vocal tract models, on the other hand, evolved from electric circuit models (Miller, 1959) and acoustic tube models (Fant, 1960) to all-pole autoregressive representations (Markel and Gray, 1976). For vowels, the vocal tract is typically modeled as an all-pole filter, where each complex-conjugate pole-pair represents a resonance frequency (formant) and its bandwidth.

To recover glottal source parameters from the acoustic speech signal, vocal tract resonances need to be removed by an "inverse filtering" process. The linear source-filter model assumes that the vocal tract is a linear filter and that it is independent and linearly separable from the source, all of which facilitates inverse filtering. Inverse filtering was first presented by Miller (1959), who applied analog electronic filters to cancel the two lowest formants and the lip radiation effect from the speech pressure waveform captured by a microphone. Rothenberg (1973) introduced a different inverse filtering technique that measures the airflow at the mouth and nose with a special mask. This method allows the estimation of absolute flow values, including the dc component, as opposed to the inverse filtering of the pressure signal captured by a microphone, which loses the absolute zero level of flow due to the lip radiation effect. The flow measurement mask is also less sensitive to low-frequency noise and the mask's frequencies are band limited at approximately 1.6 kHz (Hertegård and Gauffin, 1992). For all recording equipment, be it mask or microphone, it is important that its frequency magnitude response is flat and its phase response is linear from very low frequencies up to high frequencies. Compared to analog filtering, digital sampling, storage, and filtering techniques provide obvious advantages over analog techniques, since they are flexible, repeatable, easy to implement, and cause no phase distortion. Because of these advantages, today, digital inverse filtering methods are almost always used.

To find vocal tract filter parameters, typically a linear predictive coding based analysis is applied (Hertegård and Gauffin, 1992). However, more accurate results can usually be achieved with the method of discrete all-pole modeling (DAP) introduced by El-Jaroudi and Makhoul (1991). DAP uses a cost function which is based on the Itakuro-Saito distance evaluated at the discrete frequencies of the signal power spectrum. A recent publication which uses the DAP method in combination with a code book of source functions, generated with the LF model, and an iterative optimization algorithm is described in Fröhlich *et al.* (2001). These approaches to obtaining the glottal flow waveform are computationally expensive, and often need manual correction and tuning. Instead of trying to estimate the time domain parameters of the source models, researchers can study acoustic measures which are correlated with these parameters. This typically involves analyzing the harmonic frequencies in the speech spectrum, such as the magnitudes of the first two spectral harmonics of the source spectrum, located at the fundamental frequency F_0 and at $2F_0$, and the spectral magnitude of various formant peaks. This is less computationally intensive and less prone to error than finding the glottal flow waveform, and is therefore suited for analyzing the extensive amount of data needed for a reliable statistical evaluation. Spectral harmonics, however, are affected by both the source characteristics and by vocal tract resonances (formants). Hence, if one needs only to characterize the source signal properties, then the influence of vocal tract resonances, or formant frequencies, need to be compensated for (Fant, 1982, 1995; Hanson, 1995; Mártony, 1965). The correction in this paper is done using both formant frequencies and their bandwidths (Iseli and Alwan, 2004). The formula can be applied to voices produced with high fundamental frequency and/or low first formant frequency.

Holmberg *et al.* (1995) showed that the difference between the corrected (denoted by as asterisk hereafter) magnitudes of the first two harmonics ($H_1^*-H_2^*$) is correlated with the open quotient (OQ). On the other hand, Henrich *et al.* (2001) showed that $H_1^*-H_2^*$ is dependent on both OQ and glottal flow asymmetry. Hanson (1997) found that $H_1^*-A_3^*$, where A_3^* is the corrected spectrum level at the frequency of the third formant, is correlated with the source spectral tilt. The correction accounted for the first two formants (F_1 and F_2) and the bandwidth of the third formant (F_3), and tokens were normalized with respect to a neutral vowel. In addition, Hanson and Chuang (1999) showed that the acoustic characteristics of the glottal excitation signal are gender dependent. Their study compared the effects of gender on voice source parameters for about 21 adult male and adult female talkers for the three vowels /eh/, /ae/, and /ah/. It analyzed three acoustic cues—open quotient (as shown by the magnitude difference between the first two spectral harmonics), first formant bandwidth, and source spectral tilt (as shown by the difference between the magnitude of the first spectral harmonic and the corrected spectrum level at F_3)—and showed that open quotient and source spectral tilt are generally higher for adult female than for adult male talkers. Speech acoustics are also affected by age, which was shown in a study by Lee *et al.* (1999). It analyzed fundamental frequency (F_0) and formant frequencies for a large speech database (Miller *et al.*, 1996) with about 490 subjects in the age range of 5–50 years. The study showed that children have higher F_0 and formant frequencies, and greater temporal and spectral variability than adults. These findings are attributed to vocal-tract anatomical differences and possible differences in the ability to control speech articulators.

This paper has two specific aims. The first is to introduce and evaluate, through error analysis, a spectral magnitude correction formula, which uses both bandwidth and frequency estimates of the resonant frequencies of the vocal tract. This formula can be used to reliably estimate acoustic measures related to the voice source signal, such as the difference between the magnitude of the first two source spectral harmonics. The second aim is to use the correction formula to uncover age, sex, and vowel dependencies for the source parameter F_0 (fundamental frequency) and two acoustic measures: $H_1^*-H_2^*$ (related to open quotient), and $H_1^*-A_3^*$ (related to source spectral tilt). The dependencies are analyzed using speech signals recorded from 335 people (185 males, 150 females) in ten age groups from the CID database (Miller *et al.*, 1996).

The paper is organized as follows: In Sec. II a spectral magnitude correction formula is presented and its accuracy is evaluated through error analysis. Results on age, sex, and vowel dependencies of the three acoustic measures $(F_0, H_1^*-H_2^*, H_1^*-A_3^*)$ are presented in Sec. III. A summary in Sec. IV concludes the paper.

II. CORRECTION FORMULA AND ERROR ANALYSIS

In the following, a spectral magnitude correction formula, which uses both bandwidth and formant frequencies, is presented and evaluated. The formula can be used to reliably estimate acoustic measures related to the voice source signal such as the difference between the magnitude of the first two spectral harmonics.

A. A correction formula to compensate for the effects of formant frequencies in the speech spectrum

The spectral magnitude of the speech signal is the result of interactions from both the voice source and the vocal tract. The spectral magnitude formant correction formula (Iseli and Alwan, 2004), which requires no explicit inverse-filtering techniques, assumes the linear source-filter model of speech production (Fant, 1960) and is derived in the Appendix. The purpose of this correction formula is to "undo" the effects of the formants on the magnitudes of the source spectrum. This is done by subtracting the amount by which the formants boost the spectral magnitudes. Theoretically, if the formant frequencies and their respective bandwidths were known exactly and the linear source-filter model is applicable, then the corrected spectral magnitudes should represent the actual magnitudes of the source spectrum. For example, the corrected magnitude of the first spectral harmonic located at frequency ω_0, where $\omega_0=2\pi F_0$ and F_0 is the fundamental frequency, is given by

$$H^*(\omega_0) = H(\omega_0) - \sum_{i=1}^{N} 10\log_{10}\frac{(1-2r_i\cos(\omega_i)+r_i^2)^2}{(1-2r_i\cos(\omega_0+\omega_i)+r_i^2)(1-2r_i\cos(\omega_0-\omega_i)+r_i^2)} \quad (1)$$

with $r_i=e^{-\pi B_i/F_s}$ and $\omega_i=2\pi F_i/F_s$ where F_i and B_i are the frequencies and bandwidths of the ith formant, F_s is the sampling frequency, and N is the number of formants to be corrected for. $H(\omega_0)$ is the magnitude of the first harmonic from the speech spectrum and $H^*(\omega_0)$ represents the corrected magnitude and should coincide with the magnitude of the source spectrum at ω_0. Note that all magnitudes are in decibels. A less general form of this equation ($N=2$) is used in Sec. III, where only the first two formants are corrected for.

B. Error analysis of the correction method

To evaluate the accuracy of the correction formula (with and without bandwidth information) in estimating harmonic spectral magnitudes, error analysis is performed. In Secs. II B 1 and II B 2 error analysis is done using synthetic single-, and three-formant vowels, respectively. Specifically, error analysis is evaluated for the H_1-H_2 parameter. For the synthetic stimuli, the LF voice source signal is filtered with an all-pole model of the vocal tract. The LF shape is defined by $T_p=0.48$, $T_e=0.6$, and $T_a=0.05$, with $T_c=T_o=1$.

Analysis errors are calculated without using correction (NoC); with correction for the influence of only the first formant, F_1, without using bandwidth information, that is, by setting B_1 in Eq. (1) to zero, (F1noB1); and correction for the influence of F_1 using exact bandwidth information (F1B1). It is important to note that when $\omega=\omega_i$ for the F1noB1 case ($B_i=0$), the correction yields an infinite value [see Eq. (A5)].

1. Error analysis for single-formant synthetic signals

Formant correction is applied to single-formant synthetic signals with F_0 varying between 100 and 300 Hz, and F_1 between 200 and 800 Hz with constant bandwidth (B_1) of 75 Hz. Since the signals are synthetic, the actual values for H_1 and H_2 are known and the correction error between the actual and estimated harmonics' magnitudes can be calculated.

Figure 2 compares the H_1-H_2 error at $F_0=250$ Hz for the cases NoC, F1noB1, and F1B1. Maximum errors for NoC and F1noB1 occur at $F_1=F_0$ and $F_1=2F_0$, where the absolute NoC error is about 24 dB and the F1noB1 error is infinite. The error for F1B1 is zero, which is expected.

2. Error analysis for three-formant synthetic vowels

The vowels /a/, /i/, and /u/, are synthesized using the first three formant frequencies specified in Peterson and Barney (1952). Formant bandwidths are calculated according to the formula in Mannell (1998):

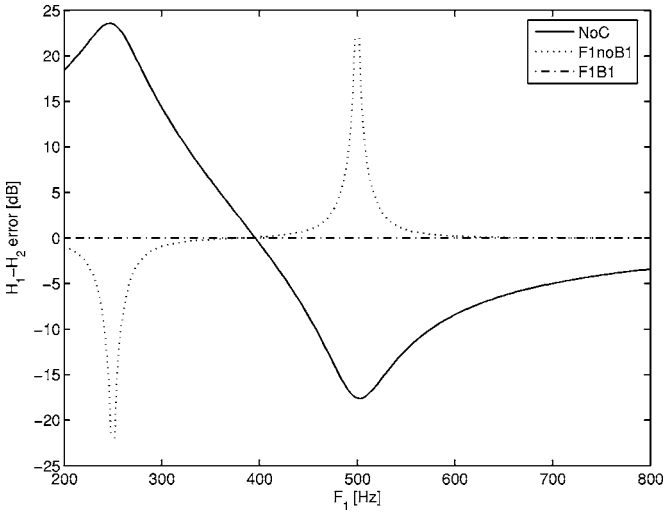

FIG. 2. H_1-H_2 error in decibels with $F_0=250$ Hz and $B_1=75$ Hz for synthetic one-formant signals. The three curves represent: NoC, no correction (solid line); F1noB1, correction for F_1 not using bandwidth information (dotted line); and F1B1, correction for F_1 using exact bandwidth information (dash-dotted line). The maximum NoC error is about 24 dB. The absolute error for the F1noB1 correction at $F_1=F_0$ and $F_1=2F_0$ is infinite, and the F1B1 error is zero.

$$B_i = (80 + 120F_i/5000). \quad (2)$$

These values are shown in Table I.

F_0 is chosen from the ranges provided by Baken (1987): For male talkers, F_0 ranges between 85 and 154 Hz, for female talkers F_0 is between 164 and 256 Hz, and for children F_0 is between 208 and 256 Hz. The sampling frequency (F_s) is at 10 kHz.

For each sex, vowel, and correction method, the minimum, average, and maximum absolute estimation errors for $|H_1-H_2|$ are calculated over the appropriate range of F_0. The results are shown in Table II. F1noB1 introduces the highest errors especially when F_1 is close to F_0 or $2F_0$. For the vowel /a/, on the other hand, F1noB1 performs similarly to F1B1 because /a/ has a very high F_1, which is greater than $2F_0$, and hence, the influence of F_1 on the first two harmonics is small. The errors for F1B1 are lower but are not zero since F1B1 does not correct for F_2 and F_3. The highest F1B1 errors are measured for /u/, which has the lowest F_2 of the three vowels.

TABLE II. Min/Mean/Max $|H_1-H_2|$ error in decibels without correction (NoC), correction for F_1 without bandwidth information (F1noB1), and correction for F_1 using bandwidth information (F1B1). Synthesis included three formants. As a reference, F_1 is given in parentheses for each of the vowels. It can be seen that the errors for NoC and F1noB1 are high when F_1 is close to F_0 or $2F_0$. The error for F1noB1 where $F_1=F_0$ or $F_1=2F_0$ is infinite.

Vowel (F_1 in Hz)	Min/Mean/Max Error in decibels		
	NoC	F1noB1	F1B1
	Male talkers (F_0: 85–154 Hz)		
/a/ (730)	0.57/1.06/1.99	0.20/0.38/0.69	0.20/0.38/0.69
/i/ (270)	3.04/5.58/8.15	0.41/∞/∞	0.07/0.13/0.23
/u/ (300)	2.66/5.61/9.67	0.00/∞/∞	0.30/0.56/1.04
	Female talker (F_0: 164–256 Hz)		
/a/ (850)	1.71/2.84/4.73	0.64/1.02/1.63	0.63/1.02/1.63
/i/ (310)	0.14/5.31/12.20	0.08/1.90/7.82	0.21/0.33/0.52
/u/ (370)	0.05/7.29/11.47	0.03/∞/∞	0.98/1.62/2.67
	Child talkers (F_0: 208–256 Hz)		
/a/ (1030)	2.05/2.59/3.25	0.86/1.07/1.33	0.83/1.04/1.30
/i/ (370)	0.36/2.91/5.97	0.01/0.73/2.15	0.29/0.36/0.44
/u/ (430)	4.60/9.59/12.48	0.23/∞/∞	1.08/1.36/1.71

Figures 3 and 4 show the absolute $|H_1-H_2|$ error as a function of F_0 for the methods NoC, F1noB1, and F1B1 for synthetic /a/ and /u/ vowels, respectively. Figure 3 shows the error for the synthetic female /a/ ($F_1=850$ Hz) where correction without using bandwidth information (F1noB1) works well. As mentioned earlier, this is due to F_1 being much higher than F_0 or $2F_0$, hence, the first formant does not have a significant effect on the magnitudes of the first two harmonics. However, for the female /u/ (Fig. 4), bandwidth information becomes important in the correction since $F_1=2F_0=370$ Hz when $F_0=185$ Hz. Hence, F1B1 yields significantly better results than F1noB1.

Since it is difficult to estimate bandwidths accurately (Hanson and Chuang, 1999), we also compare these results with another case, F1B50, which applies the correction formula using a constant bandwidth, $B_1=50$ Hz. The average

TABLE I. Formant frequencies (Peterson and Barney, 1952) and bandwidths (Mannell, 1983) in Hertz used to synthesize the three corner vowels appropriate for male, female, and child talkers.

Vowel	F_1	F_2	F_3	B_1	B_2	B_3
			Male talker			
/a/	730	1090	2440	98	106	139
/i/	270	2290	3010	86	135	152
/u/	300	870	2240	87	101	134
			Female talker			
/a/	850	1220	2810	100	109	147
/i/	310	2790	3310	87	147	159
/u/	370	950	2670	89	103	144
			Children			
/a/	1030	1370	3170	105	113	156
/i/	370	3200	3730	89	157	170
/u/	430	1170	3260	90	108	158

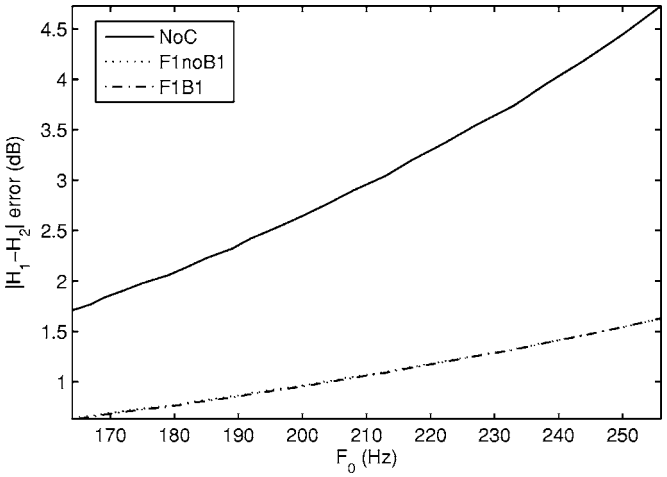

FIG. 3. $|H_1-H_2|$ error in decibels for a three-formant synthetic female /a/ (F_1=850 Hz, F_2=1220 Hz, F_3=2810 Hz) as a function of F_0. Error using NoC (solid line), with F1noB1 correction (dotted line), and with F1B1 (dash-dotted line). In this case, using bandwidth information is not critical since F_1 is much higher than $2F_0$.

absolute errors for the four cases NoC, F1noB1, F1B1, and F1B50, are shown in Fig. 5. It can be seen that the largest error occurs for the high back vowel /u/, since there is no correction for the low F_2. Using exact bandwidth information (F1B1) or using a fixed B_1 of 50 Hz improves significantly over F1noB1 for /i/ and /u/, which have low F_1. Interestingly, using a bandwidth estimate of 50 Hz (F1B50) yields similar results to using exact bandwidth information. These results imply that for reducing errors, it is better to use some bandwidth information, even if it is only an educated guess of the true bandwidth.

III. ESTIMATION OF ACOUSTIC MEASURES FOR NATURALLY PRODUCED SOUNDS

In the following we apply the correction formula to estimate age, sex, and vowel dependencies of three acoustic measures, F_0, $H_1^*-H_2^*$, $H_1^*-A_3^*$, on a relatively large speech database.

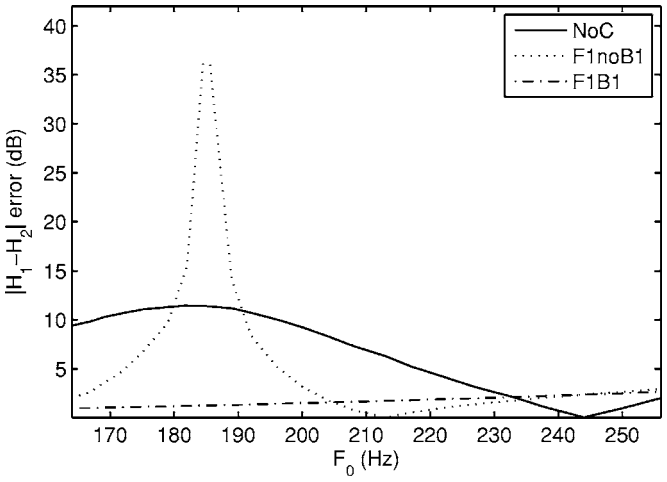

FIG. 4. $|H_1-H_2|$ error in decibels for a three-formant synthetic female /u/ (F_1=370 Hz, F_2=950 Hz, F_3=2670 Hz) as a function of F_0. Error using NoC (solid line), with F1noB1 correction (dotted line), and with F1B1 (dash-dotted line). F1B1 performed significantly better than F1noB1 since F_1 is quite low. The error for F1noB1 where $F_1=2F_0$ is infinite.

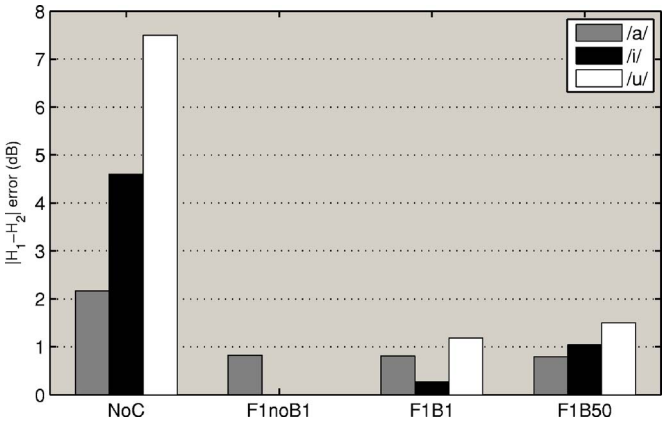

FIG. 5. (Color online) A bar diagram comparison of average $|H_1-H_2|$ error measurements for the three synthetic, three-formant vowels (averaged over both sexes, age groups, and corresponding F_0 values.) Results for NoC, F1noB1, F1B1, and F1B50, which is a correction for F_1 with B_1=50 Hz. No error bars are shown for F1noB1 for /i/ and /u/ since for some values of F_0 they can be infinite.

A. Speech data

Speech signals recorded from 335 people (185 males, 150 females) in ten age groups, ages 8, 9, 10, 11, 12, 13, 14, 15, 18, and 20–39, from the CID database (Miller *et al.*, 1996) were analyzed. The carrier sentence was "I say uh, bVt again," where the vowel was /ih/ (bit), /eh/ (bet), /ae/ (bat), and /uw/ (boot). "uh" was used before the target word to maximize vocal tract neutrality. The corner vowel /iy/ in "bead" was also analyzed. Most utterances were repeated twice by each speaker. Recordings were made at normal habitual speaking levels with a sampling frequency of 16 kHz. In total, 3145 utterances were analyzed. The age and sex distribution of the analyzed talkers is shown in Table III.

B. Methods

The voice source parameter F_0 and the acoustic cues $H_1^*-H_2^*$, and $H_1^*-A_3^*$ were estimated. As mentioned earlier, these measures are of significant importance in the areas of voice perception and voice synthesis (Fant and Kruckenberg, 1996; Holmberg *et al.*, 1995). $H_1^*-H_2^*$, the difference between the spectral magnitudes of the first two source harmonics, is related to the OQ (Holmberg *et al.*, 1995). $H_1^*-A_3^*$, the difference between the spectral magnitudes of the first harmonic and the third formant peak, is related to the source spectral tilt (Holmberg *et al.*, 1995). The asterisk denotes that spectral magnitudes (H_1,H_2,A_3) were corrected for the effects of formants. For H_1^* and H_2^*, the correction was for the first and second formant (F_1 and F_2) influence with

TABLE III. Number of analyzed talkers in each age group separated by sex (males: M; females: F).

Age	M	F	Age	M	F
8	25	11	13	16	13
9	24	25	14	11	10
10	25	14	15	11	11
11	24	19	18	10	10
12	22	21	20–39	17	16

$N=2$ in Eq. (A5). For A_3^*, the first three formants were corrected for ($N=3$) and there was no normalization to a neutral vowel; recall that our correction accounts for formant frequencies and their bandwidths.

The calculation of the three acoustic measures requires the estimation of the first three formant frequencies (F_1,F_2,F_3), their respective bandwidths (B_1,B_2,B_3), and F_0. Formant frequencies F_1, F_2, and F_3, as well as F_0 were estimated using the "SNACK SOUND TOOLKIT" software (Sjölander, 2004). The main parameters that can be changed in SNACK are frame length, frame shift, and analysis methods. For formant estimation, the covariance method was chosen because of its accuracy. F_0 can be extracted with either the ESPS (Entropic Signal Processing System), or the AMDF (Average Magnitude Difference Function (Ross *et al.*, 1974)) method. Both methods are based on conventional autocorrelation analysis. Since no significant estimation differences between the two methods were found, the ESPS method was used. Additional settings were: The preemphasis coefficient was 0.9, the length of the analysis window was 25 ms, and the window shift was 10 ms. Using the values extracted with SNACK, the amplitudes H_1, H_2, and A_3 were estimated from the speech spectrum. Since the SNACK bandwidth estimates varied greatly within the analysis segments and were sometimes unrealistic, all bandwidths were calculated from their corresponding formant frequency using Eq. (2). This reduced the bandwidth variance and therefore the variance of bandwidth-dependent results. Analysis segments were chosen at the steady-state part of the vowel, where the context influence was smaller than in other segments.

The estimates of F_0, F_1, F_2, and F_3 were manually checked for every utterance by viewing the spectrogram, time waveform, and LPC spectral slices. Most formant estimation errors occurred with child speech. For example, for high pitched /iy/, SNACK typically allocated two formants to the first spectral peak resulting in a much lower second formant frequency. The number of formant estimate corrections in percent, for 8 year old children, was: 86% for /iy/, 44% for /eh/, 32% for /ih/, and 2% for /uw/. The formant values are not listed here as the results are similar to those reported in Lee *et al.* (1999).

C. Results

In this section, we refer to males and females from ages 8 to 14, and females 15 years and older as "Group 1," and to male talkers age 15 and older as "Group 2." Group 1 talkers were typically high-pitched (with $F_0 > 175$ Hz) and Group 2 talkers were generally low-pitched (with $F_0 \leqslant 175$ Hz), although there were F_0 outliers within both groups as can be seen in the minimum/maximum F_0 values in Table VIII. The source parameter F_0, and acoustic measures $H_1^*-H_2^*$ and $H_1^*-A_3^*$ were analyzed as a function of age, sex, and vowel type, and their intercorrelations were studied.

1. Analysis of variance of the three acoustic measures

Statistical analysis was performed on the extracted acoustic measures by using the three-way analysis of variance (ANOVA) test in the software package SPSS (v13.0). The factors age (ages 8, 9, 10, 11, 12, 13, 14, 15, 18, and 20–39), sex (M, F) and vowel-type (/iy/, /ih/, /eh/, /ae/, and /uw/) were tested against the variables F_0, $H_1^*-H_2^*$ and $H_1^*-A_3^*$. These factors were tested with: (a) all the talkers, (b) the talkers separated by sex, and (c) the talkers separated into Group 1 and Group 2. Tests where the null hypothesis had a probability of $p<0.05$ were considered to be statistically significant. In addition, Pearson correlation coefficients were calculated to test for statistically significant intercorrelations between the three acoustic measures.

Table IV shows results for all the talkers for the F value (ratio of the model mean square to the error mean square) and partial η^2 (calculated as $SS_{effect}/(SS_{effect}+SS_{error})$, where SS_{effect} is the sum of squares of the effect and SS_{error} is the sum of squares of the error). Partial η^2 is a measure of effect size. For all three measures the effect size is greatest with age. For $H_1^*-H_2^*$ and $H_1^*-A_3^*$, the effect size of age is followed by vowel and sex, while for F_0, vowel type shows the smallest effect size.

Table V shows the ANOVA results when the talkers were separated by sex. It can be seen that across all three acoustic measures, the effect size of age is greater for males than for females. This was expected since speech acoustics, for example F_0 (Lee *et al.*, 1999), vary more significantly with age for male talkers. However, for vowel type, the effect size is greater for females than for males. This may suggest a greater vocal tract-source interaction for female talkers.

The results are also interesting when viewed in terms of the Group 1 (children and females, generally high-pitched) and Group 2 (older males, generally low-pitched) talkers. Table VI shows the F and partial η^2 results for Group 1 and Group 2 talkers. For Group 1 talkers, it can be seen that nearly all the entries are statistically significant except when sex is tested against $H_1^*-H_2^*$ and $H_1^*-A_3^*$. This result is interesting, since it suggests that females of all age groups have a similar OQ and source spectral tilt compared to boys (ages 8–14). More notable are the results for the Group 2 talkers which only have one significant entry: vowel type versus $H_1^*-A_3^*$. The lack of any age effect for Group 2 talkers is

TABLE IV. ANOVA results for all talkers showing F and partial η^2 values (in parentheses). All entries are statistically significant.

	F_0	$H_1^*-H_2^*$	$H_1^*-A_3^*$
Age	235.0 (0.410)	23.9 (0.066)	35.0 (0.094)
Sex	1012.3 (0.250)	57.7 (0.019)	4.1 (0.001)
Vowel	28.0 (0.036)	52.7 (0.065)	68.9 (0.083)

TABLE V. ANOVA results for female and male talkers showing F and partial η^2 values (in parentheses). All entries are statistically significant.

	F_0	$H_1^*-H_2^*$	$H_1^*-A_3^*$
Age (F)	26.4 (0.145)	2.8 (0.018)	8.8 (0.058)
Age (M)	310.3 (0.0618)	30.7 (0.138)	32.4 (0.144)
Vowel (F)	19.2 (0.052)	48.2 (0.121)	38.2 (0.098)
Vowel (M)	4.8 (0.011)	16.6 (0.037)	35.8 (0.076)

TABLE VI. ANOVA results for Group 1 (children and females) and Group 2 (older males) talkers showing F and partial η^2 values (in parentheses) for statistically significant entries. "⋯" denotes a nonsignificant entry. Sex is not included in the analysis for Group 2 since that group contains only male talkers.

	F_0	$H_1^*-H_2^*$	$H_1^*-A_3^*$
	Group 1		
Age	78.7 (0.208)	3.9 (0.013)	17.2 (0.054)
Sex	167.9 (0.059)	⋯	⋯
Vowel	26.1 (0.037)	75.9 (0.101)	65.1 (0.088)
	Group 2		
Age	⋯	⋯	⋯
Vowel	⋯	⋯	6.5 (0.069)

likely due to the fact that source characteristics for males do not change significantly with age above 15 years old; this has been shown for F_0 in Lee *et al.* (1999). Sex was not included for the Group 2 analysis since all the talkers in that group were male.

Table VII shows the Pearson correlation coefficients (PCCs) when the three acoustic measures were tested against each other. Although the intercorrelations are statistically significant, there is only one PCC greater than 0.7, indicating a strong correlation. This occurs for the relationship between $H_1^*-H_2^*$ and F_0 for Group 2 talkers.

2. F_0

Table VIII shows the range of F_0 values for all talkers. Note that F_0 was not normalized for stress. For males the mean F_0 drops by about 130 Hz between ages 8 and 20 with the largest drop between ages 12 and 15 (105 Hz), while the change is less dramatic for female talkers (overall about 50 Hz). These changes are reflected in Table V which shows that age has a greater effect size on F_0 for males (F/partial $\eta^2=310.3/0.618$) than for females (F/partial $\eta^2=26.4/0.145$). As expected, adult females exhibit higher F_0 values than adult male talkers: The difference in the means is about 110 Hz. These trends agree with the results in Lee *et al.* (1999). We noticed that a few very high F_0 values (above 300 Hz) were due to high stress on the target word. In those cases, F_0 was around 300 Hz for the rest of the sentence, but increased for the target word.

Average F_0 values are highest for /uw/, and higher for /iy/ than for /eh/ and /ae/. The trend of increasing F_0 as the tongue moves from a front to a back position and from open to closed vowels has been reported for German talkers by Marasek (1996). This trend can be seen for all ages and genders for the vowels in this study and may partly be explained by vowel-dependent intrinsic pitch (Lehiste and Peterson, 1961). ANOVA results in Table V indicate that although these trends are statistically significant for both males and females, the partial η^2 values, and hence the effect sizes of vowel type, are relatively small for both sexes: F/partial $\eta^2=19.2/0.052$ for females and $4.8/0.011$ for males. Interestingly, the vowel effect size on F_0 is five times higher for females. A further analysis into the vowel dependency was done by performing an ANOVA test on the effects of high and low formant frequencies (thresholds at the formant means) on F_0. It was found that F_0 was positively correlated only with F_3 for all talkers and this correlation was statistically significant (F/partial $\eta^2=133.1/0.041$); again the effect size was relatively small. This positive correlation can be explained by the fact that F_3 is typically correlated with vocal tract length (Wakita, 1977). Hence, a higher F_3, which typically results from a shorter vocal tract, coincides with a higher F_0.

TABLE VII. Pearson correlation coefficients (PCCs) for F_0, $H_1^*-H_2^*$ and $H_1^*-H_3^*$ for Group 1 and Group 2 talkers. Correlation coefficients greater than 0.7 indicate strong correlations. All results are statistically significant.

	F_0	$H_1^*-H_2^*$	$H_1^*-A_3^*$
	Group 1		
F_0	1	−0.471	−0.356
$H_1^*-H_2^*$	−0.471	1	0.532
$H_1^*-A_3^*$	−0.356	0.532	1
	Group 2		
F_0	1	0.767	0.268
$H_1^*-H_2^*$	0.767	1	0.473
$H_1^*-A_3^*$	0.268	0.473	1

TABLE VIII. Min/Mean/Max of F_0 (in Hz) per age group for vowels in the target syllables.

Age	F_0 males (Hz)	F_0 females (Hz)
8	170/255/420	152/283/423
9	160/264/454	187/267/437
10	141/256/407	146/266/367
11	167/256/378	185/254/494
12	125/230/328	178/236/338
13	119/190/285	180/251/394
14	101/177/272	169/228/293
15	95/125/251	179/228/310
18	84/129/239	199/246/310
20–39	88/127/191	156/235/356

3. $H_1^*-H_2^*$

The effects of age and sex on $H_1^*-H_2^*$ (related to open quotient) are shown in Fig. 6. Comparing the values, it is interesting to observe that the $H_1^*-H_2^*$ (mean value) separation between the genders is the clearest at age 15 (5.8 dB). Between ages 8 and 20–39, the mean $H_1^*-H_2^*$ value drops by about 4 dB for male talkers, whereas for female talkers it remains relatively unchanged. Having smaller changes in $H_1^*-H_2^*$ with age is reflected in the statistical analysis of Table V where the effects of age are less pronounced for females: F/partial $\eta^2=2.8/0.018$ vs $30.7/0.138$ for male talkers. The difference between genders may be related to the fact that F_0 drops significantly between age 12 and 15 for males while it does not change as much for females (Lee *et al.*, 1999). Adult females exhibit higher mean $H_1^*-H_2^*$ values (about 3.4 dB) than adult male talkers. A similar difference (3.1 dB) between adult male and adult female talkers was found in Hanson and Chuang (1999). When the talkers are split into Group 1 and Group 2 categories (see Table VI), it is interesting to note that the dependence on sex is not significant for Group 1 talkers (children and females).

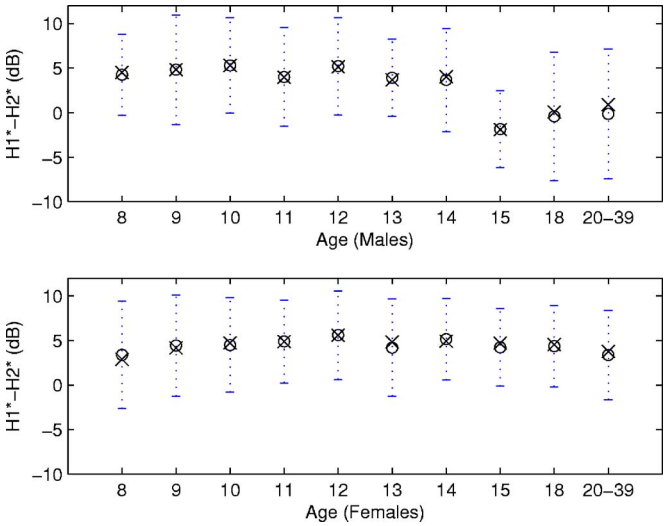

FIG. 6. (Color online) $H_1^*-H_2^*$ vs age, separated by sex. Between age 8 and 20–39, $H_1^*-H_2^*$ drops by about 4 dB for males, while for females there is little change. The largest difference between the sexes appears at age 15 where the difference in the means approaches 6 dB. Mean, median, and standard deviation are represented by circles, crosses, and whiskers, respectively.

Vowel effects are larger for female talkers than for males as shown in Table V (F/partial η^2=48.2/0.121 for females vs 16.6/0.037 for males). When analyzed against Group 1 and Group 2, the results in Table VI indicate that only Group 1 talkers exhibit a dependence on vowel (F/partial η^2 =75.9/0.101) whereas Group 2 (older male) talkers do not exhibit a significant dependence on vowel or on age.

ANOVA tests were also done to study the effects of formant values (thresholds at the formant means). The only statistically significant result is for F_1 with Group 1 talkers (F/partial η^2=91.4/0.034). No significant correlation between $H_1^*-H_2^*$ and F_1 (vowel height) can be observed for Group 2, or can a correlation with F_2 and F_3 be shown for any group. This effect can be seen in Fig. 7, which depicts

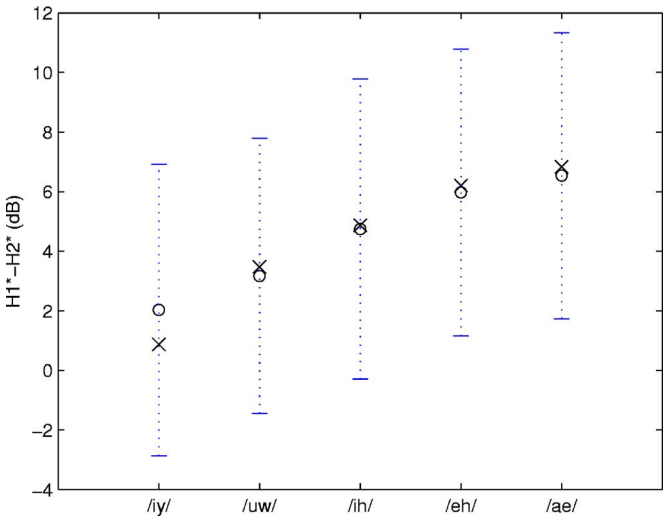

FIG. 7. (Color online) $H_1^*-H_2^*$ as a function of vowel for Group 1 talkers (females and children). Vowels are sorted according to their F_1 value from low to high. Note that the lowest values occur for the high and tense vowels /iy/ and /uw/.

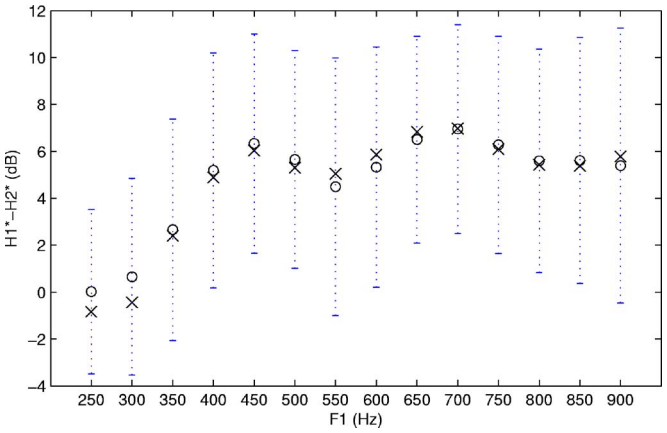

FIG. 8. (Color online) $H_1^*-H_2^*$ vs F_1 for Group 1 talkers. $H_1^*-H_2^*$ monotonically increases, on average, by about 6 dB when F_1 increases between 250 and 450 Hz.

$H_1^*-H_2^*$ as a function of vowel for the Group 1 talkers. Vowels are sorted from left to right as a function of their average F_1 value. $H_1^*-H_2^*$ values for /iy/ and /uw/ are the lowest, suggesting that high vowels have lower OQ. As F_1 increases for /iy/, /uw/, /ih/, /eh/, and /ae/, $H_1^*-H_2^*$ becomes larger. Figure 8 shows $H_1^*-H_2^*$ as a function of F_1 and agrees with Fig. 7 trends. Hanson (1997) showed that, for adult female voices, the mean value of $H_1^*-H_2^*$ was slightly lower for /eh/ than /ae/ which agrees with our results.

The lack of significant trends of $H_1^*-H_2^*$ values with F_1 for Group 2 talkers may be due to the physiology associated with voice production in different genders. This difference could be due to increased vocal tract-source interaction when F_0 or its harmonics are close to F_1 (Titze, 2004), which is often the case for low F_1 and high F_0.

For both sexes $H_1^*-H_2^*$ for /iy/ is about 3 dB lower than for /ih/. This could be due to the tense/lax difference. For four minority languages in China, Maddieson and Ladefoged (1985) reported that the amplitude difference between the first two harmonics was smaller for tense vowels than lax ones, which would agree with our findings.

Relationship of $H_1^*-H_2^*$ with F_0 and $H_1^*-A_3^*$

Figure 9 shows the relationship between $H_1^*-H_2^*$ and F_0 for both groups. As can be seen in Table VII, the PCC between $H_1^*-H_2^*$ and F_0 yields a value of 0.767 for Group 2 and a weak negative correlation (PCC=−0.471) for Group 1. An approximate mapping for $H_1^*-H_2^*$ and F_0 for Group 2 is

$$H_1^* - H_2^* \approx 0.22F_0 - 28 \quad \text{for } F_0 \text{ between 80 and 175 Hz.} \tag{3}$$

A possible interpretation for this result is that the Group 1 talkers (females and children, generally high-pitched) and the Group 2 talkers (older males, generally low-pitched) use OQ differently during the phonation of vowels. In a study by Esposito (2005) utilizing electroglottography of Zapotec talkers, females were shown to produce phonation differences by altering OQ while males did not. It has also been observed in Koreman (1996) that increased tension of the cricothyroid muscle in the larynx induces a simultaneous increase of F_0 and OQ, and therefore also of $H_1^*-H_2^*$. However,

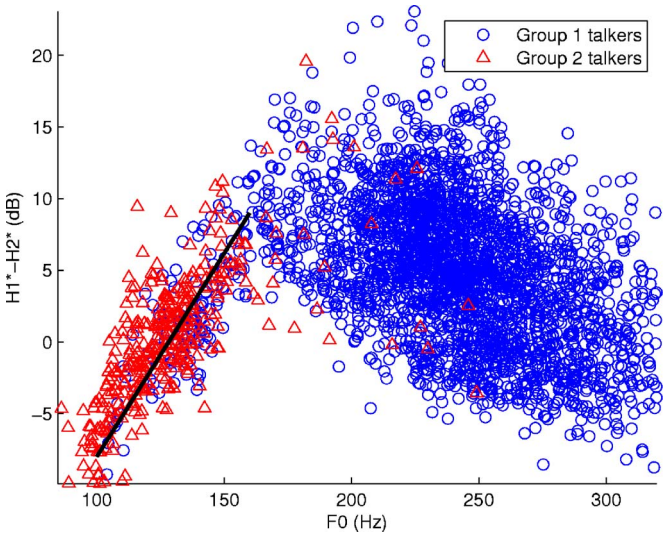

FIG. 9. (Color online) $H_1^*-H_2^*$ vs F_0 for Group 1 and Group 2 talkers. A linear relationship for F_0 between 80 and 175 Hz is observed.

we observed a strong positive correlation only for low F_0 values. Swerts and Veldhuis (2001) also found similar results for some of their speakers.

As seen in Table VII, the intercorrelation between $H_1^*-H_2^*$ and $H_1^*-A_3^*$ for both groups is weak: 0.532 (Group 1), 0.473 (Group 2). A weak correlation was also reported in Hanson (1997) for adult female talkers.

4. $H_1^*-A_3^*$

The age and sex effects on $H_1^*-A_3^*$ (related to source spectral tilt) are shown in Fig. 10. Between ages 8 and 20–39, the mean $H_1^*-A_3^*$ value drops for male talkers by about 10 dB, whereas for female talkers it drops by about 4 dB resulting in higher values (by about 4 dB) for adult females than for adult males. The higher effect size for males (F/partial η^2=32.4/0.144) compared to females

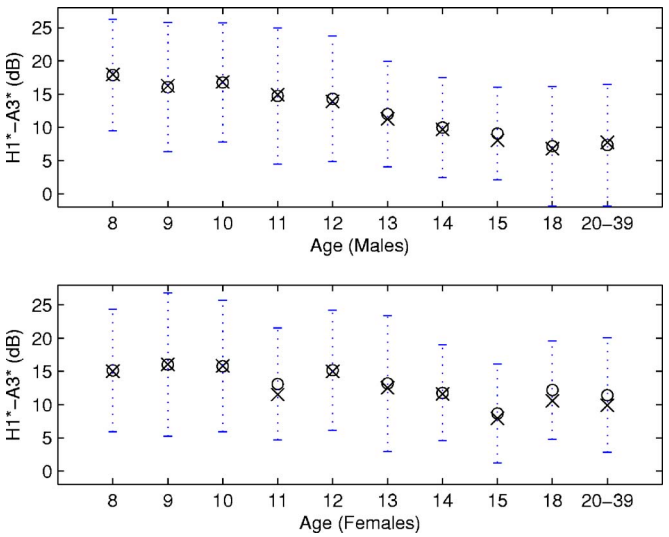

FIG. 10. (Color online) $H_1^*-A_3^*$ vs age; the top panel represents data for male talkers and the lower panel represents data for female talkers. For both sexes there is a drop of $H_1^*-A_3^*$ between age 8 and age group 20–39: The drop is about 4 dB for females, and 10 dB for males.

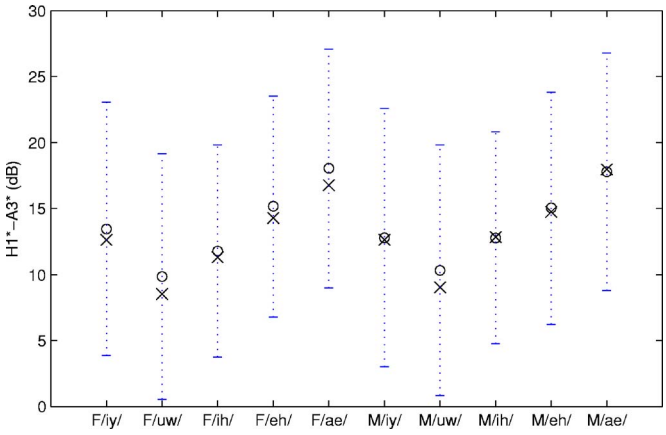

FIG. 11. (Color online) $H_1^*-A_3^*$ as a function of vowel for all talkers; M and F indicate data from male and female talkers, respectively. /ae/ and /eh/ have the highest values, while /uw/ has the lowest value. This result might be related to the dependence of the parameter on formant values.

(F/partial η^2=8.8/0.058) in Table V confirms this result. When the talkers are split into groups (see Table VI), Group 1 shows a dependence on age (F/partial η^2=17.2/0.054), whereas Group 2 does not. It is also interesting to note that the dependence on sex is not significant for Group 1. These trends are similar to those shown for $H_1^*-H_2^*$ (see Sec. III C 3), thus they can be interpreted similarly. That is, females (children and adults) and young males (8–14 years old) exhibit statistically similar OQ and source spectral tilt characteristics.

In Fig. 11, $H_1^*-A_3^*$ is depicted as a function of vowel and sex. The largest difference is observed between the vowels /ae/ and /uw/ where /ae/ is a low front vowel (high F_1, high F_2) and /uw/ is a high back vowel (low F_1, low F_2). Values for $H_1^*-A_3^*$ for /ae/ and /eh/ are the highest, and for /uw/ they are the lowest. These trends are similar for both sexes and indeed it can be seen from ANOVA analysis that the effect sizes of vowel are similar when male talkers are compared with females (Table V).

To find the effects of formants on $H_1^*-A_3^*$, an ANOVA analysis based on high and low values of F_1, F_2, and F_3 (thresholds at the formant means) yielded F/partial η^2 values of 210/0.063, 42.7/0.013, and 100.0/0.031, respectively. Thus, the first three formants have an effect on $H_1^*-A_3^*$ for all talkers. To visualize these effects, Figs. 12–14 show $H_1^*-A_3^*$ gradually rising for increasing F_1, F_2, and F_3. Since /uw/ on average has lower F_2 and F_3 compared to the other vowels used in this study, this can explain why $H_1^*-A_3^*$ values for /uw/ are lowest.

The dependency of $H_1^*-A_3^*$ on F_1 is somewhat similar to the dependency of $H_1^*-A_3^*$ on $H_1^*-A_1$ (related to F_1) which was observed in Hanson and Chuang (1999). The dependency of the measure on F_2 and F_3 was expected since a high F_2 is normally associated with a high F_3, which in term will affect the source spectral tilt. Since A_3^* represents the magnitude of the source spectrum at F_3, it is affected by the position of F_3 due to the source spectral tilt. A_3^* can also be influenced by the presence of higher formants, such as F_4, for which the parameter was not corrected for, and which would boost the value of A_3^* when evaluated close to F_4.

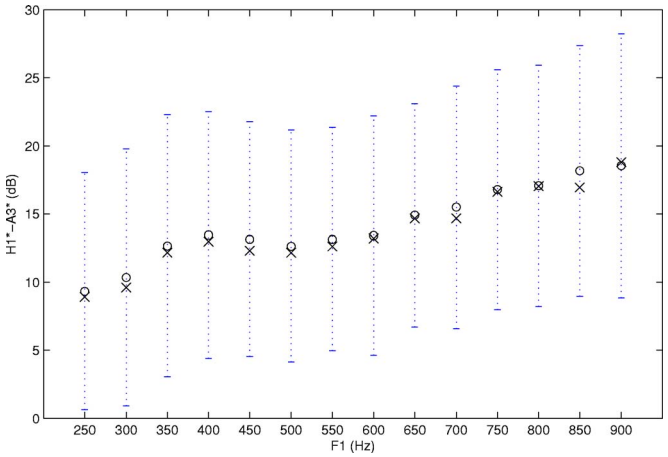

FIG. 12. (Color online) $H_1^*-A_3^*$ vs F_1 for all talkers. $H_1^*-A_3^*$ increases for increasing F_1.

IV. SUMMARY

In this paper the effects of age, sex, and vocal tract configuration on three acoustic measures related to voice source parameters: F_0, $H_1^*-H_2^*$, and $H_1^*-A_3^*$ are studied.

In order to estimate the acoustic measures $H_1^*-H_2^*$, and $H_1^*-A_3^*$, the vocal tract influence on the source spectrum needs to be compensated for. A correction formula which corrects for the influence of the vocal tract resonances is presented in Sec. II A. The importance of using the correction formula to estimate the magnitudes of the first two harmonics, H_1 and H_2, for vocal-tract influences, especially for high vowels and for high-pitched voices, is shown in Sec. II B. Synthetic speech is produced with formant frequencies from Peterson and Barney (1952) data and formant bandwidths are estimated from corresponding formant frequencies using Mannell's (1998) formula.

For synthetic speech, analysis errors are calculated without correction and with correction for the influence of only the first formant: (a) with bandwidth information, (b) without bandwidth information, and (c) with a bandwidth estimate of 50 Hz.

Error analysis results show that it is better to use an educated guess of formant bandwidth when correcting for the vocal tract influence, rather than using no bandwidth information [i.e., setting $B_i=0$ in Eq. (A5)] as in Hanson (1995). Examples of synthetic vowels show that correction without using bandwidth information can yield larger errors than no correction at all.

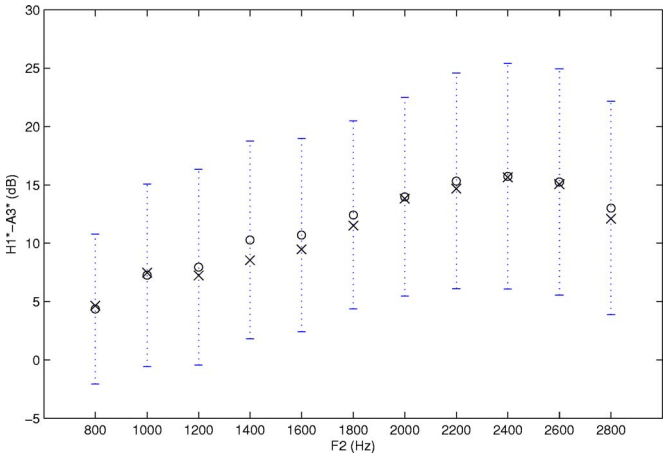

FIG. 13. (Color online) $H_1^*-A_3^*$ vs F_2 for all talkers. $H_1^*-A_3^*$ monotonically increases for F_2 increasing between 800 and 2400 Hz.

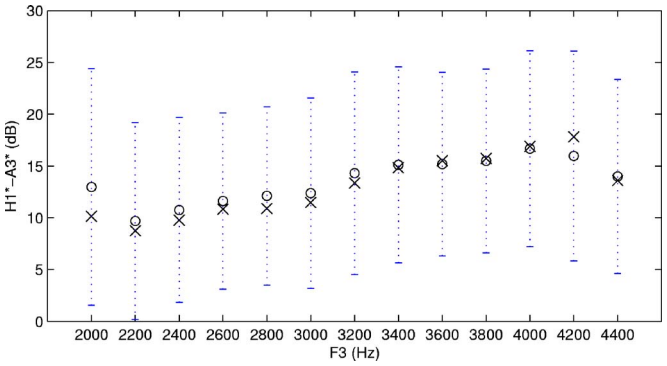

FIG. 14. (Color online) $H_1^*-A_3^*$ vs F_3 for all talkers. $H_1^*-A_3^*$ monotonically increases for F_3 increasing between 2200 and 4000 Hz.

The correction formula is then used to analyze acoustic measures related to source parameters for a relatively large speech database. The five vowels /iy/, /ih/, /eh/, /ae/, and /uw/ recorded from 335 people (185 males, 150 females) in ten age groups, ages 8, 9, 10, 11, 12, 13, 14, 15, 18, and 20–39, from the CID database (Miller *et al.*, 1996) are analyzed. F_0, as well as the formant frequencies, are extracted using the "SNACK SOUND TOOLKIT" software (Sjölander, 2004) and manually corrected if necessary. Bandwidth values are estimated from their corresponding SNACK formant frequencies again using Mannell's 1998 formula.

Statistical analysis of variance (ANOVA) is performed for all three acoustic cues and the three factors, age, sex, and vowel. These factors are tested with: (a) all talkers, (b) talkers separated by sex, and (c) talkers separated into Group 1 (children ages 8–14 and females ages 15 and older: generally high-pitched) and Group 2 (males ages 15 and older: generally low-pitched). In addition, where applicable, Pearson correlation coefficients are calculated for the different measurements. For Group 1, all effects are statistically significant except when sex is tested against $H_1^*-H_2^*$ and $H_1^*-A_3^*$. This result suggests that females of all age groups and boys (ages 8–14) have similar OQ and source spectral tilt values. For Group 2 the only significant result occurs when $H_1^*-A_3^*$ is tested against vowel type.

F_0 for male talkers drops between ages 8 and 20–39 (by about 130 Hz), whereas the overall drop for females is only about 50 Hz. F_0 is shown to be vowel dependent, with the highest values for /uw/, and higher for /iy/ than for /eh/ and /ae/. This trend may be attributed to intrinsic pitch. Furthermore, F_3 is shown to have a statistically significant relationship with F_0 which can be explained by the dependency of F_3 on vocal tract length.

$H_1^*-H_2^*$ (hence, the open quotient) is age dependent and for male talkers a drop by about 4 dB between the ages of 9 and 20–39 is found. For females, there is less dependency on age. On average, $H_1^*-H_2^*$ values are higher by about 3 dB for adult female compared to male talkers. There is no significant dependency on age and vowel for Group 2 talkers. H_1^*

TABLE IX. Summary of key results.

	Age (from 8 to 39 years old)		Vowel dependencies and intercorrelations
	Females	Males	
F_0	↓50 Hz	↓130 Hz	Linearly related to $H_1^*-H_2^*$ for low-pitched talkers, and to F_3 for all talkers
$H_1^*-H_2^*$	⋯	↓4 dB	Linearly related to F_0 for low-pitched talkers, and to F_1 for high-pitched talkers
$H_1^*-A_3^*$	↓4 dB	↓10 dB	Dependent on F_1, F_2, and F_3 for all talkers

$-H_2^*$ is proportional to F_0 for F_0 below 175 Hz. Above that frequency a weak negative correlation with F_0 could be found. For Group 1 talkers and for F_1 below 450 Hz, $H_1^*-H_2^*$ is proportional to F_1, resulting in low $H_1^*-H_2^*$ values for high vowels. For Group 2 talkers, on the other hand, no significant correlations between the $H_1^*-H_2^*$ values and vowel height could be observed. The different OQ dependencies between females and children (ages 8–14), and older males (ages 15 and older) could be due to physiological differences, to phonological differences, where females alter OQ to signal acoustic differences while males do not (Esposito, 2005), and/or to vocal tract-source interaction when F_0 or its harmonics are close to F_1 (Titze, 2004), which is often the case for low F_1 and high F_0 values. For both sexes $H_1^*-H_2^*$ for /iy/ is about 3 dB lower than for /ih/ which could be due to a tense/lax difference.

$H_1^*-A_3^*$ (hence source spectral tilt) values drop by about 10 dB between ages 8 and 20–39 for males, whereas for females the values drop by only about 4 dB within the same age period. This results in generally lower values for adult males (by about 4 dB) compared to adult females. Until age 10, the values are similar for both sexes. Statistical analysis shows a high dependence of the measure on age and vowel for all talkers. Also, $H_1^*-A_3^*$ shows a strong dependence on all formant frequencies for all talkers and age groups: Increasing F_1, F_2, or F_3 yields an increase in $H_1^*-A_3^*$. These findings imply that source spectral tilt is vowel dependent and, in fact, it can be seen that tilt values are highest for /ae/ and /eh/ and lowest for /uw/. The dependence of $H_1^*-A_3^*$ on F_3 can be explained by the dependence of A_3^* on F_3: Increasing F_3 will yield decreasing A_3^*.

Key dependencies are summarized in Table IX. Results show that all three acoustic measures are dependent to varying degrees on age and vowel. Age dependencies are more prominent for males than for females while vowel dependencies are more prominent for female talkers suggesting a greater vocal tract-source interaction. For $H_1^*-H_2^*$ vowel dependencies are only significant for Group 1 (generally high-pitched) talkers. F_0 shows a dependency on sex and on F_3, $H_1^*-H_2^*$ on F_1 (Group 1 talkers only), and $H_1^*-A_3^*$ on all three formants. For Group 2 (generally low-pitched) talkers F_0 is positively correlated with $H_1^*-H_2^*$.

The methods and results presented in this paper may contribute to a better understanding of speech production and may be useful for applications such as speech synthesis, speech recognition, and speaker identification.

Future work will study the effects of context, prosody, stress, and accentedness on acoustic measures related to source parameters.

ACKNOWLEDGMENTS

We thank Dr. Story and two anonymous reviewers for their helpful suggestions. This material is based upon work supported by NSF Grant No. 0326214 and by a Radcliffe Fellowship to Abeer Alwan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

APPENDIX: DERIVATION OF THE CORRECTION FORMULA

The derivation of the spectral magnitude formant correction formula presented in this appendix is based on the linear source-filter model of speech production (Fant, 1960). Assuming a vocal tract all-pole model, the normalized transfer function $T(s)$ with N formants can be written as

$$T(s)=\prod_{i=1}^{N}\frac{\sigma_i^2+\Omega_i^2}{(s-(\sigma_i+j\Omega_i))(s-(\sigma_i-j\Omega_i))}. \tag{A1}$$

The numerator is chosen such that $T(s=0)=1$. $s_i=\sigma_i+j\Omega_i$, $\sigma_i=-\pi B_i$, $\Omega_i=2\pi F_i$, where B_i and F_i are the ith formant bandwidth and frequency, respectively.

Assuming that the axis $s=j\Omega$ lies in the region of convergence (ROC), the Fourier transform of the magnitude of Eq. (A1) becomes

$$|T(j\Omega)|=\prod_{i=1}^{N}\left|\frac{\sigma_i^2+\Omega_i^2}{\sigma_i^2+\Omega_i^2-\Omega^2+j2\sigma_i\Omega}\right|,$$

$$|T(j\Omega)|^2=\prod_{i=1}^{N}\frac{(\sigma_i^2+\Omega_i^2)^2}{(\sigma_i^2+\Omega_i^2-\Omega^2)^2+(2\sigma_i\Omega)^2}.$$

Using the definitions of σ_i and Ω_i produces

$$|T(f)|^2=\prod_{i=1}^{N}\frac{(\pi^2B_i^2+4\pi^2F_i^2)^2}{(\pi^2B_i^2+4\pi^2F_i^2-4\pi^2f^2)^2+16\pi^4B_i^2f^2}.$$

Finally, the total contribution of N formants to the vocal tract power spectrum at frequency f is

$$|T(f)|^2=\prod_{i=1}^{N}\frac{((B_i/2)^2+F_i^2)^2}{((B_i/2)^2+F_i^2-f^2)^2+B_i^2f^2}. \tag{A2}$$

Note: For $B_i \ll F_i$ the terms $(B_i/2)^2$ can be neglected (Fant, 1995). In this paper, however, we will account for these terms.

The aforementioned analysis was done in the continuous frequency domain. For sampled signals (sampling frequency F_s) the contribution of N formants to the vocal tract transfer function can be written in the z domain as

$$T(z) = \prod_{i=1}^{N} \frac{1 - 2\Re(z_i) + |z_i|^2}{(z - z_i)(z - z_i^*)}, \tag{A3}$$

where $T(z)$ is normalized so that $|T(z=1)|=1$. $z_i = r_i e^{j\omega_i}$ with $\omega_i = 2\pi F_i / F_s$.

Assuming that the unit circle $z=e^{j\omega}$ lies in the ROC, the Fourier transform of the squared magnitude of Eq. (A3) becomes

$$|T(\omega)|^2 = \prod_{i=1}^{N} \frac{(1 - 2r_i \cos(\omega_i) + r_i^2)^2}{(1 - 2r_i \cos(\omega - \omega_i) + r_i^2)(1 - 2r_i \cos(\omega + \omega_i) + r_i^2)}, \tag{A4}$$

with $r_i = e^{-\pi B_i / F_s}$ and $\omega_i = 2\pi F_i / F_s$.

Equation (A4) specifies the amount by which the spectral magnitude at a particular frequency, ω, is boosted by the effects of formants located at frequencies ω_i. Therefore, to obtain the source spectral magnitudes, the effects of the formants need to be subtracted from the magnitudes of the speech spectrum. For example (Iseli and Alwan, 2004),

$$H^*(\omega) = H(\omega) - \sum_{i=1}^{N} 10 \log_{10} \frac{(1 - 2r_i \cos(\omega_i) + r_i^2)^2}{(1 - 2r_i \cos(\omega + \omega_i) + r_i^2)(1 - 2r_i \cos(\omega - \omega_i) + r_i^2)}, \tag{A5}$$

where $H(\omega)$ is the magnitude of the original signal spectrum (in dB) at frequency ω, N is the number of formants, and $H^*(\omega)$ is the corrected magnitude (i.e., the magnitude of the source spectrum) at frequency ω. Note that for $B_i=0$ and $\omega=\omega_i$ this formula is undefined.

Ananthapadmanabha, T. V. (**1984**). "Acoustic analysis of voice source dynamics," STL-QPSR **25**, 1–24.

Baken, R. J. (**1987**). *Clinical Measurement of Speech and Voice* (Taylor and Francis, London).

Doval, B., and d'Alessandro, C. (**1999**). "The spectrum of glottal flow models," Technical Report, LIMSI-CNRS, Orsay, France.

El-Jaroudi, A., and Makhoul, J. (**1991**). "Discrete all-pole modeling," IEEE Trans. Signal Process. **39**, 411–423.

Esposito, C. (**2005**). "An acoustic and electroglottographic study of phonation in Santa Ana del Valle Zapotec," Poster at the 79th Meeting of the Linguistic Society of America, 2005.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, Paris).

Fant, G. (**1982**). "The voice source-acoustic modeling," STL-QPSR **23**, 28–48.

Fant, G. (**1995**). "The LF model revisited. Transformations and frequency domain analysis," STL-QPSR **36**, 119–156.

Fant, G., and Kruckenberg, A. (**1996**). "Voice source properties of speech code," TMH-QPSR **37**, 45–56.

Fant, G., Kruckenberg, A., Liljencrants, J., and Hertegård, S. (**2000**). "Acoustic-phonetic studies of prominence in Swedish," TMH-QPSR **41**, 1–52.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). "A four-parameter model of glottal flow," STL-QPSR **26**, 1–13.

Fröhlich, M., Michaelis, D., and Strube, H. W. (**2001**). "Sim-Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," J. Acoust. Soc. Am. **110**, 479–488.

Hanson, H. M. (**1995**). "Glottal characteristics of female speakers," Ph.D. dissertation, Harvard University, Cambridge, MA.

Hanson, H. M. (**1997**). "Glottal characteristics of female speakers: Acoustic correlates," J. Acoust. Soc. Am. **101**, 466–481.

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**, 1064–1077.

Hedelin, P. (**1984**). "A glottal LPC-vocoder," in Proc. IEEE 1.6.1–1.6.4.

Henrich, N., d'Alessandro, C., and Doval, B. (**2001**). "Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data," in Proceedings of EUROSPEECH, Scandinavia, pp. 47–50.

Hertegård, S., and Gauffin, J. (**1992**). "Acoustic properties of the Rothenberg mask," STL-QPSR **33**, 9–18.

Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P., and Goldman, S. L. (**1995**). "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," J. Speech Hear. Res. **38**, 1212–1223.

Holmes, J. N. (**1973**). "Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," IEEE Trans. Audio Electroacoust. 298–305.

Iseli, M., and Alwan, A. (**2004**). "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in Proceedings of ICASSP, Montreal, Canada, Vol. **1**, pp. 669–672.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Koreman, J. (**1996**). "Decoding linguistic information in the glottal airflow," Ph.D. thesis, University of Nijmegen.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.

Lehiste, I., and Peterson, G. E. (**1961**). "Some basic considerations in the analysis of intonation," J. Acoust. Soc. Am. **33**, 419–425.

Maddieson, I., and Ladefoged, P. (**1985**). "Tense and lax in four minority languages of China," J. Phonetics **13**, 433–454.

Mannell, R. H. (**1998**). "Formant diphone parameter extraction utilising a labelled single speaker database," in *Proceedings of the ICSLP* (ASSTA, Sydney, Australia), Vol. **5**, pp. 2003–2006.

Marasek, K. (**1996**). "Glottal correlates of the word stress and the tense-lax opposition in German," in Proceedings ICSLP, Philadelphia, PA, pp. 1573–1576.

Markel, J. D., and Gray, A. H., Jr. (**1976**). *Linear Prediction of Speech* (Springer, New York).

Mártony, J. (**1965**). "Studies of the voice source," STL-QPSR **6**, 4–9.

Miller, J., Lee, S., Uchanski, R., Heidbreder, A., Richman, B., and Tadlock, J. (**1996**). "Creation of two children's speech databases," in Proceedings of ICASSP, Vol. **2**, pp. 849–852.

Miller, R. L. (**1959**). "Nature of the vocal cord wave," J. Acoust. Soc. Am. **31**, 667–677.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice Hall, Englewood Cliffs, NJ).

Rosenberg, A. E. (**1971**). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49**, 583–590.

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., and Manley, H. (**1974**). "Average magnitude difference function pitch extractor," IEEE Trans. Acoust., Speech, Signal Process. **22**, 353–362.

Rothenberg, M. (**1973**). "A new inverse-filtering technique for deriving the glottal airflow during voicing," J. Acoust. Soc. Am. **53**, 1632–1645.

Sjölander, K. (**2004**). "Snack sound toolkit," KTH Stockholm, Sweden, http://www.speech.kth.se/snack/ (last viewed January 2007).

Sluijter, A., and Van Heuven, V. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress," J. Acoust. Soc. Am. **100**, 2471–2485.
Sluijter, A., Van Heuven, V., and Pacilly, J. (**1997**). "Spectral balance as a cue in the perception of linguistic stress," J. Acoust. Soc. Am. **101**, 503–513.
Swerts, M., and Veldhuis, R. (**2001**). "The effect of speech melody on voice quality," Speech Commun. **33**, 297–303.
Titze, I. R. (**2004**). "A theoretical study of f0-f1 interaction with application to resonant speaking and singing voice," J. Voice **18**, 292–298.
Wakita, H. (**1977**). "Normalization of vowels by vocal-tract length and its application to vowel identification," IEEE Trans. Acoust., Speech, Signal Process. **25**, 183–192.