Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation[a]

Jody Kreiman[b]
Department of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, California 90095-1795

Yen-Liang Shue
Dolby Australia, Level 16, 233 Castlereagh Street, Sydney, NSW 2000 Australia

Gang Chen
Department of Electrical Engineering, UCLA, 66-147 G Engineering IV, Los Angeles, California 90095-1594

Markus Iseli
UCLA Graduate School of Education, 1400 Ueberroth Building, Los Angeles, California 90095-7150

Bruce R. Gerratt and Juergen Neubauer
Department of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, California 90095-1795

Abeer Alwan
Department of Electrical Engineering, UCLA, 66-147 G Engineering IV, Los Angeles, California 90095-1594

(Received 14 February 2011; revised 23 July 2012; accepted 30 July 2012)

Increases in open quotient are widely assumed to cause changes in the amplitude of the first harmonic relative to the second (H1*–H2*), which in turn correspond to increases in perceived vocal breathiness. Empirical support for these assumptions is rather limited, and reported relationships among these three descriptive levels have been variable. This study examined the empirical relationship among H1*–H2*, the glottal open quotient (OQ), and glottal area waveform skewness, measured synchronously from audio recordings and high-speed video images of the larynges of six phonetically knowledgeable, vocally healthy speakers who varied fundamental frequency and voice qualities quasi-orthogonally. Across speakers and voice qualities, OQ, the asymmetry coefficient, and fundamental frequency accounted for an average of 74% of the variance in H1*–H2*. However, analyses of individual speakers showed large differences in the strategies used to produce the same intended voice qualities. Thus, H1*–H2* can be predicted with good overall accuracy, but its relationship to phonatory characteristics appears to be speaker dependent.
[http://dx.doi.org/10.1121/1.4747007]

PACS number(s): 43.70.Gr [CHS] Pages: 2625–2632

I. INTRODUCTION

Studies of voice typically focus on either production or perception, but not usually both. In production studies, investigators image or otherwise measure movements of the phonatory apparatus, or approximate them with physical or computational models, while in perception research investigators seek to measure what listeners perceive, or to uncover acoustic correlates of particular vocal qualities. Few studies have examined the perceptual consequences of changes in glottal vibratory patterns or the physical precursors of changes in perceived quality. One partial exception is the relationship between the relative duration of the open part of the glottal vibratory cycle [the open quotient (OQ)] and a quality dimension ranging from "pressed" to "breathy" (e.g., Klatt and Klatt, 1990). An increase in OQ is widely assumed to correspond to an increase in breathiness. This relationship between OQ in the physical realm and perceived quality is further assumed to occur because of changes in the relative amplitudes of the first two harmonics of the voice source, denoted H1–H2, or H1*–H2* when harmonic amplitudes are measured from the audio signal recorded at the mouth and then corrected for the effects of vocal tract resonances (Hanson, 1997; Iseli *et al.*, 2007).[1] As OQ increases, energy in the first harmonic (and thus H1*–H2*) is assumed to increase, and this increase is the presumptive cause of the change in vocal quality (e.g., Klatt and Klatt, 1990). Thus, the relationship between H1*–H2* and OQ potentially provides a linkage between descriptive levels along the "speech chain."

Reasonable empirical support exists for the relationship between changes in H1*–H2* and changes in perceived

[a] Preliminary versions of this paper were presented at the International Conference on Voice Physiology and Biomechanics (ICVPB), Tampere, Finland, August 2008, and at the 156th Meeting of the Acoustical Society of America, Miami, FL, November 2008.

[b] Author to whom correspondence should be addressed. Electronic mail: jkreiman@ucla.edu

quality (e.g., Hillenbrand and Houde, 1996; cf. Kreiman et al., 2010b, who found that listeners could tell voices apart—i.e., the voices differed in quality (ANSI, 1960)—as H1–H2 changed in very small steps). The case is less clear for the link between amplitude at the first harmonic frequency (H1) and OQ. The longer the vocal folds remain open, the more closely matched the open phase becomes to the period, leading to a stronger fundamental component in the signal spectrum (assuming all other influences, including pulse skewness, are constant) (Fant, 1995). However, empirical support for this relationship is rather limited. Most data come from electroglottographic (EGG) or inverse filtering studies of small groups of speakers, and reported relationships are not in general especially strong. In the most frequently cited study, Holmberg et al. (1995) correlated harmonic amplitudes, estimated from acoustic spectra and then log transformed, with EGG- and airflow-based measures of the adduction quotient (defined as vocal fold contact time/period, or 1-OQ) for 20 female speakers. The two sets of adduction quotient measures were only modestly correlated ($r = 0.57$), and did not strongly predict H1*–H2* measures ($r = -0.46$ for EGG measures, and $r = -0.69$ for airflow data, indicating less than 50% shared variance). Similarly, Swerts and Veldhuis (2001) inverse-filtered four tokens of the vowel /a/ recorded from seven speakers, fitted a Liljencrants-Fant (LF) source model (Fant et al., 1985) to the data, and then measured OQ and glottal pulse skewing (the LF model parameter RK) from the best-fitting model. H1–H2 and OQ were positively correlated for 17/28 tokens, negatively correlated for 4/28 tokens, and uncorrelated for 7/28 tokens (precise correlation values are not reported). [See Huffman (1987) and Sundberg et al. (1999), for additional examples.] The same study also reported that while $F0$ is positively correlated overall with H1–H2, correlations for individual speakers are variable, with 18 cases positively correlated, 6 negatively correlated, and 4 with zero correlation (Swerts and Veldhuis, 2001).

Many more studies have examined the relationship between harmonic amplitudes and glottal configuration in the context of models of the voice source (e.g., Fant, 1995, 1997; Fant et al., 1985; Klatt and Klatt, 1990; Fujisaki and Ljungqvist, 1986; Rosenberg, 1971). Several parametric source models with varying complexities have been proposed, most of which model the shape of the glottal airflow or its derivative in the time domain, based on observations from airflow masks, electroglottographs, mechanical systems, and/or inverse filtering of speech signals. [Frequency-domain representations exist for only a subset of source models; see, e.g., Fant (1995) or Doval et al. (2006)]. The nature and extent of the relationship that exists between H1–H2 and OQ varies across models. For example, in the LF model the relationship is expressed as $\text{H1–H2} = -6 + 0.27\exp(0.055\ \text{OQ})$ [Fant (1995), although Sundberg et al. (1999) report that Fant later stated that the fit of a linear equation to the data was nearly equivalent]. In contrast, in the simpler KGLOTT88 source model H1–H2 is perfectly correlated with OQ (Klatt and Klatt, 1990). The relationship between H1–H2 and OQ has also been shown to vary *within* a single source model. Several authors (e.g., Swerts and Veldhuis, 2001; Doval and D'Alessandro, 1997; Fant, 1997; Henrich et al., 2001) have demonstrated that in the LF (Fant et al., 1985) and R++ source models (Veldhuis, 1998), the relationship between H1–H2 and OQ depends on the extent to which the modeled glottal pulse is symmetrical or asymmetrical. For example, Henrich et al. (2001) showed that in the LF model H1–H2 is minimally affected by pulse skewness when OQ is small; but as OQ increases, the influence of skewness increases as well, so that the range of possible H1–H2 values is more than 6 dB when $\text{OQ} = 0.9$; the precise value depends on the extent of pulse skewness. The R++ model shows a similar effect (Swerts and Veldhuis, 2001). Analyses of acoustic data (Swerts and Veldhuis, 2001) confirmed the relationship between H1–H2 and pulse skewness, which was a better predictor of H1–H2 than OQ was (25/28 tokens positively correlated). In contrast, as noted previously, H1–H2 is perfectly correlated with OQ in the KLGLOTT88 source model (Klatt and Klatt, 1990), in which pulse skewness is a constant (Hanson, 1997; Henrich et al., 2001). Finally, current source models are limited in their ability to match estimated pulse shapes from a broad range of speakers and phonatory modes (Henrich et al., 2001), so that the existence (or lack) of a relationship in the context of a particular model does not necessarily imply that the same relationship will exist in other models or in natural data.

Finally, definitions of OQ vary from study to study, because of differences in the manner of treating cases in which the glottis never closes fully. Such cases may be assigned an OQ of 100% (Hirano and Bless, 1993); OQ may be calculated using the most-closed phase of the glottal cycle (as in the present work); or the case may simply be omitted from analyses (e.g., Fex et al., 1991). Because most women produce phonation with a persistent glottal gap (e.g., Södersten and Lindestad, 1990), these differences produce substantial variance in the range and pattern of glottal pulse shapes examined from study to study (e.g., Fant, 1995; Klatt and Klatt, 1990; Holmberg et al., 1995; Hanson, 1997; Fex et al., 1991).

In summary, despite the insights that modeling studies have provided regarding the relationship between some aspects of pulse shapes and acoustic attributes, questions remain about the relationships that exist in natural data, so that it is difficult to assess the adequacy of different source models for explaining linkages between production and perception in voice. Studies to date suggest that the relationship between H1–H2 and OQ may be variable, but model fit to empirical pulse shapes is not such that definitive conclusions can be drawn. Because existing empirical data are not sufficient to clarify this situation, this study examined the relationship between H1*–H2* (measured from recorded acoustic signals), OQ, and the asymmetry coefficient (the length of the opening phase relative to the open phase, e.g., Henrich et al., 2001; Shue and Alwan, 2010),[2] measured synchronously from high-speed video images of the vibrating vocal folds. Note that previous work on this topic has used models of the glottal flow, which may differ in pulse skewness from the glottal area functions measured here (e.g., Howe and McGowan, 2007). However, these differences in skewness should not affect measures of OQ, which depend on glottal opening and closing instants, and not on precise pulse shapes. By gathering multiple tokens from male and female speakers who varied $F0$ and voice quality quasi-orthogonally, it is possible to compare measures across a range of glottal area waveform shapes.

II. EXPERIMENT 1

A. Methods

1. Subjects and recording procedures

Six phonetically knowledgeable speakers (three female and three male) with perceptually normal voices (as assessed by a speech-language pathologist) participated in this experiment. They were asked to sustain the vowel /i/ (Draper *et al.*, 2007) for approximately 10 s while holding voice quality, $F0$, and loudness as steady as possible, although vowel quality ranged from /I/ to approximately cardinal vowel /ε/ due to speaker response to the positioning of the laryngoscope in the mouth. Across tokens, the speakers were directed to vary $F0$ (low, normal, and high) and voice qualities (pressed, normal/modal, and breathy) quasi-orthogonally, resulting in a minimum of nine tokens from each speaker. Because the purpose of the quality and pitch variations was simply to generate a variety of glottal configurations, no attempt was made to ensure that they were produced in the same manner across speakers.

High-speed video images of the vocal folds were recorded during each utterance at 3000 frames/s (512 × 512 pixels resolution). A 70° rigid laryngoscope (KayPentax, Lincoln Park, NJ) with a 300 W xenon light source (KayPentax, Lincoln Park, NJ) and a FASTCAM-ultima APX camera (Photron Ltd., San Diego) were used to capture the images. Audio recordings were collected synchronously with the high-speed imaging. Synchronization was managed by the signal acquisition software. Voice signals were recorded with a Brüel & Kjær microphone (1.27 cm [1/2 in.] diameter; type 4193 -L-004), held approximately 7 cm from the corner of the speaker's mouth, and were directly digitized at a sampling rate of 60 kHz (conditioning amplifier: NEXUS 2690, Brüel & Kjær, Denmark; bandpass filtering of microphone signal between 20 Hz and 22.4 kHz; analog-to-digital converter: voltage resolution 16 bits, input range ± 5 V). One second of phonation was excerpted from the most auditorily stable and representative portion of each token for subsequent analysis.

2. OQ and asymmetry coefficient calculations

Glottal events for each cycle were identified via frame-by-frame examination of the high-speed images. For each token, in cases when glottal closure was complete, the first instants of glottal opening and closing were identified and marked by hand by the first author, with an approximate interpolated resolution of 1/2 frame (0.17 ms). When the glottis did not close completely, the moment when glottal area began to increase and the onset of maximum closure were treated as opening and closing instants, respectively. For each individual cycle of phonation, OQ was calculated as the time from the first opening instant to the onset of maximum closure, divided by the time from the opening instant to the opening instant of the following cycle. OQ equaled 1 only when the glottis never closed at all anywhere along its length. To reduce measurement errors (which are assumed to be random), OQ values were averaged over 100 ms windows, yielding ten estimates of OQ for each 1 s sustained utterance.

Glottal area waveform skewness was calculated for each cycle of phonation in the first 150 images of the glottal area waveform data. Glottal area waveforms were generated from high-speed images using custom software incorporating a series of edge-detection and region-growing algorithms. Factors including random noise in images, variations in contrast levels, and multiple glottal gaps made area estimation from complete 1 s recordings impractical; hence our decision to limit these analyses to only 150 frames, which could be checked by hand and manually adjusted if necessary for accuracy. After area calculations, the first instants of glottal opening, the instants of maximum opening, and the onsets of maximum closure were located, and the asymmetry coefficient was calculated.[2]

3. Acoustic measurements

Audio recordings were low-pass filtered at 8 kHz and downsampled to 16 kHz prior to acoustic analysis. H1*–H2* measures were extracted automatically from the audio signals using $F0$ values obtained from the STRAIGHT algorithm (Kawahara *et al.*, 1999). Harmonic amplitudes were calculated pitch synchronously using VOICE SAUCE software (Shue, 2009), with an analysis window of eight periods with a 1 ms shift, corresponding to the resolution of the STRAIGHT output. Values were not averaged across windows. The harmonic magnitudes were then corrected for the effects of the first two formant frequencies (and their respective bandwidths, estimated for the entire 1 s window as described next) using the formula in Iseli *et al.* (2007). Output was aligned with the 50 ms window from the imaging signal for subsequent statistical analysis.

Linear predictive coding (LPC) analysis systems produced significant errors in formant frequency estimation, especially when the voice source was near-sinusoidal, when a persistent glottal gap was present, and/or when $F0$ was high. These factors, singly or in combination, increased the prominence of H1 and led to its misidentification as $F1$. For this reason, formant frequencies and bandwidths were derived through analysis-by-synthesis. Using the UCLA voice synthesizer (Kreiman *et al.*, 2010a), the original voice samples were copy-synthesized using the following procedure. Because apparent harmonic amplitudes can vary depending on formant frequencies and bandwidths as well as source characteristics, vowel quality was first matched to the target, before the source was altered. Next, because LPC-based analysis techniques typically yielded variable (and sometimes unrealistic) bandwidth values, all bandwidths were calculated using the formant frequency-to-bandwidth mapping function given by Hawks and Miller (1995) when OQ (measured from the high-speed images) was ≤ 0.7.

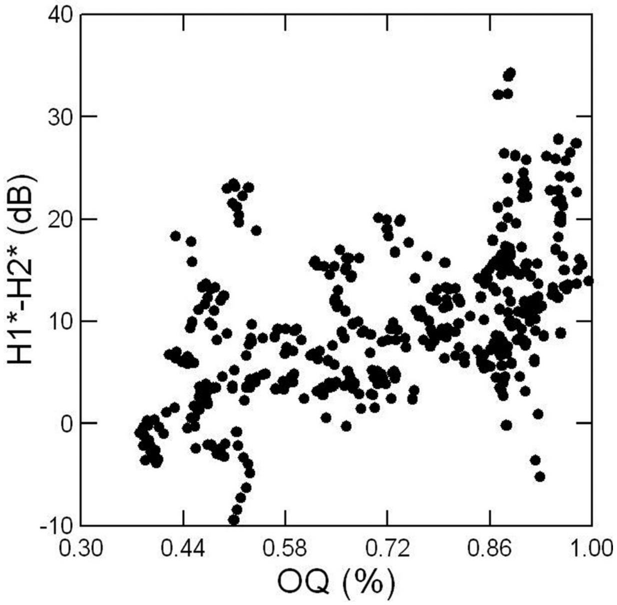

FIG. 1. H1*–H2* (in dB) vs open quotient (in percent), for all speakers and utterances in experiment 1.

When OQ exceeded 0.7, an additional “open glottis” correction was applied. In these cases, following a suggestion by Stevens (1998), calculated $B1$ was multiplied by a factor ranging from 1 (when $\mathrm{OQ} = 0.7$) to 3 (when $\mathrm{OQ} = 1.0$). Finally, the amplitudes of the first two harmonics were adjusted to match the original target voice. Once the synthesized vowel matched the original recorded sample in vowel quality and voice quality (so that in a pilot study using a same/different task they were perceptually indistinguishable to the first author, a phonetician), and the spectra of the original and synthetic tokens matched precisely as well, formant frequencies and bandwidths were recorded and H1*–H2* was recalculated using these values.[3] Finally, H1*–H2* values were smoothed over time by fitting Legendre polynomials of degree three to the raw H1*–H2* values to reduce the impact of occasional spurious values that arose in utterances with high noise levels and weak harmonic structures (e.g., breathy, high-pitched tokens).

B. Results

Cycle marking from images and audio files was validated by examining the correlation between $F0$ measures for the two data sets. Values agreed almost perfectly across speakers and tokens ($r = 0.997$, $p < 0.01$). In the case of a single token for speaker 2, acoustic $F0$ values showed a period-doubling error. These data were excluded from all subsequent analyses.

Figure 1 shows the relationship between H1*–H2* and OQ for the complete subject group. These measures were significantly but not strongly correlated ($r = 0.5$, $p < 0.01$).

As noted previously, no effort was made to ensure that the different speakers produced breathy, modal, and/or pressed phonation in comparable ways, so that (for example) one person’s modal phonation might resemble another’s breathy or pressed. Despite this, previous studies relating H1*–H2* and OQ to a voice quality continuum from breathy to pressed predict that both these measures should decrease with changes in quality along this continuum. Separate two-way analyses of variance tested this hypothesis for H1*–H2* and OQ.[4] A significant main effect of voice quality on H1*–H2* was observed [$F(2, 428) = 121.50$, $p < 0.01$]. However, a significant interaction between speaker and quality [$F(10, 428) = 19.03$, $p < 0.01$] also occurred, because no speaker showed the entire predicted pattern of pairwise differences in H1*–H2* (breathy > modal > pressed) across qualities [*post hoc* Tukey tests; $p < 0.01$; Table I(a)]. OQ also varied significantly with vocal quality [$F(2, 428) = 534.92$, $p < 0.01$), and interacted significantly with speaker [$F(10, 428) = 36.58$, $p < 0.01$]. Results of *post hoc* Tukey comparisons for this interaction ($p < 0.01$) are given in Table I(b). Note that for speaker 6, OQ for pressed phonation was significantly *greater* than that for modal phonation ($p < 0.01$), contrary to predictions.

Figure 2 illustrates the differences among speakers that underlie these variable relationships and the modest overall correlations between H1*–H2* and OQ. As Fig. 2 shows, despite some outliers, H1*–H2* was approximately linearly related to OQ for speakers 1, 2, and 5. The relationship was categorical for speaker 6 (and possibly for speaker 5 as

TABLE I. (a) H1*–H2* and (b) open quotient for each target voice quality for the six individual speakers in experiment 1. Standard deviations are given parenthetically. Statistically significant differences between qualities ($p < 0.01$) are indicated with greater than (>) and less than (<) symbols.

Speaker	Breathy vs modal	Breathy vs pressed	Modal vs pressed
(a) H1*–H2* (dB)			
1	9.92 (5.81) = 13.23 (5.42)	9.92 (5.81) > 1.21 (4.24)	13.23 (5.42) > 1.21 (4.24)
2	11.01 (2.75) > 3.80 (6.25)	11.01 (2.75) = 7.29 (1.26)	3.80 (6.25) = 7.29 (1.26)
3	14.57 (8.12) = 13.07 (2.20)	14.57 (8.12) = 12.16 (3.29)	13.07 (2.20) = 12.16 (3.29)
4	20.64 (5.02) > 5.08 (1.35)	20.64 (5.02) > 3.71 (0.66)	5.08 (1.35) = 3.71 (0.66)
5	10.33 (6.25) = 6.82 (2.72)	10.33 (6.25) > −2.12 (3.79)	6.82 (2.72) > −2.12 (3.79)
6	20.96 (7.09) > 9.06 (4.88)	20.96 (7.09) > 12.57 (6.85)	9.06 (4.88) = 12.57 (6.85)
(b) Open quotient (%)			
1	0.94 (0.02) > 0.73 (0.12)	0.94 (0.02) > 0.44 (0.05)	0.73 (0.12) > 0.44 (0.05)
2	0.88 (0.08) > 0.66 (0.14)	0.88 (0.08) > 0.69 (0.14)	0.66 (0.14) = 0.69 (0.14)
3	0.93 (0.03) > 0.82 (0.11)	0.93 (0.03) > 0.67 (0.04)	0.82 (0.11) > 0.67 (0.04)
4	0.91 (0.04) > 0.80 (0.10)	0.91 (0.04) > 0.57 (0.05)	0.80 (0.10) = 0.57 (0.05)
5	0.89 (0.07) > 0.79 (0.07)	0.89 (0.07) > 0.56 (0.08)	0.79 (0.07) > 0.56 (0.08)
6	0.89 (0.02) > 0.47 (0.02)	0.89 (0.02) > 0.59 (0.16)	0.47 (0.02) < 0.59 (0.16)

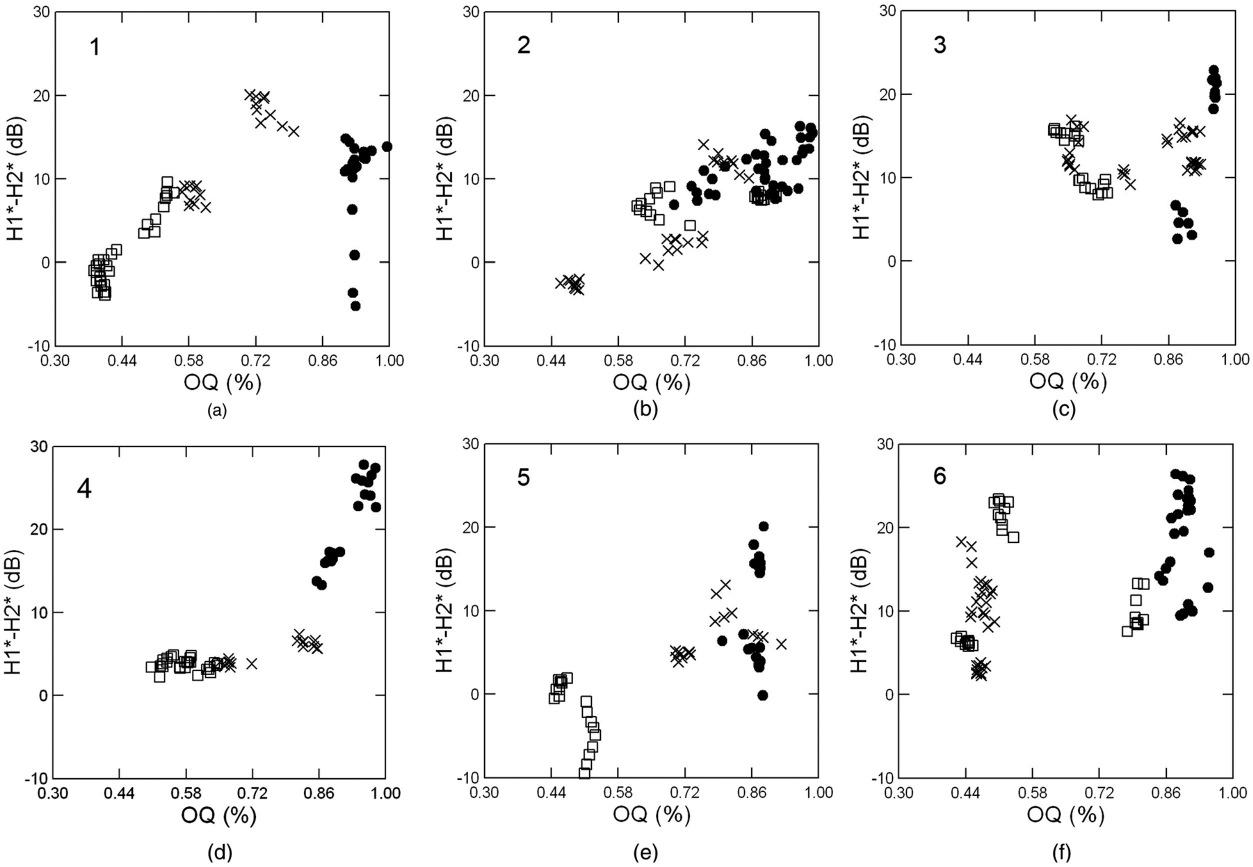

FIG. 2. H1*–H2* (in dB) vs open quotient for the six individual speakers in experiment 1. Speakers 1–3 were females; speakers 4–6 were males. Breathy utterances are plotted with filled circles, modal phonation by crosses, and pressed phonation by open squares.

well). For speaker 4, H1*–H2* varied with OQ only when OQ exceeded about 0.8; and no significant relationship between variables was observed for speaker 3 (Table II).

Next, multiple linear regression was applied to examine the extent to which OQ, glottal area waveform skewness (measured by the asymmetry coefficient), and $F0$ contributed jointly to predicting H1*–H2* values. Because theoretical work by Henrich *et al.* (2001) suggests that the importance of the asymmetry coefficient in predicting H1*–H2* should vary with OQ, three regression models were fitted to each speaker's data: One including phonatory cycles with OQ less than an individually determined cutpoint, one including cycles with OQ greater than this cutpoint, and one including all data. Cutpoints corresponded to observed gaps in the distribution of OQ values for that speaker (OQ = 0.7, 0.65, 0.82, 0.75, 0.65, and 0.7 for speakers 1–6, respectively; see Fig. 2). These analyses included only data corresponding to the first 150 images of each utterance, because the asymmetry coefficient was only measured for this interval.

Results of these regressions are shown in Table III, which lists the standardized regression coefficients and R^2 values for each speaker. Regression coefficients reflect the relative importance of the different factors in predicting H1*–H2* in that analysis. As Table III shows, speakers differed substantially in the extent to which each variable contributed to the prediction of H1*–H2* values. Speakers fell roughly into three groups: Those for whom H1*–H2* was best predicted by a weighted sum of OQ + $F0$ (speakers 1, 3, and 5), those for whom OQ alone provided the best predictive model (speakers 2 and 4), and one speaker for whom a weighted sum of the asymmetry coefficient + $F0$ provided the best model (speaker 6). OQ was the most important predictor for speakers 2 and 4, but was not significantly associated with H1*–H2* for speaker 6; and $F0$ was the most important predictor for speakers 3 and 6, but was not a significant predictor for speaker 4. Finally, the theoretical prediction that asymmetry coefficients would be more strongly predictive of H1*–H2* when OQ was large was true for speakers 1, 3, and 6, but not for speakers 2, 4, and 5.

TABLE II. Correlations between H1*–H2* and open quotient across all utterances for the six individual speakers in experiment 1.[a]

Speaker	1	2	3	4	5	6	Combined group
Correlation	0.65*	0.83*	0.20	0.85*	0.75*	0.53*	0.50*

[a]Values marked with an asterisk (*) are significant at $p < 0.05$.

C. Discussion

These findings are consistent with theoretical work (e.g., Henrich *et al.*, 2001; Swerts and Veldhuis, 2001) showing that H1*–H2* variations cannot in general be predicted by OQ alone. Contrary to predictions, however, the role of pulse asymmetry varied from speaker to speaker, and $F0$ proved the most important predictor of H1*–H2* for two of our six speakers. These results suggest that speakers have several different strategies available to them when varying voice

TABLE III. Standardized regression coefficients and R^2 values for multiple linear regression analyses relating open quotient (OQ), asymmetry coefficients, and $F0$ to H1*–H2*. Columns show coefficients for separate analyses including cases with OQ less than the cutpoint (OQ small), greater than the cutpoint (OQ large), and the complete set of data for that speaker. Coefficients in cells marked with an em dash (—) were not statistically significant. All other coefficients are significant at $p < 0.05$.

	OQ			Asymmetry coefficient			$F0$			R^2		
Speaker	OQ small	OQ large	Combined data	OQ small	OQ large	Combined data	OQ small	OQ large	Combined data	Below cutpoint	Above cutpoint	Complete data set
1	0.40	−0.59	—	0.10	0.29	0.44	0.67	0.29	0.76	0.96	0.93	0.69
2	0.76	−0.43	0.77	0.25	0.26	0.18	0.18	−0.32	0.13	0.84	0.24	0.64
3	−0.33	0.31	—	0.08	—	—	0.81	−0.82	—	0.95	0.91	—
4	−0.74	0.67	0.50	—	−0.31	−0.52	—	—	—	0.52	0.85	0.81
5	−0.16	0.75	0.58	0.07	—	−0.21	−0.84	—	−0.24	0.98	0.54	0.75
6	—	—	0.36	−0.32	0.77	−0.36	0.65	0.77	—	0.48	0.68	0.35

quality, and that the relationships between specific attributes of the source spectrum and the voice production process are more complex than sometimes assumed. We return to this topic in the general discussion to follow.

This experiment raised a number of methodological issues relevant to the study of spectral attributes of the voice source. In particular, estimation of H1*–H2* was highly sensitive to errors in formant and bandwidth estimation. Current LPC-based formant estimation methods cannot consistently detect $F1$ correctly if strong harmonics are present at low frequencies. In addition, when $F1$ is near H1, H1*–H2* cannot be unambiguously determined without knowledge of bandwidths, but interactions between the source and vocal tract make it difficult to estimate bandwidths without knowledge of glottal configuration. This issue is further complicated by the lack of data regarding bandwidths during open glottis conditions, which also limits modeling efforts.

In this study, we addressed these problems pragmatically by verifying formant frequencies using an analysis-by-synthesis procedure that constrained the order in which steps were undertaken. Although this process was theoretically motivated and resulted in consistent measurements, ambiguities remain in measured H1* and H2* values, particularly when the glottis is open (as it often is in female or non-modal speech). Empirical data regarding bandwidth modulations when the glottis is open are required to further clarify the relationships among the variables studied in this experiment.

In summary, despite these measurement issues, the present data suggest that listeners use a variety of strategies to control H1–H2 (or H1*–H2*), so that the particular correlates vary across speakers and productions. However, in this experiment, speakers varied voice quality statically across tokens, and not within tokens. The demands of changing from one phonatory state to another differ from those of initiating a particular kind of phonation from rest and then sustaining it, in ways that might affect the manner in which phonation is controlled. To examine how variables co-vary during dynamic changes in voice quality, in experiment 2 we measured H1*–H2*, asymmetry coefficients, and OQ in a speaker who varied voice quality from breathy to pressed across a single utterance.

III. EXPERIMENT 2

A. Methods

Speaker 1 from experiment 1 (a female phonetician who is experienced at manipulating voice quality) participated in this experiment. Over the course of a 3.3 s utterance, she changed phonation slowly from breathy to pressed while holding $F0$ and vowel quality as steady as possible. High-speed images and audio signals were recorded synchronously and analyzed for this utterance in the manner described previously, except that asymmetry coefficients were calculated for the entire duration of the utterance.

B. Results and discussion

Figure 3 shows how OQ changed over time for this speaker and utterance, along with selected frames from the corresponding high-speed images, and Fig. 4 shows how H1*–H2* and the asymmetry coefficient varied with OQ over time. As shown in Fig. 3(b), the speaker phonated with a persistent glottal gap for approximately the first 2/3 of the utterance (time 0–2.3 s). During this portion of the utterance, OQ changed relatively little with changing quality, but the size of the glottal gap gradually decreased, as shown in the first three glottal images in Fig. 3(b). At the same time, glottal area skewness increased across this segment of phonation (Fig. 4). When the glottal gap during the closed phase disappeared (as shown in the last two glottal images in Fig. 3(b)], the speaker adjusted OQ more markedly as vocal "pressedness" continued to increase from time 2.4 to 3.3 s. Glottal area skewness decreased sharply at this point (Fig. 4). Regression analyses showed that in the presence of a glottal gap (time 0–2.3 s, OQ = 0.91–0.85), H1*–H2* was best predicted by the asymmetry coefficient, with no significant contribution of OQ [Table IV; $F(3, 19) = 15.45$, $p < 0.05$; $R^2 = 0.66$]. In the absence of a glottal gap (time 2.4–3.3 s, OQ = 0.84–0.58), H1*–H2* was best predicted by OQ, with skewness making no significant contribution to prediction [$F(3, 6) = 4.62$, $p < 0.05$; $R^2 = 0.55$].[5]

IV. GENERAL DISCUSSION

The data in these two experiments underscore the dynamic interplay between OQ, the asymmetry coefficient,

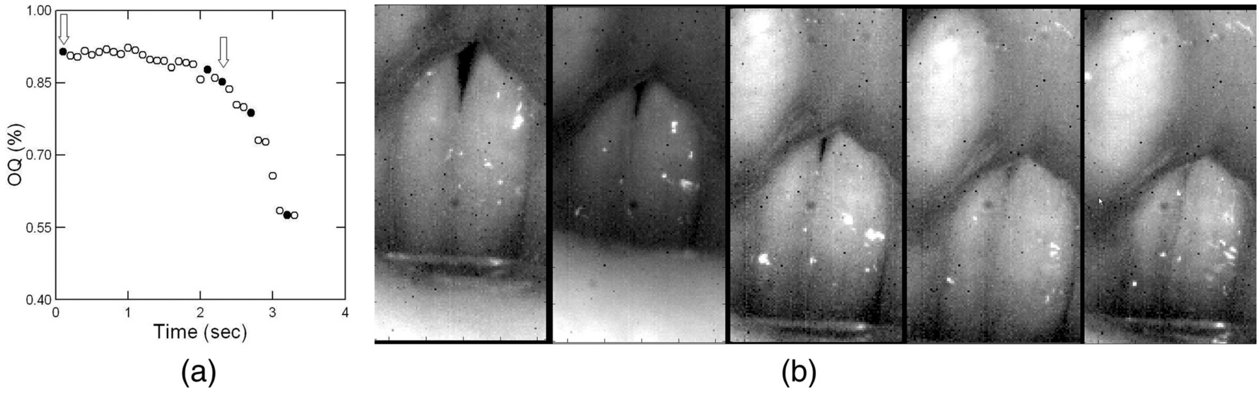

FIG. 3. (A) Changes in open quotient over time as quality changed from breathy to pressed, for experiment 2. Each point represents the mean of values in a window 0.1 s long. (B) Video images corresponding to the most-closed portion of the glottal cycle, at the times indicated by filled circles in (A). Arrows indicate the region over which glottal gap size decreased. The anterior portion of the glottis is shown at the bottom of each frame.

and $F0$ associated with H1–H2 values for a given vocal production, both across and within speakers and utterances. Although causation cannot be determined directly from these correlational analyses, in the context of the theoretical analyses and modeling results reviewed in Sec. I, it seems reasonable to conclude that speakers have a variety of strategies available to them when varying voice quality. Although the precise physiological mechanisms for achieving these goals remain unknown, these include manipulations of glottal pulse skewness, OQ, and/or glottal gap. Thus, the relationship between specific attributes of the source spectrum and the voice production process is a complex one. Despite this variety of phonatory configurations across speakers and utterances, *within* speakers OQ, the asymmetry coefficient, and/or $F0$ accounted for at least 57%, and as much as 93%, of variance in H1*–H2* across utterances. These results indicate that H1*–H2* is predictable, but that the predictive models are speaker dependent.

It is not surprising that speakers would have a variety of phonatory strategies available to them for manipulating H1–H2 in speech. Listeners are highly sensitive to the relative amplitudes of the lowest harmonics (Kreiman *et al.*, 2010b), which convey both paralinguistic information about a variety of personal and interpersonal attributes [see Kreiman and Sidtis (2011) for review] and linguistic information in languages like Gujarati (Fischer-Jorgensen, 1967), Chong (DiCanio, 2009), and White Hmong (Huffman, 1987). The ability to use different movements to produce the same speech sound has been described for the oral articulators (e.g., Guenther, 1994), and in the case of phonation may arise from attempts to produce a particular quality, whether for linguistic or paralinguistic reasons, in the context of different combinations of simultaneous pitch and/or loudness goals.

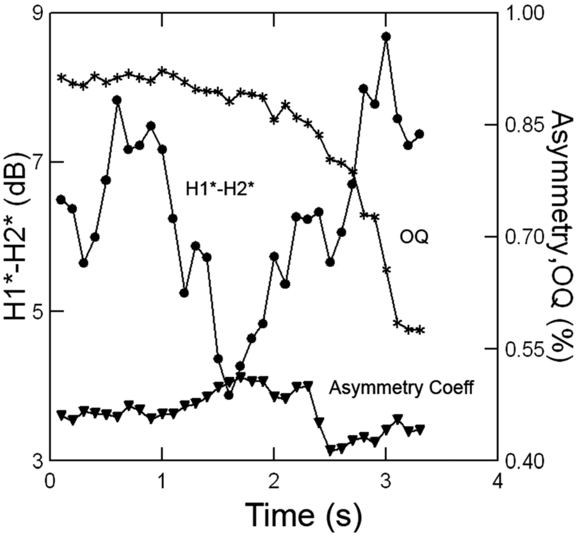

FIG. 4. H1*–H2* (filled circles), open quotient (stars), and asymmetry coefficients (open squares) for the quality glide in experiment 2, all plotted against time. Each sample represents the mean of values for that parameter in a window 0.1 s long.

However, the variety of strategies implied by the results also suggests that speakers may not directly control single spectral attributes out of the context of the overall source spectrum. H1–H2 can be manipulated by changing the amplitude of H1, but also by altering the slope of the harmonic spectrum above H1 (so that H2 changes while H1 remains constant). This suggests that examinations that focus solely on H1 and H2 out of the context of other co-occurring spectral changes may risk misrepresenting the mechanisms speakers employ to control overall spectral shape to reach intended voice quality goals.

One significant limitation to this research is the accuracy/inaccuracy of bandwidth estimates. As noted previously, it is difficult to estimate formant bandwidths with accuracy, particularly when the glottis is open. Although the relationship between formant frequencies and bandwidths is somewhat understood when the glottis is fully closed (Hawks and Miller, 1995), a large percentage of phonation occurs with glottal gaps, for both male and female speakers.

TABLE IV. Standardized regression coefficients relating open quotient, the asymmetry coefficient, glottal gap presence, and H1*–H2* in experiment 2. All values except those with an asterisk (*) are significant at $p < 0.05$.

	Standardized regression coefficients		
Time	Open quotient	Asymmetry coefficient	R^2
< = 2.3 s (glottal gap present)	−0.25*	−0.84	0.66
>2.3 s (no glottal gap present)	−1.40	0.32*	0.55

More research on the empirical effects of such gaps on bandwidths is essential for developing accurate techniques for measuring harmonic amplitudes in natural speech.

In conclusion, this study provides data that both support and contradict descriptions of the relationships among voice quality, H1*–H2*, OQ, and glottal area waveform skewness. Speakers appear to have several strategies available for varying voice quality along the breathy–pressed continuum, including manipulating glottal gap, changing OQ, varying *F*0, and altering the skewness of glottal pulses. Despite this observed variability in phonatory configuration, OQ, the asymmetry coefficient, and *F*0 accounted for the majority of variance in H1*–H2* across utterances. Thus, H1*–H2* can be predicted with good accuracy, but its relationship to phonatory characteristics is complex and speaker dependent.

ACKNOWLEDGMENTS

This research was supported in part by Grant No. DC01797 from the National Institutes of Health (NIDCD) and by Grant Nos. BCS-0720304 and IIS-1018863 from the National Science Foundation. The speech synthesizer was written by Norma Antoñanzas-Barroso, who also assisted with acoustic analyses.

[1] H1–H2 can be measured in two ways—directly from the spectrum of the voice source (usually obtained by inverse filtering or from a computational model), or from the audio signal at the mouth after canceling the influence of vocal tract resonances. This second method produces a measure designated H1*–H2*. No typographical convention exists to distinguish measures made directly from the spectrum of the glottal pulse from uncorrected measures made from spectra of unfiltered speech, both of which are usually designated H1–H2. In this paper, we use H1–H2 to indicate measures made directly from the spectrum of a glottal pulse, and H1*–H2* to indicate corrected measures made from the voice signal at the mouth. Note that 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.

[2] Also sometimes called the speed quotient or pulse skewness and defined as $t_o/(t_o + t_c)$, where t_o is the duration of the opening phase and t_c is the duration of the closing phase.

[3] The analysis-by-synthesis process produced only a few estimates of H1–H2 for each token. Calculated H1*–H2* values were used in their place to provide better resolution for comparison to OQ measures. The two sets of values were very well correlated ($r = 0.92$, $p < 0.01$).

[4] Preliminary analyses indicated no significant univariate relationship between *F*0 and either OQ or H1–H2. *F*0 was therefore excluded as a covariate in these analyses.

[5] The same result occurs if data are divided at sample = 17, which is the breakpoint in the H1*–H2* data.

ANSI S1.1-1960: *Acoustical Terminology* (American National Standards Institute, New York, 1960).

DiCanio, C. T. (**2009**). "The phonetics of register in Takhian Thong Chong," J. Int. Phonetic Assoc. **39**, 162–188.

Doval, B., and D'Alessandro, C. (**1997**). "Spectral correlates of glottal waveform models: An analytic study," in *Proceedings of ICASSP 1997*, pp. 446–452.

Doval, B., D'Alessandro, C., and Henrich, N. (**2006**). "The spectrum of glottal flow models," Acta Acust. Acust. **92**, 1226–1246.

Draper, M. R., Blagnys, B., and Premachandra, D. J. (**2007**). "To 'EE' or not to 'EE,' " J. Otolaryngol. **36**, 189–193.

Fant, G. (**1995**). "The LF model revisited. Transformations and frequency domain analysis," Speech Transm. Lab. Q. Prog. Status Rep. **2–3**, 119–156.

Fant, G. (**1997**). "The voice source in connected speech," Speech Commun. **22**, 125–139.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). "A four-parameter model of glottal flow," Speech Transm. Lab. Q. Prog. Status Rep. **4**, 1–13.

Fex, S., Lofqvist, A., and Schalen, L. (**1991**). "Videostroboscopic evaluation of glottal open quotient, related to some acoustic parameters," in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, edited by J. Gauffin and B. Hammarberg (Singular, San Diego), pp. 273–278.

Fischer-Jorgensen, E. (**1967**). "Phonetic analysis of breathy (murmured) vowels in Gujarati," Indian Linguist. **28**, 71–139.

Fujisaki, H., and Ljungqvist, M. (**1986**). "Proposal and evaluation of models for the glottal source waveform," in *Proceedings of ICASSP 1986*, pp. 1605–1608.

Guenther, F. H. (**1994**). "A neural network model of speech acquisition and motor equivalent speech production," Biol. Cybern. **72**, 43–53.

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

Hawks, W., and Miller, J. D. (**1995**). "A formant bandwidth estimation procedure for vowel synthesis," J. Acoust. Soc. Am. **97**, 1343–1344.

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*, pp. 47–50.

Hillenbrand, J., and Houde, R. A. (**1996**). "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Hear. Res. **39**, 311–321.

Hirano, M., and Bless, D. M. (**1993**). *Videostroboscopic Examination of the Larynx* (Singular, San Diego), pp. 29–33.

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

Howe, M., and McGowan, R. (**2007**). "Sound generated by aerodynamic sources near a deformable body, with application to voiced speech," J. Fluid Mech. **592**, 367–392.

Huffman, M. K. (**1987**). "Measures of phonation type in Hmong," J. Acoust. Soc. Am. **81**, 495–504.

Iseli, M., Shue, Y.-L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**, 2283–2295.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis and perception of voice quality variations among male and female talkers," J. Acoust. Soc. Am. **87**, 820–856.

Kreiman, J., Antoñanzas-Barroso, N., and Gerratt, B. R. (**2010a**). "Integrated software for analysis and synthesis of voice quality," Behav. Res. Methods **42**, 1030–1041.

Kreiman, J., Gerratt, B. R., and Khan, S. D. (**2010b**). "Effects of native language on perception of voice quality," J. Phonetics **38**, 588–593.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell, Malden, MA).

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice Hall, Englewood Cliffs, NJ,), pp. 116–164.

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

Shue, Y.-L. (**2009**). "VoiceSauce: A program for voice analysis," http://www.ee.ucla.edu/~spapl/voicesauce/index.html (Last viewed June 23, 2011).

Shue, Y.-L., and Alwan, A. (**2010**). "A new voice source model based on high-speed imaging and its application to voice source estimation," in *Proceedings of ICASSP 2010*, pp. 5134–5137.

Souderton, M., and Lindesay, P.-Å. (**1990**). "Glottal closure and perceived breathiness during phonation in normally speaking subjects," J. Speech Hear. Res. **33**, 601–611.

Stevens, K. M. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA), pp. 163–167.

Sundberg, J., Andersson, M., and Hultqvist, C. (**1999**). "Effects of subglottal pressure variation on professional baritone singers' voice sources," J. Acoust. Soc. Am. **105**, 1965–1971.

Swerts, M., and Veldhuis, R. (**2001**). "The effect of speech melody on voice quality," Speech Commun. **33**, 297–303.

Veldhuis, R. (**1998**). "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation," J. Acoust. Soc. Am. **103**, 566–571.