Linguistic Voice Quality

Patricia A. Keating[1] and Christina Esposito[2]

[1]Department of Linguistics, University of California, Los Angeles, CA, USA
[2]Department of Linguistics, Macalester College, St.Paul, MN, USA
keating@humnet.ucla.edu, esposito@macalester.edu

Abstract

Contrasting phonation types in languages can differ along several acoustic dimensions, depending on whether the contrast involves differences in open quotient, in glottal closing velocity, and in noise excitation/periodicity. Listeners thus potentially have multiple perceptual cues to such contrasts. Two perception experiments, classification and similarity rating, show that listeners from different language backgrounds attend to different acoustic correlates of modal and breathy voice vowels. Contrastive tones in tone languages may also vary in phonation quality.

1. Introduction

One of the less-studied aspects of speech is phonation (or, loosely, voice) quality, within and across languages. Phonation is the production of sound by the vibration of the vocal folds. Some languages have segmental phonation contrasts, on vowels and/or on consonants. One of Peter Ladefoged's contributions to linguistic phonetics was to encourage the study of such phonation contrasts (most recently, Gordon and Ladefoged, 2001). Ladefoged believed that segmental contrasts can provide a testbed for developing the study of voice quality, because within a language, every speaker intends to produce acceptable instances of the contrasting phonemic voice categories. This is not necessarily the case with voice quality variation due to prosody, to emotion, to speaker identity, to clinical pathology, etc. – all areas in which voice is important, but hard to pin down *a priori*.

Ladefoged focused on a simplified model of possible phonations, the glottal constriction continuum (Ladefoged, 1971; Ladefoged and Maddieson; 1996, Gordon and Ladefoged, 2001), shown in Figure 1. On this model, the size of the glottis, which depends on the distance between the vocal folds, can vary from zero, without phonation (for a glottal stop), to that used for voicelessness, again without phonation. While his original discussion gave 9 states along this continuum, in the version here the range of glottal positions that allows phonation is divided into three categories, corresponding to the three most common contrasting phonation types: **creaky voice**, produced with a constricted glottis; **breathy voice**, produced with a more open glottis; and **modal voice**, in between these two. Furthermore, since glottal constriction is a continuum, there are degrees of creakiness and of breathiness, and indeed, the modal voice category itself varies from more constricted to more open. Ladefoged also stressed that these categories, like other phonetic categories, are not absolute. Not only might they differ somewhat across languages (so that what counts as breathy voice in one language might count as modal in another), but they are sure to differ somewhat across speakers (so that what counts as breathy voice for one speaker might count as modal for another, within the same language).

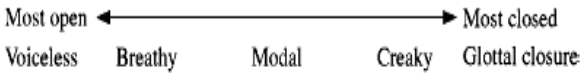

Figure 1. *Continuum of glottal contrictions (after Ladefoged, 1971), reproduced from Gordon and Ladefoged (2001).*

Other proposals, e.g. Laver (1980), are more complex; Laver for example makes fundamental distinctions between different types of glottal constriction, and between glottal constriction and overall muscular tension. He stresses that flow through the posterior glottis ("whisper") provides a noise excitation source that can combine with any other phonation. However, considering only those categories generally used in linguistic contrasts, the differences between Laver and Ladefoged are not so great.

The simplified glottal constriction continuum, with its three contrasting categories of voicing and gradient phonetic variation, is similar to the more familiar Voice Onset Time continuum (Blankenship, 2002). Another way in which the two are similar is that in both cases the articulations involved are more complex than is indicated by the simple names "glottal constriction" or "voice onset time". As a result, there can be many

acoustic differences among the categories. Furthermore, just as there are different ways of producing, e.g., voiceless unaspirated short lag VOTs, there are different ways of producing breathy voice and creaky voice, so that speakers will differ in the acoustic details of their contrasts.

In this paper we first revisit this question of the acoustic dimensions along which phonation contrasts can be produced, and how languages can differ in their use of these dimensions. We then move on to a question not addressed in Ladefoged's work, or in any cross-language studies: which dimensions listeners use to *perceive* phonation contrasts. Finally, we briefly look at one way in which phonation might vary with other linguistic contrasts.

2. Acoustic measures

Most linguists, including Ladefoged, have focused on the acoustic measure H1-H2, the difference between the amplitudes of H1 (the first harmonic, i.e. the fundamental) and H2 (the second) in the Fourier spectrum, shown in Figure 2. This measure is related to the Open Quotient (OQ), the proportion of a glottal cycle in which the glottis is open (Holmberg, Hillman, Perkell, Guiod, and Goldman, 1995; Ni Chasaide and Gobl, 1997). Since the OQ arguably in turn relates to overall glottal stricture, H1-H2 is an acoustic measure well-suited to characterizing differences along the glottal constriction continuum, and has been applied to many languages. Figure 3 shows means from 10 language/dialects samples of breathy vs. modal phonations; H1-H2 distinguishes 8 out of these 10 samples.

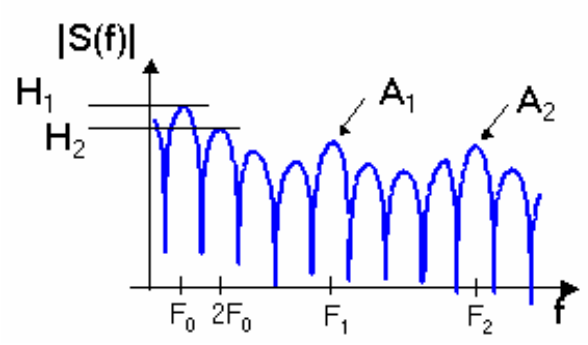

Figure 2. *Speech spectrum |S(f)| in dB, showing harmonics H1 = |S(F0)|, H2 = |S(2F0)|, and the magnitudes of the first and second formant, A1 and A2, respectively*

Two other aspects of phonation distinctions are better captured by other measures. The strength of higher frequencies in the spectrum is thought to be related to the closing velocity of the vocal folds and perhaps to muscle tensions. This property can be quantified in many ways, including the amplitude of H1 relative to some higher frequency component, such as one of the formant frequencies (or the strongest harmonic near it). Figure 2 also shows A1 (amplitude of F1) and A2, giving e.g. H1-A1. In addition, the presence of noise excitation is an important component of breathy voice (or, in Laver's terms, of whispery voice). One measure that reflects a harmonics-to-noise ratio is the Cepstral Peak Prominence of Hillenbrand, Cleveland, and Erickson (1994). The stronger the cepstral peak, the stronger the harmonics above the floor, noise or otherwise, of the Fourier spectrum, and the more periodic the signal. Various such measures – Cepstral Peak Prominence, H1-A1, H1-A2, H1-A3 – distinguish the modal and breathy categories of the two languages seen in Figure 3 where H1-H2 does not work.

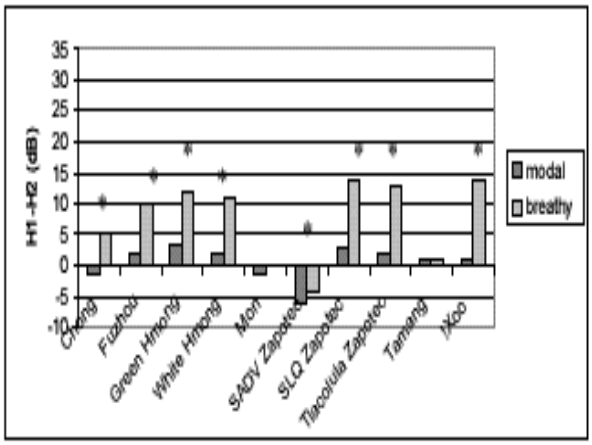

Figure 3. *Mean H1-H2 (in dB) for 10 dialects/languages with a modal vs. breathy voice distinction. Higher values are breathier. An asterisk indicates that the measures are significantly different for that sample. From Esposito (2006).*

In our work we sometimes use H1-H2 etc., but sometimes corrected measures notated H1*-H2* etc.. The problem with measures based on H1 and taken from output spectra rather than source spectra, is that H1 is very sensitive to the influence of F1. Inverse filtering to recover the source spectrum would get around this, but it has its own problems. In practice most researchers have been limited to either careful matching of vowel identity, or more commonly, only very open vowels with very high F1s. Iseli and Alwan (2004), building on a proposal by Hanson (1995), developed an algorithm to automatically correct spectral magnitudes for the effect of the F1 and F2 vocal tract resonances (frequencies and bandwidths) on harmonic amplitudes. Their H1*-H2* measure thus incorporates a sort of post-hoc substitute for inverse filtering which

can be used across vowel qualities. While Iseli and Alwan (2004) found that the measure was quite accurate for synthesized point vowels, Iseli Shue, and Alwan (2006) did find some vowel-dependence, so the algorithm continues to be developed, e.g. taking into account higher formants. In their current implementation, F0 and estimates of F1 through F4 are obtained through Snack (Sjolander, 2004) and used to correct the harmonic amplitudes. Various measures from harmonic amplitudes (e.g. H1*-A3*) are then made automatically every 10 ms over 25 ms windows, then smoothed.

3. Multiple dimensions of contrast

In earlier work, Esposito (2003, 2005) compared the phonation categories of Santa Ana del Valle Zapotec. Unexpectedly, the acoustic analysis suggested that men and women produce the contrast differently: the men's categories were distinguished by H1-A3 while the women's were distinguished by H1-H2. That this acoustic difference reflects a real difference in production was confirmed in an electroglottographic study. The men's categories differed in the symmetry of the glottal opening and closing phases, while the women's differed in the closed quotient (the EGG analog of OQ) of the glottal pulses.

Thus, to the extent that different measures reflect different aspects of production, they may or may not all distinguish the phonation categories of a given language. Typically, a set of measures is made and ANOVAs are used to determine which measures distinguish the contrasts. For example, Wayland and Jongman (2003) tested five measures in Khmer, and found that H1-H2, H1-A1, and vowel RMS amplitude successfully distinguished breathy and modal vowels. Blankenship (2002) tested 10 measures in Mazatec and found three measures that worked for its three categories (H1-H2, H1-A2, and Cepstral Peak Prominence, the latter needed only for modal vs breathy). She then applied these three measures to some other languages with varying results. Finally, Esposito (2006) tested 10 languages/dialects that contrast breathy and modal phonations, and found that Cepstral Peak Prominence, H1-A3, H1-H2, and H1-A2 were the most successful measures.

Within a language, the several successful measures define a many-dimensional phonetic space for linguistic voice quality. The question then moves from "yes/no" for each measure, to "how much" of each – the relative independence and importance of the dimensions in characterizing the differences within each language. Methods like discriminant analysis or PCA provide one sort of answer. For example, Andruski & Ratliff (2000) report very good results from a discriminant analysis of 3 Mong tones using 15 predictors.

4. Perception of breathiness

This same question can be asked of how people *perceive* a contrast. With multiple dimensions of contrast, listeners doubtless pay attention to multiple cues. We can ask which correlates listeners attend to and how they weight them.

Esposito (2006) used this approach for the first time with phonation contrasts, specifically breathy vs. modal voice. Her study comprised three experiments, in all cases involving three groups of listeners: native speakers of Gujarati (12, all screened for fluency in Gujarati and for ability to perceive the modal-breathy vowel distinction in Gujarati, and all bilingual in Indian English), American English (18), and Mexican Spanish (18, all bilingual in English). Here we'll describe two of her experiments.

4.1. Classification

Samples were selected of breathy and modal vowels spoken by males from the UCLA Phonetic Archive, or from new recordings: 2 breathy and 2 modal tokens from each language, 40 tokens total. For each token, duration was normalized to 250 msec and F0 was normalized to a slight fall 115-110 Hz. Eight acoustic measures were made from these tokens. Several of the measures distinguished pairs of vowels within languages. However, discriminant analysis, on all 40 samples combined, accounted for 91% of the variance in the data with only Cepstral Peak Prominence (46%), H1-H2 (27%), H1-A2 (10%), and H1-A3 (8%).

The three groups of listeners then categorized these stimuli in a free sort task. In a free sort task, the listeners do not have to provide labels for the tokens, but instead sort them according to perceived similarity, and thus this method is useful for cross-language designs. In this experiment, listeners sorted into two categories, by moving stimulus icons into on-screen boxes. Details of the method are given in Esposito (2006). Because duration, F0, vowel, and speaker sex were all controlled, it was expected that voice quality would be the likely basis for sorting.

Listeners' responses were scored in terms of their correspondence to the vowels' true phonation categories in their source languages, the agreement across listeners, and, crucially, the relation of listener judgments to the acoustic measures. To determine this relation, for each pair of tokens the difference on each acoustic measure (i.e. the physical difference) was calculated, as was the proportion of listeners by whom the pair was sorted into the same box (i.e. classified as more similar than not). The sorting measure was then correlated against each of the 8 physical difference measures.

The results for the English and Spanish listeners were similar: there was little cross-listener consistency

and their sortings did not correspond to the source language categories. However, interestingly the two groups of listeners attended to different cues. Spanish listeners' judgments correlated with both H1-H2 and H1-A1, while English listeners used only H1-H2. The Gujarati listeners, in contrast, were highly consistent as a group, were generally though not always correct in their sorting of the tokens, and relied strongly on H1-H2. The relation of sorting and H1-H2 across all the Gujarati listeners is shown in Figure 4. The larger the H1-H2 difference between two stimuli, the more likely that those stimuli were sorted differently. It is not surprising that Gujarati listeners should rely on H1-H2, since that is the dimension along which Gujarati breathy and modal vowels differ; the Gujaratis sorted these vowels from these other languages as if they were Gujarati (which they were not). Note that none of the three listener groups seemed to attend to Cepstral Peak Prominence, though the discriminant analysis relied on it most strongly.

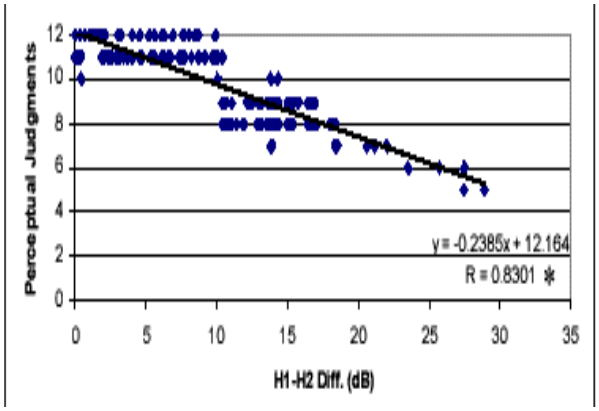

Figure 4. *From Esposito 2006 p. 70, relation of Gujarati listeners' responses to stimuli in the free sort task, and the physical H1-H2 differences between items.*

4.2. Similarity spaces

In this experiment, all the stimuli came from a single language, Mazatec. Breathy and modal vowel tokens – 16 total for each speaker - were selected from three male speakers and manipulated as described above. The listeners' task was the more traditional similarity judgment (using an on-screen slider) for all possible pairs of tokens. This experiment was randomly ordered with respect to the experiment presented above. These judgments were analyzed by MultiDimensional Scaling, and the perceptual dimensions extracted were related to the acoustic measures by correlations.

Mazatec is a good language to look at cue weighting since its vowels differ in several measures. Discriminant analysis of the set of tokens used in this experiment showed that H1-A2 is the best discriminator for these 48 tokens (53% variance), then H1-A1 (20%), and only then H1-H2 (14%) (leaving 13% variance accounted for by all other measures, including Cepstral Peak Prominence, which had been important in Blankenship (2002)'s much smaller sample).

As before, Spanish and English listeners were inconsistent as groups, while the Gujarati listeners were consistent. Also as before, the Gujaratis relied solely on H1-H2. However, the Gujarati individual and group perceptual spaces show the tokens clustering into three, not two, clusters; the breathiest of the Mazatec tokens formed a separate cluster. English listeners relied weakly on H1-H2 and sometimes Cepstral Peak Prominence, while the Spanish listeners relied weakly on H1-A1 and H1-H2. Note that none of the listeners used H1-A2 (the best predictor), and the Spanish listeners were the only ones using H1-A1.

In the sorting experiment described in section 4.1 above, which used the 10-language/dialect stimulus set, Cepstral Peak Prominence was the best discriminant of the contrasts but was not used by any listener group. This failure can be interpreted as indicating that the size of the auditory Cepstral Peak Prominence distinction present in the stimuli was too small for listeners. This interpretation is supported by the fact that in this second experiment on similarity, using stimuli from a different language, the Cepstral Peak Prominence difference is larger and was used by English listeners.

The most basic result of these perception experiments is not surprising: Gujarati listeners, who have experience in their own language with a contrast between modal and breathy voice vowels, responded consistently to stimuli from several languages, using as their sole cue the acoustic correlate of their own contrast. The English and Spanish listeners were not consistent, relying only weakly on a variety of correlates, not necessarily the strongest, of the contrasts.

5. Tones and voice

There are languages where tone and phonation co-vary, either on all tones or on some. In such cases it may not be clear if it is the F0 or the phonation that is contrastive, and whether listeners rely on one, the other, or both (e.g. Abramson, Thongkum and Nye 2004). Even when the nature of the contrast is clear, it is possible – indeed likely – that listeners attend to secondary dimensions as well. For example, Mandarin clearly uses F0 contrasts, but the low-dipping third tone, and even the falling fourth tone, is often produced with creaky voice (Davison 1991, Belotel-Grenié & Grenié 2004). It is possible that Mandarin listeners are sensitive to such correlations. Beyond such audible relations, though, is it possible that all tones (in all languages) have some correlated variation in phonation, more subtle because within the modal range? If so, perhaps listeners in many languages use the phonation

information in recognizing their tones, especially out of context. They could do so in either of two ways. They could use the phonation information directly: "creak means a low tone". Or they could use it indirectly: "creak means this is a low F0 in this speaker's range which means a low tone". Honorof & Whalen 2005 discuss the plausibility of using phonation quality to calibrate the speaker's F0 range even in a non-tone language like English. (They showed that listeners are pretty good at locating a pitch within a speaker's pitch range, even without prior experience or context.) This point underscores the desirability of systematically comparing F0 and phonation across languages, with special focus on tone categories in tone languages.

Is there any general relation of F0 to voice quality? On the one hand, Holmberg et al. (1989) did not find a strong correlation between any glottal parameters and F0, and likewise Epstein (2002) found no statistical relation between LF measures of voice and F0; but on the other hand Iseli et al. (2006), who separated male from female voices, found that H1*-H2* increases with increasing F0 for F0 < 175 Hz. We do not know whether such correlations are general or are limited to certain languages.

As a very preliminary foray into this topic, we looked at some minimal sets and other samples of tones in a few languages from the UCLA Phonetic Database. These are very limited samples – a few words from one speaker - so this is not an experiment, more a proof of concept. The languages presented here are Mandarin and Bura. The speech files were run through the current version of the Iseli et al. (2006) automated routine which extracted and smoothed these measures: F0, H1*-H2*, H2*-H4* and H1*-A3* (Cepstral Peak Prominence is not one of the measures made by this routine), and from the displayed tracks of the measures, some hand measurements were taken.

5.1. Mandarin

The Mandarin sample is the standard minimal set for the 4 tones on /ma/, and the creak is quite audible on the third tone example and visible at the end of the fourth tone example. We took values at four timepoints through the vowels. Figure 5 shows these measurements for F0 and H1*-H2*. In these four tokens H1*-H2* is the measure most clearly related to F0; it somewhat follows F0, with low pitches creakier and high pitches breathier, in accord with Iseli et al.'s result in English. What is notable is that the particular values that H1*-H2* takes on at tone onsets could be useful to listeners: it is high-positive (breathy) for the tones that start with a high F0; it is near 0 for the tone that starts with a mid F0; and it is low-negative (tense) for the tone that starts with a low F0. H1*-A3* (not shown) is not obviously related to either F0 or to the tone categories except at tone end, where it reflects the final F0.

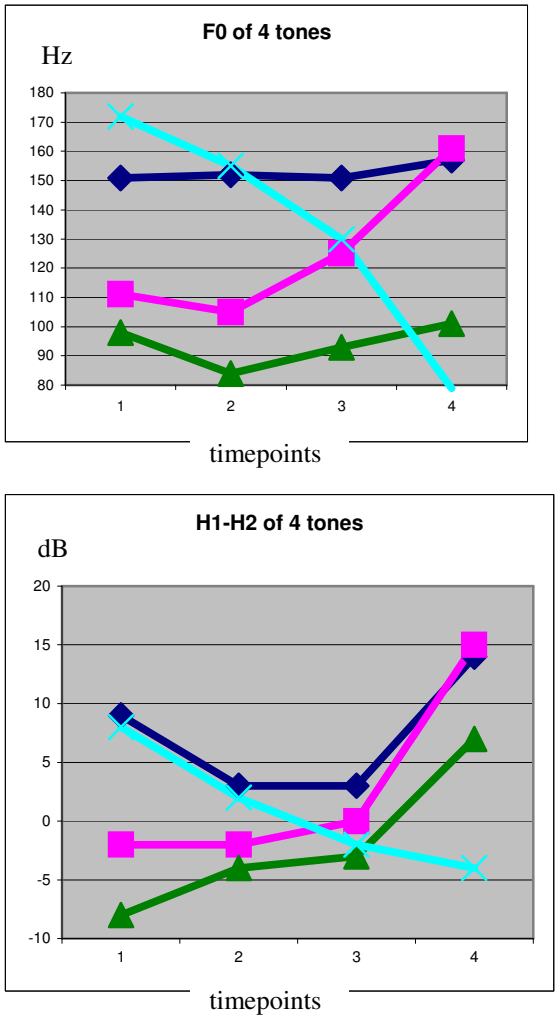

Figure 5. *F0 and H1*-H2* measurements taken at 4 timepoints in single tokens of 4 Mandarin tones. Dark blue diamonds: high level tone; pink squares: mid rising tone; green triangles: low dipping tone; aqua circles: high falling tone.*

5.2. Bura

A larger sample of words is available from one male speaker of Bura (a Chadic language of Nigeria), so that multiple tokens of High and Low tones (just these 2 level tones, no Mid tones or contours are included here) could be compared quantitatively. One measure per vowel was made at mid-vowel, since the values were relatively steady over time. The measures were compared in a exploratory factorial ANOVAs, with factors Tone and Position in word, both with and without F0 as a covariate. In Bura we see a different pattern of results compared to Mandarin: H1*-H2* is generally fairly high and does not distinguish the tones,

even though the speaker's F0 is under 175 Hz. On the other hand, both H2*-H4* and H1*-A3* are higher – breathier – in Low tones, though neither is correlated with F0. Furthermore, a discriminant analysis on these data on just the voice measures, without F0, classifies 57% of the tones correctly, using only H1*-A3*.

6. Conclusions

Linguistic voice quality is a rich yet relatively understudied area. Phonation contrasts are multidimensional, and listeners with different language experience attend to different dimensions. A promising research strategy is to better understand linguistic contrasts, so that the knowledge obtained that way can be applied to other areas in which voice quality is important.

7. Acknowledgements

We gratefully acknowledge the earlier work of Barbara Blankenship and Melissa Epstein, and collaboration with Abeer Alwan, Jody Kreiman, Markus Iseli, and Yen Shue.

8. References

Abramson, A., T. Thongkum, and P. Nye (2004). Voice register in Suai (Kuai): An Analysis of Perceptual and Acoustic Data. *Phonetica* 61: 147-71

Andruski, J., and M. Ratliff (2000). Phonation types in production of phonological tone: The case of Green Mong. *JIPA*, 30:37-61.

Belotel-Grenié, A., and M. Grenié (2004). The creaky voice phonation and the organization of Chinese discourse. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing 2004.

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics*, 30:163-191.

Davison, D. S. (1991). An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics*, 78:50-57.

Esposito, C. M. (2003). *Santa Ana Del Valle Zapotec Phonation*. MA thesis, UCLA.

Esposito, C. M. (2005). An Acoustic and Electroglottographic Study of Phonation in Santa Ana del Valle Zapotec. Poster presented at the *79th meeting of the Linguistic Society of America*, San Francisco, CA.

Esposito, C.M. (2006). *The Effects of Linguistic Experience on the Perception of Phonation.* Ph.D. dissertation, UCLA.

Gordon, M., and P. Ladefoged (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29:383-406.

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

Hillenbrand, J.M., R. A. Cleveland, and R.L. Erickson (1994). Acoustic correlates of breathy vocal quality. *JSHR*, 37:769-778.

Honorof, D., and D.H. Whalen (2005). Perception of pitch location within a speaker's F0 range. *JASA*,117:2193–2200.

Holmberg, E. B., R. E. Hillman and J. S. Perkell (1989). Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch. *J.Voice*, 3:294-305.

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

Iseli, M., Y. Shue, and A. Alwan (2006). Age- and Gender-Dependent Analysis of Voice Source Characteristics, *Proc. ICASSP*, Toulouse, May 2006, I-389.

Iseli, M., Y. Shue, M. Epstein, P. Keating, J. Kreiman, A. Alwan (2006). Voice source correlates of prosodic features in American English: A pilot study. *Proc. Interspeech 2006*, Pittsburgh, October 2006

Iseli, M., and A. Alwan (2004). An Improved Correction Formula for The Estimation of Harmonic Magnitudes and Its Application to Open Quotient Estimation. *Proc. ICASSP*, Montreal, May 2004, pp. 669-672.

Ladefoged, P. (1971) *Preliminaries to linguistic phonetic*s. University of Chicago Press, Chicago.

Ladefoged, P., and I. Maddieson (1996). *Sounds of the World's Languages.* Blackwells, Oxford.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Ni Chasaide, A., and C. Gobl (1997). Voice source variation. In W. J. Hardcastle and J. Laver (Eds.), *The Handbook of Phonetic Sciences*, pp. 428-461. Oxford: Blackwells.

Sjolander, K. (2004). Snack sound toolkit, KTH Stockholm, Sweden. Available at http://www.speech.kth.se/snack.

Wayland, R., and A. Jongman (2003). Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics*, 31: 181-201.