Analysis, synthesis, and perception of voice quality variations among female and male talkers

Dennis H. Klatt[a] and Laura C. Klatt[b]
Research Laboratory of Electronics, Room 36-523, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 10 October 1988; accepted for publication 10 October 1989)

Voice quality variations include a set of voicing sound source modifications ranging from laryngealized to normal to breathy phonation. Analysis of reiterant imitations of two sentences by ten female and six male talkers has shown that the potential acoustic cues to this type of voice quality variation include: (1) increases to the relative amplitude of the fundamental frequency component as open quotient increases; (2) increases to the amount of aspiration noise that replaces higher frequency harmonics as the arytenoids become more separated; (3) increases to lower formant bandwidths; and (4) introduction of extra pole zeros in the vocal-tract transfer function associated with tracheal coupling. Perceptual validation of the relative importance of these cues for signaling a breathy voice quality has been accomplished using a new voicing source model for synthesis of more natural male and female voices. The new formant synthesizer, KLSYN88, is fully documented here. Results of the perception study indicate that, contrary to previous research which emphasizes the importance of increased amplitude of the fundamental component, aspiration noise is perceptually most important. Without its presence, increases to the fundamental component may induce the sensation of nasality in a high-pitched voice. Further results of the acoustic analysis include the observations that: (1) over the course of a sentence, the acoustic manifestations of breathiness vary considerably—tending to increase for unstressed syllables, in utterance-final syllables, and at the margins of voiceless consonants; (2) on average, females are more breathy than males, but there are very large differences between subjects within each gender; (3) many utterances appear to end in a "breathy-laryngealized" type of vibration; and (4) diplophonic irregularities in the timing of glottal periods occur frequently, especially at the end of an utterance. Diplophonia and other deviations from perfect periodicity may be important aspects of naturalness in synthesis.

PACS numbers: 43.70.Dn, 43.70.Gr, 43.71.Bp

INTRODUCTION

Voice quality is a term that subsumes a wide range of possible meanings (Abercrombie, 1967; Laver, 1980). In this paper, the topic will be restricted to perceptual and acoustic correlates of changes in the breathiness or pressed/laryngealized nature of the voicing sound source. Additional possible aspects of voice quality, not considered here, include harshness and other pathological voice qualities, soft/weak/whispered voice, falsetto, and habitual settings of the vocal-tract configuration, such as a tendency toward an overall nasality quality.

The present study examines vowel spectra obtained from a fairly wide sample of female and male voices under conditions of natural variation over a sentence, while minimizing the confounding influences of variable consonantal contexts. Reiterant speech is used to control consonantal context. Both voiceless [h] and voiced glottal stop [ʔ] consonants are employed in reiterant imitations of two five-syllable sentences with differing stress patterns. A pilot study of a single female speaker and employing a similar speech sample (Klatt, 1986b) has guided the choice of analysis techniques for the larger corpus of subjects studied here. A preliminary report of the results of this study was given at the 114th Meeting of the Acoustical Society of America (Klatt, 1987a).

Women and children have been somewhat neglected groups in the history of speech analysis by machine. One reason is that most acoustic studies tend to focus on formant frequencies as cues to phonetic contrasts. The higher fundamental frequencies of women and children make it more difficult to estimate formant-frequency locations. Furthermore, informal observations hint at the possibility that vowel spectra obtained from women's voices do not conform as well to an all-pole model, due perhaps to tracheal coupling and source/tract interactions (Fant, 1985; Klatt, 1986b).

The acoustic analyses to be described isolate several factors that distinguish a breathy from a normal vowel. In order to determine the perceptual importance of each factor, two types of perception tests have also been performed. In the first type of test, edited samples of natural speech from each subject were played to listeners and judgments of breathiness were obtained. Correlations were then performed between the subjective and acoustic data. Of nine acoustic parameters examined, only two were significantly correlated with subject responses—degree of aspiration noise intruding at higher frequencies in the vowels and relative strength of the fundamental component. These results are consistent with prior research, as discussed in the literature review below. Second,

[a] Deceased 30 December 1988.
[b] Summer research assistant, 1987.

 0001-4966/90/020820-38$00.80

a formant synthesizer was used to systematically manipulate several acoustic parameters separately and in combination. A new voicing source was used that is characterized by more flexible control over open quotient (i.e., proportion of a period during which the glottis is open) and spectral tilt. Judgments were obtained of breathiness, naturalness, and nasality of 14 stimuli. Results confirm the importance of aspiration noise and show that, for a female voice, increases to the strength of the fundamental component are not always a sign of perceptual breathiness, but rather may induce a sensation of increased nasality unless accompanied by aspiration noise.

Reproduction of a female voice, using either a formant synthesizer or linear prediction analysis/resynthesis has not been particularly successful in the past. For example, John Holmes (1961, 1973) has produced excellent imitations of male speakers, in which the synthesis is largely indistinguishable from the natural recording. However, he and others have not been nearly as convincing in attempts to mimic a female voice (note examples 7 and 8 in the recording accompanying Klatt, 1987b). Our efforts to synthesize reiterant utterances from several of our female speakers indicate that the new voicing source waveform, coupled with a deeper understanding of the acoustic manifestations of variations in voice quality accruing from this study, results in highly successful imitations. The new synthesizer voicing source is fully documented below, and examples are provided of control parameter values required to match two utterances.

The data that we will present indicate that an utterance is often terminated in such a way that the arytenoid cartilages begin to separate in preparation for breathing, leading to a breathy-voiced offset to the final syllable. However, our interpretation of the acoustic evidence suggests that there are two alternative ways in which an early abduction gesture is implemented: (1) a general "relaxed" separation of the arytenoids or (2) a "laryngealized" mode in which the abduction is accompanied by a rotational motion of the arytenoids such that some medial compression is applied to keep the folds vibrating in a nearly normal way in spite of the opening at the posterior. In both cases, there is increased noise in the spectrum, but first-harmonic amplitude is increased and the harmonic spectrum tilts down with frequency only in the first case. The second breathy-laryngealized strategy, while deduced from acoustic evidence to be described, is consistent with what is known about the degrees of freedom of the arytenoids and their associated musculature (Sonesson, 1960). A breathy-laryngealized termination is characteristic of many male speakers in our sample and may be a social marker of maleness.

Although this study concentrates on the spectral manifestations of the contrast between breathy and normal voice qualities, the corpus provides evidence of widespread occurrence of irregularities in the *timing* of glottal pulses over portions of many reiterant sentences. This timing variation has led to the creation of two new synthesizer control parameters involving: (1) a small slowly varying f_0 pseudorandom flutter and (2) diplophonic double pulsing, in which pairs of glottal pulses migrate toward one another, and the first of the pair is usually attenuated in amplitude (Timke *et al.*, 1959; Ward *et al.*, 1969). Diplophonia tends to occur when subglottal pressure is falling and when the fundamental frequency is low, i.e., when voicing is somewhat unstable. Variations in the timing of glottal pulses such as these no doubt lend a kind of naturalness to utterances that is missing in synthetic speech generated by most models.

A. Phonetic theory and physiological mechanisms

Examinations of the physics of larynx behavior (Stevens, 1981) suggest that the possible modes of sound generation fall into a small number of natural categories. Similarly, cross-language comparisons of phonemic contrasts involving differing laryngeal modes suggest the existence of only a few distinctive contrasts (Ladefoged, 1973). According to Ladefoged, languages use the larynx in three distinctive ways: (1) by varying laryngeal tension so as to produce changes in fundamental frequency of voicing; (2) by adjusting the separation between the arytenoid cartilages to realize different phonation types such as glottal stop, creaky voice, modal voice, breathy voice, voiceless, and fully spread for breathing (see also Catford, 1964, 1977; Halle and Stevens, 1971; and Laver, 1980, for similar categorizations of phonation types[1]); and (3) by varying the timing of the onset of voicing relative to supraglottal articulatory movements to realize, for example, prevoiced, voiceless-unaspirated, and voiceless-aspirated consonants.

Ladefoged proposes a set of multivalued distinctive features to capture linguistic contrasts along each of these continua and provides examples of languages that use each category distinctively. A similar set of laryngeal states has been identified by Halle and Stevens (1971) and described using binary distinctive features, called spread glottis, constricted glottis, stiff vocal cords, and slack vocal cords. The best set of distinctive features for characterizing the phonological/physiological behavior of the larynx continues to be an area of some controversy. For our purposes, though, it is sufficient to note the contrastive use of laryngealized versus normal vowels in languages such as Jalapa Maxatec (Kirk *et al.*, 1984), the phonemic use of glottal-stop/glottalization gestures in Danish, or laryngealization as one of the characteristic properties of tone 3 in Mandarin Chinese, and the contrastive use of breathy versus normal vowels in languages such as Gujarati (Pandit, 1957; Fischer-Jorgensen, 1967), Hmong (Huffman, 1987), and !Xóõ (Ladefoged and Antoñanzas-Barroso, 1985). There is also the related use in some languages, such as Hindi, of voiced-aspirated stops, such as [b^h], which are characterized by an interval of simultaneous voicing and aspiration following release (Dixit, 1987).

While laryngealization and breathiness are not used phonemically in English, our data show clearly that there is considerable variation between speakers of English, and, more importantly, there is variation within an utterance on these dimensions. This variation has important implications for speech analysis (for example, making formant tracking more difficult), speech synthesis (absence of variation in voice quality during an utterance seems to lead to decrements in perceived naturalness of synthetic speech), and speech perception (creating an interesting perceptual para-

dox—one cue, an increase in the amplitude of the first harmonic, is interpreted either as signaling nasality or breathiness depending on values of other cues present in the signal).

Voice quality variation associated with changes in glottal opening is illustrated in physiological terms in the A row of Fig. 1, which shows a schematic view of the glottis from above. The positions of the arytenoid cartilages (triangles) and vocal processes are illustrated for laryngealized, modal, and breathy phonation. The characteristics of a modal voice are illustrated in column 2 of Fig. 1. The vocal folds are nearly approximated, leading to a typical volume velocity waveform, panel (2B), with an open quotient of about 50% to 60% of the period and a waveshape during the open phase that is slightly skewed (closure is more rapid than opening). The spectrum of the normal voicing source, panel (2C), has an average falloff of about -12 dB per octave of frequency increase.

In preparation for laryngealized phonation (column 1 of Fig. 1), the arytenoids are positioned so as to close off the glottis, and perhaps even apply some medial compression to the vocal processes. When lung pressure is applied to the system, the vocal folds vibrate, producing a glottal volume velocity waveform as shown in panel (1B) of the figure. The glottal pulse is relatively narrow; i.e., the duration of the open portion of a fundamental period is relatively short. In addition, the fundamental frequency is substantially lowered during laryngealization, and there may be period-to-period irregularities in both the duration of the period and the amplitude of the glottal volume velocity pulse (Timke *et al.*, 1959). Possible perceptual cues to laryngealization (associated with changes to the source spectrum) are a reduction in the relative amplitude of the fundamental component in the source spectrum, panel (1C), and a lowered fundamental frequency contour.

The glottal configuration during a breathy vowel is shown in panel (3A) of Fig. 1. The arytenoid cartilages are well separated at the back, but the vocal processes are sufficiently approximated so that the vocal folds vibrate when a lung pressure is applied to the system. Since the glottis is never completely closed at the back over the vibratory period, there is considerable dc airflow (panel 3B). This increased airflow results in the generation of turbulent aspiration noise, which is combined with the periodic voicing component to form a source spectrum consisting of both harmonics and random noise [panel (3C)]. Being relatively weak in amplitude, the aspiration noise might not be audible were it not for the fact that the vibratory behavior of the vocal folds is modified in a breathy vowel (Fant, 1980, 1982a). Ordinarily, as illustrated in the middle column, the vocal folds close simultaneously along their length, leading to an abrupt cessation of airflow and relatively strong excitation of higher harmonics at the instant of closure. In a breathy vowel, however, the folds close first at the front, and then closure propagates posteriorly, leading to a volume velocity waveform with a rounded corner at closure [panel (3B)]. The implications of this behavior for the harmonic components of the source spectrum are twofold—the waveform is more nearly sinusoidal and thus has a very strong

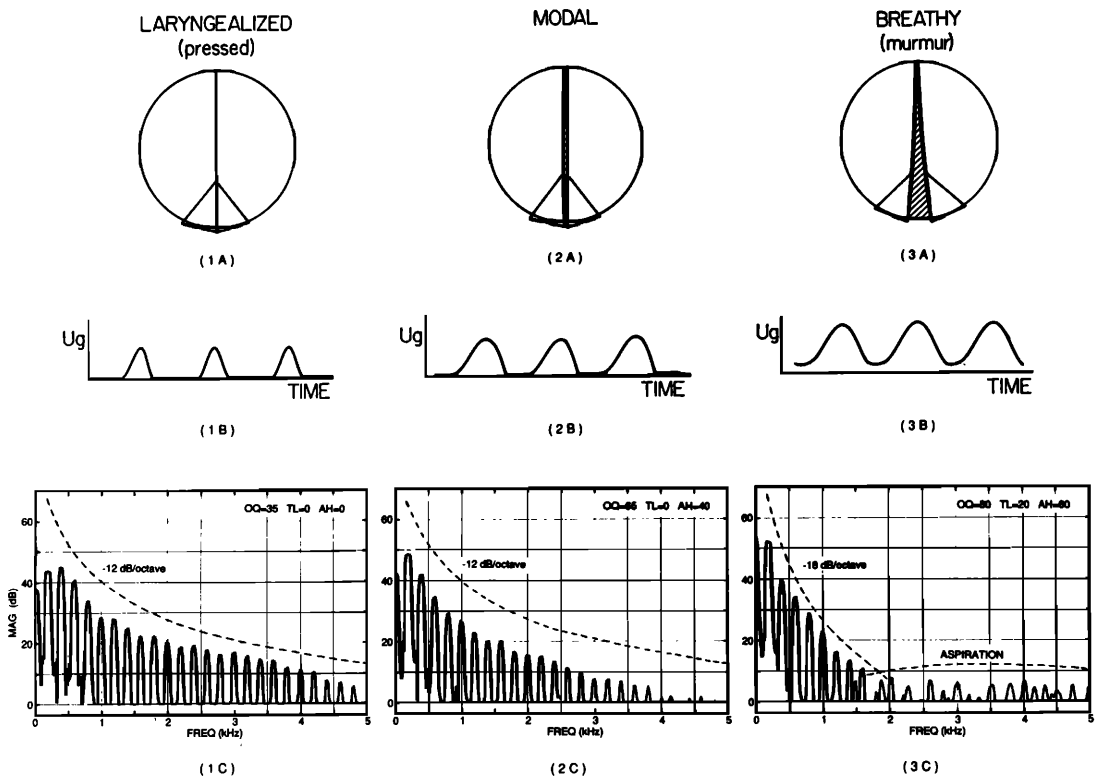


FIG. 1. Glottal configurations (row A) for (1) laryngealized, (2) modal, and (3) breathy vowels. An increased opening at the arytenoids results in glottal volume velocity waveforms (row B) with a progressively longer duration open period, an increased dc flow, and a less abrupt closure event. The source spectra (row C) have a more intense fundamental component from left to right, and the breathy configuration results in a spectrum with weaker high-frequency harmonics being replaced by aspiration noise. Figure adapted from Stevens (1977).

fundamental component, and the amplitudes of higher harmonics are attenuated substantially due to nonsimultaneous closure [panel (3C)]. Fant (1985) has termed the spectral prominence associated with the first or second harmonic in breathy phonation the "glottal formant" because it tends to show up as an extra formantlike peak in vowel spectra. Possible perceptual cues to a breathy vowel are thus an increase in the relative amplitude of the fundamental component in the spectrum and replacement of higher harmonics by aspiration noise.

B. Review of previous research on voice quality

Previous research of relevance to the area of voice quality variation for female and male speakers is divided into sections concerned with: (1) acoustical and perceptual correlates of laryngealization, (2) acoustical and perceptual correlates of breathy voice quality, and (3) acoustical and physiological studies of the voicing source, with particular emphasis on possible differences between men and women.

1. Laryngealization/creak

Laryngealization or "pressed voice" refers to a mode of vocal-fold vibrations where the volume velocity pulse is narrow (small open quotient) due to the application of medial compression through rotation of the arytenoid cartilages (Catford, 1964; Stevens, 1977). This type of physiological gesture is usually accompanied by a decrease in voicing fundamental frequency (Hollien, 1974). The extreme is a glottal stop, wherein the lung pressure is insufficient to force air through the tightly closed glottis.

Several previous studies have examined acoustic differences between laryngealized, normal and breathy vowels in languages where the contrast is phonemic. For example, Kirk *et al.* (1984) note that the amplitude of the fundamental component in the spectrum H1, relative to first-formant amplitude A1, distinguishes between the three-way phonation-type contrast (creaky, normal, and breathy) in Jalapa Maxatec; i.e., there is a more than 6-dB average difference between creaky and modal voice (H1 is 6 dB stronger for modal voice), and a further 6-dB increase in H1 between modal and breathy voice. Javkin and Maddieson (1983) used an inverse-filtering technique to show that the glottal pulse is shorter in duration for creaky voice in spite of the concomitant lowering of f_0. The expected acoustic manifestations of a narrow glottal pulse is a reduction in the relative amplitude of the fundamental component. The authors found no difference between modal and creaky phonation in a waveform jitter measure in spite of claims in the literature that creak is characterized by irregular pitch (e.g., Fourcin, 1981). Some forms of jitter, such as the diplophonic pulses (to be discussed later in connection with Fig. 12) appear to be an optional characteristic of creaky phonation.

In Danish, Abercrombie (1967) notes that two words such as *hun* "she" and *hund* "dog" are identical phonetically, except that the latter includes an interval of creaky voice. Laryngealization is also found extensively in speech materials from languages where contrasts in voice quality are not phonemic. For example, Laver (1980) suggests that speakers of British received pronunciation (RP) use a low falling f_0 accompanied by creak as a signal of completion of their turn as speaker. More extensive use of creak throughout RP speech is said to indicate bored resignation.

Creak was defined by Henton and Bladon (1987) as an irregular, very low mode of vibration in which the fundamental frequency fell below 60 Hz, and thus each pulse became audible as a separate event not unlike running a stick along a picket fence. As studied in two dialects of British English, creak was found to occur primarily, but not exclusively, near the ends of utterances. Creak occurred much more frequently for males, and much more frequently in one dialect than in another. In some males of a Northern dialect of British English, creak was observed in over 65% of the syllables. When creak was detected in a female voice, the f_0 range was observed to be essentially the same as for a male,[2] suggesting that in creaky mode the fundamental frequency is not governed by factors related to the size of the larynx or the mass of the vocal folds.

In a detailed acoustic study of reiterant sentence imitations by a single female speaker, Klatt (1986a) found that laryngealized vowel onsets were accompanied by some frication noise before the vowel [ɑ], implying the existence of a pharyngeal constriction accompanying the laryngeal adduction gesture. The noise component was fairly strong and primarily excited the second formant; this acoustic pattern is consistent with spectral data on pharyngeal fricatives in Arabic (Klatt and Stevens, 1969).

In summary, pressed voice, laryngealization, and creak refer to a mode of vocal-fold vibration in which the glottal pulse is narrower, the fundamental frequency is lowered, and there may be diplophonic irregularities to the fundamental period. It is not known whether the narrower glottal pulse or the reduction in f_0 is more important in perceptual determinations of laryngealized voice quality, but the few available perceptual data suggest that a lowered f_0 is a powerful cue. The contrast between laryngealized and modal voice qualities is used phonemically in relatively few languages of the world, but laryngealization is a very common phenomenon in all languages, where its use appears to mark word onsets, add variety to speech, signal turn taking in conversations, identify the dialect group of the speaker, indicate maleness, and function in other ways that are not yet fully understood.

2. Breathiness

A breathy vowel is produced by adjusting the glottis such that the average airflow during the vowel increases by perhaps 60% over that of normal vowels (Fischer-Jorgensen, 1967). Fischer-Jorgensen used laryngoscopic examination of one informant under stroboscopic illumination to confirm the existence of a wider opening at the rear part of the glottis for breathy vowels. Furthermore, inverse filtering of normal and breathy vowels showed that the voicing source volume velocity waveform for a breathy vowel has increased average flow, a more sinusoidal waveshape, and a relatively longer open period (Fischer-Jorgensen, 1967; see

also, Holmberg *et al.*, 1988).

Acoustic cues to phonemically breathy vowels in Gujarati have been studied by a number of authors. Pandit (1957) found that fundamental frequency tended to be lower in a breathy vowel, and there was often noise observable at higher frequencies in sound spectrograms. Presumably, the lower f_0 is due to the need to slacken the folds to promote continued voicing in spite of the static posterior separation (Halle and Stevens, 1971; Stevens, 1977), and the noise is presumably turbulent aspiration noise generated near the posterior glottal opening.

In a wide-ranging physiological, acoustic, and perceptual study of seven Gujarati informants, Fischer-Jorgensen (1967) examined acoustic dimensions related to the voicing source, as well as formant frequencies and duration cues that might help to characterize the phonemic contrast between so-called breathy and nonbreathy vowels. No consistent formant differences were detected. There was a small (4%) lowering of f_0 at the onset of a breathy vowel, the inconsistent appearance of noise in higher formants, an increase in first-formant bandwidth for low vowels, and, most notably, an increase of about 3 dB in the level of the fundamental component (H1) in the spectrum. This latter increase, while rather variable on a token-to-token basis, was present in the average data concerning H1, and also when comparing the level of H1 to the level of the first formant (A1), or when comparing the level of H1 to that of the adjacent second harmonic (H2).

A listening test revealed that the level of the fundamental component H1 and the f_0 cues were most important to the listeners. Aspiration noise, if present, seemed to override other cues, but Fischer-Jorgensen concluded on the basis of its inconsistent visual appearance in spectrograms that noise in higher formants cannot be very important. Unfortunately, her attempts to synthesize a breathy vowel based on average acoustic data from the analysis of Gujarati were not successful. The reason may be due to deficiencies in the voicing source of the formant synthesizer available at that time, but it is also possible that the synthesis strategy should include some attempt to add aspiration noise for a breathy vowel.

Ladefoged (1983) examined the contrast between breathy and modal vowels for ten speakers of !Xóõ; he found that the spectral amplitude of the fundamental component, measured relative to the amplitude of the first formant,[3] was consistently greater for breathy vowels. However, some speakers are inherently more breathy than others according to this measure, so that a breathy vowel of one speaker might overlap with a normal vowel of another speaker, and listeners must make use of some sort of speaker normalization process to determine the threshold distinguishing breathy from modal for a given speaker. Ladefoged also noted greater noise in the spectrum at higher frequencies for breathy vowels, but it was difficult to develop a measure quantifying this impressionistic difference.

In a later paper (Ladefoged and Antoñanzas-Barroso, 1985), the authors sought to define and perceptually validate two objective acoustic measures of vowel breathiness: (1) the amplitude of H1 relative to H2 (see footnote 4) and (2) the amount of noise present in the waveform, as reflected in the extent to which waveform periods from the middle of the vowel are identical or slightly different due to the statistical variability inherent in noise-excited sounds (Yumoto *et al.*, 1982; Kasuya and Ugawa, 1986).[5] Both measures correlate well with the phonemic distinction between breathy and modal vowels in several languages. In addition, it appears that some informants prefer to use one cue to signal the contrast, while others primarily employ the second. However, for American listeners, perceptual ratings of degree of breathiness correlated best with H1/H2 ($r = 0.93$) and less well with the noise measure ($r = 0.57$).

Bickley (1982) independently confirmed the increase in H1/H2 for breathy vowels in a study of the same ten speakers of !Xóõ and four speakers of Gujarati. In a synthesis experiment, she showed that Gujarati listeners are not influenced in their judgments of vowel breathiness by the amount of aspiration noise added into the synthesized vowel spectrum,[6] but switch from nearly 100% judgments of modal vowels to nearly 100% judgments of breathy vowels if the fundamental component is increased in amplitude by 15 dB. This result compares with an average difference of 6 dB between normal and breathy vowels in her contrastive words from Gujarati, and 9.7 dB in her spectral analysis of data from !Xóõ. It should not be necessary to exaggerate a cue to achieve consistent responses from listeners. Thus it may be the case that first-harmonic amplitude is not the whole story.

Huffman (1987) found that in Hmong, the breathy/modal distinction was realized by a longer open quotient (0.8 vs 0.6, as revealed by inverse filtering) and stronger fundamental component relative to the second harmonic (+ 7-dB increase for breathy). There is a simple cause and effect relationship between these two observations—an increased open quotient results in a relatively stronger fundamental component of the source spectrum, all else being equal.

The transition from a voiceless consonant to a vowel often includes a short interval of breathy voicing in which the first-harmonic amplitude is increased (Chasaide, 1987; Chasaide and Gobl, 1987). Chapin-Ringo (1988) has shown that listeners are aware of, and expect, this type of onset in the sense that a voice-onset-time continuum produced slightly more voiceless responses when first-harmonic amplitude was increased. The latter result is all the more surprising when compared with a related VOT continuum experiment (Stevens and Klatt, 1974) in which low-frequency energy was increased at voicing onset by lowering $F1$, and this resulted in more voiced responses. Both results are consistent with production data, but it is surprising that perceptual strategies are so sophisticated as to be able to identify the cause of an increase in low-frequency energy at voicing onset before assigning a phonetic value to it.

A voiced-aspirated plosive in a language such as Hindi is characterized physiologically by an airflow trace that increases substantially during the 50- to 100-ms voiced-aspirated portion following plosive release; airflow is over twice the value for a typical vowel, and about half of that found for a voiceless-aspirated [p^h] (Dixit, 1987). Typically, voicing energy is seen at or below $F1$, while significant noise excitation appears in higher formants. Fischer-Jorgensen (1967)

calls the voiced-aspirated plosive of Gujarati similar to a breathy vowel, but with more noise. Thus there appears to be a continuum from modal voicing to breathy vowels to voiced-aspirated [ɦ] vocalic intervals, with stronger cues to breathiness appearing as glottal opening increases.

In summary, breathy phonation is characterized by a glottal source with (1) an increased open quotient, resulting in an increased relative amplitude of the fundamental component in the spectrum and (2) a tendency for higher harmonics to be replaced by aspiration noise. Additional characteristics of a breathy vowel include the possibility of increased first-formant bandwidth and/or the appearance of tracheal poles and zeros in the vocal-tract transfer function due to the greater glottal opening. Perceptual data in the literature, based on synthesized and natural tokens of male vowels, suggest that the relative amplitude of the fundamental component is the most important cue to breathiness. Our experiments utilizing male and female utterances will challenge this conclusion.

3. The voicing source: Male/female differences

Many vocal characteristics differ between men and women. Some are due to anatomical differences such as a larger larynx and slightly longer vocal tract for the average man. The female vocal tract is about 15% shorter, although most of the difference is in the pharynx (Goldstein, 1980), so formant-frequency differences between the genders are vowel specific (Peterson and Barney, 1952; Fant, 1975). The average female f_0 is about 1.7 times that of the average male (Peterson and Barney, 1952; Cooper and Sorenson, 1981). Cooper and Sorenson also note that, for read sentences, a 1.7 ratio tends to map male and female contours onto one another with remarkable fidelity. On the other hand, Brend (1975) presents data suggesting systematic male/female differences in stereotypical f_0 contours under more spontaneous conditions. Other gender differences, such as the use of a breathy voice quality, the deployment of a more dynamic intonation contour for females (Thorne *et al.*, 1983), or differential dialects for the two genders (Kahn, 1975) appear to be learned behaviors. Some male/female differences are adopted well before puberty (Sachs *et al.*, 1973; Meditch, 1975; Karlsson, 1987). Transsexuals attempting to imitate male or female stereotypes have been found to speak louder, with a lower pitch, a reduced pitch range (measured in semitones), and faster when "male" (Gunzburger, 1987). In female mode, the reverse is true, and there is a slight tendency for $F3$ to be raised, as might be explained by a slightly raised larynx posture.

Theoretically, vowels produced with a higher fundamental frequency should be less intelligible due to the fewer harmonics present to define the shape of the vocal-tract transfer function. Experimental evidence for this tendency has been obtained using synthesis (Ryalls and Lieberman, 1982), and it is certainly the case that automatic formant trackers have more difficulty with the higher f_0 of a female voice. However, it appears that other factors dominate the determination of overall intelligibility of speech. Margulies (1979) compared five male and five female readers for intelligibility of sentence materials in various noise conditions, and obtained a significant (73% vs 56%) intelligibility advantage for female speakers. In a study of vowel formant data from several speaking conditions, Koopmans-van Beinum (1980) found that female speakers typically display a tendency toward more careful articulation. This may be due in part to a slower average speaking rate for women, which is known to be one of the most effective factors to enhance intelligibility (Picheny *et al.*, 1985, 1986). In the remainder of this section, we concentrate on differences in voicing source characteristics between men and women.

4. Measurement techniques

The volume velocity waveform and spectrum of the voicing source must be inferred by indirect techniques. One procedure is to use high-speed photography to measure glottal area as a function of time (Farnsworth, 1940), and then employ a simple model of glottal impedance to infer the volume velocity waveform (Flanagan, 1958). Recent modeling efforts have incorporated more realistic circuits to simulate the impedances of the subglottal and supraglottal tracts, leading to rather complex time-varying relationships between area and glottal flow (Fant, 1986). Another technique is to record a sound-pressure waveform at some distance from the lips and then formulate an inverse filter that cancels the effects of the vocal-tract transfer function, resulting in a waveform analogous to the derivative of the glottal volume velocity. A third method employs a reflectionless tube to neutralize the effects of the vocal-tract transfer function (Sondhi, 1975). However, reflectionless tubes interface to the subject in an artificial way, and auditory feedback is unnatural, which may cause subjects to act less naturally. The following paragraphs review what has been learned from these and other related approaches.

Monsen and Engebretson (1977) had subjects phonate a neutral vowel while placing their lips over a reflectionless tube, and thus were able to obtain voicing source waveforms and spectra directly for five male and five female speakers under various conditions. They found harmonic spectra to have rather irregular amplitudes, but, on the average, a male voice had a spectrum that fell off at about 12 dB/oct initially, and about 15 dB/oct at higher frequencies. Females had a somewhat greater tilt measured in dB/oct, but basically a female source waveform was about the same shape as a male waveform except (1) the fundamental frequency is higher and (2) the open quotient is slightly larger. The authors observed changes to waveform open quotient as a function of syllable stress, final f_0 fall, and question rise. Generally, these f_0 maneuvers were performed with a relatively constant open quotient, except for the question rise gesture, during which open quotient increased. In some final falls with glottalized offsets, the open quotient decreased. All of these tendencies are consistent with data that we will report below.

Sundberg and Gauffin (1979) used an inverse-filtering technique to examine the glottal waveforms of five male speakers. Though limited to frequencies below 1 kHz, they were about to quantify open quotient, which remained remarkably constant with changes to f_0, and they were able to

measure the relative intensity of the fundamental component in the harmonic spectrum, noting that it was weaker in the type of pressed voice characterized by a small open quotient. A range of more than 15 dB in the intensity of the first harmonic was observed from pressed to breathy phonation types, which is again consistent with our observations described below. Cleveland and Sundberg (1983) studied the open quotients of a tenor, baritone, and bass singer over an octave of notes sung at low, medium, and high vocal effort. All singers were consistent in maintaining an open quotient of about 0.5 at 165 Hz, while the open quotient generally grew to about 0.7 as f_0 increased.

Karlsson (1985) examined airflow and subglottal pressure for six female speakers at three values of f_0 and three levels of vocal effort using an airflow mask (Rothenberg, 1973). She found some dc flow during the nominally closed phase of the glottal cycle for only two speakers. The open quotient was determined from visual examination of the flow traces, and it was found to increase with increased effort, but to be relatively constant over changes to f_0. Impressionistically, female speakers of Swedish are judged to be less breathy than American women. This may account for the lack of an expected flow leakage for most of her female speakers—and may also indicate that breathy voice quality is a learned behavior in female speakers of English.

In a study of low vowels produced by 20 male and 16 female talkers of RP English, Henton and Bladon (1985) found that the amplitude of the first harmonic (relative to the amplitude of the second harmonic) was about 6 dB stronger on average for the female speakers. Since the vocal-tract transfer function is essentially flat in this frequency region for low vowels, the difference must be attributed to voicing source characteristics—in particular, a greater open quotient for the female speakers. The implication is that RP females ought to sound more breathy than males, which the authors argue conforms with subjective stereotypes

Stroboscopic motion pictures by Bless *et al.* (1986) suggest that about 80% of normal females and 20% of males have a visible posterior glottal aperture during the nominally closed portion of a vocal period.

Examples of inverse-filtered glottal-flow waveforms from several representative male and female speakers of English, obtained using a Rothenberg (1973) flow mask, have been reported by Holmberg *et al.* (1988). Representative periods were sampled from the middle of the vowels contained in a string of [pa] syllables for three vocal effort conditions (soft, normal, and loud). These conditions were realized in part by changes to subglottal pressure, and in part by laryngeal adjustments such that softer vocal effort is often more breathy (increased dc flow, increased open quotient, less abrupt closure), and louder is often somewhat laryngealized (reduced open quotient). These data indicate that, in normal voice, female subjects tend to be more breathy than males, but that both populations have a not-insignificant dc flow under many conditions when a vowel is surrounded by voiceless consonants.

In summary, the similarities and differences between typical male and female glottal volume velocity waveforms have been revealed by inverse-filtering techniques and other data. On average, the female fundamental frequency is about 1.7 times that of a male, and the open quotient is slightly larger, but otherwise the general shape and spectra of the two source waveforms are similar. These data do not address the question of whether a typical female source spectrum contains more aspiration noise at high frequencies—an issue addressed in the present study. Data on source changes over the course of an utterance suggest that, in general, the open quotient remains remarkably constant as f_0 varies but may increase for an utterance-final question rise, and may decrease slightly for a laryngealized offset.

This literature review reveals an extensive list of prior publications on many aspects of the nature of voice quality variations. The field seems ripe for a synthesis of these ideas by (1) examining a relatively large corpus of speech obtained from male and female speakers for evidence of variation in as many parameters as possible related to laryngealization and breathiness and (2) formulating observations into a single testable synthesis model that can then be used to investigate the relative perceptual importance of each potential cue to voice quality variation. These are the main objectives of the present paper.

I. SPEECH ANALYSIS EXPERIMENTS

The database for analysis consisted of two sentences having differing patterns of stressed and unstressed syllables, together with reiterant imitations of these sentences using the syllables [ʔV] and [hV], where V = [i, æ, ɑ, o, ɝ]. Reiterant materials were chosen so as to be able to quantify acoustic correlates of breathiness in different sentence positions, while holding constant other potentially confounding segmental effects such as the voicing feature of bounding consonants. The two underlying sentences were:

(S1) "Steve eats candy cane"

[ʔV́ ʔV̀ ʔV́ ʔV ʔV]

[hV́ hV̀ hV́ hV hV]

and

(S2) "The debate hurt Bob"

[ʔV ʔV ʔV́ ʔV̀ ʔV́]

[hV hV hV́ hV̀ hV́].

The stress notation provided above suggests secondary stress for the sentence verbs, although subjects tended to produce fully stressed verbs in this "deliberate" style of speaking. Only data involving the vowel [ɑ] will be described in the present paper.

Subjects were recruited from the Speech Communication Laboratory at MIT and were recorded in a sound-isolated chamber. An Altec model 684A omnidirectional dynamic microphone was placed approximately 12 in. in front of the lips and two in. above the breath stream. Recordings were made on a Yamaha K-1000 cassette recorder (with Dolby and dbx disabled in order to ensure that onsets were not distorted) and then were digitized at 10 000 12-bit samples/s and stored on a VAX-750 computer disk for subsequent analysis.

Ten female and six male subjects were recorded. The age

TABLE I. Dialect history of ten female and six male subjects.

	Age	Dialect history
Females		
KK	24	0–3 Cincinnati OH; 4–6 Hiroshima Japan; 7–13 Albuquerque, NM
LK	18	0–13 Cambridge, MA
CB	37	0–13 St. Louis, MO
LG	23	0–13 southeastern MA
SS	39	0–4 Missouri; 5–6 Connecticut; 7–13 Tennessee
LL	29	0–13 Long Island, NY
ND	23	0–3 upstate NY; 4–5 New Jersey; 6–13 Rhode Island
SH	45	0–3 Syracuse, NY; 4–13 Philadelphia, PA
CE	30	0–13 Atlanta, GA
JW	24	0–13 Connecticut
Males		
KS	62	0–13 Toronto, Canada
MR	29	0–13 Cincinnati, OH
JG	26	0–13 Ottawa, Canada
JP	47	0–6 Long Island, NY; 7–13 Miami, FL
MP	25	0–6 Los Angeles, CA; 7–13 South Carolina
TW	30	0–13 Nashville, TN

and dialect history of each subject (where they spent the first 13 years of their lives) are listed in Table I. A wide range of dialects is represented, but, subjectively, the vowel qualities produced are quite similar to one another, with only a few exceptions for [o].

In order to illustrate some of the hypothesized acoustic cues to breathiness, broadband sound spectrograms for two female speakers, one perceived to be nonbreathy and the other perceived to be breathy, are shown in Fig. 2. Reiterant imitations of a sentence, involving the syllables [ʔɑ] and [hɑ] are shown in the middle and lower panels of the figure, respectively.

Based on pilot analyses of data from a single female subject, DB, we isolated three primary candidate cues to perceived breathiness (Klatt, 1986b). The first is the relative strength of the first harmonic, which is known to increase as the open quotient increases, as might be expected for a breathy voice quality (Bickley, 1982; Ladefoged, 1983). This contrast is quite evident in the reiterant spectrograms of Fig. 2; there is a strong energy component at low frequencies (at about 200 Hz) during the [ɑ] vowels of the reiterant imitations for the breathy speaker CB, but not for the nonbreathy speaker SH.

The second potential cue to breathiness is the presence of aspiration noise in the vowel spectrum, particularly at higher frequencies where the noise may actually replace harmonic excitation of the third and higher formants (Ladefoged and Antoñanzas-Barroso, 1985; Klatt, 1986b). The presence or absence of noise excitation is difficult to determine from the spectrographic data illustrated in Fig. 2 (see Footnote 7), but we will show by other means (a plot of the waveform bandpass filtered to include only the $F3$ region) that there is a significant difference between CB and SH in amount of noise excitation of $F3$.

The third class of potential cues to breathiness has to do with changes to the vocal-tract transfer function when the glottis is partially abducted. One such cue is the presence of extra poles (formants) and zeros (energy gaps) in the vowel spectrum due to acoustic coupling to the trachea (Fant *et al.*,

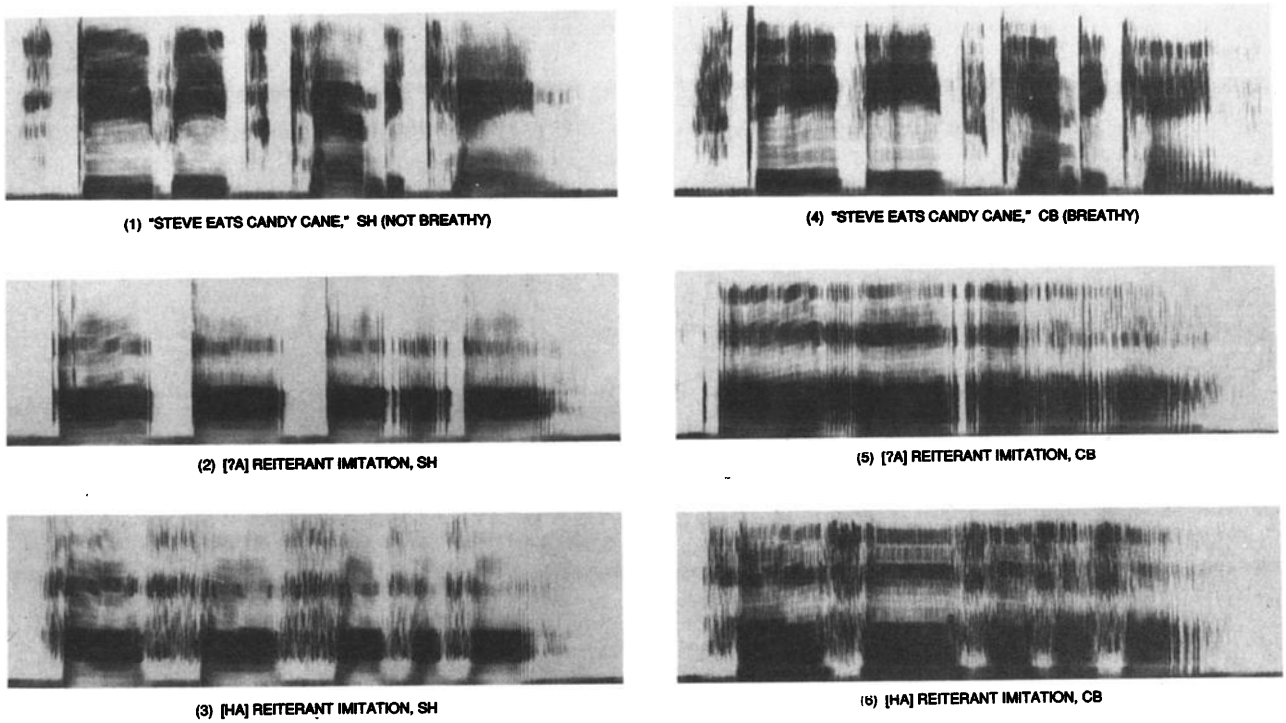


FIG. 2. Broadband spectrograms (0–5 kHz) of the target sentence S1 "Steve eats candy cane" and two reiterant imitations from a nonbreathy female speaker SH and a breathy female speaker CB.

1972; Klatt, 1986b). An extra resonance, at about 2200 Hz, can be seen in the [h] spectra of Fig. 2 for both SH and CB, and this tracheal formant continues to be visible in the initial portion of the following vowel. Another effect of an open glottis on the transfer function of the vocal tract is an increase in the bandwidth of the first formant. This increase can be quite large for a low vowel (Fant and Ananthapadmanabha, 1982).

A. Analysis techniques

The recordings were low-pass filtered at 4.8 kHz using a TTL passive seven-pole elliptical low-pass filter, and then digitized at 10 000 samples/s using a VAX-750 computer. All subsequent analysis was done by computer software (Klatt, 1984). Analysis techniques employed in the study include:

(1) display of the *waveform* useful for determining the nature of voicing onsets and offsets or large deviations from perfect periodicity;

(2) *bandpass-filtered waveform*: useful to isolate a single formant in order to determine if a formant is excited more than once during a period, or is noise-excited;

(3) *digital spectrogram*: an approximation to a broadband sound spectrogram, limited in dynamic range by a one-bit gray scale, useful for determining times at which to measure short-term spectra in a reiterant utterance;

(4) f_0 *versus time*: an estimate of voicing fundamental frequency, derived by a harmonic sieve technique (Duifhuis *et al.*, 1982), useful in determining the time course of f_0 as well as deviations from a smooth contour;

(5) *spectral cross sections*: a short-term discrete Fourier transform (dft) magnitude spectrum of a windowed waveform segment, as well as a smoothed spectrum similar to that obtained by a bank of 256 critical band filters with bandwidths of 70 Hz at low frequencies, 160 Hz at 1 kHz, and bandwidths that increase in proportion to filter center frequency thereafter, useful for estimating the auditory–perceptual representation of vowels, and for determination of the frequency locations and the relative amplitudes of higher formant peaks; and

(6) *average spectrum*: sum of the energy in a sequence of overlapping short-term dft magnitude spectra, useful in the analysis of statistically fluctuating noise-excited speech sounds such as [h].

Inverse filtering, a technique commonly employed to examine the details of glottal volume velocity waveforms (Sundberg and Gauffin, 1979; Fant, 1979, 1982a; Karlsson, 1985) was not used here because (1) the recordings would have required a better low-frequency phase response in order to preserve the relative phases of low-frequency harmonics and (2) inverse filtering usually restricts the frequency region to below about 1.5 kHz, whereas we are particularly interested in the characteristics of the source spectrum above 1.5 kHz.

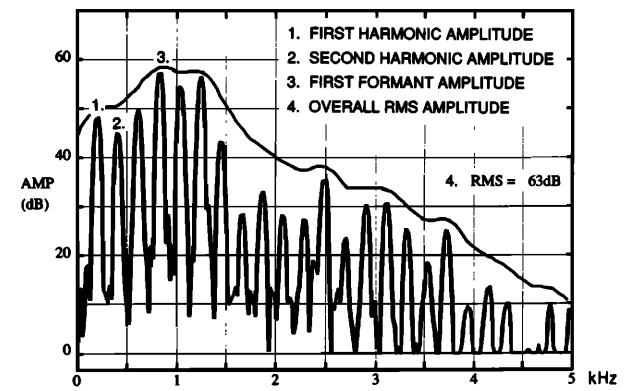
The next three sections describe our efforts to infer source characteristics related to breathiness from the reiterant data. In the first section, the relative amplitude of the first harmonic is quantified and interpreted in terms of the open quotient. Following that, the third formant region of the spectrum is examined to see if the waveform is essentially periodic, or is noise excited. Finally, we quantify some of the effects of tracheal coupling on the vocal-tract transfer function, as revealed by detailed examination of [h] noise spectra.

B. Results I: Amplitude of first harmonic

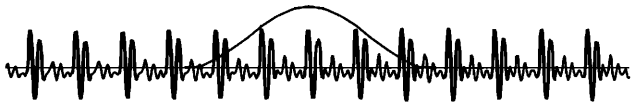
Figure 3 illustrates several methods for quantifying the relative amplitude of the first harmonic in a vowel spectrum. The spectrum has been computed without the usual first-difference operation so as to measure the true amplitude of the first harmonic, H1, relative to other frequency components of the spectrum.

The perceptual importance of H1 as an auditory cue is difficult to estimate a priori because (1) typical background noises have much of their energy at low frequencies, which could mask the detection of H1 in many normal listening situations (as well as over the phone) and (2) psychological equal-loudness contours indicate some attenuation of low frequencies relative to $F1$; for example, see Robinson and Dadson (1956). At the typical speech level of 60 dB SPL, low frequencies (below 300 Hz) are subjectively less loud by about 4 dB per octave of frequency decrease in a free field. The usual first-difference operation employed in speech-processing algorithms attenuates low frequencies by 6 dB per octave and extends the attenuation above 300 Hz, so the most appropriate speech-processing transformation to apply when preparing figures to visualize the strength of H1 is not clear.

As it can be seen in the dft spectrum of Fig. 3, the first-harmonic amplitude is about 48 dB. In order to determine whether this number is large or small, it must be compared with some reference that takes into account recording level,



(a) SHORT-TERM DFT SPECTRUM AND 330-HZ BANDWIDTH SMOOTHED SPECTRUM



(b) DIGITIZED WAVEFORM AND 25.6 MS HAMMING-WINDOW

FIG. 3. A dft spectrum and 300-Hz bandwidth smoothed spectrum illustrating the method used to quantify the amplitude of the first-harmonic (1) relative to (2) second-harmonic amplitude, (3) first-formant amplitude as estimated from the smoothed dft spectrum, or (4) rms amplitude; see text. The waveform and the time window used to calculate the dft are shown below.

such as: (a) rms amplitude of the vowel (63 dB); (b) amplitude of the second harmonic (45 dB) (Bickley, 1982); or (c) amplitude of the first formant (58 dB) (Ladefoged, 1983). We have compiled data on all of these potential references, although, for theoretical reasons, we prefer to use second-harmonic amplitude as a reference in the following analysis.[8] However, it appears that the choice of reference does not matter very much insofar as group averages are concerned; in all but one case, there is, at most, a 1-dB difference in the average relationships across conditions between the three choices for a reference.

The amplitude in dB of the first harmonic has been measured in the middle of each syllable of both reiterant utterances, spoken with either [ʔɑ] or [hɑ] replacing each syllable. The results of comparisons of ten female speakers and six male speakers are listed in Table II. Vowel midpoint was chosen so that measurements would be least affected by the voicing feature of adjacent consonants. The values in Table II indicate the difference in dB between the first-harmonic amplitude and second-harmonic amplitude, scaled up by 10 dB so that most numbers are positive. Averaged across all female data, the first-harmonic amplitude defined in this way is 11.9 dB. A comparable grand average for males is 6.2 dB, indicating that the first harmonic is weaker on average for males. The difference between the sexes is about 5.7 dB. To the extent that the first-harmonic amplitude is an acoustic correlate of breathiness, females are more breathy than males, in agreement with prior research (Henton and Bladon, 1985).

Comparing the [ʔɑ] data with the [hɑ] data, we see from the averages in Table II that there is very little difference in first-harmonic amplitude between these two versions of the sentence. Since the measurements were made in the midpart of the vowel, it must be the case that any extra influence associated with [hɑ] does not extend into the middle of the vowel. Even so, perceptual data described in Sec. E below indicate that [hɑ] sentences are perceived to be significantly more breathy than [ʔɑ] sentences.

TABLE II. Amplitude of the first harmonic plus 10 dB, relative to the amplitude of the second harmonic, as averaged across speakers of a given gender for each position in several five-syllable reiterant sentences. The average male–female difference is 5.7 dB.

			Female averages			
Sentence	Syll1	Syll2	Syll3	Syll4	Syll5	Av
S1 [ʔɑ]	12.8	12.3	11.6	12.6	13.1	12.5
S1 [hɑ]	13.2	12.5	12.2	11.4	9.3	11.7
S2 [ʔɑ]	11.6	11.8	11.8	12.1	11.1	11.7
S2 [hɑ]	13.7	12.5	12.3	12.1	8.2	11.8
Av	12.8	12.3	12.0	12.0	10.6	11.9
			Male averages			
Sentence	Syll1	Syll2	Syll3	Syll4	Syll5	Av
S1 [ʔɑ]	6.7	5.5	4.3	3.9	4.4	5.0
S1 [hɑ]	7.2	5.3	6.8	6.1	6.4	6.4
S2 [ʔɑ]	6.4	5.8	8.0	7.6	4.4	6.4
S2 [hɑ]	9.8	6.3	8.3	5.8	4.5	6.9
Av	7.5	5.7	6.9	5.9	5.0	6.2

Comparing the first and last vowels, we observe from the averages of Table II that the first harmonic is about 2 dB weaker (*re*: the second-harmonic amplitude) in the last syllable for female speakers, and about 2.5 dB weaker in the last syllable for male speakers. Thus it would appear that both groups tend to laryngealize slightly during the f_0 fall of the final syllable of an utterance, and this presumably causes the open quotient to be slightly reduced. Male speakers appear to laryngealize only slightly more in this set of data, although previous research by Henton and Bladon (1987) suggests that, in general, males laryngealize far more often than females.

There is considerable subject-to-subject variability in the measurement of the first-harmonic amplitude, especially for the last syllable of each utterance. Individual variation is quantified in Table III. Among the female speakers, there is a wide range of values for relative first-harmonic amplitude. Speaker CB, who is perceived to be breathy in voice quality, and presumably employs a speaking mode with a large open quotient, has a relative first-harmonic amplitude of 17.1 dB. Speaker SH, on the other hand, is perceived to have a laryngealized voice quality and has a relative first-harmonic amplitude of 8.4 dB. In order to explain this difference of 8.7 dB, a fairly large difference in open quotient must be postulated.[9]

Other female speakers fall in a continuum between these two extremes. The large range offers the opportunity to investigate the perceptual salience of first-harmonic amplitude by obtaining judgments of breathiness from these sentence materials; results of such a test will be presented in Sec. E below. These average H1-amplitude data suggest that males

TABLE III. Individual data concerning amplitude of the first harmonic plus 10 dB, relative to the amplitude of the second harmonic, as averaged across the five syllables of reiterant sentences S1 and S2. A data point followed by an asterisk indicates unexplained variability in that the value is more than 2 dB different from the mean for the speaker.

Females	S1 [ʔɑ]	S1[hɑ]	S2[ʔɑ]	S2[hɑ]	Av
KK	13.0	12.4	13.0	13.6	13.0
LK	13.6	10.6	14.8*	11.2	12.6
CB	17.0	17.8	16.8	16.8	17.1
LG	12.8	13.6	11.6	12.6	12.6
SS	10.6	9.2	7.6	10.2	9.4
LL	10.4	11.8	9.2	10.8	10.3
ND	14.2*	9.4*	12.2	11.8	11.9
SH	8.8	7.6	8.2	9.0	8.4
CE	12.0	13.4	12.4	12.2	12.5
JW	12.8	11.8	11.0	11.6	11.8
Av					11.9
Males					
KS	6.2	5.0	7.0	5.0	5.8
MR	2.8*	5.8	4.2	8.4*	5.3
JG	3.4	4.6	4.6	6.0	4.6
JP	2.6*	4.8	5.8	6.2	4.9
MP	5.2	8.8*	4.0	5.8	6.0
TW	9.4	9.8	9.2	10.4	9.7
Av					6.2