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

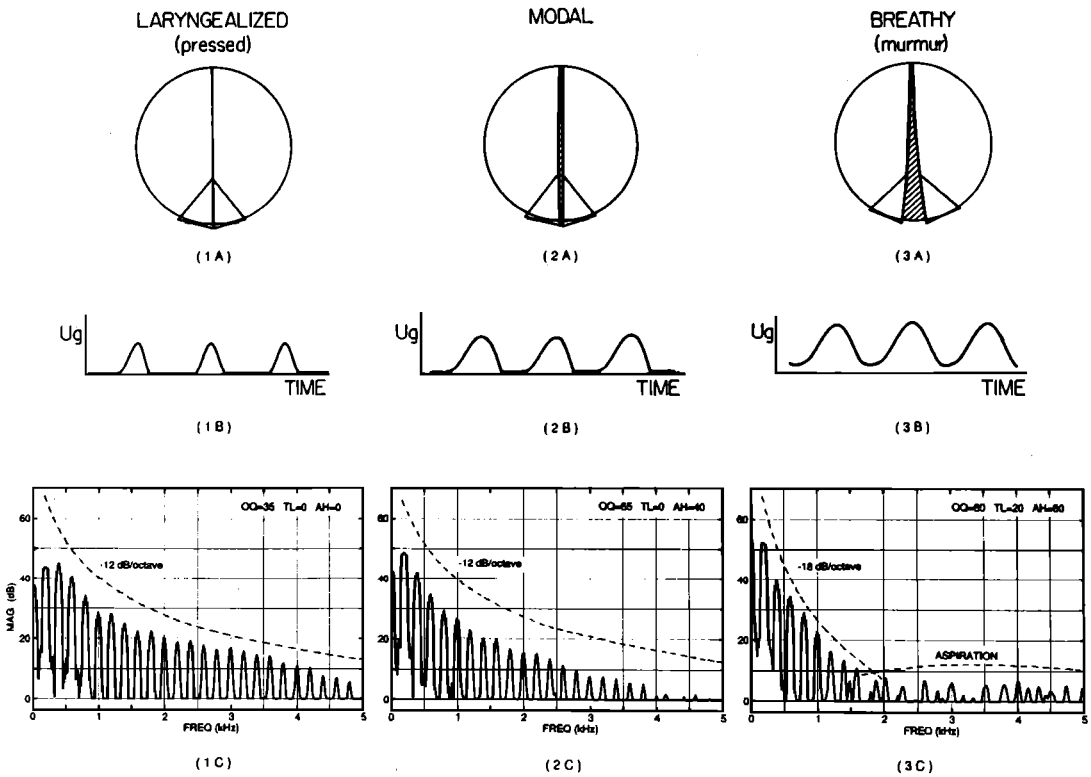


FIG. 1. Glottal configurations (row A) for (1) laryngealized, (2) modal, and (3) breathy vowels. An increased opening at the arytenoids results in glottal volume velocity waveforms (row B) with a progressively longer duration open period, an increased dc flow, and a less abrupt closure event. The source 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 [fi] 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, ɜ˞]. 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.*,

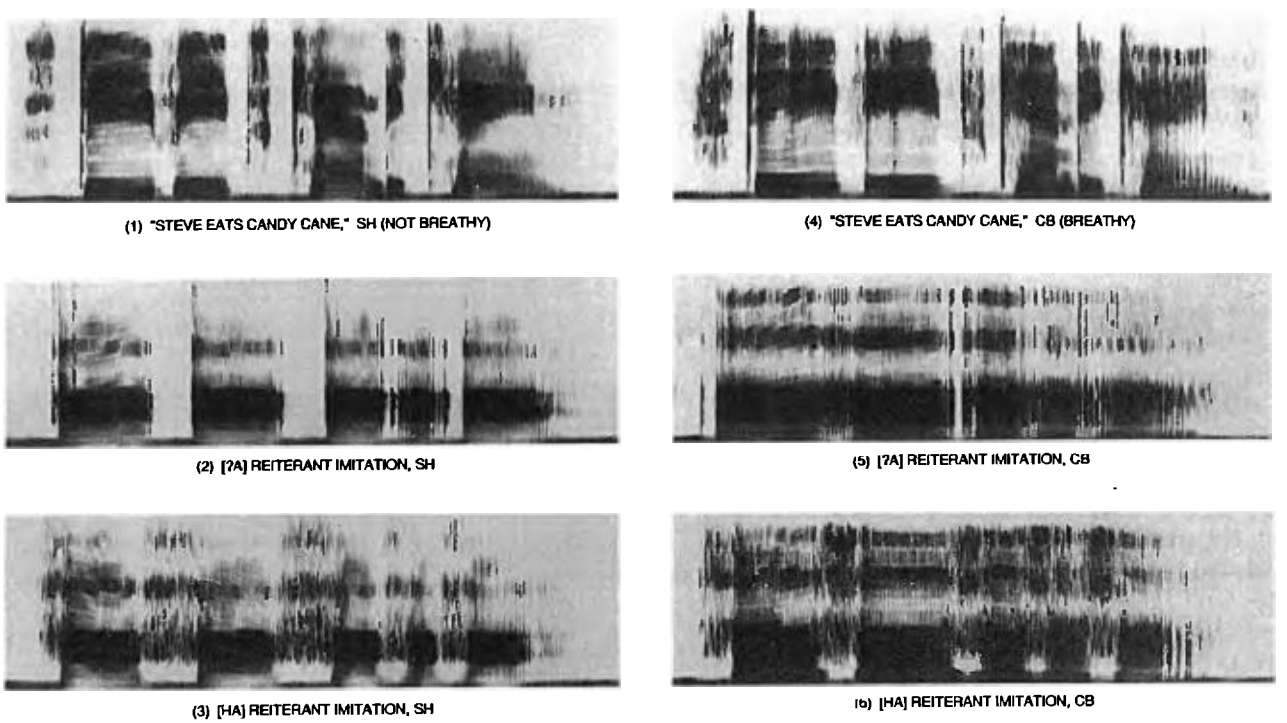


FIG. 2. Broadband spectrograms (0–5 kHz) of the target sentence S1 "Steve eats candy cane" and two reiterant imitations from a nonbreathy female speaker SH and a breathy female speaker CB.

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.

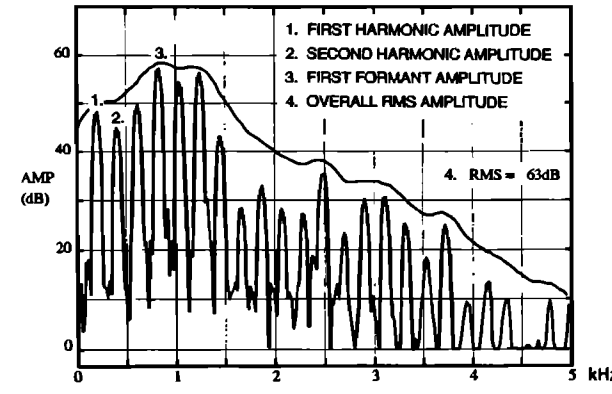
The next three sections describe our efforts to infer source characteristics related to breathiness from the reiterant data. In the first section, the relative amplitude of the first harmonic is quantified and interpreted in terms of the open quotient. Following that, the third formant region of the 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

have a shorter open quotient than females, except for male TW who is more breathy than two females according to this measure.[10] The range of H1 amplitude across subjects is considerable—from a maximum of 17.1 dB for CB to a minimum of 4.6 dB for male JG—i.e., the range is 12.5 dB. Within-subject variability is generally low, but there are clear examples of particular sentences that are uttered with a different average first-harmonic amplitude. A value in Table III is followed by an asterisk if it differs from the average for that subject by more than 2 dB. There are only seven such examples in the database, suggesting that subjects are free to select differing modes of vibration and degrees of breathiness at will but tend to stay at one mode during an experiment of this sort.

1. Stress and syllable amplitude

Data on rms amplitude measured at syllable midpoint with a 25-ms Hamming window are presented in Table IV. Stressed syllables, on average, are 3.0 dB more intense than unstressed syllables in this corpus for females, and 4.3 dB more intense for males. Some of the difference between stressed and unstressed syllables is simply due to differences in fundamental frequency (a vowel uttered at a higher fundamental has more pulses per second within the analysis window, all else being equal), while part of the difference is presumably due to vocal effort. Other confounding factors are the possibility of vowel reduction and a lowered $F1$ for unstressed vowels, and the possible increased breathiness of unstressed vowels that might, among other effects, increase the first-formant bandwidth and thereby reduce the amplitude of the vowel. Thus, while we can say that there are moderately large differences in intensity between stressed and unstressed vowels, it is not easy to quantify the contribution of the various factors that may be involved.

In addition to a stress effect, there appears to be a general tendency for syllable amplitude to be less at the end of a sentence, even when terminated by a nominally stressed syllable. Across all 16 subjects, the average rms amplitude of syllables 1, 3, and 5 for sentence S1 (ɑ̀ɑ̀ɑ́sɑ) is 58.1, 56.7, and 49.9 dB, and for sentence S2 (sɑ́sɑ̀ɑ́) is 55.8, 58.2, and 53.4 dB. If we first subtract 3 dB from each stressed syllable to normalize out the previously described stress effect on syllable amplitudes, the utterance position effect is found to be a 1.0-dB fall from syllable position 1 to position 3, and a more pronounced 4.3-dB fall from syllable position 3 to final position 5. Presumably, this amplitude reduction in utterance-final position is associated with reduced lung volume and is a natural consequence of lowered subglottal pressure at the end of a breath group (Lieberman, 1967). In addition, there appears to be a general relaxation of muscular activity and a preparation for breathing in the larynx musculature that could also contribute to the reduction in voicing source amplitude.

TABLE IV. Rms amplitude in dB at the midpoint of each syllable in several five-syllable reiterant sentences, as averaged across speakers of a given gender.

Female averages Sentence	Syll1	Syll2	Syll3	Syll4	Syll5	Av
S1 [ʔɑ] (ɑ̀ɑ̀ɑ́sɑ)	58.6	57.0	56.7	52.2	50.8	55.0
S1 [hɑ] (ɑ̀ɑ̀ɑ́sɑ)	58.8	58.4	57.8	54.0	51.6	56.1
S2 [ʔɑ] (sɑ́sɑ̀ɑ́)	57.7	54.8	59.1	58.0	55.7	57.1
S2 [hɑ] (sɑ́sɑ̀ɑ́)	57.2	53.3	58.2	56.2	52.6	55.5
Male averages						
Sentence	Syll1	Syll2	Syll3	Syll4	Syll5	Av
S1 [ʔɑ] (ɑ̀ɑ̀ɑ́sɑ)	56.8	55.7	55.3	50.7	48.0	53.3
S1 [hɑ] (ɑ̀ɑ̀ɑ́sɑ)	57.5	55.2	55.8	50.7	47.7	53.4
S2 [ʔɑ] (sɑ́sɑ̀ɑ́)	53.0	51.6	58.2	56.2	54.4	54.7
S2 [hɑ] (sɑ́sɑ̀ɑ́)	52.8	51.7	56.3	55.0	50.0	53.2

The fundamental frequency was also measured at vowel midpoint. The average f_0 of females, relative to the male data, was found to be remarkably systematic—i.e., about 1.7 times that of males for each vowel position in each sentence. Within each gender, there was considerable variation in average f_0. It is possible that f_0 could be a secondary perceptual cue to judged breathiness. For example, a high f_0 might indicate a desire to be "feminine" and thus breathy, or a very low f_0 might be indicative of laryngealization. However, correlation data to be presented below indicate that, within each sex, average f_0 of a talker is not at all correlated with perceived breathiness rating.

In summary, the most striking aspect of first-harmonic amplitude is the large variation across speakers of a particular gender. On average, the relative amplitude of the first harmonic for females is about 6 dB greater than that for males, but, within the class of female speakers, the range is over 10 dB. As an indirect measure of open quotient, first-harmonic amplitude indicates that the final vowel of a reiterant utterance is likely to be slightly laryngealized (reduced open quotient), and lower in overall amplitude.

C. Results II: Aspiration noise in the *F*3 region of the spectrum

A second potential acoustic correlate of the degree of breathiness of a vowel is the amount of noise present at higher frequencies in the spectrum. One way to estimate the relative strength of noise components is to isolate the third formant, using a bandpass filter.[11] The filtered waveform can be displayed, as in Fig. 4, and examined visually to determine whether the waveform is periodic (repeating itself identically) or has indications of random variation due to the introduction of noise. Note that it is difficult to determine whether noise is present by examining the unfiltered speech waveform shown in parts (2a) and (3a) in Fig. 4 because the periodic components at low frequencies dominate the visual impression due to their greater energy.

Reiterant sentences involving [hɑ] were processed in three steps: (1) a broadband spectrogram was produced, and the frequency of the third formant, $F3$, was estimated visually (see column 2 of Table V); (2) a four-pole Butterworth bandpass filter, having a center frequency set equal to $F3$ and a bandwidth of 600 Hz, was used to create a filtered version of the original digitized waveform; and (3) a plot of the waveform was examined subjectively to determine the degree of random noise present. A four-step scale, described in Table V, was used to quantify the presence or absence of

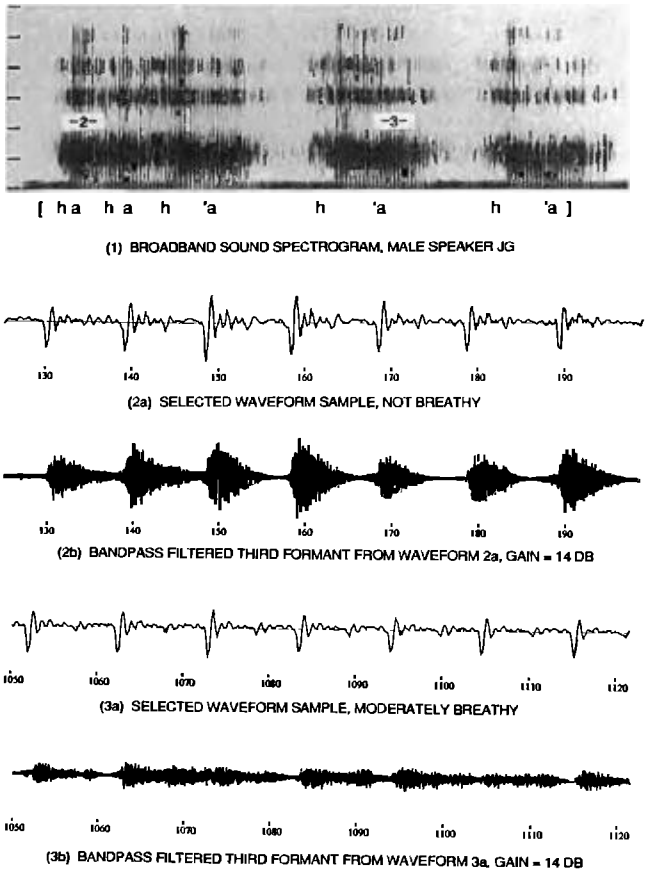


FIG. 4. (1) Broadband spectrogram of male speaker JG indicating locations where waveform samples have been extracted to show (2) a vowel with little or no aspiration noise, and (3) a vowel with appreciable aspiration noise, as evidenced by the presence/absence of noise in the third formant region of the spectrum.

noise over the course of each vowel. If the filtered $F3$ waveform consisted of a periodic damped sinusoid, in synchrony with the unfiltered waveform, the vowel was judged to be periodic and free from aspiration noise. If there was no visible periodicity in synchrony with the original waveform, the vowel was judged to have strong aspiration noise. While it has been assumed that the noise source is at the glottis (aspiration), it is possible that some speakers employ a pharyngeal constriction to augment [h] noise with the frication noise of a pharyngeal fricative before a vowel such as [ɑ].

The degree of aspiration noise in each of the five syllables of the reiterant utterance S1, "Steve eats candy cane," is summarized in the middle five columns of Table V. This rating of noise presence in $F3$ for each syllable, as estimated by the first author, is tabulated separately for female and male subjects.

Comparing the noisiness rating across syllable position, we observe that noise increases toward the end of the utterance. This tendency may be a characteristic of all utterances, or it may be due to the fact that this particular utterance ends with two unstressed syllables. We will return to the question of whether a syllable tends to have more aspiration noise if unstressed and/or if in utterance-final position after examining data from the utterance S2, "The debate hurt Bob."

TABLE V. Degree of periodicity versus noise excitation of $F3$ for [hɑ] reiterant imitations of sentence S1, "Steve eats candy cane." The 4-point subjective scale ranges from: (1) periodic, no visible noise, (2) periodic but occasional noise intrusion, (3) weakly periodic, clear evidence of noise excitation, and (4) little or no periodicity, noise is prominent.

Females	$F3$	Syll1	Syll2	Syll3	Syll4	Syll5	Av
KK	2300	1	1	1	2	2	1.4
LK	2900	3	3	3	4	4	3.4
CB	2950	2	3	3	3	4	3.0
LG	2650	3	2	3	3	3	2.8
SS	2800	2	3	3	3	4	3.0
LL	2550	1	2	1	2	2	1.6
ND	2400	2	3	3	4	3	3.0
SH	2700	2	2	2	2	2	2.0
CE	2950	3	2	3	4	4	3.2
JW	2900	3	3	4	4	3	3.4
Av		2.2	2.4	2.6	3.1	3.1	2.7
Males							
KS	2700	1	1	1	1	2	1.2
MR	2200	1	1	2	2	3	1.8
JG	2700	1	2	2	3	3	2.2
JP	2350	1	1	1	1	2	1.2
MP	2600	1	1	1	1	1	1.0
TW	2200	1	1	2	4	4	2.4
Av		1.0	1.2	1.5	2.0	2.5	1.7

Comparing the noisiness rating between female and male subjects, we see that females generate more noise, on average, than males. The average noisiness rating for females is 2.7, while for males it is 1.7. However, there is a wide range of degrees of noise presence within each gender. The noise measurements for the female speakers KK, LL, and SH are approximately the same as those for the average male. The remaining female speakers exhibit even more noise than the average difference between the sexes would imply.

A reiterant sentence with a different stress pattern, "The debate hurt Bob," was analyzed in the same way as the sentence in Table V. The results are summarized in Table VI, which indicated the noisiness rating for each syllable as tabulated separately for female and male subjects.

Comparing the average noisiness rating across syllable position, we see that noise presence increases for unstressed syllables, and that noise presence increases slightly toward the end of the utterance. Factoring out these two effects from the S1 and S2 data sets, we find a stress difference (unstressed more noisy) of 0.6, and a final position effect (more noisy relative to utterance-initial position) of 0.55 subjective units. While these tendencies are not large compared with individual differences, they reflect effects that may be important for the synthesis of natural variation in voicing source characteristics over a sentence.

There seems to be a paradox when comparing the "noise-in-$F3$" measure at utterance offset with the "first-harmonic-amplitude" measure described in the previous section. In this section, we find that there is more noise, indicating a greater glottal airflow in an utterance-final syl-

lable, but in the previous section we found a weaker first-harmonic amplitude in an utterance-final syllable, indicative of a pressed voice with a slightly shorter duration open quotient. We conjecture that the natural tendency to open the larynx in preparation for breathing at the end of an utterance indeed occurs in virtually all cases, resulting in an increase in posterior glottal chink size and an increase in aspiration noise. However, most speakers simultaneously rotate the anterior tips of the arytenoid cartilages inward, presumably to maintain voicing, but actually partially laryngealize, leading to a somewhat novel breathy-laryngealized mode of vibration.

The average difference between males and females is not as great in Table VI (0.4 subjective units) as it was in Table V (1.0 subjective units). However, those female speakers who show less $F3$ noise in Table V also have less noise for the second sentence. Again, individual variation within a gender is large compared with average differences between genders.

In summary, aspiration noise is very commonly present in the waveforms extracted from the third-formant region throughout the vowel portion of [hɑ] reiterant utterances from both sexes. The aspiration noise can strongly dominate harmonic excitation, implying a changed voicing vibration pattern in which the harmonic spectrum is tilted down at high frequencies with respect to normal voicing. On average, there is more noise infusion in female than male utterances, but three of the females are not very breathy by this measure. Variation in amount of aspiration noise in $F3$ across an utterance appears to be systematic in that there is more noise; i.e., the glottis can be inferred to be slightly more spread, in unstressed syllables and toward the end of a breath group.

TABLE VI. Degree of periodicity versus noise excitation of $F3$ for [hɑ] reiterant imitations of sentence S2, "The debate hurt Bob." The four-point subjective scale ranges from: (1) periodic, no visible noise, (2) periodic but occasional noise intrusion, (3) weakly periodic, clear evidence of noise excitation, and (4) little or no periodicity, noise is prominent.

Females	$F3$	Syll1	Syll2	Syll3	Syll4	Syll5	Av
KK	2500	2	1	2	1	1	1.4
LK	2900	4	3	2	4	3	3.2
CB	3000	2	4	2	2	3	2.6
LG	2600	3	3	3	3	2	2.8
SS	2750	3	4	2	3	3	3.0
LL	2550	2	3	2	2	2	2.2
ND	2500	1	4	2	2	2	2.2
SH	2450	1	2	1	1	1	1.2
CE	2700	4	4	2	3	3	3.2
JW	2800	3	3	3	3	2	2.8
Av		2.5	3.1	2.1	2.4	2.2	2.5
Males							
KS	2700	1	1	1	1	1	1.0
MR	2100	4	3	2	3	2	2.8
JG	2750	1	3	2	3	4	2.6
JP	2400	2	2	2	1	2	1.8
MP	2500	2	1	1	1	1	1.2
TW	2250	2	3	3	3	4	3.0
Av		2.0	2.2	1.8	2.0	2.3	2.1

D. Results III: Tracheal coupling

The acoustic effects of tracheal coupling on the normal transfer function of the vocal tract for a vowel include (1) possible addition of poles and zeros associated with the tracheal and lung system below the glottis and (2) increased losses at the glottal termination, which primarily affect the first-formant bandwidth. It is difficult to objectively quantify the extent to which these potential perturbations are present in the data from our 16 speakers. Therefore, we will present as much of the primary data as possible when describing our interpretations of the data in the following analysis.

1. Extra poles and zeros

The best way to get an idea of possible locations of tracheal poles and zeros is to choose a speech sound in which the glottis is as open as possible, subject to the constraint that there is still sound generation at the larynx. Therefore, aspiration spectra during the production of the [h] portions of the [hɑ] reiterant sentence "Steve eats candy cane" were obtained and analyzed. An example of the analysis process is shown in Fig. 5. Aspiration spectra were produced using a 51.2-ms rectangular window in order to obtain a stable estimate of the spectrum of a random process (Shadle, 1987). Note the extra pole at about 2100 Hz in Fig. 5, which is prominent in the aspiration noise spectrum and is visible as a local spectral maximum between $F2$ and $F3$ in the harmonic spectrum of the following vowel.

Samples of individual [h] noise data are presented in Fig. 6, which summarize spectra of the [h] aspiration before the second vowel of the [hɑ] reiterant imitation of "Steve eats candy cane." Peaks associated with $F2$ and $F3$ of the following vowel are identified in the figure. A clear $F1$ peak is usually not visible due to increased losses and the falloff in

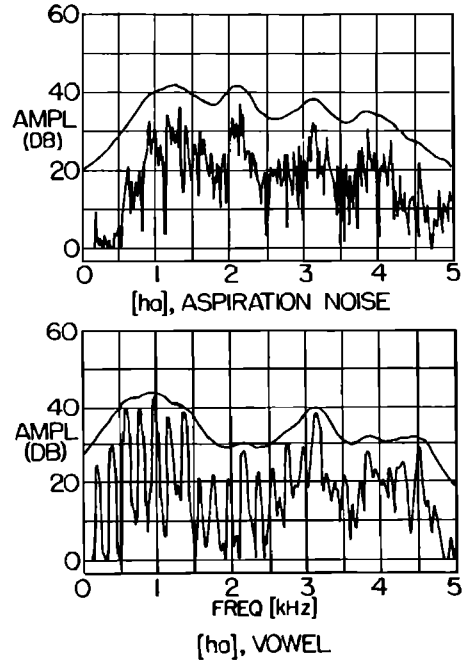


FIG. 5. A tracheal resonance at 2100 Hz is identified in an aspiration spectrum (top) and in the initial portion of the following vowel (bottom).

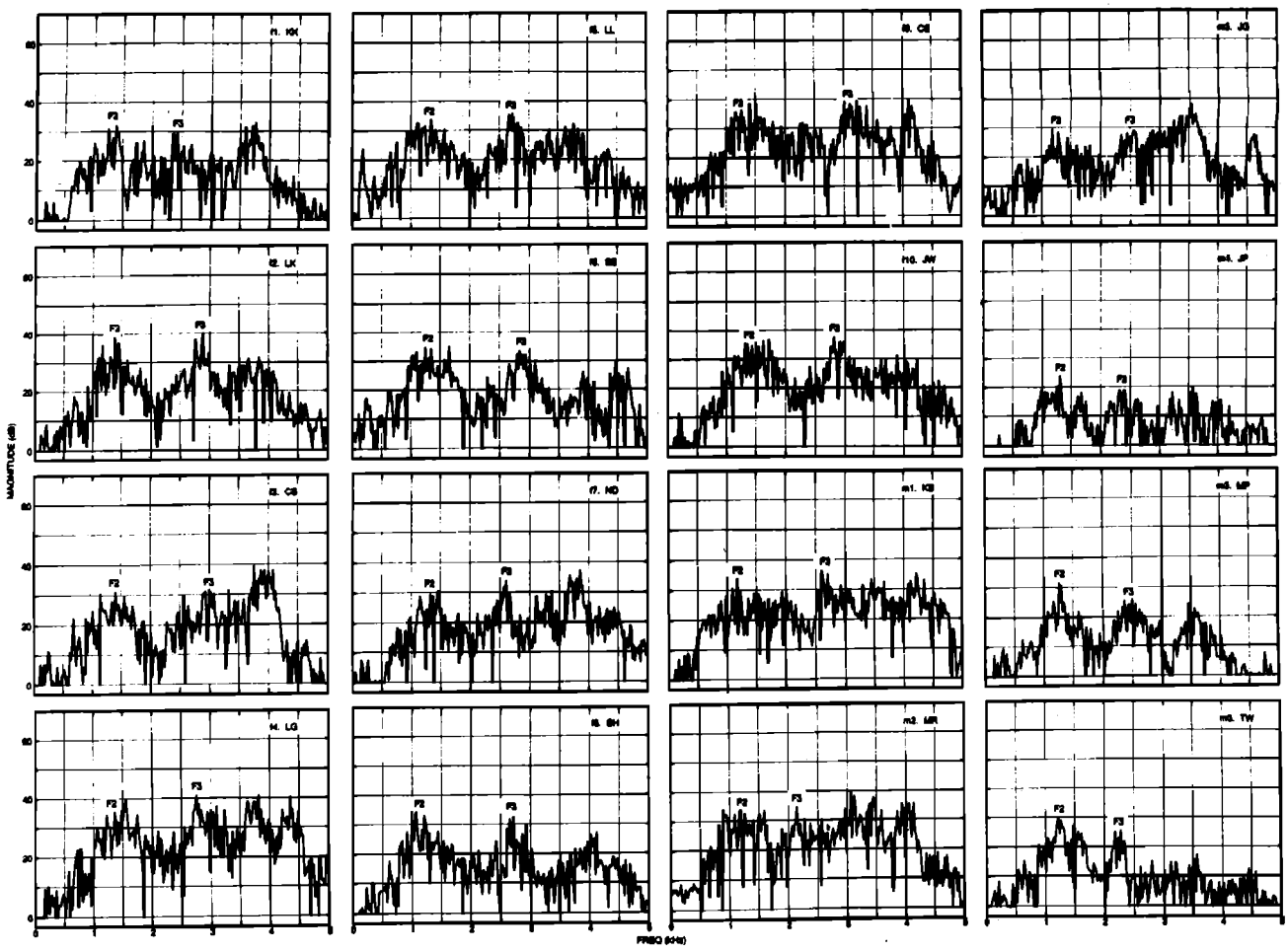


FIG. 6. Fifty-ms dft magnitude spectra of [h] aspiration noise from 16 speakers, as obtained just prior to the second vowel in [hɑ] reiterant imitations of the sentence "Steve eats candy cane."

aspiration source spectral energy at low frequencies. An appreciation for the importance of cross-speaker variability in developing an understanding of speech perception can be obtained by study of the figure. For example, a single spectral template representative of [h] before [ɑ] is unlikely to be able to account for the presumed perceptual similarity between these spectra even if powerful normalization procedures are applied to the spectral data prior to comparison with the idealized template.

Examination of aspiration spectra has revealed extra poles and zeros that often creep into adjacent vowel spectra whenever the glottis is partially spread and tracheal pole-zero coupling is possible. Therefore, as a second step in the analysis, each vowel following [h] was examined for evidence of extra poles and zeros. A 25-ms windowed dft spectrum obtained at [ɑ] vowel midpoint following the [h] is plotted for each subject in Fig. 7. The figure indicates that it is sometimes possible to see an extra formantlike peak at about 2000 Hz. When present, this peak could easily be confused with a normal formant. Other expected peaks at lower frequencies were not as easy to detect, perhaps due to the presence of close-by formants, as well as the fact that the high f_0 of females provides relatively few harmonics from which to determine spectral shape of the vocal-tract transfer function. At vowel midpoint, the effects of tracheal coupling are not great for most speakers, but during the initial and final parts of the vowel, greater departures from normal all-pole spectra were more frequently seen.

Peaks in the aspiration spectrum were compared with peaks in the spectrum of the following vowel; any peak that could not be associated with a formant of a normal [ɑ] vowel was assumed to be a tracheal resonance.[12] These extra resonances are listed in Table VII for each speaker. Also listed are the locations of prominent dips in the aspiration spectrum, which are presumably caused by tracheal zeros. Extra peaks and dips of uncertain status (either weak or present in only one or two of the five [h] tokens analyzed in the reiterant sentence) are indicated in the table by a parentheses notation. In general, the results are fairly consistent across speakers. If a pole or zero is visible, it tends to be at about the same frequency location for each speaker. An average (median) calculation across speakers of each sex yields clear poles at 1650 and 2350 Hz, as well as weak indications of poles at 750 and 3150 Hz for the females. These values are consistent with results published previously on the frequency locations of tracheal poles (Ishizaka *et al.*, 1976; Cranen and Boves, 1987). Values for our male speakers are slightly lower in frequency, as would be expected given the larger body size of males. The frequency locations of zeros associated with tracheal coupling are usually close

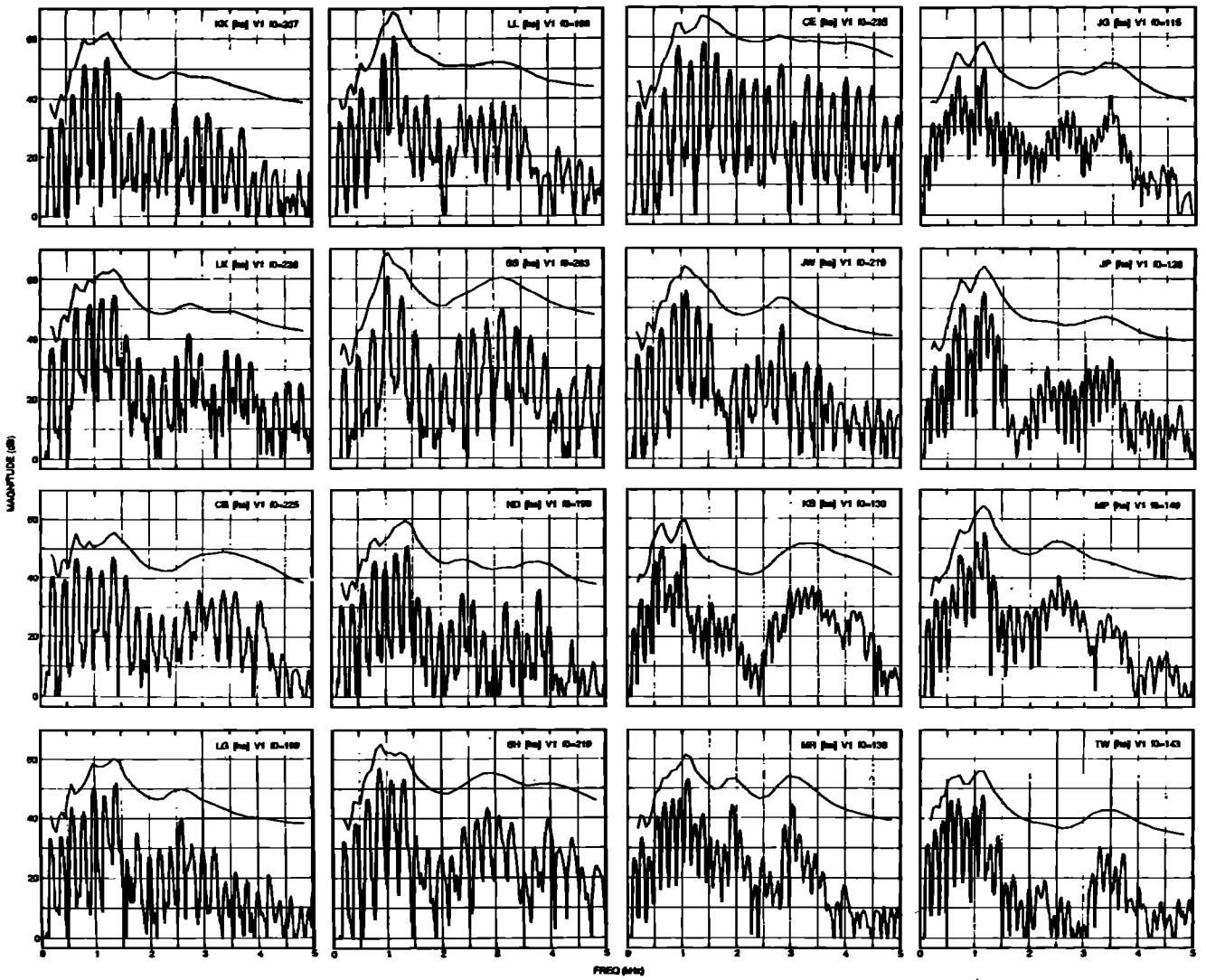


FIG. 7. Twenty-five-ms dft magnitude spectra of [ɑ] obtained from 16 speakers. Spectra were sampled at the midpoint of the vowel in the first syllable in [hɑ] reiterant imitations of the sentence "Steve eats candy cane." The smooth curve was produced by averaging dft energy over a critical band.

TABLE VII. Frequency locations of extra (tracheal) poles and zeros in aspiration spectra for ten female and six male talkers. Values in parentheses are of uncertain status. See text.

Females	Poles P1	P2	P3	P4	Zeros Z1	Z2	Z3	Z4
KK	(750)	1800	2650	(3150)	(900)	1550	2100	2900
LK	700	(1650)	(2500)		850	(1800)	2100	3200
CB		1700	(2600)		900	(1850)	2200	
LG	750	(1600)		3100	850	1800	(2300)	
SS		1650	2350		900	1950	2400	
LL	700	1650	2400	3250	850	(1800)	2200	3050
ND	750				(900)			
SH		1800	2600			1550	2500	3200
CE	800	(1650)	2300	2600	950	1750	2700	
JW		1700	2400		900		2200	3100
Median:	(750)	1650	2350	(3150)	900	1800	2200	(3100)
Males								
KS		1500	2000			1300	1750	2400
MR		1500		3400		1800		3200
JG		1400	1700	3250			2100	3050
JP		1650	2550	(3300)		1800	(2050)	(3000)
MP		(1600)	(2200)			(1800)	(2050)	
TW		1600	2800	(3200)		(1400)	1900	2400
Median:		1550	2200	3275		1800	2050	3000

to an observed extra pole; the average frequency locations for females are 900, 1800, 2200, and 3100 Hz.[13] Data for males are similar.

In summary, extra tracheal poles often distort vowel spectra to varying degrees. A pole at about 2100 Hz is frequently seen in the [ɑ] vowel for our ten female talkers. When present, tracheal resonances tend to be located at frequencies consistent with previous measurements of pole locations.

2. Bandwidth of F1

The bandwidth of the first formant of the vocal-tract transfer function determines several aspects of the acoustic output. It determines the width of the resonance peak, and it determines the relative strength or prominence of the first-formant peak. An extreme example of the effect of increased B1 on the spectrum of a vowel in our corpus is shown in Fig. 8. There is virtually no indication of the presence of $F1$ in the spectrum for the example of [ɑ] in the right panel. In an attempt to quantify the variation in prominence of the $F1$ spectral peak across speakers of our database, we have defined two measures. The first is the amplitude of the $F1$ peak relative to some reference amplitude. The second is a subjective estimate of how easy it is to see the location of $F1$ in the spectrum. Both measures were applied to the spectrum sampled at the midpoint of the first vowel in [hɑ] reiterant imitations of "Steve eats candy cane" (sentence S1). $A1_{re2}$ is an indirect estimate of first-formant bandwidth, as indicated by the amplitude in dB of the first formant, relative to amplitude of second formant, as measured at the beginning, middle, and end of the first vowel in [hɑ] reiterant imitations of sentence S1. $F1_{vis}$ is an indirect estimate of the first-formant bandwidth, as indicated by the visibility of a local $F1$ maximum in spectrum: 2 = obvious local spectral maximum, 1 = inflection point in smoothed spectrum, 0 = no evidence of $F1$ peak. $F1_{vis}$ is the sum of this measure at the beginning, middle, and end of the first vowel in [hɑ] reiterant imitations of S1.

Results of this analysis are presented in Table VIII. Due to natural variability in the levels of formants, it appears that $A1_{re2}$ is not a particularly useful measure of the distinctiveness of the $F1$ peak. In any case, there is no difference between the genders in this measure. It also appears that we might have been better off looking at a short unstressed vowel, because it is more likely to be influenced by glottal opening for adjacent consonants and thus have a less distinct first-formant peak.

TABLE VIII. Two measures of the prominence of the $F1$ peak in vowel spectra adjacent to [h]: the amplitude of the first formant relative to the amplitude of the second formant, and the subjective visibility of $F1$ as a distinct local spectral maximum.

Speaker	$A1_{re2}$	$F1_{vis}$
KK	−1.0	5
LK	−2.0	6
CB	−1.3	5
LG	−3.3	5
LL	−8.0	5
SS	−6.0	2
ND	−4.0	6
SH	3.7	6
CE	−2.3	6
JW	−0.3	3
Av	−2.5	5
KS	0.7	6
MR	−4.0	5
JG	−0.3	6
JP	−3.0	6
MP	−5.7	5
TW	−3.3	5
Av	−2.7	5.5

Hawkins and Stevens (1985) have shown that vowel nasalization can have a similar flattening effect on the spectrum of $F1$, due to increased losses in the nasal tract and due to the splitting up of $F1$ into a pole–zero–pole complex. Thus it is not clear what perceptual attributes to assign to the change in the spectrum associated with an increase in B1 (breathiness or nasality). We will return to this issue in the design of a perceptual experiment.

In summary, the partially open glottis of a breathy vowel can cause the first-formant bandwidth to increase, sometimes obliterating the spectral peak at $F1$ entirely. This effect and the appearance of extra tracheal pole-zero pairs can cause serious problems for formant trackers and for models of perception that presuppose a formantlike representation of speech sounds for men, women, and children.

E. Correlation analysis of acoustic and perceptual data

A listening test was prepared in which reiterant imitations of "Steve eats candy cane" for the 16 speakers were randomized and played to a panel of eight listeners who rated the breathiness of the speakers' vowels on a seven-point scale. The scale is defined at the top in Table IX. Judgments were obtained for reiterant imitations using both [ʔɑ] and

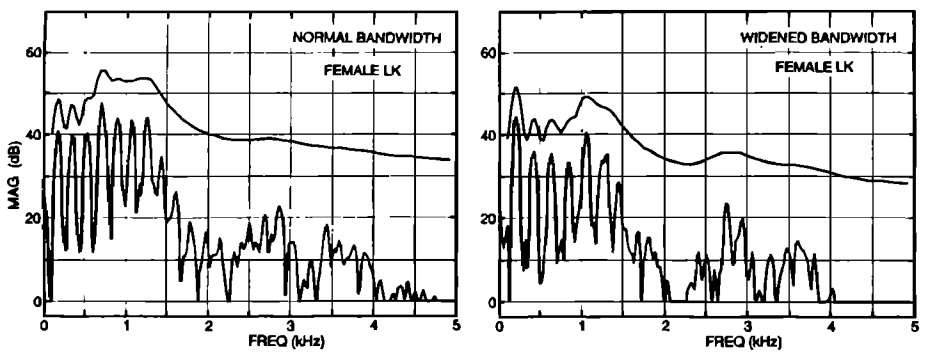


FIG. 8. Spectra of normal and breathy versions of the vowel [ɑ] are compared. The breathy token (right panel) provides an example in which the first formant bandwidth is increased to a point where it is difficult to see any evidence of a spectral peak in the expected location of $F1$ (about 750 Hz).

TABLE IX. Average ratings of perceived breathiness in vowel portions of the [ʔɑ] and [hɑ] reiterant imitation of "Steve eats candy cane," using the 7-point scale defined below. The rightmost column indicates average breathiness ratings for a vowel excised from the [hɑ] imitation.

Rating	Description		
1	not breathy		
2			
3	slight breathiness for some or all syllables		
4			
5	moderate breathiness for some or all syllables		
6			
7	strongly breathy		
Speaker	[ʔɑ] sentence	[hɑ] sentence	Excised vowel
KK	2.6	2.7	3.6
LK	5.2	5.0	4.6
CB	5.5	5.8	6.0
LG	3.8	5.1	5.0
LL	3.1	4.3	3.5
SS	2.1	3.7	2.8
ND	3.6	4.0	3.8
SH	1.8	2.7	2.9
CE	4.6	6.0	3.7
JW	4.6	5.4	3.2
Av	3.7	4.4	3.9
KS	2.7	3.5	2.7
MR	2.0	5.0	2.3
JG	3.0	4.7	3.0
JP	3.1	3.4	2.8
MP	2.2	2.0	2.3
TW	4.0	4.9	5.3
Av	2.8	3.9	3.1

[hɑ] syllables; average results based on four separate randomizations of each block of 16 trials are presented in the Table.

As general trends, it can be seen that females are judged, on average, to be slightly more breathy than males, and that sentences involving the [hɑ] syllable are perceived to be more breathy than sentences involving [ʔɑ] (even though instructions to the subjects were to rate the breathiness of the vowel portions of the utterances). Subjects judged to be most breathy include females CE, CB, JW, LG, and LK, and males TW, MR, and JG.

The perceptual judgment data may be contaminated by several factors not directly related to the breathiness of individual vowels. For example, the level of the aspiration noise could influence vowel judgments, as could the details of transitions between voicing and voicelessness. For this reason, we performed a second listening test in which the first vowel was excised from the [hɑ] reiterant imitation of "Steve eats candy cane." The first three glottal pulses and the last three pulses were deleted from the vowel and the remaining vowel had its onset and offset modified by a 10-ms half-Hanning window. A panel of three listeners produced four judgments of breathiness for each speaker. The averages are shown in the last column of Table IX.

The pattern of breathiness rankings for subjects changes a bit in this new test; CE, CW, and MR are not perceived to be as breathy as before. The correlation between the two sets of perceptual data in columns 2 and 3 of Table IX is only 0.55. Informal inspection of the acoustic data suggests that the level of the aspiration noise may well account for the perceptual difference.

Correlations between subjective breathiness ratings and a number of acoustic measures are presented in Table X. It is impossible to determine causation from such an analysis, but the few correlations reaching significance are easily interpreted in familiar terms. The acoustic measures are defined below:

PRC_1: group results of breathiness judgments for the first vowel excised from [hɑ] reiterant imitation of S1;

PRC_5: group results of breathiness judgments for [hɑ] reiterant imitation of S1;

$F0_{mid}$: fundamental frequency in Hz, as measured at the midpoint of the first vowel in [hɑ] reiterant imitation of S1;

$H1_{re2}$: amplitude of first harmonic, in dB, at first vowel midpoint in [hɑ] reiterant imitation of S1, relative to second-harmonic amplitude, offset by 10 dB to make positive;

$NOIS_1$: degree of breathiness noise visually present in $F3$ waveform, judged by DK in first vowel of [hɑ] reiterant imitation of S1;

$NOIS_5$: degree of breathiness noise visually present in $F3$ waveform, judged by DK, average of all five vowels of [hɑ] reiterant imitation of S1;

ASP_1: rms level, in dB, of aspiration noise during the [h] preceding the first vowel, relative to overall rms level of the first vowel, as measured at vowel midpoint, in the [hɑ] reiterant imitation of S1;

ASP_2: rms level, in dB, of aspiration noise during the [h] preceding the second vowel, relative to overall rms level of the second vowel, as measured at vowel midpoint, in the [hɑ] reiterant imitation of S1;

$A1_{re2}$: indirect estimate of first-formant bandwidth, as indicated by the amplitude in dB of the first formant, relative to amplitude of second formant, as measured at the beginning, middle, and end of the first vowel in [hɑ] reiterant imitation of S1, average, in dB;

$F1_{vis}$: another indirect estimate of the first-formant bandwidth, as indicated by the visibility of a local $F1$ maximum in spectrum: 2 = obvious local spectral maximum, 1 = inflection point in smoothed spectrum, 0 = no evidence of $F1$ peak; sum of measurements at beginning, middle, and end of first vowel in [hɑ] reiterant imitation of S1;

$A3_{re2}$: estimate of general spectral tilt above 1 kHz, as indicated by the amplitude of the third formant, in dB, relative to the amplitude of the second formant, average of measurements at beginning, middle, and end of the first vowel in [hɑ] reiterant imitation of S1;

$A5_{re2}$: another estimate of general spectral tilt above 1 kHz, as indicated by the amplitude of the third, fourth, or fifth formant (whichever is the greatest), in dB, relative to the amplitude of the second formant, average of measurements at beginning, middle, and end of the first vowel in [hɑ] reiterant imitation of S1.

Only two correlations with the perceptual judgments reach statistical significance: The amplitude of the first harmonic relative to H2 ($H1_{re2}$) is closely tied to subjective breathiness of the isolated first vowel (PRC_1), and the

TABLE X. Selected acoustic correlates of breathiness for ten female and six male talkers (top), and correlation coefficients between two subjective measures of perceived breathiness and these acoustic measurements (bottom).

SPKR	PRC_1	PRC_5	$F0_{mid}$	$H1_{re2}$	$NOIS_1$	$NOIS_5$	ASP_1	ASP_2	$10A1_{re2}$	$F1_{vis}$	$A3_{re2}$	$A5_{re2}$
KK	3.6	2.7	207	13.0	1	1.4	−14	−19	−10	5	−13	−13
LK	4.6	5.0	228	12.6	3	3.4	−23	−14	−20	6	−11	−11
CB	5.9	5.8	225	17.1	2	3.0	−15	−7	−13	5	−6	−6
LG	4.9	5.1	199	12.6	3	2.8	−8	−5	−33	5	−13	−13
LL	3.5	4.3	199	10.3	2	3.0	−23	−17	−80	5	−15	−15
SS	2.8	3.7	263	9.4	1	1.6	−14	−18	−60	2	−13	−13
ND	3.8	4.0	199	11.9	2	3.0	−8	−15	−40	6	−15	−14
SH	2.9	3.7	219	8.4	2	2.0	−19	−19	37	6	−8	−8
CE	3.7	6.0	235	12.5	3	3.2	−14	−12	−23	6	−9	−9
JW	3.2	5.4	219	11.8	3	3.4	−12	−10	−3	3	−14	−14
Av:	3.9						−15	−14	−25	5	−12	−12
KS	2.7	3.5	130	5.8	1	1.2	−7	−9	7	6	−14	−8
MR	2.3	5.0	138	5.3	1	1.8	−7	−8	−40	5	−8	−7
JG	3.0	4.7	115	4.6	1	2.2	−12	−9	−3	6	−9	−4
JP	2.8	3.3	128	4.9	1	1.2	−26	−23	−30	6	−21	−21
MP	2.3	2.0	149	6.0	1	1.0	−24	−20	−57	5	−17	−17
TW	5.3	4.9	143	9.7	1	2.4	−9	−13	−33	5	−20	−15
Av:	3.1						−14	−14	−27	5.5	−15	−12

	PRC_1	PRC_5	$F0_{mid}$	$H1_{re2}$	$NOIS_1$	$NOIS_5$	ASP_1	ASP_2	$A1_{re2}$	$F1_{vis}$	$A3_{re2}$	$A5_{re2}$
PRC_1	···	0.55	0.30	0.83	0.41	0.59	0.14	0.39	0.01	0.08	0.09	0.07
PRC_5		···	0.28	0.57	0.63	0.81	0.35	0.72	0.11	−0.01	0.47	0.45
$F0_{mid}$			···	0.50	0.60	0.50	−0.10	−0.06	−0.03	−0.43	0.36	0.03
$H1_{re2}$				···	0.54	0.66	0.24	0.45	0.08	−0.03	0.35	0.23
$NOIS_1$					···	0.84	−0.03	0.34	0.12	0.03	0.30	0.06
$NOIS_5$						···	0.11	0.45	−0.01	0.00	0.31	0.18
ASP_1							···	0.70	0.19	−0.12	0.22	0.44
ASP_2								···	0.19	0.00	0.49	0.63
$A1_{re2}$									···	0.33	0.40	0.50
$F1_{vis}$										···	0.07	0.20
$A3_{re2}$											···	0.88
$A5_{re2}$												···

amount of noise replacing harmonics in the $F3$ region of the spectrum for the five vowels of the [hɑ] reiterant imitation of S1 ($NOIS_5$) is closely tied with subjective breathiness of the entire [hɑ] reiterant imitation of S1 (PRC_5).

Low correlations with f_0, in spite of a known tendency for females to be heard as more breathy than males, must mean that, within each gender, f_0 is a poor predictor. Low correlations with $NOIS_1$, the estimate of noise in $F3$ for the first vowel of the sentence, may be due, in part, to the highly quantized nature of the data; for example, all males were assigned a value of 1 (periodic with no evidence of noise). Stronger correlations with $NOIS_5$, the average over five vowels of the same measure, support this conjecture. Low correlations with ASP_1, the level of aspiration noise in the [h] preceding the first vowel, appear to indicate that this level is highly variable and not at all predictive of the noise level at the middle of the first vowel ($NOIS_1$). The correlation of $NOIS_1$ with ASP_2, the level of the aspiration noise in the [h] preceding the second vowel, is somewhat higher, reinforcing a general observation that the first consonant of an utterance varies more in level than do utterance-internal consonants.

The correlations suggest that both (1) the relative amplitude of the first harmonic and (2) the presence of noise during voicing affect judgments of breathiness, but we cannot conclude that these are the only factors involved. Instead, we will use a speech synthesis experiment to establish the perceptual importance of manipulations to a large set of variables.

We turn next to a description of a new speech synthesizer that was developed on the basis of recent research on source mechanisms in speech production. Perceptual experiments using this synthesizer will then be described.

II. SOURCE MODELS FOR SPEECH SYNTHESIS

As indicated in Fig. 9, the effective sound source during vowel production is the glottal volume velocity waveform, $U_g(t)$. This waveform acts as the input to the vocal-tract transfer function, which introduces resonant structure to the output lip volume velocity. Sound pressure measured some distance from the lips is then proportional to the temporal derivative of lip volume velocity (Fant, 1960).

Recent efforts to characterize the essential features of the voicing source waveform $U_g(t)$ for different male and

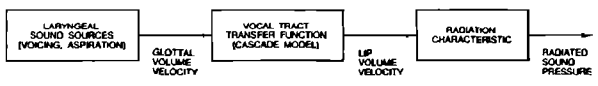


FIG. 9. Block diagram illustrating the acoustic theory of speech production.

female voices have led to several new parametric models of glottal output (Ananthapadmanabha, 1984; Fant, 1979, 1982b; Titze, 1984; Fant *et al.*, 1985; Allen and Strong, 1985; Fujisaki and Ljungqvist, 1986; Klatt, 1987b; Rosenberg, 1971, 1975).

In an updated version of a laboratory formant synthesizer (which we have called KLSYN88, to distinguish it from the older version KLSYN), two different new models of the glottal source have been incorporated. One of these is a slightly modified version of the Liljencrants–Fant (LF) model (Fant *et al.*, 1985). The other model, which has some characteristics in common with the LF model but incorporates some additional features, is called KLGLOTT88.

In the models, the characteristics of the waveform are described by conventional parameters such as **F0**, the fundamental frequency of voicing, and **AV**, the peak amplitude of the glottal pulse, as well as new parameters: (1) **OQ**, the open quotient—or ratio of open time to total period duration and (2) **TL**, spectral tilt—or the additional spectral change associated with "corner rounding" in which closure is nonsimultaneous along the length of the vocal folds.

A. The modified LF model

The LF model was chosen over other candidates because both Fant *et al.* (1985) and Fujisaki and Ljungqvist (1986) have shown it to be superior to other models of the same complexity when the objective is to model natural speech with minimum rms error. The model was originally formulated in terms of a set of times of waveform events, but it can easily be recast in terms of familiar parameters **AV** (amplitude of voicing), **F0** (fundamental frequency), **OQ** (open quotient), **SQ** (speed quotient), and **TL** (spectral tilt).[14]

The LF model does not consider the possible importance of turbulence noise generation at the glottis due to a constant flow leakage between partially spread arytenoid cartilages. Data from Holmberg *et al.* (1988) indicate that dc flows are very common for male and female speakers of English in the environment of aspirated stops. If the dc flow component causes the generation of simultaneous aspiration noise at the glottis, as is very likely, then the modeling of $U_g(t)$ should include provisions for introduction of aspiration noise (Pandit, 1957; Fujimura, 1968; Dolansky and Tjernlund, 1968; Rothenberg, 1974; Rothenberg *et al.*, 1975; Ladefoged and Antoñanzas-Barroso, 1985; Klatt, 1986b, 1987b; Hunt, 1987). Voicing source models have been devised for a formant synthesizer that are intended to increase the naturalness of the output speech by permitting a mixture of an impulse train and noise as the source waveform (Kato *et al.*, 1967; Holmes, 1973). The strategy is to specify a cutoff frequency below which the source consists of harmonics, and above which the source is flat-spectrum noise. Similar strategies for mixed-excitation synthesis have been described by Rothenberg *et al.* (1975) and Makhoul *et al.* (1978). The KLGLOTT88 voicing source model, to be described next, includes a parameter, **AH**, which controls the amplitude of aspiration noise that can be added to the $U_g(t)$ waveform. The aspiration noise has a spectrum that falls off at about 6 dB/oct of frequency increase, but, when the radiation characteristic is folded into the source models, the result is a relatively flat spectrum. Thus, as noise is added to the voicing source, it is most noticeable at high frequencies where the harmonic spectrum of voicing is weaker. If the glottis is partially spread, as in a breathy vowel, **TL** and **AH** will be increased, and higher harmonics of the source spectrum will be replaced by aspiration noise.

B. The KLGLOTT88 voicing source model for KLSYN88

A block diagram of the new voicing source model for the synthesizer originally described in Klatt (1980) is presented in Fig. 10. In order to distinguish it from its predecessor, the new model will be called the KLGLOTT88 model. Source control parameters identified in the figure include: **AV**, amplitude of voicing, in dB; **F0**, voicing fundamental frequency, in tenths of a Hz; **OQ**, open quotient of the glottal waveform, in percent of a full period; **TL**, tilt of the voicing source spectrum, in dB down at 3 kHz; **FL**, period-to-period flutter (quasirandom fluctuations) in f_0, in percent of maximum; **DI**, degree of diplophonic double-pulsing irregularity in f_0, in percent of maximum; and **AH**, amplitude of aspiration (breathiness) noise, in dB.

During the open phase of a glottal cycle, the volume velocity waveform has been parametrized to obey a relationship first proposed by Rosenberg (1971); i.e., $U_g(t) = at^2 - bt^3$, where a and b are constants whose values depend on the amplitude of voicing and the duration of the open period. The spectral consequences of varying open quotient, spectral tilt, and aspiration level are shown in Fig. 11. The aspiration source provides noise energy with a relatively flat spectrum, as indicated in the last panel of the figure. The behavior of the remaining control parameters of the KLGLOTT88 voicing source will be explicated in the next few sections.

Current acoustic models of the $U_g(t)$ waveform are rather simple, capturing only the first-order shapes and spectral aspects of observed natural glottal waveforms. It is hoped that, for speech synthesis purposes, the models will turn out to be useful. However, the following qualifications suggest that several additional modeling complications may be necessary to achieve high-quality synthesis of male and female voices.

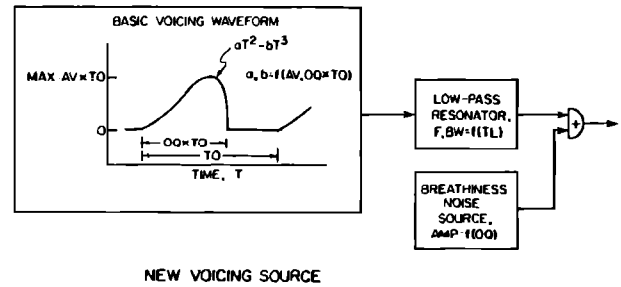


FIG. 10. Block diagram of the voicing source for the KLSYN88 formant synthesizer. The effects of the radiation characteristic have also been folded into the source models, resulting in a voicing source spectral output (Fig. 11) that falls off at about 6 dB/oct [corresponding to $U_g'(t)$] and an aspiration source spectrum that is essentially flat over the frequency range of interest.

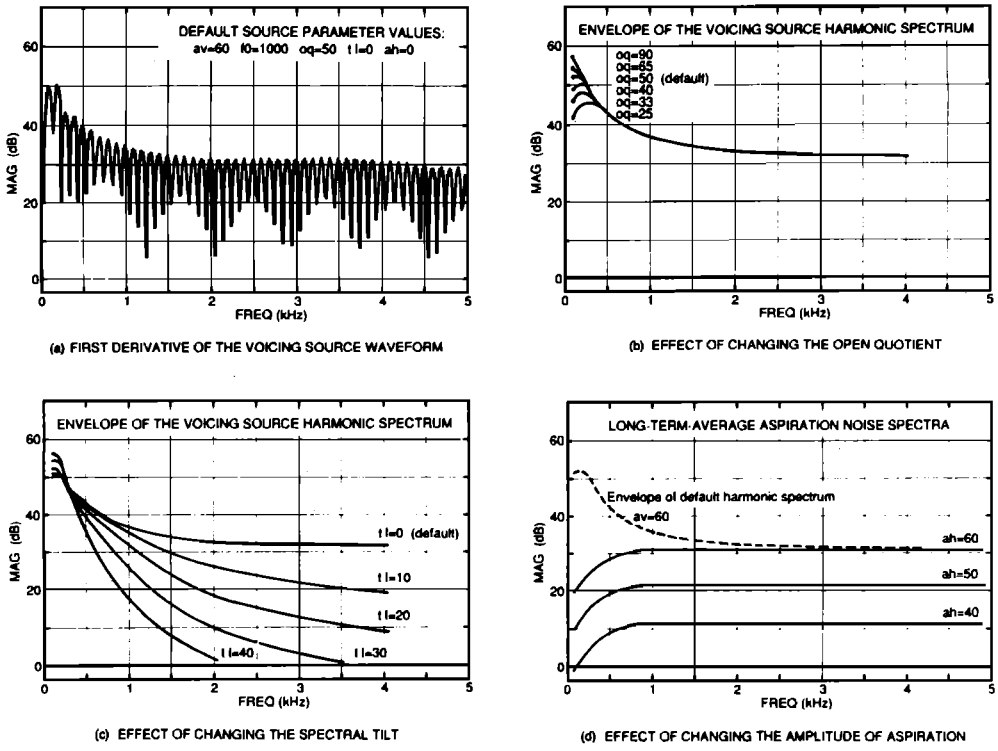


FIG. 11. Dft magnitude spectra are shown of $U_g'(t)(U_g(t)$ modified by a first difference approximation to the radiation characteristic), as synthesized by the KLGLOTT88 voicing source model at several values for each of three control parameters. (a) Spectrum of a train of pulses with $f_0 = 100$ Hz; (b) and (c) spectral envelope of a harmonic spectrum that would result from synthesizing a train of such pulses; (d) spectra of aspiration noise source.

1. Complications I: Glottal pulse timing irregularities

The waveshape of successive periods of $U_g(t)$ in a sustained vowel need not be identical. The literature includes terms such as "jitter," the period-to-period random fluctuations in period durations (Horii, 1979), "shimmer," the period-to-period random fluctuations in glottal-pulse amplitude (Horii, 1980), and "diplophonic double pulsing," the tendency for a voice to sometimes vibrate in a mode where pairs of glottal pulses move toward one another, with the first often being attenuated in amplitude (Timke *et al.*, 1959). We consider each of these deviations from perfect periodicity in turn. The discussion then shifts to acoustic interactions between the source and vocal tract that may contribute to naturalness.

a. Jitter and shimmer. It is well known that a constant f_0 is to be avoided in speech synthesis because the result is a peculiarly mechanical sound quality. An example of an analysis of fundamental frequency of a female subject attempting to hold a constant pitch is shown in Fig. 12. While the wavering nature of the f_0 trace may in small part be due to analysis artifacts,[15] it is known that normal physiological mechanisms can impart these kinds of fluctuations. In an insightful correlational analysis of f_0 and EMG data, Baer (1978) was able to show that a single muscle fiber twitch in the cricothyroid causes a predictable not-insignificant local increase in f_0, and that normal statistical variations in fiber firing can be expected to produce fluctuations in f_0 not unlike those observed in the figure.

The mechanical quality of synthesis at constant f_0 can be reduced or eliminated simply by introducing a normal intonation contour to the synthesis (Rosenberg, 1968), but there are often time intervals where the f_0 is nearly constant, and some sort of simulation of the f_0 flutter or jitter seen in Fig. 12 would be desirable. Jitter, defined as the period-to-period variability in f_0, has been measured in sustained vowels for both normal and pathological voices (Lieberman, 1961, 1963; Horii, 1979, 1980; Hollien *et al.*, 1973; Askenfelt and Hammarberg, 1981, 1986). If the appropriate parameter to characterize jitter and shimmer is the standard deviations of a presumed Gaussian distribution of periods or pulse amplitudes, respectively, then normal voices sustaining the vowel [a] contain a jitter of about 0.5% to 1.0% (Hollien *et al.*, 1973). This is slightly less than the detectability threshold—perceptual data indicate a detectability threshold for jitter of about 2% and for shimmer of about 10% or 1 dB (Pollack, 1971)—calling into question the utility of adding this kind of Gaussian jitter to synthesis. It is also likely that the jitter and especially shimmer measured by these techniques is, in part, a measurement artifact due to superposition effects (Milenkovic, 1987).

The nature of a better random component for the synthesis of jitter has been a subject of debate, since most efforts to introduce audible random jitter to the pitch period in synthesis have led to a harsh voice quality (Rozsypal and Millar, 1979). The KLGLOTT88 voicing source model includes a mechanism for introducing a slow quasirandom drift to the f_0 contour through the FL flutter control parameter (shimmer is not modeled). The term "flutter" has been adopted since jitter has a well-defined meaning that differs from our synthesis strategy. Instead of using a random process to simulate jitter, we add to the nominal f_0 a quasirandom component that is, in fact, the sum of three slowly varying sine waves:

$$\Delta f_0 = (\mathrm{FL}/50)(F0/100)[\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)]\ \mathrm{Hz}. \quad (1)$$

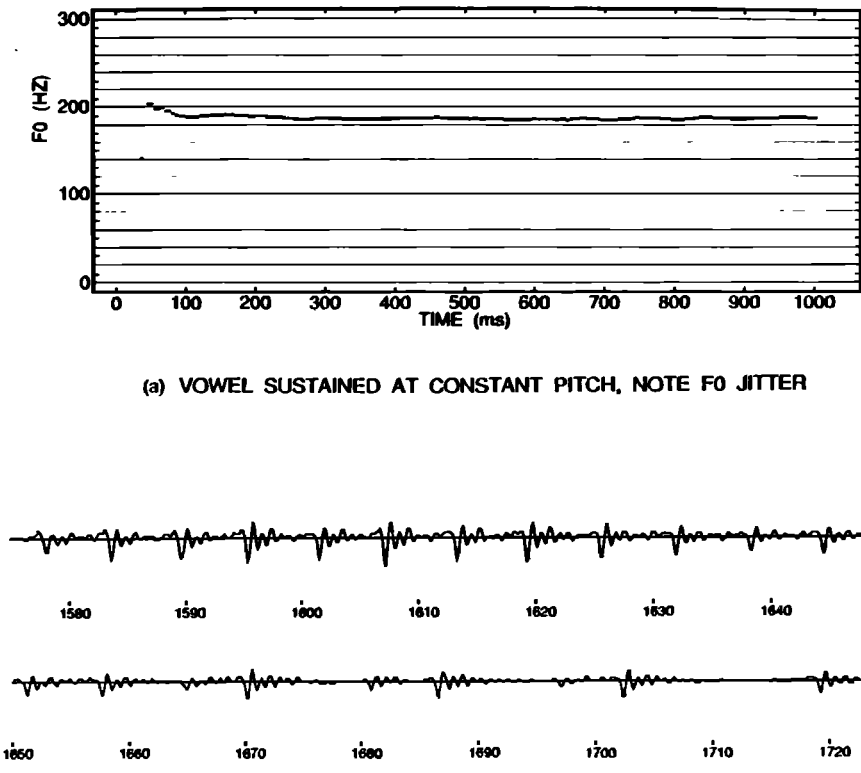


FIG. 12. Examples of deviations from perfect periodicity: (a) fundamental frequency contour of a female subject sustaining a vowel at constant pitch (note the waver, or inability to hold pitch constant) and (b) a speech waveform in which various degrees of diplophonic double pulsing are present [normal voicing ($t = 1580$ to 1660) suddenly changes to a vibration mode where the first of a pair of periods is delayed and reduced in amplitude ($t = 1660$ to 1710) and the first pulse may disappear entirely ($t = 1710$ to 1720)]. (Diplophonic example extracted from the final syllable of female speaker LK [hɑ] imitation of "Steve eats candy cane.")

Sine-wave frequencies of 12.7, 7.1, and 4.7 Hz were chosen so as to ensure a long period before repetition of the perturbation that is introduced. A value of FL = 25% results in synthetic vowels with a quite realistic deviation from constant pitch. It is unlikely that this slowly varying flutter component is the only deviation from constant pitch in normal voicing, but it appears to be sufficient for synthesis purposes.

b. Diplophonic double pulsing. An example of diplophonic double pulsing is shown in Fig. 12(b). In the extreme, the alternate pulses may actually disappear, in which case f_0 is halved. Obvious examples of double pulsing were observed sporadically, usually near the termination of an utterance, for more than a quarter of the speakers that we have examined. Less extreme diplophonia may occur more often. The KLGLOTT88 voicing source model includes a mechanism for simulating double pulsing using the **DI** (diplophonic double pulsing) control parameter. Alternate pulses are modified whenever **DI** is greater than zero. A modified pulse is delayed in time and attenuated in amplitude by an amount that is specified in terms of the maximum allowed in percent, where the maximum delay is such as to time the closure of the first pulse to be simultaneous with the opening of the next unaltered pulse, and the amplitude attenuation goes from one to zero on a linear scale as **DI** ranges from 0%–100%. For example, if **OQ** is at 50%, setting **DI** to a value of 50% results in a first pulse of each pair that is delayed by a quarter of a period and is attenuated by half (-6 dB).

2. Complications II: Source–tract interactions

According to the original classical formulation of the acoustic theory of speech production (Fant, 1960; Flanagan, 1972), the voicing source can be characterized as a "current source" because the volume velocity waveform $U_g(t)$ was said to depend very little on the shape or impedance of the vocal tract, at least for vowels. Similarly, the vocal-tract transfer function was assumed to be modeled well by a (succession of) time-invariant linear filter(s) because the terminating impedance at the glottis, while varying over a period, is nonetheless high compared with the vocal-tract impedance. These assumptions are illustrated in Fig. 13(a).

Recent work by Fant and his associates suggests that some of the original simplifying assumptions of the classical theory are not really valid. First of all, the presumed direct relationship between glottal area and glottal flow is perturbed by standing wave-pressure fluctuations in the pharynx, which invalidate an assumed constant transglottal pressure over a cycle. The pharyngeal pressure variations cause the glottal source flow waveform to take on ripple components at the frequency of $F1$, and may even be large enough to have a direct influence on the mechanical behavior of the vocal folds (Fant, 1985). Furthermore, as the glottis opens and closes, the vocal-tract transfer function undergoes rapid changes over a single period that may be of perceptual importance. The essential characteristics of an "interactive source-filter model" that takes into account these complications are shown in Fig. 13(b). Four phenomena not satisfac-

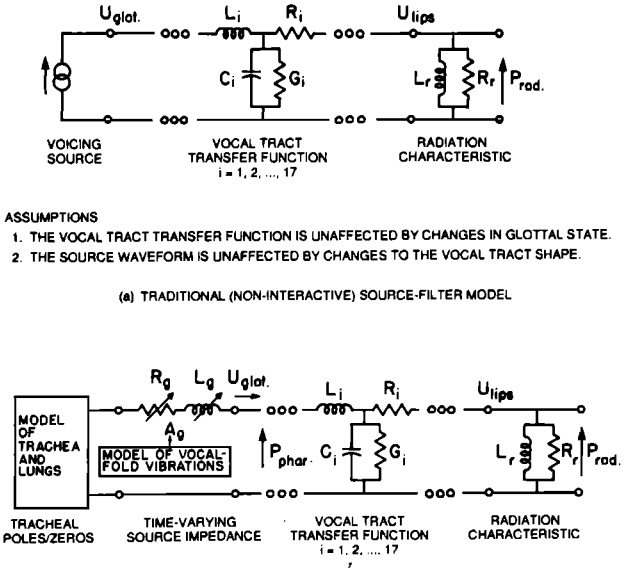


FIG. 13. The simplifications implicit in the classical acoustic theory of speech production, which constitute a noninteractive source-filter model shown in the top part of the figure, are contrasted with an interactive model in which voicing source glottal volume velocity $U_g(t)$ is influenced by pressure fluctuations above the glottis, and the vocal-tract transfer function changes over a period due to the time-varying glottal impedance.

torily modeled by conventional formant synthesizers can be identified from examination of the interactive source-filter model: The first two affect the source waveform $U_g(t)$, and the second pair of phenomena affect the vocal-tract transfer function.

a. F1 ripple in the source waveform. The transglottal pressure is an important variable in determining glottal volume velocity from the time variation in glottal area. However, transglottal pressure is not constant over a period as was originally assumed, but rather varies due to pressure fluctuations associated with the $F1$ standing wave in the lower pharyngeal portion of the vocal tract (Fant, 1982b; Fant *et al.*, 1985). The interaction is nonlinear in that volume velocity through an orifice is proportional to square root of the pressure drop (Stevens, 1971). Assuming a constant glottal area function from period to period, the noninteractive model predicts a succession of smooth identical volume velocity pulses, whereas the interactive model predicts a buildup of "$F1$ ripple" interaction as a standing wave is developed in the vocal tract. The resulting $U_g(t)$ waveform includes significant "ripples" associated with the $F1$ standing wave, and the source waveform actually changes over the first two to three periods of voicing onset (Fant and Lin, 1987).

These effects, which Fant calls nonlinear superposition effects, result in an overall boost in the spectral amplitude of $F1$ relative to other formants because an $F1$ component is contained in the source waveform. They are of unknown perceptual importance, but presumably could be crudely approximated by first-formant bandwidth changes and perhaps an increase in the source spectral tilt using the KLGLOTT88 voicing source model. As is the case with most synthesis parameters, it is the change in parameter over time rather than its static value that has the greatest perceptual importance for improving naturalness. Thus, to mimic the buildup of an $F1$ ripple over the first few periods at voicing onset, a simultaneous decrease of **B1**, and increase in spectral tilt **TL**, might be performed. Such a change in relative formant amplitudes has been observed for laryngealized voicing onsets (Klatt, 1986b), but, in a transition between a voiceless consonant and a following vowel, the observed change is typically in the opposite direction, calling into question the generality and/or perceptual importance of this effect.

b. Nonlinear F1–f_0 interaction. The pharyngeal pressure standing waves may actually influence the mechanical behavior of the vocal folds. One nonlinear effect that could be associated with acoustical-to-mechanical coupling is an increase in glottal source strength whenever $F1$ is near an integral multiple of f_0 (Fant and Mártony, 1963; Fant and Ananthapadmanabha, 1982). Perhaps the pressure changes associated with the $F1$ standing wave induce a stronger closure if the phase is favorable (Rothenberg, 1985), and this occurs whenever $F1 = n \times f_0$. It should be possible to simulate the essential characteristics of this type of interaction by causing **AV** to increase whenever a harmonic is close to the frequency of $F1$. However, informal attempts to replicate this phenomenon with two speakers have been unsuccessful. Perhaps the interaction occurs only under some glottal conditions and not others.

In addition to the $F1$ standing wave induced in the vocal tract, it is possible that the increased impedance of a constricted vocal tract could influence source characteristics. The effect of a vocal-tract constriction on the vibratory behavior of the larynx has been studied by Bickley and Stevens (1986). Using both spectral analysis of natural speech produced under conditions of various suddenly applied oral constrictions and a subsequent modeling simulation, the authors found little change in source spectral characteristics until the constriction size became comparable to or smaller than that of a typical fricative. In this case, the glottal open time increased slightly, as did the bandwidth of the first formant, both being evidence of an increased average opening at the glottis. The authors conclude that, during the production of vowels and sonorant consonants, the effect of changes to oral constriction size on the mode of vibration of the larynx is probably negligible.

c. Truncation of the F1 damped sinusoid. The time-varying glottal impedance affects the vocal-tract transfer function primarily by causing losses at low frequencies to increase when the glottis is open. The first-formant bandwidth may increase substantially, leading to a truncation of the damped sinusoid corresponding to $F1$ during the open portion of the period (Fant and Ananthapadmanabha, 1982). Effects of time-varying formant bandwidths can be approximated in a formant synthesizer either by employing a perceptually equivalent constant bandwidth, or by varying bandwidth over a period. Perceptual data indicate that it is

difficult but not impossible to hear the difference between a time-varying first-formant bandwidth and an appropriately chosen constant bandwidth (Nord *et al.*, 1986). Some time variation in formant frequencies may also be desirable; $F1$ has been observed to increase by as much as 10% during the open phase of a glottal cycle. A method for changing first-formant bandwidth and first-formant frequency pitch-synchronously is included in KLSYN88.

The variables **DF1**, "delta frequency of $F1$," the incremental increase in first-formant frequency during the open portion of each period, and **DB1**, "delta bandwidth of $F1$," the incremental increase in first-formant bandwidth during the open portion of each period, have been created in order to allow pitch-synchronous changes to **F1** and **B1** in KLGLOTT88. The change to first-formant frequency and bandwidth occurs in "square-wave" fashion, increasing at the instant of glottal opening, and decreasing at the instant of glottal closure, as determined by the open quotient. For example, to have **F1** = 500 Hz during the closed phase and 550 Hz during the open phase of each period, one would set **F1** = 500 and **DF1** = 50. In a low vowel, the time variation in first-formant bandwidth might be approximated by setting **B1** = 50 and **DB1** = 400. A perceptually nearly equivalent constant first-formant bandwidth (equal spectral level of $F1$) corresponds to a first-formant bandwidth setting of about 90 Hz. The default values for the **DF1** and **DB1** incremental parameters are set to zero because most users will not need to resort to this kind of detail during synthesis.

d. Tracheal poles and zeros. Tracheal resonances may show up as additional pole-zero pairs in the vocal-tract transfer function, especially for breathy phonation where the glottis is presumably open over its posterior portion throughout the glottal cycle (Fant *et al.*, 1972; Klatt, 1986b). An example has been shown in Fig. 5. Effects of tracheal coupling can be modeled in a formant synthesizer by adding one or more paired pole-zero resonators to the vocal-tract transfer function (Fant *et al.*, 1972; Ishizaka *et al.*, 1976; Cranen and Boves, 1987). Berg (1960) originally observed a lowest tracheal resonance of 300 Hz. Ishizaka *et al.* (1976) found a much higher value of 640 Hz. Cranen and Boves (1987) measured values of the lowest three tracheal resonances of 510, 1350, and 2290 Hz from one male speaker. Fant *et al.* (1972) concluded from a modeling study that the latter estimates were more reasonable. The recent modeling work of Ananthapadmanabha and Fant (1982) and Rothenberg (1985) indicates that the actual effect on the vocal-tract transfer function of tracheal resonances is a complex function of the glottal configuration over time. A tracheal pole-zero pair has been added to the cascade model of the vocal-tract transfer function of the new KLSYN88 synthesizer in order to improve the synthesis of breathy vowels.

The variable **FTP**, "frequency of the tracheal pole," in consort with the variable **FTZ**, "frequency of the tracheal zero," can mimic the primary spectral effects of tracheal coupling in breathy vowels. A cascaded pole-zero pair is provided in the cascade branch of the synthesizer to mimic the addition of a "spurious" resonant peak due to this tracheal coupling interaction. Tracheal resonances are often seen in breathy vowels at frequencies of about 550, 1300, and/or 2100 Hz (slightly higher for female voices). The best synthesis strategy is to pick the most prominent one for synthesis (or use the nasal pole-zero pair to simulate a second[16]).

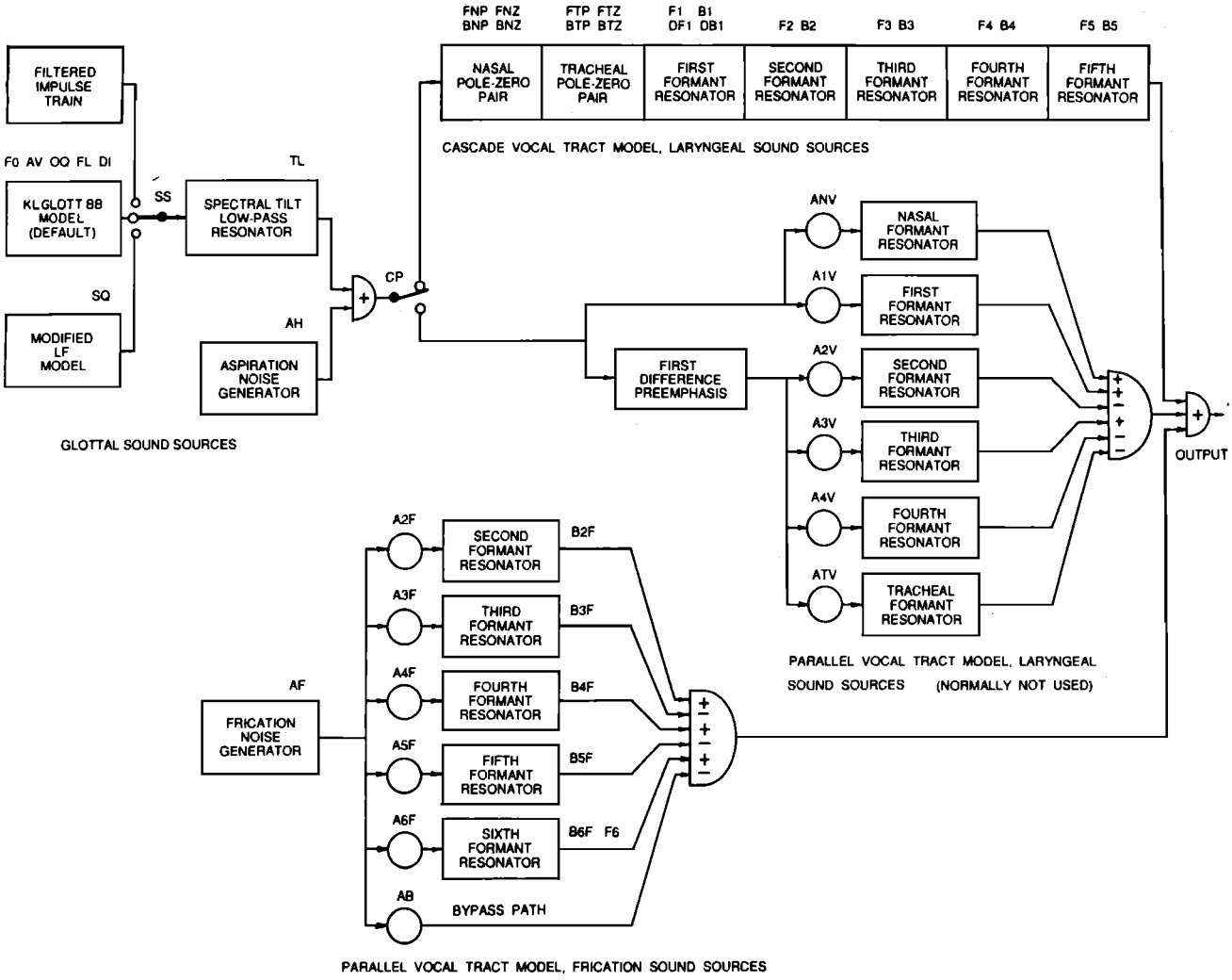
Normally, the spectral dip or zero corresponding to the selected tracheal resonance is immediately below it in frequency. Tracheal coupling usually begins and ends gradually as the glottis is opened or closed, which suggests a synthesis strategy in which both the tracheal pole and zero are usually moved together to the frequency location of an observed tracheal pole, and then the frequency of the tracheal zero **FTZ** is gradually moved down over perhaps 50 ms prior to glottal abduction to an appropriate value, as revealed by spectral analysis of the breathy interval.

The variables **BTP**, "bandwidth of the tracheal pole," and **BTZ**, "bandwidth of the tracheal zero," have default values of 180 Hz. It is difficult to determine appropriate synthesis bandwidths for individual tracheal resonances, but, fortunately, one can achieve good synthesis results without changing these default values in most cases. If the location of a tracheal zero is not clear from analysis of a breathy vowel, one possible synthesis strategy is to leave the frequencies of the tracheal pole and zero overlapped and simply increase the bandwidth of the zero (and/or decrease the bandwidth of the pole) in order to reveal the presence of the tracheal pole as a resonance peak in the synthesis. Each doubling of zero bandwidth will increase the strength of the tracheal resonance by about 6 dB.

C. Summary

In summary, the old cascade/parallel formant synthesizer, KLSYN (Klatt, 1980), has been modified to incorporate: (1) a version of the Liljencrants–Fant (LF) model of the glottal source; (2) a new voicing source model having flexible control of open quotient, spectral tilt, aspiration noise of breathiness, flutter to the timing of individual glottal pulses, and diplophonic double pulsing; (3) an extra pole-zero pair for simulating the introduction of a tracheal resonance in the vocal-tract transfer function; and (4) an ability to change the first-formant bandwidth pitch-synchronously to simulate one of the interactions between source and vocal tract identified by Fant.

The resulting KLSYN88 synthesizer, summarized in Fig. 14, consists of circuits to generate voicing, aspiration and/or frication, and circuits to approximate the sound source filtering performed by the vocal tract. The radiation characteristic has been folded into the sound sources for computational efficiency. There is a cascade formant model of the vocal-tract transfer function for laryngeal sound sources, and a parallel formant model with formant amplitude controls for frication excitation. A third vocal-tract model in which the vocal-tract transfer function for laryngeal sound sources is approximated by formants configured in parallel is useful for some specialized synthesis applications, but is normally not used. As was the case in the original formant synthesizer, the aspiration and frication noise sources are amplitude modulated, to simulate the effect of vocal-fold vibration, if **AV** is nonzero.



THE KLSYN88 CASCADE / PARALLEL FORMANT SYNTHESIZER

FIG. 14. Block diagram of the KLSYN88 formant synthesizer. Three voicing source models are available: (1) the old KLSYN impulsive source, (2) the KLGLOTT88 model (the default), and (3) the modified LF model. Also added are a tracheal pole-zero pair and control parameters allowing the first-formant frequency and bandwidth to vary over a fundamental period.

Control parameters are identified above each block in Fig. 14. Some control parameter names have been changed slightly from Klatt (1980) in order to accommodate the new components and to be more mnemonic. There are, in addition, several constants that the user can modify; a complete list of synthesizer control parameters is identified in Tables XI and XII. New constant control parameters **RS** and **SB** permit the selection of a particular noise sample with a maximally flat spectrum, and duplication of that spectrum at every noise onset if desired. The parameters **GV**, **GH**, and **GF** are used to set scale factors for the individual sources, as indicated.

In the synthesis and perception experiments to be described below, some sentence-length reiterant utterances were generated using the KLSYN88 synthesizer. Various parameters of the synthesizer were manipulated to produce the stimuli, and the KLGLOTT88 source model was used in all of these experiments.

III. SYNTHESIS OF REITERANT UTTERANCES

One way to determine whether the acoustic correlates of breathiness identified in this study are perceptually important and sufficient cues to signal various voice qualities is to define a speech synthesizer that can manipulate these acoustic variables, and then attempt to mimic in detail some of the voices observed. Reiterant speech provides a significant advantage in these circumstances because one does not have to spend an inordinate amount of time attempting to deduce variations in the vocal-tract transfer function over each sentence to be synthesized. A tape recording played at the fall 1987 Meeting of the Acoustical Society of America demonstrated our success in this endeavor in that the synthesized and natural versions of [ʔɑ] and [hɑ] reiterant sentences from several female speakers are virtually indistinguishable.

In this section, strategies are presented for selecting optimal values over time for each glottal source synthesizer

TABLE XI. Constant control parameters for the KLSYN88 synthesizer configuration. Each control parameter is assigned a two-letter name, an indication of whether it is a constant or can be made to vary over time, a minimum value, a default value that applies if the user makes no changes, a maximum value, and an English description of its effect on the synthesis.

SYM	V/C	MIN	VAL	MAX	Description
DU	C	30	500	5000	duration of the utterance, in ms
UI	C	1	5	20	update interval for parameter reset, in ms
SR	C	5000	10000	20000	output sampling rate, in samples/s
NF	C	1	5	6	number of formants in cascade branch
SS	C	1	2	3	source switch (1 = impulse, 2 = natural, 3 = LF model)
RS	C	1	8	8191	random seed (initial value of random number generator)
SB	C	0	1	1	same noise burst, reset RS if AF = 0 and AH = 0 (0 = no, 1 = yes)
CP	C	0	0	1	0 implies Cascade, 1 implies Parallel tract excitation by AV
OS	C	0	0	20	output selector (0 = normal, 1 = voicing source, ...)
GV	C	0	60	80	overall gain scale factor for AV, in dB
GH	C	0	60	80	overall gain scale factor for AH, in dB
GF	C	0	60	80	overall gain scale factor for AF, in dB

control parameter on the basis of spectral comparisons between synthesis and a natural recording. Finally, generalizations are made concerning typical synthesis parameter values for breathy and glottalized onsets and offsets.

A. Copying reiterant utterances

Application of the procedures described in this section to two female utterances resulted in the control parameter values that are shown in Figs. 15 and 16. The first utterance is a [ʔɑ] reiterant imitation of the five-syllable sentence "Steve eats candy cane" spoken by female LK. The second is a [hɑ] reiterant imitation of the same sentence by LK. The new voicing source has seven control parameters that must be specified:

F0 fundamental frequency,

AV amplitude of voicing,

OQ open quotient (ratio of open period to total period),

TL extra spectral tilt of the source (dB down at 3 kHz),

AH amplitude of turbulent aspiration noise added to voicing,

FL flutter (slowly varying statistical fluctuations to the fundamental period),

DI double pulsing (temporal offset and reduced amplitude of alternate periods).

1. Step 1: Set f_0 contour

The fundamental frequency parameter **F0** is used to reset the fundamental period $T0$ at the beginning of each pitch period.[17] A harmonic sieve-pitch-tracking algorithm (Duifhuis *et al.*, 1982), employing a 25-ms Hamming window, was used to determine f_0 every 10 ms in the original recording. The analysis data were transferred to the **F0** synthesis parameter track as the very first step in synthesis specification. Some trial and error adjustment based on comparison of synthesis and natural waveforms was necessary to match the irregular periods of glottalized attacks and offsets. A good match to the f_0 contour facilitates spectral comparisons needed to optimize other synthesis parameters since natural and synthesis spectra then have the same harmonic locations.

2. Step 2: Set AV, the amplitude of voicing

The amplitude of voicing **AV**, in dB, controls the height of each glottal pulse. Initial values for **AV** were determined from a plot of overall rms energy in the waveform, as estimated every 10 ms[18] using a 25-ms Hamming window, combined with a knowledge of when voicing was present in the waveform. Incremental adjustments were then made to the **AV** parameter track on the basis of trial and error comparison of synthesis and natural spectral levels until there was a good match (within about 1 to 2 dB) in spectra sampled about every 30 ms. This adjustment was done several times, the last being after all other synthesis parameters had been optimized.

3. Step 3: Set formant frequencies and bandwidths

Formant-frequency locations were estimated from spectra sampled at regular 30-ms time intervals throughout the vocalic portions of the utterance. An attempt was made to find a single constant value for each formant that would be a satisfactory approximation to the (possibly time varying) formant position. It is likely that slightly better synthesis matches could have been achieved by varying formant frequencies over time, but we wished to concentrate efforts on the glottal source parameter behavior, and we wanted to be sure that we were not covering up possible deficiencies in source flexibility by substituting unrealistic formant-frequency changes over time. After specifying formant frequencies, formant bandwidth values were then adjusted so as to give the appropriate level and shape to each spectral peak. Formant bandwidths were generally set to constants for this synthesis so as to minimize possible bandwidth compensation for source deficiencies (bandwidths were increased when the glottis was open for [h]).

4. Step 4: Set OQ, the open quotient

The open quotient **OQ**, in percent of the total period, determines the time during which the waveform is nonzero.

TABLE XII. Control parameters that can be varied over time in the KLSYN88 synthesizer configuration. Each control parameter is assigned a name, an indication of whether it is a constant or can be made to vary over time, a minimum value, a default value that applies if the user makes no changes, a maximum value, and an English description of its effect on the synthesis.

SYM	V/C	MIN	VAL	MAX	Description
F0	v	0	1000	5000	fundamental frequency, in tenths of an Hz
AV	v	0	60	80	amplitude of voicing, in dB
OQ	v	10	50	99	open quotient (voicing open-time/period), in %
SQ	v	100	200	500	speed quotient (rise/fall time of open period, LF model only), in %
TL	v	0	0	41	extra tilt of voicing spectrum, dB down @ 3 kHz
FL	v	0	0	100	flutter (random fluct in $f0$), in % of maximum
DI	v	0	0	100	diplophonia (pairs of periods migrate together), in % of max
AH	v	0	0	80	amplitude of aspiration, in dB
AF	v	0	0	80	amplitude of frication, in dB
F1	v	180	500	1300	frequency of the 1st formant, in Hz
B1	v	30	60	1000	bandwidth of the 1st formant, in Hz
DF1	v	0	0	100	change in $F1$ during open portion of a period, in Hz
DB1	v	0	0	400	change in B1 during open portion of a period, in Hz
F2	v	550	1500	3000	frequency of the 2nd formant, in Hz
B2	v	40	90	1000	bandwidth of the 2nd formant, in Hz
F3	v	1200	2500	4800	frequency of the 3rd formant, in Hz
B3	v	60	150	1000	bandwidth of the 3rd formant, in Hz
F4	v	2400	3250	4990	frequency of the 4th formant, in Hz
B4	v	100	200	1000	bandwidth of the 4th formant, in Hz
F5	v	3000	3700	4990	frequency of the 5th formant, in Hz
B5	v	100	200	1500	bandwidth of the 5th formant, in Hz
F6	v	3000	4990	4990	frequency of the 6th formant, in Hz (frication excited, or if NF = 6)
B6	v	100	500	4000	bandwidth of the 6th formant in Hz (only applies if NF = 6)
FNP	v	180	280	500	frequency of the nasal pole, in Hz
BNP	v	40	90	1000	bandwidth of the nasal pole, in Hz
FNZ	v	180	280	800	frequency of the nasal zero, in Hz
BNZ	v	40	90	1000	bandwidth of the nasal zero, in Hz
FTP	v	300	2150	3000	frequency of the tracheal pole, in Hz
BTP	v	40	180	1000	bandwidth of the tracheal pole, in Hz
FTZ	v	300	2150	3000	frequency of the tracheal zero, in Hz
BTZ	v	40	180	2000	bandwidth of the tracheal zero, in Hz
A2F	v	0	0	80	amplitude of frication-excited parallel 2nd formant, in dB
A3F	v	0	0	80	amplitude of frication-excited parallel 3rd formant, in dB
A4F	v	0	0	80	amplitude of frication-excited parallel 4th formant, in dB
A5F	v	0	0	80	amplitude of frication-excited parallel 5th formant, in dB
A6F	v	0	0	80	amplitude of frication-excited parallel 6th formant, in dB
AB	v	0	0	80	amplitude of frication-excited parallel bypass path, in dB
B2F	v	40	250	1000	bandwidth of frication-excited parallel 2nd formant, in Hz
B3F	v	60	320	1000	bandwidth of frication-excited parallel 3rd formant, in Hz
B4F	v	100	350	1000	bandwidth of frication-excited parallel 4th formant, in Hz
B5F	v	100	500	1500	bandwidth of frication-excited parallel 5th formant, in Hz
B6F	v	100	1500	4000	bandwidth of frication-excited parallel 6th formant, in Hz
ANV	v	0	0	80	amplitude of voicing-excited parallel nasal formant, in dB
A1V	v	0	60	80	amplitude of voicing-excited parallel 1st formant, in dB
A2V	v	0	60	80	amplitude of voicing-excited parallel 2nd formant, in dB
A3V	v	0	60	80	amplitude of voicing-excited parallel 3rd formant, in dB
A4V	v	0	60	80	amplitude of voicing-excited parallel 4th formant, in dB
ATV	v	0	0	80	amplitude of voicing-excited parallel tracheal formant, in dB

This waveform is later filtered by the "tilt" low-pass filter, which may increase the effective open time by rounding the waveform corner at closure. An appropriate value for **OQ** is very difficult to determine directly from spectral or waveform characteristics, so a typical default value of 50% for a male voice or 60% for a female voice is usually chosen as a departure point. The primary acoustic effect of changes in open time is to increase and decrease the amplitude of the first harmonic relative to adjacent harmonics. Thus trial and error adjustment of **OQ** to match first-harmonic amplitude was attempted. In some cases, locations of spectral zeros can be used to confirm the correct value for this parameter, but spectral zeros were not evident at low frequencies in the data of LK. Matching of first-harmonic amplitude by adjustments to **OQ** was generally successful and resulted in physiologically reasonable values for **OQ** over most of the utterance. As would be expected, **OQ** had to be decreased for glottalization, and increased for breathiness associated with adjacent voiceless consonants.There was one problem situation. Frequently, at the end of a voicing interval followed by silence, the first harmonic gained a large relative prominence that could not be matched only by changes to **OQ**; it was necessary to also increase spectral tilt **TL**, increase **AV**, and increase the bandwidths of the lower formants.

5. Step 5: Set TL, the spectral tilt

The spectral tilt **TL**, in dB, determines the amount of extra attenuation of higher harmonics of the voicing source spectrum. Initial values for the **TL** parameter were determined by observing whether the dft spectrum was perfectly periodic in the frequency region of the third and fourth formants. In cases where the spectrum is essentially periodic, **TL** is set to zero, while indications of random aspiration noise in place of harmonics implies a **TL** value of about 20 dB. Transitions between these two states can be gradual, but are sometimes quite abrupt. Simultaneous trial and error matching of both **TL** and **AH** may be necessary to get the right balance of harmonics and aspiration noise in each part of the spectrum.

6. Step 6: Set AH, the amplitude of aspiration noise

The amplitude of aspiration noise, **AH** in dB, that is used to generate [h] and the aspiration of [p,t,k] is also used to add breathiness noise to voiced portions of utterances. Initial values for the **AH** parameter were obtained on a trial and error basis by matching levels of $F3$ and $F4$ excitation in spectra that were clearly inharmonic (after the **TL** parameter had been optimized so as to attenuate higher harmonics). The process is complicated by the statistical variability of noise processes, but if one averages visually over several 25-ms windowed spectra, the results are more stable and reliable. An appropriate value for **AH** in an essentially perfectly harmonic vowel is either zero, or perhaps some value about 10 dB lower than in a clearly breathy vowel (i.e., some small degree of aspiration noise may be present even if no visible manifestation appears in the spectrum).

7. Step 7: Set DI, waveform of diplophonic double pulsing

The double-pulsing parameter is provided in order to be able to mimic a special form of laryngealization that is occasionally seen for some talkers—alternate periods are delayed and attenuated [recall Fig. 12(b)]. The **DI** parameter is calibrated in percent such that 100% means that the first of two pulses is delayed maximally and is attenuated to zero, while 50% would delay only half as much and attenuate to half its unperturbed amplitude. For speaker LK, several offsets were laryngealized in this way, as indicated by the **DI** parameter in Figs. 15 and 16. Appropriate initial values for **DI** were determined by studying the natural waveform for obvious examples of doubling pulsing. Refined estimates of appropriate synthesis values for **DI** were obtained by comparing waveforms of synthetic and natural speech.

In summary, the seven steps enumerated above provide a reasonably straightforward procedure for synthesizing a close imitation for reiterant [hV] or [ʔV] versions of any sentence. Several female and male voices have been successfully imitated by this process. It may even be possible to automate parts of the analysis (Ananthapadmanabha, 1984; Fujisaki and Ljungqvist, 1986).

B. Synthesis generalizations concerning breathiness and laryngealization

The two reiterant sentences synthesized using values shown in Figs. 15 and 16 are representative of the larger corpus of speakers in that, in general, breathiness noise increased toward the end of a vowel (see Fig. 15) and near voiceless consonants (see Fig. 16), and was greater in unstressed syllables and in the final syllable of a sentence. Similar results have recently been described by Chasaide and Gobl (1987). Typical voicing source parameter values for normal voicing and breathy voicing for this speaker were:

Par	Normal	Breathy
AV	60	60
OQ	60	80
TL	8	24
AH	0–40	52
DI	0	0
FL	25	25

It is clear from listening, and from the **TL** value for the normal voice, that LK has a somewhat breathy version of normal phonation, but it is also clear from the time variation in Figs. 15 and 16 that her voice becomes more breathy in specified predictable circumstances.

In order to match the reiterant spectra of the [ʔɑ] and [hɑ] utterances of LK, it was also necessary to make some modifications to the vocal-tract transfer function. The formant frequencies were left roughly constant (exactly constant in the [ʔɑ] sentence and constant except for a slight rise during the [h]'s of the [hɑ] sentence). Formant bandwidths **B1** and **B2** were increased whenever formant spectral peaks were flattened; these times corresponded very well with times at which source parameters indicated that the

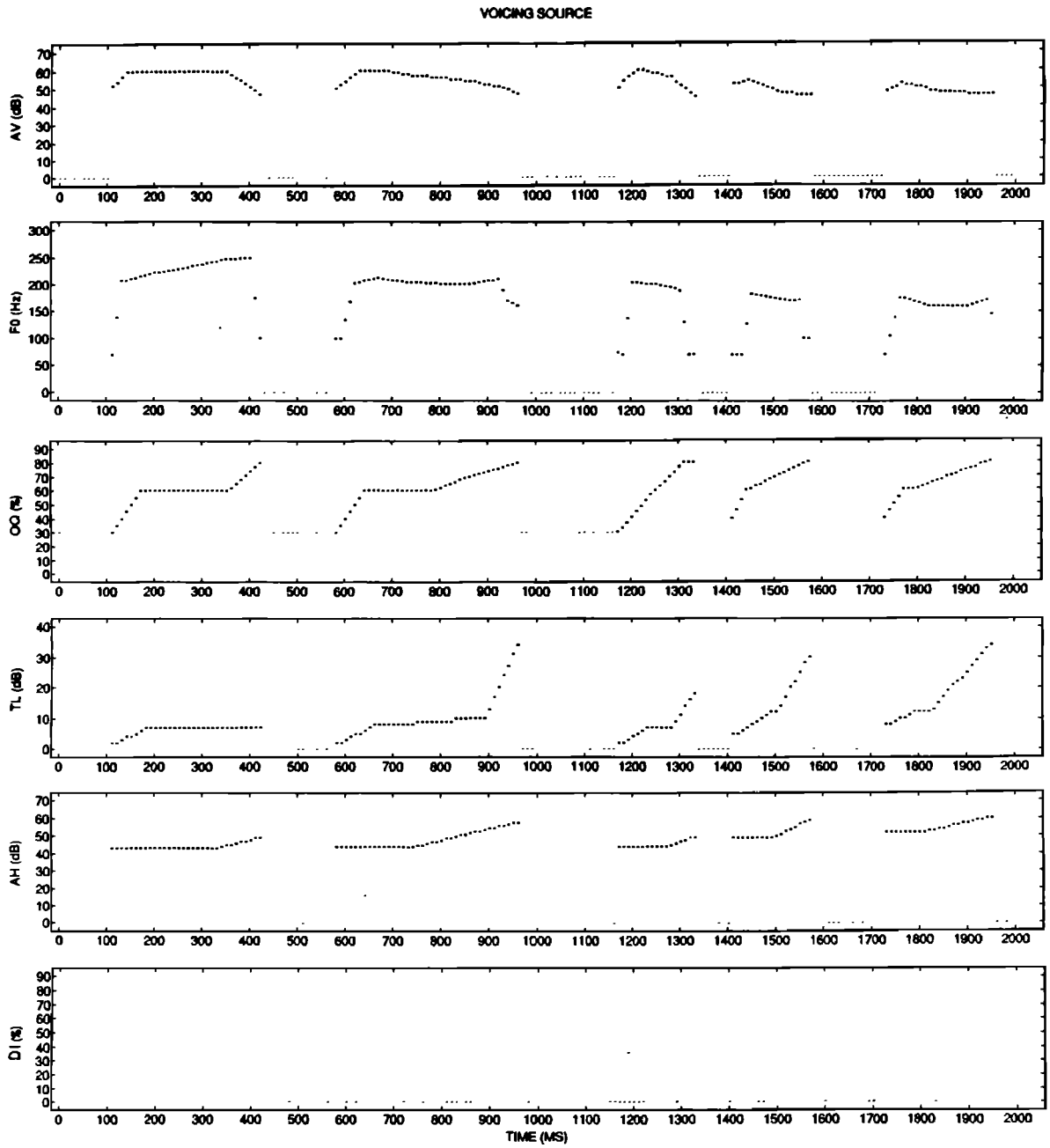


FIG. 15. Synthesis parameter values as a function of time for the [ʔɑ] reiterant imitation of "Steve eats candy cane" by female speaker LK. Bold data points indicate time intervals when voicing is on.

glottal opening was significantly increased. For example, there were many cases of vowel terminations where the first-formant peak essentially disappeared from the spectrum, and the first-formant bandwidth had to be increased substantially. For this speaker, acoustic coupling to the resonances of the trachea during breathy vowels and aspiration was minimal, so no use was made of the synthesis parameters available to introduce tracheal pole and zero pairs into the vocal-tract transfer function.

The reiterant speech data also reveal changes to a typical vowel induced by a glottal-stop onset or offset. There is a rapid (30 ms) fall in f_0 accompanying a glottalized onset or offset, and the amplitude of voicing **AV** is reduced by about 6 dB when f_0 is at its lowest. In a glottalized onset, the primary waveshape-controlling parameter to be affected is the open quotient, **OQ**, which is reduced to perhaps 30% in the vicinity of the glottal stop. It is likely that **AH** will be reduced and **TL** reduced in this interval if the vowel is otherwise somewhat breathy. Similar changes occur at a glottalized offset, but there is less of a reduction in breathiness correlates.

There were three examples of temporally offset glottal pulses (double pulsing) in the two sentences from LK. At these times, it was observed that the **TL** and **AH** parameters were not changed as much as for other vowel offsets, suggest-

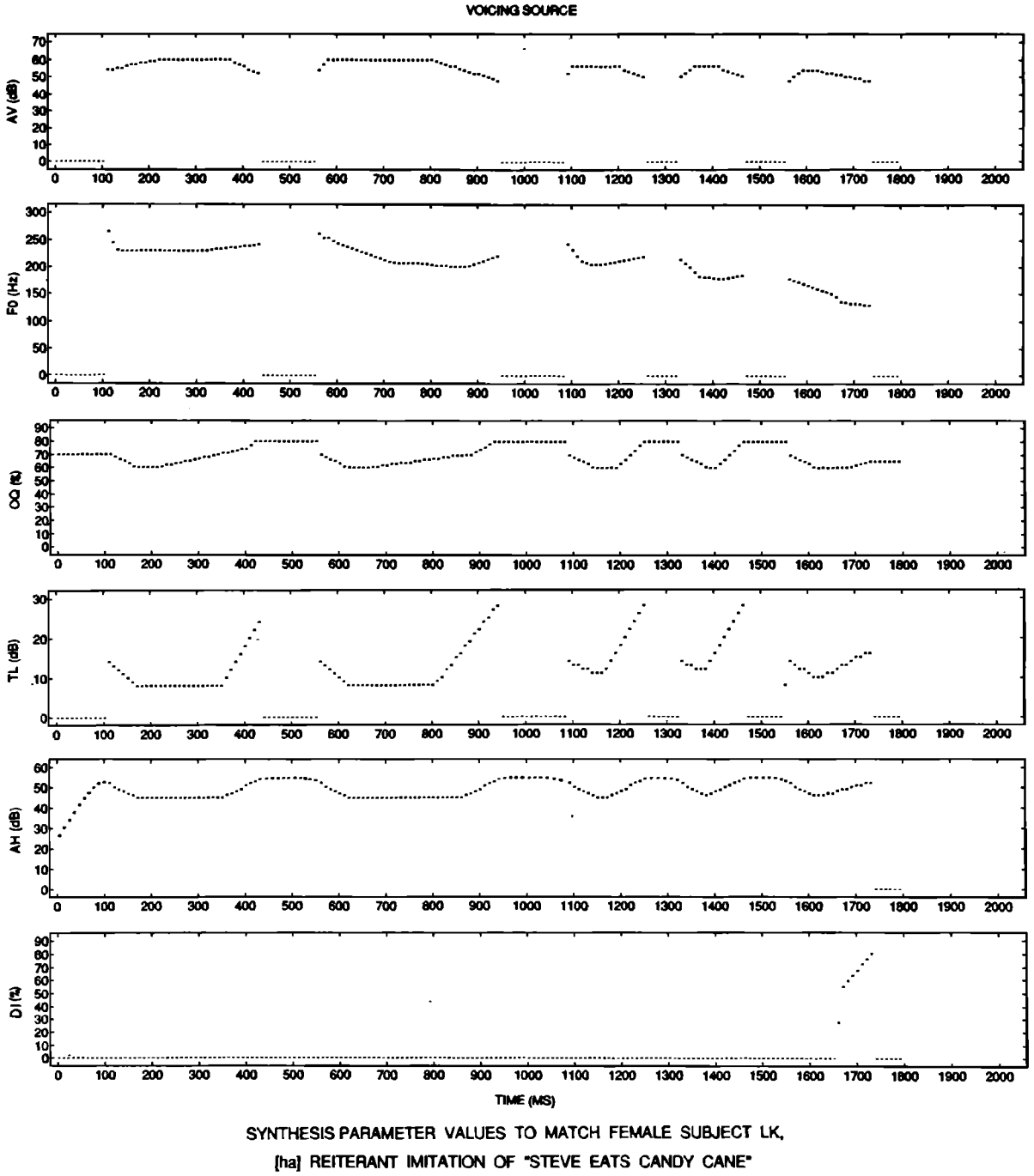


FIG. 16. Synthesis parameter values as a function of time for the [hɑ] reiterant imitation of "Steve eats candy cane" by female speaker LK.

ing a possible distinction between a laryngealized offset accompanied by double pulsing, and a more typical "breathy falling-pitch offset." It is not known whether the ability to mimic details of the sort exemplified by a few periods of double pulsing at voicing offset has much perceptual importance, but it is clear that there exist some voices that display a degree of double pulsing most of the time (Lieberman, 1963). For these voices, it is very likely that the parameter is of perceptual importance.

In summary, during breathy voicing, several of the voicing source parameters change together in such a way as to: (1) increase the relative strength of the first harmonic (by increasing the open quotient **OQ**), (2) reduce the strength of higher harmonics (by increasing spectral tilt **TL**), and (3) add in some aspiration noise at mid and high frequencies (by increasing the amplitude of turbulent aspiration noise **AH**). In addition, the partial abduction of the vocal folds associated with breathiness has effects on the vocal-tract transfer function; the bandwidths of lower formants, especially $F1$, are increased due to increased glottal losses, and additional formantlike spectral peaks and valleys may be introduced into the transfer function due to acoustic coupling to the trachea. These five acoustic effects can be observed not only in a breathy vowel, but also in transitions between normal vowels and voiceless consonants such as [h], or at the offset of a vowel into silence. There are regular

TABLE XIII. Values for synthesis constants that differ from default values for the reference stimulus used in the breathiness perception test.

Parameter	Value
NF	4
DU	300
OQ	65
TL	3
AH	40
F1	800
B1	200
F2	1300
B2	110
F3	2850
B3	180
F4	3700
B4	250

acoustic manifestations of breathiness and laryngealization that occur in predictable locations over the course of typical English sentences. These manifestations have been parametrized and mimicked using a simple synthesis model. Utilization of these types of voice quality cues in speech synthesis by rule should enhance the naturalness of the speech so produced, especially for female voices.

IV. BREATHINESS PERCEPTION TEST USING SYNTHESIS

The stimuli used in the breathiness perception test included 12 synthetic vowels: a reference stimulus and a set of 11 modified stimuli. The reference stimulus was generated by using the default values in Tables XI and XII, except for the constant parameters in Table XIII, and the two time-varying parameters in Fig. 17. The modified stimuli are defined by indicating parameters that change from the reference values, Table XIV, and by showing spectral cross sections of each stimulus in Fig. 18. The reference stimulus was patterned after the LK female voice, but with no cues to breathiness included. Other stimuli add various individual breathiness cues (increased first harmonic amplitude, or bandwidth increases, or added aspiration noise, or spectral tilt at high frequencies, or tracheal pole-zero added) and several combinations of such cues (noise plus tilt, all cues together).

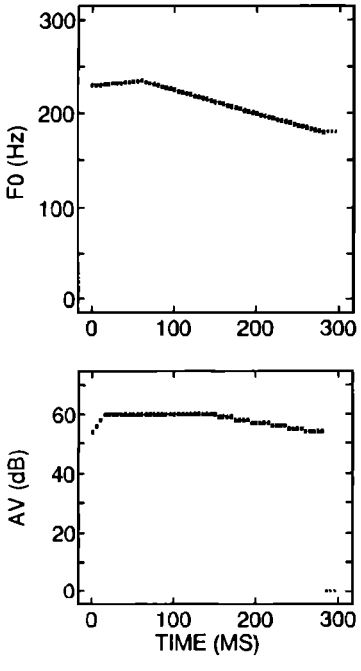


FIG. 17. Values for time-varying parameters for the reference synthetic stimulus used in the breathiness perception test.

TABLE XIV. Synthesis parameters that have been changed for each of the stimuli contrasted with the reference vowel in the breathiness perception test. GO is the overall gain scale factor, in dB.

(1) Fundamental component boosted 6 dB	FTP and FTZ follow f_0, BTP = 50, BTZ = 100
(2) Fundamental component boosted 10 dB	FTP and FTZ follow f_0, BTP = 50, BTZ = 150
(3) Fundamental frequency lowered initially	f_0 down 10 Hz over first 100 ms
(4) Formant bandwidths increased	B1 = 500, B2 = 170
(5) Spectral tilt down 15 dB at 3 kHz	TL = 15, GO = 65
(6) Spectral tilt down 25 dB at 3 kHz	TL = 25, GO = 67
(7) Aspiration noise of 54 dB added	AH = 54
(8) Aspiration noise of 60 dB added	AH = 60
(9) Spectral tilt of 15 dB and aspiration of 55 dB	TL = 15, GO = 65, AH = 55
(10) Spectral tilt of 20 dB and aspiration of 50 dB	TL = 20, GO = 67, AH = 50
(11) Ditto, plus bandwidth widening and OQ increase	TL = 15, GO = 65, AH = 55, B1 = 400, B2 = 170, OQ = 75

The instructions to the subjects were as follows:

"You will hear pairs of synthetic vowels, in which the first vowel, the reference, is always the same. The second vowel of the pair has been modified in some way. Most of the modifications were intended to increase the perceived breathiness of the second member of the pair, but we were clearly not always successful. In fact, some of the changes result in no change, in a nasalized vowel, or in vowel sound qualities that would be difficult for a human to produce. We are interested in two things: (1) the degree to which the second vowel is perceived to be more breathy than the first, and (2) whether the change in sound quality is natural (could be produced by the same human talker who produced the reference, rather than, for example, by computer processing). Since these two judgments are more-or-less independent, we will play the 6-min tape to you twice, first asking for breathiness judgments, and then for naturalness judgments. Because some of the changes seem to have made some of the vowels sound nasalized, during the naturalness rating process, we would also like you to star any item that sounds nasalized.

Instructions before first playing: The breathiness scale will go from 0 (no change in breathiness between reference

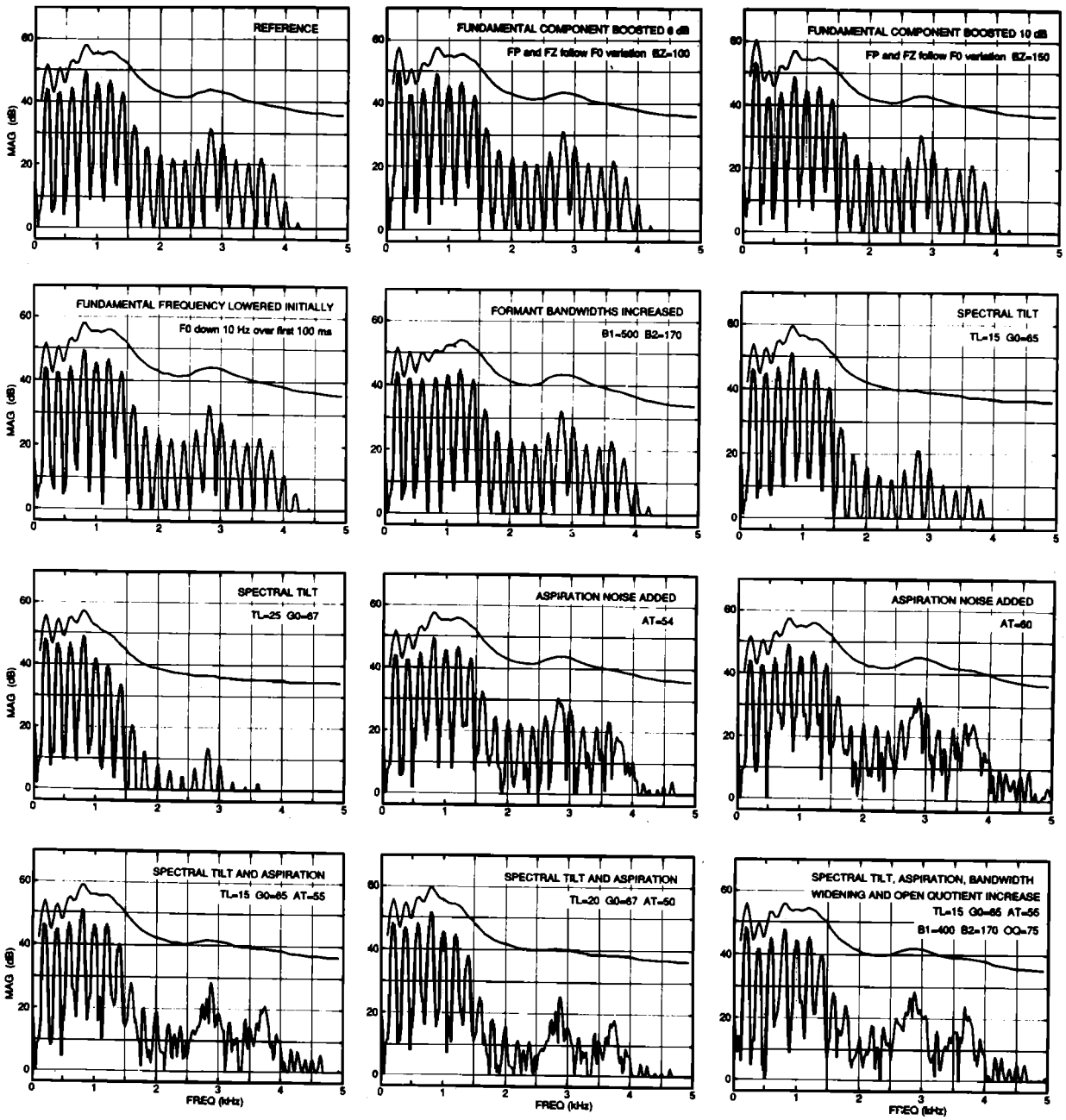


FIG. 18. Spectral cross sections (no pre-emphasis) for each of the synthetic stimuli used in the breathiness perception test.

and second vowel) to 5 (maximal increase in breathiness). If a vowel actually sounds less breathy that the reference, a negative number, such as -1, can be used as a response. Try not to downgrade ratings of breathiness for unnatural stimuli; simply wait until the second playing of the tape to express dissatisfaction with the stimulus. The first ten trials will be treated as practice trials in order to allow you to hear the range of stimuli to be encountered; nevertheless, please respond and write down a practice answer for these trials.

Instructions before second playing: The naturalness rating scale to be used will go from 0 to 5, with 5 being perfectly natural, and zero being very unnatural or machinelike (not possible for a human to imitate through any natural speech production process, probably produced by artificial means). Remember that we would also like you to star any item that sounds nasalized."

Subjects listened over TDH model HD-420 earphones to a randomized tape recording that included a practice run of 14 trials to familiarize subjects with the expected range of breathiness, and five blocks of trials that were scored and combined to obtain average ratings of breathiness for each stimulus.

Five subjects participated individually in the perception test. The results are summarized in Table XV, which gives the average breathiness rating response of each listener, the average naturalness rating, and the fraction of times a listener indicated nasality was present. Also included in the table are group averages. These experienced listeners were re-

TABLE XV. Listener ratings of: (1) change in breathiness relative to the reference stimulus, (2) naturalness, and (3) nasality for each of the 11 synthetic stimuli that were contrasted with the reference vowel in the breathiness perception test.

Change in breathiness (− 5.0 to 5.0): Condition	CB	SH	SM	SB	KS	Av
(1) Fundamental component boosted 6 dB	1.2	0.0	1.6	0.4	1.4	0.92
(2) Fundamental component boosted 10 dB	2.0	− 0.7	2.2	0.2	1.6	1.26
(3) Fundamental frequency lowered initially	0.0	0.0	− 0.4	0.0	0.6	0.04
(4) Formant bandwidths increased	0.8	− 0.1	0.6	0.4	0.6	0.46
(5) Spectral tilt down 15 dB at 3 kHz	1.8	− 1.2	2.2	0.2	2.0	1.00
(6) Spectral tilt down 25 dB at 3 kHz	2.8	− 1.8	2.0	0.6	3.2	1.36
(7) Aspiration noise of 54 dB added	0.4	1.3	2.0	0.0	2.0	1.14
(8) Aspiration noise of 60 dB added	1.8	2.4	3.0	2.6	4.4	2.88
(9) Spectral tilt of 15 dB and aspiration of 55 dB	3.4	1.3	3.2	1.4	4.2	2.70
(10) Spectral tilt of 20 dB and aspiration of 50 dB	3.8	1.4	2.8	0.8	4.4	2.64
(11) Ditto, plus bandwidth widening and OQ increase	4.6	3.0	3.4	3.0	4.8	3.76
Naturalness (0 to 5.0): Condition	CB	SH	SM	SB	KS	Av
(1) Fundamental component boosted 6 dB	4.8	5.0	2.8	4.0	2.4	3.80
(2) Fundamental component boosted 10 dB	3.8	5.0	2.6	2.4	1.6	3.08
(3) Fundamental frequency lowered initially	5.0	5.0	2.6	4.8	4.4	4.36
(4) Formant bandwidths increased	2.4	4.2	1.2	5.0	2.8	3.12
(5) Spectral tilt down 15 dB at 3 kHz	3.4	5.0	3.6	4.0	3.4	3.88
(6) Spectral tilt down 25 dB at 3 kHz	1.8	5.0	3.6	3.2	2.4	3.20
(7) Aspiration noise of 54 dB added	5.0	5.0	3.8	5.0	3.8	4.52
(8) Aspiration noise of 60 dB added	4.0	5.0	4.2	5.0	2.4	4.12
(9) Spectral tilt of 15 dB and aspiration of 55 dB	4.4	5.0	4.4	3.4	4.2	4.28
(10) Spectral tilt of 20 dB and aspiration of 50 dB	3.4	5.0	4.8	4.6	4.2	4.40
(11) Ditto, plus bandwidth widening and OQ increase	5.0	5.0	5.0	2.4	3.0	4.08
Nasality (0 to 1.0): Condition	CB	SH	SM	SB	KS	Av
(1) Fundamental component boosted 6 dB	0.4	0.8	0.8	0.2	0.6	0.56
(2) Fundamental component boosted 10 dB	0.4	1.0	1.0	0.6	0.6	0.72
(3) Fundamental frequency lowered initially	0.0	0.0	0.2	0.0	0.0	0.04
(4) Formant bandwidths increased	1.0	0.8	1.0	0.2	1.0	0.80
(5) Spectral tilt down 15 dB at 3 kHz	0.4	0.0	0.2	1.0	0.0	0.32
(6) Spectral tilt down 25 dB at 3 kHz	0.6	0.2	0.0	1.0	0.0	0.36
(7) Aspiration noise of 54 dB added	0.0	0.0	0.0	0.0	0.0	0.00
(8) Aspiration noise of 60 dB added	0.0	0.0	0.2	0.0	0.0	0.04
(9) Spectral tilt of 15 dB and aspiration of 55 dB	0.0	0.0	0.0	0.4	0.0	0.08
(10) Spectral tilt of 20 dB and aspiration of 50 dB	0.6	0.0	0.0	1.0	0.0	0.32
(11) Ditto, plus bandwidth widening and OQ increase	0.0	0.0	0.0	0.4	0.0	0.08

markably self consistent; except for SB, they rarely differed by more than 1 in assigning an integer to ratings of breathiness and naturalness of five repetitions of each condition.

While there are some intersubject differences, the pooled results point to aspiration amplitude as the dominant factor in eliciting judgments of increased breathiness. An increase in first-harmonic amplitude, by itself, does not induce the sensation of breathiness for most listeners. The reason is probably the high fundamental frequency employed in the synthesis; an increase in the spectrum at about 200 Hz is consistent with the appearance of a nasal pole indicative of nasalization. In fact, nasalization judgments are quite common for stimuli in which only the fundamental component has been increased. We tentatively conclude that either breathiness is signaled differently for men and women, or that the increases in the first harmonic observed in production data from women must be accompanied by other cues to be interpreted by the listener as cues to breathiness (see below).

An increase in the first-formant bandwidth, stimulus 4, is also by itself ineffectual in suggesting breathiness. The stimulus sounds both nasal and unnatural. Again, it is either the case that production data (indicative of increased bandwidths for breathy vowels) do not comport with perceptual strategies, or that the bandwidth increases must be accompanied by certain other cues before they are unambiguously interpreted as relating to breathiness.

A lowering of f_0 had no effect on perceived breathiness. Tilting down the spectrum to reduce the amplitudes of higher frequency harmonics increased breathiness judgments a little bit for some listeners, but was heard as unnatural or as an increase in nasality for others. By itself, spectral tilt does not appear to be a strong cue to breathiness, perhaps again because it does not occur naturally by itself, but rather only in conjunction with certain other cues to breathiness.

The addition of aspiration noise increases the number of breathiness judgments significantly for most listeners. If the harmonic spectrum is also attenuated at high frequencies by

tilting the spectrum down (as is observed in natural breathy vowels), then less noise is required to achieve the same degree of perceived breathiness. Stimuli with aspiration noise were all judged to be natural.

The stimulus perceived to be most breathy was the one in which all of the various cues observed in natural stimuli were present: aspiration noise, spectral tilt, longer open quotient, and increased bandwidths of $F1$ and $F2$. The increase in first-harmonic amplitude and the increases in bandwidths that, by themselves, induced the sensation of nasality for many listeners, did not produce nasality judgments for this stimulus. It appears that nasality perception is a rather complex function of acoustic properties that are attributed to breathiness under some circumstances and attributed to nasality in others. This outcome is troublesome for many simple models of speech perception but is in agreement with the philosophy known as the motor theory of speech perception (Liberman *et al.*, 1967) and other models that permit learning of complex cue interactions (Klatt, 1986a, 1989).

V. DISCUSSION AND CONCLUSIONS

The analysis of reiterant speech from ten female and six male speakers has revealed a number of acoustic cues related to breathiness. A perception test has established the importance of aspiration noise as a component of a breathy voice quality and has shown the complexity of some cue interaction perceptual strategies. Synthesis efforts using a new version of the Klatt formant synthesizer, KLSYN88, which has a new voicing source model, verify that it is possible to mimic several female voices with an accuracy that makes it difficult to distinguish between the original recording and the synthesis. The following sections go into greater detail on each of these topics and speculate as to the implications of several of our results.

A. Acoustic analysis

A breathy voice quality is signaled by a surprisingly large number of diverse acoustic cues, all related to the presumed posterior glottal opening posture shown in panel (3A) of Fig. 1. First of all, the posterior opening leads to a dc flow component and the generation of aspiration noise throughout a period, with noise intensity perhaps increasing during the open phase. The open quotient is increased, and this leads to a relative increase in the amplitude of the first harmonic, H1, by 6 dB or more. In addition, there is nonsimultaneous closure of the folds over their length, with the posterior portion of the folds making contact somewhat later than the anterior edges; this pattern of closure leads to a reduction in the relative amplitudes of higher harmonics in the source spectrum. As a result of these two factors, aspiration noise tends to replace harmonics at frequencies above about 1.5 kHz in a breathy vowel. The posterior glottal opening increases low-frequency losses in the vocal-tract transfer function, resulting in an increased first-formant bandwidth and a less distinct first-formant peak in the spectrum. The posterior glottal opening also provides acoustic coupling to the subglottal system, resulting in the possible appearance of tracheal poles and zeros in the vocal-tract transfer function. The poles tend to appear at predictable frequency locations—about 600, 1400, and 2200 Hz for a female voice and somewhat lower for males.

The degree to which an individual vowel takes on the cues of breathiness can vary considerably over the course of an utterance. In our data, cues to breathiness tend to increase for unstressed syllables, for final syllables, and at the margins of voiceless consonants. In a stressed vowel with a relatively high fundamental frequency, the spectrum may be perfectly periodic, even for the most breathy of our speakers. It is almost certainly this time variation that contributes to naturalness and highlights a breathy voice.

Males and females differ on average in the two perceptually most important acoustic measures of breathiness—amount of aspiration noise in the $F3$ region of the spectrum and relative amplitude of the first harmonic. Females are more breathy than males to a significant degree. However, within each gender, there is much greater variation in acoustic manifestations of breathiness, with some males being more breathy than many females. In addition, it is likely that any individual is capable of adopting a fairly wide range of speaking styles that differ in degree of breathiness. Thus it is dangerous to make sweeping generalizations with regard to sex typing, as well as the behavior of particular individuals.

We have discovered evidence for a breathy-laryngealized mode of vibration that is employed by many speakers at the ends of utterances. There is increased noise in the $F3$ region of the spectrum, but the open quotient does not increase, as it normally would whenever the posterior portions of the folds are spread in preparation for a voiceless consonant or for breathing. We speculate that the arytenoid cartilages are rotated inward to facilitate the maintenance of voicing in the face of a developing posterior glottal chink, a lowering subglottal pressure in anticipation of the end of speaking, and a slack vocal-fold posture appropriate for low f_0. This breathy-laryngealized vibration pattern will have to be incorporated into representational schemes for the phonetic description of speech and may possibly require changes to the distinctive features used to represent language. For example, the feature system of Halle and Stevens (1971) might be revised to allow simultaneous plus values for the features *spread* and *constricted* to represent a breathy-laryngealized voice quality, where spread is redefined to refer to the posterior interarytenoid separation, and constricted to refer to the medial compression resulting from rotational motion of the arytenoids.

A two-mass model of vocal-fold vibrational behavior was devised by Ishizaka and Matsudaira (1968) in order to better explain the transfer of energy from static lung pressure to dynamic vibratory motions of the vocal folds. While successful in these terms (Stevens, 1977), the model does not have the flexibility to generate output volume velocity waveforms with corner rounding associated with nonsimultaneous closure. Neither does the model permit a static posterior glottal opening necessary to simulate breathy voicing. Therefore, use of the model in synthesis of speech (Flanagan *et al.*, 1975) is likely to lead to suboptimal synthetic imitations of female voices until such time as these features of laryngeal behavior are incorporated.

A new measure of the amount of aspiration noise in a voiced sound has been proposed. By filtering out a region of the spectrum in the vicinity of $F3$, it is possible to isolate a waveform that can be visually interpreted in terms of whether the source is periodic or random. While we have used subjective rating scales to quantify this judgment, further research, such as quantifying the peak-to-valley ratio in the energy contour over a period, may make it possible to automate its measurement and thereby remove the subjective component.

Formant-frequency measurement can be very difficult in a breathy vowel, particularly in a voice with a high fundamental frequency. The strong first harmonic can be confused with a formant; the first-formant bandwidth increase can make it difficult to detect a local maximum in the spectrum corresponding to $F1$, and the tracheal coupling can cause extra peaks in the spectrum that are easily confused with formants. When harmonics are widely spaced in the spectrum, each of these complications becomes more difficult to deal with because the vocal-tract transfer function is essentially sampled only at harmonic locations. The remarkable ability of the human perceptual system to deal with these problems calls into question the degree to which formants are actually perceptual dimensions (Bladon, 1982). Alternatives include whole-spectrum template-matching approaches (Klatt, 1982), and complex "spectrum-interpretive" strategies, a discussion of which goes beyond the scope of the present paper.

Vowels in natural utterances are rarely perfectly periodic. But it has been difficult to characterize exactly how individual periods differ from perfect periodicity. A popular theory has been that there is a jitter component, or Gaussian random fluctuation in individual periods, but efforts to turn this into an effective synthesis strategy have failed. Such jitter is either imperceptible if added in amount corresponding to prior literature on the measurement of jitter, or sounds like a harsh pathological voice quality when jitter is increased. Our observations on deviations from perfect periodicity in this database suggest two new methods of synthesizing natural deviations: (1) a slowly varying "pseudorandom flutter" consisting of a sum of sine waves and (2) an optional diplophonic double pulsing that occurs in certain fairly predictable situations. The perceptual importance of temporal variability and the success of our proposed synthesis strategy have yet to be established. Also, in order to better account for diplophonia in terms of the physics of larynx behavior, a more complex physiological model may be required in which the three-dimensional nature of the vibration pattern is considered (Titze, 1974; Titze and Talkin, 1979).

Our results must be qualified by the limited scope and artificial nature of the data base. We have analyzed only one vowel in a sample of reiterant speech approximating two sentences. The speech is read with regular pauses between utterances. In the future, other vowels should be employed, and the analysis techniques should be extended to more natural databases of spontaneous speech.

B. Perception of breathiness

The perception test using natural speech samples edited from the reiterant sentences revealed that females, on average, are perceived to be slightly more breathy than males. The perception data also are in agreement with the acoustic data in showing wide variation within each gender; two females are judged less breathy than the male average, and one male is judged more breathy than the female average. A correlation analysis with ten acoustic parameters related to breathiness revealed only two statistically significant correlations with the perceptual responses—one with the degree of aspiration noise seen in the $F3$ region of the spectrum, and one with the relative strength of the first harmonic H1.

The second type of perception test used synthetic speech in order to be able to determine the perceptual importance of individual acoustic cues in isolation and in combination. This test was somewhat unusual in that the standard comparison stimulus was patterned after a female voice rather than using a male voice as has been the practice so often in the past. Stimuli were judged as to degree of breathiness, naturalness, and nasality. The strongest single cue to breathiness was found to be the amplitude of the aspiration noise added to the spectrum. However, the stimulus that was most preferred in terms of breathiness and naturalness was one in which all cues (add aspiration noise, increase spectral tilt, increase open quotient to increase H1, and widen first-formant bandwidth) were present. It is as if the perceiver is aware of all of the systematic changes that go into breathy phonation and uses these expectations during perception in such a way that no single cue is as effective as all in combination.

When only the first-harmonic amplitude was increased, some subjects heard an increase in breathiness, but many others heard an increase in nasality. While this result has never been reported in previous perceptual studies of H1, it may well be due to the fact that we simulated a female voice in which the first harmonic, on average, was about 200 Hz. This is close to the frequency of the lowest pole in the transfer function of a nasalized vowel. The same increase in amplitude of the first harmonic, if accompanied by aspiration noise, is never heard as nasalized. The implication is probably not that different perceptual strategies are employed for male and female voices. Rather, it is more likely that single-cue manipulations can create somewhat unnatural stimuli that result in perceptual ambiguities. This is also observed for first-formant bandwidth increases, which by themselves increase the nasality of a stimulus, but taken in conjunction with other cues to breathiness are not heard as nasalized.

It is interesting to speculate on the detailed nature of perceptual strategies, given the ambiguity introduced by a strong first-harmonic amplitude. Is first-harmonic amplitude a perceptual cue whose interpretation involves a complex interaction with other perceptual cues such as the degree of periodicity at high frequencies, or are spectra interpreted in a wholistic fashion against templates representing nasalized or breathy versions of various speech sounds? Neither alternative is very compelling to us as a perceptual strategy. The first option seems undesirable because, from our perspective, speech perception ought to be simple and direct at the lowest levels, with little if any cognitive processing. The second option involving a template approach may not be workable for various reasons having to do

with natural variability within and across speakers (Klatt, 1986a, 1989).

C. Synthesis of a female voice

The standard acoustic theory of speech production has been reviewed in light of several source–tract interactive phenomena that have been identified by Gunnar Fant and his co-workers. Based on this theoretical background and the experimental data presented here, we conclude that the old cascade/parallel formant synthesizer model described in Klatt (1980), while still useful for most perception experiments involving cues to various consonant and vowel distinctions, is not capable of mimicking a female voice very accurately. In this paper, we have identified certain cues related to a breathy voice quality that must be modeled in order to closely mimic most female voices. The new version of the Klatt cascade/parallel synthesizer, described here, has been augmented with: (1) a new voicing source model that has control parameters **F0, AV, OQ, TL, AH, FL, DI,** (2) an ability to change first-formant bandwidth dynamically over a period, using **DB1** to simulate the rapid change in glottal losses as the glottis opens and closes, and (3) an additional tracheal pole-zero pair with control parameters **FTP, FTZ, BTP,** and **BTZ.** This new version, KLSYN88, has been used to copy reiterant utterances from several female and male speakers with very good perceptual fidelity.

Experience with the new model suggests that the ability to make dynamic changes to the degree of aspiration noise intruding in the $F3$ region of the spectrum through use of the **AH** and **TL** parameters is the most important aspect of the model for improving the quality of female voice synthesis, but that all of the new parameters are useful in optimizing the match to individual spectra. If analysis by synthesis is used, and a good initial match to the f_0 contour is achieved, then successful matching of spectra sampled throughout an utterance has always resulted in perceptually successful synthesis using the new model. General rules for synthesizing breathiness variations over a sentence have been proposed based on the voice analysis and voice-matching efforts reported on here. Cues to breathiness are strongest at the end of an utterance, in unstressed syllables, and at the margins of voiceless consonants. Utilization of these rules may lead to improved naturalness of both male and female voices in synthesis-by-rule programs for text-to-speech devices such as DECtalk.

Variation in the timing of glottal pulses, using the new voicing source control parameters **FL** and **DI** may also increase the naturalness of synthetic speech, but at this point, we have not conducted the appropriate perceptual experiments to determine the importance of the timing variation informally observed in the reiterant corpus. Another random variation that might be added to synthesis by rule is a small random change in successive values of the **TL** parameter to simulate the observed tendency for rapid period-to-period variation in harmonic strengths above 2 kHz, even in spectra that do not have a strong noise component in this frequency region.

The KLSYN88 cascade/parallel formant synthesizer has been programmed in floating point in C on a Digital Equipment Corporation Microvax-II. This software will be made available to the speech community at cost.[19] It is hoped that the synthesizer can serve as a standard in preparing stimuli for perceptual experiments, and the fact that the algorithms are documented here will make it easier for researchers to replicate and extend the findings of others.

ACKNOWLEDGMENTS

This research was supported by Grant NS04332 from the National Institutes of Health. We are very grateful to Kenneth Stevens for many helpful comments and editorial assistance.

[1] Unfortunately, there is at present no standard terminology or agreement on the meaning of many of these terms.

[2] Should creak be defined to begin at a higher threshold for females, say 1.7 times 60 Hz, or about 100 Hz, and, if so, would they show as much creak as males? The decision depends on whether the definition of creak is perceptual (such as "picket fence" percept) or physiological (so much below normal pitch). In this paper, we will assume that creak refers to the absolutely low pitch sensation where individual pulses are audible, and laryngealization, on the other hand, occurs at some percentage change below a speaker's normal f_0 range (or perhaps when the glottal pulse becomes very narrow such that H1 is reduced in amplitude).

[3] The amplitude of the strongest harmonic near $F1$ was used to define A1, the level of $F1$, for reference purposes. This definition may underestimate the level of the vocal-tract transfer function at $F1$ by as much as 6 to 9 dB when two harmonics straddle $F1$ (Klatt, 1986a).

[4] In order to be able to compare breathiness of vowels with different phonetic quality, and thus different $F1$ values, the authors decided to abandon the use of A1, the level of $F1$, as a reference, in spite of the fact that it should tend to somewhat accentuate the difference between breathy and nonbreathy phonation due to the increase in first-formant bandwidth associated with a partially open glottis.

[5] There are two problems with the noise measure chosen: (1) Lower formants dominate waveform characteristics because they are more intense, but aspiration noise, whose presence is to be detected, tends to be restricted to higher formants and (2) when sampling speech, frequency components near the high-frequency sampling limit are represented by only a few sample points per cycle so that, if the fundamental frequency is not an exact multiple of the sampling interval, even a perfectly periodic waveform will appear to be different from period to period, and thus contaminated with noise according to this measure. The first problem could be solved by high-pass filtering at about 1.5 kHz, but the second problem is less amenable to a straightforward solution other than to significantly increase the sampling rate and thus decrease the granularity with which one compares periods.

[6] The analysis data that we will present suggest that aspiration noise might be a much more effective perceptual cue if the higher harmonics are simultaneously attenuated, as occurs for nonsimultaneous closure of the vocal folds along their length.

[7] In a spectrogram, the vertical striations indicative of periodic excitation would be replaced by a more random pattern if the excitation were exclusively turbulence noise. However, for relatively high-pitched female voices, the vertical striations are not as evident as for a male voice even when excitation is completely periodic.

[8] The amplitude of the second harmonic depends, in part, on locations of zeroes in the source spectrum and thus is not entirely satisfactory for reference purposes. However, the problems with the other alternatives seem more serious. The amplitude of the first formant (i.e., the peak in the underlying vocal-tract transfer function) is hard to determine from a harmonic spectrum because measured values depend to a considerable extent on whether a harmonic is centered on the formant frequency or two harmonics straddle the formant (Klatt, 1986a). Since the overall rms amplitude depends primarily on first-formant amplitude, it too suffers from this source of unpredictable variability, which can be up to 6–9 dB.

[9] Extreme changes to the amplitude of the fundamental component may be due to factors other than simply the open quotient. For example, a general reduction in the tilt of the spectrum would result if the vocal-fold closure event became nonsimultaneous due to glottal abduction. This effect can be simulated in a speech synthesizer by either the spectral tilt parameter TL

or the open quotient parameter OQ, as will be discussed in Section II below.

[10] Other measures described below will rate TW as even more breathy.

[11] Alternatively, one might use a two-pole inverse filter, but this requires fairly precise detection of the frequency and bandwidth of the third formant versus time. Such precision is not as critical when employing a 600-Hz bandwidth bandpass filter.

[12] Fant and Lin (1987) have attributed the extra pole between $F2$ and $F3$ to nonlinear superposition effects, but it seems to us to be more likely that the pole is due to tracheal resonance coupling, especially since it is seen both in aspiration and voiced excitation.

[13] Theoretically, the tracheal poles and zeros should come in pairs, where the zero is lower in frequency than the corresponding pole. However, it is difficult to detect evidence of a zero below the first tracheal pole due to the rapid falloff in energy at low frequencies.

[14] Detailed documentation of the characteristics of this source will not be given here since it was not used in the experiments to be described later in Secs. III and IV. The characteristics of the LF source are described in documentation available in the Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology.

[15] The fundamental frequency extraction algorithm that was employed to produce this contour is of the harmonic sieve type, which probably averages out some rapid period-to-period changes within the 25-ms analysis window rather than accentuating them.

[16] In an analogous fashion, one can use the newly defined tracheal pole-zero pair to simulate any observed second nasal resonance in a nasalized vowel.

[17] The f_0 is specified in tenths of a hertz in order to minimize perceptual "staircase" effects for slowly changing fundamental frequency contours. The period that is computed is quantized into quarter-of-a-sample increments, again to avoid "staircase" effects; this is accomplished by running the glottal source code at four times the regular sampling rate of 10 000 samples per second, and then low-pass filtering and downsampling the resulting voicing waveform.

[18] For these utterances, the update interval was set to 10 ms, but internal to the synthesizer, AV is reset only at the beginning of each pitch period so as to avoid waveform discontinuities.

[19] Information about the availability of this software can be obtained from Kenneth Stevens, Room 36-517, Massachusetts Institute of Technology, Cambridge, MA 02139.

Abercrombie, D. (1967). *Elements of General Phonetics* (Edinburgh U.P., Edinburgh).

Allen, D. R., and Strong, W. J. (1985). "A Model for the Synthesis of Natural Sounding Vowels," J. Acoust. Soc. Am. **78**, 58–69.

Ananthapadmanabha, T. V. (1984). "Acoustic Analysis of Voice Source Dynamics," Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 1–24.

Ananthapadmanabha, T. V., and Fant, G. (1982). "Calculation of True Glottal Flow and Its Components," Speech Commun. **1**, 167–184.

Askenfelt, A., and Hammarberg, B. (1981). "Speech Waveform Perturbation Analysis Revisited," Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 49–68.

Askenfelt, A., and Hammarberg, B. (1986). "Speech Waveform Perturbation Analysis: A Perceptual-Acoustical Comparison of Seven Measures," J. Speech Hear. Res. **29**, 50–64.

Baer, T. (1978). "Effect of Single-Motor-Unit Firings on Fundamental Frequency of Phonation," J. Acoust. Soc. Am. Suppl. 1, **64**, S90.

Berg, J. W. van den (1960). "An Electrical Analog of the Trachea, Lungs and Tissues," Acta Physiol. Parmacol. Neerandica **9**, 1–24.

Bickley, C. (1982). "Acoustic Analysis and Perception of Breathy Vowels," Speech Commun. Group Work. Papers I, Research Laboratory of Electronics, MIT, Cambridge, MA, 71–82.

Bickley, C., and Stevens, K. N. (1986). "Effect of a Vocal Tract Constriction on the Glottal Source: Experimental and Modeling Studies," J. Phon. **14**, 373–382.

Bladon, R. A. W. (1982). "Arguments against Formants in the Auditory Representation of Speech," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granstrom (Elsevier Biomedical, Amsterdam), pp. 95–102.

Bless, D. M., Biever, D., and Shaikh, A. (1986). "Comparisons of Vibratory Characteristics of Young Adult Males and Females," Proceedings of International Conference on Voice, Kurume, Japan, Vol. 2, 46–54.

Brend, R. M. (1975). "Male–Female Intonation Patterns in American English," in *Language and Sex: Difference and Dominance*, edited by B. Thorn and N. Henley (Newbury House, Rowley, MA), pp. 84–87.

Catford, J. C. (1964). "Phonation Types: The Classification of Some Laryngeal Components of Speech Production," in *In Honour of Daniel Jones*, edited by D. Abercrombie, D. B. Fry, P. A. D. McCarthy, N. C. Scott, and J. L. M. Trim (Longmans, London), pp. 26–37.

Catford, J. C. (1977). *Fundamental Problems in Phonetics* (Indiana U.P., Bloomington, IN).

Chapin-Ringo, C. (1988). "Enhanced Amplitude of the First Harmonic as a Correlate of Voicelessness in Aspirated Consonants," J. Acoust. Soc. Am. Suppl. 1 **83**, S71.

Chasaide, A. (1987). "Glottal Control of Aspiration and of Voicelessness," Proceedings of Eleventh International Congress of Phonetic Sciences, Tallinn, Estonia, Vol. 6, 28–31.

Chasaide, A., and Gobl, C. (1987). "Cross Language Study of the Effects of Voiced/Voiceless Consonants on the Vowel Voice Source Characteristics," J. Acoust. Soc. Am. Suppl. 1 **82**, S116.

Cleveland, T., and Sundberg, J. (1983). "Acoustic Analysis of Three Male Voices of Different Quality," Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 27–38.

Cooper, W. E., and Sorenson, J. (1981). *Fundamental Frequency in Sentence Production* (Springer, New York).

Cranen, B., and Boves, L. (1987). "On Subglottal Formant Analysis," J. Acoust. Soc. Am. **81**, 734–746.

Dixit, R. P. (1987). "In Defense of the Phonetic Adequacy of the Traditional Term 'Voiced Aspirated,' " Proceedings of the Eleventh International Congress of Phonetic Sciences, Tallinn, Estonia, Vol. 2, 145–148.

Dolansky, L., and Tjernlund, P. (1968). "On Certain Irregularities of Voiced Speech Waveforms," IEEE Trans. Audio Electroacoust. **AU-16**, 51–56.

Duifhuis, H., Willems, L. F., and Sluyter, R. J. (1982). "Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception," J. Acoust. Soc. Am. **71**, 1568–1580.

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

Fant, G. (1975). "Non-Uniform Vowel Normalization," Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 1–19.

Fant, G. (1979). "Glottal Source and Excitation Analysis," Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 85–107.

Fant, G. (1980). "Voice Source Dynamics," Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 17–37.

Fant, G. (1982a). "Preliminaries to Analysis of the Human Voice Source," Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 1–25.

Fant, G. (1982b). "The Voice Source: Acoustic Modeling," Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 28–48.

Fant, G. (1985). "The Voice Source: Theory and Acoustic Modeling," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, edited by I. R. Titze and R. C. Scherer (The Denver Center for the Performing Arts, Denver, CO), pp. 453–464.

Fant, G. (1986). "Glottal Flow: Models and Interaction," J. Phon. **14**, 393–400.

Fant, G., and Ananthapadmanabha, T. V. (1982). "Truncation and Superposition," Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 1–17.

Fant, G., Ishizaka, K., Lindqvist, J., and Sundberg, J. (1972). "Subglottal Formants," Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 85–107.

Fant, G., Liljencrants, J., and Lin, Q. G. (1985). "A Four-Parameter Model of Glottal Flow," Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 1–13.

Fant, G., and Lin, Q. G. (1987). "Glottal Source—Vocal Tract Acoustic Interaction," Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 13–27.

Fant, G., Lin, Q. G., and Gobl, C. (1985). "Notes on Glottal Flow Interaction," Speech Trans. Lab. Q. Prog. Stat. Rep. **2**, Royal Institute of Technology, Stockholm, 18–24.

Fant, G., and Mártony, J. (1963). "Speech Analysis," Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 1–5.

Farnsworth, D. W. (1940). "High Speed Motion Pictures of the Human Vocal Cords," Bell Lab. Rec. **18**, 203–208.

Fischer-Jorgensen, E. (1967). "Phonetic Analysis of Breathy (Murmured) Vowels in Gujarati," Indian Linguistics **28**, 71–139.

Flanagan, J. L. (1958). "Some Properties of the Glottal Sound Source," J. Speech Hear. Res. 1, 99–116.
Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception* (Springer, New York).
Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1975). "Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract," Bell Sys. Tech. J. 54, 485–506.
Fourcin, A. J. (1981). "Laryngographic Assessment of Phonatory Function," in *Proceedings of the Conference on the Assessment of Vocal Pathology, ASHA Reports 11*, edited by C. L. Ludlow and M. O. Hart (American Speech and Hearing Association, Rockville, MD); reprinted in J. Phon. 14, 435–442.
Fujimura, O. (1968). "Approximation to Voice Aperiodicity," IEEE Trans. Audio Electroacoust. AU-16, 68–73.
Fujisaki, H., and Ljungqvist, M. (1986). "Proposal and Evaluation of Models for the Glottal Source Waveform," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-86, 1605–1608.
Goldstein, U. (1980). "An Articulatory Model for the Vocal Tracts of Growing Children," unpublished Sc.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Gunzburger, D. (1987). "Duality in Vocal Gender Roles," Prog. Rep. Institute Phon., Utrecht 12, (2), 1–10.
Halle, M. and Stevens, K. N. (1971). "A Note on Laryngeal Features," Res. Lab. Electron. Q. Prog. Rep. 101, MIT, Cambridge, MA, 198–213.
Hawkins, S., and Stevens, K. N. (1985). "Acoustic and Perceptual Correlates of the Non-Nasal/Nasal Distinction for Vowels," J. Acoust. Soc. Am. 77, 1560–1575.
Henton, C. G., and Bladon, R. A. W. (1985). "Breathiness in Normal Female Speech: Inefficiency Versus Desirability," Lang. Commun. 5, 221–227.
Henton, C. G., and Bladon, R. A. W. (1987). "Creak as a Sociophonetic Marker," in *Language, Speech and Mind: Studies in Honor of Victoria Fromkin*, edited by L. Hyman and C. N. Li (Routledge, London), pp. 3–29.
Hollien, H. (1974). "On Vocal Registers," J. Phon. 2, 125–143.
Hollien, H., Michel, J., and Doherty, E. T. (1973). "A Method for Analyzing Vocal Jitter in Sustained Phonation," J. Phon. 1, 85–91.
Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1988). "Glottal Air Flow and Pressure Measurements for Soft, Normal and Loud Voice by Male and Female Speakers," J. Acoust. Soc. Am. 84, 511–529.
Holmes, J. N. (1961). "Research on Speech Synthesis," Joint Speech Research Unit Report JU 11–4, British Post Office, Eastcote, England.
Holmes, J. N. (1973). "Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," IEEE Trans. Audio Electroacoust. AU-21, 298–305.
Horii, Y. (1979). "Fundamental Frequency Perturbation Observed in Sustained Phonation," J. Speech Hear. Res. 22, 5–19.
Horii, Y. (1980). "Vocal Shimmer in Sustained Phonation," J. Speech Hear. Res. 23, 202–209.
Huffman, M. K. (1987). "Measures of Phonation Type in Hmong," J. Acoust. Soc. Am. 81, 495–504.
Hunt, M. J. (1987). "Studies of Glottal Excitation Using Inverse Filtering and an Electroglottograph," Proceedings of Eleventh International Congress of Phonetic Sciences, Tallinn, Estonia, Vol. 3, 23–26.
Ishizaka, K., and Matsudaira, M. (1968). "What Makes the Vocal Cords Vibrate?," in *The Sixth International Congress on Acoustics, Vol. II*, edited by Y. Kohasi (Elsevier, New York), B9–B12.
Ishizaka, K., Matsudaira, M., and Kaneko, T. (1976). "Input Acoustic Impedance Measurements of the Subglottal System," J. Acoust. Soc. Am. 60, 190–197.
Javkin, H. R., and Maddieson, I. (1983). "An Inverse Filtering Study of Burmese Creaky Voice," Work. Papers Phon. 57, U. California at Los Angeles, 115–125.
Kahn, M. (1975). "Arabic Emphatics: The Evidence for Cultural Determinants of Phonetic Sex-Typing," Phonetica 31, 38–50.
Karlsson, I. (1985). "Glottal Waveforms for Normal Female Speakers," Speech Trans. Lab. Q. Prog. Stat. Rep. 1, Royal Institute of Technology, Stockholm, 31–36.
Karlsson, I. (1987). "Sex Differentiation Cues in the Voices of Young Children of Different Language Backgrounds," J. Acoust. Soc. Am. Suppl. 1 81, S68.
Kasuya, H., and Ogawa, S. (1986). "Normalized Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice," J. Acoust. Soc. Am. 80, 1329–1334.
Kato, Y., Ochiai, K., Fujimura, O., and Maeda, S. (1967). "A Vocoder Excitation with Dynamically Controlled Voicedness," 1967 Conference on Speech Communication and Processing, Cambridge, MA, 288–291.
Kirk, P., Ladefoged, P., and Ladefoged, J. (1984). "The Linguistic Use of Different Phonation Types," Work. Papers Phon. 59, U. California at Los Angeles, 102–113.
Klatt, D. H. (1980). "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am. 67, 971–995.
Klatt, D. H. (1982). "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step," Proceedings of the International Conference of Acoustics on Speech and Signal Processing, ICASSP-82, 1278–1281.
Klatt, D. H. (1984). "The New MIT Speechvax Computer Facility," Speech Communication Group Working Papers IV, Research Laboratory of Electronics, MIT, Cambridge, MA, 73–82.
Klatt, D. H. (1986a). "Representation of the First Formant in Speech Recognition and in Models of the Auditory Periphery," in *Proceedings of Montreal Symposium on Speech Recognition*, edited by P. Mermelstein (McGill University, Montreal, 1986), pp. 5–7.
Klatt, D. H. (1986b). "Detailed Spectral Analysis of a Female Voice," J. Acoust. Soc. Am. Suppl. 1 81, S80.
Klatt, D. H. (1987a). "Acoustic Correlates of Breathiness: First Harmonic Amplitude, Turbulence Noise, and Tracheal Coupling," J. Acoust. Soc. Am. Suppl. 1, 82, S91.
Klatt, D. H. (1987b). "Review of Text-to-Speech Conversion for English," J. Acoust. Soc. Am. 82, 737–793.
Klatt, D. H. (1989). "Review of Selected Models of Speech Perception," in *Lexical Representation and Process*, edited by W. Marslen-Wilson (MIT, Cambridge, MA), pp. 169–226.
Klatt, D. H., and Stevens, K. N. (1969). "Pharyngeal Consonants," Res. Lab. of Electron. Q. Prog. Rep. 93, MIT, Cambridge, MA, 207–216.
Koopmans-van Beinum, F. J. (1980). "Vowel Contrast Reduction: An Acoustic and Perceptual Study of Dutch Vowels in Various Speech Conditions," Ph.D. dissertation, Academic, Amsterdam.
Ladefoged, P. (1973). "The Features of the Larynx," J. Phonetics 1, 73–83.
Ladefoged, P. (1983). "The Linguistic Use of Different Phonation Types," in *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, edited by D. Bless and J. Abbs (College Hill, San Diego), pp. 351–360.
Ladefoged, P., and Antoñanzas-Barroso, N. (1985). "Computer Measures of Breathy Phonation," Work. Papers Phon. 61, U. California at Los Angeles, 79–86.
Laver, J. (1980). "*The Phonetic Description of Voice Quality* (Cambridge U.P., Cambridge).
Liberman, A. M., Cooper, F. S., Shankweiler, D. S., and Studdert-Kennedy, M. (1967). "Perception of the Speech Code," Psychol. Rev. 74, 431–461.
Lieberman, P. (1961). "Perturbation in Vocal Pitch," J. Acoust. Soc. Am. 33, 597–603.
Lieberman, P. (1963). "Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathologic Larynges," J. Acoust. Soc. Am. 35, 344–353.
Lieberman, P. (1967). *Intonation, Perception and Language* (MIT, Cambridge, MA).
Makhoul, J., Vishwanathan, R., Schwartz, R., and Huggins, A. W. F. (1978). "A Mixed-Source Model for Speech Compression and Synthesis," J. Acoust. Soc. Am. 64, 1577–1581.
Margulies, M. K. (1979). "Male–Female Differences in Speaker Intelligibility: Normal versus Hearing Impaired Listeners," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 363–366.
Meditch, A. (1975). "The Development of Sex-Specific Speech Patterns in Young Children," Anthropol. Linguistics 17, 421–465.
Milenkovic, P. (1987). "Least Mean Square Measures of Voice Perturbation," J. Speech Hear. Res. 30, 529–538.
Monsen, R. B., and Engebretson, A. M. (1977). "Study of Variations in the Male and Females Glottal Wave," J. Acoust. Soc. Am. 62, 981–993.
Nord, L., Ananthapadmanabha, T. V., and Fant, G. (1986). "Signal Analysis and Perceptual Tests of Vowel Responses with an Interactive Source-Filter Model," J. Phon. 14, 401–404.
Pandit, P. B. (1957). "Nasalization, Aspiration and Murmur in Gujarati," Indian Linguistics 17, 165–172.
Peterson, G. E., and Barney, H. L. (1952). "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. 24, 175–184.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985). "Speaking Clearly for the Hearing Impaired I: Intelligibility Differences between Clear

and Conversational Speech," J. Speech Hear. Res. **28**, 96–103.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking Clearly for the Hearing Impaired II: Acoustic Characteristics of Clear and Conversational Speech," J. Speech Hear. Res. **29**, 434–449.
Pollack, I. (**1971**). "Amplitude and Time Jitter Thresholds for Rectangular Wave Trains," J. Acoust. Soc. Am. **50**, 1133–1142.
Robinson, D. W., and Dadson, M. A. (**1956**). "A Redetermination of Equal-Loudness Relations for Pure Tones," Br. J. Appl. Phys. **7**, 166–181.
Rosenberg, A. (**1968**). "Effect of Pitch Averaging on the Quality of Natural Vowels," J. Acoust. Soc. Am. **44**, 1592–1595.
Rosenberg, A. (**1971**). "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," J. Acoust. Soc. Am. **49**, 583–590.
Rothenberg, M. (**1973**). "A New Inverse Filtering Technique for Deriving the Glottal Air Flow Waveform during Voicing," J. Acoust. Soc. Am. **53**, 1632–1645.
Rothenberg, M. (**1974**). "Glottal Noise During Speech," Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 1–10.
Rothenberg, M. (**1985**). "Source-Tract Acoustic Interactions in Breathy Voice," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, edited by I. R. Titze and R. C. Scherer (Denver Center for the Performing Arts, Denver, CO), pp. 155–165.
Rothenberg, M., Carlson, R., Granstrom, B., and Lindqvist-Gauffin, J. (**1975**). "A Three-Parameter Voice Source for Speech Synthesis," in *Speech Communication*, edited by G. Fant (Almqvist and Wiksell, Uppsala, Sweden), Vol. 2, pp. 235–243.
Rozsypal, A. J., and Millar, B. F. (**1979**). "Perception of Jitter and Shimmer in Synthetic Vowels," J. Phon. **7**, 343–355.
Ryalls, J., and Lieberman, P. (**1982**). "Fundamental Frequency and Vowel Perception," J. Acoust. Soc. Am. **72**, 1631–1634.
Sachs, J., Lieberman, P., and Erickson, D. (**1973**). "Anatomical and Cultural Determinants of Male and Female Speech," in *Language Attitudes: Current Trends and Prospects*, edited by R. W. Shuy and R. W. Fasold (Georgetown U.P., Washington, DC).
Shadle, C. (**1987**). "The Acoustics of Fricative Consonants," Ph.D. thesis, MIT, Cambridge, MA.
Sondhi, M. M. (**1975**). "Measurement of the Glottal Waveform," J. Acoust. Soc. Am. **57**, 228–232.
Sonesson, B. (**1960**). "On the Anatomy and Vibratory Pattern of Human Vocal Folds," Acta. Oto-Laryngol. Suppl. 156.
Stevens, K. N. (**1971**). "Airflow and Turbulence Noise for Fricative and Stop Consonants," J. Acoust. Soc. Am. **50**, 1180–1192.
Stevens, K. N. (**1977**). "Physics of Larynx Behavior and Larynx Modes," Phonetica **34**, 264–279.
Stevens, K. N. (**1981**). "Vibration Modes in Relation to Model Parameters," in *Vocal-Fold Physiology*, edited by K. N. Stevens and M. Hirano (University of Tokyo, Tokyo), pp. 291–301.
Stevens, K. N., and Klatt, D. H. (**1974**). "The Role of Formant Transitions in the Voiced–Voiceless Distinction for Stops," J. Acoust. Soc. Am. **55**, 653–659.
Sundberg, J., and Gauffin, J. (**1979**). "Waveform and Spectrum of the Glottal Voice Source," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, New York), pp. 301–322.
Thorne, B., Kramearae, C., and Henley, B. (Eds.) (**1983**). *Language, Gender and Society* (Newbury House, Rowley, MA).
Timke, R., von Leden, H., and Moore, P. (**1959**). "Laryngeal Vibrations: Measurements of the Glottic Wave. Part II: Physiological Considerations," A. M. A. Arch. Otolaryng. **69**, 438–444.
Titze, I. R. (**1974**). "The Human Vocal Cords: A Mathematical Model," Phonetica **29**, 1–21.
Titze, I. R. (**1984**). "Parametrization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Areas," J. Acoust. Soc. Am. **75**, 570–580.
Titze, I. R., and Talkin, D. (**1979**). "A Theoretical Study of the Effects of the Various Laryngeal Configurations on the Acoustics of Phonation," J. Acoust. Soc. Am. **66**, 60–74.
Ward, P. H., Sanders, J. W., Goldman, R., and Moore, G. P. (**1969**). "Diplophonia," Ann. Otol., Rhinol., and Laryngol. **78**, 771–777.
Yumoto, E., Gould, W. J., and Baer, T. (**1982**). "Harmonics-to-Noise Ratio as an Index of the Degree of Hoarseness," J. Acoust. Soc. Am. **71**, 1544–1550.