

# Vocal quality factors: Analysis, synthesis, and perception

D. G. Childers

*Department of Electrical Engineering, University of Florida, Gainesville, Florida 32611-2024*

C. K. Lee

*Department of Electrical Engineering, Tatung Institute of Technology, Taipei, Taiwan, Republic of China*

(Received 28 September 1990; accepted for publication 25 July 1991)

The purpose of this study was to examine several factors of vocal quality that might be affected by changes in vocal fold vibratory patterns. Four voice types were examined: modal, vocal fry, falsetto, and breathy. Three categories of analysis techniques were developed to extract source-related features from speech and electroglottographic (EGG) signals. Four factors were found to be important for characterizing the glottal excitations for the four voice types: the glottal pulse width, the glottal pulse skewness, the abruptness of glottal closure, and the turbulent noise component. The significance of these factors for voice synthesis was studied and a new voice source model that accounted for certain physiological aspects of vocal fold motion was developed and tested using speech synthesis. Perceptual listening tests were conducted to evaluate the auditory effects of the source model parameters upon synthesized speech. The effects of the spectral slope of the source excitation, the shape of the glottal excitation pulse, and the characteristics of the turbulent noise source were considered. Applications for these research results include synthesis of natural sounding speech, synthesis and modeling of vocal disorders, and the development of speaker independent (or adaptive) speech recognition systems.

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

## INTRODUCTION

The quality of voice may be referred to as the total auditory impression the listener experiences upon hearing the speech of another talker. There is no generally accepted definition of vocal quality and the term has been used in different contexts, e.g., a phonetician might use quality in the context of articulatory differences; a singer might refer to the quality of vocal registers, which are related to vocal fold vibration; and quality might be used to describe voice types such as breathy, hoarse, or harsh. In this study, we investigated aspects of vocal quality related to vocal fold vibratory patterns, i.e., laryngeal vocal quality. We did not address two other common characteristics of vocal quality, namely, loudness, and resonance.

Our purpose was to improve existing or develop new speech and electroglottographic (EGG) analysis techniques to assist the assessment of vocal quality and to design new voice source models for synthesizing natural sounding speech with a selectable vocal quality. We investigated the nature and extent of the glottal excitation variations of four voice types, namely breathy voice and the three registers, modal, vocal fry, and falsetto. Our analysis of results led to the development of a new source model for speech synthesis that contains some factors that appear perceptually important for characterizing vocal quality. The knowledge gained from this research may improve the development of natural sounding speech synthesizers and assist the advancement of speaker-independent speech recognition systems.

Laver and Hanson (1981) have defined six major types of phonation, modal voice, vocal fry, falsetto, breathy voice,

harshness, and whisper. We eliminated harshness and whisper from our study because harshness involved large variations from cycle-to-cycle in the vocal fold vibratory patterns (Coleman, 1960; Moore, 1975; Wendahl, 1963; Wendahl, 1966) and whisper is characterized by little or no vocal fold vibratory motion. To help us achieve our goal, we reviewed vocal quality with respect to its physiological, perceptual, and acoustical aspects for the four voice types of modal, vocal fry, falsetto, and breathy.

## A. Physiological characteristics

Vocal fold length and thickness have been shown to affect the three voice registers of modal, vocal fry, and falsetto. For modal voice vocal fold length may be considered to be medium with the length increasing as fundamental frequency ( $F_0$ ) increases (Damste *et al.*, 1968; Hollien, 1974). For vocal fry and falsetto the vocal fold length is short and long, respectively (Boone, 1971; Hollien, 1974; Ladefoged, 1975). Vocal fold thickness varies with  $F_0$ , being medium for modal voice, while the vocal folds become thick for vocal fry and thin for falsetto (Hollien *et al.*, 1968; Hollien and Colton, 1969; Boone, 1971; Allen and Hollien, 1973; Hollien, 1974).

Some characteristics of the vocal fold vibratory pattern are also known. For modal voice the vocal folds have a speed quotient greater than one with the glottis having a long opening phase and a rapid closing phase (Hollien, 1974; Monsen and Engebretson, 1977). Vocal fry is characterized by a glottal area function that has sharp, short pulses followed by a long closed glottal interval. The glottal opening phase may

have one, two, or three opening/closing pulses (Moore and von Leden, 1958; Timcke *et al.*, 1959; Hollien, 1974; Monsen and Engebretson, 1977; Hollien *et al.*, 1977; Whitehead *et al.*, 1984). Falsetto voice has gradual glottal opening and closing phases with a short or no closed phase (Hollien, 1974; Monsen and Engebretson, 1977; Kitzing, 1982). Breathy voice may have a slight vibratory excursion of the vocal folds with incomplete glottal closure (Boone, 1971; Hollien, 1974; Ladefoged, 1975).

## B. Perceptual characteristics

Generally, the pitch for modal, vocal fry, falsetto, and breathy voice is characterized as medium, low, high, and wide ranging, respectively (Hollien and Michel, 1968; Colton, 1969; Boone, 1971; Colton and Hollien, 1973; Hollien, 1974; Laver, 1980). Loudness for modal voice can vary over a wide range while vocal fry, falsetto, and breathy voice are typically soft (Boone, 1971; Hollien, 1974; Laver, 1980). The quality of these four voice types has been described as normal (modal), a low pitch, rough-sounding voice (vocal fry), a flutelike tone that is sometimes breathy (falsetto), and a friction noiselike sound (breathy) (Wendahl *et al.*, 1963; Boone, 1971; Colton and Hollien, 1973; Hollien, 1974; Laver, 1980).

## C. Acoustical characteristics

The  $F_0$  of modal voices ranges over 94–287 Hz for males and 144 to 538 Hz for females, while the  $F_0$  range for vocal fry is 24–52 Hz for males and 18–46 Hz for females; the falsetto  $F_0$  range for males is 275–634 Hz and 495–1131 Hz for females, while breathy is described as wide ranging (Hollien and Michel, 1968; Colton, 1969; Boone, 1971; Hollien, 1974). Vocal intensity characteristics are less specific and have been described as wide ranging, low, low to medium, and low for modal, vocal fry, falsetto and breathy, respectively (Boone, 1971; Colton, 1973a; Hollien, 1974; Laver, 1980). Likewise, the source spectral slope (tilt) for these same voice types has been described as medium but  $F_0$  dependent, relatively flat, steep (–20 dB/oct), and steep (Colton, 1973b; Hollien, 1974; Monsen and Engebretson, 1977; Hurme and Sonninen, 1985). Turbulent noise has been associated with only modal (low) and breathy (high) voices (Yanagihara, 1967; Isshiki *et al.*, 1978; Hiraoka *et al.*, 1984). Pitch perturbation is generally low for modal voice, high for vocal fry and breathy voices while amplitude perturbation is low for modal and high for breathy voices (Monsen and Engebretson, 1977; Deal and Emanuel, 1978; Davis,

1979; Horii, 1980; Heiberger and Horii, 1982; Hiller *et al.*, 1983; Hirano *et al.*, 1985; Askenfelt and Hammarburg, 1986; Kasuya *et al.*, 1986; Wolfe and Steinfatt, 1987).

Many more details are available on the four voice types we studied. We selected and summarized those features we thought would be analyzable from the speech and EGG signals and that might be useful in designing a source excitation model for synthesizing natural sounding speech and vocal disorders. Recently, several issues in synthesizing natural sounding voices have been considered, including the design of source excitation models (Fant, 1979; Fant *et al.*, 1985; Klatt, 1987; Lee and Childers, 1989; Klatt and Klatt, 1990; Childers and Wu, 1990; Childers and Wu, 1991; Wu and Childers, 1991). We will discuss these issues later when we present our source excitation model. This study considered two types of experiments: (1) voice quality factors that might be influenced by changes in vocal fold vibratory patterns and (2) the perceptual validation of the effects of acoustic parameter variations on the quality of the synthesis of a particular voice type.

## I. EXPERIMENTAL PROCEDURES

A summary of our research scheme appears in Fig. 1. The speech and electroglottographic data were digitized simultaneously using Digital Sound Corp. DSC-240 preamplifiers and a DSC-200 digitizer. We sampled each signal at 10 kHz with 16-bits precision. The microphone was an Electro-Voice RE-10 held six inches from the lips. The EGG device was a model from Synchrovoice, Inc. All data were collected in a professional IAC single wall sound booth. A Digital Equipment Corp. VAX 11/750 computer system managed the data collection.

The subjects for this study consisted of 23 (8 male, 15 female) patients with a vocal disorder or pathology and 52 (27 male, 25 female) subjects with a normal larynx. We denote the patients as  $P_n$ , where  $n$  goes from 1 to 23, while the normal subjects are denoted as  $N_n$  where  $n$  goes from 1 to 52. The subject's ages ranged from 20 to 80 years old. The complete speech protocol consisted of 27 tasks, including ten sustained vowels /i, I, ε, æ, a, ɔ, U, u, ʌ, ɜ/, two sustained diphthongs /ou, ei/, five sustained unvoiced fricatives /h, f, θ, s, ʃ/, and four sustained voiced fricatives /v, ð, z, ʒ/. The subjects were instructed to pronounce and sustain each vowel as it would be pronounced in the following words, respectively: bet, bit, bet, bat, Bob, bought, book, boot, but, Bert. Similar instructions were given for the diphthongs, for which the cue words were boat and bait, while for the fricatives we used the following cue words: hat, fix, thick, sat,

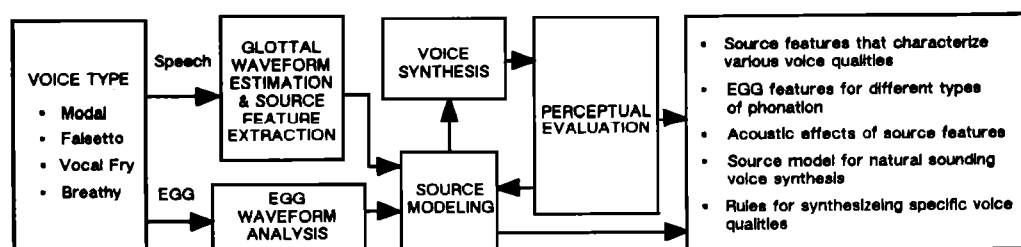


FIG. 1. Block diagram of basic research scheme.

ship, van, this, zoo, and azure. The duration of each vowel, diphthong, and fricative approximated 2 s. The additional tasks included counting from one to ten with a comfortable pitch and loudness, counting from one to five with a progressive increase in loudness, singing the musical scale using "la," and speaking three sentences. (We were away a year ago. Early one morning a man and a woman ambled along a one mile lane. Should we chase those cowboys?) This complete protocol has provided a data base for several studies on voice quality and speech synthesis, including this one. For this study, we analyzed only data for the two vowels /i/ and /a/, the three sentences, and selected data records from the counting task. The two vowels were selected because two recent studies had found acoustic correlates of vocal quality and vocal disorders using these two vowels (Prosek *et al.*, 1987; Eskenazi *et al.*, 1990). We analyzed the sentences and the counting task to give us some indication of the glottal factors in a word and sentence context.

## II. ANALYSIS FOR EXTRACTING SOURCE FEATURES

### A. Inverse filtering

Figure 2 illustrates that the peaks in the linear prediction (LP) error function occur nearly simultaneously with the negative peaks of the differentiated EGG (DEGG) signal, which correspond to the instants of glottal closure (Childers *et al.*, 1983; Childers and Krishnamurthy, 1985; Childers *et al.*, 1990). From these observations, we developed the "two-pass method" for accurate, automatic glottal inverse filtering, which works as well as the two-channel (speech and EGG) method (Krishnamurthy and Childers, 1986).

The two-pass method first identifies the locations of the main pulses of the LP error signal derived during the first pass of the inverse filtering procedure. These main pulses are then used as indicators of glottal closure, and a "pseudo-closed phase" is selected as the analysis interval for a pitch-synchronous covariance LP analysis to estimate the vocal tract filter, which in turn is used to obtain the desired glottal

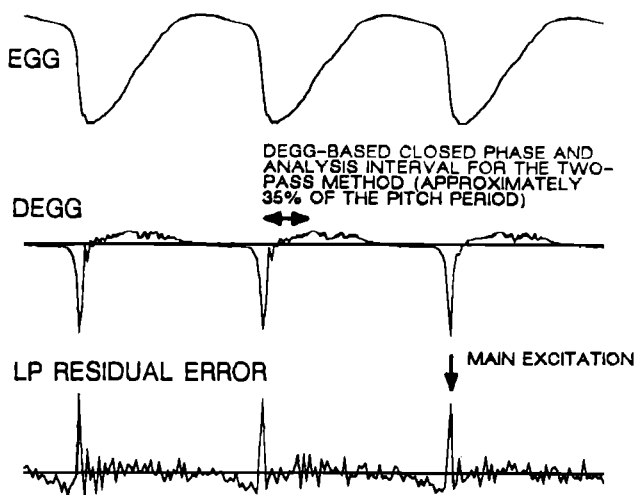


FIG. 2. Synchronized EGG, DEGG, and LP residual error.

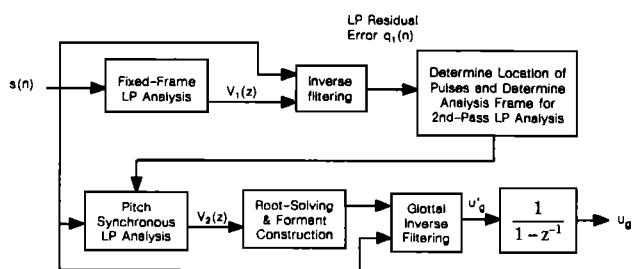


FIG. 3. Block diagram of the two-pass method for glottal inverse filtering analysis.

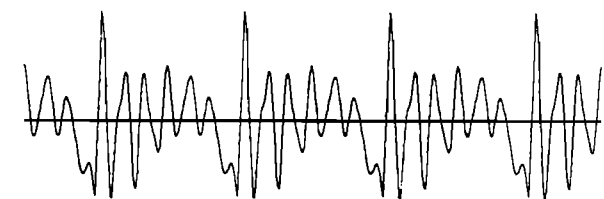
volume-velocity waveform. As with the EGG signal, the estimated LP pulses may not provide an exact indication for the instant of glottal closure. The key feature of the two-pass method is that it ensures the exclusion of the main pulse of the LP error signal from the analysis interval. This tailoring of the analysis interval increases the accuracy of the LP analysis. For a definition of the signal processing terms used in this paper, the reader is referred to Rabiner and Schafer (1978).

A block diagram of the two-pass method is shown in Fig. 3. During the first pass, a pitch-asynchronous (fixed frame) LP analysis is performed on the input speech signal  $s(n)$ . The estimated LP filter,  $V_1(z)$ , is used to derive the LP error signal,  $q_1(n)$ , by inverse filtering. For a voiced speech signal, the LP error function is characterized by a pulse train with the appropriate pitch period. The locations of these pulses are detected by a peak-picking method and are used as indicators of glottal closure. In the second pass of the procedure, a pitch-synchronous covariance LP analysis is used to estimate an improved LP filter,  $V_2(z)$ . For each pitch period, the criterion for determining the analysis interval is to pick the samples starting one point after the instant of the main pulse. The formant resonances of the vocal tract are estimated by solving the roots of the LP polynomial. The formant structure is then shaped by empirical rules, which include: (1) discarding the roots with center frequencies below 250 Hz, (2) discarding the roots with bandwidths greater than 500 Hz, and (3) merging two adjacent roots. The refined formant resonances are then used to construct the vocal tract transfer function, which is used in the final (second-pass) glottal inverse filtering procedure. The direct output of the glottal inverse filtering operation is a differential glottal volume-velocity  $u'_g(n)$  (i.e., the equivalent driving function to the vocal tract filter), which represents the combined effect of the lip radiation and the glottal volume-velocity. A glottal volume-velocity waveform,  $u_g(n)$ , is derived by integration. The validity of the two-pass method was verified by testing it with synthetic speech signals (Lee, 1988). The synthetic speech signals were produced by a cascade formant synthesizer (Klatt, 1980) excited by stylized glottal pulses generated by the LF (Liljencrants and Fant) model (Fant *et al.*, 1985). A typical result appears in Fig. 4.

### B. Source features

Using the inverse filtered waveforms (glottal pulses) as the excitation for voice types, modal, vocal fry, falsetto, and

## SPEECH WAVEFORM



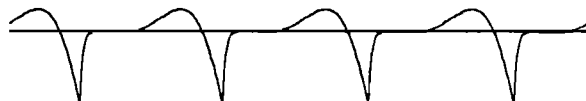
## LP RESIDUAL ERROR



## INTEGRAL OF ORIGINAL EXCITATION WAVEFORM



## ORIGINAL EXCITATION WAVEFORM



## INTEGRAL OF ESTIMATED EXCITATION WAVEFORM



## ESTIMATED EXCITATION WAVEFORM



FIG. 4. An example of two-pass glottal inverse filtering.

breathy, we measured the following features: (1) instant of maximum closing slope of the glottal pulse, (2) glottal pulse width, and (3) glottal pulse skewness (ratio of duration of glottal opening phase to duration of glottal closing phase). For modal and vocal fry phonations, the instant of the maximum closing slope occurs near the instant of glottal closure, resulting in an abrupt termination of the glottal airflow. For falsetto and breathy phonations, the instant of the maximum closing slope occurs near the middle of the glottal closing phase, followed by a residual phase of progressive closure.

The glottal pulse widths were moderate (65%–70% of the pitch period) for modal phonations and small (25%–45% of the pitch period) for vocal fry. Falsetto and breathy voices had large pulse widths (90%–100% of the pitch period), often making it appear there was no closed phase. Glottal pulse skewness also varied with voice type. The ranking of voice type according to decreasing skewness was vocal fry, modal, falsetto, and breathy.

## C. Glottal spectral characteristics

The literature (e.g., Mosen and Engbreton, 1977; Pinto *et al.*, 1989) suggested that the glottal waveforms of different voice types could be distinguished by two factors: (1) the general spectral slope (tilt or trend), and (2) the relationship between the intensity of the fundamental frequency and its harmonics. For normal phonations, Flanagan (1957) gave an average value of  $-12$  dB/octave for the slope of the glottal spectra. This led to the commonly accepted two-pole model for approximating the general spectral trend of a normal glottal volume flow:

$$U_g(z) = K / (1 - z_a z^{-1})(1 - z_b z^{-1}), \quad (1)$$

where  $K$  is a constant related to the amplitude of the glottal flow and  $z_a, z_b$  are real poles inside the unit circle. A point  $z_a$  is said to be a pole of a function  $U_g(z)$ , if  $U_g(z_a)$  equals infinity. The number of poles a function contains determines the slope of the spectrum of the function as the frequency increases. A single pole produces a spectral slope of  $-6$  dB/octave, two poles  $-12$  dB/octave, and so on. This discussion applies provided there are no zeros of the function, for which there are none in the all pole models we consider here.

Figure 5 shows the Fourier spectra and the corresponding two-pole approximations for the glottal pulses of various voice types. The results show that the two-pole model is appropriate for modal phonations and is reasonably good for vocal fry except for a small low-frequency interval (below the third harmonic). However, the two-pole model is not adequate for falsetto and breathy phonations, which have spectral roll-off rates higher than 12 dB/octave. For these latter two types of phonation, we found that a three-pole model provides a better fit to the data, e.g.,

$$U_g(z) = K / (1 - z_a z^{-1})(1 - z_b z^{-1})(1 - z_c z^{-1}), \quad (2)$$

where  $z_c$  is the third real pole inside the unit circle. For both the two- and three-pole models, we found that we could set  $z_a$  equal to unity and then calculate  $z_b$  and  $z_c$  from the preemphasized glottal waveform using linear prediction analysis of the waveform estimated by inverse filtering.

Table I lists typical values of  $z_b$  and  $z_c$  for the inverse filtered glottal waves for the four voice types for selected patients and normal subjects. The results confirmed that the general spectral slope (tilt) of a modal or vocal fry phonation can be modeled by two real poles. On the other hand, to simulate the spectral slope of a falsetto or breathy phonation, one extra real pole was required to account for its steeper roll-off rate. Figure 6 shows the improved spectral matching using the three-pole model for the falsetto and breathy phonations.

Our two- and three-pole models approximate the high-frequency spectral slope of a glottal pulse better than the low-frequency characteristics. An explanation for this is that the glottal pulse must be of finite duration. Therefore, an exact model would be a finite impulse response filter, and hence, would contain only zeros. Our results showed that the use of an all-pole model will cause mismatch of the model spectrum with the data at low frequencies (see Figs. 5 and 6). We reasoned that the low-frequency mismatch would not be as important perceptually as the high-frequency characteristics provided the harmonics in the speech signal were reasonably well matched by the model. This point was to be examined using a perceptual evaluation of synthesized tokens.

The spectral tilt or slope of the glottal pulse can be measured without estimating the waveshape of the glottal pulse.

The general spectral slope for a voice phonation is determined by the combined contribution of the spectra of the glottal pulse and the lip radiation. This spectral slope may often be approximated as a two-pole spectrum, i.e., two real poles inside the unit circle. These poles are estimated using LP analysis of preemphasized speech. The results obtained are consistent with those obtained using inverse filtered

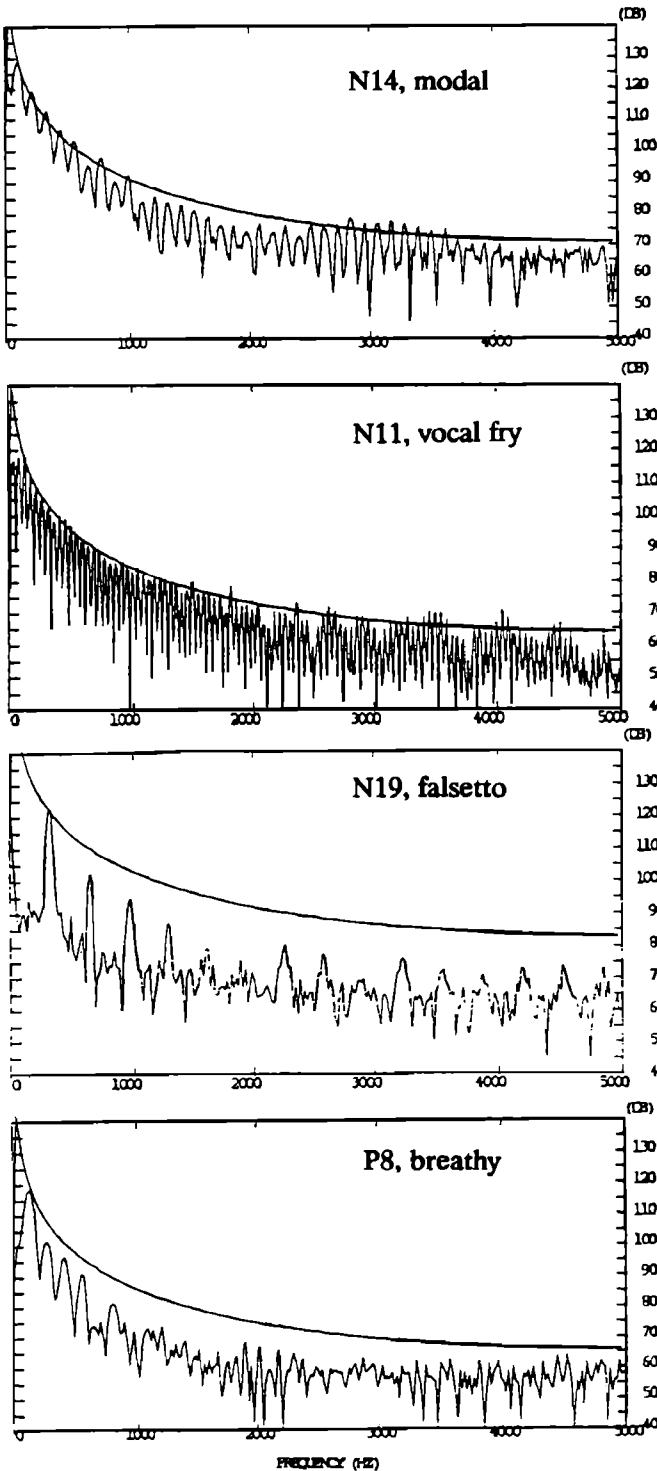


FIG. 5. Glottal source spectra and superimposed two-pole model approximation for modal, vocal fry, falsetto, and breathy voice types. The subjects are indicated in Table I.

TABLE I. The estimated real poles for the low-pass filter model of the glottal waveforms for a few typical male subjects (N = normal subject; P = disordered subject).

Subject	Voice type	$F_0$ (Hz)	$z_b$	$z_c$
N11	vocal fry /a/	46	0.83	...
N24	vocal fry /i/	45	0.81	...
N20	slight vocal fry /i/	90	0.83	...
N3	modal /a/	155	0.92	...
N14	modal /i/	106	0.92	...
N27	modal /i/	126	0.96	...
N7	falsetto /a/	210	1.00	0.60
N19	falsetto /a/	312	1.00	0.73
P8	breathy /i/	137	1.00	0.25
P13	breathy /i/	200	1.00	0.87

speech. For falsetto and breathy voices, one pole is on the unit circle (or nearly so) and the other varies roughly from 0.2 to 0.9. The pole values are, however, affected by the formants of the phonation. For example, the location of the poles within the unit circle for the two pole model varied depending on the type of vowel (/i/ or /a/) for a given speaker and a given voice type. The poles were usually closer to the origin for /i/ than for /a/.

#### D. Fundamental frequency and harmonics

The harmonics in the speech signal below the first formant are often considered important for the perception of

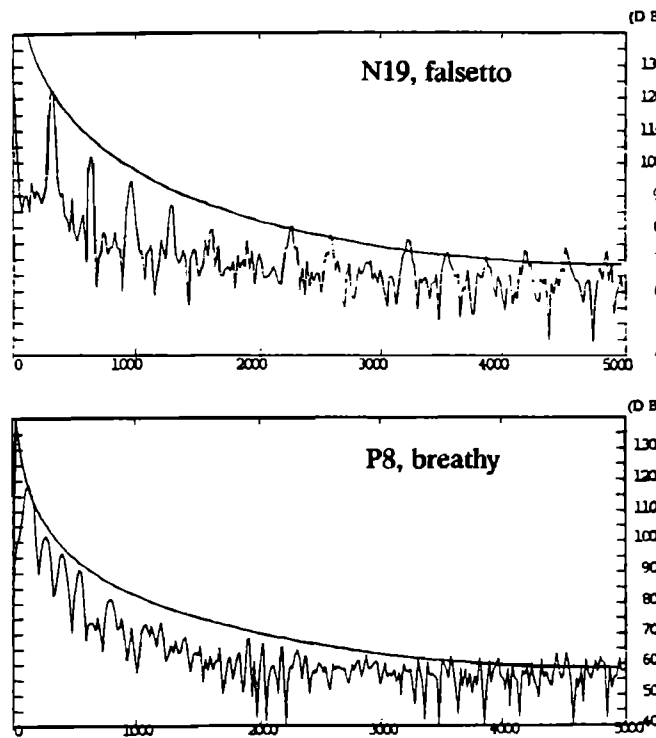


FIG. 6. Glottal source spectra and superimposed three-pole model approximation for falsetto and breathy voice types for the same corresponding subjects as in Fig. 5.

vocal quality (Holmes, 1973). This is presumably due to the high energy in these harmonics. We found that the glottal spectra of different voice types showed distinctive amplitude relationships between the fundamental and higher harmonics. Figure 7 shows the glottal spectra for various voice types, displaying only the first ten harmonics. There are distinctive amplitude relationships between the fundamental and the higher harmonics aside from the differences in the spectral slope as discussed above. We defined a parameter called the "harmonic richness factor" (HRF) to measure this relationship

$$\text{HRF} = \frac{\sum_{i=2}^{10} H_i}{H_1}, \quad (3)$$

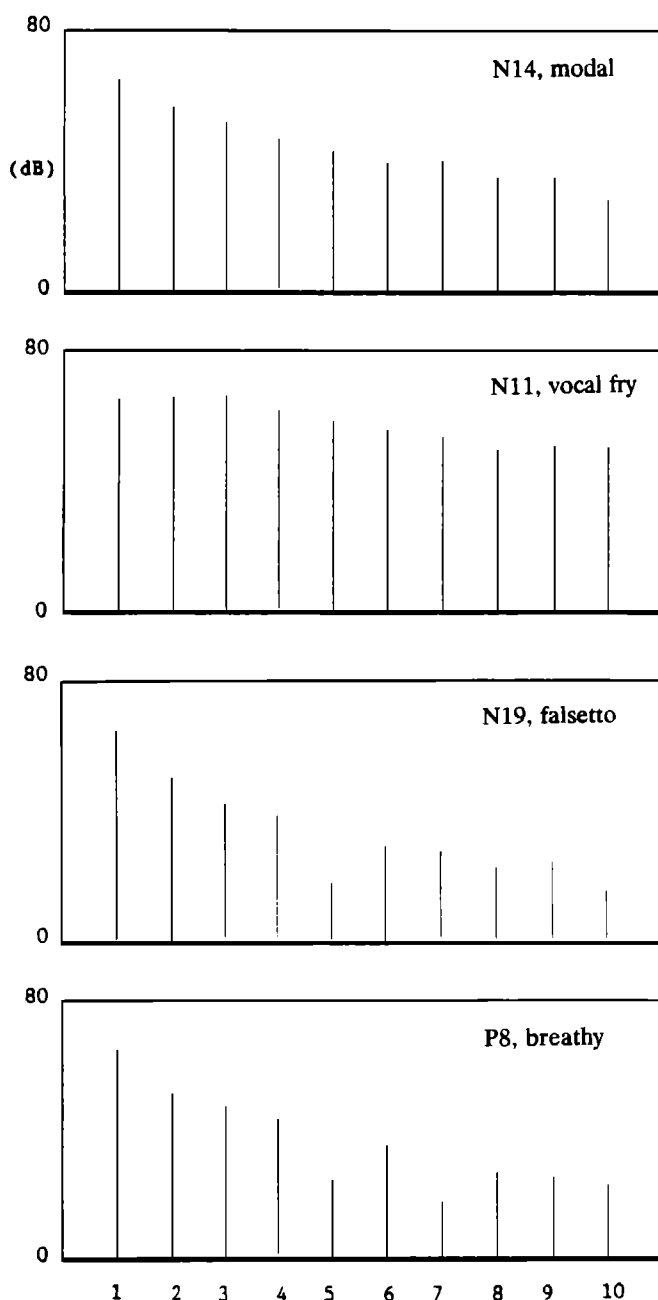


FIG. 7. The spectrum (dB) for the first ten harmonics of the glottal waveforms for modal, vocal fry, falsetto, and breathy voice types. All data are for vowel /i/ except falsetto which was /a/.

where  $H_i$  is the amplitude of the  $i$ th harmonic and  $H_1$  is the amplitude of the fundamental frequency. We found that vocal fry was approximately characterized by a high HRF (2.1 dB) followed by modal (−9.9 dB), breathy (−16.7 dB), and falsetto (−19.1 dB). Falsetto and breathy voices had a high intensity fundamental as well. These variations in harmonic relations appear to have little significance for speech intelligibility (phonetic identifiability) but affect the perceived vocal quality.

## E. Interharmonic noise

Turbulence at the level of the glottis has been noted to contribute to the perceptual quality of breathy voice (Isshiki *et al.*, 1978; Yumoto *et al.*, 1982; Hillman *et al.*, 1983; Hiraoka *et al.*, 1984). We modified the frequency-domain method of Hiraoka *et al.* (1984) to measure the noise-to-harmonic ratio (NHR). Our method uses an adaptive procedure to estimate  $F0$ , which is crucial for identifying the higher harmonics. This was accomplished by a two-step procedure. We estimate the pitch periods of the speech signal by using the synchronous EGG signal. The accuracy of this initial estimation is restricted by the signal sampling rate (in our case, 10 kHz). For example, if the accuracy of  $F0$  is within one sample point, an  $F0$  estimate of 100 sample points means an  $F0$  of  $100 \pm 1$  Hz and an  $F0$  estimate of 50 sample points mean an  $F0$  of  $200 \pm 4$  Hz. Such estimation errors can cause a severe problem in identifying high-order harmonics. For example, a 4-Hz error of  $F0$  could lead to a 40-Hz shift in locating the tenth harmonic.

An adaptive  $F0$  correction procedure was used to reduce the error in the initial  $F0$  estimate. We defined  $F0_n = F0 \pm n\Delta F$ , where  $n = 1, 2, \dots, 10$ , and  $\Delta F$  is one tenth of the maximum possible error of  $F0$ , as described above. Based on each  $F0_n$ , the energy of the harmonic components over the frequency range of 0 to 2 kHz are computed. The value  $F0_n$  giving the maximum harmonic energy is selected as the final estimate of  $F0$ . This criterion is based on our experimental results, which showed that over the low-frequency range of 0 to 2 kHz, the harmonic energy is considerably higher than the interharmonic noise energy.

Once  $F0$  is determined, the  $i$ th harmonic amplitude,  $H_i$ , and the interharmonic noise,  $N_i$ , are computed in the frequency region  $iF0 \pm F0/2$ , where  $H_i$  represents the energy in the subregion centered at  $iF0$  with a bandwidth of the Hamming analysis window. The symbol  $N_i$  represents the energy in the remaining frequency region (Fig. 8). And the noise-to-harmonic ratio at the  $i$ th harmonic region ( $\text{NHR}_i$ ) is defined as:

$$\text{NHR}_i = N_i/H_i. \quad (4)$$

The distribution of the interharmonic noise can be observed by plotting  $\text{NHR}_i$  along the frequency axis as in Fig. 9, where we see that a voice judged to be breathy has higher interharmonic noise above 2 kHz than modal or falsetto. Based on this observation, we defined a noise-to-harmonic ratio over a high-frequency range to be an indicator for the vocal quality of breathiness, namely,

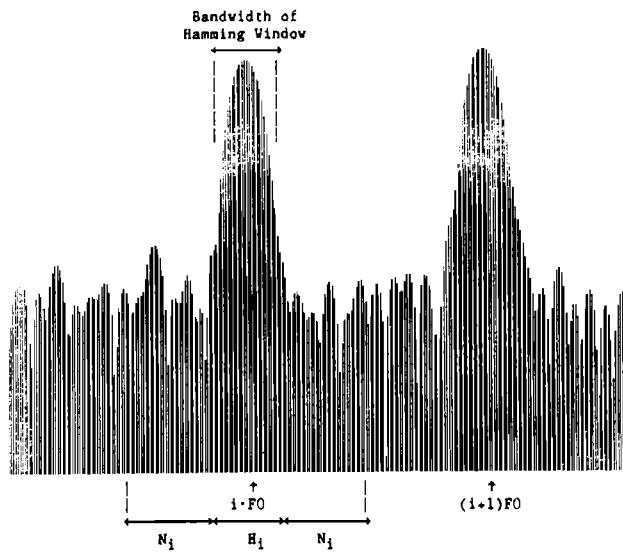


FIG. 8. Frequency intervals for harmonic and interharmonic noise components.

$$\text{NHR}_h = \frac{\sum N_i}{\sum H_i}, \quad (5)$$

where  $(N_i)$  and  $(H_i)$  are the noise and harmonic components above 2 kHz, respectively. The overall NHR, which was defined over the whole frequency range (0–5 kHz), was also computed for comparison. Table II lists the analysis of results for three voice types. The  $\text{NHR}_h$  (noise-to-harmonic ratio above 2 kHz) is a good indicator for the existence of the vocal quality of breathiness. We were unable to measure reliably  $\text{NHR}_h$  for vocal fry.

## F. Temporal energy distribution

The temporal energy distribution of a speech waveform is related to the glottal excitation and has long been thought to affect vocal quality. Wendahl *et al.* (1963) and Coleman (1963) established that the perception of vocal fry is related to the damping of the speech signal between glottal excitations. To measure the decay characteristics of a speech waveform during a single pitch period, a parameter called the waveform peak factor (WPF) was defined as

$$\text{WPF} = \frac{\text{peak amplitude}}{\text{rms value}} = \frac{\max(|x_i|)}{[(1/N)\sum_{i=1}^N x_i^2]^{1/2}}, \quad (6)$$

where  $x_i$  is the amplitude in the  $i$ th sample point and  $N$  is the total number of sample points in one pitch period. Theoretically, the WPF has a minimum value of 1 when the waveform is flat, and a maximum value  $N^{1/2}$  when the waveform is an impulse. The WPF value of a speech waveform is related to the underlying glottal waveshape. For glottal waves with narrow pulses separated by a long glottal closure, the WPF value is large and for pulses of long duration the WPF is near unity.

Although the average WPF values for sustained vowels vary somewhat with the type of vowel for the same subject and voice type, a general rule is that vocal fry, modal and falsetto registers are characterized by WPF values that are

high (4.0), medium (2.8), and low (1.8), respectively. We were unable to measure reliably the WPF for breathy voices. Our results imply that vocal fry register is characterized by a pulselike excitation waveform with a long glottal closure, while falsetto register is characterized by a short glottal closure.

## G. EGG waveform features

The electroglottographic signal (EGG) is generally believed to be representative of the amount of the lateral contact between the vocal folds (Childers and Krishnamurthy, 1985; Childers *et al.*, 1986; Childers *et al.*, 1990; Titze, 1990). The EGG is useful for registering glottal events during vocal fold vibration. When the EGG waveform is arranged such that an upward deflection reflects the opening of the glottis and a downward deflection depicts the closing of the glottis, the sharp negative peaks in the differentiated EGG (DEGG) waveform have been found to be very close to the instants of glottal closure (if closure exists), and the maxima of the DEGG are indications of the instant of glottal opening as in Fig. 10 (Childers *et al.*, 1990). In this study, we investigated the EGG waveform features for various types of phonation.

Figure 11 shows typical EGG and DEGG waveforms of modal, vocal fry, falsetto, and breathy phonations. All of the EGG waveforms exhibit a steeper slope (implying a rapid change in vocal fold contact area) during the glottal closing phase than during the opening phase. This characteristic phenomenon of vocal fold vibration results in a sharp negative pulse in the DEGG waveform at or near the instant of maximum glottal closure. Aside from this common feature,

TABLE II. Noise-to-harmonic ratios for several typical subjects and three voice types (N = normal subject; P = disordered subject).

Subject	Sex	Voice type	Hi-Freq. NHR	Overall NHR
N12	M	modal /i/	− 5.1 dB	− 13.8 dB
N13	M	modal /a/	− 4.2	− 16.5
N28	M	modal /i/	− 4.2	− 14.8
N1	M	modal /i/	− 5.4	− 14.8
N2	M	modal /a/	− 5.6	− 16.6
N30	M	modal /i/	− 4.8	− 16.9
N31	M	modal /a/	− 6.2	− 15.1
N32	M	modal /i/	− 7.1	− 14.9
N33	M	modal /a/	− 5.1	− 16.2
N16	M	falsetto /i/	− 8.5	− 20.4
N17	M	falsetto /o/	− 8.2	− 20.3
N18	M	falsetto /i/	− 9.9	− 23.2
N19	M	falsetto /a/	− 8.4	− 22.1
N4	M	falsetto /i/	− 4.3	− 21.7
N5	M	falsetto /a/	− 3.9	− 20.1
N6	M	falsetto /i/	− 6.1	− 22.1
N7	M	falsetto /a/	− 3.3	− 18.5
N25	M	breathy /i/	− 0.7	− 12.6
N29	M	breathy /i/	5.5	− 12.7
P7	M	breathy /i/	− 0.4	− 19.6
P12	M	breathy /i/	0.2	− 19.8
P20	F	breathy /i/	5.9	− 14.8
P21	F	breathy /i/	6.2	− 8.9

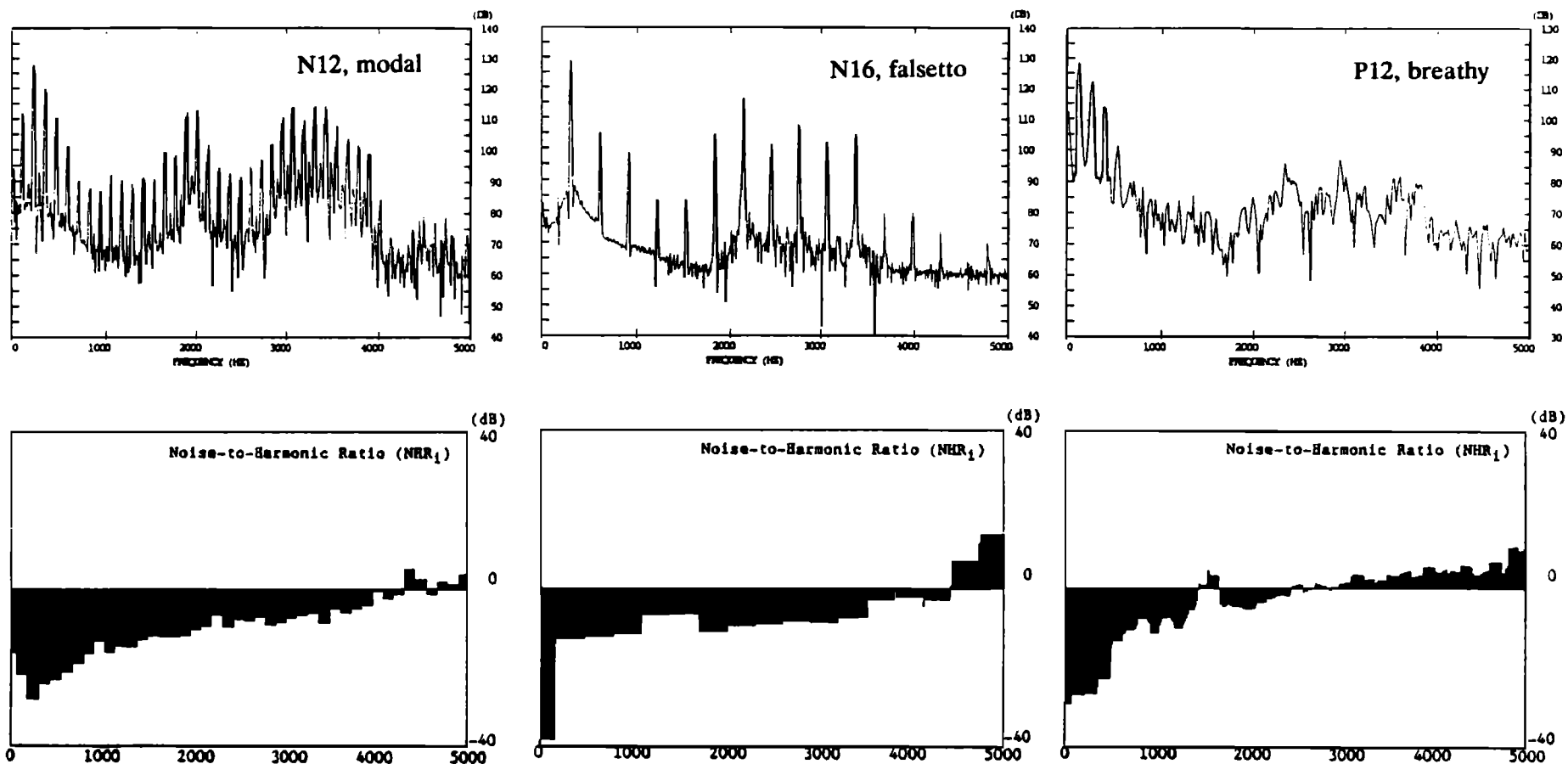


FIG. 9. The spectra and noise-to-harmonic ratio (NHR<sub>1</sub>) for three male subjects indicated in Table II for the vowel /i/ for modal, falsetto, and breathy voice types.



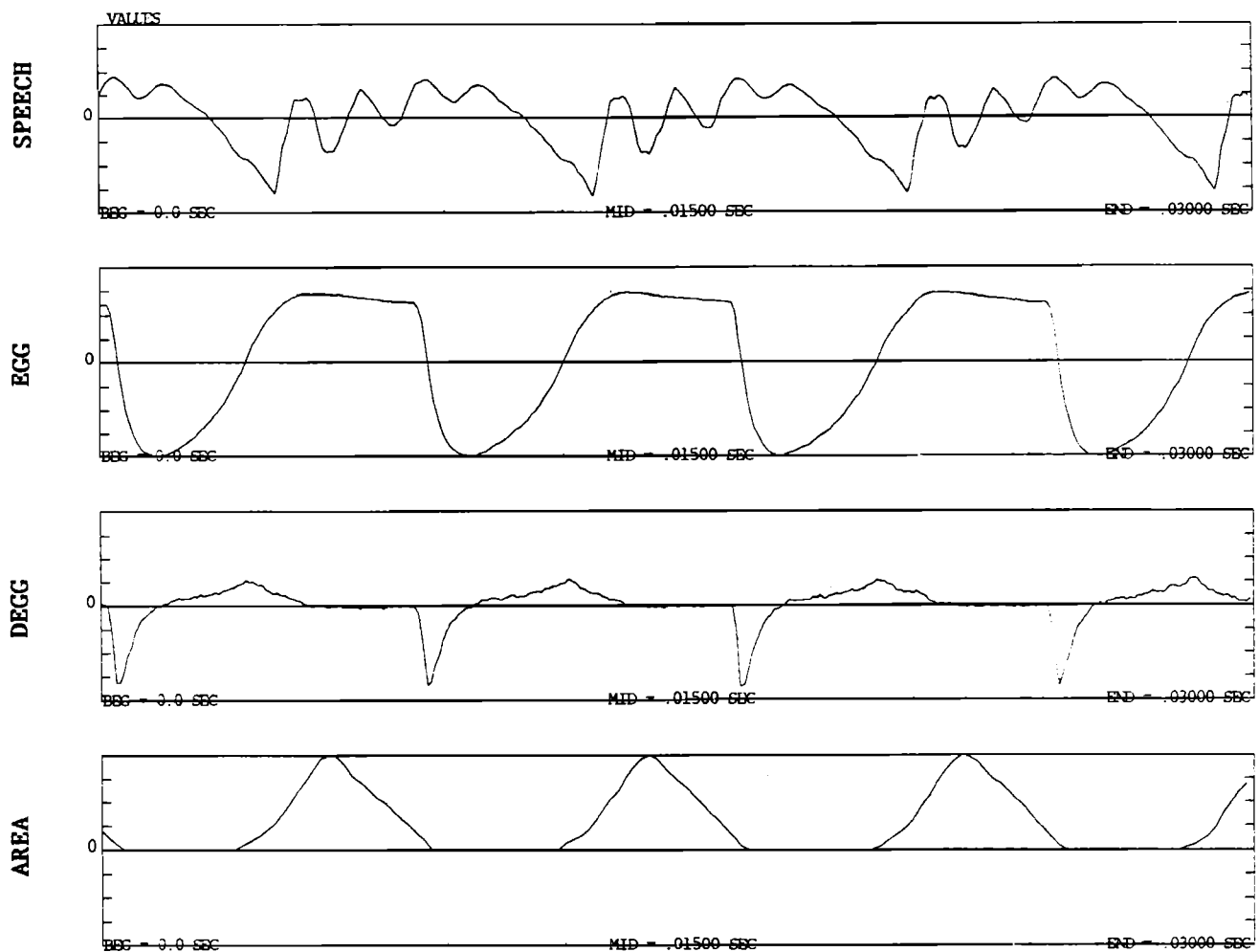


FIG. 10. The speech, EGG, DEGG, and glottal area waveforms for the vowel /i/. The glottal area waveform was measured from ultra high-speed laryngeal films and was used as the basis for validating the estimation of the opening, closing and closed glottal phases from the EGG and DEGG waveforms. The speech signal was monitored simultaneously with the EGG. All waveforms were synchronized. The closed phase was measured as the interval on the DEGG between the negative and following positive peak.

distinctive EGG and DEGG waveform patterns (implying different vocal fold contact phenomena) were found to characterize the different types of phonation. For modal and vocal fry phonations the instant of the maximum closing slope in the EGG waveform is close to its minimum extension, and thus results in a very narrow negative pulse in the DEGG waveform. This implies that the instant of the DEGG negative peak is near the occurrence of glottal closure, as observed by Childers *et al.* (1983) and Childers *et al.* (1990). The DEGG waveforms for falsetto, on the other hand, show much wider negative pulses, and thus do not provide reliable indications for glottal closure. The EGG waveforms for breathy voice have a rapid closing segment closely followed by the next opening segment, implying a momentary glottal closure or even the absence of glottal closure. The vocal fry EGG waveform also shows the double opening/closing pattern during an individual glottal cycle, as observed by other researchers (Moore and von Leden, 1958; Timcke *et al.*, 1959; Whitehead *et al.*, 1984; Klatt and Klatt, 1990).

Figure 11 indicates that the vocal fold contact varies during glottal closure for the four voice types. We defined parameters of the EGG waveform to predict these different glottal closure phenomena and found we could use these features along with additional features such as pitch, double opening/closing pattern of the volume velocity waveform, and sinusoidal-like pattern for falsetto to distinguish the four voice types (Lee, 1988). In addition, we estimated the pitch period (PP) and the glottal open quotient (OQ) for various voice types (Childers *et al.*, 1983). The pitch period was estimated as the time duration between two successive negative peaks in the DEGG waveform. The open quotient was defined as

$$\text{OQ} = \text{duration of the open phase/pitch period}, \quad (7)$$

where the open phase was estimated by the time duration between a positive peak and the next adjacent negative peak in the DEGG waveform.

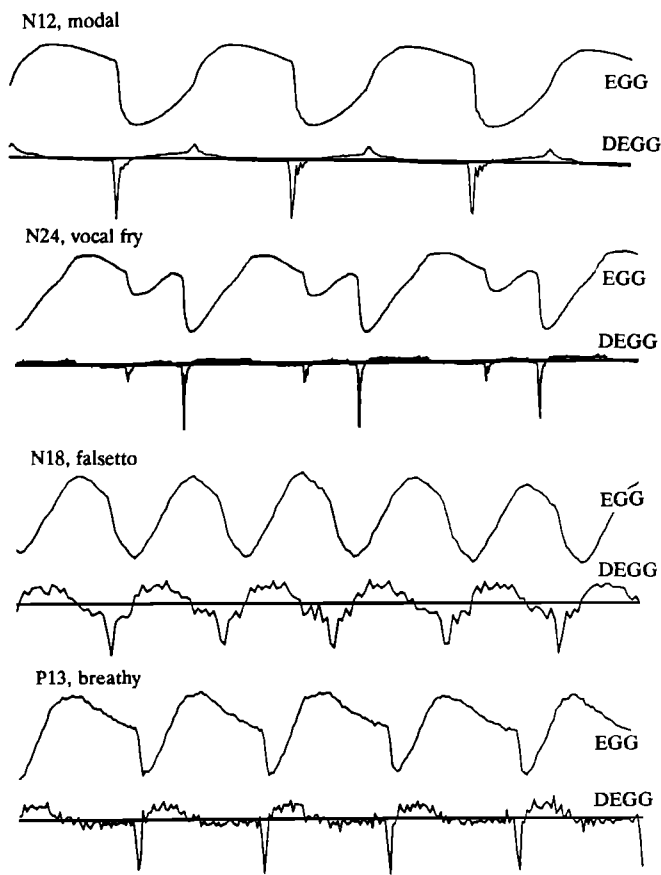


FIG. 11. The EGG and DEGG waveforms for the vowel /i/ for modal, vocal fry, falsetto, and breathy voice types.

The average PP and OQ values estimated from the DEGG waveforms of various voice types were generally lower than those estimated from the inverse filtered glottal waves. This was particularly so for falsetto and breathy voices, which have progressive glottal closures, consequently, the EGG-based technique underestimated the OQ values. Nevertheless, the results still showed that the ranking in order of increasing OQ values were vocal fry, modal, falsetto, and breathy voices, respectively. One of the important advantages of using the EGG is that it can register the dynamic characteristics of vocal fold movements of continuous speech.

### III. A NEW EXCITATION MODEL

The overall spectral envelope for voiced speech is determined by three factors (Fant, 1960; Flanagan, 1972): (1) the spectrum of the glottal excitation pulse, (2) the vocal tract transfer function, and (3) the radiation of the lips and nostrils. For this paper, we are focusing on the spectrum of the source. In recent years researchers have come to recognize the importance of accurately modeling the source excitation if a formant synthesizer is to faithfully reproduce natural vocal characteristics (Rosenberg, 1971; Hedelin, 1984; Fant *et al.*, 1985; Fujisaki and Ljungquist, 1986; Klatt, 1987; Lee and Childers, 1989; Pinto *et al.*, 1989; Klatt and Klatt, 1990; Childers and Wu, 1990). Both the cascade and parallel

formant synthesizers have been found to produce natural sounding speech when the source excitation waveform attempts to replicate both temporal and spectral characteristics of the human source. (See Childers and Wu, 1990, for a review.) Holmes (1973) recommended an excitation signal based on the second time derivative of a typical glottal volume-velocity waveform for the parallel formant synthesizer. For a glottal pulse with a  $-12$ -dB/oct spectral slope such a differentiated signal would give an approximately flat spectrum. But as we have shown in the previous section, human glottal excitation characteristics vary for different types of voice production. To synthesize or model natural sounding speech (for normal speech or for a vocal disorder) a more sophisticated approach is called for. Linear predictive coders are using more sophisticated excitation waveforms as well (Schroeder and Atal, 1985; Trancoso and Atal, 1990). But these models still require extensive computation, e.g., Schroeder and Atal (1985) reported that their coding procedure took 125 s of Cray-1 CPU time to process 1 s of speech.

### A. Source excitation classifications

Basically, researchers have used three types of excitation for speech synthesis, especially for formant synthesizers (Childers and Wu, 1990). These are (1) impulse excitation with a glottal shaping filter (Flanagan, 1957; Rabiner, 1968; Klatt, 1980; Klatt, 1987); (2) glottal waveforms obtained by inverse filtering (Rosenberg, 1971; Holmes, 1973) or glottal area waveforms (Yea *et al.*, 1983); and (3) excitation waveform models (Rosenberg, 1971; Fant, 1979; Ananthapadmanabha, 1984; Hedelin, 1984; Fant *et al.*, 1985; Fujisaki and Ljungqvist, 1986; Klatt, 1987; Childers *et al.*, 1989; Lee and Childers, 1989; Pinto *et al.*, 1989; Klatt and Klatt, 1990). For various reasons, excitation waveform models are flexible, easy to use, and produce natural sounding speech (Childers and Wu, 1990).

The source model (1) above is simple and can easily yield a source spectrum with  $-12$ -dB/oct slope (or any other slope in increments of  $-6$  dB/octave). However, the phase and the spectral notches that are often characteristics of natural voicing are not reproduced well and the vocal quality is poor (Childers and Wu, 1990). The fault with model (2) is that the glottal waveform is difficult to estimate and generally requires a microphone with a high-quality, low-frequency response (Childers and Wu, 1990). Furthermore, such an excitation source is not suitable for application in text-to-speech systems that require prestored source parameters.

A well-designed excitation waveform model along the lines of model (3) above, is easy to use and capable of producing natural sounding speech as shown by the researchers cited above. Klatt and Klatt (1990) have recently suggested one model. The LF model has also found acceptance (Fant *et al.*, 1985). An advantage of the LF model is that parameters of the model can be measured using the speech and EGG signals. For example, the LF model (Fant *et al.*, 1985) requires four model parameters for a differential glottal wave [Fig. 12(b)]:

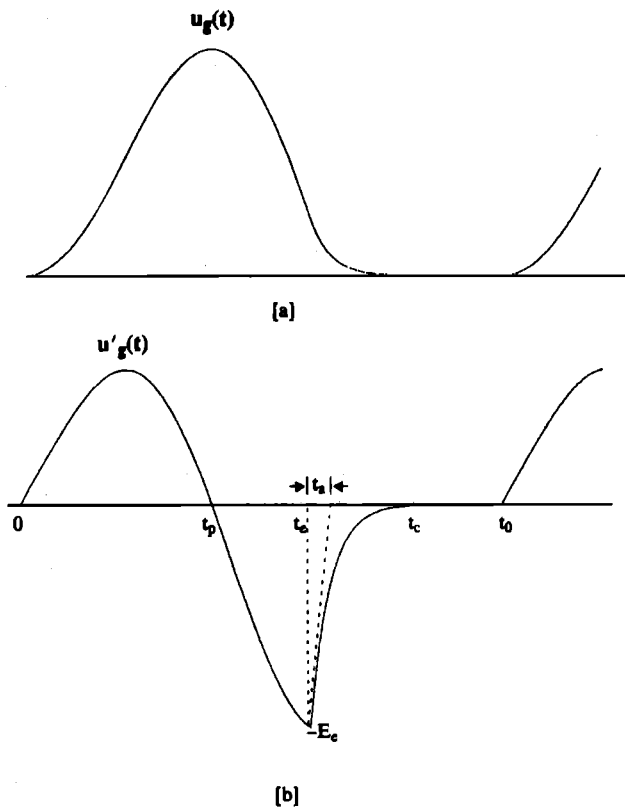


FIG. 12. The glottal waveform,  $u_g(t)$ , is shown in (a), while the Liljencrants and Fant (LF) model for the differential glottal waveform,  $u'_g(t)$ , appears in (b).

$t_p$  = glottal flow peak position,

$t_c$  = instant of maximum closing rate,

$t_a$  = time constant of an exponential recovery, i.e., return phase, from the point of maximum closing discontinuity towards the maximum closure.

The equations governing the derivative of the volume velocity in the model are

$$u'_g(t) = Ae^{\alpha t} \sin(\omega_g t), \quad \text{for } 0 \leq t \leq t_c \quad (8)$$

and

$$u'_g(t) = [u'_g(t_c)/\epsilon t_a] (e^{-\epsilon(t-t_c)} - e^{-\epsilon(t_0-t_c)}), \quad \text{for } t_c < t < t_0, \quad (9)$$

where  $\omega_g = \pi/t_p$  is the frequency of the sinewave in Eq. (8) and  $t_0$  is the pitch period. Parameters  $\alpha$  and  $\epsilon$  are defined for computational use. The value of  $\alpha$  can be derived by using the four basic parameters listed above and by solving the equation

$$\int_0^{t_0} u'_g(t) dt = 0. \quad (10)$$

Similarly, the value of  $\epsilon$  can be derived by solving the equation

$$\epsilon t_a = 1 - e^{-\epsilon(t_0-t_c)}. \quad (11)$$

To relate the waveshape parameters of the LF model to measured data, we expressed the open quotient (OQ) and speed quotient (SQ) for this model as

$$\text{OQ} = \text{open phase/pitch period} = \text{OQ}_{\text{LF}} = (t_c + kt_a)/t_0, \quad \text{for LF model,} \quad (12)$$

and

$$\begin{aligned} \text{SQ} &= \frac{\text{opening phase}}{\text{closing phase}} \\ &= \text{SQ}_{\text{LF}} = t_p/(t_c + kt_a - t_p), \quad \text{for LF model.} \end{aligned} \quad (13)$$

For the LF model, the instant of “glottal closure” was defined as the instant at which the glottal flow amplitude drops to 1% of its peak value. Based on this definition, the value of  $k$  is a function of the parameter  $t_a$ . Our data show that  $k$  has values in the range of 2.0 to 3.0 when  $0\% < t_a \leq 10\%$  ( $k = 0$  when  $t_a = 0$ ), where  $t_a$  is represented by a percentage of the pitch period  $t_0$ .

## B. Glottal factors for source modeling

As shown in the previous section, four major factors were found to be important for characterizing different types of voice production: namely, the glottal pulse width, the glottal pulse skewness, the abruptness of glottal closure, and the turbulent noise component. These factors are important in modeling the source excitation as shown below. Furthermore, these factors are related to characteristics or features of the glottal volume-velocity waveform,  $u_g(t)$ , its derivative,  $u'_g(t)$ , the EGG, its derivative, DEGG, and the speech waveform. One characteristic is depicted in Fig. 13. The main excitation of the vocal tract occurs at the instant of the maximum negative peak of  $u'_g(t)$ , which corresponds to the instant of glottal closure. Following glottal closure, the speech signal decays due to vocal tract damping. During the glottal opening phase, a secondary excitation occurs that alters the vocal tract damping trend. Although not as significant as the main excitation pulse, which occurs every pitch period, the secondary excitations also contribute to vocal quality, e.g., when such excitations are absent as in LPC synthetic speech, the voice sounds buzzy (Sambur *et al.*, 1978). The duration of glottal closure is also related to the voice type, e.g., falsetto and breathy voices have a brief glottal closure or perhaps none at all. Vocal fry, on the other hand, has a long closed phase between glottal excitation pulses.

Another feature that characterizes the four voice types, which we examined in the previous section, is the “sharpness” of the main excitation pulse in  $u'_g(t)$ , which, in turn, is caused by the steepness of the closing phase of the volume velocity and by the abruptness of glottal closure in the volume-velocity waveform. These time domain characteristics of the waveform affect the slope of the spectrum (Fant *et al.*, 1985).

The extent of the glottal waveshape variations is illustrated in Fig. 14, which shows inverse-filtered differential

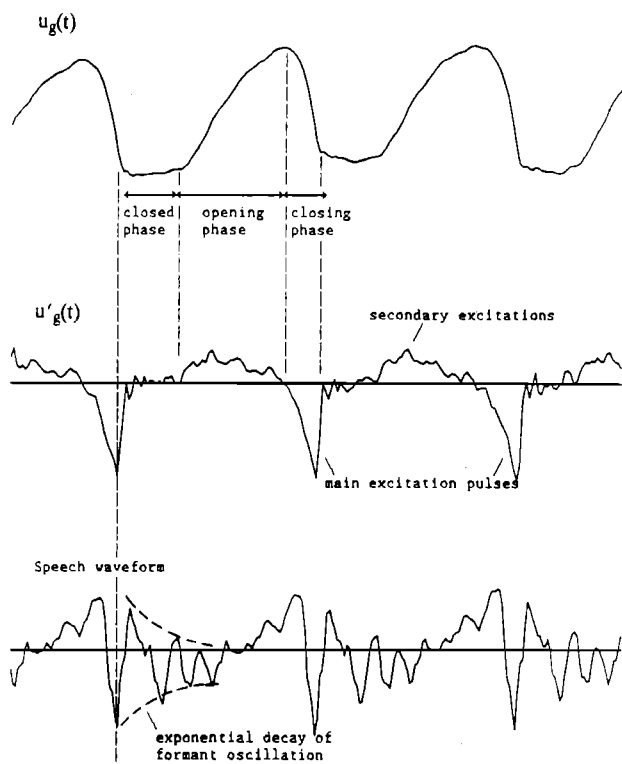


FIG. 13. Three synchronized waveforms: the glottal waveform, its derivative, and the speech waveform. Indicated on the figure are the instants of excitation of the vocal tract, the opening, closed, and closing glottal phases, and the exponential decay of the speech waveform.

glottal waveforms for the four voice types we examined, each waveform matched with an LF model waveform. The matching LF model was derived by measuring the initial estimates of the amplitude parameter  $u'_g(t_e)$  and the wave-shape parameters  $t_p$ ,  $t_c$ , and  $t_a$  in the corresponding pitch period of the differential glottal waveform. These parameters were then adjusted using a least-mean-squared error criteria that minimized the error between the signal and the matching waveform<sup>1</sup> (Lee, 1988). Some typical measured waveform parameters (represented in percentages of the pitch period) and the corresponding factors for the glottal pulse width ( $OQ_{LF}$ ) and the glottal pulse skewness ( $SQ_{LF}$ ) are given in Table III. For the four voice types we found that these glottal factors varied widely, e.g.,  $OQ_{LF}$  ranged from 0.26 to 1.0,  $SQ_{LF}$  from 1.3 to 3.6, and  $t_a$  from 0.5% to 13.3%.

### C. Turbulent noise component

When the glottis has an imperfect closure and the air-flow rate is high, turbulent airflow is produced. This phenomenon has been examined by others (Flanagan and Ishizaka, 1976; Isshiki *et al.*, 1978; Pinto *et al.*, 1989; Childers and Ding, 1991). The sound pressure of the turbulent noise is approximately proportional to the square of the volume velocity of the airflow and inversely proportional to the cross-sectional area of the constriction. Furthermore, the energy of the turbulent noise is distributed over a wide range

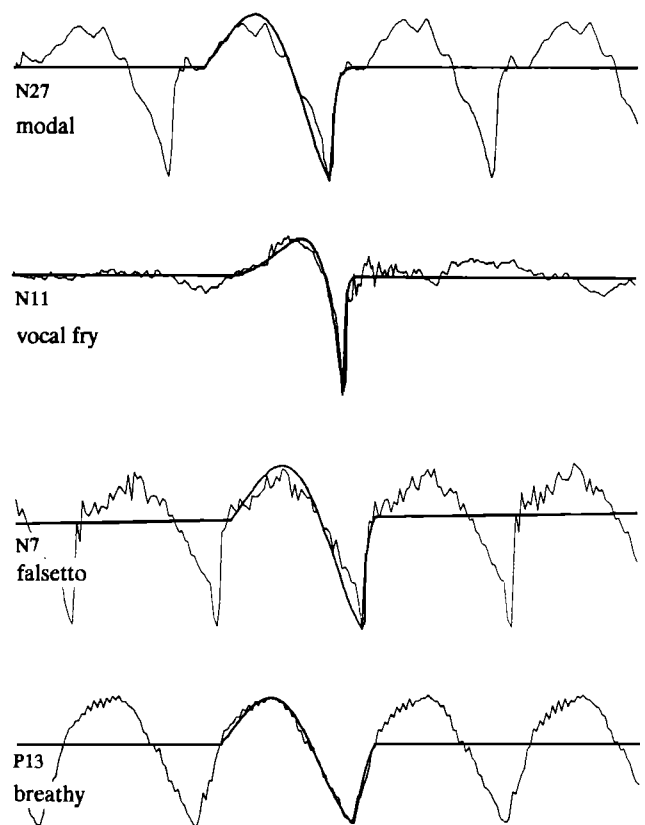


FIG. 14. The inverse filtered differential glottal waveforms and the superimposed (least-mean-square error) waveforms produced by the Liljencrants and Fant (LF) model for selected subjects, voice types, and phonations for the subjects indicated in Table III.

of frequencies (2–8 kHz) with some accentuation at about 4 kHz. Our data for breathy voices agreed with these results, showing a high interharmonic noise level above 2 kHz.

Although the existence of high-frequency turbulent noise has been shown to be an important feature for breathiness (Yanagihara, 1967; Yumoto *et al.*, 1982; Hiraoka *et al.*, 1984; Hurme and Sonninen, 1985), the role of a turbulent noise source in synthetic speech has only recently received

TABLE III. Typical waveshape parameters (based on the LF model) for glottal waveforms for the four voice types for a few male subjects (N = normal subject; P = disordered subject),  $t_0$  (pitch period) in number of sample points (sampling period = 0.1 ms).  $t_p$ ,  $t_c$ , and  $t_a$  in percentage of the pitch period.

Subject	Voice type	$t_0$	$t_p$	$t_c$	$t_a$	$OQ_{LF}$	$SQ_{LF}$
N14	modal /i/	94	49%	64%	2.1%	0.67	2.7
N27	modal /i/	79	53%	71%	2.5%	0.75	2.4
N3	modal /a/	65	51%	68%	1.5%	0.69	2.8
N23	slight vocal fry /a/	119	49%	63%	0.8%	0.64	3.3
N11	vocal fry /a/	220	20%	25%	0.5%	0.26	3.6
N19	falsetto /a/	29	57%	77%	13.3%	1.00	1.3
N7	falsetto /a/	47	62%	89%	4.3%	0.98	1.7
P8	breathy /i/	73	48%	68%	6.8%	0.81	1.5
P13	breathy /i/	50	58%	84%	10.0%	1.00	1.4

serious consideration (Pinto *et al.*, 1989; Lee and Childers, 1989; Klatt and Klatt, 1990).

#### D. Existing source models

To synthesize natural-sounding speech with various quality characteristics, a voice source model must have controllable parameters that are important for perception. As discussed above, we considered glottal pulse width, glottal pulse skewness, the abruptness of glottal closure, and a turbulent noise component as important features for a source excitation model. Several typical glottal waveform models and voice source models already exist (Rosenberg, 1971; Fant, 1979; Hedelin, 1984; Ananthapadmanabha, 1984; Fant *et al.*, 1985; Fujisaki and Ljungqvist, 1986; Klatt, 1987; Klatt and Klatt, 1990). All of these models allow variable glottal pulse width and skewness. Only one model uses a turbulent noise component (Klatt, 1987; Klatt and Klatt, 1990). But few details of the turbulent noise generator are given. Three glottal waveform models (Fant *et al.*, 1985; Ananthapadmanabha, 1984; Fujisaki and Ljungqvist, 1986) can vary the abruptness of the glottal closure, while the others (Rosenberg, 1971; Fant, 1979; Hedelin, 1984; Klatt, 1987) always generate an abrupt glottal closure after the instant of maximum closing slope. No source model incorporates all of the four factors we considered important for characterizing vocal quality. Except for the turbulent noise component, the LF model has the capability of varying the three glottal waveshape factors with only four model parameters, and as we have shown can approximate a wide range of glottal excitation waveforms.

#### E. A new experimental source model

Based on the data from this study, we designed the excitation model shown in Fig. 15, which consists of two components: (1) a glottal pulse generator, and (2) a turbulent noise generator.

##### 1. Glottal pulse generator

We adapted the LF model (Fant *et al.*, 1985) for the glottal pulse generator because the LF model can approximate a wide range of pulse widths, pulse skewness, and abruptness of glottal closure. Furthermore, the parameter values for the LF model can be derived by using the inverse filtered differential glottal waveform and/or the EGG signal.

##### 2. Turbulent noise generator

The turbulent noise generator consists of a random number generator, a spectrum-shaping filter, and an amplitude modulator. The random number generator produces random noise with a normal distribution and a flat spectrum. The amplitude level of the random noise was controlled by a parameter,  $A_n$ , that specified the ratio of the energy of the noise to the energy of the glottal pulse waveform. This ratio is small, typically about 0.25%. The spec-

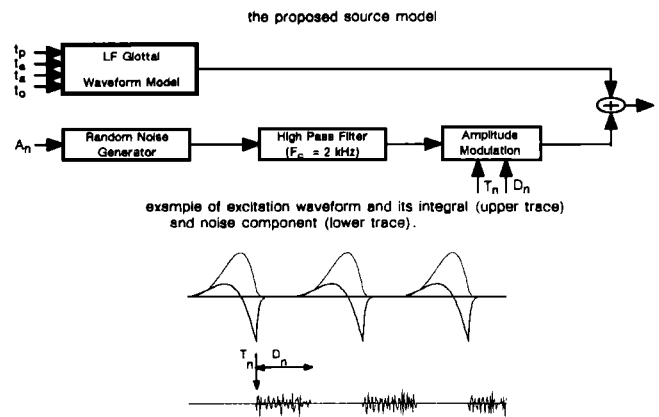


FIG. 15. Block diagram of the new source model and an example of the generated excitation wave.

trum-shaping filter was designed to simulate the spectral characteristics of the glottal turbulent noise. The present design, based on our data and that of Isshiki *et al.* (1978) uses a high-pass filter with a 3-dB frequency of 2 kHz. This differs from that suggested by Klatt and Klatt (1990). The amplitude modulator simulates the amplitude fluctuations of the glottal turbulent noise due to the variations in airflow and glottal area during vocal fold vibration. A pitch-modulated square wave with an adjustable duty cycle is presently used, but is easily modified. Two parameters are used to control the starting position ( $T_n$ ) of the turbulent noise and its duty cycle ( $D_n$ ). The parameter  $T_n$  designates the location of the onset of the turbulent noise source relative to the beginning of the pitch period, e.g.,  $T_n = 75\%$  denotes a noise onset at 3/4 of the pitch period of the glottal excitation waveform. The parameter  $D_n$  controls the duration of the noise source, e.g.,  $D_n = 50\%$  denotes that the noise source is on the 1/2 of the pitch period of the glottal excitation waveform. At present, these parameters are specified by the designer, but in the future they might be measured from the speech signal. Besides simulating breathiness, the turbulent noise generator is also suitable for producing excitation signals for voiced fricatives, where the turbulent noise is generated at a constriction in the vocal tract. Our observations suggest that the amplitude modulation factor is important for voiced fricatives.

#### IV. PERCEPTUAL EVALUATION

##### A. Method

The tokens for the listening tests were synthesized using a cascade formant synthesizer of our own design (Pinto *et al.*, 1989) but based on Klatt's synthesizer (Klatt, 1980). For the listening tests reported here, all tokens were the sustained vowels /i/ and /a/. All tokens were 2 s in duration; all were adjusted to have the same energy level. While we were able to synthesize sentences, this involved additional factors such as phoneme transitions that were not part of this study. The tokens for the listening tests were generated by comput-

er via a digital-to-analog converter and presented via headphones in an IAC single-wall sound booth. The tokens were arranged into groups, each group consisting of four vowel tokens. For each glottal source factor under investigation, a group of synthetic vowels was produced by progressively changing the selected source parameter while maintaining the formant structure fixed. A group of vowel tokens was created by varying one glottal excitation source parameter over a range of four values while all other parameters were held fixed. The vowels /i/ and /a/ were synthesized using known formant structures while the glottal excitation source characteristics were varied. For example, we varied  $t_p$  (the glottal opening time),  $t_c$  (the instant of maximum closing rate), and  $t_a$  (the time factor for controlling the abruptness of glottal closure). Each of these parameters was progressively varied in four steps while the other parameters were held fixed. Thus we created a group of four synthesized vowels with only one parameter being varied in four steps. The SQ and OQ were varied in a similar manner and the same vowels were also synthesized.

The judges for the listening tests were three professors from the Department of Speech at the University of Florida. Each judge was familiar with synthetic speech. Furthermore, each judge was an expert in voice evaluation in a clinical setting. Only three judges were used for the listening tests since we sought preliminary guidance with respect to the effectiveness of our glottal source model for synthesizing three voice qualities. Each judge was asked to perceptually rate the various tokens in the manner described below. The judges were not informed as to the manner by which the tokens were synthesized; i.e., they did not know the excitation source parameters that were being varied nor the order in which the synthesized tokens were presented, although this was not a factor since they could listen to the vowel tokens within a group as many times as they desired before indicating a rating of the vowel tokens. The judges were given three terms to describe and, thereby, rate the three voice qualities we synthesized, namely, naturalness, breathiness, and hypo-/hyperfunction. Naturalness was defined as "human sounding." The judges were familiar with breathiness and generally agreed that a breathy voice is usually associated with an incomplete glottal closure during vocal fold vibration, suggesting that the important audible component of the sound is the noise that is produced by the escape of turbulent airflow at the glottis. The judges also agreed that the voice quality of hypo- or hyperfunction is related to the perceptual sensation of vocal effort. Hyperfunction was used to describe a strained or tense voice quality, as if the vocal folds were compressed during phonation. Hypofunction was described as too little tension in the vocal folds, thereby resulting in an thin or lax voice quality. For each group of four vowel tokens the judges rated the stimuli on a 7-point scale from zero to six (Prosek *et al.*, 1987; Eskenazi *et al.*, 1990). Only one type of rating was made at a time, e.g., naturalness or breathiness or hypo-/hyperfunction. If a particular group of tokens was to be rated for two voice qualities, then the judges listened to the group of tokens on two separate occasions. A rating of six represented the highest degree of quality for naturalness and breathiness. A rating of six on the

hypo-/hyperfunction scale represented the extreme hyperfunction while zero represented the extreme hypofunction and three represented a normal voice quality. The judges were allowed to train on both synthesized and spoken tokens for the three voice qualities of naturalness, breathiness, and hypo-/hyperfunction. This allowed each judge to establish a correspondence between their own perceptual evaluation and the rating scale they were to use.

## B. Results

For the rating of the degree of hypo-/hyperfunction, i.e., the lax/tense vocal quality of the vowel tokens, the results were as follows. With  $t_a$  and  $t_c$  fixed, as  $t_p$  varied from 58% to 36% the synthesized voice was judged to go from hyperfunctional to hypofunctional. The perceptual ratings by the three judges were in general agreement on the seven point scale. The rating for  $t_p = 58\%$  was (5, 4, 6) for judges 1, 2, and 3, respectively. When  $t_p = 36\%$ , the rating was (0, 1, 0). With  $t_p$  and  $t_a$  fixed and  $t_c$  varying from 55% to 85%, the synthesized voice was judged to go from slightly hyperfunctional to hypofunctional. The judges' ratings for  $t_c = 55\%$  was (4, 4, 6), while that for  $t_c = 85\%$  was (0, 0, 0). A result similar to the latter was obtained with  $t_p$  and  $t_c$  fixed but with  $t_a$  varying from 0% to 10%. The ratings for  $t_a = 0\%$  was (5, 4, 5) and was (0, 1, 0) for  $t_a = 10\%$ .

The perceptual sensation of vocal effort was closely related to the speed quotient, SQ. A high SQ (7.3) produced a broad peak in the spectrum at high frequencies, which was perceived by the judges as a tense or hyperfunctional vocal quality. The judges' ratings were (5, 4, 6). On the other hand, a small SQ (1.2) produced a steeply declining spectral slope for the source, which resulted in a lax or hypofunctional vocal quality. The judges' rating were (0, 1, 0). An SQ of 3.0 produced tokens that were judged as normal in voice quality. The rating was (3, 2, 3). The open quotient, OQ, was not very useful for predicting the vocal quality of hypo-/hyperfunctional voice.

A number of experiments were conducted synthesizing a breathy voice by varying aspects of the turbulent noise component. In one experiment four types of turbulent noise were tested: (1) continuous noise with a flat spectrum, (2) the noise time waveform set to 50% of the pitch period with a flat spectrum, (3) continuous high-pass filtered noise, and (4) high-pass filtered noise with the time duration being set to 50% of the pitch period. Other experiments evaluated the perceptual effect of varying the duration (in percentage of the pitch period) of the noise component as well as varying the location of the noise source within the pitch period. We also examined the correlation between the degree of perceptual breathiness and the noise-to-harmonic ratio. The findings from these experiments were as follows.

(1) Amplitude modulation of the turbulent noise source is important for achieving a natural sounding synthetic breathy voice. Our results suggest that a duty cycle,  $D_n$ , of about 50% (or slightly greater) is preferred.

(2) Although the location within a pitch period of the noise production was not critical, the perception of naturalness was improved when the noise source was located near

the time at which maximum glottal closure occurred or slightly after, i.e.,  $T_n$  occurred at about 75% of the excitation pitch period (Lee and Childers, 1990; Childers and Ding, 1991).

(3) High-pass filtering of the turbulent noise is not critical for the perception of breathiness since the effect of noise in the low-frequency region is masked by strong harmonics of the fundamental frequency.

(4) The degree of perceptual breathiness was primarily determined by the noise-to-harmonic ratio at frequencies above 2 kHz. The noise pulse energy,  $A_n$ , was about 0.25%. The  $NHR_n$  was controlled by varying the harmonics of the glottal excitation pulse. This was done by varying  $t_p$ ,  $t_e$ , and  $t_a$  being the major parameter.

Other experiments were conducted to synthesize vocal fry and falsetto. These experiments, however, were found to require parameters that were not the primary interest of this study, e.g.,  $F_0$ ,  $F_0$  perturbations, and multiple excitation pulses for vocal fry. Consequently, we do not report these results here.

## V. DISCUSSION AND CONCLUSIONS

The four voice types (modal, vocal fry, falsetto, and breathy) were found to be characterized by four major factors: pulse width, pulse skewness, the abruptness of glottal closure, and turbulent noise. The effectiveness of the four factors for synthesizing a particular vocal quality was evaluated using a new source excitation model with a formant synthesizer. Other factors included the glottal spectral slope, the harmonic richness factor, and the waveform peak factor. Typical measured values for each of these factors are indicated in Table IV.

From the perceptual listening evaluations of the synthesized vowel tokens, we found that the major distinguishing

features in the frequency domain included the slope of the spectrum and the relationships between the fundamental frequency and higher harmonics as well as the noise-to-harmonic ratio. The results showed that the sensation of vocal effort is closely related to two parameters of the glottal waveform: the speed quotient (a time domain parameter) and the spectral slope (a frequency domain parameter). A large SQ (7.3) produces energy at higher frequencies, thereby causing the perceptual sensation of tense or hyperfunctional voice quality. On the other hand, a small SQ (1.2) causes a steep-falling spectral slope resulting in a lax or hypofunctional voice quality. The results of the listening tests also revealed that the degree of perceptual breathiness was strongly correlated with the noise-to-harmonic ratio above 2 kHz. The temporal characteristics of the turbulent noise source were also found to be important for producing a natural-sounding voice or a breathy voice. The three major parameters of the glottal excitation model are: (1)  $T_n$ —the location of the onset of the turbulent noise source as a percentage of the pitch period (typically 75%); (2)  $D_n$ —the duty cycle of the turbulent noise source (typically 50%); and (3)  $A_n$ —the ratio of the noise energy of the turbulent noise source to the energy of the glottal excitation pulse (typically 0.25%). The  $NHR_n$  for the model is controlled by the shape of the glottal pulse generated by the model. The key parameter is  $t_a$ .

Since the judges' ratings for the various listening tasks were found to generally agree with one another, we feel that other similarly skilled judges would have rated the tasks in a similar manner. The listening test results also imply that on the whole the glottal excitation model and its associated parameters were effective in synthesizing the three voice qualities of naturalness, breathiness, and hypo-/hyperfunction. However, the listening test results should not be interpreted as determining the best choice of glottal excitation param-

TABLE IV. Summary of the source-related features and their typical values.

Voice types		Modal voice	Vocal fry	Falsetto	Breathy voice
Features					
Glottal waveform	Pulse width (OQ)	medium (0.70)	short (0.45)	long (0.99)	long (0.91)
	Pulse skewing (SQ)	medium (2.6)	high (3.5)	low (1.5)	low (1.4)
	Abruptness of closure ( $t_a$ )	abrupt closure (2.0%)	very abrupt closure (0.7%)	progressive closure (8.8%)	progressive closure (8.4%)
Glottal spectrum	Spectral slope (dB/oct)	medium (− 12)	slight (− 6)	steep (− 18)	steep (− 18)
	Harmonic richness factor (dB)	medium (− 9.9)	high (2.1)	low (− 19.1)	low (− 16.7)
Speech features	Turbulent noise ( $NHR_n$ , in dB)	low (− 5.3)	low (unable to measure)	low (− 6.6)	high (2.8)
	Waveform peak factor	medium (2.8)	high (4.0)	low (1.8)	low (unable to measure)

eters or even the best glottal model for the three voice qualities we considered. Our results are only valid for the excitation model and associated parameters we used. Another model may yield different listening test results.

The currently available methods for estimating glottal waves are restrictive in one way or another. This study has shown that it is possible to measure parameters that are sensitive to source features from both the speech and EGG signals. However, these procedures need further development especially for connected speech, where various prosodic patterns are used to express different types of statements. Thus an important extension of this study would be to develop methods for measuring voice source dynamics for connected speech. For example, various intonation and stress patterns may be correlated to source parameters other than fundamental frequency and timing. The ultimate practical goal is to develop a source model and synthesis rules that can produce natural-sounding synthetic speech with desired vocal and tonal characteristics. [Only preliminary experiments have been conducted on synthesizing connected speech using the methods reported here (Pinto *et al.*, 1989; Childers and Wu, 1990).] We also note that differences in physiological structure, social customs, gender, and age may give rise to distinctive phonation patterns that result in various voice characteristics.

As a final note, we mention that the knowledge gained from this study might benefit applications of speech recognition and speaker identification. For example, one application would be to examine ways to study the extraction and use of source parameters for speaker adaptive signal processing to improve the reliability of a speaker-independent speech recognition system.

## ACKNOWLEDGMENTS

This work was supported in part by NIH Grant No. NIDCD DC 00577 with additional support from the University of Florida Center for Excellence Program in Information Transfer and Processing and also from the Mind-Machine Interaction Research Center.

<sup>1</sup>This is similar to a procedure adopted by Titze (1984). Note, however, that a least-mean-square error criterion is not guaranteed to produce a glottal waveform excitation model that may generate perceptually the most natural sounding synthetic speech.

- Allen, E. L., and Hollien, H. (1973). "A laminagraphic study of pulse (vocal fry) register phonation," *Folia Phoniatr.* **25**, 241-250.
- Ananthapadmanabha, T. V. (1984). "Acoustic analysis of voice source dynamics," in *Quarterly Progress and Status Report* (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden), Parts 2 and 3, pp. 1-24.
- Askenfelt, A. G., and Hammarberg, D. (1986). "Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures," *J. Speech Hear. Res.* **29**, 50-64.
- Boone, D. R. (1971). *The Voice and Voice Therapy* (Prentice-Hall, Englewood Cliffs, NJ).
- Childers, D. G., Hicks, D. M., Moore, G. P., and Alsaka, Y. A. (1986). "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *J. Acoust. Soc. Am.* **80**, 1309-1320.
- Childers, D. G., Hicks, D. M., Moore, G. P., Eskenazi, L., and Lalwani, A. L. (1990). "Electroglottography and vocal fold physiology," *J. Speech Hear. Res.* **33**, 245-254.
- Childers, D. G., and Krishnamurthy, A. K. (1985). "A critical review of electroglottography," *CRC Crit. Rev. Bioeng.* **12**, 131-164.
- Childers, D. G., Naik, J. M., Larar, J. N., Krishnamurthy, A. K., and Moore, G. P. (1983). "Electroglottography, speech and ultra-high speed cinematography," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonotory Control*, edited by I. R. Titze and R. C. Scherer (Denver Center for the Performing Arts, Denver, CO), pp. 202-220.
- Childers, D. G., and Wu, K. (1990). "Quality of speech produced by analysis-synthesis," *Speech Commun.* **9**, 97-117.
- Childers, D. G., and Ding, C. (1991). "Articulatory synthesis: nasal sounds and male and female voices," *J. Phon.* **19**, 453-464.
- Childers, D. G. and Wu, K. (1991). "Gender recognition from speech, Part II: Fine analysis," *J. Acoust. Soc. Am.* **90**, 1828-1840.
- Coleman, R. F. (1960). "Some acoustic correlates of hoarseness," Master's thesis, Vanderbilt University, Nashville, TN.
- Colton, R. F. (1969). "Some acoustic and perceptual correlates of the modal and falsetto registers," Ph. D. dissertation, University of Florida, Gainesville, FL.
- Colton, R. H. (1973a). "Vocal intensity in the modal and falsetto registers," *Folia Phoniatr.* **25**, 62-70.
- Colton, R. H. (1973b). "Some acoustic parameters related to the perception of modal-falsetto voice quality," *Folia Phoniatr.* **25**, 302-311.
- Colton, R. H., and Hollien, H. (1972). "Phonational range in the modal and falsetto registers," *J. Speech Hear. Res.* **15**, 708-713.
- Colton, R. H., and Hollien, H. (1973). "Perceptual differentiation of the modal and falsetto registers," *Folia Phoniatr.* **25**, 270-280.
- Damste, H., Hollien, H., Moore, G. P., and Murry, T. (1968). "An x-ray study of vocal fold length," *Folia Phoniatr.* **20**, 349-359.
- Davis, S. B. (1979). "Acoustic characteristics of normal and pathological voices," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. Lass (Academic, New York), Vol. 1, pp. 271-335.
- Deal, R. E., and Emanuel, F. W. (1978). "Some waveform and spectral features of vowel roughness," *J. Speech Hear. Res.* **21**, 250-264.
- Eskenazi, L., Childers, D. G., and Hicks, D. M. (1990). "Acoustic correlates of vocal quality," *J. Speech Hear. Res.* **33**, 298-306.
- Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, Paris).
- Fant, G. (1979). "Glottal source and excitation analysis," *Quarterly Progress and Status Report* (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden), Vol. 1, pp. 85-125.
- Fant, G., Liljencrants, J., and Lin, Q.-G. (1985). "A four parameter model of glottal flow," *STL-OPSR* **4**, 1-13.
- Flanagan, J. L. (1957). "Note on the design of terminal-analog speech synthesizers," *J. Acoust. Soc. Am.* **29**, 306-310.
- Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, New York), 2nd ed.
- Flanagan, J. L., and Ishizaka, K. (1976). "Acoustic generation of voiceless excitation in a vocal cord-vocal tract speech synthesizer," *IEEE Trans. Acoust. Speech Signal Process.* **24**, 163-170.
- Fujisaki, H., and Ljungqvist, M. (1986). "Proposal and evaluation of models for the glottal source waveform," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1605-1608.
- Hedelin, P. (1984). "A glottal LPC-vocoder," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1.6.1-1.6.4.
- Heiberger, V. L., and Horii, Y. (1982). "Jitter and shimmer in sustained phonation," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. Lass (Academic, New York), Vol. 7, pp. 299-322.
- Hiller, S. M., Laver, J., and Mackenzie, J. (1983). "Automatic analysis of waveform perturbations in connected speech," *Work in Progress*, Dept. of Linguistics, Edinburgh University, **16**, 40-68.
- Hillman, R. E., Osterle, E., and Feth, L. L. (1983). "Characteristics of the turbulent noise source," *J. Acoust. Soc. Am.* **74**, 690-694.
- Hirano, M., Hibi, S., Terasawa, R., and Fujiu, M. (1985). "Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia," *Symposium on Voice Acoustic and Dysphonia*, Gotland, Sweden.
- Hiraoka, N., Kitazoe, Y., Ueta, H., Tanaka, S., and Tanabe, M. (1984). "Harmonic-intensity analysis of normal and hoarse voices," *J. Acoust. Soc. Am.* **76**, 1648-1651.
- Hollien, H. (1974). "On vocal register," *J. Phon.* **2**, 125-144.
- Hollien, H., Coleman, R. F., and Moore, P. (1968). "Stroboscopic laminagraphy of the larynx during phonation," *Acta Oto-laryngol.* **LXV**, 209-215.
- Hollien, H., and Colton, R. H. (1969). "Four laminagraphic studies of vocal fold thickness," *Folia Phoniatr.* **21**, 179-198.



- Hollien, H., Girard, G. T., and Coleman, R. F. (1977). "Vocal fold vibratory patterns of pulse register phonation," *Folia Phoniatr.* **29**, 200-205.
- Hollien, H., and Michel, J. F. (1968). "Vocal fry as a phonational register," *J. Speech Hear. Res.* **11**, 600-604.
- Holmes, J. N. (1973). "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.* **AU-21**, 298-305.
- Horii, Y. (1980). "Vocal shimmer in sustained phonation," *J. Speech Hear. Res.* **23**, 202-209.
- Hurme, P., and Sonninen, A. (1985). "Normal and disordered voice quality: listening tests and long-term spectrum analyses," Symposium on Voice Acoustics and Dysphonia, Gotland, Sweden.
- Isshiki, N., Kitajima, K., Kojima, H., and Harita, Y. (1978). "Turbulent noise in dysphonia," *Folia Phoniatr.* **30**, 214-224.
- Kasuya, H., Ogawa, S., and Kikuchi, Y. (1986). "An acoustic analysis of pathologic voice and its application to the evaluation of laryngeal pathology," *Speech Commun.* **5**, 171-181.
- Kitzing, P. (1982). "Photo- and electrophysiological recording of the laryngeal vibratory pattern during different registers," *Folia Phoniatr.* **34**, 234-241.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971-995.
- Klatt, D. H. (1987). "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.* **82**, 737-793.
- Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.* **87**, 820-857.
- Krishnamurthy, A. K., and Childers, D. G. (1986). "Two-channel speech analysis," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**, 730-743.
- Ladefoged, P. (1975). *A Course in Phonetics* (Harcourt Brace Jovanovich, New York), pp. 121-124.
- Laver, J. (1980). *The Phonetic Description of Voice Quality* (Cambridge U. P., New York), pp. 115-117.
- Laver, J., and Hanson, R. (1981). "Describing the normal voice," in *Evaluation of Speech in Psychiatry*, edited by J. Darby (Grune and Stratton, New York), pp. 51-78.
- Lee, C. K. (1988). "Voice quality: Analysis and synthesis," Ph.D. dissertation, University of Florida, Gainesville, FL.
- Lee, C. K., and Childers, D. G. (1989). "Some acoustical perceptual and physiological aspects of vocal quality," in *Vocal Fold Physiology*, edited by B. Hammarberg and J. Gauffin (Raven, New York) (in press).
- Monsen, R., and Engebretson, M. (1977). "Study of variations in the male and female glottal wave," *J. Acoust. Soc. Am.* **62**, 981-993.
- Moore, G. P. (1975). "Observation on the physiology of hoarseness," Proceedings of the 4th International Congress of Phonetic Science, Helsinki, Finland, pp. 92-95.
- Moore, P., and von Leden, H. (1958). "Dynamic variations of the vibratory pattern in the normal larynx," *Folia Phoniatr.* **10**, 205-238.
- Pinto, N. B., Childers, D. G., and Lalwani, A. L. (1989). "Formant speech synthesis: improving production quality," *IEEE Trans. Acoust. Speech Signal Process.* **37**(12), 1870-1887.
- Prosek, A. R., Montgomery, B. E., and Hawkins, D. B. (1987). "An evaluation of residue features as correlates of voice disorders," *J. Commun. Disord.* **20**, 105-117.
- Rabiner, L. R. (1968). "Digital-formant synthesizer for speech synthesis studies," *J. Acoust. Soc. Am.* **43**, 822-828.
- Rabiner, L. R., and Schafer, R. W. (1978). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).
- Rosenberg, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.* **49**, 583-590.
- Sambur, M. R., Rosenberg, A. E., Rabiner, L. R., and McGonegal, C. A. (1978). "On reducing the buzz in LPC synthesis," *J. Acoust. Soc. Am.* **63**, 918-924.
- Schroeder, M. R., and Atal, B. S. (1985). "Code-excited linear prediction (CELP): high quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 937-940.
- Timcke, R., von Leden, H., and Moore, P. (1959). "Laryngeal vibrations: measurements of the glottic wave," *Arch. Otolaryngol.* **69**, 438-444.
- Titze, I. R. (1984). "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc. Am.* **75**(2), 570-580.
- Titze, I. R. (1990). "Interpretation of the electroglottographic signal," *J. Voice* **4**, 1-9.
- Trancoso, I. M., and Atal, B. S. (1990). "Efficient search procedures for selecting the optimum innovation in stochastic coders," *IEEE Trans. Acoust. Speech Signal Process.* **38**, 385-396.
- Wendahl, R. W. (1963). "Laryngeal analog synthesis of harsh voice quality," *Folia Phoniatr.* **15**, 241-250.
- Wendahl, R. W. (1966). "Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness," *Folia Phoniatr.* **18**, 98-108.
- Wendahl, R. W., Moore, P., and Hollien, H. (1963). "Comments on vocal fry," *Folia Phoniatr.* **15**, 251-255.
- Whitehead, R. L., Mets, D. E., and Whitehead, B. H. (1984). "Vibratory patterns of the vocal folds during pulse register phonation," *J. Acoust. Soc. Am.* **75**, 1293-1297.
- Wolfe, V. I., and Steinfatt, T. M. (1987). "Prediction of vocal severity within and across voice types," *J. Speech Hearing Res.* **30**, 230-240.
- Wu, K., and Childers, D. G. (1991). "Gender recognition from speech, Part I: Coarse analysis," *J. Acoust. Soc. Am.* (in press).
- Yanagihara, H. (1967). "Significance of harmonic changes and noise components in hoarseness," *J. Speech Hear. Res.* **10**, 531-541.
- Yea, J. J., Krishnamurthy, A. K., Naik, J. M., and Childers, D. G. (1983). "Glottal sensing for speech analysis and synthesis," *Int. Conf. Acoust. Speech Signal Process.* 1332-1335.
- Yumoto, E., Gould, W., and Baer, T. (1982). "Harmonic-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.* **71**, 1544-1550.