Measuring and Modeling Vocal Source-Tract Interaction

Donald G. Childers, *Fellow, IEEE,* and Chun-Fan Wong, *Member, IEEE*

Abstract—**The quality of synthetic speech is affected by two factors: intelligibility and naturalness. At present, synthesized speech may be highly intelligible, but often sounds unnatural. Speech intelligibility depends on the synthesizer's ability to reproduce the formants, the formant bandwidths, and formant transitions, whereas speech naturalness is thought to depend on the excitation waveform characteristics for voiced and unvoiced sounds. Voiced sounds may be generated by a quasiperiodic train of glottal pulses of specified shape exciting the vocal tract filter. It is generally assumed that the glottal source and the vocal tract filter are linearly separable and do not interact. However, this assumption is often not valid, since it has been observed that appreciable source-tract interaction can occur in natural speech. Previous experiments in speech synthesis have demonstrated that the naturalness of synthetic speech does improve when source-tract interaction is simulated in the synthesis process. The purpose of this paper is two-fold: 1) to present an algorithm for automatically measuring source-tract interaction for voiced speech, and 2) to present a simple speech production model that incorporates source-tract interaction into the glottal source model. This glottal source model controls: 1) the skewness of the glottal pulse, and 2) the amount of the first formant ripple superimposed on the glottal pulse. A major application of the results of this paper is the modeling of vocal disorders.**

I. Introduction

A LINEAR model of speech production was developed by Fant in the late 1950's [16]. Acoustic theory demonstrates that the transmission characteristics of the vocal tract may be approximated by a cascade of resonators and antiresonators whose bandwidths and center frequencies may be independently controlled. Hence, the vocal tract may be represented as a linear short-time invariant filter. For voiced sounds, the glottal source model is represented by a train of quasiperiodic pulses with controllable pitch period and amplitude. For unvoiced sounds, the source model is represented by random noise. Speech signals are synthesized by the source exciting the vocal tract filter. A discrete-time version of the speech production model is given in Fig. 1 [50]. The source-filter theory of speech production is the basis of the majority of today's speech synthesizers. This theory states that the glottal source and the vocal tract filter are linearly separable and do not interact, implying that a time varying vocal tract configuration has no effect on the shape of the glottal source pulses. Speech synthesized using such a model is generally very intelligible, but often sounds unnatural. It is believed that the intelligibility of synthetic speech depends on the ability of the synthesizer to reproduce dynamic (transient) sounds such as stops, whereas the naturalness of the synthetic speech is mainly determined by the reproduction of voiced segments [8], [10], [11].

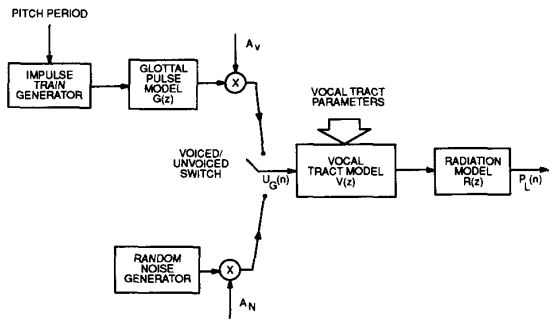


Fig. 1. Discrete time model of speech production [50].

For most speech synthesizers, the properties of the vocal fold excitation source are only coarsely approximated. It is customary to specify the pitch contour as a smooth, continuous function and to use a fixed glottal waveform whose amplitude spectrum falls at -12 dB/octave [26]. A typical source waveform is produced by repeatedly exciting a fixed, spectral-shaping, two pole filter with a train of impulses, producing the correct magnitude spectrum; however, the phase may be incorrect. One difficulty with this excitation model is that, depending on the impulse train and the spectral-shaping filter, the primary excitation presented to the vocal tract filter may occur at the instant at which the vocal folds open. This excitation model is incorrect since for real speech the primary excitation occurs when the vocal folds close, with a secondary excitation occurring as the vocal folds open. Furthermore, this excitation model does not generally introduce zeros (antiresonances) in the source spectrum, which occur in real speech. A consequence of this improper modeling of the source excitation characteristics is that synthetic speech may not sound natural and may not properly simulate the excitation for various voice types, such as breathy, hoarse, creaky, etc. [8], [11].

The shape and periodicity of the volume-velocity waveform in natural speech can vary considerably [34], [47]. The extent to which variability in the period and the shape of

Manuscript received July 23, 1992; revised February 24, 1994. This work was supported in part by NIH Grant No. NIDCD DC 00577 and NSF grant IRI-9215331 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.

D. G. Childers is with the Department of Electrical Engineering, University of Florida, Gainesville, FL 32611-2024 USA.

C.-F. Wong is with Centigram Communications Corporation, San Jose, CA 95134 USA.

IEEE Log Number 9401301.

the waveform in speech synthesis affects speech naturalness and quality is an important research question [8], [11]. It has been shown that the glottal pulse waveform can have important effects on the quality of synthetic speech, i.e., the type of voice produced. Rosenberg [51] studied the effect of varying the glottal pulse shape on speech naturalness. He used simulated waveforms that varied the number and location of slope discontinuities. His results indicated that simulated excitations with pulse shapes with a single slope discontinuity at glottal closure were preferred by listeners. Holmes [31] pursued this research and showed that under certain listening conditions, the use of glottal pulses derived from actual speech significantly improved the naturalness of synthesized speech compared to speech generated using fixed glottal waveform models. Wong and Markel [61] showed that retaining the phase characteristics of a typical glottal pulse can improve linear predictive coded (LPC) speech synthesis quality. Childers *et al.* [9] demonstrated that simulating source-tract interaction can improve the quality of synthetic speech. The synthesis of female speech has also proved problematical since we know little about speech prosody, especially pitch, individual voice qualities, speaking habits, and, the glottal closed-phase interval is short for female speech, making pitch synchronous, closed-phase analysis more difficult than that for male speech [4], [8], [11], [35]–[41]. Voice pathology and the theory of singing are related topics, where the need for improved voice source models are apparent [11], [17]. The purpose of this paper is to show that a simple excitation model can vary the glottal pulse skewness and simulate the effect of the interaction of the first formant with the source by introducing a first formant ripple on the glottal pulse. The parameters for this excitation model are measured from the speech signal using inverse filtering.

II. Source-Tract Interaction

A. Summary of Effects

Two major and several minor source-tract interaction effects have been observed. Inverse filtering of various vowel tokens spoken by the same speaker has demonstrated that glottal pulse shapes occur with varying skewness [22], [25], [32]–[34], [46], [47], [52]. Another aspect of source-tract interaction is that a ripple may be superimposed on the opening phase of the glottal pulse. This ripple is thought to be caused by the interaction of the glottal pulse with the vocal tract first-formant frequency. We may summarize the source-tract interaction effects as follows.

1) *Skewing:* Source-tract interaction causes the glottal volume-velocity waveform to be skewed to the right with respect to the glottal area (Fig. 2) [22], [25], [32]–[34], [46], [47], [52]. This effect is caused by the vocal tract loading the source, causing the volume-velocity pulse to become skewed compared to the glottal area function. Skewing depends on the input inductance of the load as seen by the source. Acoustically, the main effect of skewing is to uniformly increase the dB level of the formants.

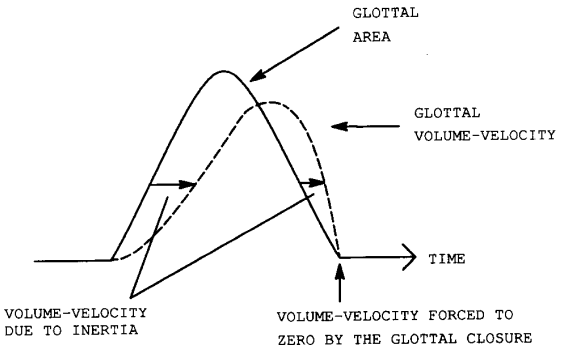


Fig. 2. Glottal volume-velocity skewed to the right.

2) *Ripples: Truncation and Dispersion:* Ripples have been observed superimposed on the opening phase segment of the glottal volume-velocity waveform (Fig. 3) [1], [14], [15], [18]–[20]. These ripples have been attributed to the first formant (F1) energy, which is dissipated by the glottis during the open phase of the glottal cycle [53]. This absorption of the first formant energy is called glottal damping, which causes: 1) a truncation of the first formant waveform, 2) a shift in the formant frequencies, and 3) an increase in the formant bandwidths during the glottal open interval. These effects are especially apparent for high F1 sounds, e.g., the vowel /A/, as in Bob. The truncation effect manifests itself as a ripple or oscillation on the glottal volume-velocity waveform. The main perceptual effect of truncation is a reduction of the loudness level of the first formant [21].
3) *Superposition–Linear:* In those cases where truncation is not significant, e.g., high-pitched voices, energy will be carried over from one glottal period to the next. The degree of carry-over depends on the relationship between the frequency of voicing and the formant frequency [1].
4) *Superposition–Nonlinear:* The superposition of the F1 oscillation of one glottal interval into the next may affect the derivative of the volume-velocity at the instant of glottal closure, thereby changing the amplitude of the excitation level [42], [48].
5) *Superposition–Mechanical:* It is conjectured that when superposition occurs, it may affect the mechanical vibratory characteristics of the vocal folds [27].
6) *Supraglottal:* A supraglottal constriction will affect the transglottal pressure drop and, thus, the pattern of the vibratory motion of the vocal folds. This is typical of voiced fricatives. Bickley and Stevens [3] have studied the effects of a vocal tract constriction on the glottal source. Their results determined that constricted vocal tract configurations resulted in longer glottal open intervals, while vocal tract configurations for open vowels tended to have shorter glottal open intervals.

B. Modeling

The methods used to simulate the effects of source-tract interaction may be classified as either interactive or nonin-

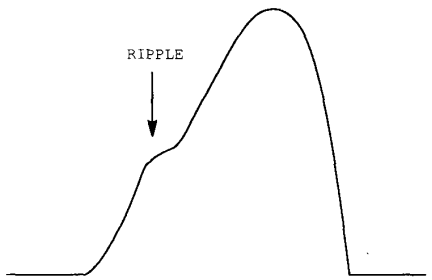


Fig. 3. Ripple superimposed on the glottal volume-velocity waveform.

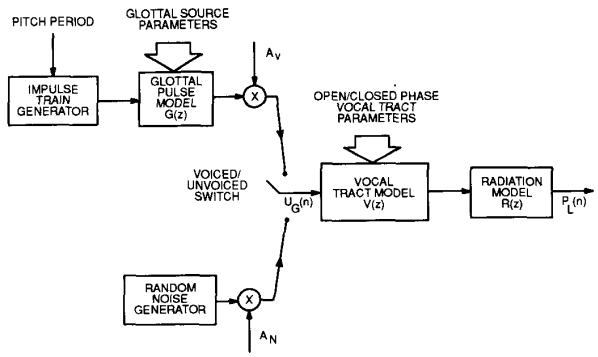


Fig. 4. Speech production model with source-tract interaction.

teractive. The interactive model does not separate the glottal source and the vocal tract. The interaction of these two systems is modeled as a nonlinear, time-varying model. For example, a model may use the lung pressure as the source and derive the speech output from a set of differential equations representing the state of the volume-velocity and pressure in the subglottal, glottal, and supraglottal systems [17]. This is the principle of the Ishizaka-Flanagan [34] model. The interactive approach for simulating source-tract interaction requires a knowledge of parameters that are not easily measured. [44], [53]–[56]. This approach is not readily implemented in a speech synthesizer.

The noninteractive approach models the source and the vocal tract filter as linearly separable systems with time-varying parameters that approximate the source-tract interaction. For example, the shape of the glottal volume-velocity waveform may be varied for different voiced sounds. Although this approach does not simulate source-tract interaction itself, it does simulate some of the major effects of source-tract interaction in a simple manner. For example, one may implement a model for nonnasalized voiced sounds with a cascade of parallel resonant electrical circuits [1]. The parallel resonant circuits act as a load on the source, with the circuit and current parameters chosen to represent the formants and to approximate the impedances of the subglottal and the supraglottal systems. The cascade arrangement of resonators yields vowel sounds with the proper formant characteristics.

Guerin *et al.* [28] developed a simple glottal source model to generate a volume-velocity waveform that simulates interaction with the vocal cavities. The load of the model is characterized by the equivalent circuit for the driving point impedance of the vocal tract. This circuit is controlled dynamically by the first two formant frequencies. Later, Cheng and Guerin [5] developed a strategy for controlling the parameters for male and female glottal source models.

Nearly all previous approaches to source-tract interaction modeling have used inverse filtering to measure the glottal velocity waveform from which source-tract interaction effects were estimated. The various circuit model parameters were calculated from these data. Since inverse filtering provides an estimate for the glottal volume velocity waveform, we reasoned that a model for source-tract interaction could be based solely on a model for the glottal volume velocity waveform, thereby eliminating the need for a circuit model for the glottal source.

Our model for speech production is shown in Fig. 4, which models the skewness of the glottal volume-velocity waveform using the Liljencrants-Fant (LF) model [24]. The properties of the LF model for the differentiated glottal volume-velocity waveform are shown in Fig. 5 [23], [24]. The LF model uses four parameters to model the differentiated glottal volume-velocity. The model consists of two parts. The first part is an exponentially growing sinusoid represented by three of the four parameters of the model.

$$\frac{dU_g(t)}{dt} = E(t) = E_o \cdot e^{\alpha t} \sin \omega_g t \qquad t_0 \le t \le t_e. \quad (1)$$

This portion of the model represents the volume-velocity from the opening of the glottis to the instant at which the main excitation occurs, which is the instant at which the maximum discontinuity in the glottal volume-velocity occurs. This discontinuity normally coincides with the instant of the maximum negative derivative.

The three parameters pertaining to the first segment of the LF model are

1) E_o is a scale factor.
2) $\alpha = B\pi$ where B is the bandwidth of the exponentially growing amplitude.
3) $\omega_g = 2\pi F_g$ where $F_g = 1/(2t_p)$ and t_p is the rise-time (the time from glottal opening to maximum volume-velocity).

The second part of the model is an exponential segment that allows a residual volume-velocity (dynamic leakage) following the main discontinuity, at time t_e, when the vocal folds close. This segment is the "return phase," which is represented as

$$E(t) = -\frac{E_e}{\varepsilon t_a}[e^{-\varepsilon(t-t_c)} - e^{-\varepsilon(t_c-t_e)}] \qquad t_e \le t \le t_c \quad (2)$$

where t_a is the fourth parameter of the model. The parameter E_e is the magnitude of the maximum negative amplitude of the differentiated glottal volume-velocity waveform and t_c is the instant of complete glottal closure. The parameter t_a is the time constant of the exponential curve and is determined by extrapolating the tangent of the derivative at time t_e until it intersects with the time axis. The parameter ε can be determined from (2), where at $t = t_e, E(t) = -E_e$. Thus,

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)}. \quad (3)$$

For small values of t_a, ε is approximately equal to $1/t_a$.

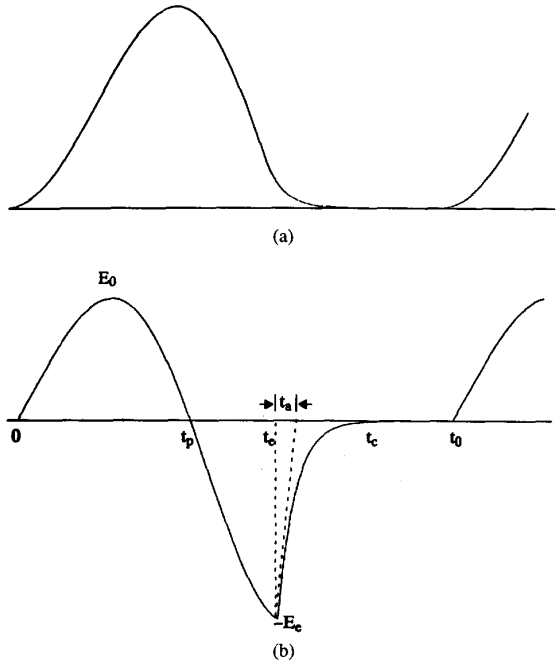


Fig. 5. The LF model of (a) the glottal volume-velocity and (b) the differentiated glottal volume-velocity waveform.

The effect of the return phase in (2) on the source spectrum may be approximated as a first order low-pass filter with a cutoff frequency $F_a = 1/(2\pi t_a)$. This means that the longer the return phase, the lower the cutoff frequency.

By convention, t_0 denotes the time of the glottal opening for the next pulse. If $t_c = t_0$, then this implies that the model has no glottal closed phase. In practice, this does not represent a problem, since for small values of t_a, which are typical, the model effectively simulates a glottal closed phase.

In addition to the four parameters, there is the requirement that the area of the differentiated LF model shown in Fig. 5 must be zero, i.e.,

$$\int_0^{t_o} E(t) = 0. \tag{4}$$

This condition maintains a constant baseline for the volume-velocity.

The ripple on the volume-velocity waveform is caused by an increase in the formant bandwidths and a shift in the formant frequencies during the open glottal phase. We will approximate the ripple effect as follows. The first formant bandwidth will be calculated using the vocal tract filter estimated for the glottal closed phase interval. This bandwidth will be increased during the glottal open phase, thereby creating a first formant ripple on the volume-velocity waveform.

III. Measurement of Model Parameters

The most direct method for measuring the source parameters is from the glottal volume-velocity waveform, which may be estimated by inverse filtering [4], [6], [11], [30], [33], [35]–[39], [43], [45]–[47], [49], [52], [56], [58], [59], [62]. The inverse filter must be estimated from the speech signal during the glottal closed phase to avoid source-tract interaction. The vocal tract filter for vowel sounds is typically modeled as an all-pole filter using linear predictive coding techniques, shown as $V(z)$ in Figs. 1 and 4. The inverse of the vocal tract filter, i.e., $(V(z))^{-1}$, therefore, contains only zeros or antiresonances.

A major difficulty with inverse filtering is locating the closed glottal interval. Another problem occurs if the voice is high-pitched, then the closed glottal interval may be short, making it difficult to estimate the coefficients of the inverse filter. If a closed glottal interval does not occur, then the volume-velocity waveform estimate may be inaccurate. Similarly, if zeros are present in the vocal tract system, as during nasalized speech, then the zeros of the glottal volume-velocity cannot be separated from the vocal tract zeros. For the latter reason, an all-pole model of the vocal tract is used for glottal inverse filtering, and nasalized speech tokens are avoided in the analysis. Since the lip radiation impedance introduces a zero at zero frequency, the glottal volume-velocity baseline cannot be recovered.

To recover the glottal volume-velocity waveform with reasonable accuracy the original speech signal should be sampled without phase distortion. Because standard tape recorders introduce phase distortion, the recording should be made with an FM system or by directly sampling the signal. Some studio microphones also introduce considerable phase distortion, so the speech pressure wave should be recorded with a condenser microphone. However, even a condenser microphone may not pick up the continuous (dc) airflow in a free field. This means that the continuous part of the airflow that occurs when the vocal folds vibrate without making a complete closure will not be properly monitored. Other drawbacks include the fact that the amplitude of the glottal airflow cannot be calibrated and that the microphone is sensitive to very low frequency changes in pressure in the environment in which the recording is made. If the data are monitored with phase distortion, then it may be corrected [2], [60].

Inverse filtering may be accomplished using only the speech signal [62], or with a two channel method that uses both the speech and electroglottographic (EGG) signals [11], [43]. Most inverse filtering techniques are not automatic, requiring adjustment by an operator. We implement two automatic glottal inverse filtering programs. One is a two channel method using both the speech and electroglottographic (EGG) signals [60]. The other is based on the speech signal alone [7]. The speech is parsed into four segments: silence, voiced, unvoiced or mixed (voiced and unvoiced excitation) [13]. The analysis of the voiced speech segments is pitch synchronous. The inverse filter is determined by a linear prediction covariance analysis of the closed phase interval of the speech waveform. The beginning and the ending points of the closed phase interval are determined from either the EGG signal or directly from the speech signal. The closed phase interval defines the maximum frame size for LPC analysis.

The LF model is matched to the measured differentiated inverse filtered waveform as follows. The values of t_e and E_e are measured from the inverse filtered differentiated glottal

flow waveform for one pitch period. The parameter t_a is determined using a least square error criterion between the inverse filtered differentiated glottal flow waveform and the LF model given by (3). Possible candidates for the parameter t_p are located in the interval from the instant of the opening of the glottis to the instant t_e. For each candidate of t_p, the parameters E_o, α, ω_g, and ε are calculated. These parameters are used to produce a model of the differentiated glottal flow waveform. The total squared error between the model waveform and the data waveform is calculated for each set of parameters. The values of the parameter set that give the minimum total squared error are selected as the best LF model for that pitch period.

The timing parameters of the LF model are closely related to the glottal waveshape factors, e.g., t_c is related to the glottal pulse width, t_a to the abruptness of glottal closure, t_e to the instant of the main excitation and the glottal pulse skewness may be represented by the speed quotient, which for the LF model is defined as

$$SQ_{LF} = \frac{\text{opening phase}}{\text{closing phase}} = \frac{t_p}{t_c - t_p}. \tag{5}$$

The speech and electroglottographic data are digitized simultaneously using Digital Sound Corp. DSC-240 preamplifiers and a DSC-200 digitizer. We sample each signal at 10 kHz with 16-bits precision. The microphone is an Electro-Voice RE-10 held six inches from the lips. The EGG device is a model from Synchrovoice, Incorporated. All data are collected in a professional IAC single wall sound booth. A Digital Equipment Corporation VAX11/750 computer system managed the data collection.

The subjects for this study consist of 52 (27 male, 25 female) subjects with normal larynges. The subject's ages ranged from 20 to 80 years old. The complete speech protocol consists of 27 tasks, including ten sustained vowels /IY, I, E, AE, A, OW, U, OO, UH, ER/, two sustained diphthongs, five sustained unvoiced fricatives /H, F, THE, S, SH/, and four sustained voiced fricatives /V, TH, Z, ZH/. This notation is adopted from Rabiner and Schafer [50]. The subjects are instructed to pronounce and sustain each vowel as it would be pronounced in the following words, respectively: beet, bit, bet, bat, Bob, bought, book, boot, but, Berr. Similar instructions are given for the diphthongs, for which the cue words were boat, and bait, while for the fricatives we use the following cue words: hat, fix, thick, sat, ship, van, this, zoo, and azure. The duration of each vowel, diphthong, and fricative approximated 2 s. The additional tasks include 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?) For this study, we analyzed the vowel data in detail. In addition, we analyzed some of the sentences to give us some indication of the glottal factors in a word and sentence context.

The analysis procedures use a 12th order linear predictor, while the closed phase interval is 28 data samples. Thus, the data window size is 40 data samples. The software varied these values to fit the closed phase interval measured from the EGG data on the speech signal. The location of the LPC analysis window is determined by minimizing the total squared error, as described above. The inverse filter is then determined. If real poles occur at the origin, then they are removed. However, real poles at the sampling frequency are retained. Extraneous poles with very low frequencies or with very large bandwidths are also removed. To provide a stable filter, poles outside the unit circle are reflected inside the unit circle, even though such a procedure might distort the glottal volume velocity waveform. Pole reflection is seldom necessary, with the primary exceptions occurring at voicing onset and offset.

The speech signal is then inverse filtered to yield the differentiated glottal volume velocity, which is integrated to obtain the glottal volume velocity. The dc level of the differentiated glottal volume velocity is removed prior to integration.

All software algorithms are extensively tested using synthesized speech data prior to being used to process real speech data. All speech data are corrected for microphone distortions by deriving a microphone correction transfer function [60].

IV. Results

Typical glottal volume-velocity waveforms obtained by inverse filtering sustained vowel tokens uttered by a male speaker are shown in Fig. 6. A closed phase interval is clearly visible. As mentioned above, the base line is arbitrary. In Fig. 6, the skewing to the right of the glottal volume-velocity waveforms is apparent. Ripples can be seen on some of the waveforms, especially /OO/. The glottal volume-velocity waveforms for the sustained vowel /A/ for both male and female speakers are shown in Figs. 7 and 8, respectively. The glottal closed phase is generally apparent for the male speakers, but less so for the female speakers. The glottal volume-velocity waveforms are more sinusoidal in shape for the female speakers than for the male speakers. Generally, the glottal volume-velocity waveforms differ among the various speakers. The modeled differentiated glottal volume-velocity and glottal volume-velocity waveforms for the vowel /A/ are shown in Fig. 9 for a male speaker. Fig. 9(a) illustrates that the instant of the opening of the vocal folds is difficult to measure from either the data or the model. The optimum location of the peak of the volume-velocity waveform for the model may not occur at the location of the peak of the volume-velocity of the measured volume-velocity waveform. This inability to align the peaks of the two waveforms is partially caused by source-tract interaction. In another study, we have compiled some statistics for the average values and the standard deviations for the four LF model timing parameters (t_a, t_c, t_e, t_p) for three voice types: modal, vocal fry, and breathy [6].

The data synthesized using the modeled glottal volume-velocity waveforms are shown in Fig. 10. The results in Fig. 10(b) are synthesized using the same vocal tract filter for both the open and closed glottal intervals. A comparison of the original (natural speech) data in Fig. 10(a) with the synthesized

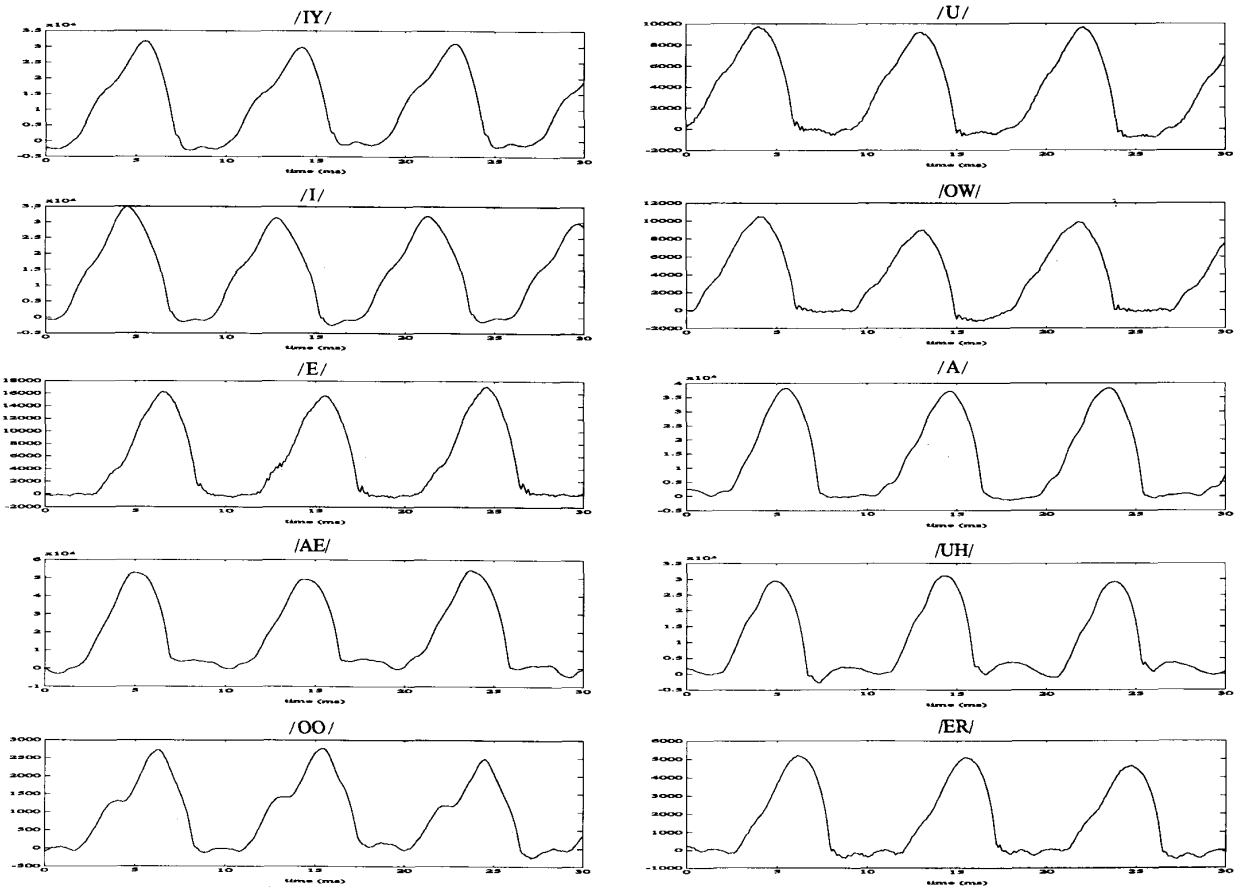


Fig. 6. Glottal volume-velocity waveforms for different sustained vowels for a male speaker.

speech data in Fig. 10(b) shows that the first formant energy of the synthesized speech does not decay during the open phase. The data in Fig. 10(c) are synthesized by increasing the first and second formant bandwidths of the vocal tract filter by a factor of four for the open interval over that for the closed interval. The damping of the formants is apparent.

V. Discussion

The results show that glottal volume-velocity waveforms vary considerably among speakers and vary among vowels for the same speaker, agreeing with previous results discussed in Section I. The glottal volume-velocity waveform model is obtained by adjusting the LF model parameters to minimize the least squared error between the data and the LF model. For synthesis, source-tract interaction is simulated by creating two effects reflected in the model waveform: a first formant (F1) ripple and skewing. The ripple effect is simulated by adjusting the first formant bandwidth of the vocal tract filter to have different values during the open and closed glottal intervals. The skewing effect is created by adjusting the LF model parameters to minimize the least squared error between the measured volume velocity waveform and the model waveform.

It has long been noted that some "ringing" or other noise-like activity may occur in the closed phase of the estimated glottal waveform obtained by inverse filtering [32], [33], [38], [39], [52]. On occasion, we see such activity in some of our data as well (Figs. 6 and 7). Several possible explanations of this phenomenon have been suggested in the literature, including acoustic interaction with the glottis [52], mucosal wave motion across the surface of the vocal folds [32], displaced glottal air [52], laryngeal adjustments [52], and nasalization of vowels [33], [52]. While this matter has not been resolved, it is likely due in part to several or all of these phenomena. However, a most common factor is acoustic interaction with the glottis, wherein, the first formant may not be completely removed during the inverse filtering procedure, thereby leaving a first formant remnant in the inverse filtered glottal waveform, which appears as a ringing type of activity in the closed phase region. This phenomenon is readily reproduced using simulated glottal waveforms in synthesized speech (Fig. 10). The activity in the closed phase region of the glottal waveform can be eliminated by adjusting the parameters of the inverse filter through user interaction with the software [32], [33]. On occasion, we found that we had to also make such adjustments to the inverse filter parameters to minimize the activity in the closed phase region. We feel that any remaining activity in the closed phase region, after such adjustments, is probably due to one of the other

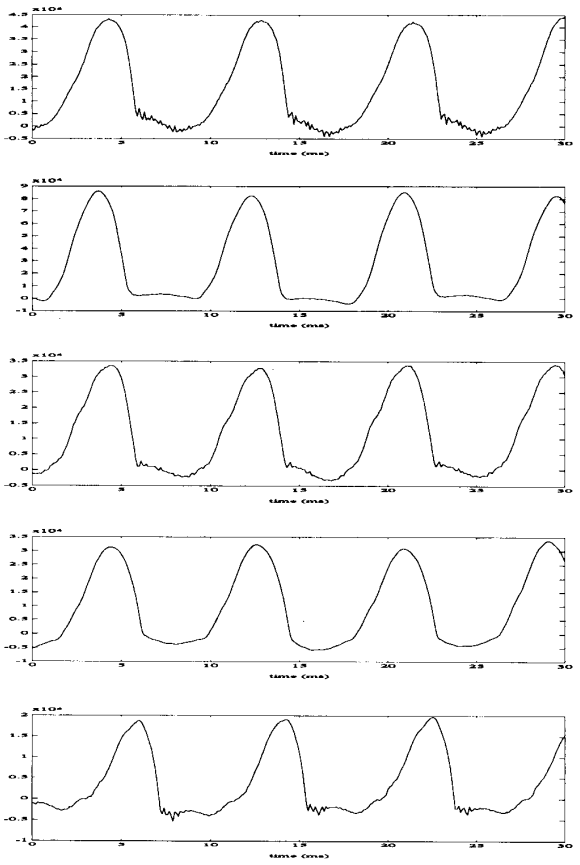


Fig. 7. Glottal flow waveforms for sustained vowel /A/ for different male speakers.

causes mentioned above, as also suggested by Holmes [32] and Hunt *et al.* [33]. However, such activity had no effect on the measurement of the glottal pulse parameters for the glottal pulse model, since the activity in the measured closed phase region of the glottal waveform is not matched to the model waveform, and, therefore, does not influence the mean squared error between the LF model pulse and the pulse determined by inverse filtering.

We use a waveform matching technique with a minimum mean squared error criterion for determining the LF model parameter values, rather than a spectrum based criterion. If a magnitude spectrum approach is used, one can obtain errors in the time domain waveform parameter values. For example, the magnitude spectrum of a pulse with a slow rise time and a fast fall time is the same as that for a pulse that is reversed in the time domain. The spectrum features that distinguish these two pulses are contained in the phase, which is not represented in the magnitude spectrum. Furthermore, if the time domain waveform features of the model are correct, then the spectral features of the model will be correct.

Informal listening tests compared the original (natural) speech with speech synthesized using the source-tract interaction model, shown in Fig. 4. The skewness of the glottal pulse and the first formant ripple were varied. The comparison of

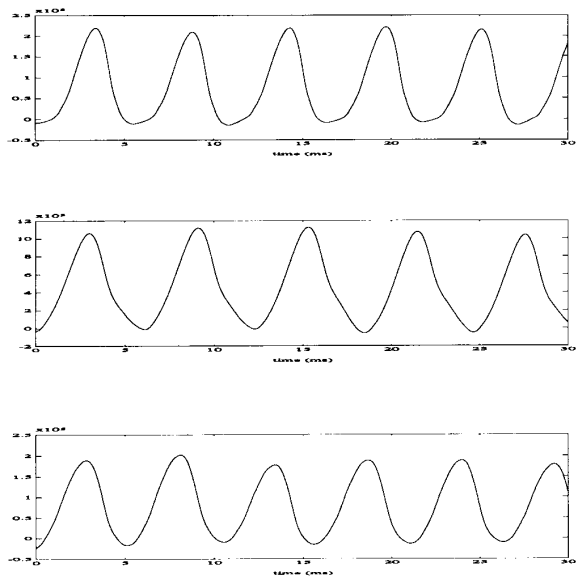


Fig. 8. Glottal flow waveforms for sustained vowel /A/ for different female speakers.

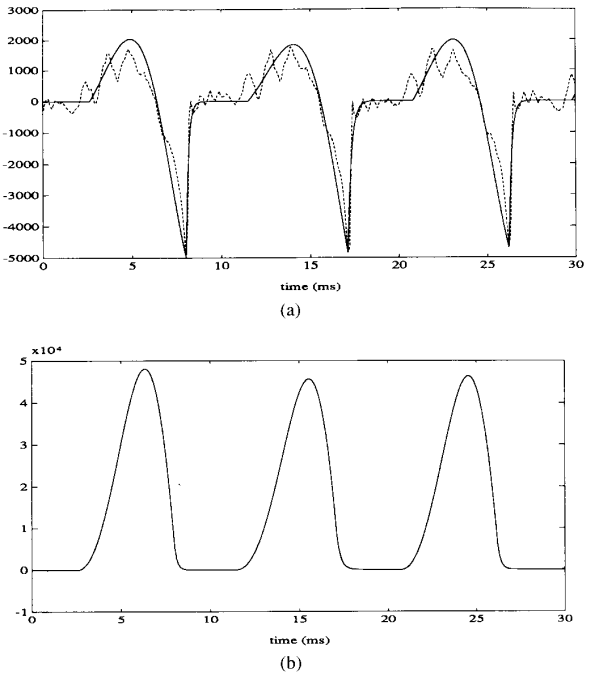
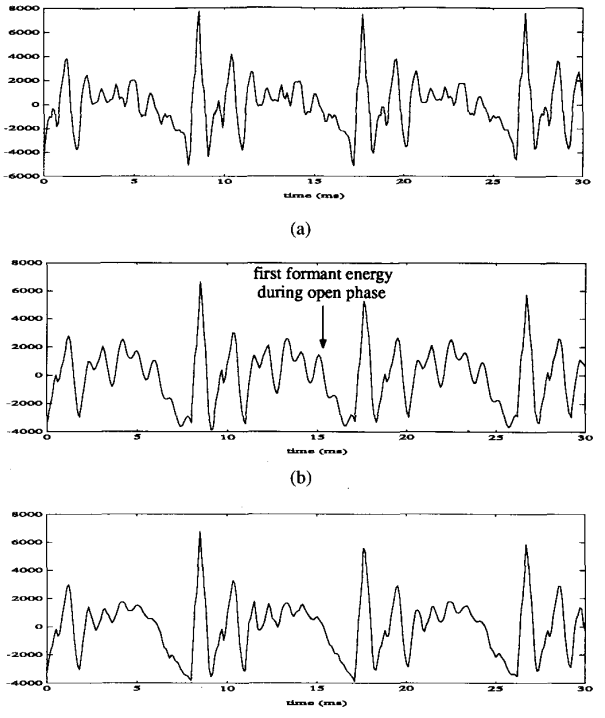


Fig. 9. LF modeled data for the vowel /A/ for a male speaker. (a) Differentiated glottal volume-velocity, where the dotted line represents the waveform to be modeled. (b) Glottal volume-velocity waveform.

original and synthesized tokens convinced us that the model is able to produce more natural sounding speech than that synthesized without the model [6]. The most important effect is the skewing of the glottal pulse, which must be modeled adequately. The effect of the ripple does not seem to be as significant, but more analysis/synthesis experiments are needed



(a) (b) (c)

Fig. 10. Sustained vowel /A/ for a male speaker. (a) Original data. (b) Synthesized data without including the ripple effects. (c) Synthesized data with simulation of ripple effects.

to confirm this observation. The analysis/synthesis procedure was tested on sentences as well as vowels. However, for sentences, minor adjustments must be made to the software program to handle the dynamics of the speech signal present in a sentence. More discussion on various aspects of the synthesis procedure appears in [6].

Our analysis procedures assume that a closed glottal interval exists and is of sufficient duration to perform a closed phase covariance LPC analysis. If the closed phase interval is short, then analysis errors may occur. This often happens with female voices or with breathy voices. Consequently, the glottal volume velocity waveforms estimated by inverse filtering may not accurately represent the true data.

A natural extension of this work would be to use the techniques developed here to quantify the first formant bandwidth differences that occur between the closed and open glottal intervals. This should be done on a large data base of speakers. The results presented here should be extended to measure and model the glottal volume velocity variations that occur in continuous speech. These procedures should be applicable to speech produced by speakers with and without a vocal disorder.

REFERENCES

[1] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal volume-velocity and its components," *Speech Commun.*, vol. 1, pp. 167–184, 1982.

[2] M. B. Berouti, D. G. Childers, and A. Paige, "Correction of tape recorder distortion," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, 1977, pp. 397–400.

[3] C. A. Bickley and K. N. Stevens, "Effects of a vocal tract constriction on the glottal source: data from voiced consonants," in *Laryngeal Function in Phonation and Respiration*, T. Baer, C. Sasaki, and K. Harris, Eds. Boston, MA: College-Hill Press, 1987, pp. 239–253.

[4] R. Carlson, B. Granstrom, and I. Karlsson, "Experiments with voice modeling in speech synthesis," *Speech Commun.*, vol. 10, pp. 481–489, 1991.

[5] Y. M. Cheng and B. Guerin, "Control parameters in male and female glottal sources," in *Laryngeal Function in Phonation and Respiration*, T. Baer, C. Sasaki, and K. Harris, Eds. Boston, MA: College-Hill Press, 1987, pp. 219–238.

[6] D. G. Childers and C. Ahn, "Modeling the glottal volume velocity waveform for three voice types," submitted to *J. Acoust. Soc. Am.*, 1994.

[7] D. G. Childers and J. C. Principe, Y. T. Ting, and K. Lee, "Adaptive WRLS-VFF for speech analysis," submitted to *IEEE Trans. Speech, Audio Processing*, 1993.

[8] D. G. Childers and K. Wu, "Quality of speech produced by analysis-synthesis," *Speech Commun.*, vol. 9, pp. 97–117, 1990.

[9] D. G. Childers, J. J. Yea, and E. L. Bocchieri, "Source/vocal-tract interaction in speech and singing synthesis," in *Proc. Stockholm Music Acoust. Conf..*, Stockholm, Sweden, 1983, pp. 125–141.

[10] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Commun.*, vol. 8, pp. 147–158, 1989.

[11] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[12] D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis," *J. Acoust. Soc. Amer.*, vol. 90, pt. 1, pp. 1841–1856, 1991.

[13] D. G. Childers, M. Hahn, and J. N. Larar, "Silent and voiced/unvoiced/mix excitation (four-way) classification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1771–1774, 1989.

[14] B. Cranen and L. Boves, "On subglottal formant analysis," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 734–746, 1987.

[15] B. Cranen and L. Boves, "On the measurement of glottal volume-velocity," *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 888–900, 1988.

[16] G. Fant, *Acoust. Theory Speech Production*, Mouton, Hague, 1960.

[17] G. Fant, "The source filter concept in voice production," *STL-QPSR* 1/1981, Royal Institute of Technology, Stockholm, Sweden, pp. 21–37, 1981.

[18] G. Fant, "Preliminaries to analysis of the human voice source," *STL-QPSR* 4/1982, Royal Institute of Technology, Stockholm, Sweden, pp. 1–27, 1982.

[19] G. Fant, "The voice source-acoustic modeling," *STL-QPSR* 4/1982, Royal Institute of Technology, Stockholm, Sweden, pp. 28–48, 1982.

[20] G. Fant and T. V. Ananthapadmanabha, "Truncation and superposition," *STL-QPSR* 2–3/1982, Royal Institute of Technology, Stockholm, Sweden, pp. 1–17, 1982.

[21] G. Fant and J. Liljencrants, "Perception of vowels with truncated intraperiod decay envelopes," *STL-QPSR* 1/1979, Royal Institute of Technology, Stockholm, Sweden, pp. 79–84, 1979.

[22] G. Fant and Q. G. Lin, "Glottal source-vocal tract acoustic interaction," *STL-QPSR* 1/1987, Royal Institute of Technology, Stockholm, Sweden, pp. 13–27, 1987.

[23] G. Fant and Q. G. Lin, "Frequency domain interpretation and derivation of glottal volume-velocity parameters," *STL-QPSR* 2–3/1988, Royal Institute of Technology, Stockholm, Sweden, pp. 1–21, 1988.

[24] G. Fant, J. Liljencrants, and Q. G. Lin, "A four-parameter model of glottal volume-velocity," *STL-QPSR* 4/1985, Royal Institute of Technology, Stockholm, Sweden, pp. 1–13, 1985.

[25] G. Fant, Q. G. Lin, and C. Gobl, "Notes on glottal volume-velocity interaction," *STL-QPSR* 2–3/1985, Royal Institute of Technology, Stockholm, Sweden, pp. 21–45, 1985.

[26] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed. New York: Springer-Verlag, 1972.

[27] B. Guerin, "Effects of the source—tract interaction using vocal fold models," *Vocal Fold Physiology: Biomechanics, Acoust. and Phonatory Control*, I. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 482–499.

[28] B. Guerin, M. Mrayati, and R. Carre, "A voice source taking account of coupling with the supraglottal cavities," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, 1976, pp. 47–50.

[29] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers

in soft, normal, and loud voice," *J. Acoust. Soc. Amer.*, vol. 84, pp. 511–529, 1988.
[30] J. N. Holmes, "An investigation of the volume-velocity waveform at the larynx during speech by means of an inverse filter," in *Proc. 4th Int. Congr. Acoust.*, Copenhagen, Denmark, 1962, pp. 1–4.
[31] J. N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 298–305, 1973.
[32] J. N. Holmes, "Formant excitation before and after glottal closure," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 39–42, 1976.
[33] M. J. Hunt, J. S. Bridle and J. N. Holmes, "Interactive digital inverse filtering and its relation to linear prediction methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, OK, 1978, pp. 15–18.
[34] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Tech. J.*, vol. 51, pp. 1233–1268, 1972.
[35] I. Karlsson, "Glottal wave forms for normal female speakers," *J. Phon.*, vol. 14, pp. 415–419, 1986.
[36] I. Karlsson, "Glottal waveform parameters for different speaker types," in *1988 Proc. Speech, 7 FASE Symposium*, vol. 1, pp. 225–231, 1988.
[37] I. Karlsson, "Voice source dynamics for female speakers," in *Proc. Int. Conf. Spoken Language*, vol. 1, pp. 69–72, 1990.
[38] I. Karlsson, "Female voices in speech synthesis," *J. Phon.*, vol. 19, pp. 111–120, 1991.
[39] I. Karlsson, "Modeling voice variations in female speech synthesis," *Speech Commun.*, vol. 11, pp. 491–495, 1992.
[40] D. H. Klatt, "Acoustic correlates of breathiness: first harmonic amplitude, turbulence noise, and tracheal coupling," *J. Acoust. Soc. Amer.*, suppl. 1, vol. 82, p. S91, 1987.
[41] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 80, pp. 820–857, 1990.
[42] T. Koizumi, S. Taniguchi, and S. Hiromitsu, "Glottal source-vocal tract interaction," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1541–1547, 1985.
[43] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 730–743, 1986.
[44] Q. G. Lin, "Nonlinear interaction in voice production," *STL-QPSR* 1/1987, Royal Institute of Technology, Stockholm, Sweden, pp. 1–12, 1987.
[45] P. Milenkovic, "Glottal inverse filtering by joint estimation of an AR system with a linear input model," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 28–42, 1986.
[46] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Amer.*, vol. 31, pp. 667–677, 1959.
[47] R. B. Monsen and A. M. Engebretson, "Study of variations in the male and female glottal wave," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 981–993, 1977.
[48] L. Nord, T. V. Ananthapadmanabha, and G. Fant, "Signal analysis and perceptual tests of vowel responses with an interactive source filter model," *STL-QPSR* 2–3/1984, Royal Institute of Technology, Stockholm, Sweden, pp. 25–52, 1984.
[49] P. J. Price, "Male and female voice source characteristics: inverse filtering results," *Speech Commun.*, vol. 8, pp. 261–277, 1989.
[50] L. R. Rabiner and R. W. Schafer, *Digital Processing Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
[51] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, 1971.
[52] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air volume-velocity waveform during voicing," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1632–1645, 1973.
[53] M. Rothenberg, "Acoustic interaction between the glottal source and the vocal tract," *Vocal Fold Physiology*, K. N. Stevens and M. Hirano, Eds. Tokyo: Univ. of Tokyo Press, 1981, pp. 305–323.
[54] M. Rothenberg, "An interactive model for the voice source," *STL-QPSR* 4/1981, Royal Institute of Technology, Stockholm, Sweden, pp. 1–17, 1981.
[55] M. Rothenberg, "Source-tract acoustic interaction in breathy voice," *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, I. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 465–481.
[56] M. Rothenberg, "Nonlinear source-tract acoustic interaction in the soprano voice and some implications for the definition of vocal efficiency," *Laryngeal Function in Phonation and Respiration*, T. Baer, C. Sasaki, and K. Harris, Eds. Boston, MA: College-Hill, 1987, pp. 254–269.
[57] M. Rothenberg and S. Zahorian, "Nonlinear inverse filtering technique for estimating the glottal-area waveform," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1063–1071, 1977.
[58] M. M. Sondhi, "Measurement of the glottal waveform," *J. Acoust. Soc. Amer.*,, vol. 57, pp. 228–232, 1975.
[59] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 369–377, 1985.
[60] C. F. Wong, "The incorporation of glottal source-vocal tract interaction effects to improve the naturalness of synthetic speech," Ph.D. dissertation, Univ. Florida, Gainesville, 1991.
[61] D. Y. Wong and J. D. Markel, "An excitation function for LPC synthesis which retains the human glottal phase characteristics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, OK, 1978, pp. 171–174.
[62] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 350–355, 1979.
[63] K. Wu and D. G. Childers, "Gender recognition from speech. Part I: Coarse analyses," *J. Acoust. Soc. Amer.*, vol. 90, pt. 1, pp. 1828–1840, 1991.

Donald G. Childers (S'56–M'59–SM'65–F'76) received the B.S., M.S. and Ph.D. degrees in 1958, 1959, and 1964, respectively, from the University of Southern California.

He has seven years experience in industry and has been a Consultant continuously since becoming a Professor. His research and teaching have been recognized through several awards. For the last 28 years, he has been a Professor at the University of Florida where he directs the multidisciplinary Mind-Machine Interaction Research Center. Research activities within the center include speech analysis, synthesis, and recognition. He is interested in all aspects of signal processing. He has published over 100 papers, numerous book chapters, authored or coauthored three books.

Dr. Childers is a Fellow of the Acoustical Society of America and a member of several other technical societies. He has served the IEEE in a number of capacities including four years on the editorial board of the IEEE Press and four years as Chief Editor of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING.

Chun-Fan Wong (S'90–M'91) was born in Hong Kong July 15, 1961. He graduated from the University of Hong Kong, Hong Kong, in November 1983, with the B.Sc. degree in physics. His minor was computer science. He went on to get the M.Sc. degree in modern electronics from the University of Nottingham, United Kingdom, in December 1984. Since January 1985, he has been with the department of electrical engineering, University of Florida, Gainesville. He expects to receive the Ph.D. degree in May 1991.

Mr. Wong is a member of the Tau Beta Pi Engineering Honor Society and he is also an associate member of the IEE (UK).