

# Modeling the glottal volume-velocity waveform for three voice types

D. G. Childers<sup>a)</sup>

405 CSE, Department of Electrical Engineering, University of Florida, Gainesville, Florida: 32611-2024

Chieteuk Ahn

Electronics Telecommunications Research Institute, Video Communications, P. O. Box 8, Daeduk Science Town, Daejeon, 305-606 Korea

(Received 18 January 1994; accepted for publication 17 August 1994)

The purpose of this study was to model features of the glottal volume-velocity waveform for three voice types: modal voice, vocal fry, and breathy voice. The study analyzed data measured from two sustained vowels and one sentence uttered by nine adult, male subjects who represented examples of the three voice types. The primary analysis procedure was glottal inverse filtering, which estimated the glottal volume-velocity waveform. The estimated glottal volume-velocity waveform was then fit to an LF model waveform. Four parameters of the LF model were adjusted to minimize the mean-squared error between the estimated glottal waveform and the LF model waveform. Statistical averages and standard deviations of the four parameters of the LF glottal waveform model were calculated using the data for each voice type. The four LF model parameters characterize important low-frequency features of the glottal waveform, namely, the glottal pulse width, pulse skewness, abruptness of closure of the glottal pulse, and the spectral tilt of the glottal pulse. Statistical analysis included ANOVA and multiple linear regression analysis. The ANOVA results demonstrated that there was a difference in three of the four LF model parameters for the three voice types. The linear regression analysis between the four LF model parameters and a formal rating by a listening test of the quality of the three voice types was used to determine the most significant LF model parameters for each voice type. A simple rule was devised for synthesizing the three voice types with a formant synthesizer using the LF glottal waveform model. Listener evaluations of the synthesized speech tended to confirm the results determined by the analysis procedures.

PACS numbers: 43.70.Dn, 43.70.Gr

## INTRODUCTION

Fant's linear model of speech production has contributed to the advancement of speech analysis, synthesis, and coding (Fant, 1960). Until recently, the primary research interest in this model focused on the vocal tract filter characteristics. Models for the source were not given as much attention, even though early work showed that the glottal pulse shape was important for synthesizing natural sounding vowels (Rosenberg, 1971; Holmes, 1973). In the last few years more emphasis has been given to the characteristics of the glottal source waveform, both for speech synthesis and for modeling voice types and vocal disorders (Carlson *et al.*, 1991; Childers *et al.*, 1989b; Childers and Wu, 1990; Childers and Lee, 1991; Childers and Wong, 1994; Fant, 1993; Fant and Lin, 1988; Fujisaki and Ljungqvist, 1986; Klatt and Klatt, 1990; Karlsson, 1986, 1988, 1990, 1991, 1992; Pinto *et al.*, 1989). One study reported that four factors were important for characterizing the glottal excitations for four voice types (Childers and Lee, 1991). The four factors were the glottal pulse width, the glottal pulse skewness, the abruptness of glottal closure, and the turbulence noise component. The significance of these factors for voice synthesis was examined in that study and a voice source model was developed that

could account for certain glottal volume-velocity waveform features that were considered characteristic of the different voice types.

For this study we hypothesized that a simple glottal waveform model could characterize attributes of the glottal volume-velocity waveform for three voice types, namely, modal (a vocal register), vocal fry (a vocal register), and breathy voice. The subjects for this study were selected from an earlier study that examined some acoustic correlates of vocal quality (Eskenazi *et al.*, 1990). To validate the hypothesis, we estimated the glottal volume-velocity waveform (glottal pulse or glottal flow) for each subject for each voice type by inverse filtering. The estimated waveform was then compared to an LF glottal model waveform (Fant *et al.*, 1985; Fant, 1993) for each pitch period of analyzed data. Four parameters of the LF model were adjusted to minimize the mean-squared error between the LF model waveform and the glottal waveform estimated by inverse filtering. These four parameters model the glottal source low-frequency features, which are primarily determined by the glottal pulse width, pulse skewness, abruptness of closure of the glottal pulse, and the spectral tilt of the glottal pulse. While the LF pulse model does contribute to the high-frequency portions of the spectrum, the turbulence noise (such as aspiration) often dominates in this region. This study focused on the four LF model parameters and did not investigate other fea-

<sup>a)</sup>E-mail: childers@drwho.ee.ufl.edu, Telephone: (904) 392-2633.

tures of the glottal excitation, such as,  $f_0$ , turbulence noise, jitter, and shimmer. ANOVA determined that three of the four LF model parameters were different among the three voice types. Multiple linear regression analysis determined the most significant of the four LF model parameters. This was achieved by predicting the listener's ratings of the quality of each of the three voice types from the four LF model parameters. Statistical averages and standard deviations for the four LF model parameters were determined for each voice type. In summary, the purpose of the paper is to illustrate that the methodology of using the LF source model and speech synthesis techniques is a useful procedure for modeling and synthesizing aspects of three voice types.

## I. PROCEDURES

### A. Database

All data recordings were performed in an Industrial Acoustics Company single-wall sound room. The speech and electroglottographic (EGG) signals were monitored simultaneously. One of two microphones was used: an Electro-Voice RE-10 dynamic cardioid or a Bruel and Kjaer model 4113 condenser. The selected microphone was located 6 in. from the speaker's lips. The electroglottograph was a Synchro-voice, Inc. model. All data were directly digitized, thereby avoiding any low-frequency distortions that may have been introduced through the use of audio tape recordings. The speech and EGG signals were bandlimited to 5 kHz by anti-aliasing elliptic filters with a minimum stop-band attenuation of  $-55$  dB and a passband ripple of  $\pm 0.2$  dB. Both signals were amplified by a Digital Sound Corp. DSC-240 audio control console. The two signals were sampled at 10 kHz per channel by a Digital Sound Corp. DSC-200 analog-to-digital system with 16-bit resolution. The data that were recorded using the Electro-Voice microphone were corrected for microphone distortions by deriving a microphone correction transfer function (Childers and Wong, 1994). The data that were recorded with the Bruel and Kjaer microphone did not require correction since its bandwidth characteristics were sufficiently broad that no frequency distortions were introduced into the data. The experimental speaking tasks were two sustained vowels: /i/ and /a/ and the all-voiced sentence "We were away a year ago." The vowel tokens were about 2 s in duration, while the sentence was approximately 1.5 s. All data were analyzed in this study. The vocal intensity was not controlled. Each subject phonated at a comfortable pitch and intensity level. Intensity was not considered a factor because all signals analyzed were approximately the same magnitude after digitization. No recording nor postrecording amplification adjustments of gain were made. However, to help insure that the data recording level was not a factor in this study, we normalized the energy of all the inverse filtered differentiated glottal volume-velocity waveforms to unity prior to fitting these waveforms to the LF model waveform.

A factor in a study of this nature is to establish the representatives of the voice qualities (or voice types) selected for analysis. We addressed this issue in Eskenazi *et al.* (1990), where a panel of seven listeners (four males and three females) served as judges to rate the quality of both

pathological and normal voices. From that study we selected three male subjects each for modal, vocal fry, and breathy voice that the seven judges had rated as representative of these three voice types.

## B. Inverse filtering

### 1. Overview of the algorithm

The speech signal was parsed into voiced and unvoiced segments using the EGG signal. Only the voiced speech segments were inverse filtered for this study. Next, the closed phase region for each pitch period was identified using the EGG signal. Then pitch synchronous, closed phase, covariance linear prediction (LP) analysis was performed over the closed phase interval. The inverse filter was derived from the LP coefficients by selecting only the appropriate poles and zeros (Krishnamurthy and Childers, 1986; Childers and Lee, 1991; Childers and Wong, 1994). Modifications (discussed below) to the above procedure were undertaken if there was no closed phase interval or if it was too short.

### 2. Details of the algorithm

The inverse filtering algorithm was implemented to use both the speech and EGG signals (Krishnamurthy and Childers, 1986; Childers and Lee, 1991; Childers and Wong, 1994). A frame of the speech signal was first identified as voiced or unvoiced through the use of the differentiated EGG (DEGG) signal (Childers *et al.*, 1989a). Since the analysis was pitch synchronous, each frame corresponded to a pitch period. For each voiced frame, the pitch period, the instant of the opening of the glottis (the starting point of the frame), the instant of the peak of the glottal flow, the instant and the maximum magnitude of the negative minimum of the differentiated glottal flow, the instant of the closing of the glottis, and the beginning and ending of the closed phase interval were computed.

It is known that voiced sounds have large negative minima in the DEGG corresponding to the instant of vocal fold closure, so a negative threshold was used to locate these minima. The interval between the minima of the DEGG waveform gives the pitch period (Childers *et al.*, 1989a; Childers *et al.*, 1990). Voicing was considered to start when two successive minima fell below the negative threshold and the pitch period was in a range of 25–200 samples at a 10-kHz sampling rate (frequency range of 50–400 Hz). When the above two conditions were not met, the corresponding segments were considered as unvoiced. The instant of the opening of the glottis was the positive peak of the DEGG located between two negative minima. The interval between the negative minimum peak and the positive maximum peak was the closed glottal interval. These points are illustrated in Fig. 1. From a pilot study of the DEGG data records for each subject, the negative threshold for the DEGG minima was determined empirically to be approximately 1/6 of the peak-to-peak value of the DEGG signal. This is approximately at the  $-1000$  level for the DEGG in Fig. 1. Since the positive peak of the DEGG is sometimes noisy, a false detection of this peak can occur. The criterion we use to determine if a false detection has occurred is based

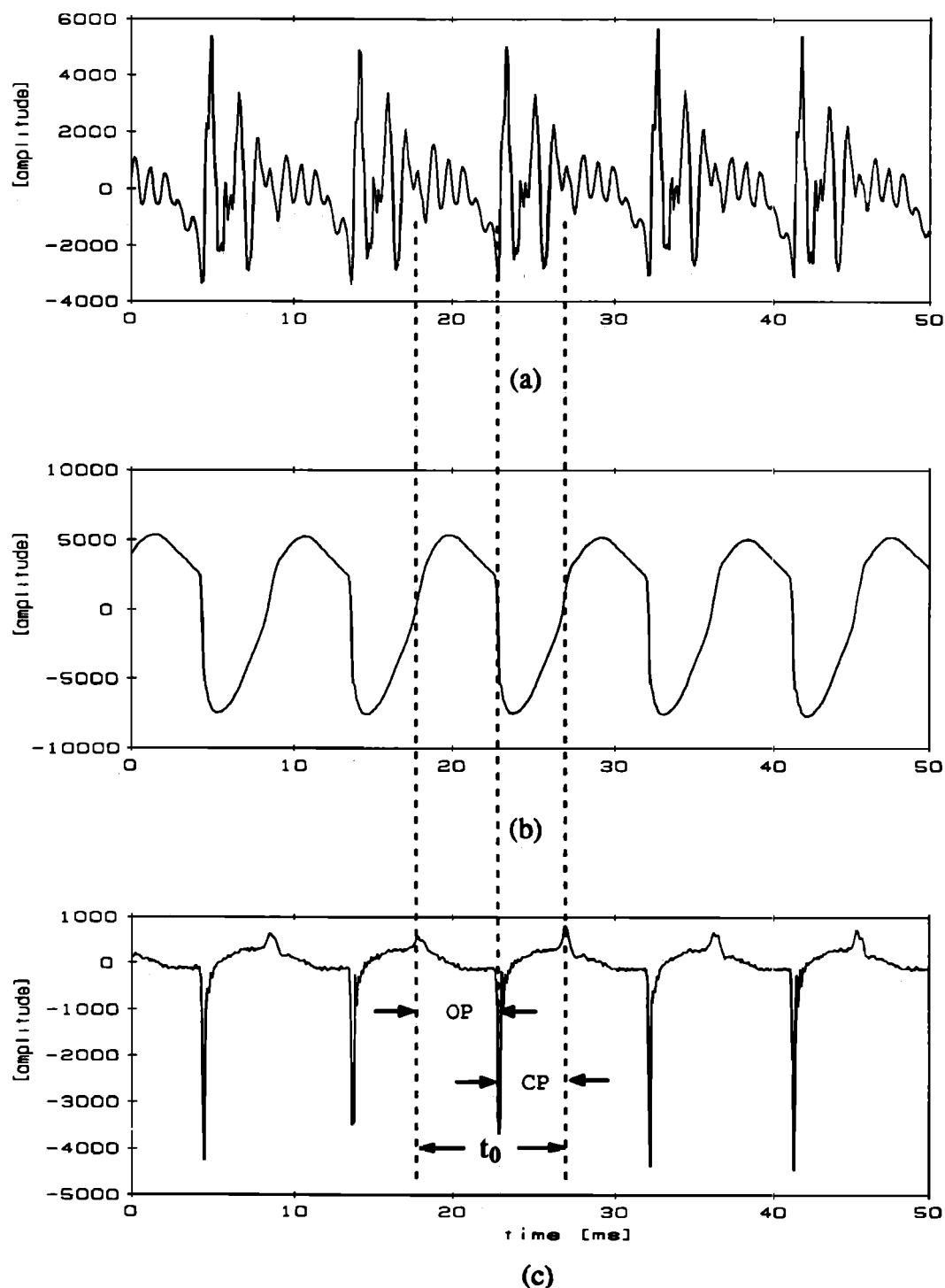


FIG. 1. Speech signal (top trace), EGG Signal (middle trace), differentiated EGG (DEGG) signal (bottom trace).  $t_0$ : pitch period, OP: open phase, and CP: closed phase.

on the pitch period. Within the interval between two successive positive peaks of the electroglottograph lies the largest positive peak and the largest negative peak of the DEGG. If the interval between these latter two peaks was less than one fourth of the average pitch period, then the data for that pitch period was discarded. This did not occur very often and we felt that since there was sufficient data for each subject, it was better to discard the data for a few pitch periods than to search for the correct peak values within that interval.

If a sufficient closed phase interval did not exist to allow

a covariance LP analysis, then we used the LP coefficients calculated from the previous frame. This frequently occurred for breathy voices. If no previous LP coefficients were available (such as might occur at the initiation of processing), we performed an autocorrelation LP analysis over the entire pitch period.

The typical order of the LP analysis was 12; however, this value was adjustable through an interactive user command option designed into the software. The minimum window size for the covariance LP analysis was 28 samples. The

fixed-frame LP analysis procedure was repeated numerous times as follows. The LP analysis was initiated with the window being placed at the beginning of the closed glottal interval. Provided that the LP analysis window did not exceed the bounds of the closed glottal interval, the first LP analysis calculation was performed. Next the analysis window was shifted one sample value along the closed glottal interval, and the next LP analysis was performed, and so on. This repeated analysis procedure was terminated when the data within the LP analysis window was less than 28 samples. Numerous sets of LP coefficients were generated for each frame using this procedure. The set of LP coefficients that provided the minimum total squared prediction error was selected for that particular analysis frame. These LP coefficients determined the linear prediction polynomial. To determine the inverse filter, the formant frequencies and bandwidths of the poles were calculated by factoring the linear prediction polynomial. The real poles at the origin were removed, because the vocal tract was assumed to consist of resonators only. Real poles at one-half the sampling frequency, however, was retained. Extraneous resonances at very low frequencies, or with very large bandwidths, were removed. Finally, the inverse filter was then reconstructed from the poles that remained. The minimum variance differentiated glottal volume-velocity waveform over the closed phase interval was obtained by inverse filtering the nonpre-emphasized speech signal. No inverse filtered waveforms obtained by this method were rejected. We also compared our method to one that used a closed phase flatness measure, which provides the minimum variance glottal volume-velocity waveform over the closed phase region bounded by the EGG (Childers and Wong, 1994). Both methods gave similar results.

Since we did a frame-by-frame analysis, it was possible for a particular frame to have a dc component because each frame was only a fragment of the total signal. The dc level of the differentiated glottal flow within each frame was removed and the resulting differentiated glottal flow was normalized to have unity energy over the frame interval. Thus the sum of squares of the data samples for each frame were set equal to one for each pitch interval.

We tested the algorithm with synthesized speech and found that the mean-squared error was, for all practical purposes, equal to zero (Ahn, 1991; Krishnamurthy and Childers, 1986; Childers and Lee, 1991; Childers and Wong, 1994). For natural speech the results of the algorithm are illustrated in Fig. 2, which shows an inverse filtered glottal flow waveform for a sustained vowel /a/ phonated by several subjects for the three voice types.

### C. Measurement of LF model parameters

The LF model (summarized in Appendix I) was fit to the measured differentiated inverse filtered waveform in a manner similar to that shown of Fig. 3 for a sustained vowel /a/. The procedure was as follows. First, the parameter  $t_c$  was set to a first approximate value. The parameter  $t_c$  of the LF model was defined for this study as the instant at which the amplitude of the differentiated glottal flow falls to 1% of its

maximum negative value. Thus the parameter  $t_c$  was an approximation to the closing instant that was measured from the data in a reliable manner. Next, the first approximate values for  $t_e$  and  $E_e$  were measured from the inverse filtered differentiated glottal flow waveform for each pitch period. The remaining parameters were determined using an iteration procedure as follows. An estimate for  $t_a$  and  $\epsilon$  was obtained by minimizing the total squared error between the inverse filtered differentiated glottal flow waveform and the LF model waveform given by Eq. (A2) over the interval from  $t_e$  to  $t_c$ . This was done by using Eq. (A2) with  $t = t_e$ , so that  $\epsilon t_a = 1 - \exp[-\epsilon(t_c - t_e)]$ . (For small values of  $t_a$ ,  $\epsilon$  is approximately equal to  $1/t_a$ .) Thus we adjusted the parameters  $t_a$  and  $\epsilon$  repeatedly until the exponential model approximated the data with the least total squared error over the interval  $t_e$  to  $t_c$ . Next, we adjusted the model to the data for the interval 0 (the opening of the glottis) to the instant  $t_e$ . This was done by searching for the best values for the parameters  $t_p$ ,  $E_0$ ,  $\alpha$ , and  $\omega_g$ . [Note that  $\omega_g = \pi/t_p$ ,  $E_0$  is found from Eq. (A1), and  $\alpha$  can be found from  $E_0 = -E_e/e^{\alpha t_e} \sin(\omega_g t_e)$ .] Trial values for these parameters were used to produce a first approximation model of the differentiated glottal flow waveform for the interval 0 to  $t_e$ . The total squared error between this approximate LF model waveform and the measured data waveform was calculated for the first set of parameters. The values of the parameter set were then repeatedly varied until the total squared error between the LF model and the data was minimized.

### D. Estimation of spectral tilt

By convention the spectral tilt or slope for a voiced phonation was determined by the combined contribution of the spectrum of the glottal pulse and the lip radiation. The general spectral tilt of the glottal flow waveform can be represented as a low-pass filter with multiple real poles. While the glottal spectral characteristics for modal and vocal fry voices could be modeled by a two-pole model ( $-12$  dB/octave), an extra pole was usually required for breathy phonations (Klatt and Klatt, 1990; Childers and Lee, 1991). The extra pole resulted in a steeper spectral slope ( $-18$  dB/octave). We adopted the following three-pole model to estimate the spectral tilt for the glottal volume-velocity (flow) waveform:

$$U_g(z) = \frac{K}{(1 - z_0 z^{-1})(1 - z_b z^{-1})(1 - z_c z^{-1})}, \quad (1)$$

where  $K$  is a constant related to the amplitude of the glottal flow and  $z_a$ ,  $z_b$ , and  $z_c$  are real poles inside the unit circle in the  $z$  domain, where each  $z$  parameter contributes  $-6$  dB/octave slope to the spectral tilt. We may simplify this representation by noting that the value of  $z_a$  is approximately the same value as the zero that is used to represent the lip radiation. Consequently, the  $z_a$  pole was canceled by the lip radiation zero for this study. The values for the two remaining real poles were estimated using the procedure given in Childers and Lee (1991). For this study we did not examine the possibility that the spectral tilt might change with fundamental frequency for a given voice type.

## II. RESULTS

### A. Spectral tilt

Table I shows the coefficients for the real-pole glottal models estimated for the three voice types for both the inverse filtered and the LF modeled differentiated glottal flow waveforms. The data for the three subjects for each voice type were combined for this table and for all subsequent calculations. The ranking of voice type according to increas-

ing spectral tilt was vocal fry, modal, and breathy. Thus the larger the spectral tilt, the steeper the spectral slope. Our results show that the spectral tilt estimated from the LF modeled differentiated glottal flow waveform is, on the average, larger than that calculated from the inverse filtered waveform. This is attributed to the lack of high-frequency energy in the LF model as compared to the real data. We compensated the model by adjusting the parameter  $t_a$  of the LF

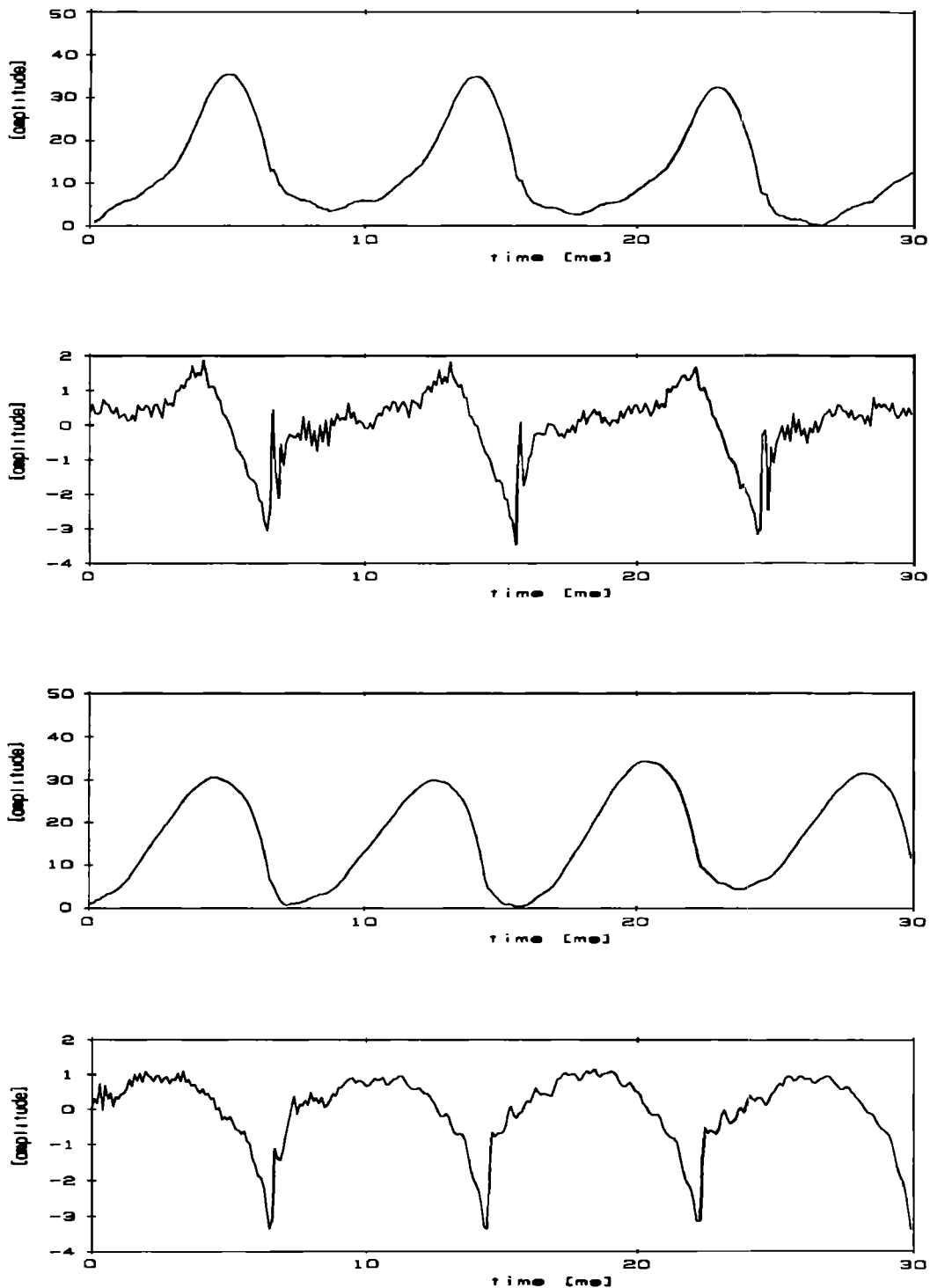


FIG. 2. Pairs of glottal flow and normalized differentiated glottal flow waveforms for different voice types and different male subjects: (a) modal voice (first two pairs, subjects 1 and 2), (b) vocal fry (second two pairs, subjects 4 and 5), and (c) breathy voice (third two pairs, subjects 7 and 9).

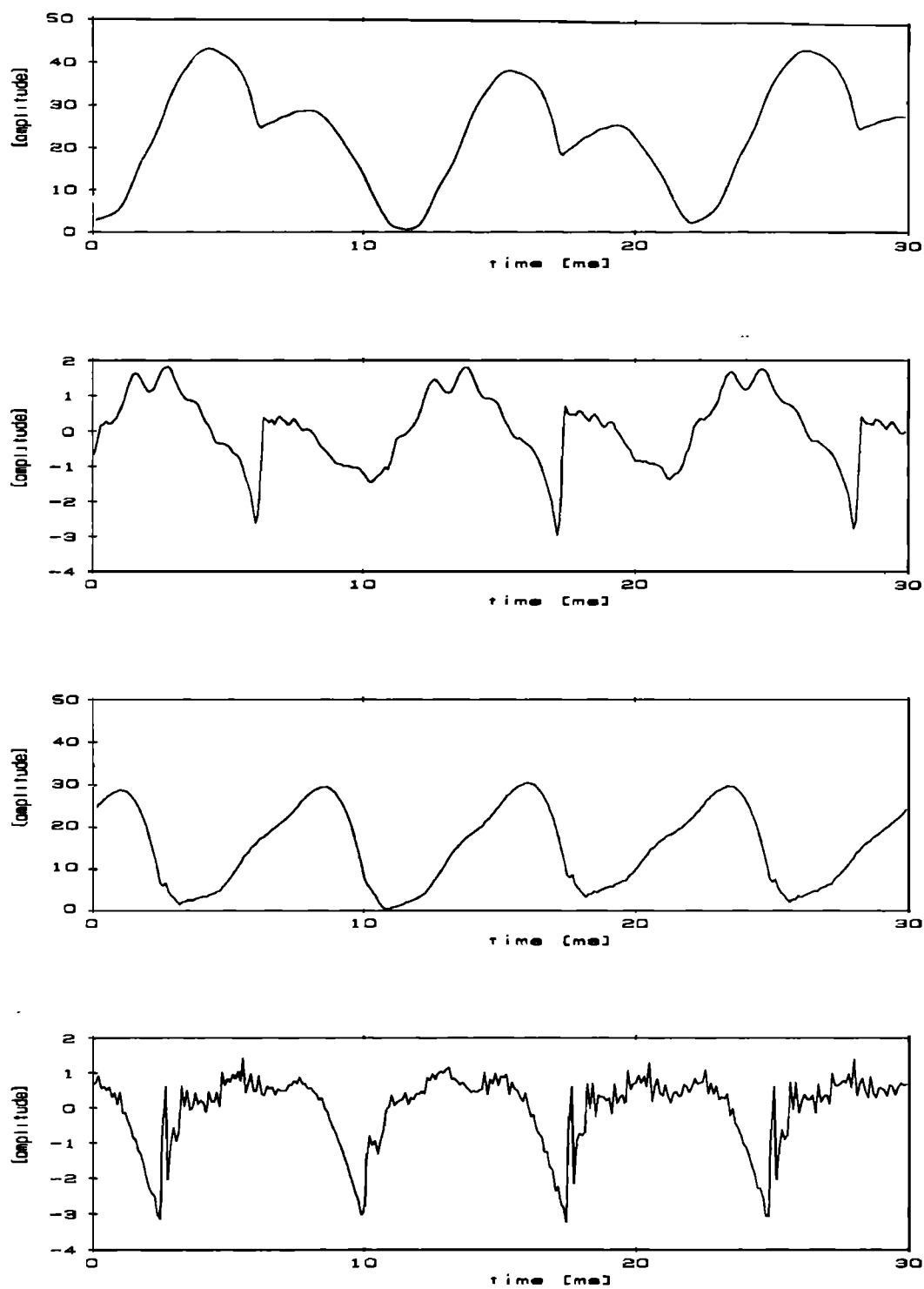


FIG. 2. (Continued.)

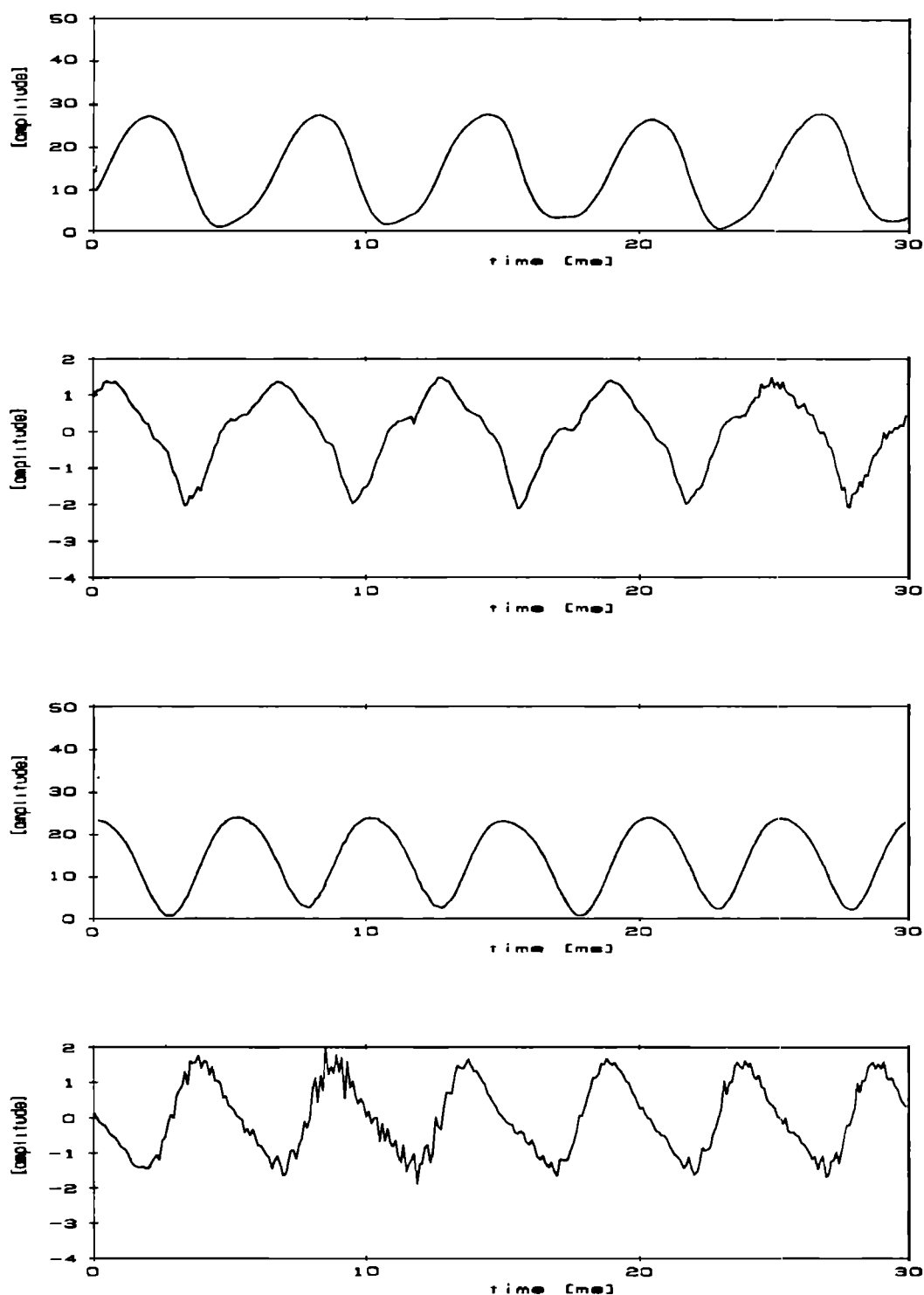


FIG. 2. (Continued.)

model. The algorithm for compensating the glottal model is described in Appendix II and is used throughout the remainder of the paper.

### B. LF model parameter values

The mean values and standard deviations (s.d.) of the spectral tilt compensated normalized LF model parameters for each voice type are tabulated in Table II and shown in Fig. 4. Note that (1) all mean values and standard deviations

are expressed as a percentage (%) of the pitch period (pp), which is denoted as normalized, (2)  $t_c$  was computed from the LF model, (3) the speed quotient  $SQ_{LF}$  for the LF model (the ratio of the glottal open phase to the closing phase) was computed as  $SQ_{LF} = t_p / (t_c - t_p)$ , and (4)  $f_0$  was computed as  $f_0 = 1/pp$ . The total number of frames (pitch periods) analyzed is specified in Table II. In Table II,  $t_c$  is an approximation to the open quotient (the ratio of the open phase to the pitch period) for the LF model, which was computed as  $OQ_{LF} = t_c/pp$ , because all timing parameters were normal-

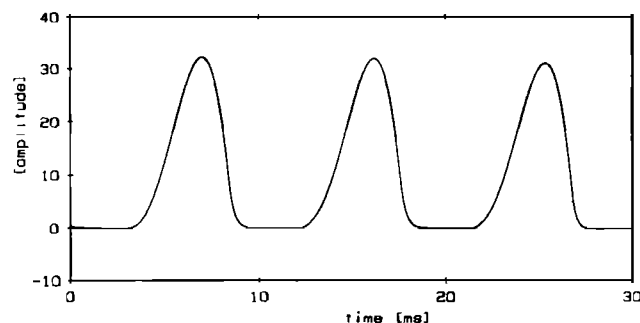
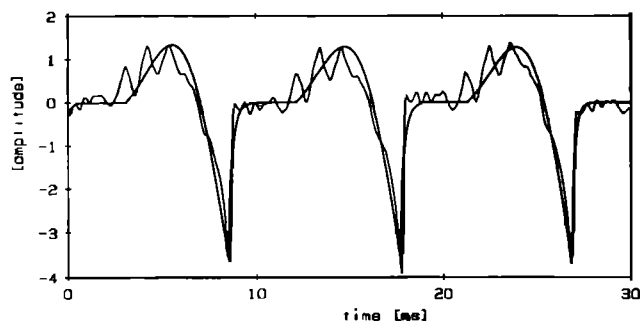


FIG. 3. LF-modeled data for the vowel /a/ for male subject 1: normalized differentiated glottal flow (top trace), glottal flow (bottom trace).

ized with respect to the corresponding pitch period. The  $SQ_{LF}$  was also computed for each analysis frame. We verified the measurements for the pitch period ( $pp$ ) using the EGG signal. It appears that the mean value of the normalized parameter  $t_c$ , which is equivalent to the open quotient ( $OQ_{LF}$ ) in this study, is a potential feature for distinguishing the three voice types. The mean value of the speed quotient ( $SQ_{LF}$ ) is comparable for modal and vocal fry phonations, while it is smaller for a breathy voice. 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, and  $t_e$  to the instant of the main excitation. Glottal pulse skewness may be represented by the speed quotient.

TABLE I. Mean values and standard deviations (s.d.) for the spectral tilt estimated for the three voice types from both the inverse filtered differentiated glottal (DG) flow and the LF modeled differentiated glottal (LFMDG) flow waveforms.

Voice type (number of pitch periods analyzed)	DG		LFMDG	
	$z_b$ (s.d.)	$z_c$ (s.d.)	$z_b$ (s.d.)	$z_c$ (s.d.)
Modal (1294)	0.884 (0.207)	0.070 (0.195)	0.959 (0.049)	0.372 (0.338)
Vocal fry (1708)	0.797 (0.269)	0.047 (0.160)	0.941 (0.064)	0.198 (0.294)
Breathy (848)	0.887 (0.299)	0.396 (0.339)	0.978 (0.042)	0.690 (0.241)

TABLE II. Mean values and standard deviations (s.d.) for the compensated LF model parameters for the three voice types.  $t_p$ ,  $t_e$ ,  $t_a$ , and  $t_c$  are in percentage of pitch period (pp),  $t_c$  was computed from the LF model,  $SQ_{LF} = t_p / (t_c - t_p)$ , and  $f_0 = 1/pp$ .

Voice type <sup>a</sup>	$t_p$ (%)	$t_e$ (%)	$t_a$ (%)	$t_c$ (%)	$SQ_{LF}$	$pp$ (ms)	$f_0$ (Hz)
Modal (1294)	41.34 (5.49)	55.30 (7.77)	0.41 (0.92)	58.17 (8.84)	2.80 (1.33)	8.51 (0.92)	118.63 (11.16)
Vocal fry (1708)	48.08 (17.81)	59.55 (17.76)	2.69 (2.20)	72.00 (21.66)	2.34 (1.08)	10.63 (2.55)	101.26 (30.82)
Breathy (848)	46.21 (11.01)	66.04 (16.14)	2.70 (2.08)	77.12 (15.27)	1.62 (0.71)	9.12 (1.81)	114.28 (27.96)

<sup>a</sup>Number of pitch periods analyzed.

## C. Statistical analysis

Our hypothesis is that there is a significant difference in at least one parameter of the LF model among the three voice types. To demonstrate that this hypothesis is valid we did an analysis of variance (ANOVA) with a nested design. The independent variable is voice type with three subjects for each type. For modal voice there are 1294 observations, 1708 observations for vocal fry, and 848 observations for breathy, giving a total of 3850 observations. Four one-way ANOVAs were run, one for each of the four LF parameters. The results are as follows: for  $t_p$ ,  $F(2,6)=2.64$ ,  $p<0.1508$ ; for  $t_e$ ,  $F(2,6)=3.75$ ,  $p<0.0880$ ; for  $t_a$ ,  $F(2,6)=11.40$ ,  $p<0.0090$ ; for  $t_c$ ,  $F(2,6)=12.60$ ,  $p<0.0071$ . Thus three ( $t_e, t_a, t_c$ ) LF model parameters are statistically different across the three voice types, while  $t_p$  is marginally different.

Since the ANOVA did not consider the average quality rating provided by the seven judges, we also conducted a multiple linear regression analysis. The selection criterion was the prediction sum of squares (PRESS) (Allen and Cady, 1982; Eskenazi *et al.*, 1990). The four LF model parameters ( $t_p, t_a, t_e, t_c$ ) were the predictors and the average quality rating provided by the seven judges for each voice type was the criterion. The best PRESS model (i.e., the model with the lowest PRESS value) was adopted. We also calculated the square of the multiple linear correlations ( $R^2$ ) for each model

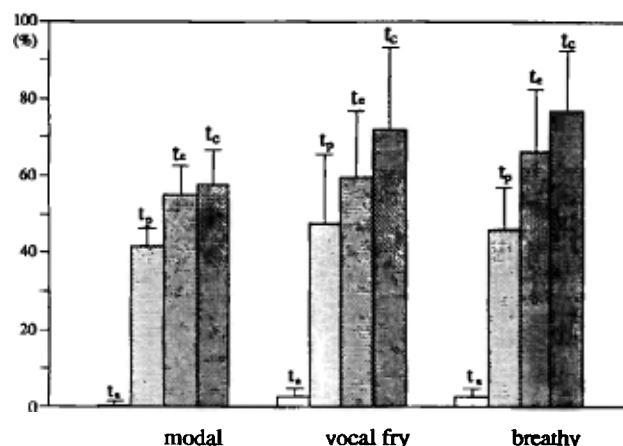


FIG. 4. Mean values and standard deviations for the compensated, normalized LF model parameters for the different voice types.



TABLE III.  $R^2$  values for multiple linear regression for the LF model parameters for the prediction of voice type with PRESS as the selection criterion.

Voice type	LF parameters			
	$t_p$	$t_e$	$t_a$	$t_c$
Modal	0.70	0.70	0.72	0.69
Vocal fry	0.42	0.60	0.36	0.37
Breathy	0.26	0.34	0.39	0.30

in multiple linear regression. The LF model parameters were calculated for successive pitch periods (i.e., over time), and therefore may be considered to represent a time-varying vector of LF parameter values. However, the average quality ratings provided by the seven judges represented a single value judgement for the entire data record for each subject and each voice type. Consequently, for the multiple linear regression analysis we averaged the LF parameter values for all three subjects for each voice type. Thus we calculated a multiple linear regression model for each voice type. The results for the one parameter multiple linear regression models for modal, vocal fry, and breathy voice types are summarized in Table III. These results were obtained by calculating four successive first-order models, i.e., we calculated  $y = b_0 + b_1 t_i$  four times with  $i = p, e, a, c$  for the LF model parameters for each voice type. For modal voice all four LF model parameters have high  $R^2$  values with all four parameters being nearly the same value, and therefore are of comparable significance. For breathy voice the LF model parameter  $t_a$  is the most significant, while for vocal fry the LF parameter  $t_e$  is the most significant. Higher-order regression models were also calculated. However the grouping of the most significant LF model parameters did not differ greatly from what one would predict from the first-order model. For example, from Table III one would predict that the two most significant LF model parameters for breathy voice are  $t_a$  and  $t_e$ . The results for the second-order linear regression model confirmed this. Although the  $R^2$  always increased with higher-order models, the increase was not very great over that calculated for the first-order model, suggesting that the parameter values are highly correlated.

## D. Synthesis

To synthesize a particular voice type the LF model timing parameters must have  $0 \leq t_p \leq t_e \leq t_c$  and  $t_a \geq 0$ . Note also that  $t_p$  and  $t_c$  are related to  $SQ_{LF}$  by

$$SQ_{LF} = \frac{t_p}{t_c - t_p}. \quad (2)$$

Thus a rule to select a set of LF model parameters to synthesize a particular voice type could be either

- (1) specify  $f_0$ , then select  $t_p$ ,  $t_e$ ,  $t_a$ , and  $t_c$  from Table II, or
- (2) specify  $f_0$ , then compute  $t_p$  from (2) and select  $t_e$ ,  $t_a$ ,  $t_c$  and  $SQ_{LF}$ , from Table II.

During the synthesis, we reproduced in the synthesized speech the perturbation (jitter and shimmer) measured from the natural speech, as well as  $f_0$ . One must ensure that the choice of  $f_0$  is compatible with the value for  $t_c$ , since  $t_c$  must be less than or equal to the pitch period.

The glottal pulse characteristics, along with the spectral tilt, generally modeled the low-frequency characteristics of the excitation well. However, the high-frequency characteristics of the excitation were not accounted for by these glottal pulse characteristics. In fact, most of the inverse filtered glottal waveforms exhibited some high-frequency "noise" superimposed on the volume-velocity waveform. This noise component is called "turbulent noise," and was modeled in our synthesis procedure.

To partially verify the analysis results and to validate our rule for synthesizing various voice types, speech tokens were synthesized with a formant synthesizer, similar in design to that of Klatt's (Klatt, 1980; Klatt and Klatt, 1990; Pinto *et al.*, 1989). The synthesized speech tokens were evaluated perceptually by listening tests by ten listeners who were faculty or graduate students from the University of Florida Speech Dept. The listeners were familiar with various voice types, including the three we studied, but were generally unfamiliar with synthesized speech. The listening tests were conducted with headphones in an IAC room. The speech tokens were played back directly from the computer using a digital-to-analog converter and an audio amplifier. The speech tokens consisted of the sentence "We were away a year ago." Three listening tests were conducted, one for each voice type. For example, a token of a modal voice was analyzed for the sentence "We were away a year ago," extracting the necessary parameter values for the formant synthesizer and the inverse filtered glottal waveform. The parameter values for the glottal pulse model for a modal voice were then selected from Table II and a glottal excitation waveform was constructed using these values as well as the measured  $f_0$ , the measured jitter and shimmer. Simulated turbulence noise was added as per Childers and Lee (1991). The excitation waveform was simulated for the entire sentence and was used to excite the formant synthesizer. This procedure was repeated for each voice type. The tokens were presented in an A-B manner for each voice type. The tokens for A and B were selected in a random manner to be either the natural (original) speech token or the synthesized speech token. Thus the A-B presentations included tokens in all possible orders: natural-natural, synthesized-synthesized, natural-synthesized, and synthesized-natural. The listeners were asked to judge which token in the A-B presentation sounded the most like the voice type that was being judged. Nine out of the ten judges agreed that the tokens sounded like the voice type being judged. For approximately 40% of the cases the judges could not distinguish the synthesized token from the natural token. In summary, the synthesized speech tokens were judged by the listeners to mimic the three voice types, when compared with the corresponding natural voice type.

We also found that we could "convert" one voice type to another by using the appropriate glottal excitation pulse

model (Childers *et al.*, 1989b). For example, the procedure described above was repeated, except the parameter values appropriate for the glottal pulse model for a vocal fry (or breathy) voice were taken from Table II and the glottal excitation waveform was constructed using these values as well as the measured  $f_0$ , the measured jitter and shimmer. Again simulated turbulence noise was added (Childers and Lee, 1991). This excitation waveform was simulated for the entire sentence and was used to excite the formant synthesizer using the parameter values measured for the modal voice. Nine out of ten of the listeners agreed that the synthesized voice sounded like the original speaker, but with a vocal fry (or breathy) voice, instead of a modal voice. However, the listeners were unable to judge the quality of the synthesized voice type since speech tokens for the speakers were available for modal voice only.

It was noted that as  $t_a$  was increased, the synthesized voices informally sounded more lax (or hypofunctional). However, the parameter  $SQ_{LF}$  gave the opposite results for all voice types, i.e., as  $SQ_{LF}$  was increased (or, equivalently,  $t_p$  was increased for a fixed  $t_c$ ), then the synthesized voices sounded more tense (or hyperfunctional). As  $t_c$  was increased, the synthesized voices sounded softer, i.e., were less loud and more breathy. Among the three parameters ( $t_a, t_p, t_c$ ),  $t_c$  (or equivalently  $OQ_{LF}$ ) appeared to be the most important with respect to characterizing the three voice types we examined, which agrees with the results obtained for the LF model parameters. Other informal findings included:

- (1) The glottal pulse width, the abruptness of glottal closure, and the spectral tilt were useful factors for differentiating a breathy voice from modal and vocal fry, and
- (2) in general, the incorporation of turbulence noise in the excitation enhanced the naturalness of the synthesized speech.

In summary, our listening tests of the synthesized speech verified that the glottal pulse width, pulse skewness, the abruptness of glottal closure, and the spectral tilt were useful factors for differentiating the three voice types: modal, vocal fry, and breathy. The incorporation of the appropriate  $f_0$ , jitter and shimmer, and a glottal turbulence noise source enhanced the naturalness of the synthesized speech.

### III. DISCUSSION

As we discussed previously, it is known that the shape of the glottal pulse varies greatly from speaker to speaker for different speaking tasks (Monsen and Engebretson, 1977) and affects the quality and naturalness of synthetic speech (Carlson *et al.*, 1991; Childers and Lee, 1991; Childers and Wu, 1990; Fant, 1979; Fujisaki and Ljungqvist, 1986; Holmes, 1973, 1983; Karlsson, 1988, 1990, 1991, 1992; Klatt, 1987; Klatt and Klatt, 1990; Pinto *et al.*, 1989; Rosenberg, 1971). A wide variation of the glottal waveform shape, its rms (root mean squared) intensity, fundamental frequency, phase spectrum, and intensity spectrum have been reported to occur across subjects (Sondhi, 1975). Thus the purpose of this study was to verify and quantify the degree that the glottal flow waveforms for the three voice types might differ.

*Data and subjects.* One of the liabilities of the data for vocal fry and breathy voices for this study was that these subjects were patients. Consequently, the quality ratings for their voices reflect the subject's symptoms, which could have been determined by several factors, including both functional (physiological) and structural (pathological) factors. Another liability was that the ratings of the voices provided by the seven judges were based on evaluations of a sustained vowel of approximately 2 s in duration. A longer speech token would probably have produced more consistent ratings. In addition, the subject population was small as were the number of speech tokens analyzed for each subject and each voice type. This latter weakness certainly contributed to the large variance in the values calculated for the LF model parameters. However, the factor that we felt influenced the variance the most was the fitting of the LF model to the measured volume-velocity waveform data. This fitting process was most difficult for the vocal fry and breathy voices because the measured volume-velocity waveforms for these two voice types often deviated from the more "typical" waveforms measured for modal voices. Consequently, the standard deviation for the LF model parameters was greater for vocal fry and breathy voice types than for modal voice. This can be observed in Table II and Fig. 4. Another factor, but one of less importance, was the high-frequency ripple activity that appeared on the inverse filtered differentiated volume-velocity waveforms, as seen in Fig. 2 and the top of Fig. 3. This type of activity is not modeled by the LF waveform. Rather, due to the fluctuating nature of this activity, the LF model tends to arrive at an average waveform that represents a "smoothed" version of the data. This type of activity was not as much of a problem as one might first suspect. Another factor that may have contributed to the variance in the measured LF model parameters was that we did not reject any inverse filtered waveforms as being inferior according to some criterion. This is in agreement with Milenkovic (1986, 1993). We feel that retaining all the inverse filtered waveforms was not as important a problem as the other factors we have discussed, and contributed very little to the observed variance in the results. The average fundamental frequency for the vocal fry data was approximately 100 Hz. This seems high for vocal fry. However, the fundamental frequency of voicing calculated from the average pitch period data was lower than 100 Hz. The reason for the high  $f_0$  shown in Table II was that we calculated  $f_0$  for every pitch period. This calculation resulted in a large variance in the estimate for  $f_0$  due to calculating the reciprocals of the values for the pitch period. Stated another way, the numbers for the pitch period were small; therefore, their reciprocals were large. Thus small differences in successive values for the pitch period become large differences in the successive values for the fundamental frequency of voicing. (This argument applies to all voice types, but was particularly noticeable for vocal fry.) Consequently, we feel the values for  $f_0$  in Table II are biased on the high side because of the manner by which they were calculated. However, this is of little or no importance since it had no effect on the estimation procedure used to determine the LF model parameters.

## A. Glottal waveform characteristics

We used a waveform matching technique with a minimum mean-squared error criterion for determining the LF model parameter values rather than a spectrum based criterion. There are two reasons for taking this approach. First, if the time domain waveform features of the model are correct, then the spectral features of the model will be correct. Second, if a magnitude spectrum approach is used, then 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.

The average values for the LF model parameters showed that the glottal pulse width was approximately 60% of the pitch period for modal voices, 72% for vocal fry, and 80% for breathy voices. For the three voice types examined (modal, vocal fry, and breathy), the glottal closing phase of the volume-velocity waveform exhibits a steeper slope than the slope for the opening phase. Thus the glottal flow waveforms are skewed to the right. Glottal pulse skewness varied with voice type. For modal and vocal fry phonations waveform skewing was more apparent than for breathy phonations. Typically, the waveforms for modal and vocal fry voices showed a more distinct closed phase. The closed phase was not always apparent for breathy voices, and, in addition, the glottal flow for breathy voice waveforms was approximately sinusoidal. Overall, our results agreed with the findings reported in Childers and Lee (1991).

Due to glottal pulse skewness, i.e., an increase in pulse slope during glottal closure, the main excitation for the vocal tract occurs at the point of vocal fold closure. This excitation can be controlled by the talker (Miller, 1959). In many cases we noted that there were well defined instants of excitation of the second and higher formants at other points in the volume-velocity waveform; one such point occurred at the instant of the opening of the glottis. This agrees with Holmes (1962). For modal voice, the instant of the maximum closing slope occurs near the instant of glottal closure, resulting in an abrupt termination of the glottal airflow. Vocal fry has appreciable excitation at both the beginning and end of the glottal open phase. Frequently an alteration in the spectral content of the excitation may occur from cycle to cycle for vocal fry. This causes the relative intensities of the formants to vary (Childers and Lee, 1991; Hunt, 1987). For breathy voice, the instant of the maximum glottal pulse closing slope occurs near the middle of the glottal closing phase, followed by a residual phase of progressive closure. Thus there is the expectation that vocal fry and breathy voice may have appreciable formant excitation at various locations within the flow waveform.

The results of the ANOVA demonstrated that there is a difference in the four LF model parameters among the three voice types, with the possible exception of  $t_p$ , which was only marginally significant. This seems reasonable since this parameter identifies the location of the peak of the glottal volume-velocity waveform, which is typically broad for all voice types. Therefore it is not unreasonable that  $t_p$  should

be the least significant of the four parameters. However, the ANOVA did not consider the average quality rating provided by the seven judges. Consequently, a multiple linear regression analysis was performed, the results of which predicted that all four of the LF model parameters were nearly equally significant for modal voice, while  $t_a$  was most significant for breathy voice, and  $t_e$  was most significant for vocal fry. These results generally agree with the results from the listener evaluation of the synthesized speech and with Childers and Lee (1991), Fant (1993), Fant and Lin (1988), and Karlsson (1988). While the multiple linear regression analysis only predicted the most significant parameters, we must rely on other results and inferences to assess the importance of such predictions. The parameter  $t_a$  has been determined to be a potential measure of breathiness (Childers and Lee, 1991; Fant, 1993; Fant and Lin, 1988; Karlsson, 1988). The larger  $t_a$ , the greater the tendency for the voice to be breathy, since the larger  $t_a$ , the less abrupt the glottal closure becomes. The reason that  $t_e$  may be significant for vocal fry is that this parameter marks the instant of the maximum glottal closing rate, which is the time for the primary glottal excitation for the LF model. Vocal fry tends to have two glottal closure events within one pitch period; one event is usually a well defined abrupt glottal closure, while the other event is usually a secondary glottal closure. Thus it seems reasonable that  $t_e$  could be significant for vocal fry. The fact that all four LF parameters appear significant for modal voice is reasonable based on the assumption that such a voice usually has a reasonably abrupt glottal closure, is not breathy, and has a well defined peak glottal flow. Finally, recall that there was only a small increase in the  $R^2$  values for higher-order linear regression models, suggesting that the parameter values for the LF model are highly correlated. Thus while the ANOVA results determined that three out of the four LF model parameters were significant, the linear regression analysis predicted that the parameters were possibly highly correlated. One interpretation of this is that the four LF parameters cannot be selected in an arbitrary manner when modeling glottal volume-velocity waveforms or the derivative of the volume velocity. The parameters apparently are restricted to ranges of values, if they are to properly model actual data. This is why we believe the tables of mean values and their corresponding standard deviations are of some importance for modeling the three voice types.

It has long been noted that some "ringing" activity may occur in the closed phase of the estimated glottal waveform obtained by inverse filtering (Childers and Wong, 1994; Hillman and Weinberg, 1981; Holmes, 1976; Hunt *et al.*, 1978; Karlsson, 1991, 1992; Milenkovic, 1993; Rothenberg, 1973). On occasion we see such activity in some of our data as well (Fig. 2). Several possible explanations of this phenomenon have been suggested in the literature, including acoustic interaction with the glottis (Rothenberg, 1973), mucosal wave motion across the surface of the vocal folds (Holmes, 1976), displaced glottal air (Rothenberg, 1973), laryngeal adjustments (Rothenberg, 1973), and nasalization of vowels (Rothenberg, 1973; Hunt *et al.*, 1978). 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 in-

teraction 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 (Childers and Wong, 1994). 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 (Holmes, 1976; Hunt *et al.*, 1978; Childers and Wong, 1994). 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 causes mentioned above, as also suggested by Holmes (1976) and Hunt *et al.* (1978). However, such activity had little or no effect on the measurement of the glottal pulse parameters for the glottal pulse model, since the activity in the closed phase region was small relative to the glottal pulse activity during the open phase region, and therefore did not influence the mean-squared error between the LF model pulse and the pulse determined by inverse filtering. Our inverse filtered waveforms compare favorably with those of others, including the recent results of Milenkovic (1993).

In addition to the ringing sometimes observed in the closed phase of the inverse filtered waveform, there may also be a ripple component in the open phase of the glottal flow waveform due to source–tract interaction (Fant and Ananthapadmanabha, 1982). This ripple is attributed to first formant interaction with the source within the glottal open phase. Source–tract interaction may change the formant amplitudes, the formant frequencies, and the formant bandwidths during the glottal open phase. While we did not investigate source–tract interaction in this study, we do address this problem elsewhere (Childers and Wong, 1994).

Fant (1993) has summarized the above remarks concerning inverse filtering by noting that inverse filtering does not necessarily determine the true glottal flow waveform. Rather, it is a compromise attained by adjusting the filter parameters. Usually the inverse filter is adjusted to provide maximum formant cancellation over the closed glottal interval, which must be estimated. Such a setting may cause errors since there may in fact be a finite glottal opening with some subglottal coupling (Fant, 1993). However, if the filter is adjusted to account for a glottal opening, then the formants will not be completely cancelled. The filter is usually set for maximum formant cancellation over the estimated closed glottal interval since this gives good results for formant synthesis (Fant, 1993).

## B. Synthesis

To partially validate our results obtained by analysis, we developed a simple rule to synthesize a particular voice type using parameter values selected from Table III. We found that one voice type could be converted to sound like that of another voice type (Childers *et al.*, 1989b). Using this approach we found that as  $t_a$  was increased, the synthesized

voice sounded more lax (or hypofunctional). As  $t_c$  was increased, the synthesized voice sounded softer, i.e., was less loud and more breathy. Among the three parameters ( $t_a, t_p, t_c$ ),  $t_c$  appears to be the most important with respect to characterizing the three voice types we examined. This agrees with the conclusions we reached concerning the average values for the LF model parameters and with the ANOVA results, which determined that  $t_c$  was the most statistically significant. We also noted that the glottal pulse width, the abruptness of glottal closure, and the spectral tilt were useful factors for differentiating a breathy voice from modal and vocal fry.

## C. Summary

Several glottal source factors for three different voice types were investigated. The procedures for this research were glottal inverse filtering and glottal source modeling. The glottal inverse filtering was achieved using both the speech and EGG signals. Our inverse filtering method was able to process sentences as well as sustained vowels. For sentences, the inverse filtering was performed on voiced segments only. A range of parameter values was determined for the glottal source model along with a simple rule which was used to synthesize the three voice types.

The inverse filtered glottal flow waveforms for the three voice types showed typical patterns that could be characterized by pulse width, pulse skewness, and abruptness of closure. The spectral characteristics of the glottal flow waveforms for the three voice types also differed in spectral tilt. Therefore the low-frequency characteristics for each of the three voice types could be synthesized by specifying the appropriate glottal pulse characteristics and the spectral tilt. The high-frequency characteristics for each voice type were accounted for as described in Childers and Lee (1991).

One aspect of this study showed that the LF model waveform may result in an overestimation of the spectral tilt of the glottal excitation because the LF model has a lack of high-frequency energy. We have suggested one algorithm that compensates the LF glottal model to correct this feature. The algorithm only affects the parameter values for  $t_a$  and  $t_c$ .

Recall that for our study the subjects for vocal fry and breathy voice were taken from a patient population. Thus we stress caution in generalizing from our results, since it is likely that a vocal disorder may affect more than one dimension of the voice simultaneously. This makes the task of estimating specific attributes of the voice using acoustic parameters, such as we examined here, difficult. A further weakness is the dependence on the classification of voices by listener evaluations. To be more conclusive, more information concerning the classification of specific voice types and a more comprehensive study of larger speech samples is needed. Furthermore, a larger set of subjects should be used so that the statistics would be more meaningful. Despite these weaknesses, the results obtained in this study did generally agree with previous results and did prove sufficiently specific that aspects of the three voice types could be synthesized.

The results obtained with and the methods developed for this study may serve as the basis for further study of (1) the estimation of parameters for excitation source models for a broad range of voice types including falsetto, hoarse, and harsh voices, (2) the quantification of severity of voice quality or vocal dysfunction, (3) speaker normalization to improve the performance of a speaker-independent speech recognition systems, or the development of an objective distortion measure that would incorporate dynamic features of speech signals, (4) a database of different voice types to be used in training a speech recognition system, and (5) the effects of variability caused by variations in the vocal tract parameters.

## ACKNOWLEDGMENTS

This work was supported in part by NIH Grant No. NIDCD DC 00577 and NSF Grant No. IRI-9215331. The authors wish to express their appreciation to the referees for their conscientious and thorough review of the manuscript and its revisions.

## APPENDIX A

The LF model is shown in Fig. A1 (Fant *et al.*, 1985; Fant, 1993). This model describes the differentiated glottal flow rather than the glottal flow itself. The differentiated flow is commonly used in speech synthesis, and includes the effect of radiation at the lips. The LF model consists of two segments. The first segment is an exponentially growing sinusoid, and the second segment is an exponential decaying function. Each segment may be expressed as follows:

$$\frac{dU_g(t)}{dt} = E(t) = E_0 e^{\alpha t} \sin \omega_g t, \quad 0 \leq t \leq t_e, \quad (\text{A1})$$

$$E(t) = -\frac{E_e}{\epsilon t_a} [e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], \quad t_e \leq t \leq t_c < t_0, \quad (\text{A2})$$

where  $t_0$  is the pitch period interval over which the waveform of the LF model is defined. At time  $t_e$  both segments have the same value  $E_e$ . Besides the above relationships, the model requires that the positive and negative areas of the differentiated glottal flow must be equal so that the base line of the glottal flow does not drift. Thus the integral of the LF model over the glottal period is zero,

$$\text{i.e., } \int_0^{t_0} E(t) dt = 0.$$

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

- (1)  $E_0$ , which 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 flow).

In the second part of the LF model, the parameter  $t_a$  is the time constant of the exponential curve and is the time interval from  $t_e$ , the location of the negative peak of the LF

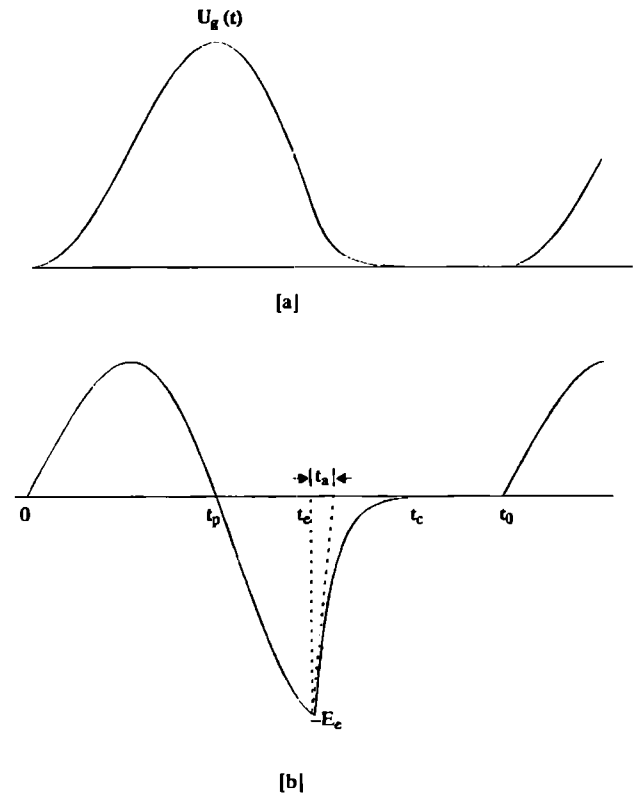


FIG. A1. The LF model for both (a) the glottal flow  $U_g(t)$  and (b) the differentiated glottal flow  $U'_g(t)$ . (Not drawn to scale.)

model, to its intercept of the projected derivative of the model at time  $t_e$ . The parameter  $-E_e$  is the negative amplitude of the model at time  $t_e$ . The parameter  $t_c$  is the instant at which the model returns to zero and therefore represents the time of glottal closure. The parameter  $\epsilon$  is the decay constant of the recovery phase of the exponential. The four parameters  $E_0$ ,  $\alpha$ ,  $\omega_g$ , and  $\epsilon$  are called the “direct synthesis parameters” of the LF model, while the time parameters  $t_p$ ,  $t_e$ ,  $t_a$ , and  $t_c$  are called the “timing parameters.” These parameters may be considered as independent of one another since a unique waveform may be created with each combination of parameters. Methods for calculating the various parameters of the model from the timing parameters  $t_p$ ,  $t_e$ ,  $t_a$ ,  $t_c$ , and  $-E_e$  are discussed in Fant *et al.* (1985) and Childers and Lee (1991). In summary, the parameter  $t_p$  marks the position of peak glottal flow,  $t_e$  is the instant of the maximum glottal closing rate,  $t_a$  is the time constant of the exponential recovery as well as an indication of the abruptness of glottal closure (the larger  $t_a$ , the less abrupt the closure), and  $t_c$  marks the instant of glottal closure, which is less than or equal to the pitch period  $t_0$ , which is denoted in the tables as  $pp$ .

## APPENDIX B

The frequency response of the LF model has a zero at dc, a complex pole pair at  $\alpha \pm j\omega_g$ , and a real pole at  $-\epsilon$  (Fant and Lin, 1988). The zero is due to the fact that the integral of the LF model time function is equal to zero. The complex pole pair is attributed to the first segment of the LF model.

The real pole is due to the return phase  $t_a$ , which determines the second part of the LF model. The zero and complex pole pair result in a spectral roll-off of  $-6$  dB/oct, and the return phase provides a spectral roll-off of about  $-6$  dB/oct, depending on the details of the return phase. Thus the return phase can be used to control the spectral tilt of the model. The effect of the return phase on the source spectrum is equivalent to that of a first-order low-pass filter with a cutoff frequency  $F_a = 1/(2\pi t_a)$  in hertz (Fant and Lin, 1988; Fant, 1993). Thus the power spectral density function attributed to the return phase can be expressed as:

$$|S(\Omega)|^2 = \frac{1}{1 + t_a^2 \Omega^2}, \quad (\text{B1})$$

where  $\Omega$  is the analog frequency in radians. The digital impulse invariant realization of Eq. (B1) is in the form (Childers and Durling, 1975):

$$S(z) = \frac{1/t_a}{1 - e^{-T/t_a} z^{-1}}, \quad (\text{B2})$$

where  $T$  is the sampling period. Equation (B2) can be interpreted as a real-pole model of the form  $(1/t_a)/(1 - z_c z^{-1})$ , where the coefficient  $z_c$  is given by

$$z_c = e^{-T/t_a}. \quad (\text{B3})$$

Equation (B3) implies that the longer the return phase, the steeper the spectral tilt, and the greater the reduction of the high frequencies in the spectrum.

Our compensation algorithm first compares the spectral tilt estimated (by using the three-pole source model) for both the inverse filtered, differentiated glottal waveforms and the modeled, differentiated glottal waveforms. Then, using the relationship in Eq. (B3), the return phase  $t_a$  of the LF model is adjusted to approximate the spectral tilt of the inverse filtered differentiated glottal flow waveforms (Ahn, 1991). One side effect of adjusting the return phase  $t_a$  is that the settling time  $t_c$  is also changed. Hence there is the possibility that such a change in  $t_c$  may cancel out the effect of adjusting  $t_a$  in the low-frequency region of the compensated LF model. However, this is less important than achieving the desired general spectral tilt in the high-frequency region of the glottal excitation spectrum.

Ahn, C. (1991). "A study of voice types and acoustic variability: Analysis-by-synthesis," Ph. D. dissertation, University of Florida, Gainesville, FL.  
 Allen, D. M., and Cady, F. B. (1982). *Analyzing Experimental Data by Regression* (Wadsworth, Belmont, CA).  
 Carlson, R., Granstrom, B., and Karlsson, I. (1991). "Experiments with voice modeling in speech synthesis," *Speech Commun.* **10**, 481–489.  
 Childers, D. G., and Durling, A. (1975). *Digital Filtering and Signal Processing* (West Publishing, St. Paul, MN).  
 Childers, D. G., Hahn, M., and Larar, J. N. (1989a). "Silent and voiced/unvoiced/mixed excitation (four way) classification of speech," *IEEE Trans. Acoust. Speech Signal Process.* **37**, 1771–1774.  
 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 Lee, C. K. (1991). "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.* **90**, 2394–2410.

Childers, D. G., and Wu, K. (1990). "Quality of speech produced by analysis-synthesis," *Speech Commun.* **9**, 97–117.  
 Childers, D. G., Wu, K., Hicks, D. M., and Yegnanarayana, B. (1989b). "Voice conversion," *Speech Commun.* **8**, 147–158.  
 Childers, D. G., and Wong, C. F. (1994). "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.* **41**, 663–671.  
 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, The Hague).  
 Fant, G. (1979). "Glottal source and excitation analysis," STL-QPSR (Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) **1**, 85–107.  
 Fant, G. (1993). "Some problems in voice source analysis," *Speech Commun.* **13**, 7–22.  
 Fant, G., and Ananthapadmanabha, T. V. (1982). "Truncation and superposition," STL-QPSR **2–3**, 1–17.  
 Fant, G., Liljencrants, J., and Lin, Q. G. (1985). "A four-parameter model of glottal flow," STL-QPSR **4**, 1–13.  
 Fant, G., and Lin, Q. G. (1988). "Frequency domain interpretation and derivation of glottal flow parameters," STL-QPSR **2–3**, 1–21.  
 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.  
 Hillman, R. E., and Weinberg, B. (1981). "Estimation of glottal volume velocity waveform properties: A review and study of some methodological assumptions," in *Speech and Language*, edited by N. J. Lass (Academic, New York), Vol. 6, pp. 411–473.  
 Holmes, J. N. (1962). "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," *Proceedings of the Fourth International Congress on Acoustics, Copenhagen, Denmark* (unpublished), pp. 1–4.  
 Holmes, J. N. (1973). "The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.* **21**, 298–305.  
 Holmes, J. N. (1976). "Formant excitation before and after glottal closure," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* **1**, 39–42.  
 Holmes, J. N. (1983). "Formant synthesizers: Cascade or parallel?" *Speech Commun.* **2**, 251–273.  
 Hunt, M. J., Eridle, J. S., and Holmes, J. N. (1978). "Interactive digital inverse filtering and its relation to linear prediction methods," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* **1**, 15–18.  
 Hunt, M. J. (1987). "Studies of glottal excitation using inverse filtering and an electroglottograph," *Proceedings of the 13th International Congress of Phonetic Science, Tallinn, Estonia, August 1987* (unpublished).  
 Karlsson, I. (1986). "Glottal wave forms for normal female speakers," *J. Phon.* **14**, 415–419.  
 Karlsson, I. (1988). "Glottal waveform parameters for different speaker types," *Proc. Speech '88, 7 FASE Symposium*, Vol. 1, pp. 225–231.  
 Karlsson, I. (1990). "Voice source dynamics for female speakers," *International Conference Spoken Language Proceedings*, Vol. 1, pp. 69–72.  
 Karlsson, I. (1991). "Female voices in speech synthesis," *J. Phon.* **19**, 111–120.  
 Karlsson, I. (1992). "Modeling voice variations in female speech synthesis," *Speech Commun.* **11**, 491–495.  
 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.* **34**, 730–743.  
 Milenkovic, P. H. (1986). "Glottal inverse filtering by joint estimation of AR system with a linear input model," *IEEE Trans. Acoust. Speech Signal Process.* **23**, 28–42.  
 Milenkovic, P. H. (1993). "Voice source model for continuous control of pitch period," *J. Acoust. Soc. Am.* **93**, 1087–1096.  
 Miller, J. D. (1959). "Nature of the vocal cord wave," *J. Acoust. Soc. Am.* **31**, 667–677.  
 Monsen, R. B., and Engebretson, A. M. (1977). "Study of variations in the male and female glottal wave," *J. Acoust. Soc. Am.* **62**, 981–993.

- Pinto, N. B., Childers, D. G., and Lalwani, A. L. (1989). "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust. Speech Signal Process.* **37**, 1870–1887.
- Rosenberg, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.* **49**, 583–590.
- Rothenberg, M. (1973). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Am.* **53**, 1632–1645.
- Sondhi, M. M. (1975). "Measurement of the glottal waveform," *J. Acoust. Soc. Am.* **57**, 228–232.