Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification

Michael D. Plumpe, *Member, IEEE,* Thomas F. Quatieri, *Fellow, IEEE,* and Douglas A. Reynolds, *Senior Member, IEEE*

***Abstract*—An automatic technique for estimating and modeling the glottal flow derivative source waveform from speech, and applying the model parameters to speaker identification, is presented. The estimate of the glottal flow derivative is decomposed into coarse structure, representing the general flow shape, and fine structure, comprising aspiration and other perturbations in the flow, from which model parameters are obtained. The glottal flow derivative is estimated using an inverse filter determined within a time interval of vocal-fold closure that is identified through differences in formant frequency modulation during the open and closed phases of the glottal cycle. This formant motion is predicted by Ananthapadmanabha and Fant to be a result of time-varying and nonlinear source/vocal tract coupling within a glottal cycle. The glottal flow derivative estimate is modeled using the Liljencrants–Fant model to capture its coarse structure, while the fine structure of the flow derivative is represented through energy and perturbation measures. The model parameters are used in a Gaussian mixture model speaker identification (SID) system. Both coarse- and fine-structure glottal features are shown to contain significant speaker-dependent information. For a large TIMIT database subset, averaging over male and female SID scores, the coarse-structure parameters achieve about 60% accuracy, the fine-structure parameters give about 40% accuracy, and their combination yields about 70% correct identification. Finally, in preliminary experiments on the counterpart telephone-degraded NTIMIT database, about a 5% error reduction in SID scores is obtained when source features are combined with traditional mel-cepstral measures.**

***Index Terms*—Frequency modulation, Gaussian mixture model, glottal flow derivative, inverse filtering, nonlinear glottal/vocal tract interaction, speaker recognition.**

I. Introduction

VIDEOS of vocal fold vibration [21] show large variations in the movement of the vocal folds from one individual to another. For certain speakers, the vocal folds may close completely, while for others, the folds may never reach full closure. The manner and speed in which the vocal folds close also varies across speakers. For example, the cords may close in a zipperlike fashion, or may close along the length of the vocal folds at approximately the same time. Differences in fold vibration correspond to differences in the time-varying area of the slitlike opening between the folds, referred to as the *glottis*, and therefore in volume velocity air flow through the glottis, i.e., the *glottal flow*. The flow may be smooth, as when the folds never close completely, corresponding perhaps to a "soft" voice, or discontinuous, as when they closed rapidly, giving perhaps a "hard" voice. The flow at the glottis may be turbulent, as when air passes near a small portion of the folds that remains partly open. Turbulence at the glottis is referred to as *aspiration* which, when occurring during vocal cord vibration, can result in a "breathy" voice. In order to determine quantitatively whether such glottal characteristics contain speaker dependence, we must extract features such as the timing of vocal fold opening and closing, the general shape of the glottal flow, and the extent and timing of turbulence at the vocal folds.

This paper describes a technique to automatically estimate and temporally model the glottal flow derivative waveform from voiced speech, and then uses the model parameters for speaker identification. A block diagram of the approach is given in Fig. 1. Our first goal of estimating the derivative of the glottal flow, rather than the glottal flow itself, stems from the availability of pressure measurements of the speech waveform, pressure being the derivative of volume velocity airflow. Estimation of the glottal flow derivative relies on inverse filtering the speech waveform with an estimate of the vocal tract transfer function. This estimation is typically performed during the glottal closed phase within which the vocal folds are in a closed position and there is no dynamic source/vocal tract interaction. Wong *et al.* [34] and Cummings and Clements [9] perform, for example, a sliding covariance analysis with a one sample shift, using a function of the linear prediction error to identify the glottal closed phase. This method, relying on the prediction error, has been observed to have difficulty when the vocal folds do not close completely or when the folds open slowly. The approach of this paper estimates the glottal closed phase, relying also on a sliding covariance analysis, but, rather than using the prediction error from this analysis, uses vocal tract formant modulation which is predicted by Ananthapadmanabha and Fant [1] to vary more slowly in the glottal closed phase than in its open phase and to respond quickly to a change in glottal area. A "stationary" region of formant modulation gives a closed-phase time interval, over which we estimate the vocal tract transfer function; a stationary region is present even when the vocal folds remain partly open. For high-pitched speakers, where the closed phase over a single pitch period is small, a method is proposed which uses two consecutive pitch periods

Manuscript received December 22, 1997; revised November 19, 1998. This work was sponsored by the Department of the Air Force. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

M. Plumpe is with Microsoft Research, Redmond, WA 98052 USA (e-mail: mplumpe@microsoft.com).

T. F. Quatieri and D. A. Reynolds are with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173 USA (e-mail: tfq@sst.ll.mit.edu).

Publisher Item Identifier S 1063-6676(99)06562-1.

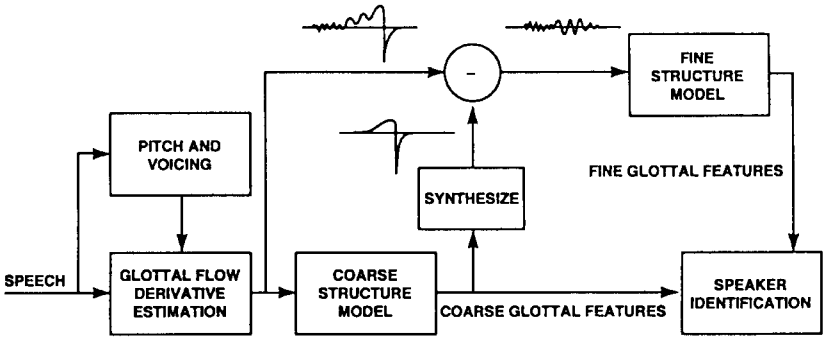


Fig. 1. Approach to glottal flow derivative estimation and modeling, and its use in speaker identification.

to improve estimation of the formant modulation and closed phase, and thus of the glottal flow derivative. The glottal flow derivative waveform that results from inverse filtering is characterized by temporal structure that is consistent with that predicted by Ananthapadmanabha and Fant [1].

With the objective of modeling the temporal structure of the glottal flow, we then develop a time-domain feature representation of the glottal flow derivative during voicing. The *coarse structure* of the flow derivative is represented by the piecewise-functional Liljencrants–Fant (LF) model [14] consisting of seven parameters, obtained by a nonlinear estimation method of the Newton–Gauss type which allows for invoking physically-motivated solution bounds, as well as for possibly large residual error. As illustrated in Fig. 1, the coarse structure is then subtracted from the glottal flow derivative estimate to give its *fine-structure* component, reflecting characteristics not captured by the general flow shape such as aspiration and a perturbation in the flow referred to as *ripple.* Ripple is associated with first-formant modulation and is due to the time-varying and nonlinear coupling of the source and vocal tract cavity [1]. Five energy measures and a ripple measure are obtained from the fine structure.

The coarse- and fine-structure features are then applied to a speaker identification (ID) task using a Gaussian mixture model speaker ID system [27]. The results represent the first demonstration, using an automatic speaker ID system, of the speaker dependence of glottal-flow derivative feature estimates for voiced speech. Early use of source information in automatic speaker ID systems was accomplished mainly through pitch features [3]. More recently, source information was used in a few speaker ID systems [16], [31]. These methods, however, do not use an explicit temporal model of the glottal flow derivative, but rather use cepstral representations of a linear prediction residual, and, moreover, without the aid of glottal open/close timing and voiced speech identification. As such, the residual is a representation of the source, primarily in the form of pitch and voicing information, and not of a glottal flow derivative measure. In this paper, on the other hand, the importance of the glottal flow derivative shape in voicing is isolated by not using pitch in the speaker ID task and by making glottal flow measurements only during voiced speech.

The paper is organized as follows. In Section II, properties of the glottal flow, its derivative, and the Ananthapadmanabha and Fant glottal flow physiological model are reviewed, and then a functional model that captures the important features of the glottal flow derivative is developed. Section III describes the technique used to estimate the glottal flow derivative waveform. Estimation of the features of the coarse structure of the glottal flow derivative is given in Section IV, while Section V develops estimation of its fine-structure parameters. Section VI describes the use of the model parameters for speaker identification. Finally, Section VII gives conclusions and ideas for future directions.

II. Glottal Flow Model

This section first describes qualitatively the properties of the components of glottal flow and its derivative, then briefly reviews Ananthapadmanabha and Fant's [1] theory of formant frequency-modulation and associated ripple due to source/vocal tract interaction, and ends with a glottal flow derivative model for extracting features to be used in speaker identification.

A. Properties of the Glottal Flow

Speech production is typically viewed as a linear filtering process which can be considered time invariant over short time intervals. The glottal flow volume velocity, denoted by $u_g(t)$, acts as the source, sometimes also referred to as the "excitation," to the vocal tract with impulse response $h(t)$. The volume velocity output of the vocal tract is then modified by the lip impedance. Because the pressure/volume velocity relation at the lips can be approximated by a differentiator [26], the speech pressure waveform $s(t)$ measured in front of the lips can be expressed as $s(t) \approx d[u_g(t) * h(t)]/dt = [du_g(t)/dt] * h(t)$. The effect of radiation is typically included in the source function [26]; the source to the vocal tract, therefore, becomes the derivative of the glottal flow volume velocity, which we henceforth denote by $v_g(t)$, i.e., $v_g(t) = \dot{u}_g(t)$. Following the approach of Ananthapadmanabha and Fant [1], we assume that the glottal flow and its derivative consists of coarse- and fine-structure components.

1) Coarse Structure: The relation between the coarse structure of the glottal flow, denoted by $u_{gc}(t)$, and its derivative, $v_{gc}(t)$, is shown in Fig. 2 for an idealized[1] glottal flow function. In obtaining the glottal flow derivative, applying the

[1] Typically, the vocal folds do not fully close and some air flow may always be present.

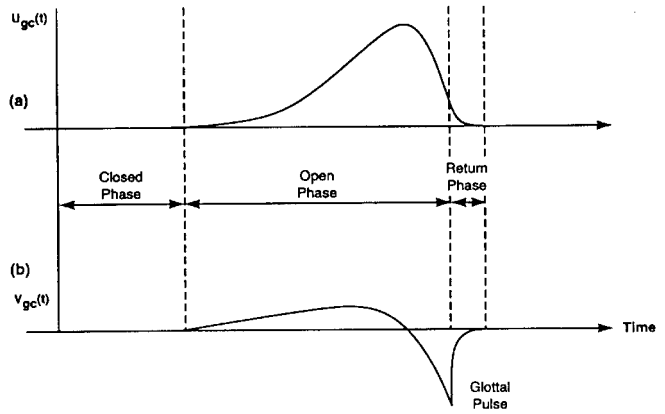


Fig. 2. Relation between glottal flow and its derivative: (a) glottal volume velocity (flow); (b) glottal flow derivative.

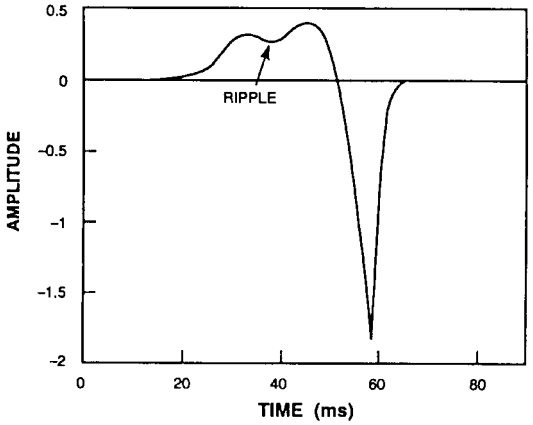


Fig. 3. Glottal flow derivative waveform showing coarse and ripple component of fine structure due to source/vocal tract interaction.

lip radiation effect to the source flow, rapid closing of the vocal folds results in a large negative impulse-like response at glottal closure, called the *glottal pulse*, as shown in Fig. 2. The coarse structure represents the general shape of the glottal flow. The time interval during which the vocal folds are closed, and during which no flow occurs, is referred to as the glottal *closed phase*. The time interval over which there is nonzero flow and the vocal folds are fully or partially open is referred to as the glottal *open phase*. The time interval from the most negative value of the glottal flow derivative to the time of glottal closure is referred to as the *return phase*. The asymmetry of the glottal flow shape during the open phase, sometimes referred to as *skewness* in the glottal flow, is due approximately in part to the manner in which the glottis changes in time, and in part to the loading by the vocal tract during the glottal open phase[2] [1]. In this glottal flow model, the return phase is particularly important, as this determines the amount of high-frequency energy present in both the source and the speech. The more rapidly the vocal folds close, the shorter the return phase, resulting in more high-frequency energy and less spectral tilt.

2) Fine Structure: Fine structure of the glottal flow derivative, denoted by $v_{gf}(t)$, is the residual waveform obtained by subtracting the coarse structure from the glottal flow derivative, i.e., $v_{gf}(t) = v_g(t) - v_{gc}(t)$. Two contributions of fine structure are considered in this paper, ripple and aspiration. As illustrated in Fig. 3, ripple is a sinusoidal-like perturbation that overlays the coarse glottal flow, and thus the glottal flow derivative, and arises from the time-varying and nonlinear coupling of the glottal flow with the vocal tract cavity, due to primarily the vocal tract first formant[3] [1]. Based on a physical model, a more quantitative description of ripple and its correspondence to first-formant modulation is given in Section II-B. The timing and amount of ripple is dependent on the configuration of the glottis during both the open and closed phases [1], [14], [24]. For example, with folds that open in a zipperlike fashion, ripple may begin at a low level early into the glottal cycle, and then grow as the vocal folds open more completely.

Our second form of fine structure, aspiration at the glottis, arises when turbulence is created as air flows through constricted vocal folds, and is also dependent on the glottis for its timing and magnitude. For example, a long, narrow opening, which constricts the air flow along the entire glottal length, tends to produce more aspiration than, for example, a triangular-shaped opening with partial constriction. The creation of turbulence at the glottis is highly nonlinear and a satisfactory physical model has yet to be developed. A simplification is to model aspiration as a random noise process, which is the source to the (linear) vocal tract. The complete fine-structure source is modeled as the addition of the aspiration and ripple source components.

B. Physical Model of Source/Vocal Tract Interaction

Ananthapadmanabha and Fant [1] proposed an equivalent circuit to approximately model the glottal flow, accounting for the time-varying opening and closing approximately of the vocal folds, and accounting for the nonlinear relation between the pressure drop across the glottis and the glottal flow as found empirically by van den Berg for a static glottis [33]. Time variation in this equivalent circuit enters through the time-varying area function of the glottis which is assumed known. This model is described by a set of equations which were simultaneously solved with a numerical iterative algorithm, yielding the glottal flow, assuming a constant vocal tract cavity. The resulting numerical solution gives an asymmetric (skewed) glottal flow with an overriding ripple component.

The numerical simulation of Ananthapadmanabha and Fant also revealed that formants above the first vocal tract formant do not significantly affect glottal flow. Ananthapadmanabha and Fant, therefore, proposed an approximate equivalent circuit with a single resonance. This simplified nonlinear, time-varying circuit was approximated by a Norton equivalent circuit with an ideal source given by the coarse glottal flow,

[2] Without vocal tract loading, the glottal flow would be proportional to the glottal area. While loading is somewhat influenced by the vocal tract, the coarse structure is largely free of vocal tract shape influence [1].

[3] In a qualitative sense, the time-varying pressure in the vocal tract cavity just above the glottis "backs up" against the glottal flow and interacts nonlinearly with the flow to form the ripple. In certain wind instruments, as in the trumpet, such a mechanism is even more pronounced, and indeed is essential in determining the sound of the instrument, as the vibration of the "lip reed" is strongly, and nonlinearly, coupled to the resonant frequencies of the cavity [5].

and an equivalent glottal impedance that is time-varying and controlled by the changing glottal area function. This resulting circuit was thus represented by a linear differential equation with time-varying coefficients. To obtain a frequency-domain representation of the source/vocal tract interaction, and thus a different perspective on the ripple component, the glottal impedance was assumed stationary at each time instant. A "pseudo-Laplace transform," representing the time-varying transfer function from the volume-velocity source (coarse glottal flow), with Laplace transform $U_{gc}(s)$, to output speech pressure, with Laplace transform $P_o(s, t)$, can then be written as

$$H(s, t) = \frac{P_o(s,t)}{U_{gc}(s)} = \frac{s/C}{s^2 + B_1(t)s + \omega_1^2(t)} \tag{1}$$

where the time-varying formant frequency and bandwidth

$$\begin{aligned} \omega_1(t) &= \omega_0 \sqrt{1 + \alpha \dot{g}_o(t)} \\ B_1(t) &= B_0[1 + \beta g_o(t)] \end{aligned} \tag{2}$$

are given in terms of the first formant frequency ω_0 and bandwidth B_0, where the function $g_o(t)$ is proportional to the time-varying area function, and where the constants α and β are a function of the amplitude and frequency of the first formant. We see that over a glottal cycle, the bandwidth change follows that of the area function, while the formant frequency change, being proportional to the derivative of the area function, rises at the onset of the glottal open phase and falls near the termination of this phase [1], [17], [24], thus showing that time-varying and nonlinear source/vocal tract coupling corresponds to a modulation of the first formant.

Alternatively, to obtain an approximate time-domain representation of the effect of the source/vocal tract interaction, Ananthapadmanabha and Fant held the vocal tract filter fixed and mapped all source/vocal tract interaction to the source. Using (1), an approximate expression for the glottal flow derivative can be derived as [1], [24]

$$v_g(t) \approx v_{gc}(t) + f(t) e^{-0.5tB_1(t)} \cos\left[\int_0^t \omega_1(\tau)\, d\tau\right] \tag{3}$$

where the second term contains the ripple component of the fine structure. The ripple has a frequency close to the first formant of the vocal tract, and the function $f(t)$ represents an amplitude modulation controlled by the glottal area function. These relations reveal an approximate duality of ripple in the time-domain and formant modulation in the frequency domain [1], [8], [24].

C. Feature Model

We now propose feature models for the coarse and fine structure of the glottal flow derivative.

1) Coarse Structure: The features we wish to capture through the coarse structure include the glottal open, closed, and return phases, the speeds of opening and closing, and the relationship between the glottal pulse and the peak glottal flow. To model the coarse component, $v_{gc}(t)$, of the glottal flow derivative, we use the LF model [14], expressed over a single glottal cycle by (Fig. 4)

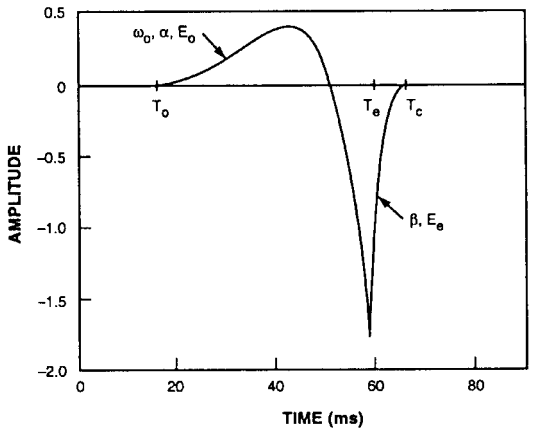


Fig. 4. LF Model for the glottal flow derivative waveform.

TABLE I
DESCRIPTION OF THE SEVEN PARAMETERS OF THE LF MODEL FOR THE GLOTTAL FLOW DERIVATIVE WAVEFORM

T_o	The time of glottal opening.
α	Factor that determines the ratio of E_e to the peak height of the positive portion of the glottal flow derivative.
ω_o	Frequency that determines flow derivative curvature to the left of the glottal pulse; also determines how much time elapses between the zero crossing and T_e.
T_e	The time of the maximum negative value of the glottal pulse.
E_e	The value of the flow derivative at time T_e.
β	An exponential time constant which determines how quickly the flow derivative returns to zero after time T_e.
T_c	The time of glottal closure.

$$\begin{aligned} v_{\mathrm{LF}}(t) &= 0 && 0 \le t < T_o \\ &= E_o e^{\alpha(t-T_o)} \sin[\omega_o(t - T_o)] && T_o \le t < T_e \\ &= -E_1[e^{-\beta(t-T_e)} - e^{-\beta(T_c - T_e)}] && T_e \le t < T_c \end{aligned} \tag{4}$$

where $E_1 = E_e/(1 - \exp(-\beta(T_c - T_e)))$, and where the time origin, $t = 0$, is the start time of the closed phase (also the end of the return phase of the previous glottal cycle which we later also denote by T_{c-1}), T_o is the start time of the open phase (also the end of the closed phase), T_e is the start time of the return phase (also the end of the open phase and time of the glottal pulse), and T_c is the end time of the return phase (also the beginning of the closed phase of the next glottal cycle). Three of the parameters, E_o, ω_o, and α, describe the shape of the glottal flow during the open phase. The two parameters E_e and β describe the shape of the return phase. Because at time $t = T_e$, E_o can be calculated from E_e using the relation $E_o = E_e/e^{\alpha(T_e - T_o)} \sin\omega_o(T_e - T_o)$, we reduce the number of waveshape parameters to four, i.e., ω_o, α, E_e, and β. Observe that we estimate E_e, not E_o or E_1; E_e is the absolute value of the negative peak for which an initial estimate is easily obtained. The resulting four waveshape parameters do not include the glottal timing parameters; therefore, the times T_o, T_e, and T_c must also be made variables. We thus have a seven-parameter model to describe the glottal flow, with the four

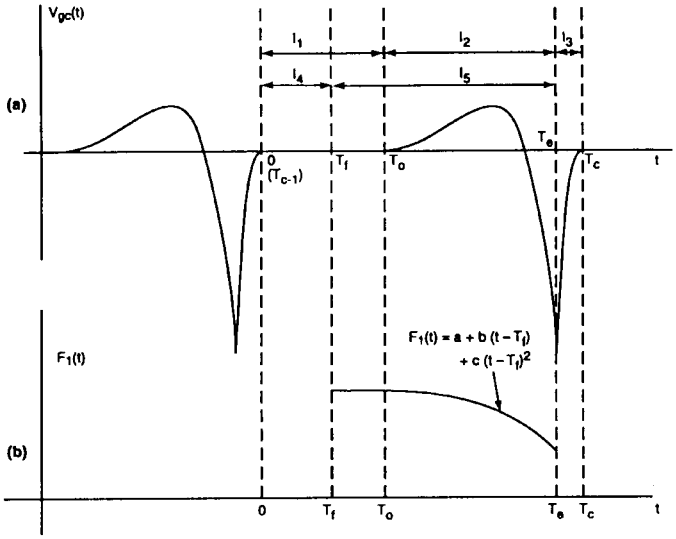


Fig. 5. Time intervals used for fine structure features: (a) glottal flow derivative and (b) frequency modulation associated with glottal ripple.

standard LF model parameters, and three parameters indicating the timing of the flow components. A descriptive summary of the seven parameters of the LF model is given in Table I.

2) Fine Structure: In determining fine-structure features, we define five time intervals within a glottal cycle, as illustrated in Fig. 5(a), over which we make energy measurements on fine structure. The first three intervals correspond to timing within the LF model of coarse structure, while the last two intervals come from timing measurements made on formant modulation to be described in Section III. The time intervals are given as follows.

1) $I_1 = [0, T_o)$ is the closed phase for the LF model;
2) $I_2 = [T_o, T_e)$ is the open phase for the LF model;
3) $I_3 = [T_e, T_c]$ is the return phase for the LF model;
4) $I_4 = [0, T_f)$ is the closed phase for formant modulation;
5) $I_5 = [T_f, T_e)$ is the open phase for formant modulation;

over which we make the energy measures $E^i = \int_{I_i} v_{fc}(t)^2\,dt = \int_{I_i} [v_g(t) - v_{gc}(t)]^2\,dt$. We have given two different pairs of open- and closed-phase estimates, one according to the LF model, and a second according to formant modulation. The latter is motivated by the observation that when the vocal folds are not fully shut during the closed phase, ripple can begin prior to the end of the closed phase as determined by the LF model. Therefore, aspiration or ripple may occur anywhere over the glottal cycle. The open- and closed-phase estimates using formant frequency modulation allow additional temporal resolution in the energy characterization of fine structure.

In addition to the energy measures E^i over the five time intervals, a frequency-domain measure of ripple is used. Observation of calculated formant motion over the open phase $I_5 = [T_f, T_e)$ led us to model the modulation of the first formant frequency through the use of a parabola of the form $F_1(t) = a + b(t - T_f) + c(t - T_f)^2$ over this interval [Fig. 5(b)]. The parameter a consists of two terms, i.e., $a = \overline{F}_1 + \Delta F_1$. The term $\overline{F}$ is the average formant value during the closed phase I_4 and thus reflects primarily the vocal tract. ΔF is the offset of the average formant value during the open phase I_5 from the average formant value during the closed phase. The change in the formant value over the open phase I_5 is given by the two remaining parameters b and c. The three parameters ΔF_1, b, and c reflect primarily source-filter interaction.

III. Estimation of the Glottal Flow Derivative Waveform

The glottal flow derivative estimate is obtained by inverse filtering the speech waveform with a vocal tract filter derived over a glottal closed-phase estimate, according to a "stationary" region of formant modulation. We first find an approximate location of the glottal pulse using an initial pass at inverse-filtering the speech waveform. This pulse location is then used to define a region over which formant modulation is computed via the covariance method of linear prediction. Finally, statistics are derived on the formant modulation function for determining a closed-phase estimate. We begin with a review of the covariance method of linear prediction.

A. Covariance Method

According to the linear filtering view of speech production, in discrete time, the speech waveform $s[n]$ is the output of the vocal tract filter with impulse response $h[n]$, excited by the glottal flow derivative $v_g[n] = \dot{u}_g[n]$ which includes lip radiation.[4] For an all-pole response $h[n]$ with transfer function $H(z) = 1/(1 - \sum_{i=1}^{p} a_i z^{-i})$, we have

$$s[n] = \sum_{i=1}^{p} a_i s[n-i] + v_g[n]. \tag{5}$$

To estimate the filter coefficients a_i from the speech signal $s[n]$, the covariance method of linear prediction is used. Covariance-based linear prediction is preferred over the autocorrelation method because, when the waveform follows the assumed all-pole model, the analysis window over which the prediction error is defined results in the correct solution for any window length M greater than the prediction order [26]. The covariance method solution is described by $\mathbf{\Phi}\vec{\alpha} = \vec{\psi}$, where the (i, j)th term of matrix $\mathbf{\Phi}$ is given by $\phi_{i,j} = \sum_{n=0}^{M-1} s[n-i]s[n-j]$: $1 \le i, j \le p$, the vector $\vec{\psi} = [\phi_{0,1}, \phi_{0,2}, \cdots, \phi_{0,p}]^T$, and the solution vector $\vec{\alpha} = [\alpha_i, \alpha_2, \cdots, \alpha_p]^T$. This matrix problem can be solved efficiently by using Cholesky decomposition [26].

B. Determination of a Closed Phase Region Through Formant Modulation

During the closed phase, the absence of source-filter interaction will result in no or little formant modulation, while during the open phase the changing glottal area will result in nonstationary formants. The glottal closed phase is identified as the time region during which formant modulation is stationary. The first step in determining the closed phase is to identify the approximate timing of glottal pulses through peak picking of a whitened speech waveform. We whiten

[4] In discrete time, the derivative operation by the lip radiation can be approximated by a first backward difference, $\delta[n] - \delta[n-1]$ [26].

the speech waveform by inverse filtering with the covariance method solution using a one pitch-period frame update and a two pitch-period analysis window during voicing. An all-pole model of order 14 was used in the covariance method, and the glottal poles are not removed in this first pass. The pitch and voicing estimates are generated with a sinusoidal-based pitch estimator [22]. The goal of estimating glottal pulses is to find a time that occurs during the open phase of the glottis. This information provides a window in which the glottal closure will occur shortly after the identified pulse, and the glottal opening will occur somewhere during the window, depending on the open quotient. As the majority of the open phase occurs before the glottal pulse, and multiple pulses sometime occur in the presence of zeros [2], [6], this estimate of the timing of the glottal pulse is somewhat biased to the left by selecting previous peaks of similar amplitude as the initial output of the peak picker as the glottal pulse estimate. This estimate of the glottal pulse will be refined during estimation of the coarse structure in Section IV. A glottal closed-phase estimate is then obtained using an estimate of the formant modulation, calculated between successive glottal pulses. Formant modulation can be exploited for this purpose because a formant change occurs in going from the glottal closed phase to open phase in which source/vocal tract interaction occurs.[5] Therefore, by tracking the first formant[6] within a glottal cycle, an approximate onset time of formant motion can be identified providing an estimate of the start time of the glottal open phase.

To measure the formant frequencies within a glottal cycle, a sliding covariance-based linear prediction analysis with a one-sample shift is used. The size of the rectangular analysis window is constrained to be slightly larger than the prediction order, while still being several times smaller than the pitch period. In particular, the length in discrete time of the sliding covariance analysis window, denoted by N_w, is chosen for each frame to be $N_w = N/4$, except constrained by upper and lower bounds of $p + 3 \leq N_w \leq 2p$, where N is the length of the pitch period as calculated by the time between the glottal pulses identified above, and p is the order of the linear prediction analysis; as in the first pass, an all-pole model of order $p = 14$ is used. Observe that the pitch period $N = N_c$ which is the sample-time counterpart to the continuous time variable T_c. Window lengths less than $p+3$ cause occasional failure of the Cholesky decomposition, while using more than $2p$ samples will not make the estimate significantly more accurate but will decrease the time resolution. The first analysis window begins immediately after the glottal pulse of the previous glottal cycle, while the last analysis window ends at the sample before the glottal pulse of the current glottal cycle. There are thus a total of $N - N_w$ windows over each glottal cycle; a vocal tract estimate is found for each window by the sliding covariance analysis. Formant tracking is performed within glottal cycles using the formants calculated from the vocal tract estimates, and provides formant trajectories through glottal closed and open phases.[7]

While a mathematical framework for calculating the expected modulation of the formant frequencies was developed in Section II-B, we have observed a variety of formant motions. Due to the possibility that the vocal folds may never completely close, the degree of formant modulation during the closed phase will vary from speaker to speaker. This varying amount of formant modulation during the closed phase makes it difficult to set a threshold on the degree of formant modulation that indicates the onset of the glottal open phase. We have therefore chosen to take a statistical approach to identifying the glottal closed phase. The approach involves first finding a small region of sequential formant samples in which the formant modulation is minimal. To do so, we define a *formant change* function as the sum of the absolute difference between successive formant estimates over a five-sample interval, i.e.,

$$D(n_o) = \sum_{i=n_o}^{n_o+4} |F(i) - F(i-1)|: 1 \leq n_o < N - N_w - 5. \tag{6}$$

The argument n_o, the first sample of the five-sample region, is varied to minimize $D(n)$; $F(n)$ represents the formant values calculated for each sample in the glottal cycle, and N is the number of samples in the glottal cycle. The size of five samples is selected to ensure meaningful statistics for determining the initial "stationary" formant region [24].

Once an initial stationary region is identified, the mean and standard deviation of the first formant within this small region are calculated, and the region is grown based on the following procedure. To extend the region to the right (Fig. 6), if the next sample is less than two standard deviations from the mean,[8] it is included in the stationary region and the mean and standard deviation are recalculated before continuing on to test the next sample. Motivation for the use of the standard deviation of formant motion in the closed phase is to obtain a statistic that indicates the onset of glottal opening. A slightly different algorithm is used to extend the region to the left. The final mean and standard deviation from extending the stable region to the right are kept constant, and the region is grown to the left until a sample is more than two of these standard deviations from this mean. The closed phase is considered to include every speech sample which was used to calculate the stable formant values.[9]

[5]The formant frequencies and bandwidths are expected to remain constant during the closed phase but will shift during the open phase. For voices in which the vocal folds never completely close, such as breathy voices, a similar formant modulation will occur. For such voices, during the nominally closed phase, the glottis should remain approximately constant, resulting in a fixed change on formant frequencies. When the vocal folds begin to open, the formants will move from their relatively stationary values during the closed phase.

[6]The first formant was found to be more stationary than higher formants during the closed phase and exhibits a more observable change at the start of the open phase [24].

[7]The first four formants are tracked by their frequency using a Viterbi search. The search space is the four lowest poles with bandwidth less than 500 Hz calculated by the sliding covariance analysis. The cost function is the variance of the formant track including the proposed pole to be added to the end of the track [24].

[8]Under a Gaussian assumption, a value of more than two standard deviations from the mean will occur with a probability of less than 4%.

[9]There are two primary reasons for the different techniques used to identify the glottal opening and closure. First, after the region has been extended to the

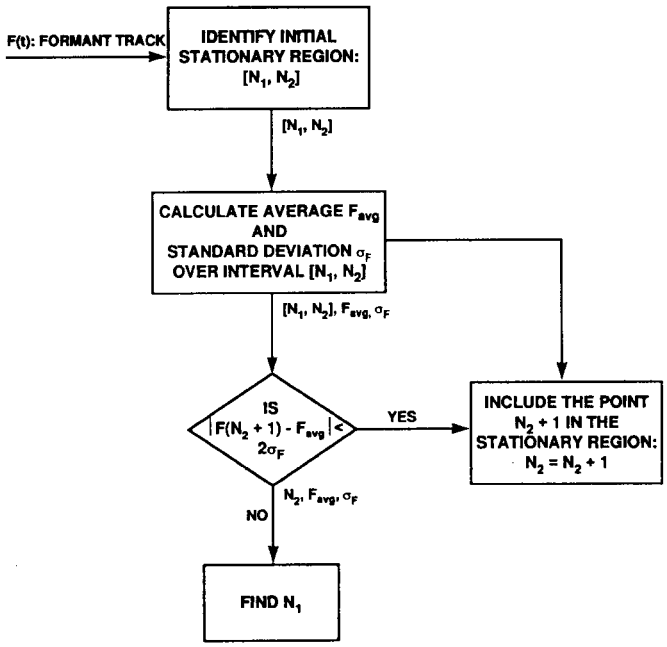


Fig. 6. Glottal closed phase is estimated by growing a small initial region in which the first formant frequency is stationary. The procedure for extending the region to the right is illustrated in this flow chart. A new sample is included if its value is greater than two standard deviations from the mean of the formant in the region.

The result of this algorithm is a region, denoted by $[N_1, N_2]$, over which the formant modulation is considered stationary, and with which a vocal tract estimate is obtained for inverse filtering. Identifying a small initial stationary region, from which a larger stationary region is grown, allows the algorithm to adapt to the variability of the formant. For example, if some ripple is present at the beginning of the glottal cycle due to source/vocal tract interaction, the initial standard deviation will reflect this variability. As the vocal folds open beyond their position during the closed phase, source/vocal tract interaction increases the degree of ripple, and the standard deviation limits will be exceeded, indicating the onset of the open phase.

Recall that in Section II-C, we defined for formant modulation closed- and open-phase intervals $I_4 = [0, T_f)$ and $I_5 = [T_f, T_e)$, which allow, over a glottal cycle, additional energy measures, beyond that of the LF model, for representing fine structure. For the two intervals, the time T_f equals the end time of the stationary formant region $[N_1, N_2]$, i.e., $T_f = N_2$, where T_f is assumed discretized in time. For the open-phase interval I_4, however, we set the starting time to zero (the beginning of the glottal cycle) because the start time N_1 of the stationary formant region falls within the previous glottal cycle.[10]

right to identify the glottal opening, the statistics have been estimated from sufficient data and extending the window to the left will not improve those estimates. More importantly, we have found that the glottal opening tends to result in sudden formant shifts, while gradual formant shifts are found when extending the region to the left toward glottal closure. If we attempted to update the statistics during a gradual change in the formant estimate, the statistics would likely incorporate this change, and glottal closure would not be identified.

[10] The stationary region includes every speech sample used in its calculation, so that its starting time N_1 is given by the starting time of the window, of length N_w, required in computing the first formant in the stationary region.

TABLE II
AVERAGE SNR'S FOR SEVERAL POTENTIAL MEASURES USED IN IDENTIFYING THE GLOTTAL OPENING. THE CLOSED PHASE WAS IDENTIFIED USING THE FIRST FORMANT FREQUENCY

SNR Measure	Males	Females
F_1 Freq	161	155
F_1 BW	8.7	4.2
F_2 Freq	8.3	5.7
F_2 BW	1.2	0.9

TABLE III
AVERAGE SNR'S FOR SEVERAL POTENTIAL MEASURES USED IN IDENTIFYING THE GLOTTAL OPENING. THE CLOSED PHASE WAS IDENTIFIED USING THE SECOND FORMANT FREQUENCY

SNR Measure	Males	Females
F_1 Freq	25.1	13.7
F_1 BW	6.3	2.4
F_2 Freq	42.7	59.4
F_2 BW	2.2	2.0

Observe that according to the theory presented in Section II, frequencies and bandwidths of all formants will exhibit modulation across the glottal open and closed phases so that any of these formant parameters may be used in determining the closed-phase region. In general, both the formant frequencies and bandwidths tend to increase at the onset of the open phase, while they remain relatively constant during the closed phase. Krishnamurthy [20], for example, has indicated that *average* formant bandwidth may exhibit a greater difference in the glottal open phase than average formant frequency. We have found, however, that the *instantaneous* first formant is most stable during the closed phase and exhibits the most observable change at the start of the open phase. This desirable property of the first formant, in light of the other formant and bandwidth features, is due perhaps to the relatively large energy of the first formant and because linear predictive analysis has greater difficulty in estimating formant bandwidths than formant frequencies. In addition, because the formant frequencies vary as the derivative of the glottal area, and bandwidths vary linearly with glottal area, the motion of the formant frequency will be quicker. A more quantitative justification for use of the first formant is illustrated in Tables II and III that show a measure of signal-to-noise ratio (SNR) for various statistics which could be used to identify the closed phase. The SNR is calculated as the ratio of the average variance at the start of the open phase, over a window five samples in duration, to the average variance over the closed phase for a large subset of TIMIT. We can think of this ratio as an SNR in the conventional sense because it compares the energy of the aspiration noise and ripple during the closed phase with the energy in the onset of the glottal flow derivative. The closed phase was determined for Table II using the frequency of the first formant as the measure of formant modulation, while for Table III the frequency of the second formant was used. The SNR for the F1 frequency in the first table is higher than the SNR for the F2 frequency in the second table. This indicates that the change in F1 frequency at the boundary of the identified closed phase is more noticeable than the change in F2 frequency at the boundary.

C. High Pitch Speakers: Using Two Analysis Windows

For high pitch speakers, it is possible that the above technique will require too large a sliding analysis window in attempting to determine a closed phase. In particular, the minimum length of the sliding covariance window is 17 samples (the lower bound of $p+3=17$ from the previous section), while the minimum size of the initial stationary region is five sequential sliding covariance windows, which will cover a total of 21 samples. At a 10 kHz sampling rate, this corresponds to a minimum closed-phase duration of 2.1 ms. A speaker with a fundamental frequency of 200 Hz and an open phase 70% of a pitch period will have a closed phase of only 1.5 ms $=$ 0.30/200 Hz. Many female speakers will accordingly have closed phases with duration less than 2.1 ms. To address this problem, we use a covariance analysis that is based on two windows across two successive pitch periods.[11]

Assuming that the rate of change of the vocal tract is dependent on time and not on the number of pitch periods, the vocal tract variation over two frames for a 200 Hz voice is approximately the same as one frame of a 100 Hz voice, since both last for 10 ms. By splitting the sliding covariance analysis window into two parts, each one need be slightly larger than half the desired linear prediction order, which results in a minimum identifiable closed-phase duration of 1.3 ms, five sequential windows each half the size of the standard minimum window length of 17 samples. Because this technique is dependent on stationarity of both the vocal tract and the source across multiple pitch periods, it is only used when the pitch period is small (chosen empirically at 6.5 ms).

As an extension to the covariance method of linear prediction of Section III-A, two windows of speech data can be used to calculate the matrix $\mathbf{\Phi}$ and the vector $\vec{\psi}$,

$$\phi_{i,j}=\sum_{n=M_1}^{M_1+L_1-1}s[n-i]s[n-j]+\sum_{n=M_2}^{M_2+L_2-1}s[n-i]s[n-j]\colon 1\le i,j\le p \quad (7)$$

where

M_1 start of the first region
L_1 length of the first region
M_2 start of the second region
L_2 length of the second region.

The only change required to convert the standard covariance linear prediction procedure into a two-window procedure is this change in the calculation of the matrix $\mathbf{\Phi}$. The properties of the matrix $\mathbf{\Phi}$ still hold as long as the windows are nonoverlapping, allowing efficient solution by Cholesky decomposition.

D. Examples

The example in Fig. 7 illustrates the waveforms obtained in deriving a glottal flow derivative estimate. The whitened speech waveform, obtained by the initial pass at inverse filtering with a vocal tract filter derived from an analysis window over multiple pitch periods, consists of a series of sharp negative-going peaks at roughly the location of glottal pulses. The flow derivative estimate, on the other hand, obtained from inverse filtering with a vocal tract estimate derived from the covariance method over the stationary region of formant modulation, shows a clear view of the closed, open, and return phases as well as the glottal pulse and ripple component overriding the coarse glottal flow derivative.

Fig. 8 shows several other examples of glottal flow derivative estimates that result from inverse filtering with a vocal tract estimate derived from the covariance method over the stationary region of formant modulation. Each panel shows

[11] The use of multiple pitch periods in analysis was independently proposed by Yegnanarayana and Veldhuis in a recent publication [35].

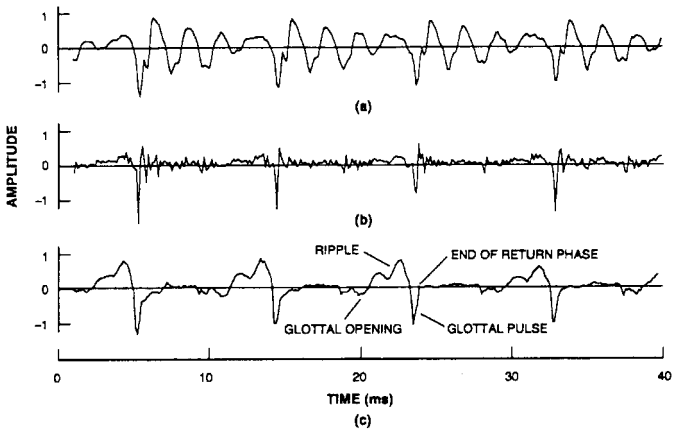


Fig. 7. Example of estimation of glottal flow derivative: (a) original speech; (b) pitch synchronous whitened speech, used to identify the closing of the glottis by searching for the largest pulse; (c) estimated glottal flow derivative from the closed phase analysis.

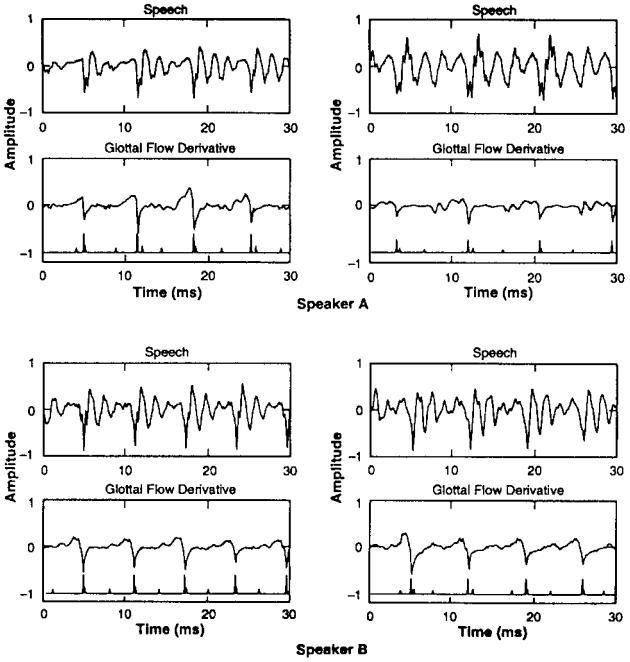


Fig. 8. Several examples of estimated glottal flow derivatives. The speech signals are above the corresponding glottal flow derivative waveforms. The top panel represents Speaker A and the bottom panel represents Speaker B. The two examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels.

two examples for a particular speaker. In each example, the speech waveform appears above the glottal flow derivative waveform. The examples in the first column come from the vowel in the word "had" while those in the second column come from two different vowels. Small pulses along the time axis of the glottal flow derivative waveforms show boundaries of the stationary formant modulation region $[N_1, N_2]$, while the large pulses represent the estimate of the time of the glottal pulse derived from peak-picking the initial inverse-filtered waveform. The examples demonstrate speaker dependence of the flow, as well as variety in the flow for each speaker, characteristics that will be further described in the following two sections on coarse and fine structure estimation.

A consequence of our inverse filtering is an approximate "separation" of the effect of the vocal tract filter out of the speech waveform. Indeed, the inverse-filtered speech is essentially unintelligible, perceived largely as a "buzz." Although we do not currently have a quantitative way of measuring this separation, a frequency-domain view confirms that negligible vocal tract formant energy is present; the inverse-filtered spectrum typically consists of a smooth lowpass function with an occasional weak peak in the vicinity of the first formant, due to the presence of a ripple component, consistent with the Ananthapadmanabha and Fant [1] theory of nonlinear source/vocal tract interaction of Section II-B.

IV. Estimating Coarse Structure

With the source waveform estimate, we now estimate parameters of the coarse model component of the glottal flow derivative.

A. Formulation of the Estimation Problem

The seven parameters of the LF model of Section II-C to be estimated for each glottal cycle were summarized in Table I. A least-squared error minimization problem can be formulated to fit the LF model of (4) to the glottal flow derivative waveform. Denoting the glottal flow derivative estimate by $\hat{v}_g[n]$, the error criterion is given as

$$\begin{aligned} E(\vec{x}) = &\sum_{n=0}^{N_o-1} \hat{v}_g^2[n] \\ &+ \sum_{n=N_o}^{N_e-1} (\hat{v}_g[n] - E_o e^{\alpha(n-N_o)} \sin[\Omega_o(n-N_o)])^2 \\ &+ \sum_{n=N_e}^{N_c-1} (\hat{v}_g[n] - E_1[e^{-\beta(N_c-N_e)} - e^{-\beta(n-N_e)}])^2 \end{aligned} \tag{8}$$

where N_o, N_e, and N_c are sample-time[12] counterparts to the continuous-time variables T_o, T_e, and T_c of (4), Ω_o is a discrete-time frequency, and $\vec{x}$ is a vector of the seven model parameters. The error $E(\vec{x})$ is a nonlinear function of the seven model parameters with no closed-form solution, and thus the problem is solved iteratively using a nonlinear least-squares algorithm with calculation of first- and second-order gradients.[13]

Standard methods for solving nonlinear least-squares problems, such as the Gauss–Newton method, are not adequate when the minimum error $E(\vec{x})$ is large [11]. This is often the case in fitting the LF model to the glottal flow derivative waveform because ripple and aspiration, not represented by the LF model, manifest themselves in $E(\vec{x})$. An algorithm more amenable to large optimization error is an adaptive nonlinear least-squares regression technique, referred to as the NL2SOL algorithm.[14] This algorithm also has the advantage of allowing bounds to enable parameters to be limited to physically reasonable values.

B. NL2SOL Algorithm

In the NL2SOL algorithm, a residue is defined as $r_i(\vec{x}) = m_i(\vec{x}) - y_i$, where $\vec{x}$ is a vector of the parameters to be solved, y_i is the data to be fitted, and $m_i(\vec{x})$ is the value of a function at point i with parameters $\vec{x}$. The summed squared residue to be minimized is expressed as

$$f(\vec{x}) = \tfrac{1}{2} \sum_{i=1}^{N} r_i^2(\vec{x}) = \tfrac{1}{2} \vec{R}(\vec{x})^T \vec{R}(\vec{x}) \tag{9}$$

with $\vec{R}(\vec{x}) = [r_1(\vec{x}), r_2(\vec{x}), \cdots, r_N(\vec{x})]$. The specific value of $\vec{x}$ that minimizes (9) is written as $\vec{x}^*$ which will be considered a local minimum when a convergence criteria is reached.

To minimize $f(\vec{x})$, we iteratively change the parameter vector $\vec{x}$, the result of which we denote by $\vec{x}_k$. The iteration is based on the Taylor series expansion of $f(\vec{x})$ around the point $\vec{x}_k$ given by

$$\begin{aligned} f(\vec{x}) = &\tfrac{1}{2} \vec{R}(\vec{x}_k)^T \vec{R}(\vec{x}_k) + (\vec{x} - \vec{x}_k)^T \nabla f(\vec{x}_k) \\ &+ \tfrac{1}{2} (\vec{x} - \vec{x}_k)^T \nabla^2 f(\vec{x}_k)(\vec{x} - \vec{x}_k) + \cdots \end{aligned} \tag{10}$$

with the first-order gradient of $f(\vec{x})$ given by $\nabla f(\vec{x}) = J(\vec{x})^T \vec{R}(\vec{x})$ where the (i, l)th element of the Jacobian matrix $J(\vec{x})$ of the vector $\vec{R}(\vec{x})$ is given by $j_{i,l}(\vec{x}) = \partial r_i(\vec{x}) / \partial x_l$, i.e., the $(i, l)^{th}$ element of $J(\vec{x})$ is the partial derivative of $\vec{R}(\vec{x})$ at the point i with respect to the lth element of the parameter vector $\vec{x}$. The second-order gradient of $f(\vec{x})$, referred to as the Hessian, is $\nabla^2 f(\vec{x}) = J(\vec{x})^T J(\vec{x}) + \sum_{i=1}^{N} r_i(\vec{x}) \nabla^2 r_i(\vec{x})$.

With the Taylor series approximated by a finite number of terms, the minimum of $f(\vec{x})$ is determined iteratively through the following procedure.

1) Start with an initial guess for $\vec{x}^*$, $\vec{x}_0$.
2) Calculate the Taylor series expansion of $f(\vec{x})$ around the point $\vec{x}_k$, where k is the current iteration number.

[12] For notational convenience, the sampling time interval is normalized to unity.

[13] An iterative approach was previously applied to estimation of LF model parameters [7]; the method does not compute LF parameter gradients, however, manually adjusting parameters on each iteration to minimize error. The approach also requires that the closed-phase estimate be determined by electroglottographic (ECG) analysis.

[14] The NL2SOL algorithm is the Association for Computing Machinery (ACM) algorithm 573 [11], [12]. The acronym derives from its being a NonLinear secant approximation To the Second-Order part of the Least squares Hessian.

3) Choose $\vec{x}_{k+1}$ as the parameter vector which minimizes the Taylor series. The Taylor series expansion, being a linear function of the powers of $\vec{x}$, can be minimized explicitly.
4) If the difference between $\vec{x}_k$ and $\vec{x}_{k+1}$ is small, or the value of $f(\vec{x})$ is small at $\vec{x}_{k+1}$, consider $\vec{x}^*$ to equal $\vec{x}_{k+1}$. If not, return to step 2 to refine the estimate of $\vec{x}^*$.

Using the first-order term of the Taylor series makes the assumption that $f(\vec{x})$ can be adequately modeled by a linear function, giving the Newton method. Including the second-order term makes the assumption that $f(\vec{x})$ can be modeled by a quadratic; in this case, using only the first component of the Hessian of $f(\vec{x})$, i.e., $J(\vec{x})^T J(\vec{x})$, gives the Gauss–Newton method. The NL2SOL algorithm uses two methods to minimize $f(\vec{x})$, the Gauss–Newton method and the Gauss–Newton method with the complete Hessian of $f(\vec{x})$. The algorithm begins with the Gauss–Newton method. On succeeding iterations, one of the two methods is selected; if the previous iteration fails to achieve a reduction in the summed squared residual (9), then the algorithm switches to the alternate method. This approach is more effective with large optimization error than either method alone [11].

In applying the NL2SOL algorithm to the glottal flow derivative waveform, the parameter vector $\vec{x}$ consists of the seven LF-model parameters, and the vector $\vec{R}$ is the difference between the model and the waveform, with one element of $\vec{R}(\vec{x})$ for each time sample. The implementation of the algorithm takes as input the vector $\vec{R}(\vec{x})$ and requires the first- and second-order gradients of $f(\vec{x})$ and thus calculation of the Jacobian and the Hessian. Calculation of the Jacobian $J(\vec{x})$ requires evaluation of the first-order partial derivatives of the LF model equations, which can be found in closed form,[15] and also provides the first component of the Hessian, $J(\vec{x})^T J(\vec{x})$. Although in theory the second component of the Hessian, $\sum_{i=1}^{n} r_i(\vec{x})\nabla^2 r_i(\vec{x})$, can also be found in closed form, the large number and complexity of the required second-order partial derivatives make this solution impractical; consequently, the second term of the Hessian was approximated using finite differences [24].

In order to ensure physically reasonable parameter values, we set bounds on the parameters. For example, if the value Ω_o is less than π, the model will have no negative flow derivative during the open phase which is inconsistent with a negative going glottal pulse. Another example of an unrealistic condition is the parameter E_e taking on a positive value or a value near zero. Therefore, π and zero are the lower bounds for estimates of the model parameters Ω_o and E_e, respectively. Such constraints are allowed internally[16] within the iterative NL2SOL algorithm [11]. When a resulting model parameter estimate is too close to its bound, in particular a constraint of 1% of its bound obtained experimentally, we consider data for that frame to be unreliable. In Section VI, we refer to such parameters as *singularities*; frames with singularities are discarded and the model parameter estimates are not used in speaker identification.

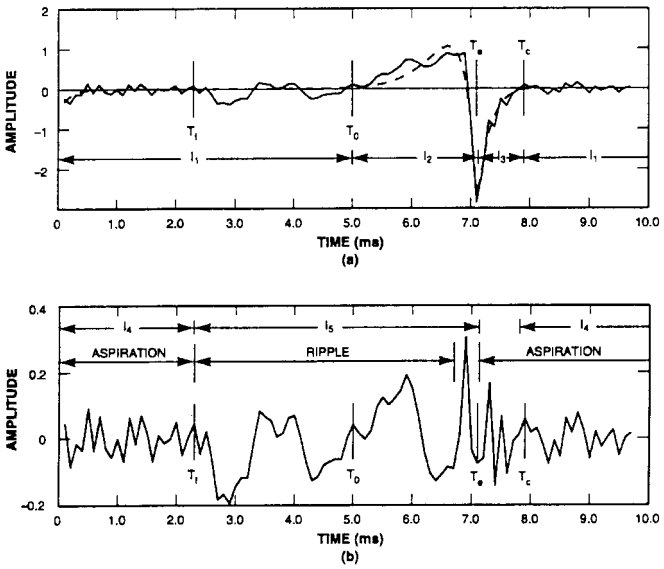


Fig. 9. Example of a glottal flow derivative estimate and its coarse and fine structure: (a) estimated glottal flow derivative (solid) and overlaid LF model, i.e., the coarse structure (dashed); (b) fine structure obtained by subtracting the coarse structure from the glottal flow derivative estimate. Aspiration and ripple are seen in different intervals of the glottal cycle.

C. Examples

Fig. 9(a) shows an example of the coarse-structure estimate (dashed) superimposed on the glottal flow derivative estimate (solid), along with the timing estimates T_o, T_e, and T_c of the LF model. The interval $I_1 = [0, T_o)$ contains both aspiration and ripple, while $I_2 = [T_o, T_e)$ contains significant flow, after which we see the occurrence of a sharp glottal pulse and a gradual return phase $I_3 = [T_e, T_c)$. The residual [Fig. 9(b)], formed by subtracting the coarse structure from the glottal flow derivative estimate, forms the fine-structure estimate to be used in the following section. The starting time of the open phase, T_f, according to formant modulation, is also shown.[17] In this case, the interval $I_4 = [0, T_f)$ consists of primarily aspiration, while the interval $I_5 = [T_f, T_e)$ appears to exhibit ripple, but as yet no significant glottal flow.

Fig. 10 shows the LF model of the coarse structure extracted from the glottal flow derivative estimates in Fig. 8 of Section III-D. In each example, the estimated glottal flow derivative appears above its modeled coarse structure. In comparing the LF model timing characteristics of the flow derivative estimates of the two speakers, the first speaker typically exhibits a longer closed phase and a shorter return phase, relative to a glottal cycle. With respect to waveshape, the second speaker shows a more gradual glottal flow derivative

[15] Algorithm refinements were necessary in identifying the times N_0, N_e, and N_c, due to discontinuities of the partial derivatives at these points because of the piecewise nature of the LF model. If the partial derivatives are discontinuous, $f(\vec{x})$ will not be adequately modeled by the first two terms of the Taylor series expansion. The result of this inadequate modeling is that the NL2SOL algorithm is slower to converge, and is more likely to find a local minimum that is not the global minimum.

[16] By imposing bounds internally, the NL2SOL algorithm can avoid a diverging solution, in contrast to simply applying bounds after a solution is found.

[17] This example also illustrates the improved temporal resolution that can be gained by the timing parameter T_f in representing fine structure.

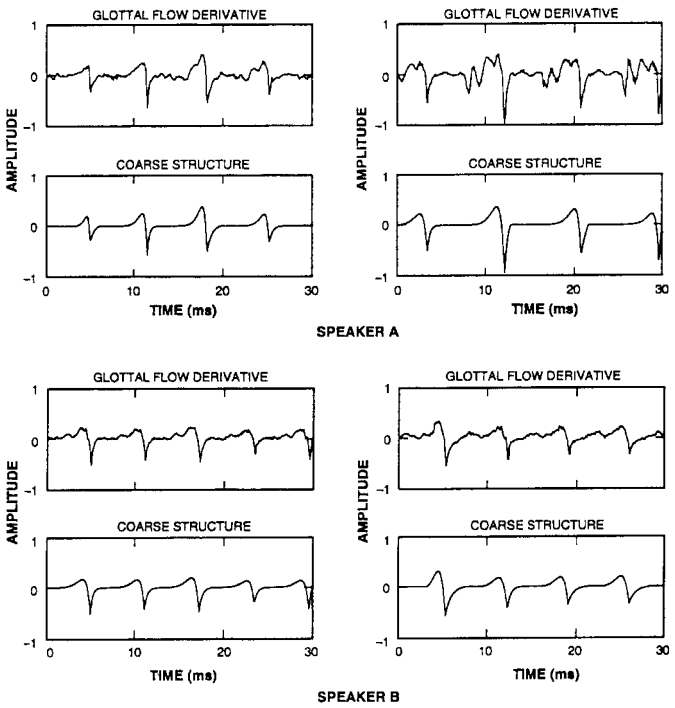


Fig. 10. Several examples of the LF model for the coarse structure in the estimated glottal flow derivative of Fig. 8. The glottal flow derivatives are shown above the corresponding model waveforms. The top panel represents Speaker A and the bottom panel represents Speaker B. The two examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels.

in entering the open phase, consistent with a shorter closed phase, and more often has a glottal pulse of smaller amplitude, relative to the peak flow derivative.

V. Estimating Fine Structure

In the previous section, we estimated the coarse structure of the glottal flow derivative waveform. Subtracting the estimated coarse structure from the glottal flow derivative waveform yields the fine structure with contributions of aspiration due to turbulence at the glottis and ripple due to source/vocal tract interaction. In Section II-C, we introduced a feature set for representing the fine structure, consisting of its energy over various time intervals, and a formant modulation function as a frequency-domain representation of ripple. This section further describes estimation of these features and illustrates them with examples.

A. Time-Domain Fine Structure

Time-domain energy measures E^i on the fine-structure source waveform were defined in Section II-C. Estimates of these energy measures are calculated over five time intervals for each glottal cycle, determined by timing measures derived from the coarse-structure model and a stationary region of formant modulation, and normalized by the total energy in the estimated glottal flow derivative waveform. The total energy is given by $E_{tot} = \sum_{n=N_{c-1}}^{N_c} \hat{v}_g^2[n]$ where $\hat{v}_g[n]$ is the glottal flow derivative estimate, and N_{c-1} and N_c are the termination time of the return phases of the previous and current glottal cycle, respectively. As an example of a normalized energy measure, the energy of the fine structure during the open phase interval $I_2 = [N_o, N_e)$, as determined by the LF model of the coarse structure, is calculated as

$$E = \frac{1}{E_{tot}} \sum_{n=N_o}^{N_e-1} (\hat{v}_g[n] - \hat{v}_{gc}[n])^2 \tag{11}$$

where $\hat{v}_{gc}[n]$ is the LF model estimate of the coarse structure. The normalized energy of the fine structure during the other four intervals is similarly calculated.

B. Modeling Ripple Through Formant Modulation

As described at the end of Section II-C, the modulation of the first formant frequency is modeled with a parabola. Parameters of the parabola, a, b, and c, are obtained by minimization of the summed squared error between the parabola and the measurement given by

$$E(a, b, c) = \sum_{n=N_f}^{N_e-1} e^2[n]$$
$$e[n] = (a + bn + cn^2) - F_1[n] \tag{12}$$

which is taken over the open-phase interval,[18] $I_5 = [N_f, N_e)$, as determined by the formant modulation, $F_1[n]$ is the measured frequency of the first formant, and N_f is the sampled-time version of T_f. Recall in Section II-C that the parameter a is considered to contain the average of $F_1[n]$, $\overline{F}_1[n]$, over the closed-phase interval $I_4 = [0, N_f)$ reflecting primarily the vocal tract. The value $\overline{F}_1[n]$ is thus subtracted from the estimate of a to reduce vocal tract influence.

The formant estimates from the sliding covariance analysis can yield occasional outliers during the open phase. In order to increase the robustness of the least-squares regression, we replace the summation in (12) by a median, giving a *least median-of-squares* estimator [29]

$$\tilde{E}(a, b, c) = \text{med}_{n=N_f}^{N_e-1} e^2[n] \tag{13}$$

where $\text{med}_{n=N_f}^{N_e-1}$ indicates the median value of the error samples from $e^2[N_f]$ to $e^2[N_e - 1]$. Half of the samples of $F_1[n]$ give a squared error less than $\tilde{E}$. This solution can be shown to be more robust in the presence of outliers, but requires a larger number of samples to accurately fit the model to the data than traditional least-squares regression [29]. To increase the accuracy of the fit with a limited amount of data, the algorithm is further refined with a weighted least-median squares version of (13) given by

$$\text{med}_{n=N_f}^{N_e-1} w_n((a + bn + cn^2) - F_1[n])^2 \tag{14}$$

where w_n are weights designed to lessen the influence of outlier points. The weights are determined using an estimate of the error of the least-median-of-squares fit

$$\hat{\sigma} = \sqrt{\text{med}(e^2[n])} \tag{15}$$

[18] Occasionally, a single noisy estimate of the first formant causes an early identification of the glottal open phase. To avoid modeling a region which includes formant values that belong in the closed phase, the start of the open phase is identified as the first of five sequential samples which are outside the standard deviation bound set for identifying the open phase.

derived from (13). In discarding outliers, the weights w_n are chosen as

$$w_n = \begin{cases} 1, & \text{if } |e[n]/\hat{\sigma}| \leq 2.5 \\ 0, & \text{if } |e[n]/\hat{\sigma}| > 2.5 \end{cases} \tag{16}$$

where the threshold 2.5 is selected to avoid loss of samples that are not outliers [29].

C. Examples

We earlier showed in Fig. 9(b) an example of fine structure obtained by subtracting the coarse-structure estimate (dashed) from the glottal flow derivative estimate (solid) of Fig. 9(a). In this case, the closed phase $I_4 = [0, T_f)$, as determined by formant modulation, consists of primarily aspiration, while the corresponding open phase $I_5 = [T_f, T_e)$ is dominated by ripple. The closed phase $I_1 = [0, T_o)$, as determined by the LF model, comprises an interval of aspiration followed by ripple, the ripple continuing into the open phase $I_2 = [T_o, T_e)$.

In the example of Fig. 11, we use the glottal flow derivative estimates of Fig. 10 of Section IV to further illustrate fine structure. In each example, the glottal flow derivative estimate is shown above the estimated fine structure, obtained by subtracting the coarse structure of Fig. 10 from the glottal flow derivative estimate. As before, each panel shows two examples for a particular speaker. The fine structure is the basis for our five energy measures, and more clearly shows the ripple and aspiration components of the flow derivative than does the flow derivative estimate. The fine-structure waveforms are scaled in amplitude to make the features more visible. In addition, the open- and closed-phase estimates, according to the stationary region of frequency modulation, are illustrated for one glottal cycle. In comparing the fine structure in the two speakers, we look for ripple, aspiration, and energy fluctuation differences within a glottal cycle.[19] In this example, the first speaker shows more prominent ripple within the open phase, while the second speaker generally shows stronger aspiration over a glottal cycle. In addition, the fine structure for the first speaker tends to have less energy in the closed-phase than in the open-phase regions, while the fine structure of the second speaker contains more steady energy across the two phases.

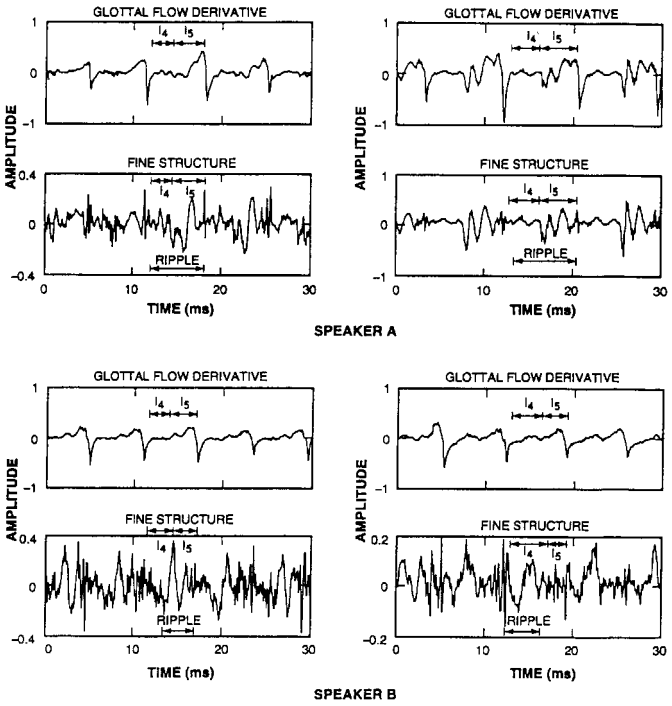


Fig. 11. Several examples of fine structure obtained from subtracting coarse structure from the estimated glottal flow derivative of Fig. 10. The glottal flow derivatives are shown above the fine structure waveforms. The top panel represents Speaker A and the bottom panel represents Speaker B. The open phase and closed phase, according to the stationary region of formant modulation, are indicated.

VI. Speaker Identification Experiments

Previous sections described estimation of the glottal flow derivative from speech and modeling the coarse and fine structure of this source waveform. We now discuss the application of the model parameters to speaker identification.

A. SID Using Gaussian Mixtures

For determining the speaker identifiability of our source features, we use a Gaussian mixture model (GMM) speaker identification (SID) system. Each Gaussian is assumed characterized by a diagonal covariance matrix [27], [28]. This choice is based on the empirical evidence that diagonal matrices outperform full matrices and the fact that the probability density modeling of an Mth-order full covariance mixture can equally well be achieved using a larger order, diagonal covariance mixture. Maximum likelihood speaker model parameters are estimated using the iterative expectation-maximization (EM) algorithm[20] [10]. Use of the GMM classifier is justified by its being an established, general classifier which acts as a hybird between standard parametric classifiers, which assume predetermined distributions, and nonparametric classifiers which typically are computationally expensive, such as k-nearest neighbors [13], [15]. It is well known that if the number of component densities in the mixture model is not limited, we can approximate virtually any "smooth" density. We have chosen 16 Gaussians in our mixture model based on experiments showing that increasing the number of Gaussians beyond 16 did not improve performance of the classifier with LF/energy features. Moreover, use of the GMM classifier is also motivated by observation of distributions of projections of individual LF feature and fine-structure energy elements (Section VI-B), indicating that a smooth distribution model such as a GMM is appropriate. Certainly, there may exist a more efficient density representation of our new features than that of a Gaussian mixture model for approximating the feature distribution. Our aim in this paper, however, is not to optimize the classifier, but rather to use an established classifier which is general enough for new features to show that the glottal features convey speaker identity information.

[19] Our energy measurements, however, do not distinguish between ripple and aspiration within a time interval.

[20] Ten iterations are sufficient for parameter convergence and a variance floor of 0.0001 was imposed.

In the use of the Gaussian mixture model, certain "unseen" or abnormal observations, that occur infrequently, are discarded in the estimation of the model mixtures. This discarding of such outliers is typically invoked in GMM-based recognition as, for example, with mel-cepstral representations of the speech spectrum. Outliers can cause problems in Gaussian mixture modeling because they will tend to "grab" Gaussians, reducing the number of Gaussians available for modeling meaningful feature values. The outlier detection in our GMM system is accomplished by detecting those input observation vectors which have an extremely low probability of being generated from the GMM currently being used [27], [28]. These detected outliers are then discarded during training and testing. While perhaps not mathematically optimal, it is an engineering solution for outlier problems. For our source feature vectors, outliers can occur with a poor estimation of the glottal flow derivative observed to result from error in the pitch or voicing estimates or in the initial glottal pulse time from peak-picking the whitened speech waveform. Outliers can also occur when the glottal flow derivative estimate is accurate, but does not follow the LF model waveshape, an example of which will be given in Fig. 14.

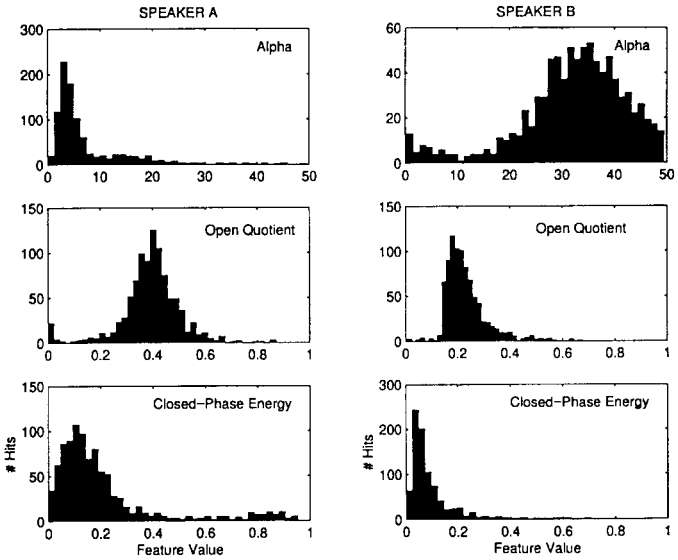


Fig. 12. Comparison of the histograms of three glottal flow features, shape parameter α, open quotient OQ, and closed-phase energy E^1 determined from the LF model, for two different male TIMIT speakers. The experiment used about 20 s of data (all the TIMIT training data) for each speaker and feature values were divided across 40 histogram bins.

B. Using Source Features for SID

Certain parameters used for SID are made a function of those used in the coarse model. For example, the parameter N_o, indicating the first sample of the open phase, as determined by the LF model, will increase without bound as we move further into an utterance. Therefore, rather than using the absolute times N_o, N_e, and N_c, we calculate the lengths of the closed, open, and return phases normalized by the length of the glottal cycle. The transformed parameters are given by

$$CQ = \frac{N_o - N_{c-1}}{N_e - N_{e-1}}; \qquad OQ = \frac{N_e - N_o}{N_e - N_{e-1}};$$
$$RQ = \frac{N_c - N_e}{N_e - N_{e-1}} \tag{17}$$

where N_{e-1} is the time of the glottal pulse and N_{c-1} is the time of the end of the return phase, both for the previous glottal cycle. This normalization also provides a means for removing pitch as a feature in SID. The waveshape parameters α, ω_o, E_e, and β are included as calculated during modeling, giving a total of seven coarse-structure parameters. The complete source feature vector is computed on voiced frames only for each pitch period. In addition, we saw in Section IV-B that the source features may reach their bounds set within the NL2SOL algorithm. As do outliers, these "singularities" can occur with a poor estimation of the glottal flow derivative or with a deviation in the flow from the typical LF model. Frames with singularities are discarded. While discarding data might generally be undesirable, we have found that it increases the accuracy of the speaker identification system by approximately 15%. Only about one third of the speech frames are used for training and testing, some being discarded by virtue of their being unvoiced. A subset of the voiced frames are then discarded due to containing singularities or outliers.[21]

Before describing the speaker identification experiments, to obtain further insight into the nature of the glottal parameters and the importance of removing frames with singularities, we give results of an initial statistical analysis. Fig. 12 shows a comparison of the histograms of three glottal flow features, the parameter α, open quotient OQ, and closed-phase energy E^1 determined from the LF model, for two different male TIMIT speakers. These parameters represent our three classes of features: glottal flow derivative shape (α), timing (OQ), and fine-structure energy (E^1). The experiment used about 20 s of data (all of the TIMIT training data) for each speaker, and feature values were divided across 40 histogram bins. In this experiment, the above frame discard procedure was invoked, thus including only those frames to be used in our speaker identification experiments, i.e., unvoiced frames and frames with singularities were discarded. We see in Fig. 12 generally "smooth" distributions with specific energy concentrations, indicating their amenity to a GMM model.[22] We also see in Fig. 12 that there is a separation of distributions of glottal features across speaker, particularly with the shape parameter α and the open quotient parameter OQ. In a second experiment, we illustrate the importance of removing frames with singularities. Fig. 13 shows a comparison of the histograms of the two glottal flow features, shape parameter

[21] The distinction between a singularity, i.e., a model parameter estimate reaching its bound, and an outlier, i.e., a model parameter estimate with very low probability, were given in Sections IV-B and VI-A, respectively. In addition, in Section V-B the term outlier is used in a different sense in reference to formant estimation.

[22] Occasionally, we have seen a strong asymmetry in a distribution, particularly with the return phase RQ and the open quotient OQ as determined by formant modulation. Both may be more efficiently modeled by sums of Rayleigh or Maxwell distributions [23], being characterized by a sharp "attack" and slow "decay."

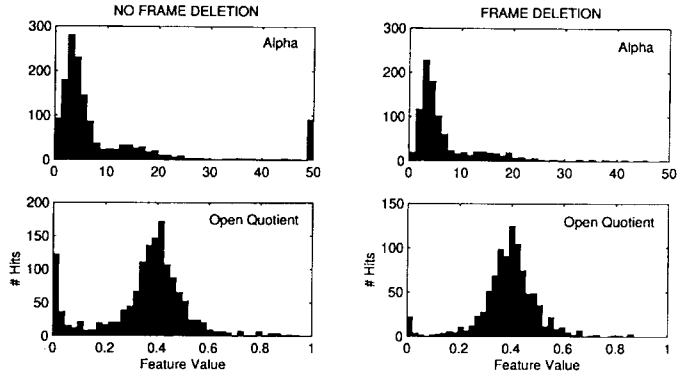


Fig. 13. Comparison of the histograms of two glottal flow features, shape parameter α and open quotient OQ determined from the LF model, with and without deletion of frames that include source feature singularities. Singularities occur when source features reach their bounds set within the NL2SOL algorithm. The experiment used about 20 s of data (all the TIMIT training data) from speaker A of Fig. 12 and feature values were divided across 40 histogram bins.

α and open quotient OQ determined from the LF model, with and without deletion of frames that include source feature singularities. The training data of Speaker A from Fig. 12 was used. As we noted earlier, singularities occur when source features reach their bounds set within the NL2SOL algorithm. In Fig. 13, we observe for the parameter α a strong component at about value 50 of the distribution when singularities are not removed; this component corresponds to frames where a NL2SOL bound is reached. Likewise, the parameter OQ has a strong singularity component at the origin. In both cases, the undesired component of the distribution is significantly reduced by not including singular frames.

In speaker identification experiments, a subset of the TIMIT database was used. This subset contains ten sentences of read speech for each speaker, recorded in a quiet room with a Sennheiser microphone. The male subset contains 112 speakers, while the female subset contains 56 speakers. For each speaker, eight of the sentences are used for training and two are used for two independent tests. As noted earlier, 16 Gaussians were used in the mixture model. Male and female sets are handled separately, as the large differences in anatomy result in cross-sex errors being very rare.

Six separate tests were conducted with the following feature sets:

1) seven LF coarse-structure model parameters;
2) five energy measures of the fine structure;
3) seven LF and five energy parameters;
4) three formant modulation parameters;
5) fourteen linear-predictive derived cepstral coefficients;
6) twelve source parameters and fourteen cepstral parameters.

In each test, the features were concatenated into one vector and passed to the SID system. The cepstral parameters consist of the first 14 coefficients of the real cepstrum as calculated by the recursion $c_i = -\alpha_i - (1/i)\sum_{k=1}^{i-1}(i-k)\alpha_k c_{i-k}$ where c_i are the cepstral coefficients, and c_0 is not used [26]. The α_k's are the estimated vocal tract parameters from the covariance method of linear prediction over a closed phase, according to the stationary region of formant modulation of Section III-B. The recursion assumes a minimum-phase filter given by the α_k's. Any maximum-phase poles, which are possible with the covariance method of linear prediction, are flipped inside the unit circle to their reciprocal location before the cepstral coefficients are calculated.

TABLE IV
SPEAKER IDENTIFICATION PERFORMANCE (PERCENT CORRECT) FOR VARIOUS COMBINATIONS OF THE SOURCE PARAMETERS FOR LARGE TIMIT DATABASE

Features		Male	Female
Coarse:	7 LF	58.3%	68.2%
Fine:	5 energy	39.5%	41.8%
Source:	12 LF & energy	69.1%	73.6%
Fine:	3 FM	7.6%	16.4%
Filter:	14 LPC cepstrum	91.0%	93.6%
Combined	26 LF, energy, cep	93.7%	92.6%

The results in Table IV show that the three categories of source parameters all contain significant speaker-dependent information. The 14 cepstral parameters which model the vocal tract, however, contain more speaker-dependent information than the 12 source parameters. For the male data subset, the combination of source and vocal tract features increased SID identification to 93.7% accuracy.[23] For the female speakers, on the other hand, outliers caused the reduction in score when source parameters are added to the vocal tract parameters.[24] The three frequency-modulation (FM) parameters show some speaker dependence as their scores of approximately 8% and 15% correct identification are well above chance (less than 1% for both cases). Including the three formant modulation parameters with the other parameters, however, lowered the scores significantly, due to a large number of outliers in the formant modulation data.

As a secondary measure of information in the glottal flow derivative waveform, we calculated the 23 mel-cepstra of these waveforms and used these coefficients as the features for SID. Both the glottal flow derivative waveform and its counterpart modeled waveform, i.e., the waveform synthesized using the LF-modeled glottal flow derivative, were processed in this manner. The results are shown in Table V. We observe that the seven LF parameters shown in the first row of Table IV better represent the modeled glottal flow derivative than the 23 mel-cepstral parameters. Thus it appears that for this one experiment, the seven LF parameters are more compatible with GMM than their mel-cepstral counterpart; however, one must consider that the mel-cepstra operation smears the spectrum

[23] In comparing accuracy rates near 100%, it is generally more instructive to compare the relative reduction in error rate. For the male subset, the error was reduced from 9% to 6.3%, a 30% reduction in error.

[24] Observe that the higher scores for females with individual feature vectors, due likely to using a smaller number of females speakers, is not reflected in the combined score. This apparent discrepancy is because, although the SID system attempts to automatically remove outliers, a larger number of outliers found for female than for male speakers resulted in more outliers being passed to SID training and testing with female speakers in the combined vectors. The relatively greater number of outliers for females, also seen to occur in a variety of GMM-based systems that use mel-cepstra, may be due to the greater difficulty in estimating the spectral envelope of females given the harmonic undersampling of the spectrum and during mixed sound classes such as voiced fricatives [4].

TABLE V
SPEAKER IDENTIFICATION PERFORMANCE (PERCENT CORRECT) FOR MEL-CEPSTRAL REPRESENTATIONS OF THE GLOTTAL FLOW DERIVATIVE (GFD) WAVEFORM AND THE MODELED GFD WAVEFORM FOR LARGE TIMIT DATABASE

Features	Male	Female
Modeled GFD	41.1%	51.8%
GFD	95.1%	95.5%

of the flow derivative and discards its phase. Table V also indicates that the mel-cepstra of the seven-parameter LF modeled glottal flow derivative contains significantly less information than that of the glottal flow derivative; this is not surprising given that the synthesized modeled glottal flow derivative is a much reduced version of the glottal flow derivative. Comparing Tables IV and V, we further see that the score of 95.1% (males) and 95.5% (females) SID for the mel-cepstra for the glottal flow derivative is a marked increase over the 69.1% and 73.6% of the 12 LF and energy measurements themselves. We might conclude from this comparison that the mel-cepstra of the glottal flow derivative estimate is more compatible to GMM than its feature-vector representation; however, the estimated glottal flow derivative contains pitch, timing of aspiration, and ripple fine structure. In addition, the ripple component of the glottal flow derivative is related primarily to the vocal tract first formant, and, therefore, as noted in Section III-D, a weak peak is occasionally seen in the glottal flow derivative spectrum in the vicinity of the first formant.

C. SID for Degraded Speech

Although the results of the previous section are fundamental for speech science, their practical importance lies in part with speaker identification in degrading environments, such as telephone speech. In preliminary experiments, to test performance of the source features on degraded speech, we first used a subset of 20 male speakers and a subset of 20 female speakers from the telephone-channel NTIMIT database [18]. For these tests, we used a 23-mel-cepstra representation of the speech signals. In one experiment, the 23-mel-cepstra representation of the synthesized LF-modeled (coarse) source waveform, rather than the LF-model parameters themselves, was used in order to provide frame synchrony and similar feature sets for speech and source. The selected 20 male and 20 female speakers are cases for which the baseline GMM SID system performs particularly poorly, achieving scores of 40.0% and 52.5% on males and females, respectively. While the LF-modeled source performs poorly on its own, i.e., 12.5% on males and 27.5% on females,[25] when combined with the mel-cepstra of the speech waveform, performance improved to 60.0% on males and 55.0% on females. Tests were performed with 32 Gaussians and by training on a 23-element vector rather than the 46-element vector that would result by combining the two vectors into one. Using this approach, each feature vector contains either speech or source information;

[25]This is in contrast to the SID results in Table V for the mel-cepstra of the LF-modeled source waveform, using the counterpart TIMIT database, of 41.1% on males and 51.8% on females.

TABLE VI
SPEAKER IDENTIFICATION RESULTS FOR MEL-CEPSTRAL REPRESENTATIONS OF THE SPEECH SIGNAL, THE GLOTTAL FLOW DERIVATIVE (GFD) WAVEFORM, THE MODELED GFD WAVEFORM, AND COMBINATIONS OF THE SPEECH AND SOURCE MEL-CEPSTRAL DATA FOR SMALL NTIMIT DATABASE

Features	Male	Female
Speech	40.0%	52.5%
GFD	25.0%	22.5%
Modeled GFD	12.5%	27.5%
Speech & GFD	57.5%	52.5%
Speech & modeled GFD	60.0%	55.0%

TABLE VII
SPEAKER IDENTIFICATION RESULTS FOR MEL-CEPSTRAL REPRESENTATIONS OF THE SPEECH SIGNAL, THE GLOTTAL FLOW DERIVATIVE (GFD) WAVEFORM, THE MODELED GFD WAVEFORM, AND COMBINATIONS OF THE SPEECH AND SOURCE MEL-CEPSTRAL DATA FOR LARGE NTIMIT DATABASE

Features	Male	Female
Speech	56.7%	66.3%
GFD	21.9%	34.5%
Modeled GFD	4.5%	16.3%
Speech & GFD	59.4%	69.0%
Speech & modeled GFD	59.8%	69.0%

we consider the mel-cepstra vector from the speech waveform and the mel-cepstra vector from the glottal flow derivative to be two independent streams of data, each vector with 23 elements. By treating them as separate vectors, we allow mel-cepstra from the speech and source to be classified separately, while also reducing the requirement of a larger training set for a 46-element vector that would result by combining the speech and source vectors. In this method, some of the 32 Gaussians are assigned to model the speech signal, while some are used to model the source signal. This same experiment was also performed on the larger NTIMIT subset used in Section VI-B, giving an improvement of roughly 3.0% SID for both males and females, representing a 5% error reduction from using only the mel-cepstra of the speech waveform. A more complete summary of the SID results for the small and large NTIMIT databases, using mel-cepstral representations, is shown in Tables VI and VII, respectively.

In closing this section, it is important to address the loss in performance with LF parameters alone in experimenting with the NTIMIT database relative to the TIMIT database. The TIMIT database was recorded with a high-quality, essentially distortionless Sennheiser microphone, while the NTIMIT database was recorded with a carbon-button microphone [18]. Because we are estimating temporal features of the glottal flow derivative, we speculate that one important source of degradation is channel phase distortion. Observe that the cepstral coefficients are nearly immune to phase change because these parameters are derived from a Fourier transform magnitude representation; phase distortion, therefore, manifests itself in the excitation function. In fact, in comparison of estimated glottal flow derivatives from the same utterance of TIMIT and NTIMIT, we have observed a change in shape in the glottal flow derivative, especially over its open phase. To test the hypothesis that phase distortion contributes to this shape change, we estimated a phase compensation by first

averaging the Fourier transform phase (computed on successive speech frames) over a paired TIMIT and NTIMIT utterance and then formed a phase difference. Applying this phase compensation to NTIMIT brings the NTIMIT glottal flow derivative closer in shape to that derived from TIMIT. Note that in spite of shape change due to phase distortion, the modeled glottal flow derivative from the NTIMIT data gives a 12.5% (males) and 27.5% (females) speaker identification accuracy. Our preliminary observations indicate that this is consistent with some preservation of LF features, particularly an approximate preservation of open- and closed-phase timing. A more comprehensive study is required of relative sensitivity of the different glottal features, as well as methods of compensation.

VII. Conclusions and Discussion

In this paper, we presented an automatic technique for estimating and modeling the glottal flow derivative waveform from speech, and applied the model parameters to speaker identification. The glottal flow derivative was estimated using an inverse filter estimated during a closed-phase estimate, determined by formant frequency modulation calculated using a sliding covariance analysis with a one-sample shift. A statistical technique, used to identify the glottal closed-phase estimate, allows this algorithm to adapt to the amount of formant modulation during the closed phase, which is dependent on the degree of glottal closure. A two-window covariance technique was developed to improve time resolution for high pitch speakers. The Liljencrants–Fant model for the glottal flow derivative was used to model the coarse structure of the glottal flow. The parameters of this model were determined for each pitch period using the NL2SOL algorithm for nonlinear least-squares regression. The fine structure of the glottal flow was represented through five energy measures and first-formant frequency modulation, modeled by a parabola using robust least-squares regression tailored to the presence of outliers. All aspects of the source model have been shown to contain speaker-dependent information on a TIMIT corpus. The coarse structure parameters contain the most information, the time domain energy measures of fine structure less information, and the frequency modulation of the first formant contains the least speaker-dependent information, though still resulting in speaker identification scores well above chance. Finally, in preliminary experiments on the NTIMIT database, the telephone-degraded counterpart to the TIMIT subset, about a 5% error reduction in SID scores is obtained when source features are added to traditional mel-cepstral measures.

There are several improvements to be made in the algorithms of this paper for estimating the glottal flow derivative and its model parameters, and using these parameters for SID. For example, a nonlinear least-squares algorithm that is better designed to handle piecewise functions should enable more accurate estimation of the LF model times T_o, T_e, and T_c. Such an algorithm, along with the reduction of singularities by more accurate glottal flow derivative estimation, would reduce the need to discard feature vectors. In terms of the fine structure, the aspiration and ripple components could be separated, perhaps using a noise/sine-wave model, and separate feature representations used in SID. The methods of this paper might also be extended in new directions. For example, the temporal change of the glottal flow derivative waveform is not included in our SID experiments. Changes in the glottal flow from period-to-period will indicate when glottal stops are used, how sudden is the onset of voicing, and the inter-period variability of the vocal fold vibration. Asymmetries in the vocal folds and jet flow through the glottis will result in a less stable pattern of vibration and flow patterns [30], [32], which would be captured through the temporal change of the glottal flow derivative. Another important direction is comparing the importance of our measured features for human and machine speaker recognition. A perception of breathiness, for example, may correspond to certain properties of coarse and fine structure in the glottal flow derivative such as a large open quotient and noisiness, respectively [19]. On the other hand, certain features useful to machines may not be useful to humans, such as the phase of the glottal ripple component. Yet, another area of research, largely unexplored, is showing the degree of correlation and separation between source and vocal tract features, and obtaining a better understanding of their relative importance. A related area is integration, with appropriate weighting and synchronization,[26] of source and vocal tract features for improved SID in degrading environments. In addition, one must overcome difficulty in estimating the glottal flow derivative from telephone-degraded speech, and in particular, as alluded to in Section VI-C, the problem posed by a time-domain algorithm, which requires phase coherence. Over the noisy, nonlinearly-degraded NTIMIT channel, it behooves us to apply channel compensation prior to estimation. It is important also to better understand the within-speaker variability of the proposed glottal features. The SID experiments thus far have used TIMIT and NTIMIT in which training and test data were collected in a single session per speaker. Robustness to intersession variability of source vs vocal tract parameters will exhibit the practicality of the approach. Toward this end, we are currently investigating source features in other databases such as switchboard.

Finally, the examples presented in this paper were chosen to illustrate certain properties of the glottal source, but are nevertheless typical examples. Fig. 14, on the other hand, shows atypical cases. The two examples show multiple points of excitation within a glottal cycle. We have found that such multiple pulses occur primarily for speakers who appear to have nearly complete glottal closure. As the vocal folds open, the change in glottal flow is large, and a "secondary" glottal pulse is generated; these secondary glottal pulses are observed to occasionally excite formants different from those corresponding to the primary excitation [24], indicating the possibility of multiple sources distributed along the vocal tract [30]. The presence of such secondary pulses may in part explain the improved SID scores achieved by measuring energy onset times in formant bands using the Teager operator

[26] The source features are calculated on a pitch period basis, while mel-cepstra are calculated using a fixed window size and shift.

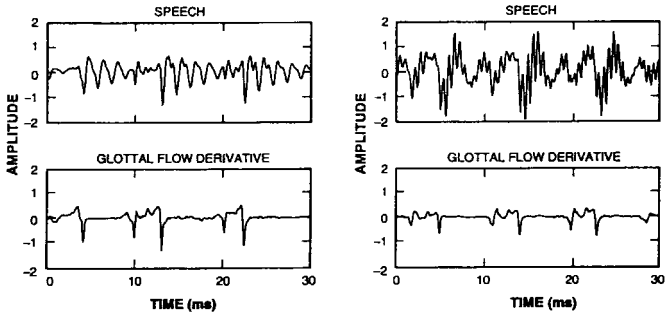


Fig. 14. Two examples of multiple pulses in the glottal flow derivative estimate. In each case, the speech waveform is shown above the glottal flow derivative waveform.

[25]. These examples, as well as other atypical cases [24], point the way to a more complete glottal flow model.

ACKNOWLEDGMENT

The authors would like to thank G. O'Leary of MIT Lincoln Laboratory, J. Kaiser of Duke University, and the three IEEE reviewers for detailed and insightful comments on the manuscript. The authors also acknowledge D. Staelin of MIT for discussions on pitch-synchronous analysis in the early stages of this work.

REFERENCES

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

[2] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-27, pp. 309–319, Aug. 1979.

[3] B. S. Atal, "Automatic recogniton of speakers from their voices, " *Proc. IEEE,* vol. 64, pp. 460–475, 1976.

[4] B. S. Atal and M. R. Schroeder, "Recent advances in predictive coding—Applications to voice speech synthesis, " in *Proc. Speech Communications Seminar,* 1974.

[5] M. Campbell and C. Greated, *The Musician's Guide to Acoustics.* New York: Schirmer, 1987.

[6] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, and Signal Processing,* vol. 37, pp. 1805–1815, Dec. 1989.

[7] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *J. Acoust. Soc. Amer.,* vol. 97, pp. 505–519, Jan. 1995.

[8] D. G. Childers and C. F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.,* vol. 41, pp. 663–671, July 1994.

[9] K. E. Cummings and M. A. Clements, "Analysis of glottal waveforms across stress styles," in *Proc. IEEE ICASSP,* Albuquerque, NM, 1990, pp. 369–372.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. R. Stat. Soc.,* vol. 39, pp. 1–38, 1977.

[11] J. E. Dennis, D. M. Gay, and R. E. Welsch, "An adaptive nonlinear least-squares algorithm," *ACM Trans. Math. Softw.,* vol. 7, pp. 348–368, Sept. 1981.

[12] ———, "Algorithm 573 NL2SOL—An adaptive nonlinear least-squares algorithm," *ACM Trans. Math. Softw.,* vol. 7, pp. 369–383, Sept. 1981.

[13] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[14] G. Fant, "Glottal flow: models and interaction," *J. Phonet.,* vol. 14, pp. 393–399, 1986.

[15] K. Fukanaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.

[16] J. He, L. Liu, and G. Palm, "On the use of features from prediction residual signals in speaker identification," in *Proc. EUROSPEECH,* 1995, pp. 313–316.

[17] C. R. Jankowski, "Fine structure features for speaker identification," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1996.

[18] C. R. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE ICASSP,* Albuquerque, NM, 1990, pp. 109–112.

[19] 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. 87, pp. 820–857, Feb. 1990.

[20] A. K. Krishnamurthy, "Glottal source estimation using a sum-of-exponentials model," *IEEE Trans. Signal Processing,* vol. 40, pp. 682–686, Mar. 1992.

[21] D. W. Farnsworth, "High speed motion pictures of the human vocal cords," *Bell Labs. Rec.,* vol. 18, pp. 203–208, 1940.

[22] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal model," in *Proc. IEEE ICASSP,* Albuquerque, NM, 1990, pp. 249–252.

[23] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1965.

[24] M. D. Plumpe, "Modeling of the glottal flow derivative waveform with application to speaker identification," S.M. thesis, Mass. Inst. Technol., Cambridge, MA, Feb. 1997.

[25] T. F. Quatieri, C. R. Jankowski, and D. A. Reynolds, "Energy onset times for speaker identification," *IEEE Signal Processing Lett.,* vol. 1, pp. 160–162, 1994.

[26] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[27] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.,* vol. 17, pp. 91–108, Aug. 1995.

[28] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing,* vol. 3, pp. 72–83, Jan. 1995.

[29] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection.* New York: Wiley, 1987.

[30] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract," in *Speech Science: Recent Advances,* R. G. Daniloff, Ed. San Diego, DA: College-Hill, 1985, pp. 73–109.

[31] P. Thévenaz and H. Hügli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Commun.,* vol. 17, pp. 145–157, Aug. 1995.

[32] I. Titze, "What's in a voice," *New Scientist,* pp. 38–42, Sept. 23, 1995.

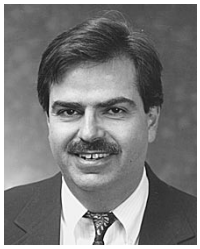
[33] J. W. van den Berg, "On the air response and the Bernoulli effect of the human larynx," *J. Acoust. Soc. Amer.,* vol. 29, pp. 626–631, 1957.

[34] 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, Aug. 1979.

[35] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Processing,* vol. 6, pp. 313–327, July 1998.

Michael D. Plumpe (M'97) was born in Arlington, VA, on March 25, 1973. He received the B.S. degree (summa cum laude) from Virginia Polytechnic Institute and State University, Blacksburg, in 1995, and the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1997.

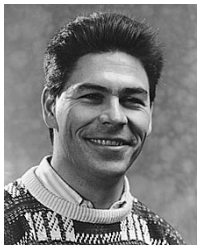
During the S.M. program, he worked on glottal flow estimation and modeling. He joined the Speech Technology Group, Microsoft Research, Redmond, WA, in 1997, working primarily on speech synthesis, focusing on improving acoustic quality. He currently works on both speech synthesis and acoustic modeling for speech recognition.



Thomas F. Quatieri (S'73–M'79–SM'87–F'98) received the B.S. degree (summa cum laude) from Tufts University, Medford, MA, in 1973, and the S.M., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1975, 1977, and 1979, respectively.

He is currently Senior Member of Research Staff, MIT Lincoln Laboratory, Lexington, involved in digital signal processing for speech, audio, and underwater sound applications, and in nonlinear signal processing. He has contributed many publications to journals and conference proceedings, written several patents, and co-authored chapters in numerous edited books. He is a Lecturer at MIT, where he has developed the graduate course in digital speech processing.

Dr. Quatieri was a member of the IEEE Digital Signal Processing Technical Committee from 1983 to 1992, served on the steering committee for the biannual Digital Signal Processing Workshop, and was Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING in the area of nonlinear systems. He is the recipient of the 1982 Paper Award of the IEEE Acoustics, Speech and Signal Processing Society for the paper, "Implementation of 2-D digital filters by iterative methods." In 1990, he received the IEEE Signal Processing Society's Senior Award for the paper, "Speech analysis/synthesis based on a sinusoidal representation," published in the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, and in 1994 won this same award for the paper "Energy separation in signal modulations with application to speech analysis" which was also selected for the 1995 IEEE W.R.G. Baker Prize Award. He is also a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, and the Acoustical Society of America.



Douglas A. Reynolds (M'86–SM'98) received the B.E.E. degree (with highest honors) in 1986 and the Ph.D. degree in electrical engineering in 1992, both from the Georgia Institute of Technology, Atlanta.

He joined the Speech Systems Technology Group, MIT Lincoln Laboratory, Lexington, MA, in 1992. Currently, he is Senior Member of Technical Staff. His research interests include robust speaker identification and verification, speech recognition, and general problems in signal classification.

Dr. Reynolds is a member of IEEE Signal Processing Society, the Speech Technical Committee, Eta Kappa Nu, and Tau Beta Pi.