Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm

Patrick A. Naylor, *Member, IEEE*, Anastasis Kounoudes, *Member, IEEE*, Jon Gudnason, *Member, IEEE*, and Mike Brookes, *Member, IEEE*

***Abstract*—We present the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) for automatic estimation of glottal closure instants (GCIs) in voiced speech. Accurate estimation of GCIs is an important tool that can be applied to a wide range of speech processing tasks including speech analysis, synthesis and coding. DYPSA is automatic and operates using the speech signal alone without the need for an EGG signal. The algorithm employs the phase-slope function and a novel phase-slope projection technique for estimating GCI candidates from the speech signal. The most likely candidates are then selected using a dynamic programming technique to minimize a cost function that we define. We review and evaluate three existing methods of GCI estimation and compare the new DYPSA algorithm to them. Results are presented for the APLAWD and SAM databases for which 95.7% and 93.1% of GCIs are correctly identified.**

***Index Terms*—Closed-phase, glottal closure, speech processing, speech segmentation.**

I. Introduction

THE classical model for speech production represents the vocal tract as an all-pole filter whose input combines a quasi-periodic source representing voiced excitation with a noise source representing unvoiced excitation. The parameters of the all-pole filter are typically estimated from the speech signal using autoregressive modelling with the assumption that the vocal tract transfer function changes sufficiently slowly to be considered constant during an analysis window of 20–30 ms. Speech production theory and modelling have been widely studied in the literature, for example [1]–[3]. In this work, we focus on voiced speech and refer to the excitation in the speech production model [1] as the voice source signal. Each cycle of voiced speech can normally be divided into a closed phase, during which air flow through the glottis is blocked by closure of the vocal folds, and an open phase during which the vocal folds are open. Our aim in this paper is to develop an automatic technique to estimate the instant of glottal closure in each cycle.

Although conventional speech analysis methods work well for many purposes, they do not attempt to deconvolve the vocal tract filter and the source signal explicitly. Consequently, the features extracted by conventional analysis methods characterize the combined effect of the excitation signal and vocal tract. In several important applications of speech processing, including speech analysis, speaker recognition, and speech coding, it is advantageous both to extract features of the vocal tract and also, separately, features of the excitation signal. This entails the blind deconvolution of the vocal tract transfer function and its input excitation signal since neither is observable individually. Such a deconvolution enables separate feature sets to be extracted for the excitation and the vocal tract in each analysis frame. These separate feature sets provide significantly more accurate analysis and modelling of speech [4]. In addition, their speaker-specific and phoneme-specific properties can be exploited in a targeted way to improve performance of several speech processing applications. For example, it has been shown [5], [6] that excitation features can be used to improve discrimination between speakers in a speaker recognition task.

Using the techniques of glottal inverse filtering [7]–[9], the voice source signal can be estimated by deconvolving the speech signal with the all-zero filter obtained from closed-phase LPC for both normal and pathological speech [10]. In this approach, least squares estimates are made of the parameters of an all-pole model of the speech signal using covariance LPC applied only to those samples of the speech signal that occur at times when the glottis is closed. Closed-phase LPC gives accurate estimates of the vocal tract transfer function, independent of the voiced excitation, provided that: 1) the temporal location of the closed-phase is known or can be estimated with sufficient accuracy; 2) an all-pole filter is a good model of the vocal tract transfer function which is assumed not to change significantly over the analysis window; and 3) sufficient data samples exist in the closed-phase to permit accurate least-squares LPC parameter estimation. Condition 2) is usually satisfied for typical closed-phase durations. The technique of multicycle closed-phase LPC (MCLPC) was introduced [9] to address cases in which condition 3) would not otherwise be satisfied. The success of MCLPC comes from the use of data samples from the closed-phase of a small number of consecutive larynx cycles, thereby obtaining good parameter estimates even when the duration, in samples, of the closed-phase is small with reference to the order of the LPC analysis. Typically, a minimum of 2 ms of data is required for covariance LPC when the order is chosen as 1/1000 of the sampling frequency in Hz [1].

Accurate segmentation of the closed-phase (condition 1) above) can be achieved using contemporaneous EGG recordings [4], [11] from which glottal closure instants (GCIs) and glottal opening instants (GOIs) can be derived. The effect of segmentation accuracy on the performance of closed-phase

Manuscript received September 24, 2004; revised January 27, 2006. This work was supported by the Engineering and Physical Sciences Research Council, U.K., under Grant GR/N01569. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rainer Martin..

P. A. Naylor, J. Gudnason, and M. Brookes are with the Electrical and Electronic Engineering Department, Imperial College London, London SW7 2AZ, U.K. (e-mail: p.naylor@imperial.ac.uk; jon.gudnason@imperial.ac.uk; mike.brookes@imperial.ac.uk).

A. Kounoudes was with Imperial College London, London SW7 2AZ, U.K. He is now with the Department of Computing and Information Systems, The Philips College, Nicosia, Cyprus (e-mail: a.kounoudes@signalgenerix.com).

Digital Object Identifier 10.1109/TASL.2006.876878

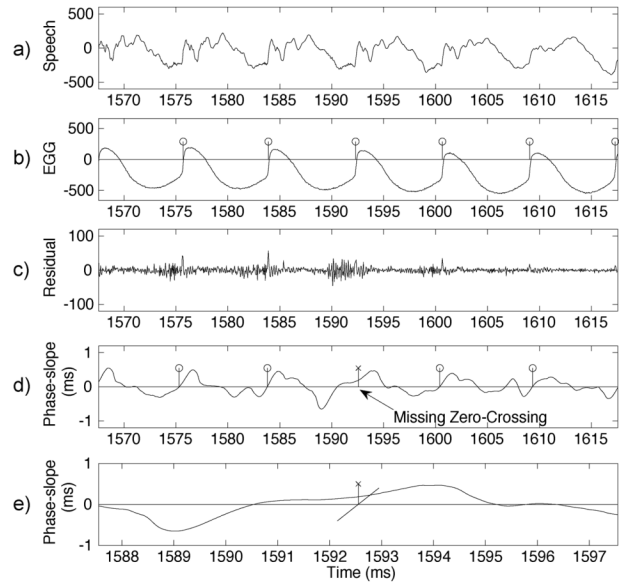


Fig. 1. Phase-slope projection. (a) Voiced speech signal. (b) EGG with reference GCIs extracted from the EGG using HQTx. (c) LPC prediction residual. (d) Phase-slope function with zero-crossings indicating GCIs (circles) and a missed GCI recovered using phase-slope projection (cross). (e) Detail showing the projection of a "missed" zero-crossing onto the horizontal axis.

LPC has been discussed in [12]. Since the EGG signal is not normally available in practical applications, there exists a strong motivation to develop techniques for determining GCIs and GOIs from the speech signal alone. Several such techniques have been presented in the literature and are considered in Section II. In this paper we present a new technique for estimating GCIs, known as the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA). We also briefly discuss approaches to estimate GOIs. An earlier version of the DYPSA algorithm was outlined in [13].

Section II of this paper briefly reviews some existing techniques for segmentation of speech signals into larynx cycles. In Section III, the DYPSA algorithm is presented and comparative results of tests on the new and existing algorithms are described in Section IV, from which conclusions are drawn in Section V.

II. Identification of Glottal Closure Instants

Although glottal closure and opening can be reliably observed in the EGG signal, there is no universally agreed definition of the precise instants [14]. The HQTx algorithm [15], [16] identifies GCIs from the EGG signal using the following definitions which we adopt in this paper. The starting points of glottal closure and opening are defined respectively as positive-going and negative-going zero crossings in the smoothed EGG time-derivative. The interval between the start of closure and the start of opening is defined as a glottal pulse if its duration and the amplitude of the EGG within the interval are within defined limits. A GCI is defined to occur at the maximum of the smoothed EGG time-derivative during a glottal pulse. An example segment of voiced speech, with the corresponding EGG signal and GCIs determined by HQTx, is shown in Fig. 1(a) and (b). Note that some authors invert the EGG signal from that shown here.

Several algorithms have been proposed for estimating glottal closure instants from a speech waveform $s(n)$ without the use of an EGG signal. The most widely used approach is to detect discontinuities in a linear model of speech production. An early algorithm [17] derived GCIs from the autocovariance matrix of the speech signal. This was developed further in [7] using the minimum energy in the LPC residual and additionally enhanced in [6], [18]. An alternative approach is to detect energy peaks in waveforms derived from the speech signal [19]–[21] or from features in its time-frequency representation [22], [23]. For example, the GCIs in [19] are identified as the maxima of the Frobenius norm of the signal matrix. Work on energy flow in the lossless-tube model [21] has suggested that the signal representing acoustic input power at the glottis could be used to determine the instants of glottal closure and opening. An approach based on the use of the group delay function was first proposed in [24] and later refined in [25] and [26]. In these methods, estimates of the time instants of excitation within an analysis frame are identified by zero-crossings of the frequency-averaged group delay over a sliding window applied to the LPC residual.

In this paper, we present the DYPSA algorithm for identifying GCIs in voiced speech from the speech signal alone. In Section IV, we evaluate DYPSA on databases that include contemporaneously recorded EGG signals. The GCIs determined by DYPSA are compared with reference GCIs obtained from the EGG signal using the HQTx algorithm [16]. We compare DYPSA's performance with three existing algorithms using the APLAWD database [27] and the SAM database [28]. The existing methods, selected to represent a cross-section of the various approaches, are Wong's LPC residual (LPCR) [7], the Frobenius Norm method (FN) [19], and the Group Delay method (GD) [24]. Details of the tests and the results obtained are discussed later in Section IV. In summary, the results show that, in our tests, the GD method is clearly the best performing of the existing methods studied. These preliminary findings provide the motivation for the use of the phase-slope function, as also employed in the GD method, as one of two indicators of excitation events in our new algorithm.

To determine the analysis window for closed-phase LPC, it is necessary to estimate GOIs as well as GCIs. It is commonly observed that the energy of excitation at GOIs is normally much weaker and more dispersed than at GCIs. Consequently, it is generally a more challenging problem to identify instants of glottal opening than closure [29]. However, the lower energy of excitation at GOIs also means that their timing estimation is less demanding of accuracy. Whereas a small timing error in a GCI might erroneously include a major excitation event in the closed-phase and consequently cause substantial errors in closed-phase LPC analysis, small timing errors in GOI estimation do not normally cause such significant effects. Increasing the closed-phase analysis period to include samples from the opening event causes a gradual corresponding increase in the spectral energy below the first format [7], [12]. Direct estimation of GOIs has not been addressed to any great extent in the current literature and remains an open research issue. It is suggested in [30], for example, that the phase-slope function may be capable of detecting glottal closures when a relatively short

averaging window is used and this issue is discussed further in Section V. Alternatively, closed-phase analysis can be performed over an interval defined as either a fixed fraction of the larynx cycle beginning at the GCI, or a fixed period following the GCI. Closed-phase intervals in the typical range of 30% to 45% have been reported for normal male speech [11], [31], [32] although the closed-phase duration can sometimes be shorter, or even totally absent in, for example, speech with breathy phonation.

A range of other studies of glottal action during the larynx cycle, and the extraction of related features, have been undertaken and are widely reported in the literature. For further information, the reader is referred to, for example, [33]–[35] and the references contained therein.

III. The DYPSA Algorithm

The DYPSA algorithm is an automatic technique for estimating GCIs in voiced speech from the speech signal alone. There are three components of the algorithm. The first component generates candidate GCIs using zero crossings of the phase-slope function [30]. The second component employs a novel phase-slope projection technique to recover candidates for which the phase-slope function does not include a zero crossing. These two components successfully detect almost all the true GCIs but also include a large number of false GCI candidates. The third component of the algorithm uses dynamic programming (DP) to identify the true GCIs from the set of candidates by minimizing a cost function that we define.

A. The Phase-Slope Function

The phase-slope function was defined in [24] to be the average slope of the unwrapped phase spectrum of the short-time Fourier transform of the linear prediction residual. More recently, alternative analytical definitions have been preferred [26]. Instants of glottal closure are identified in the phase-slope[1] function as positive-going zero-crossings.

Given the linear prediction residual signal $u(n)$ and applying a sliding M-sample Hamming window $w(m)$, we obtain a signal segment

$$x_n(m) = \begin{cases} w(m)u(m+n), & \text{for } m = 0, \ldots, M-1 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The discrete-time Fourier transform of $x_n(m)$ is

$$\tilde{X}_n(\omega) = \sum_{m=-\infty}^{\infty} x_n(m) e^{-j\omega m}. \quad (2)$$

The phase-slope $\tilde{\tau}_n(\omega)$ of $\tilde{X}_n(\omega)$ is defined as [25], [30]

$$\tilde{\tau}_n(\omega) = \frac{d \arg\left(\tilde{X}_n(\omega)\right)}{d\omega} \quad (3)$$

[1]For reasons of clarity, we choose here to use the name "phase-slope" so as to reserve the term "group delay" for the Group Delay method of [24]. Phase-slope and group delay differ only in sign.

and this can be sampled at M frequencies to give

$$\tau_n(k) = \tilde{\tau}_n(\omega)|_{\omega=2\pi k/M}$$
$$X_n(k) = \tilde{X}_n(\omega)|_{\omega=2\pi k/M}$$

for $k = 0, 1, \ldots, M-1$. It can be shown [30] that

$$\tau_n(k) = \Re\left(\frac{\breve{X}_n(k)}{X_n(k)}\right) \quad (4)$$

where $\breve{X}_n(k)$ is the discrete Fourier transform of $mx_n(m)$ and $\Re$ indicates the real part. For implementation, $x_n(m)$ is obtained from (1) with $u(n)$ computed as the LPC residual obtained using autocorrelation LPC on pre-emphasized speech with 20 ms overlapping analysis frames. The phase-slope function is found by forming an average of $\tau_n(k)$ across frequency and can be used to detect an impulse within the analysis window as follows. If $x_n(m)$ contains a noise-free impulse at $m = m_0$ then it follows directly from (3) that $\tau_n(k) \equiv m_0 \forall k$. However, when $x_n(m)$ additionally contains noise, $\tau_n(k)$ will not be constant and must be averaged over k in some manner. There are several alternative methods for forming such an average which are discussed in detail in [30]. For use in DYPSA, the energy-weighted formulation defined by

$$g(n) = \frac{\sum_{k=0}^{M-1}\left(|X_n(k)|^2 \tau_n(k)\right)}{\sum_{k=0}^{M-1}|X_n(k)|^2} \quad (5)$$

has been selected. It is shown in [30] that this formulation has relatively high performance and low computational cost since it can be rewritten as a "center of gravity" of signal energy

$$g(n) = \frac{\sum_{m=0}^{M-1} m x_n^2(m)}{\sum_{m=0}^{M-1} x_n^2(m)}. \quad (6)$$

The formulation of (5) and (6) also has the advantage that $g(n)$ is bounded, unlike several other proposed measures, and lies in the range $0 \leq g(n) \leq M-1$ provided that the denominator of (6) is nonzero. For subsequent processing in DYPSA, we employ

$$d(n) = g\left(n - \frac{M-1}{2}\right) - \frac{M-1}{2} \quad (7)$$

so that the analysis window is centered on sample n.

The nature of the phase-slope function has been studied recently in [30]. It has been shown that the choice of window size affects significantly the occurrence of zero-crossings in the phase-slope function. Ideally, the window should span exactly one impulsive event in the prediction residual so that a zero-crossing will occur to indicate a GCI candidate. When the window is larger than one larynx cycle, more than one closure event may be included and a zero-crossing in this case will typically occur mid-way between the two events, therefore giving inaccurate GCI detection. When the window is very much smaller that the larynx cycle, there will often be no impulsive event in the prediction residual within the analysis

window and spurious zero-crossings occur in the phase-slope function, giving rises to false alarms. Additionally, even when the window is one larynx cycle in duration, inaccurate detection will occur if the speech contains strong excitation at glottal opening, as occurs in a minority of talkers.

In our approach described in Section III-C, we employ DP to select GCIs from a set of candidates. Therefore, an increased number of GCI candidates is not normally problematic since spurious candidates will not be selected if, as is the intention, they are assigned a high cost within the DP. In contrast, missed candidates cause errors which are not recoverable in the DP. Our approach is therefore to minimize the number of missed candidates by using a moderately small window size of 3 ms and, in addition, the phase-slope projection technique described below to avoid the need to overly reduce the window-size. Finally, we note that the choice of window function is not critical since we are concerned with zero-crossings that occur near the center of the window where normally the window amplitude is approximately constant.

B. Phase-Slope Projection

In studying the phase-slope function [30], we observe that GCI events can go undetected because the phase-slope function fails to cross zero appropriately, even though the turning-points and general shape of the waveform are consistent with the presence of an impulsive event indicating a GCI. An example can be seen in Fig. 1, in which (a) shows a segment of speech, (b) shows the EGG signal with reference GCIs extracted from the EGG signal using HQTx, (c) shows the LPC residual signal, and (d) shows the phase-slope function with zero-crossings indicating GCIs (marked as circles). Fig. 1(e) shows the detail near 1593 ms and includes an example of a true GCI, marked by a "$\times$", for which the phase-slope function fails to cross the zero axis. A GCI candidate at this instant is indicated by successive turning points but would be undetected by methods relying only on zero-crossings. To recover such otherwise undetected GCI candidates, we introduce the phase-slope projection technique as illustrated in Fig. 1(e). In this method, whenever a local minimum is followed by a local maximum without an intervening zero-crossing, the midpoint between the two turning points is identified and its position projected with unit slope onto the time axis. This technique draws on the assumption that, in the absence of noise, the phase-slope at a zero-crossing is unity [24], [30]. The number of detection misses is significantly reduced by defining the set of GCI candidates to be the union of all positive going zero-crossings and projected zero-crossings as will be shown in tests described in Section IV.

Most often, one pulse in the prediction residual can be expected at the instant of glottal closure. However, for some talkers, LPC analysis can give a prediction residual containing additional strong pulses, possibly for example at the start of glottal opening such as at 1586 ms in Fig. 1(c), or an absence of any significantly strong pulses such as near 1593 ms.

C. Dynamic Programming

Given a set of candidate GCIs determined as described above, we now wish to choose from those candidates a subset corresponding to the true GCIs. The selection of GCIs from a set of candidates is performed by minimizing a cost function using N-best DP [36], [37]. Procedures employing N-best DP maintain information about the N most likely hypotheses at each step. The value of $N = 3$ has been chosen in this work as discussed in Section IV.

The factors used in the construction of the cost function are based on the attributes of the GD and FN methods and known characteristics of voiced speech including spectral quasi-stationarity and the periodic behavior of the vocal folds [38]. DYPSA employs DP to select a subset, Ω, of GCIs from the set of all GCI candidates generated as described above so as to minimize a cost function. We define the minimization problem as

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \boldsymbol{\lambda}^T \mathbf{c}_{\Omega}(r) \tag{8}$$

where

$$\boldsymbol{\lambda} = [\lambda_A, \lambda_P, \lambda_J, \lambda_F, \lambda_S]^T \tag{9}$$

is a vector of weighting factors, Ω is a subset of GCIs selected from all GCI candidates, $|\Omega|$ is the size of Ω, r indexes the elements of Ω, and $^\tau$ represents the transpose operation.

The elements of the cost vector evaluated for the rth GCI of subset Ω

$$\mathbf{c}_{\Omega}(r) = [c_A(r), c_P(r), c_J(r), c_F(r), c_S(r)]^T \tag{10}$$

all lie in the range [$-$0.5,0.5] and are defined below. We additionally define n_r, n_{r-1} and n_{r-2} to be the sample indices of GCI candidates r, $r-1$ and $r-2$ respectively where, for clarity of notation, we omit the explicit dependency of n_r on Ω.

1) Speech Waveform Similarity Cost: The speech waveform similarity cost uses the normalized cross-correlation estimator calculated from the speech signal as

$$c_A(r) = -\frac{1}{2}\frac{\gamma_{r-1,r}}{\sqrt{\gamma_{r-1,r-1}\gamma_{r,r}}} \tag{11}$$

where $\gamma_{r-1,r}$ is the covariance of 10 ms speech segments centered at samples n_{r-1} and n_r, and $\gamma_{r-1,r-1}$ and $\gamma_{r,r}$ are similarly computed autocovariances. During voicing, it is common that the speech waveform near an instant of excitation is well correlated to the waveform at the previous excitation. A high cost is therefore applied to any candidate that occurs where the speech signal is not well correlated with the previous candidate. This serves effectively to penalize candidates that occur, for example, part way through a larynx cycle. Additionally, c_A is insensitive to the stationary amplitude and phase distortion that can be introduced by speech input devices or during transmission since it is concerned only with relative variations between consecutive larynx cycles.

2) Pitch Deviation Cost: The pitch deviation cost is a function of the current and previous two GCI candidates under consideration by the DP and is defined as

$$c_P(r) = 0.5 - \exp\left(-\left(\psi(\Delta_P - 1)\right)^2\right) \tag{12}$$

where the pitch deviation is

$$\Delta_P = \frac{\min\left((n_r - n_{r-1}), (n_{r-1} - n_{r-2})\right)}{\max\left((n_r - n_{r-1}), (n_{r-1} - n_{r-2})\right)}. \tag{13}$$

The cost increases nonlinearly with Δ_P from -0.5 to $+0.5$, applying relatively small penalties for minor pitch changes based on an assumption of smooth variation in pitch over short segments of voiced speech. The rate of increase of cost with pitch deviation is controlled by ψ and zero cost is obtained at

$$\Delta_{P_0} = 1 + \frac{1}{\psi}\sqrt{-\ln\left(\frac{1}{2}\right)}. \tag{14}$$

In our experiments, $\psi = 3.3$ has been employed so as to obtain zero cost at pitch deviation of 25%. The DYPSA algorithm does not require a supplemental pitch estimator.

3) Projected Candidate Cost: The projected candidate cost penalizes a GCI candidate that arises from a projection of the phase-slope function onto the time-axis as described in Section III-B such that

$$c_J(r) = \begin{cases} 0.0, & \text{candidates from phase-slope zero crossings} \\ 0.5, & \text{candidates from phase-slope projection} \end{cases}. \tag{15}$$

This cost function term is included because, as well as recovering GCIs that are not detectable as zero-crossings, phase-slope projection can generate spurious GCIs due to noise in the LPC residual.

4) Normalized Energy Cost: The normalized energy cost is formulated as

$$c_F(r) = 0.5 - \frac{F(n_r)}{\breve{F}(n_r)} \tag{16}$$

where $F(n_r)$ is the energy of the speech signal $s(n)$ in the vicinity of GCI candidate r. This is computed using

$$F(n_r) = \sum_{k=-K}^{K} \min\left(H, K - |k|\right) s^2(n_r - k) \tag{17}$$

where, following [19], we take H and K to be 1 and 2 ms times the sampling frequency respectively. The term $F(n_r)$ differs only by a scale factor from the Frobenius norm measure used in [19] but is computed here more efficiently. The normalization term $\breve{F}(n_r)$ is an estimate of the local maximum of $F(n)$ in the vicinity of GCI candidate r calculated using a sliding window of size L

$$\breve{F}(n_r) = \max_k \left(F(n_r - k)\right), \quad 0 \le k < L. \tag{18}$$

The choice of L should be large enough to ensure that the window contains at least one excitation event in voiced speech and a duration corresponding to 16 ms has therefore been chosen.

The cost c_F is smallest when the GCI candidate occurs at a local maximum in the short-term signal energy. This measure is used to penalize candidates that do not correspond to high energy in the speech signal such as candidates that arise due to opening of the glottis or noise events.

5) Ideal Phase-Slope Function Deviation Cost: In the absence of noise, an impulsive event at the input of the group delay function that DYPSA employs for candidate generation gives rise to a zero-crossing with unit gradient at its output. Since the group delay function is applied to the LPC residual signal, the events are not normally true impulses and therefore the gradient at the zero-crossing will deviate from unity [18]. The ideal phase-slope function deviation cost is used to provide a measure of confidence in the LPC residual and the candidates obtained from it. Candidates arising from zero-crossings with gradients close to unity are favored. This cost is set to zero for candidates arising from phase-slope projections. We define

$$c_S(r) = 0.5 - \max\left(0, \min\left(\tilde{d}(n_r), 1/\tilde{d}(n_r)\right)\right) \tag{19}$$

where $\tilde{d}(n_r)$ is the mean value of the phase-slope calculated over a short window centered on candidate r such that

$$\tilde{d}(n_r) = \frac{1}{\nu}\left(d\left(n_r + \frac{\nu}{2}\right) - d\left(n_r - \frac{\nu}{2}\right)\right) \tag{20}$$

where ν is the even window length in samples. From our tests we have found 0.3 ms to be a satisfactory choice for the window duration and have observed that overall performance of DYPSA is insensitive to the choice of window duration over the range 0.3–1 ms.

IV. Experiments and Results

An initial evaluation of existing techniques LPCR [7], FN [19] and GD [24] was carried out in order to determine which of the variously proposed methods in the literature is most effective at generating GCI candidates. The window-size for GD was chosen as 7.5 ms so as to be in the range of approximately one to two times the average pitch period as specified in [24]. Subsequent experiments were performed to test the effectiveness and to quantify the overall performance of DYPSA in comparison to the existing techniques.

Two speech databases have been employed for evaluating the performance of DYPSA. The APLAWD database [27] contains ten repetitions of five phonetically balanced English sentences spoken by each of five male and five female talkers. The SAM

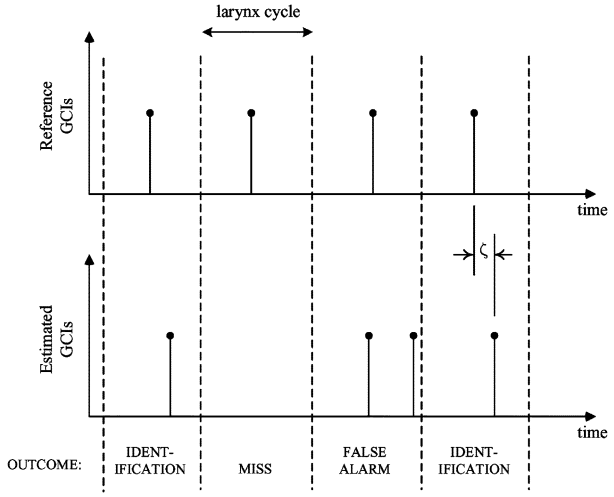


Fig. 2. Characterization of GCI Estimates showing 4 larynx cycles with examples of each possible outcome from GCI estimation. Identification accuracy is measured by ζ.

database [28] contains extended two minute passages read in English by four talkers.

A training subset of APLAWD containing 20 randomly selected files, approximately 4% of the database, was employed to determine λ in (9) as

$$\lambda = [\lambda_A, \lambda_P, \lambda_J, \lambda_F, \lambda_S]^T = [0.8, 0.5, 0.4, 0.3, 0.1]^T \quad (21)$$

using an optimization procedure which exhaustively searched each parameter over the range 0, 0.1, …, 1. The training data was subsequently excluded from all testing of DYPSA. The SAM database was used only for testing and was not used for determining λ nor for algorithm development.

Contemporaneous EGG recordings available in both databases were used to obtain reference GCIs for the purpose of evaluation. The speech and EGG signals were time-aligned to compensate for the larynx-to-microphone delay, determined for both databases to be approximately 0.95 ms. Reference GCIs were then extracted from the EGG signals corresponding to the voiced speech of APLAWD and SAM using the HQTx algorithm [16]. Performance comparisons have been made over these voiced speech segments between the reference GCIs and GCI estimates obtained from the methods studied.

To assess the performance of the algorithms we define with reference to Fig. 2: *larynx cycle*—the range of samples $(1/2)(\breve{n}_{r-1} + \breve{n}_r) \leq n < (1/2)(\breve{n}_r + \breve{n}_{r+1})$ given a reference GCI at sample $\breve{n}_r$ with preceding and following reference GCIs at samples $\breve{n}_{r-1}$ and $\breve{n}_{r+1}$, respectively; *identification rate*—the percentage of larynx cycles for which exactly one GCI is detected; *miss rate*—the percentage of larynx cycles for which no GCI is detected; *false alarm rate*—the percentage of larynx cycles for which more than one GCI is detected; *identification error*, ζ—the timing error between the reference GCIs and the detected GCIs in larynx cycles for which exactly one GCI has been detected; *identification accuracy*, σ—the standard deviation of ζ. Small values of σ indicate high accuracy of identification.

TABLE I
PERFORMANCE COMPARISON FOR GCI DETECTION METHODS ON THE APLAWD DATABASE. RESULTS FOR DYPSA WITHOUT PHASE-SLOPE PROJECTION ARE INDICATED BY "w/o PSP"

	Identification Rate (%)	Miss-Rate (%)	False Alarm Rate (%)	Identification Accuracy, σ (ms)
LPCR	40.2	53.1	6.7	1.38
FN	59.5	0.3	40.2	0.62
GD	81.7	2.3	16.0	0.52
DYPSA	95.7	1.6	2.7	0.71
DYPSA w/o PSP	94.0	3.6	2.4	0.74

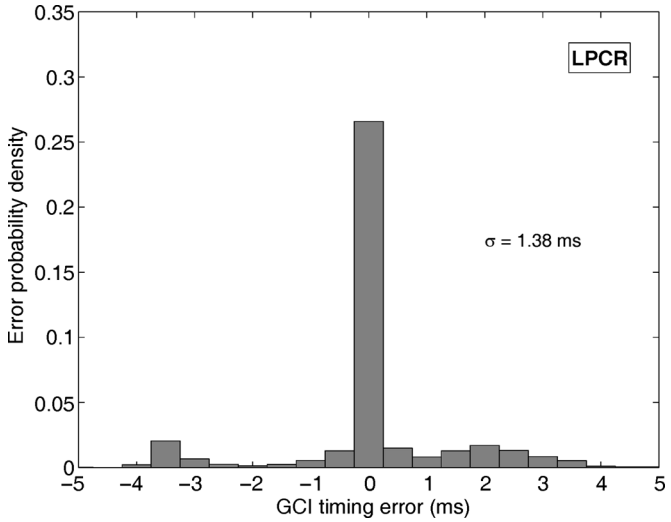


Fig. 3. GCI timing errors, ζ, for the LPCR algorithm on APLAWD.

Table I shows comparative results on the APLAWD database for identification rate, miss rate, false alarm rate, and identification accuracy σ for the three existing methods LPCR, FN and GD as well as for DYPSA. Figs. 3–5 show the distribution of timing errors in detection of GCIs, ζ, for the LPCR, FN and GD methods respectively, averaged over all five male and five female talkers in the APLAWD database. Results for the SAM database are shown in Table II.

It can be concluded from Tables I and II and Figs. 3–5 that the GD method performed best of the previously published methods in our tests. This motivated our choice of the phase-slope function, as used in the GD method, as the principal GCI candidate generator for use within DYPSA.

The tests described above were repeated using the DYPSA algorithm. Figs. 6 and 7 show the corresponding distribution of timing errors, ζ. The better performance of DYPSA over the GD method can be accounted for by considering the capability of the DP within DYPSA to reject GCI candidates generated from the phase-slope function for which the DP cost is high. This reduces false alarms in larynx cycles for which more than one candidate has been generated. Although DYPSA has no explicit knowledge of the time-range of each larynx cycle, and does not

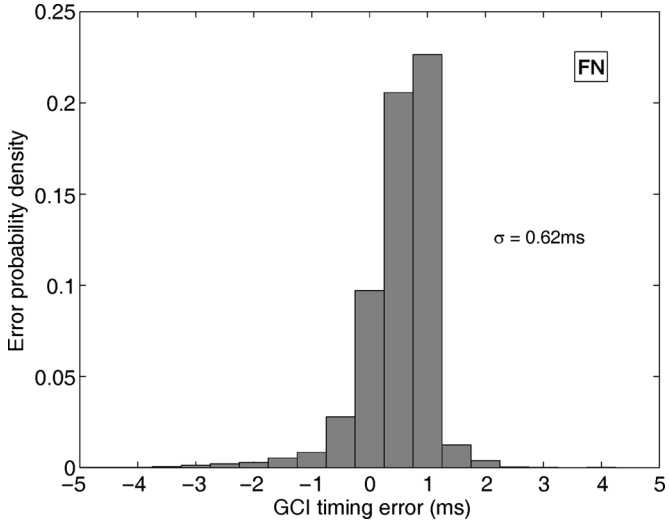


Fig. 4. GCI timing errors, ζ, for the FN algorithm on APLAWD.

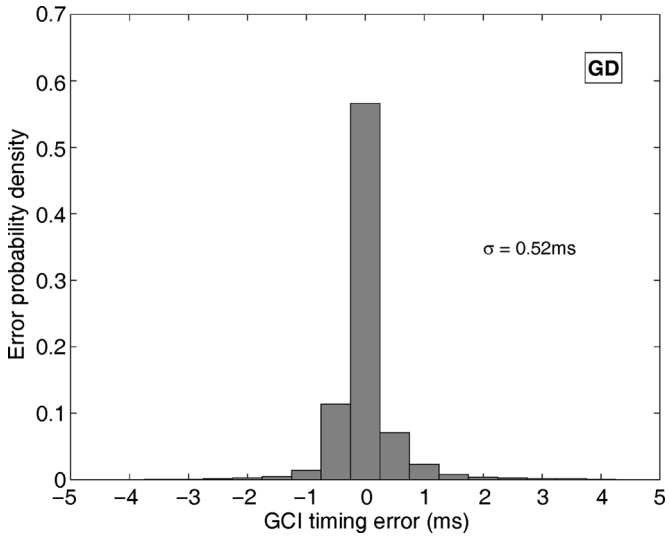


Fig. 5. GCI timing errors, ζ, for the GD algorithm on APLAWD.

TABLE II
PERFORMANCE COMPARISON FOR GCI DETECTION METHODS ON THE SAM DATABASE

	Identification Rate (%)	Miss-Rate (%)	False Alarm Rate (%)	Identification Accuracy, σ (ms)
LPCR	42.3	50.5	7.2	1.47
FN	58.7	1.4	40.0	0.59
GD	82.6	4.8	12.6	0.55
DYPSA	93.1	4.0	3.0	0.65

attempt to estimate it, the DP cost function can be seen effectively to penalize GCI candidates so as to reject all but one candidate per larynx cycle in most cases. The low value of σ indicates that the remaining GCI candidate in each larynx cycle is close in time to the reference GCI. A further factor towards the improved performance comes from the use of phase-slope

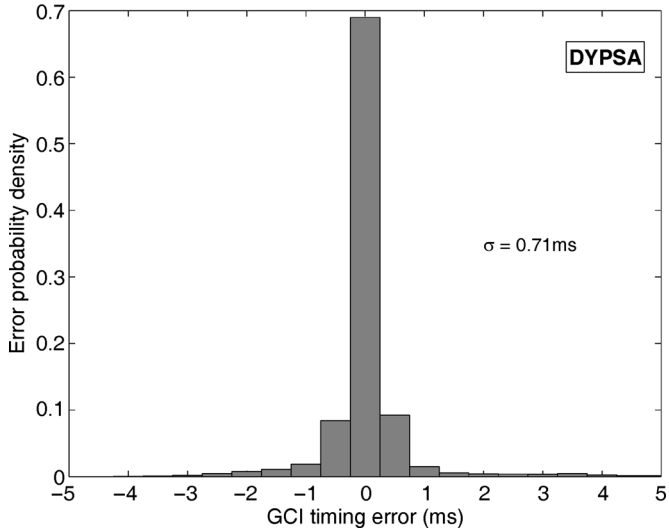


Fig. 6. GCI timing errors, ζ, for DYPSA on APLAWD.

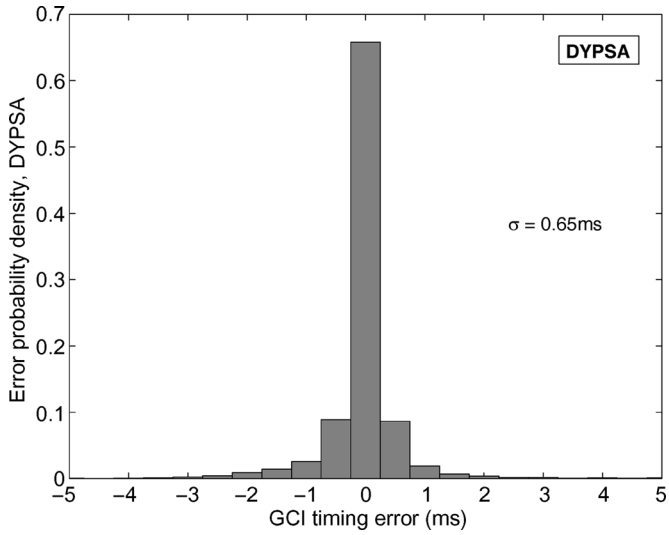


Fig. 7. GCI timing errors, ζ, for DYPSA on SAM.

projections that recovers GCI candidates that would otherwise be missed. The last row of Table I shows the performance of DYPSA without phase-slope projection and indicates that the phase-slope projection technique identifies, with good identification accuracy, GCIs that would otherwise be missed, resulting in a rise of identification rate from 94.0% to 95.7%.

In these experiments, a reasonable tradeoff between complexity and performance of DP has been found when $N = 3$. This choice is supported by Fig. 8, which results from an experiment in which N was varied and shows the frequency of selection finally made by the DP at each GCI. The result indicates that the choice $N = 3$ is adequate in 96.6% of cases, though some small improvements in overall performance could be obtained by increasing N at the cost of increased computation.

Fig. 9 shows an example of DYPSA's operation. For the utterance shown in Fig. 9(a) and the detail of the same data shown in Fig. 9(b), the lower and upper traces of ticks indicate respectively the reference GCIs obtained from the EGG using HQTx and GCIs obtained from DYPSA. This example has been chosen

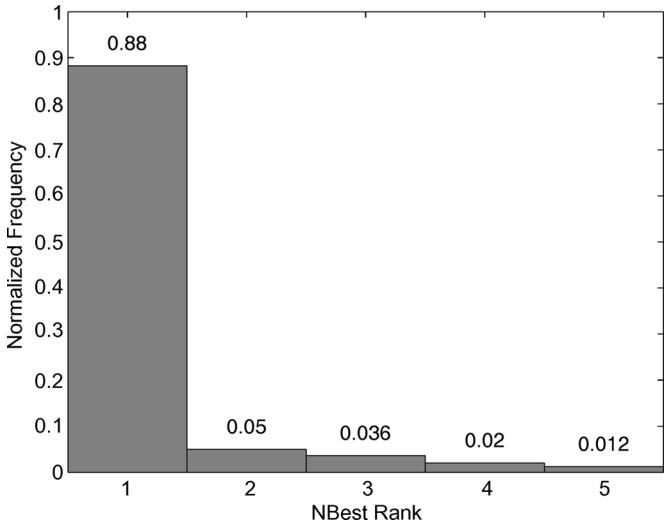


Fig. 8. Frequency of selection from the N-best paths at each GCI.

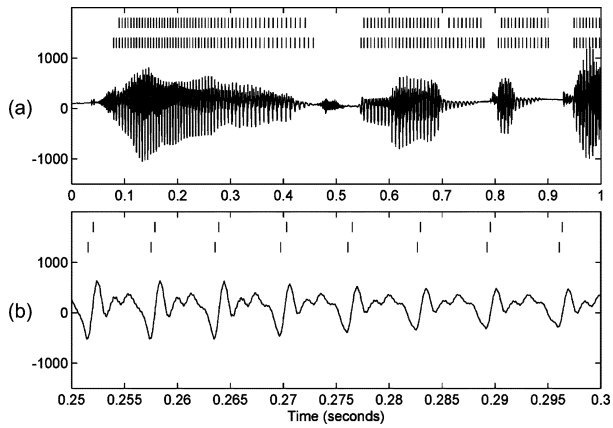


Fig. 9. GCI identification using DYPSA. (a) Speech signal. (b) Detail at 0.25 s. The lower row of ticks are reference GCIs determined from the EGG. The upper row of ticks are obtained from DYPSA. Unvoiced regions are excluded by DYPSA.

to illustrate two different types of missed GCIs. It can be seen that DYPSA's GCIs match well with the EGG-derived GCIs except near the onset and ending of voiced regions where DYPSA misses GCIs due to the use of consistency measures in the cost function. This is normally less problematic than misses well within a voiced region since the speech data at voiced-unvoiced boundaries is often less useful for speech analysis. DYPSA also misses GCIs occasionally within a voiced segment such as that illustrated in this example near 0.7 s. Fig. 9(b), showing a detail from the waveform, illustrates that the GCIs obtained from DYPSA are aligned with a consistent offset to the reference GCIs. Such an offset will arise from imperfect time-alignment between the speech and EGG channels in the test data and is therefore not included in our assessment of accuracy of any of the algorithms.

Fig. 10 presents an illustrative example of the components of the DYPSA cost function. A segment of voiced speech is shown in Fig. 10(a) in which the upper ticks represent the candidate GCIs and the vertical lines indicate the GCIs selected by the DP. Fig. 10(b) shows the time-variation of four components of the

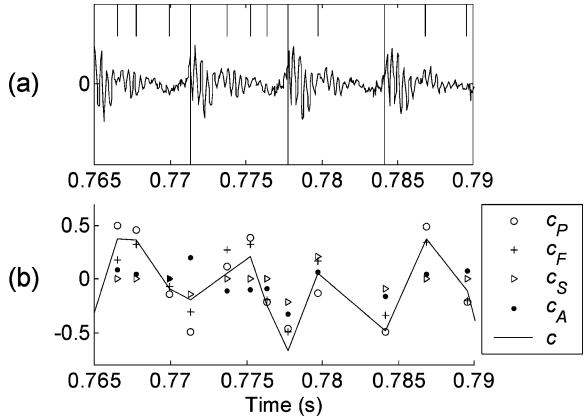


Fig. 10. Components of the DYPSA cost function. (a) Voiced speech segment with GCI candidates (upper ticks) and selected GCIs determined by DYPSA (vertical lines). (b) Components of the cost function and total cost c.

cost function and their weighted sum $\boldsymbol{\lambda}^T\mathbf{c}$ for each of the candidates. For a given candidate r, the cost function components $c_F(r)$ and $c_S(r)$ can be determined independently of any other GCI selections. However, the other cost components are dependent on the particular selection of GCIs from candidates made by the DP. Therefore, in this example, the cost of selecting a given candidate j to be the r^{th} GCI is found using DP as the optimal cost across all possible selections for which candidate j is selected to be GCI r. It can be seen that, as expected, the overall cost is higher for the rejected candidates than for the selected GCIs. The pitch deviation cost c_P can be seen to discriminate well in most cases and this is consistent with the high weighting of this cost component $\lambda_P = 0.5$. Near 0.78 s, however, its cost of zero indicates uncertainty and the successful rejection of the candidate is achieved by the other cost function components in the DP. The component with the highest weighting is the amplitude consistence cost, c_A, with $\lambda_A = 0.8$. It can be seen that during the second half of this example c_A discriminates the GCIs correctly but that during the first half it incorrectly penalizes a GCI. Nevertheless, the contributions of the other cost function components are sufficient to lead the DP to select the GCI correctly, as in the case discussed above.

V. Discussion and Conclusion

The DYPSA algorithm for estimating GCIs in voiced speech has been presented. It employs the phase-slope function in combination with a novel phase-slope projection method to determine GCI candidates, and a DP algorithm to select the most likely candidates according to a defined cost function. Its accuracy has been tested on the APLAWD and SAM speech databases in comparison to reference GCIs derived from simultaneously recorded EGG signals. It has also been compared to three other existing methods. The results show that DYPSA correctly detected 95.7% and 93.1% of GCIs in the two databases tested. Of the other three methods, FN has a consistently low miss rate but an extremely poor false-alarm rate. The GD and LPCR methods are significantly worse than DYPSA in both miss rate and false-alarm rate.

Candidate GCIs are obtained in DYPSA as zero-crossings of the phase-slope function. The choice of the analysis window

size M for calculation of the phase-slope function in (1) is important. If it is too long relative to the pitch period then it is likely to span more than one excitation event giving rise to missed GCI candidate zero-crossings as discussed in [24]. Additionally, speech from talkers with unusually high pitch such as can occur under stress, or talkers with strong excitation at opening as well as closure, increases the likelihood that the analysis window spans more than one excitation event for a chosen value M. Alternatively, if the analysis window is too short relative to the pitch period then many spurious GCI candidate zero-crossings will be generated. Noise can be expected to give rise to a similar effect, although detailed study of the effects of noise on DYPSA are outside the scope of the current study. The use of DP within DYPSA makes the algorithm robust to spurious candidates since they are penalized in the cost function. In contrast, a missed zero-crossing represents an error which the DP cannot recover. We have therefore incorporated two important features into DYPSA's candidate generation technique. Firstly, because of the introduction of DP, we have been able to employ a shorter window than proposed in [24]. Secondly, we have introduced the phase-slope projection technique. These techniques ensure the inclusion of valid GCI candidates that would otherwise be missed and result in improved robustness to the choice of analysis window size M and, importantly, its relation to the pitch.

An important use of this type of speech segmentation is in closed-phase LPC analysis and inverse filtering from which the voice source and vocal tract features can be estimated separately. Such features have potential advantages in speaker recognition, for example. Previous work has shown that GCI accuracy of around 0.25 ms is required for effective closed-phase LPC analysis [5]. DYPSA identified 64.7% of the reference GCIs in the APLAWD database to an accuracy of ± 0.25 ms. The GD, FN and LPCR methods identified respectively 53.4%, 21.8%, and 24.6% to the same accuracy.

The significantly enhanced performance offered by DYPSA in identifying GCIs in voiced speech offers opportunities for the use of closed-phase LPC analysis, inverse filtering and source-tract feature estimation in diverse applications of speech processing. A MATLAB implementation of DYPSA is available in [39].

ACKNOWLEDGMENT

The authors would like to thank the associate editor and anonymous reviewers whose suggestions have helped us to improve this paper significantly.

REFERENCES

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

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

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

[4] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 8, pp. 730–743, Aug. 1986.

[5] A. Neocleous and P. A. Naylor, "Voice source parameters for speaker verification," in *Proc. Eur. Signal Process. Conf.*, 1998, pp. 697–700.

[6] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 569–586, Sep. 1999.

[7] D. Y. Wong, J. D. Markel, and J. A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 350–355, Aug. 1979.

[8] E. L. Riegelsberger and A. K. Krishnamurthy, "Glottal source estimation: Methods of applying the LF-model to inverse filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1993, vol. 2, pp. 542–545.

[9] D. M. Brookes and D. S. Chan, "Speaker characteristics from a glottal airflow model using glottal inverse filtering," in *Proc. Institue of Acoust.*, 1994, vol. 15, pp. 501–508.

[10] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Signal Process.*, vol. 33, pp. 369–377, Apr. 1985.

[11] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic assessment of normal voice: a tutorial," *Clinical Linguistics and Phonetics*, vol. 3, pp. 281–296, 1989.

[12] J. Larar, Y. Alsaka, and D. Childers, "Variability in closed phase analysis of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1985, pp. 1089–1092.

[13] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, 2002, vol. 11, pp. 349–352.

[14] R. Scherer, V. Vail, and B. Rockwell, "Examination of the laryngeal adduction measure EGGW," in *Producing Speech: Contemporary Issues: for Katherine Safford Harris*, F. Bell-Berti and L. Raphael, Eds. Woodbury, NY: AIP, 1995.

[15] M. A. Huckvale, D. M. Brookes, L. Dworkin, M. E. Johnson, D. J. Pearce, and L. Whitaker, "The SPAR speech filing system," in *Proc. Eur. Conf. Speech Technology*, Edinburgh, U.K., Sep. 1987, vol. 1, pp. 305–308.

[16] M. Huckvale, Speech Filing System: Tools for Speech Research University College London, 2000 [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/

[17] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1625–1629, 1974.

[18] J. G. McKenna, "Automatic glottal closed-phase location and analysis by Kalman filtering," in *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Blair Atholl, Aug. 2001.

[19] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 258–265, Apr. 1994.

[20] C. R. Jankowski, Jr, T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1995, pp. 325–328.

[21] D. M. Brookes and H. P. Loke, "Modelling energy flow in the vocal tract with applications to glottal closure and opening detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1999, pp. 213–216.

[22] V. N. Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Proc. Eur. Conf. Speech Technology*, Budapest, Hungary, Sep. 1999, pp. 2805–2808.

[23] J. L. Navarro-Mesa, E. Lleida-Solano, and A. Moreno-Bilbao, "A new method for epoch detection based on the Cohen's class of time frequency representations," *IEEE Signal Process. Lett.*, vol. 8, pp. 225–227, Aug. 2001.

[24] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 9, pp. 325–333, Sep. 1995.

[25] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, 1995, pp. 776–779.

[26] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

[27] G. Lindsey, A. Breen, and S. Nevard, SPAR's Archivable Actual-Word Databases Univ. College London, London, U.K., Jun. 1987, Tech. Rep..

[28] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM—a spoken language resource for the EU," in *Proc. Eur. Conf. Speech Communication and Speech Technology*, Sep. 1995, pp. 867–870.

[29] T. Backstrom, P. Alku, and E. Vilkman, "Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 186–192, Mar. 2002.

[30] D. M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifiying glottal closures in voiced speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 456–466, Mar. 2006.
[31] J. J. Jiang, S. Tang, M. Dalal, C.-H. Wu, and D. G. Hanson, "Integrated analyzer and classifier of glottographic signals," *IEEE Trans. Rehab. Eng.*, vol. 6, pp. 227–234, Jun. 1998.
[32] S. V. Batty, P. E. Garner, D. M. Howard, P. Turner, and A. D. White, "The development of a portable real-time display of voice source characteristics," in *Proc. 26th Euromicro Conf.*, 2000, pp. 419–422.
[33] Y. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1805–1815, Dec. 1989.
[34] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 8, pp. 309–319, Aug. 1979.
[35] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, Jul. 1998.
[36] R. Schwartz and Y.-L. Chow, "The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 81–84.
[37] J.-K. Chen and F. K. Soong, "An N-best candidates-based discriminative training for speech recognition applications," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 206–216, Jan. 1994.
[38] D. Talkin, "A robust algorithm for pitch tracking," in *Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New Providence, NJ: Elsevier, 1995.
[39] M. Brookes, VOICEBOX: A speech processing toolbox for MATLAB. 2006. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Patrick A. Naylor (M'89) received the B.Eng. degree in electronics and electrical engineering from the University of Sheffield, Sheffield, U.K., in 1986 and the Ph.D. degree from Imperial College, London, U.K., in 1990.

Since 1989, he has been a Member of Academic Staff in the Communications and Signal Processing Research Group, Imperial College London, where he is also Director of Postgraduate Studies. His research interests are in the areas of speech and audio signal processing and he has worked in particular on adaptive signal processing for acoustic echo control, speaker identification, multichannel speech enhancement, and speech production modeling. In addition to his academic research, he enjoys several fruitful links with industry in the U.K., U.S., and in mainland Europe.

Dr. Naylor is an Associate Editor of IEEE SIGNAL PROCESSING LETTERS and a member of the IEEE Signal Processing Society Technical Committee on Audio and Electroacoustics.

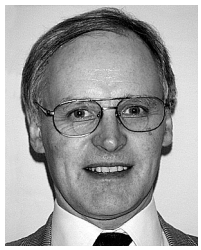
Anastasis Kounoudes (M'99) was born in Limassol, Cyprus, on March 31, 1973. He received the M.Eng. degree in computer engineering and informatics from the University of Patras, Patras, Greece, in 1997 and the Ph.D. degree in speech signal processing from Imperial College, University of London, London, U.K., in 2001.

During the period 2000–2002, he was a Postdoctoral Research Associate at the Speech Processing Group of Imperial College. From March 2002 until March 2003, he was with Domain Dynamics Ltd., U.K., as a Senior Speech Engineer. Since April 2003, he has been an Assistant Professor in Computer Engineering at The Philips College, Nicosia, Cyprus. In January 2004, he co-founded SignalGeneriX Ltd., a start-up company specializing in digital signal processing, where he acts as a Chief Technical Officer. His current research includes statistical pattern recognition, speech recognition, speaker verification, speaker adaptation, and biometrics.

Jon Gudnason (M'96) received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Iceland in 1999 and 2000, respectively. He is currently pursuing the Ph.D. degree with the Communication and Signal Processing Group, Imperial College, London, U.K.

From 1996 to 1998, he was an Intern with the Hydrology Service at the National Energy Authority in Iceland. In 1999, he was a Research Assistant for the Information and Signal Processing Laboratory, University of Iceland, working on remote sensing applications. He is currently a Research Associate with the Communication and Signal Processing Group, Imperial College, where his research has been on speaker recognition and automatic target recognition using radar.

Mr. Gudnason has been a member of the IEEE Signal Processing Society since 1996. He was the president of the IEEE Iceland Student Branch in 1998.



Mike Brookes (M'88) received the B.A. degree in mathematics from Cambridge University, Cambridge, U.K., in 1972.

Following this, he spent four years at the Massachussets Institute of Technology, Cambridge, working on astronomical instrumentation and telescope control systems. Since 1979, he has been with the Electrical and Electronic Engineering Department, Imperial College, London, U.K., where he is now a Deputy Head of Department and Head of the Communications and Signal Processing Research Group. His main area of research is speech processing, where he has worked on speech production modelling, speaker recognition algorithms, and techniques for speech enhancement using both single microphones and microphone arrays. He is currently applying techniques from speech processing to RADAR target identification and is also actively involved in computer vision research.