# Estimation of Glottal Closing and Opening Instants in Voiced Speech Using the YAGA Algorithm

Mark R. P. Thomas, *Member, IEEE*, Jon Gudnason, *Member, IEEE*, and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—Accurate estimation of glottal closing instants (GCIs) and opening instants (GOIs) is important for speech processing applications that benefit from glottal-synchronous processing including pitch tracking, prosodic speech modification, speech dereverberation, synthesis and study of pathological voice. We propose the Yet Another GCI/GOI Algorithm (YAGA) to detect GCIs from speech signals by employing multiscale analysis, the group delay function, and $N$-best dynamic programming. A novel GOI detector based upon the consistency of the candidates' closed quotients relative to the estimated GCIs is also presented. Particular attention is paid to the precise definition of the glottal closed phase, which we define as the analysis interval that produces minimum deviation from an all-pole model of the speech signal with closed-phase linear prediction (LP). A reference algorithm analyzing both electroglottograph (EGG) and speech signals is described for evaluation of the proposed speech-based algorithm. In addition to the development of a GCI/GOI detector, an important outcome of this work is in demonstrating that GOIs derived from the EGG signal are not necessarily well-suited to closed-phase LP analysis. Evaluation of YAGA against the APLAWD and SAM databases show that GCI identification rates of up to 99.3% can be achieved with an accuracy of 0.3 ms and GOI detection can be achieved equally reliably with an accuracy of 0.5 ms.**

***Index Terms*—Dynamic programming, electroglottograph (EGG), glottal closing instants (GCIs), glottal opening instants (GOIs), group delay function, multiscale analysis, speech processing.**

## I. Introduction

VOICED speech is produced when the vocal tract is excited by the vocal folds, which consists of opposing ligaments that form a constriction as it joins the lower vocal tract. When air is expelled from the lungs at sufficient velocity through this orifice—usually referred to as the glottis—the vocal folds experience a separating force. The instant of time at which the glottal folds begin to separate is termed the glottal opening instant (GOI). The vocal folds continue to open until equilibrium is reached between the separating force and the tension in the vocal folds, at which point the potential energy stored in the vocal folds causes them to begin to close. When the vocal folds become sufficiently close, the Bernoulli force results in an abrupt closure at the glottal closure instant (GCI). Elastic restoring forces during closure cause the cycle to repeat, producing a series of periodic pulses. The glottal cycle is defined as the period between successive GCIs.

The detection of GCIs in voiced speech is important for glottal-synchronous speech processing algorithms such as pitch tracking, prosodic speech modification [1], speech dereverberation [2], data-driven voice source modeling [3] and areas of speech synthesis [4]. Identification of GOIs is necessary for closed-phase linear predictive coding (LPC) [5] and the analysis of pathological speech that relies upon knowledge of the open quotient (OQ) [6]. Whereas many methods existing in the literature aim to estimate GCIs from the voiced speech signal, very few exist for the more challenging task of GOI detection. The broad applications of glottal-synchronous processing have given rise to a corresponding demand for increasingly reliable and automatic identification of GCIs and GOIs. There exists, however, no universally agreed definition of the GOI [7]. In this work, we aim to find an analysis interval that is best-suited to closed-phase LPC analysis [5] that is shown not to always correspond to the closed phase estimated from the EGG signal. An automatic reference is proposed that builds upon earlier works in [5] and [8] by iteratively refining electroglottograph (EGG)-based estimates based upon the variance of the estimated voice source signal in the closed phase.

Most existing techniques assume that the speech is stationary throughout an analysis window of 20–30 ms. During this time, a widely used approach is the detection of discontinuities in an estimation of the voice source signal with LPC that correspond closely to the GCIs and GOIs. An early example of practical applications of LPC in GCI/GOI detection can be found in [5] and has been applied to many more recent algorithms, notably [9]–[12]. Additional model-based approaches that estimate the voice source include homomorphic processing [13], in which the excitation signal is estimated as the signal components that contribute to fast changes in the speech spectrum. Model-based processing is advantageous because it exploits knowledge of the voice to provide a signal that is more straightforward to analyze than the speech signal alone, providing the model is sufficiently well-suited to the speech signal under test. The identification GCIs/GOIs by discontinuities or changes in signal energy include the Hilbert Envelope [14] and Frobenius Norm [15].

The wavelet transform can be viewed as an analysis filterbank that decomposes a signal into multiple wavelet scales. This has been used in the field of $f_0$ detection in speech signals [16], but much attention has been paid to the observation that discontinuities in a signal, such as those caused by GCIs and GOIs, are manifest as local maxima across multiple scales. The Lines

Manuscript received February 04, 2010; revised September 14, 2010 and March 06, 2011; accepted May 02, 2011. Date of publication June 07, 2011; date of current version November 04, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

M. R. P. Thomas and P. A. Naylor are with the Electrical and Electronic Engineering Department, Imperial College London SW7 2AZ, London, U.K. (e-mail: mrt102@imperial.ac.uk; p.naylor@imperial.ac.uk).

J. Gudnason is with the School of Science and Engineering, Reykjavik University, IS 101 Reykjavik, Iceland (e-mail: jg@ru.is).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2157684

of Maximum Amplitudes (LOMA) algorithm identifies local maxima that align across multiple wavelet scales [17]. The multiscale product [18] of the decomposed signal has been shown to be particularly effective for GCI/GOI detection in EGG signals [19], [20] and speech signals [21], [22]. The multiscale product is a key element in the technique proposed in this paper. Detection of periodicity in the speech has also been explored through analysis of the autocovariance matrix of the speech signal [23], zero-frequency resonator [24] and empirical mode decomposition (EMD) [25]. These non model-based approaches are advantageous because they are well-rooted in signal processing and are not constrained by any particular speech model.

Many algorithms emphasize GCIs and GOIs by transforming them into either an impulsive event (e.g., LPC residual), a local maxima or minima of a smoothly varying waveform (e.g., LOMA), or a zero crossing (e.g., zero-frequency resonator). The latter two are relatively straightforward to detect but impulsive events can often be masked by noise and neighboring events that can render them difficult to detect. A technique for the detection of impulsive events is a fixed threshold based upon a long-term measure of speech amplitude, sometimes used for GCI/GOI detection in EGG signals [26] but with limited application to speech signals due to the large dynamic range of natural conversational speech. Dynamic thresholds based on short-term averages [11] yield better results but can sit on a knife-edge between missing events or detecting false events if the threshold is too high or too low, respectively [20]. The method based upon group delay functions [27] uses a weighted average group delay calculated on a sliding window. The negative-going zero crossings of this function have been shown to reliably detect impulsive events in the LP residual [28]. Different approaches are reviewed in [27]. Phase slope projection [12] further improves estimates by detecting missed zero crossings and inserting them at the most likely time instant. In some cases the heuristics of the speech signal are used to improve quality of the estimates or suppress erroneous detections during unvoiced speech. Techniques such as $N$-best dynamic programming [29] have therefore been applied to minimize a cost function derived from features such as pitch consistency, waveform similarity, energy, multichannel correlation or goodness of fit to voice source models. Most existing approaches work well on sustained voiced phonemes but can fail on more challenging conversational speech if the heuristics of the signal are not considered [12].

In this paper, we present Yet Another GCI/GOI Algorithm (YAGA) that reliably estimates both GCIs and GOIs from speech signals. The algorithm is a combination of existing techniques including multiscale analysis, group delay functions and $N$-best dynamic programming [29]. A new technique for the detection of GOIs using the consistency of candidates' closed quotient relative to the estimated GCIs is proposed. YAGA, DYPSA [12], and the EGG-based SIGMA algorithm [20] are evaluated against the two-channel reference algorithm proposed in this paper.

The remainder of this paper is organized as follows. Section II describes the voice source signal in the context of GCI/GOI detection. A two-channel reference algorithm is described in Section III. Section IV describes the YAGA algorithm. Evaluation results of the GCI and GOI detection against the reference algorithm is presented in Section V and conclusions are drawn in Section VI.

## II. Estimation of the Voice Source Signal

We denote the GCIs $\mathbf{n}^{\mathrm{c}} = [n_1^{\mathrm{c}}\, n_2^{\mathrm{c}}\, \ldots\, n_R^{\mathrm{c}}]_{R\times 1}^T$ and GOIs $\mathbf{n}^{\mathrm{o}} = [n_1^{\mathrm{o}}\, n_2^{\mathrm{o}}\, \ldots\, n_R^{\mathrm{o}}]_{R\times 1}^T$, where $n_r^{\mathrm{c}}$ is the $r$th GCI, $n_r^{\mathrm{o}}$ is the $r$th GOI and $R$ is the total number of GCIs in a speech utterance. Glottal closed and open phases are defined by pairs of instants $\mathbf{c} = [\mathbf{c}_1\, \mathbf{c}_2\, \ldots\, \mathbf{c}_R]_{R\times 2}^T$ and $\mathbf{o} = [\mathbf{o}_1\, \mathbf{o}_2\, \ldots\, \mathbf{o}_R]_{R\times 2}^T$, respectively, where $\mathbf{c}_r = [n_r^{\mathrm{c}} + 1\; n_r^{\mathrm{o}} - 1]^T$ and $\mathbf{o}_r = [n_r^{\mathrm{o}}\; n_{r+1}^{\mathrm{c}} - 1]^T$.

### A. The Source-Filter Model

GCIs, and especially GOIs, are difficult to locate in the speech signal [12] due to the spectral shaping by the vocal tract transfer function $V(z)$. It is common to blindly estimate and equalize $V(z)$ from the observed speech signal, so as to estimate the voice source signal from which GCIs and GOIs are more straightforward to detect [12]. Let $s(n)$ be a frame of voiced speech with $z$-transform $S(z)$ such that

$$S(z) = U(z)V(z)R(z) = U'(z)V(z) \tag{1}$$

where $U(z)$ represents glottal volume velocity, $V(z)$ is an all-pole vocal tract filter, and $R(z) \simeq 1 - z^{-1}$ models lip radiation. The term $U(z)$ and the differential effect of $R(z)$ are usually combined into the glottal flow derivative $U'(z)$, often termed *voice source signal* with time-domain waveform $u'(n)$. If $V(z)$ is known, $U'(z)$ can be estimated from $S(z)$:

$$\hat{U}'(z) = S(z)/V(z) \tag{2}$$

with time-domain waveform $\hat{u}'(n)$. A whitened voice source signal (or LP residual) can be found by $E(z) = \tilde{S}(z)/V(z)$ with time-domain waveform $e(n)$, where $\tilde{S}(z) = S(z)\Phi(z)$ is preemphasized speech as discussed in the following section.

### B. Estimation by Linear Prediction

Various short-term LPC techniques have been developed that estimate $V(z)$ from the speech signal [10], [15]. Estimation of $U'(z)$ using (2) is then straightforward. Other techniques jointly estimate $V(z)$ and $U'(z)$ [30] that are not considered here. Re-writing (1) in the time domain

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + \hat{u}'(n) \tag{3}$$

where $a_i$ are the prediction coefficients, $\hat{u}'(n)$ is an estimate of $u'(n)$, and $p$ is the prediction order. The vocal tract transfer function can be approximated as

$$\hat{V}(z) = 1\Big/\left(1 + \sum_{i=1}^{p} a_i z^{-i}\right). \tag{4}$$

The prediction order $p$ for an adult male of vocal tract length 17 cm is approximately $f_{\mathrm{s}}/1000$, where $f_{\mathrm{s}}$ is the sampling frequency. The aim is to find the $a_i$ that minimize a cost function formed from (3):

$$J = E\left\{(\hat{u}'(n))^2\right\} = E\left\{\left(s(n) - \sum_{i=1}^{p} a_i s(n-i)\right)^2\right\} \tag{5}$$

where $E\{\cdot\}$ denotes expectation. Minimizing on each analysis frame by setting the derivative of $J$ to zero with respect to the LPC coefficients results in

$$\sum_{j=1}^{p} r_{i,j}a_j = r_{i,0} \text{ where } r_{i,j} = E\{s(n-i)s(n-j)\} \quad (6)$$

which can be represented in matrix form as

$$\mathbf{Ra} = \mathbf{c} \Rightarrow \mathbf{a} = \mathbf{R}^{-1}\mathbf{c}. \quad (7)$$

We consider here two methods for estimating $r_{i,j}$: pitch-asynchronous autocorrelation LPC and closed-phase covariance LPC.

### C. Pitch-Asynchronous Autocorrelation LPC

Pitch-asynchronous autocorrelation LPC calculates $r_{i,j}$ without knowledge of the temporal structure of the speech:

$$r_{i,j} = \sum_{n=-\infty}^{+\infty} \tilde{s}(n-i)\tilde{s}(n-j) \quad (8)$$

where $\tilde{s} = w(n)s(n)$ and $w(n)$ is a windowing function of typically 20–30 ms. The infinite sum leads to a Toeplitz matrix $\mathbf{R}$ that can be inverted with the Levinson–Durbin algorithm whose computational complexity scales $O(p^2)$. The fixed window includes the samples outside the glottal closed phase, which tilts the spectrum of the speech signal [31]. This has the effect of both introducing a spectral tilt into the estimated vocal tract filter $\hat{V}(z)$ and to spoil the conditioning of the matrix $\mathbf{R}$. With reference to the two-pole model of $U(z)$ [10], one pole is cancelled by the lip radiation filter $R(z)$. A common approach is to cancel the remaining pole with a first-order preemphasis filter of the form

$$\Phi(z) = 1 - \mu z^{-1} \quad (9)$$

with $\mu \simeq 1$. Using the estimate of the vocal tract filter, the voice source signal $\hat{U}'(z) = S(z)/\hat{V}(z)$ or linear prediction residual $E(z) = \tilde{S}(z)/\hat{V}(z)$ can be estimated. The linear prediction residual, though not having any physical significance, is often used in the detection of GCIs [12] and coding [32]. It is of limited use in studying glottal waveforms due to the level of high-frequency noise resulting from the preemphasis that masks some finer detail in the open phase; greater interest has therefore been shown in modeling $u'(n)$ [9], [33], [34].

The validity of the two-pole model of $U(z)$ can be questioned when phase characteristics are considered. Alternative approaches have therefore been devised to estimate and remove the spectral contribution of the voice source. The Iterative Adaptive Inverse Filtering (IAIF) method [35] imposes an additional model on $V(z)$, assuming an all-pass nature with spectral peaks caused by the formants. An iterative process first estimates a first-order AR model of the speech signal to form an initial estimate of the glottal pulse; this is removed from the speech signal by inverse-filtering. Subsequent stages estimate the glottal pulse and vocal tract filter at increasing orders. By adapting to the voice source in this way, IAIF is capable of producing superior estimation of the voice source than can be achieved with a fixed first-order model.

### D. Closed-Phase Covariance LPC

Pitch-synchronous autocorrelation LPC is a practical approach if knowledge of closed phase is unavailable. If, however, the closed phase is known, closed-phase covariance LPC can be beneficial by restricting its analysis window to the region in which the glottis is closed, i.e., $u'(n) = 0$. This circumvents the need for preemphasis and provides more accurate estimate of $V(z)$ and therefore $u'(n)$ [5], [8], [10]. Consider the covariance of a finite segment of speech

$$r_{i,j} = \sum_{n=n_r^c}^{n_r^o-1} s(n-i)s(n-j) \quad (10)$$

in which no windowing function is applied to the speech signal. The spectral resolution is therefore limited only by the number of samples in the analysis interval, and allows analysis intervals of as low as 2 ms. The resulting AR coefficients are however not guaranteed stable [10]. In some voices, particularly female, the closed phase may be less than 2 ms, rendering this approach ineffective. The problem can be addressed by multi-cycle closed phase analysis [36] that includes adjacent glottal closed phases in the calculation of the covariance matrix $\mathbf{R}$. The covariance equation in (10) can be rewritten as

$$r_{i,j} = \sum_{n=n_r^c}^{n_r^o-1} s(n-i)s(n-j) + \sum_{n=n_{r+1}^c}^{n_{r+1}^o-1} s(n-i)s(n-j) + \ldots \quad (11)$$

where the sum is often limited to include 2–3 adjacent cycles.

### E. Defining the Glottal Closed Phase

Glottal closing and opening are not truly instantaneous but phases of finite duration [37], although in general the closing phase is sufficiently short for it to be considered instantaneous. However, there is no universally agreed definition of the precise instants of GOIs [7].

There are three main definitions of the GOI in common use. Fig. 1 shows (a) an estimated voice source signal with pitch-asynchronous autocorrelation LPC, (b) the multiscale product [18] of (a), (c) the corresponding time-aligned EGG signal, and (d) the multiscale product of (c). The multiscale product is an estimate of the derivative of a signal over multiple dyadic scales and is discussed in detail in Section IV-A. The first GOI definition, defined in [5], corresponds to the instant at end of the closed phase when increased residual error is observed in the linear model of the speech signal, indicating nonstationarity caused by excitation of the vocal tract by glottal airflow. This is shown by the $(\circ)$ line in Fig. 1 and is used to define analysis intervals for closed-phase covariance LPC but may not necessarily correspond to the definition of opening in the physiological sense. Fig. 1 shows a discontinuity at this instant in plots (a) and (b) but there is little evidence in the EGG signal of plots (c) and (d). The second definition of the GOI, defined in [8] and [37], is the maximum derivative of the EGG signal as marked with the $(*)$ line in Fig. 1. This definition is used extensively to assess open quotients in pathological speech, although it corresponds solely to the maximum rate of change of glottal conductivity and not airflow. This can be seen as a discontinuity in both the estimated voice source (a), (b) and EGG signal (c), (d). The third type of GOI is the point

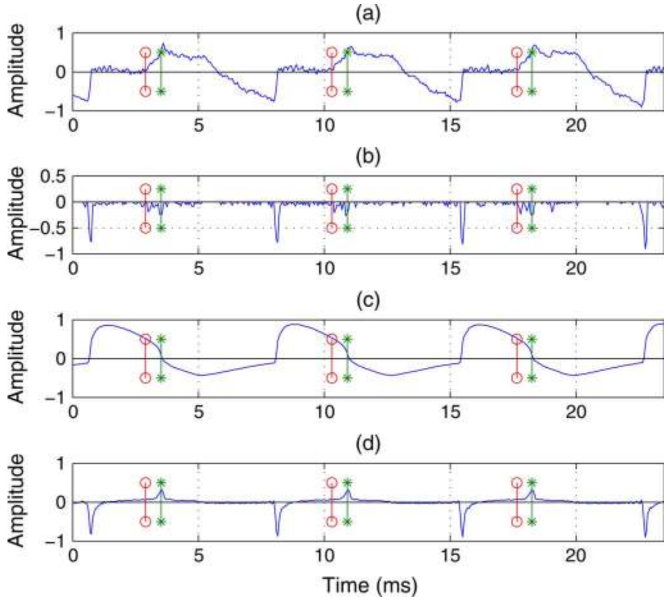

Fig. 1. Two definitions of GOI overlaid on (a) estimated voice source, (b) multiscale product of (a), (c) EGG and (d) multiscale product of (c). In the first case (red ○), the GOI marks the beginning of the opening phase, in the second (green ∗), the GOI marks the end of the opening phase.

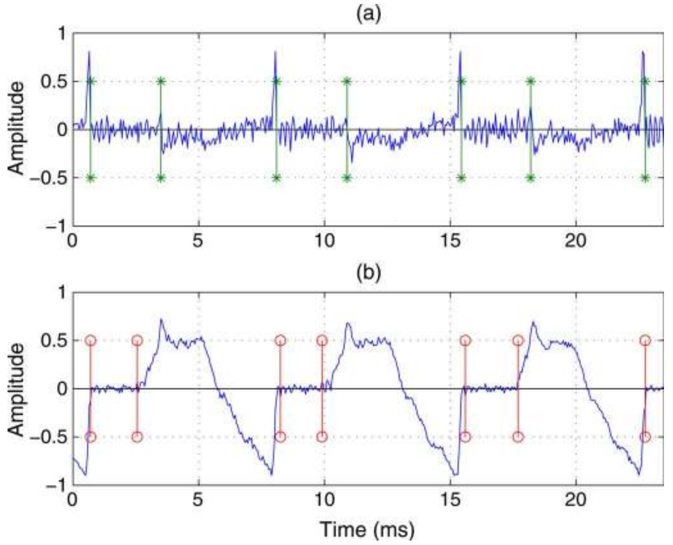

Fig. 2. Voice source estimated with closed-phase LPC. Analysis intervals from (a) EGG (green ∗) and (b) the proposed reference algorithm (red ○).

at which the amplitude of the EGG waveform is equal to a percentage of its maximum value within a cycle [38]. Each of the above definitions is limited to specific fields of interest. In this paper the aim is to find an analysis interval suitable for minimizing the modeling error in closed-phase LPC, hence the first definition is used. Put more precisely, we define the optimum closed-phase interval as that for which *the residual error of a fixed-order all-pole model of the speech signal is minimal*. The following section describes a reference algorithm that finds this interval.

## III. Evaluation Reference

Algorithms for speech-based GCI detection have been widely evaluated using EGG-based references [12], [22], [24]. It is known that the synchronization of EGG and speech signals is affected by the propagation time from the talker's lips to the recording microphone that may be estimated and subtracted to synchronize the two signals. Any residual synchronization error is expected to produce a constant *bias* in the GCI estimates throughout the utterance. However, with regard to GOIs, the difference between definitions is not guaranteed to be a constant bias alone; defining a suitable reference therefore requires careful consideration. Various approaches for finding optimal intervals for closed-phase LPC analysis have been proposed in [5], [8], and [9]. The following is a two-channel algorithm that is based upon the approaches in [5], and [8], operating upon both the EGG and speech signal.

### A. Proposed Reference Algorithm

As defined in Section II-E, the optimum closed-phase interval is defined as that for which the residual error of a fixed-order all-pole model of the speech signal is minimal. As a baseline approach, initial GCI and GOI estimates $\tilde{\mathbf{n}}^{\mathrm{c}}$ and $\tilde{\mathbf{n}}^{\mathrm{o}}$ are provided by analysis of the EGG signal with the SIGMA algorithm [20]. As there is no guarantee that this result represents an optimal analysis interval for closed-phase LPC, an exhaustive search is conducted over a range of intervals, centered around $\tilde{\mathbf{n}}^{\mathrm{c}}$ and $\tilde{\mathbf{n}}^{\mathrm{o}}$. It is assumed that the error in the GCI is significantly less than the error in the GOI so the search intervals are set accordingly at $n_r^{\mathrm{c}} \pm 0.05\Delta n_r$ and $n_r^{\mathrm{o}} \pm 0.2\Delta n_r$, where $\Delta n_r = n_{r+1}^{\mathrm{c}} - n_r^{\mathrm{c}}$. The quality of each estimate is evaluated with the following cost function

$$Q(\mathbf{c}_r, \mathbf{o}_r) = \frac{\mathrm{var}\left\{u'_{\mathrm{cp}}(\mathbf{c}_r)\right\}}{\mathrm{var}\left\{u'_{\mathrm{cp}}(\mathbf{o}_r)\right\}} \qquad (12)$$

where $u'_{\mathrm{cp}}(\mathbf{c}_r)$ and $u'_{\mathrm{cp}}(\mathbf{o}_r)$ denote the estimated voice source waveform from closed-phase analysis in the closed and open phases for each iteration at cycles $r$, respectively, and $\mathrm{var}\{\cdot\}$ denotes variance. The optimum window is defined as

$$\mathbf{c}_r^{\mathrm{opt}} = \underset{\mathbf{c}_r,\, \mathbf{o}_r}{\arg\min}\left(Q(\mathbf{c}_r, \mathbf{o}_r)\right). \qquad (13)$$

Optimum closed phase intervals are found for sets of three neighboring cycles according to (11) to improve robustness. The voice source signal is estimated according to (2) from the middle of each of the three cycle sets. Iteration through all analysis intervals for all voice source cycles produces $\mathbf{n}^{\mathrm{copt}}$ and $\mathbf{n}^{\mathrm{oopt}}$, respectively. It has been observed that the algorithm favors longer analysis intervals within the closed phase as it improves the conditioning of the covariance matrix $\mathbf{R}$. The technique is not particularly practical due to the requirement of an EGG signal and high computational demand; it is therefore best suited as an offline reference.

The result of the optimization scheme is exemplified in Fig. 2, which shows the voice source estimated with closed-phase LP analysis using intervals defined by (a) EGG and (b) the proposed reference algorithm on the same signal used in Fig. 1. The EGG GOIs are marked green ∗ and the optimized GOIs marked red ○. The result of this experiment demonstrates the sensitivity of closed-phase LP analysis to framing errors: the inclusion of glottal excitation in the opening phase in (a) does not give zero airflow during the closed phase, whereas in (b) the refined analysis interval gives a very flat closed phase in the estimated voice

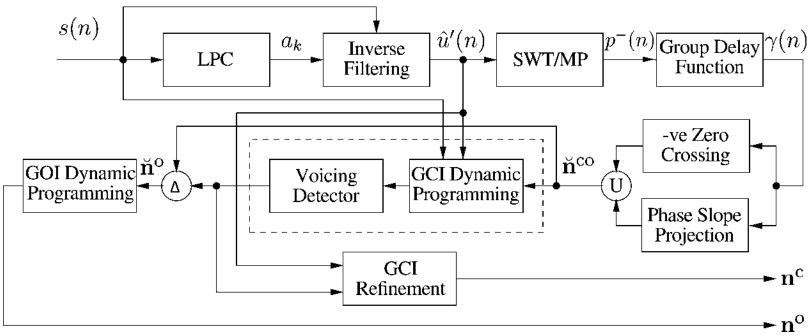

Fig. 3. System diagram. Voice source $\hat{u}'(n)$ is estimated, discontinuities reinforced with the multiscale product, $p^-(n)$ and impulsive features located with the group delay function. Candidates denoted $\breve{\mathbf{n}}^{\mathrm{co}}$. The algorithm sequentially extracts GCIs, $\mathbf{n}^{\mathrm{c}}$ and GOIs, $\mathbf{n}^{\mathrm{o}}$ with optional voicing detection.

source signal. The latter is deemed to be derived from a better estimate of $V(z)$.

Closed-phase LP analysis will generally fail if incomplete vocal fold closure occurs, such as in the case of weakly voiced speech or vocal fry. It is expected that this will cause the optimization routine in (13) to produce random closed phases, increasing the local variance of the closed quotients. In order to suppress erroneous GOIs in these regions, a sliding variance is calculated on five neighboring CQ values and those cycles in which the standard deviation exceeds 0.02 are flagged as unreliable and excluded.

## IV. The YAGA Algorithm

The *Yet Another GCI Algorithm* is a culmination of new and existing GCI/GOI detection techniques using a framework based upon the DYPSA algorithm. The aim is to find closed phase intervals that are suitable for closed phase LPC. The algorithm is split into two parts: *candidate detection* in which potential GCIs and GOIs are extracted from the speech signal and *candidate selection* in which GCIs and GOIs are selected from the candidate set. A system diagram is shown in Fig. 3.

### A. Candidate Detection

The voice source signal $\hat{u}'(n)$ is first estimated from the speech signal using the IAIF method described in Section II-B with an analysis interval of 32 ms, a frame increment of 16 ms, and a prediction order of $f_s/1000$. The multiscale product of the stationary wavelet transform (SWT) reinforces discontinuities in a signal by calculating its derivative at multiple dyadic scales and locating converging maxima [18] as previously applied to speech [22] and EGG [20] signals. A biorthogonal spline wavelet with one vanishing moment is used in this paper, with corresponding detail and approximation filters $g(n)$ and $h(n)$, respectively.

The SWT of signal $\hat{u}'(n)$, $1 \leq n \leq N$ at scale $j$ is

$$d_j^s(n) = \sum_k g_j(k) a_{j-1}^s(n-k) \tag{14}$$

where $J$ is bounded by $\log_2 N$ and $j = 1, 2, \ldots, J-1$. The approximation coefficients are given by

$$a_j^s(n) = \sum_k h_j(k) a_{j-1}^s(n-k) \tag{15}$$

where $a_0^s(n) = \hat{u}'(n)$. Detail and approximation filters are upsampled by two on each iteration to effect a change of scale. The multiscale product $p(n)$ is formed by

$$p(n) = \prod_{j=1}^{j_1} d_j(n) \tag{16}$$

where it is assumed that the lowest scale to include is always 1. The de-noising effect of the $h(n)$ at each scale in conjunction with the multiscale product means that $p(n)$ is near-zero except at discontinuities across the first $j_1$ scales of $\hat{u}'(n)$ where it becomes impulse-like. The value of $j_1$ is bounded by $J$, but in practice $j_1 = 3$ gives good localization of discontinuities [39]. Experimentation with this algorithm has shown that the performance of the subsequent group delay function-based event detector is improved by first taking the $j_1^{\text{th}}$ root of $p(n)$ and half-wave rectifying to give $p^-(n)$. This technique is further confirmed by [20].

The signal $p^-(n)$ contains sparse impulse-like features of the same sign at the location of GCIs and GOIs. In order to locate these features, the following group delay function [27] is used. Consider an $L$-sample windowed segment of $p^-(n)$ beginning at sample $n$

$$x_n(l) = w(l)p^-(n+l) \text{ for } l = 0, \ldots, L-1. \tag{17}$$

The group delay of $x_n(l)$ is given by [27]

$$\tau_n(k) = \Re\left(\frac{\tilde{X}_n(k)}{X_n(k)}\right) \tag{18}$$

where $X_n(k)$ is the discrete Fourier transform of $x_n(l)$ and $\tilde{X}_n(k)$ is the discrete Fourier transform of $lx_n(l)$. If $x_n(l) = \delta(l - l_0)$, where $\delta(l)$ is a unit impulse function, it follows from (18) that $\tau_n(k) \equiv l_0 \forall k$. For noise robustness, an averaging procedure is performed over all frequency bins as reviewed in [27]. An energy-based weighting was deemed the most appropriate [12], defined as

$$\gamma(n) = \frac{\sum_{l=0}^{L-1} l x_n^2(l)}{\sum_{l=0}^{L-1} x_n^2(l)} - \frac{L-1}{2} \tag{19}$$

which is an efficient time-domain formulation and can be viewed as the center of energy of $x_n(l)$, bounded in the range $[-(L-1)/2, (L-1)/2]$. This time-domain signal is called the *group delay function* of a signal,[1] differing from *group delay*

[1] Some authors use *phase slope function* which differs only by sign.

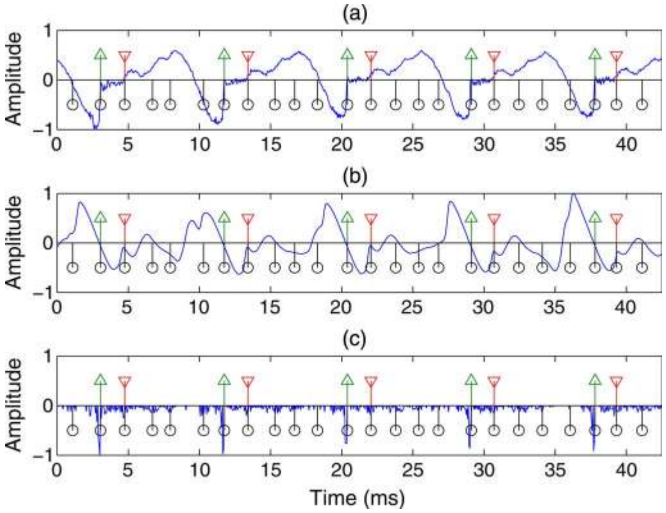

Fig. 4. (a) Estimated voice source, $\hat{u}'(n)$, (b) Group Delay Function, $\gamma(n)$, (c) Multiscale Product, $p(n)$, with overlaid candidate set (black o) and estimated GCIs (green △) and GOIs (red ▽) following the dynamic programming stage.

which is a function of frequency. The location of the negative-going zero crossings of $\gamma(n)$ give an accurate estimation of the location of impulsive features that form a set of candidate GCIs and GOIs as shown in Fig. 4(b). Additionally, if an impulsive feature is spread in time then the group delay function method will find its center of energy, which is particularly useful in the case of the "redoubled" GCI discussed in [40]. A similar approach has been applied directly to speech signals [41] in which $\tau_n(k)$ is not expected to take a constant value, nor whose mean is zero when the GCI lies in the center of the window. A suitable correction is applied that is not necessary in the case of impulsive signals [41]. The length of the group delay window is set at 2 ms, which lies within the bounds suggested in [20] and [41].

In the presence of noise, an impulsive feature may produce a local minimum that follows a local maximum without a negative-going zero crossing. The *phase slope projection* technique [12] identifies the midpoint of the local maximum and minimum and projects it onto the time axis with unit slope. The point of intersection with the time axis is added to the candidate set. The complete set of candidates for both GCIs and GOIs is denoted $\breve{\mathbf{n}}^{\mathrm{co}} = [\breve{n}_1^{\mathrm{co}}\ \breve{n}_2^{\mathrm{co}}\ \ldots\ \breve{n}_{\breve{R}}^{\mathrm{co}}]^T_{\breve{R}\times 1}$.

### B. Candidate Selection

The candidate selection applies $N$-best Dynamic Programming [29] to find a path that minimizes a set of costs in order to detect GCIs, $\mathbf{n}^{\mathrm{c}}$, only. A similar methodology is employed in [12]. A second stage detects GOIs from the remaining candidates by considering the consistency of the closed quotient of the remaining candidates relative to estimated GCIs. This sequential approach is required because both GCI and GOI candidates arise from positive-going discontinuities in the voice source signal.[2] Voicing detection removes erroneous detections during unvoiced speech. The output of the candidate selection is depicted in Fig. 4, showing candidates (black) and detected GCIs (green), GOIs (red) overlaid on (a) estimated voice source signal, $\hat{u}'(n)$, (b) the group delay function, $\gamma(n)$, and (c) the multiscale product of the voice source signal, $p^{-}(n)$.

*1) $N$-Best Dynamic Programming:* The GCI dynamic programming minimizes the following function over a finite subset of candidates, $\Omega$, of size $|\Omega|$

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \boldsymbol{\lambda}^T \boldsymbol{\zeta}_{\Omega}(r) \tag{20}$$

where $\boldsymbol{\lambda} = [\lambda_A\ \lambda_P\ \lambda_J\ \lambda_F\ \lambda_S\ \lambda_C]^T$ is a vector of weighting factors, and $\boldsymbol{\zeta}(r) = [\zeta_A(r)\ \zeta_P(r)\ \zeta_J(r)\ \zeta_F(r)\ \zeta_S(r)\ \zeta_C(r)]^T$ is a vector of cost elements evaluated at the $r$th GCI of the subset, normalized in the range $-0.5 \le \zeta_k(r) \le 0.5$, as defined in [12]. The cost vector elements are as follows.

- *Waveform similarity*, $\zeta_A(r)$, between $\hat{u}'(n)$ in neighboring candidates, where candidates not correlated with the previous candidate are penalized.
- *Pitch deviation*, $\zeta_P(r)$, between the current and the previous two candidates, where candidates with large deviation are penalized.
- *Projected candidate cost*, $\zeta_J(r)$, for the candidates from the phase-slope projection, which are sometimes erroneous. $\zeta_J(r) = 0.5$ for projected candidates and $-0.5$ otherwise.
- *Normalized energy*, $\zeta_F(r)$, which penalizes candidates that do not correspond to high energy in the speech signal.
- *Ideal phase-slope function deviation*, $\zeta_S(r)$, where candidates arising from zero-crossings with gradients close to unity are favored.
- *Closed phase energy*, $\zeta_C(r)$. The energy contained in $\hat{u}'(n)$ between successive candidates. Glottal closure causes $\zeta_C(r)$ to be low.

The first five costs are calculated with mappings defined in [12]. The closed phase energy cost is defined as

$$\zeta_C(r) = \frac{\|\hat{u}'\,(n'_r)\|_2}{\max\limits_{k} \|\hat{u}'\,(n'_k)\|_2} - 0.5,\ k = 1, 2, \ldots \breve{R} - 1 \tag{21}$$

where $\breve{n}_r^{\mathrm{c}} \le n'_r < \breve{n}_{r+1}^{\mathrm{c}}$.

*2) GCI Refinement:* The zero crossings of the group delay function correspond to local centers of energy in the voice source signal that lie in the vicinity of the maximum discontinuity in the voice source. In order to reduce small errors caused by nonideal impulsive behavior, the maximum positive-going derivatives of the voice source signal lying within 0.5 ms of the zero crossing are identified. In [41], in which the group delay function is applied to the speech signal directly, the minimum phase component of the speech signal is considered as mentioned in Section IV-A. Such an explicit model of the phase behavior of $p^{-}(n)$ is not applied in this case as the proposed correction has been found to be sufficient here.

*3) Voicing Detection:* The waveform similarity measure is useful not only for eliminating unlikely candidates but it also serves as a reliable measure of voicing. This is required to suppress erroneous GCI/GOIs during unvoiced and silent segments. The duration of voiced segments is relatively long compared with the fundamental period of voicing, $T_0$. This permits smoothing of the waveform similarity cost $\zeta_A(r)$ to help suppress sudden changes which could result in an

[2] This is dissimilar to the EGG signal in which GCI and GOI candidates correspond to discontinuities of opposite sign in the EGG waveform [37].

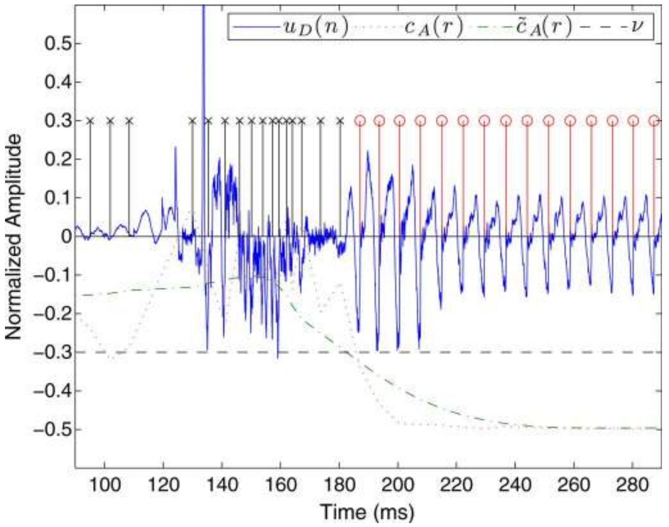

Fig. 5. Segment of $\hat{u}'(n)$ showing silence-unvoiced-voiced transitions, waveform similarity cost $\zeta_A(r)$ smoothed waveform similarity cost $\tilde{\zeta}_A(r)$ and threshold $\nu$. $\tilde{\zeta}_A(r)$ provides a good voicing detector; when less than $\nu$, GCIs are kept $(\circ)$, else they are rejected $(\times)$. GOIs not displayed for clarity.

erroneous voicing decision. Let $\tilde{\zeta}_A(r) = \zeta_A(r) * w(r)$ be a smoothed waveform similarity cost, where $w(r)$ is a Hamming window of length 1 ms. A fixed threshold $\nu$ is used to make a voiced/unvoiced decision

$$v(r) = \begin{cases} 1, & \text{if } \tilde{\zeta}_A(r) < \nu \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

The parameter $\nu$ is set empirically to $-0.3$. An example of a voiced/unvoiced decision is shown in Fig. 5, showing $\zeta_A(r)$, $\tilde{\zeta}_A(r)$ and the GCIs that are accepted or rejected. During periods of weakly voiced speech, vocal fry or registers that do not produce a discontinuity in the voice source signal, no suitable candidates will be found. The output of the voicing detector is therefore nonzero during modal voiced speech only.

*4) GOI Detection:* It was stated that the aim is to find GOIs that are best-suited to closed phase LPC analysis. It was shown in Section IV that too long an analysis interval can impair the quality of the estimated vocal tract filter; in the example of Figs. 1 and 2, there exist in the estimated voice source signal two close discontinuities of similar amplitude within each cycle, the earlier of which is shown to be best-suited to closed-phase LPC. It has been found that these discontinuities produce candidates that have similar costs $\boldsymbol{\zeta}$, and as such an alternative approach to that described in Section IV-B is required. It is proposed that a set of GOI candidates is defined as

$$\{\breve{\mathbf{n}}^{\mathrm{o}}\} = \{\breve{\mathbf{n}}^{\mathrm{co}}\} \Delta \{\mathbf{n}^{\mathrm{c}}\} \tag{23}$$

where $\breve{\mathbf{n}}^{\mathrm{o}} = [\breve{n}_1^{\mathrm{o}}\ \breve{n}_2^{\mathrm{o}}\ \dots\ \breve{n}_{\breve{R}^{\mathrm{o}}}^{\mathrm{o}}]^T_{\breve{R}^{\mathrm{o}}\times 1}$ and $\triangle$ denotes the symmetric difference (union minus intersection) of the two sets. The closed quotients (CQ) of $\breve{\mathbf{n}}^{\mathrm{o}}$ relative to $\mathbf{n}^{\mathrm{c}}$, termed $\mathbf{Q}^{\mathrm{c}}$, are calculated for all candidates $\mathbf{n}^{\mathrm{o}}$. The best path is deemed to be the lowest path of consistent CQ values. A dynamic programming algorithm finds the best path by searching for sets of three candidates with CQ within $\xi$ of one another. A state variable $\rho$ saves the previous good CQ, empirically initialized to 0.2, so that artificial GOIs may be inserted when no suitable candidates are found. Fig. 6 shows (a) a speech signal and (b) the candidates' CQ $(\circ)$ and with the best path overlaid. The examples in Figs. 1 and 2 correspond to time ~0.2 s in this figure. Visual inspection reveals multiple tracks when excitation is present at both the beginning and ending of the opening phase as discussed in Section II-E. By initializing $\rho$ to different values and using alternative search criteria different paths may be found. The estimated GOIs are denoted $\mathbf{n}^{\mathrm{o}}$.

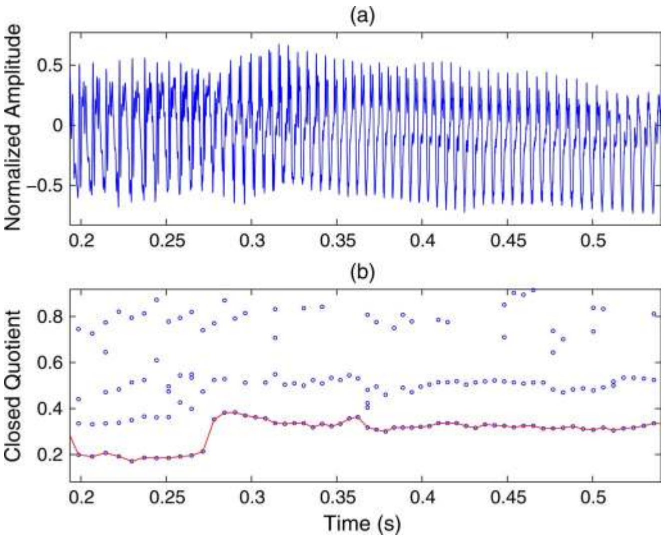

Fig. 6. (a) Speech signal and (b) CQ of GOI candidates $(\circ)$ with best path.

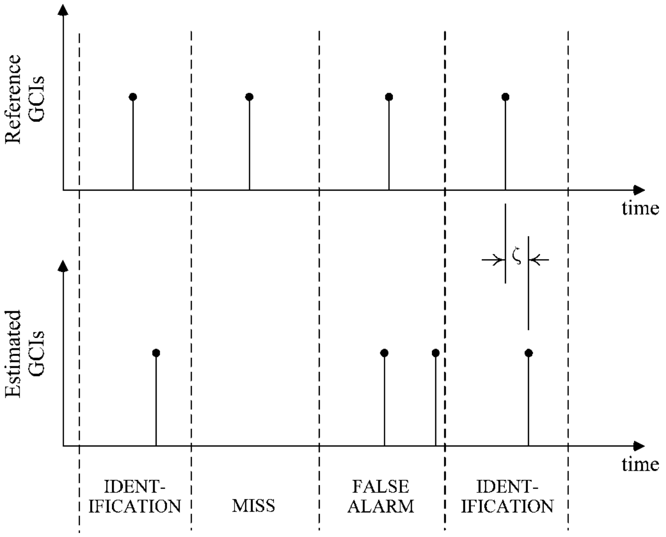

Fig. 7. Characterization of GCI Estimates showing four larynx cycles with examples of each possible outcome from GCI estimation.

## V. Performance Assessment

The YAGA algorithm was configured with cost weights $\boldsymbol{\lambda} = [\lambda_A\ \lambda_P\ \lambda_J\ \lambda_F\ \lambda_S\ \lambda_C]^T = [0.8\ 0.6\ 0.4\ 0.3\ 0.1\ 0.5]^T$ and CQ tolerance $\xi = 0.1$. The first five elements of $\boldsymbol{\lambda}$ were optimized in [12] and $\lambda_C$ and $\xi$ were trained on 10% of the APLAWD database which was omitted for the following tests.

### A. Evaluation Methodology

The APLAWD database [42] contains speech and contemporaneous EGG recordings of five short sentences, repeated ten times by five male and five female talkers. A subset of the SAM

TABLE I
GCI/GOI PERFORMANCE ON THE APLAWD DATABASE

| | ID Rate (%) | Miss Rate (%) | FA Rate (%) | FAT Rate (%) | Bias, $\mu$ (ms) | ID Acc., $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| SIGMA (EGG) GCI | 100.00 | 0.00 | 0.00 | 0.00 | -0.04 | 0.12 |
| SIGMA (EGG) GOI | ” | ” | ” | ” | 1.07 | 0.70 |
| DYPSA GCI | 96.39 | 1.54 | 2.07 | 42.81 | 0.11 | 0.65 |
| DYPSA GOI | ” | ” | ” | ” | 0.54 | 0.53 |
| YAGA GCI | 99.31 | 0.18 | 0.51 | 45.46 | -0.01 | 0.39 |
| YAGA GOI | ” | ” | ” | ” | 0.10 | 0.63 |
| YAGA GCI + V. det | 94.84 | 5.04 | 0.13 | 5.95 | -0.01 | 0.38 |
| YAGA GOI + V. det | ” | ” | ” | ” | 0.11 | 0.63 |

TABLE II
GCI/GOI PERFORMANCE ON THE SAM DATABASE

| | ID Rate (%) | Miss Rate (%) | FA Rate (%) | FAT Rate (%) | Bias, $\mu$ (ms) | ID Acc., $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| SIGMA (EGG) GCI | 100.00 | 0.00 | 0.00 | 0.00 | -0.03 | 0.08 |
| SIGMA (EGG) GOI | ” | ” | ” | ” | 0.94 | 0.53 |
| DYPSA GCI | 95.41 | 2.08 | 2.50 | 55.22 | -0.02 | 0.40 |
| DYPSA GOI | ” | ” | ” | ” | 0.06 | 0.39 |
| YAGA GCI | 98.80 | 0.43 | 0.77 | 58.98 | -0.11 | 0.31 |
| YAGA GOI | ” | ” | ” | ” | -0.21 | 0.54 |
| YAGA GCI + V. det | 90.87 | 9.00 | 0.16 | 3.67 | -0.11 | 0.27 |
| YAGA GOI + V. det | ” | ” | ” | ” | -0.24 | 0.54 |

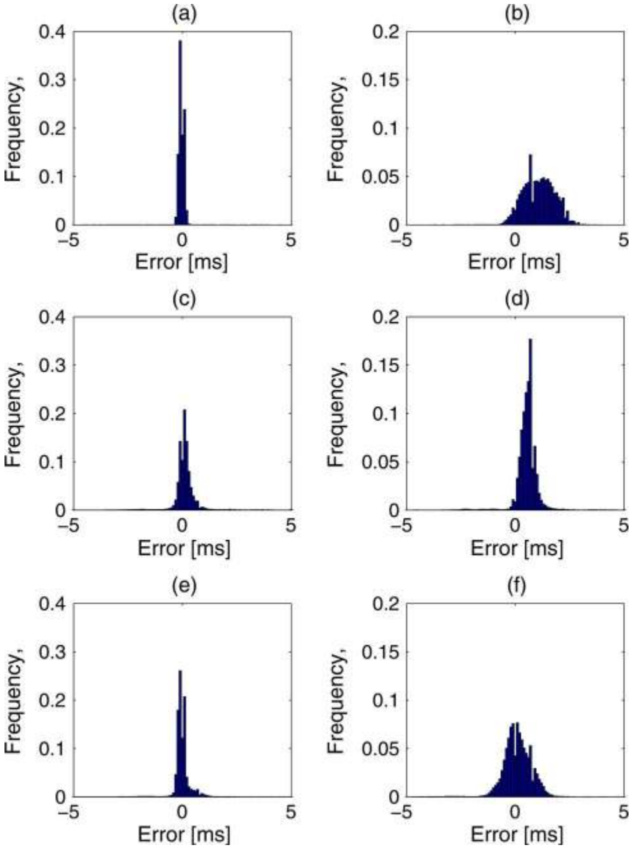

Fig. 8. Performance results on the APLAWD database for (a) SIGMA (EGG) GCI, (b) SIGMA (EGG) GOI, (c) DYPSA GCI, (d) DYPSA GOI, (e) YAGA GCI, and (f) YAGA GOI. The bin interval is 0.1 ms.

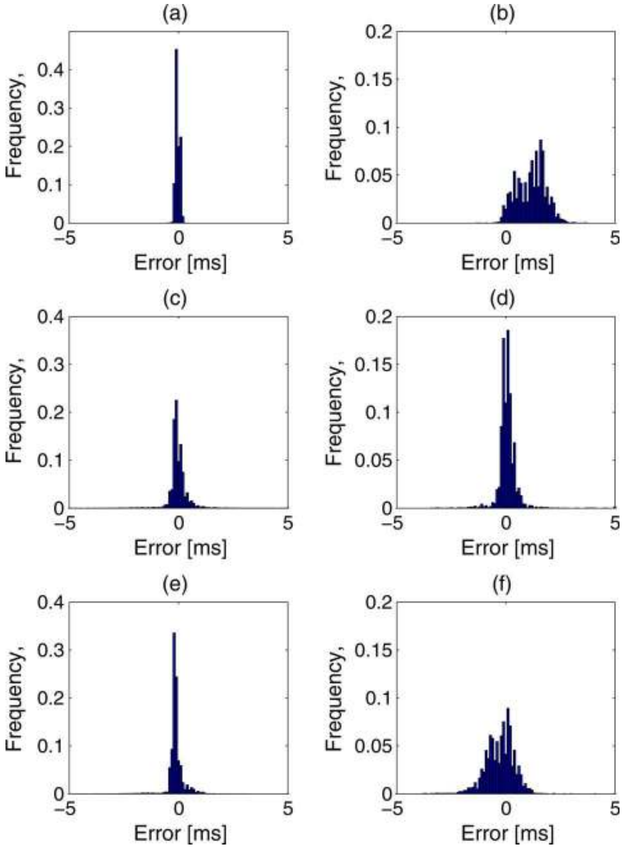

Fig. 9. Performance results on the SAM database for (a) SIGMA (EGG) GCI, (b) SIGMA (EGG) GOI, (c) DYPSA GCI, (d) DYPSA GOI, (e) YAGA GCI, and (f) YAGA GOI. The bin interval is 0.1 ms.

database [43] contains EGG and speech signals of duration approximately 150 seconds by two male and two female speakers. Estimated GCIs and GOIs were derived from the EGG signals with SIGMA and from the speech signals with DYPSA and YAGA. Using the algorithm described in Section III as a reference, the performance of these algorithms was evaluated using the strategy defined in [12] as depicted in Fig. 7. *Detection rate* is the percentage of all reference GCI periods for which exactly one GCI is estimated. *Accuracy*, $\sigma$, and *bias*, $\mu$, are respectively the standard deviation and mean of the error, $\zeta$, between estimated and reference GCIs. In the case of GOIs, accuracy and bias are measured only on those closed phases for which the reference was flagged as accurate. *False alarm rate* is the percentage of all reference GCI periods for which more than one GCI is estimated and *Miss rate* is the percentage of all reference GCI periods for which no GCIs were estimated. False alarms are not counted if they occur between voiced segments separated by more than 3 ms. *False Alarm Total (FAT)*, measures *all* false alarms as a proportion of total candidates, including those between voiced segments. This helps to assess the quality of voicing detection and the suppression of multiple false alarms within one reference cycle.

### B. Results and Discussion

Results are recorded in Tables I and II with corresponding error histograms in Figs. 8 and 9. GCI and GOI hit rates are necessarily equal and so are stated once in each case for clarity. The initial estimates given to the proposed reference algorithm were derived from EGG signal by the SIGMA algorithm. Only the positions of the GCIs and GOIs were altered so ID, miss, false alarm and FAT rate are perfect by definition.

With regard to GCI detection, the EGG-based SIGMA algorithm exhibits the lowest error standard deviation of all methods under test. There exists a small bias that can be attributed to synchronization error between speech and EGG signals. The YAGA algorithm delivers an identification rate in excess of 99.3% on APLAWD and 98.8% on SAM with negligible bias and an identification error of within 0.3–0.4 ms. The DYPSA algorithm, whose candidate generation relies upon the LPC residual as opposed to the multiscale product of the voice source signal, fairs worst with ID rate at 3% below YAGA. YAGA's high GCI accuracy can be attributed to the GCI refinement following candidate selection that is not performed in DYPSA, although both candidate selection routines have much in common. The YAGA voicing detector heavily suppresses FAT by 40%–55% at the expense of increasing misses by 5%–10%; this has little effect upon bias and accuracy. Future improvements are expected to use through dynamic, rather than static, voicing decision thresholds.

The GOI performance of SIGMA's EGG-based estimates shows a positive bias of around 1 ms on both databases, as predicted by the examples in Section III. SIGMA's relatively high error standard deviation is not necessarily indicative that SIGMA contains error in its estimates but that the difference between GOIs in the EGG signal and GOIs for the ideal closed-phase analysis interval is not a constant bias. The histogram (b) shows that the EGG GOI rarely occurs before the closed-phase GOI; the relationship between these two definitions is most likely to be related to the duration of the closed phase. DYPSA, which estimates GOIs from a fixed CQ of 0.3, shows identification accuracy of 0.4–0.5 ms, seemingly the best of all three methods under test. YAGA shows slightly worse accuracy than DYPSA; however, this statistic does not represent the results of inverse-filtering by visual inspection that are similar to the results in Fig. 2. Further refinement of the estimated GOIs, possibly by exhaustive search as in the proposed reference algorithm but over a smaller interval, may be necessary to further improve the GOI estimation.

The results indicate that the proposed method is reliable when applied to natural conversational speech signals. Informal testing with additive noise sources has shown that similar identification rates can be achieved with white Gaussian and babble noise down to about 15-dB signal-to-noise ratio. In the presence of reverberation, a significant reduction in identification rate is seen with reverberation times of greater than 100 ms. It was further observed that the accuracy of the identified GCIs/GOIs is less sensitive to such distortions than identification rate.

## VI. Conclusion

The YAGA algorithm was proposed for the detection of GCIs and GOIs from speech signals. The approach is a culmination of existing methods that estimates a set of candidate GCIs and GOIs, from which the best path through the GCI candidates is found. A new approach for detecting GOIs was proposed that finds the lowest consistent track of the candidates' closed quotients relative to the estimated GCIs. Optional voicing detection suppresses detections during unvoiced speech and silence. The precise definition of the closed phase was related to the analysis interval for closed-phase LPC analysis, for which a reference algorithm estimates optimal closed phases jointly from EGG and speech signals. An important outcome was demonstrating that closed-phase intervals from the EGG signal are not always suitable for closed-phase LPC analysis as the GOIs tend to be positively biased towards the end of the opening phase, whereas speech and EGG GCIs are highly coherent. The proposed YAGA algorithm, the DYPSA algorithm and the EGG-based SIGMA algorithm were evaluated against the reference algorithm on the APLAWD and SAM databases. YAGA achieved a GCI hit rate of ∼99% on both databases with GCI and GOI hit accuracy of 0.3–0.4 ms and 0.5–0.6 ms respectively.

## References

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–467, Dec. 1990.

[2] N. D. Gaubitch, E. A. P. Habets, and P. A. Naylor, "Multi-microphone speech dereverberation using spatio-temporal and spectral processing," in *Proc. Int. Symp. Circuits Syst.*, Seattle, WA, May 2008.

[3] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice source waveform modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3965–3968.

[4] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3793–3796.

[5] 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. ASSP-27, no. 4, pp. 350–355, Aug. 1979.

[6] P. Davies, G. A. Lindsey, H. Fuller, and A. J. Fourcin, "Variation of glottal open and closed phases for speakers of English," *Proc. Inst. Acoust.*, vol. 8, no. 7, pp. 539–546, 1986.

[7] R. C. Scherer, V. J. 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. J. Raphael, Eds. Melville, NY: Amer. Inst. of Phys., 1995, pp. 269–290.

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

[9] 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, no. 5, pp. 569–576, Sep. 1999.

[10] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[11] J. G. McKenna, "Automatic glottal closed-phase location and analyis by Kalman filtering," in *Proc. 4th ISCA Tutorial Res. Workshop Speech Synth.*, Aug. 2001.

[12] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[13] P. Chytil and M. Pavel, "Variability of glottal pulse estimation using cepstral method," in *Proc. 7th Nordic Signal Process. Symp. (NORSIG)*, 2006, pp. 314–317.

[14] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, Oct. 2007.

[15] 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, no. 2, pp. 258–265, Apr. 1994.

[16] S. K. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 917–924, Mar. 1992.

[17] N. Sturmel, C. d'Alessandro, and F. Rigaud, "Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 4517–4520.

[18] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 617–643, Mar. 1992.

[19] A. Bouzid and N. Ellouze, "Electroglottographic measures based on gci and goi detection using multiscale product," *Int. J. Comput., Commun., Control*, vol. III, pp. 21–32, 2008.

[20] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1557–1566, Nov. 2009.

[21] A. Bouzid and N. Ellouze, "Open quotient measurements based on multiscale product of speech signal wavelet transform," *Res. Lett. Signal Process.*, 2007.

[22] W. Saidi, A. Bouzid, and N. Ellouze, "Evaluation of multi-scale product method and DYPSA algorithm for glottal closure instant detection," in *Proc. 3rd Int. Conf. Inf. Commun. Technol.: From Theory to Applicat. (ICTTA)*, Apr. 2010, pp. 1–5.

[23] 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.

[24] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614–624, May 2009.

[25] A. Bouzid and N. Ellouze, "Empirical mode decomposition of voiced speech signal," in *Proc. Int. Symp. Control, Commmun., Signal Process.*, Hammamet, Tunisia, Mar. 2004, pp. 603–606.

[26] M. A. Huckvale, Speech Filing System: Tools for Speech Univ. College London, 2004 [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs, Tech. Rep.

[27] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 456–466, Mar. 2006.

[28] 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. (ICASSP)*, May 1995, pp. 776–779.

[29] 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. (ICASSP)*, 1990, pp. 81–84.

[30] H. Fujisaki and M. Ljungqvist, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1987, vol. 12, pp. 637–640.

[31] A. H. Gray and J. D. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-22, no. 3, pp. 207–217, Jun. 1974.

[32] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1985, vol. 10, pp. 937–940.

[33] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[34] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, Feb. 1971.

[35] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive filtering," *Speech Commun.*, vol. 11, pp. 109–118, 1992.

[36] D. S. F. Chan and D. M. Brookes, "Variability of excitation parameters derived from robust closed phase glottal inverse filtering," in *Proc. Eur. Conf. Speech Commun. Technol.*, Sep. 1989, vol. 33, no. 1.

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

[38] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," *J. Speech. Hear. Res.*, vol. 31, no. 3, pp. 338–351, Sep. 1988.

[39] B. M. Sadler and A. Swami, "Analysis of multiscale products for step detection and estimation," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 1043–1051, Apr. 1999.

[40] N. Henrich, C. d'Alessandro, M. Castellengo, and B. Doval, "On the use of the derivative of electroglottographic signals for characterization of nonpathological voice phonation," *J. Acoust. Soc. Amer.*, vol. 115, no. 3, pp. 1321–1332, Mar. 2004.

[41] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Beijing, China, Oct. 2000, vol. 4, pp. 664–667.

[42] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Univ. College London, Jun. 1987, Tech. Rep..

[43] 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 Commun. Technol.*, Sep. 1995, pp. 867–870.

**Mark R. P. Thomas** (S'06–M'09) received the M.Eng. degree in electrical and electronic engineering and the Ph.D. degree from Imperial College London, London, U.K., in 2006 and 2010, respectively.

His research interests include glottal-synchronous speech processing and multichannel acoustic signal processing. He has industrial experience with audio, video, and RF in the field of broadcast engineering. He is currently a Research Associate with the Communications and Signal Processing Group at Imperial College London.

Dr. Thomas has been a member of the IEEE Signal Processing Society since 2006.

**Jon Gudnason** (M'96) received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Iceland, Reykjavik, in 1999 and 2000, respectively, and the Ph.D. degree with the Communications and Signal Processing Group, Imperial College London, London, U.K., in 2007.

In 1999, he was a Research Assistant for the Information and Signal Processing Laboratory, University of Iceland, working on remote sensing applications and from 2001 to 2009 he was a Research Assistant with the Communications and Signal Processing Group, Imperial College London, where his research focused on speaker recognition and automatic target recognition using radar. From 2008 to 2009, he was a Visiting Scholar at LabROSA, Columbia University, New York. Since 2009, he has been a Member of the Academic Staff at the School of Science and Engineering, Reykjavik University.

Dr. 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.

**Patrick A. Naylor** (M'89–SM'07) received the B.Eng. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., in 1986 and the Ph.D. degree from Imperial College London, London, U.K., in 1990.

Since 1990, he has been a Member of Academic Staff in the Department of Electrical and Electronic Engineering, Imperial College London, where he is also Director of Postgraduate Studies. His research interests are in the areas of speech, audio, and acoustic signal processing. He has worked in particular on adaptive signal processing for dereverberation, blind multichannel system identification and equalization, acoustic echo control, speaker identification, single and multi-channel speech enhancement, and speech production modeling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the U.K., USA, and in mainland Europe.

Dr. Naylor is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and an Associate Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing.