DETECTION OF GLOTTAL CLOSING AND OPENING INSTANTS USING AN IMPROVED DYPSA FRAMEWORK

Mark R. P. Thomas, Jon Gudnason and Patrick A. Naylor

Communications and Signal Processing Group, Imperial College
Exhibition Road, SW7 2AB, London, UK
phone: + (44) 20 7594 6235, fax: + (44) 20 7594 6234,
email: {mark.r.thomas02, jon.gudnason, p.naylor} @imperial.ac.uk
web: www.commsp.ee.ic.ac.uk/

ABSTRACT

Accurate estimation of glottal closure instants (GCIs) and opening instants (GOIs) is important for speech processing applications that benefit from glottal-synchronous processing.

This paper proposes a novel improvement to the DYPSA framework, based upon a multiscale analysis technique and an accurate estimation of glottal volume velocity. This replaces the linear prediction residual for candidate selection and enables the reliable detection of both GCI and GOI candidates. A two-stage dynamic programming process then detects the GCIs and removes them from the candidate set, before detecting GOIs from the remaining candidates. A post-processing step improves GOI detection using the estimated GCIs.

Evaluation against hand-labelled data on a large speech database shows that GCI detection is marginally improved compared with original DYPSA at 96% but, more importantly, shows that GOI detection can be achieved to a similar accuracy of 95%.

1. INTRODUCTION

In voiced speech, the primary acoustic excitation normally occurs at the instant of vocal fold closure which is defined as the glottal closure instant (GCI). This marks the start of the closed phase, during which there is little or no airflow through the glottis. Following the closed phase, the vocal-folds open, often creating a smaller secondary excitation in the source signal at the time defined as the glottal opening instant (GOI). The GCIs, and especially GOIs, are difficult to locate in the recorded speech signal due to spectral shaping by the vocal tract.

The detection of glottal closure instants (GCIs) in voiced speech is important for glottal-synchronous speech processing algorithms such as pitch tracking, prosodic speech modification, speech dereverberation, and certain areas of speech synthesis. Identification of glottal opening instants (GOIs) is necessary for closed-phase linear predictive coding [1] and pathological speech analysis that relies on the ratio of the open phase to the cycle period, termed the open quotient (OQ) [2].

Automatic identification of GCIs has been an aim of speech researchers for many years for which numerous techniques have been proposed. A widely used approach is the detection of discontinuities in a linear model of speech production [3, 4]. The use of a group delay measure to determine the acoustic excitation instants was first proposed in [5] and later refined in [6] and [7]. The method calculates the frequency-averaged group delay over a sliding window applied to the linear prediction residual. It has been found to be an effective method for locating the GCIs from the linear prediction residual of speech. The technique was employed in the DYPSA algorithm [8] which provides improved GCI estimates by employing phase-slope projection [9] and dynamic programming (DP).

Glottal opening instants are more difficult to detect owing to the small effect they cause in both speech signals and the corresponding linear prediction residual, $e(n)$. In this paper, we propose a new preprocessor that is novel in i) the use of a new preemphasis technique in the estimation of the derivative of glottal flow, $u'(n)$, and ii) application of a multiscale product [10] for the detection of discontinuities in $u'(n)$. Using the group delay function, a candidate set is derived, then a two-stage dynamic programming is performed. The first stage detects the GCIs, which are of high amplitude compared with GOIs, that are then removed from the candidate set. The second DP stage detects a preliminary set of GOIs, upon which a postprocessing stage removes erroneous detections and inserts missing GOIs using the periodicity of the GCIs as a reference. The new algorithm's performance is evaluated against a hand-labelled database of speech signals and shows that GCI detection is marginally improved over the existing version but, more importantly, that GOIs can be detected to a similarly high degree of accuracy.

The remainder of this paper is organized as follows: Section 2 describes the enhanced inverse-filtering method for deriving the the approximate glottal flow derivative. Section 3 reviews the operation of the DYPSA algorithm application and describes the proposed preprocessing and GOI postprocessing stages. Evaluation results of the GCI and GOI detection against hand-labelled data is presented in Section 4 and conclusions are drawn in Section. 5.

2. EXCITATION IN VOICED SPEECH

Voiced excitation instants are difficult to locate in speech signals because of spectral shaping by the vocal tract transfer function, $V(z)$. Detection of excitation instants is more easily performed on an estimate of the excitation signal by removing the spectral contribution of the vocal tract.

2.1 The Source-Filter Model

Consider the source-filter model of speech production [3]. Let $s(n)$ be a frame of voiced speech with z-transform $S(z)$ such that

$$S(z) = D(z)G(z)V(z)R(z) = U'(z)V(z), \qquad (1)$$

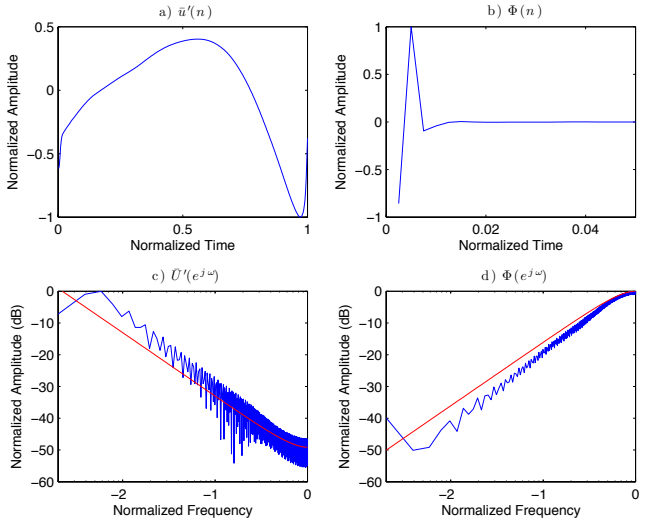

Figure 1: a) Time-domain excitation signal, $\bar{u}'(n)$, b) Time-domain inverse excitation signal, $\phi(n)$, c) Frequency-domain excitation signal, $\bar{U}'(e^{j\omega})$ (blue) and 1st order integrator (red), d) Frequency-domain inverse excitation signal, $\Phi(e^{j\omega})$ (blue) and 1st order differentiator (red).

where $D(z)$ is a periodic sequence of unit impulses, $G(z)$ is a glottal pulse shaping filter, $V(z)$ is an all-pole vocal tract filter and $R(z) \simeq 1 - z^{-1}$ models lip-radiation. The term $D(z)G(z)$ and the differential effect of $R(z)$ are usually combined and described as the glottal volume flow derivative, $U'(z)$, with time-domain waveform $u'(n)$.

In order to determine the excitation signal, the all-pole filter, $V(z)$, is estimated with LPC; however, LPC cannot distinguish between the spectral contribution of $V(z)$ and $U'(z)$. Consider then a preemphasis filter, $\Phi(z)$, which compensates for the spectral contribution of $G(z)R(z)$, such that

$$\tilde{S}(z) = S(z)\Phi(z) \simeq D(z)V(z), \tag{2}$$

where $\tilde{S}(z)$ is the preemphasised speech signal. Modelling $G(z)R(z)$ as a pole near unity, $\Phi(z)$ is usually chosen to be single zero placed below 50 Hz. As the term $D(z)$ is approximately white compared with vocal tract filter, $V(z)$ can be estimated with LPC as $\hat{V}(z) \simeq V(z)$.

It is common to derive a linear prediction residual, $e(n)$ with z-transform $E(z)$, as the result of inverse-filtering $\tilde{S}(z)$ with $\hat{V}(z)$,

$$E(z) = \frac{\tilde{S}(z)}{\hat{V}(z)} = \frac{D(z)V(z)}{\hat{V}(z)} \simeq D(z) + \eta(z), \tag{3}$$

where $\eta(z)$ is an additive noise term. $E(z)$ is useful for glottal closure detection [8] and coding [11] but the noise power of $\eta(z)$ is significant enough to mask any evidence of glottal opening.

Many models of the glottal excitation waveform such as [12, 13, 4] model $U(z)$ or $U'(z)$ where the effect of glottal opening constitutes a more significant contribution. In order to approximate $U'(z)$, an inverse-filtering operation is performed on $S(z)$ and not $\tilde{S}(z)$ as in (3),

$$\hat{U}'(z) = \frac{S(z)}{\hat{V}(z)} = \frac{U'(z)V(z)}{\hat{V}(z)} \simeq U'(z) + \eta(z). \tag{4}$$

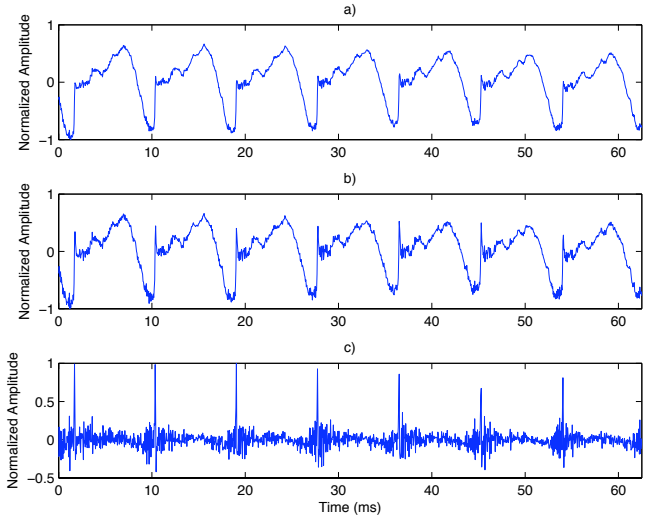

Figure 2: a) $u'(n)$ derived with proposed preemphasis filter, b) $u'(n)$ derived with standard preemphasis filter, c) $e(n)$

Though theoretically valid, this relies on $\hat{V}(z)$ being a good approximation to $V(z)$, which is in itself reliant on the preemphasis filter fully removing the spectral effects of $U'(z)$. This presents a contradiction as $U'(z)$ should therefore be known prior to the LPC analysis. The following subsection describes an approach that addresses this problem.

2.2 Improved Preemphasis

Let us assume that the single pole model of $u'(n)$, and hence the single zero preemphasis filter, are over-simplified. Consider instead a 'prototype' waveform that represents an average excitation waveform, $\bar{u}'(n)$. From this, an enhanced preemphasis filter, $\phi_{enh}(n)$ with z-transform $\Phi_{enh}(z)$, can be derived with least-squares inverse filtering that satisfies

$$\phi_{enh}(n) * \bar{u}'(n) \simeq \delta(n). \tag{5}$$

The waveform $\bar{u}'(n)$ is calculated by averaging glottal excitation cycles of $\hat{u}'(n)$ derived in (4) from a large database of continuous speech so as to attenuate noise and any remaining effects of $V(z)$ not removed by inverse filtering,

$$\bar{u}'(n) = \sum_r u'_r(n), \tag{6}$$

where r is the cycle index and $u'_r(n)$ is a glottal cycle in $u'(n)$. The averaging operation in (6) is rendered scale and amplitude independent by first resampling each cycle to a constant 20 ms in length and normalizing their A-weighted energy [16]. The speech is obtained from the APLAWD database [14] which contains contemporaneous EGG and audio recordings of ten repetitions of five phonetically-balanced English sentences spoken by five male and five female talkers, sampled at 20 kHz. The SIGMA algorithm [15] detects GCIs from the EGG signal which are refined by finding the maximum in $e(n)$ that lies in the vicinity of $\pm$ 0.5 ms of each SIGMA-derived GCI. This corrects for small deviations in the EGG-to-speech time alignment in the database. All utterances of one sentence (100 in total) were excluded from the training data for use in evaluation in Section 4.

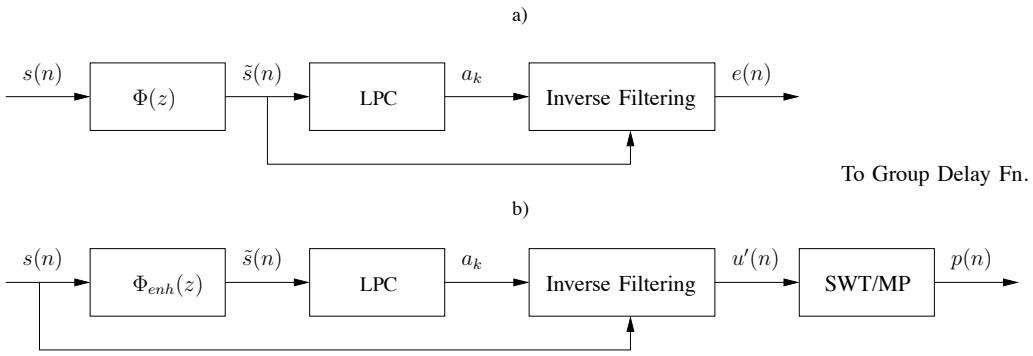

Figure 3: a) Original DYPSA preprocessor whose output is a linear prediction residual, $e(n)$. b) Enhanced DYPSA preprocessor where $\Phi_{enh}(z)$ is derived from an average glottal waveform and whose output is the multiscale product of the derivative of glottal volume velocity, $u'(n)$.

Figure 1 shows a) $\bar{u}'(n)$, b) its least-squares inverse filter and c), d) their corresponding frequency-domain plots. Plot a) exhibits little evidence of glottal opening as the varying open quotients have averaged to a smoothly-varying waveform. Strong excitation is present at the instants of glottal closure towards the ends of the cycle as they are aligned so that they coincide and reinforce. In b), only the first few taps of the inverse filter are shown as it is close to a perfect differentiator with the majority of taps close to zero. The frequency domain plots in c) and d) show straight slopes of slightly greater gradient than 6 dB/oct predicted by the traditional single pole model. $\bar{u}'(n)$ can be thought of as a representation of the average glottal pulse for all speakers and all f_0 in the training corpus.

Figure 2 shows three types of excitation signal: a) $u'(n)$ derived using $\Phi_{enh}(n)$ preemphasis, b) $u'(n)$ derived with conventional preemphasis and c) the linear prediction residual, $e(n)$. The key improvements in (a) compared with (b) are the reduced noise during the closed phase (the flat portion of the waveform) and the reduced overshoot at the glottal closure instant caused by improved estimation of the vocal tract filter coefficients. Though glottal closure instants are clearly seen as spikes in (c), little evidence of glottal opening is seen as it is buried in the noise floor.

3. CANDIDATE GENERATION AND SELECTION

The current release of DYPSA [8] derives a GCI candidate set by peak detection in the linear prediction residual, $e(n)$, using a group delay function [9]. We propose an improvement in the form of a new method for deriving a candidate set that detects peaks in the multiscale product of $u'(n)$. This section describes the main components of the DYPSA algorithm, followed by the proposed preprocessor enhancements and the GOI postprocessing stage.

3.1 The DYPSA Algorithm

The DYPSA algorithm comprises three main parts:

(i) *Group Delay Function* – defined as the average slope of the unwrapped phase spectrum of the short time Fourier transform of the prediction residual. GCI candidates are selected based on the negative-going zero crossings of the group delay function.

(ii) *Phase-Slope Projection* – introduced to generate GCI candidates when a local maximum is followed by a local minimum without crossing a zero. The midpoint between these is identified and projected onto the time axis with unit slope. In this way, GCIs whose negative-going slope does not cross the zero point (those missed by the group delay function) are identified.

(iii) *Dynamic Programming (DP)* – uses known characteristics of voiced speech (such as pitch consistency and waveform similarity) and forms a cost function to select a subset of the GCI candidates that are most likely to correspond to the true ones. The subset of candidates is selected according to the minimisation problem defined as

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \lambda^T \mathbf{c}_{\Omega}(r), \tag{7}$$

where Ω is a subset of GCIs of size $|\Omega|$, λ is a vector of weighting factors and $\mathbf{c}_{\Omega}(r)$ is a vector of cost elements evaluated at the rth GCI of the subset.

3.2 The Enhanced Preprocessor

The derivative of glottal volume velocity, $u'(n)$, is first derived from $s(n)$ as described in Section 2. The stationary wavelet transform (SWT) reinforces discontinuities in a signal by calculating its derivative at multiple dyadic scales. The technique is described in detail in [15] where it is applied to the electroglottogram (EGG) signal for the detection of glottal closure and opening instants. The same biorthogonal spline wavelet with one vanishing moment is used in this paper. Existing and proposed preprocessors are depicted in Figure 3.

The dyadic wavelet transform [17] involves iteratively decomposing a signal into decimated subbands. Let filters $g(n)$ and $h(n)$ have high- and low-pass characteristics respectively. Filterbank trees using wavelets with one vanishing moment detect discontinuities in a signal's smoothed derivative, displaying maxima at the discontinuity across multiple scales [18]. As the signal traverses deeper into the tree of filter banks, the derivative is estimated at increasing levels of smoothing as shown in Fig. 4.

The dyadic wavelet transform is dyadic in both scale and time; however, as we only wish to determine the projection

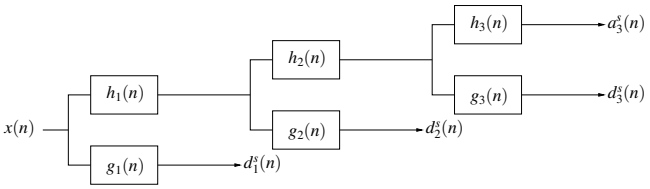

Figure 4: Stationary wavelet transform, for decomposing a signal $x(n)$ into detail and approximation components without decimation.

of $x(n)$ on different subspaces, the filters $g(n)$ and $h(n)$ are instead upsampled by 2 at each iteration to implement the change of scale to form $g_j(n)$ and $h_j(n)$ at scale j. This overcomplete representation of a signal is commonly referred to as the *Stationary Wavelet Transform (SWT)*.

Denote the wavelet $\phi_s(t) = (1/s)\phi(t/s)$, where $s = 2^j, j \in \mathbb{Z}$. The SWT of signal $x(n)$, $1 \leq n \leq N$ at scale j is

$$d_j^s(n) = W_{2^j}x(n) = \sum_k g_j(k)a_{j-1}^s(n-k), \tag{8}$$

where a_{j-1}^s are the approximation coefficients at scale $j-1$. The multiscale product, $p(n)$, is formed by

$$p(n) = \prod_{j=1}^{j_1} d_j(n) = \prod_{j=1}^{j_1} W_{2^j}x(n) \tag{9}$$

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 $x(n)$. The value of j_1 is bounded by J, but it is often no greater than $j_1 = 5$ as the region of support (RoS) of $h_i(n)$ and $g_i(n)$ becomes prohibitively large, demanding high processing resources and smoothing adjacent discontinuties. $j_1 = 3$ is a good compromise [19].

Using the same approach as existing DYPSA, the group delay function, $\tau(n)$, is determined for $p(n)$, whose negative-going zero crossings locate peaks in the multiscale product. Phase slope projection locates missed zero crossings, providing the complete candidate set n_r^{cand}.

3.3 GOI Postprocessing

The high-amplitude GCIs, n_r^c, are extracted from the candidate set by the DYPSA DP. A new candidate set is defined,

$$\{n_r^{ocand}\} = \{n_r^{ccand}\} \triangle \{n_r^c\} \tag{10}$$

where $\triangle$ denotes the symmetric difference (union minus intersection) of the two sets. This candidate set is fed into an identical DP stage to find a set of detected GOIs, n_r^o. A final post-processing stage removes erroneous GOIs and adds missing detections using the GCI periodicity as a reference.

Figure 5 shows a) the glottal volume flow derivative, $u'(n)$, b) the group delay function, $\tau(n)$, and c) the multiscale product, $p(n)$, with overlayed candidates (cyan) and detected GCIs (green), GOIs (red) following the dynamic programming stage. Candidates corresponding to GCIs show negative-going zero crossings with unit negative slope, whereas GOI candidates would not be identified from $\tau(n)$ without phase slope projection. GCIs are straightforward to identify from $p(n)$ by eye but GOIs are less apparent. The algorithm successfully identifies GOIs as they correspond to a subset of candidates with lowest cost; erroneous candidates with high cost, are removed by the dynamic programming.

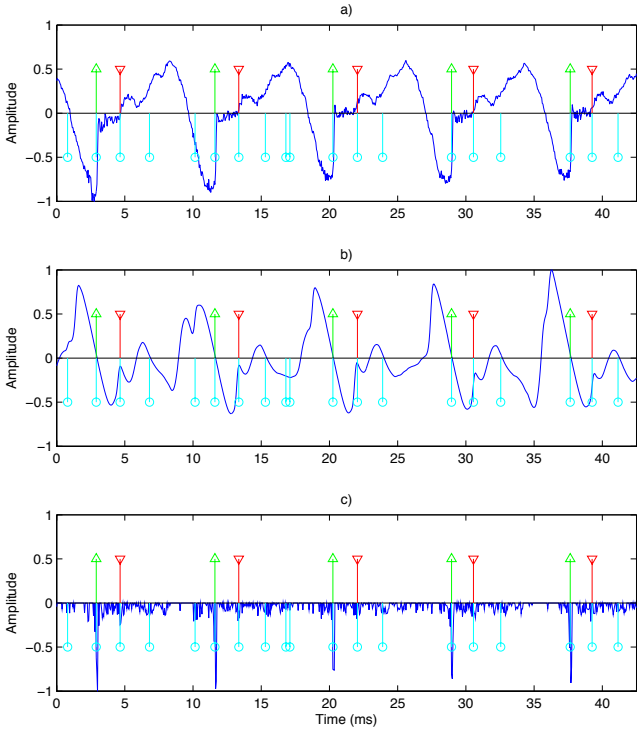

Figure 5: a) Excitation signal, $u'(n)$, b) Group Delay Function, $\tau(n)$ c) Multiscale Product, $p(n)$, with overlayed candidate set (cyan ○) and estimated GCIs (green △) and GOIs (red ▽) following the dynamic programming stage.

4. PERFORMANCE ASSESSMENT

The APLAWD sentence set excluded from the calculation of $\bar{u}'(n)$ was analysed with the proposed algorithm.

An evaluation strategy identical to that defined in [8] was employed, depicted in Figure 6. *Detection rate* is the percentage of all reference GCI periods for which exactly one GCI is estimated. *Accuracy*, σ, and *bias*, μ, are respectively the standard deviation and mean of the error, ζ, between estimated and reference GCIs, when exactly one GCI is estimated in a reference GCI period. *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.

The results are shown in Table 1, showing marginal improvement in GCI detection over the current version of DYPSA. GOI detection shows similarly high identification rates but with around half the identification accuracy.

5. CONCLUSIONS

A novel enhancement to the DYPSA algorithm has been proposed, enabling accurate detection of both glottal closure and opening instants from speech signals. A new preprocessor replaces the linear prediction residual with a signal derived

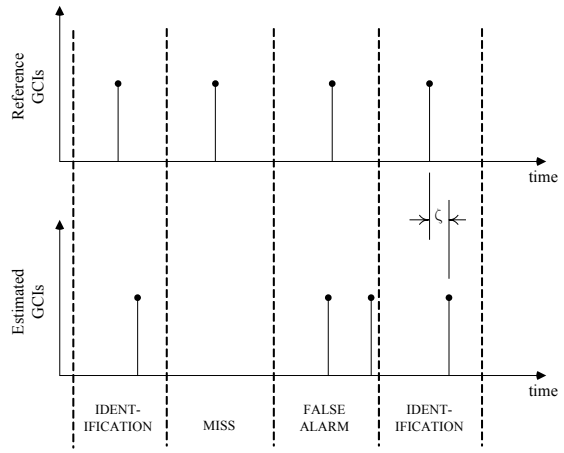

Figure 6: Characterization of GCI Estimates showing four larynx cycles with examples of each possible outcome from GCI estimation. Identication accuracy is measured as the standard deviation, σ, of the error, ζ.

Table 1: Performance comparison current and improved (I) DYPSA algorithms on the APLAWD database.

	ID Rate (%)	Miss Rate (%)	FA Rate (%)	Bias, μ (ms)	ID Accuracy, σ (ms)
DYPSA GCI	96.37	1.73	1.89	0.09	0.68
I. DYPSA GCI	96.41	1.33	2.25	0.08	0.58
I. DYPSA GOI	95.00	1.90	3.09	0.02	1.09

from the multiscale product of an estimate of glottal volume flow derivative. Using DYPSA's existing group delay function to generate a candidate set, a two-stage dynamic programming stage first detects GCIs, then removes them from the candidate set before a second dynamic programming step detects GOIs. A post-processing stage removes erroneous detections and inserts missing GOIs using the GCI periodicity as a reference. A marginal improvement in GCI detection is achieved, with a 96% detection rate and 0.68 ms identification error but, more importantly, GOI detection is achieved with a 95% detection rate and 1.09 ms error.

REFERENCES

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

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

[3] 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, no. 4, pp. 350–355, Aug. 1979.

[4] 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, Sept. 1999.

[5] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 325–333, Sept. 1995.

[6] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 1995, pp. 776–779.

[7] P. S. Murthy and B. Yegananarayana, "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.

[8] 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.

[9] Mike Brookes, Patrick A. Naylor, and Jon Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Speech Audio Process.*, vol. 14, 2006.

[10] B. M. Sadler, T. Pham, and L. C. Sadler, "Optimal and wavelet-based shock wave detection and estimation," *Journal Acoust. Soc. of America*, vol. 104, no. 2, pp. 955–963, Aug. 1998.

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

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

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

[14] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Technical report, University College London, June 1987.

[15] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm for estimation of reference-quality glottal closure instants from electroglottograph signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[16] IEC, "IEC 61672:2003: Electroacoustics – sound level meters," Tech. Rep., IEC, 2003.

[17] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 7, pp. 710–732, 1992.

[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] 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, 1999.