Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review

Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, Thierry Dutoit

***Abstract*—The pseudo-periodicity of voiced speech can be exploited in several speech processing applications. This requires however that the precise locations of the Glottal Closure Instants (GCIs) are available. The focus of this paper is the evaluation of automatic methods for the detection of GCIs directly from the speech waveform. Five state-of-the-art GCI detection algorithms are compared using six different databases with contemporaneous electroglottographic recordings as ground truth, and containing many hours of speech by multiple speakers. The five techniques compared are the Hilbert Envelope-based detection (HE), the Zero Frequency Resonator-based method (ZFR), the Dynamic Programming Phase Slope Algorithm (DYPSA), the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) and the Yet Another GCI Algorithm (YAGA). The efficacy of these methods is first evaluated on clean speech, both in terms of reliabililty and accuracy. Their robustness to additive noise and to reverberation is also assessed. A further contribution of the paper is the evaluation of their performance on a concrete application of speech processing: the causal-anticausal decomposition of speech. It is shown that for clean speech, SEDREAMS and YAGA are the best performing techniques, both in terms of identification rate and accuracy. ZFR and SEDREAMS also show a superior robustness to additive noise and reverberation.**

***Index Terms*—Speech Processing, Speech Analysis, Pitch-synchronous, Glottal Closure Instant**

I. Introduction

GLOTTAL-synchronous speech processing is a field of speech science in which the pseudoperiodicity of voiced speech is exploited. Research into the tracking of pitch contours has proven useful in the field of phonetics [1] and speech quality assessment [2]; however more recent efforts in the detection of Glottal Closure Instants (GCIs) enable the estimation of both pitch contours and, additionally, the boundaries of individual cycles of speech. Such information has been put to practical use in applications including prosodic speech modification [3], speech dereverberation [4], glottal flow estimation [5], speech synthesis [6], [7], data-driven voice source modelling [8] and causal-anticausal deconvolution of speech signals [9].

Increased interest in glottal-synchronous speech processing has brought about a corresponding demand for automatic and reliable detection of GCIs from both clean speech and speech that has been corrupted by acoustic noise sources and/or reverberation. Early approaches that search for maxima in the autocorrelation function of the speech signal [10] were found to be unreliable due to formant frequencies causing multiple maxima. More recent methods search for discontinuities in the linear production model of speech [11] by deconvolving the excitation signal and vocal tract filter with linear predictive coding (LPC) [12]. Preliminary efforts are documented in [5]; more recent algorithms use known features of speech to achieve more reliable detection [13], [14], [15]. Deconvolution of the vocal tract and excitation signal by homomorphic processing [16] has also been used for GCI detection although its efficacy compared with LPC has not been fully researched. Various studies have shown that, while linear model-based approaches can give accurate results on clean speech, reverberation can be particularly detrimental to performance [4], [17].

Methods that use smoothing or measures of energy in speech signal are also common. These include the Hilbert Envelope [18], Frobenius Norm [19], Zero-Frequency Resonator (ZFR) [20] and SEDREAMS [21]. Smoothing of the speech signal is advantageous because the vocal tract resonances, additive noise and reverberation are attenuated while the periodicity of the speech signal is preserved. A disadvantage lies in the ambiguity of the precise time instant of the GCI; for this reason LP residual can be used in addition to smoothed speech to obtain more accurate estimates [14], [21]. Smoothing on multiple dyadic scales is exploited by wavelet decomposition of the speech signal with the Multiscale Product [22] and Lines of Maximum Amplitudes (LOMA) [23] to achieve both accuracy and robustness. The YAGA algorithm [15] employs both multiscale processing and the linear speech model.

The aim of this paper is to provide a review and objective evaluation of five contemporary methods for GCI detection, namely Hilbert Envelope-based method [18], DYPSA [14], ZFR [20], SEDREAMS [21] and YAGA [15] algorithms. These techniques are evaluated against reference GCIs provided by an Electroglottograph (EGG) signal on six databases, of combined duration 232 minutes, containing contemporaneous recordings of EGG and speech. Performance is also evaluated in the presence of additive noise and reverberation. A novel contribution of this paper is the application of the algorithms to causal-anticausal deconvolution [9], which provides additional insight into their performance in a real-world problem.

The remainder of this paper is organised as follows. In Section II the algorithms under test are described. In Section III the evaluation techniques are described. Sections IV and V discuss the performance results on clean and noisy/reverberant speech respectively. Conclusions are given in Section VI.

II. Methods Compared in this Work

This Section presents five of the main representative state-of-the-art methods for automatically detecting GCIs from

speech waveforms. These techniques are detailed here below and their reliability, accuracy and robustness will be compared in Sections IV and V.

A. Hilbert Envelope-based method

Several approaches relying on the Hilbert Envelope (HE) have been proposed in the literature [24], [25], [26]. In this article, a method based on the HE of the Linear Prediction (LP) residual signal (i.e the signal whitened by inverse filtering after removing an auto-regressive modeling of the spectral envelope) is considered.

Figure 1 illustrates the principle of this method for a short segment of voiced speech (Fig.1(a)). The corresponding synchronized derivative of the ElectroGlottoGraph (dEGG) is displayed in Fig.1(e), as it is informative about the actual positions of both GCIs (instants where the dEGG has a large positive value) and GOIs (instants of weaker negative peaks between two successive GCIs). The LP residual signal (shown in Fig.1(b)) contains clear peaks around the GCI locations. Indeed the impulse-like nature of the excitation at GCIs is reflected by discontinuities in this signal. It is also observed that for some larynx cycles (particularly before 170 ms or beyond 280 ms) the LP residual also presents clear discontinuities around GOIs. The resulting HE of the LP residual, containing large positive peaks when the excitation presents discontinuities, and its Center of Gravity (CoG)-based signal are respectively exhibited in Figures 1(c) and 1(d). Denoting $H_e(n)$ the Hilbert envelope of the residue at sample index n, the CoG-based signal is defined as:

$$CoG(n) = \frac{\sum_{m=-N}^{N} m \cdot w(m) H_e(n+m)}{\sum_{m=-N}^{N} w(m) H_e(n+m)} \qquad (1)$$

where $w(m)$ is a windowing function of length $2N+1$. In this work a Blackman window whose length is 1.1 times the mean pitch period of the considered speaker was used. We empirically reported in our experiments that using this window length led to a good compromise between misses and false alarms (i.e to the best reliability performance). Once the CoG-based signal is computed, GCI locations correspond to the instants of negative zero-crossing. The resulting GCI positions obtained for the speech segment are indicated in the top of Fig.1(e). It is clearly noticed that the possible ambiguity with the discontinuities around GOIs is removed by using the CoG-based signal.

B. The DYPSA algorithm

The Dynamic Programming Phase Slope Algorithm (DYPSA) [14] estimates GCIs by the identification of peaks in the linear prediction residual of speech in a similar way to the HE method. It consists of two main components: estimation of GCI candidates with the group delay function of the LP residual and N-best dynamic programming. These components are defined as follows.

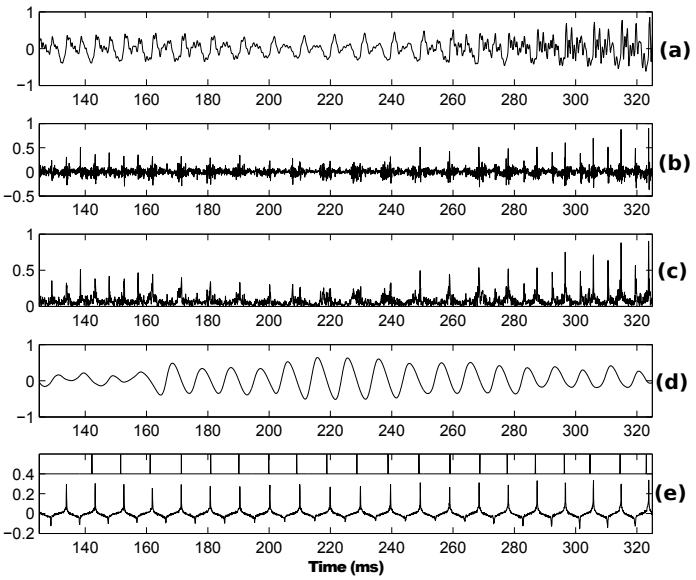

Fig. 1. Illustration of GCI detection using the Hilbert Envelope-based method on a segment of voiced speech. *(a) :* the speech signal, *(b) :* the LP residual signal, *(c) :* the Hilbert Envelope (HE) of the LP residue, *(d) :* the Center of Gravity-based signal computed from the HE, *(e) :* the synchronized differenced EGG with the GCI positions located by the HE-based method.

1) Group Delay Function: The group delay function is the average slope of the unwrapped phase spectrum of the short time Fourier transform of the LP residual [27] [28]. It can be shown to accurately identify impulsive features in a function provided their minimum separation is known. GCI candidates are selected based on the negative-going zero crossings of the group delay function. Consider an LP residual signal, $e(n)$, and an R-sample windowed segment $x_n(r)$ beginning at sample n

$$x_n(r) = w(r)e(n+r) \text{ for } r = 0, \ldots, R-1 \qquad (2)$$

where $w(r)$ is a windowing function. The group delay of $x_n(r)$ is given by [27]

$$\tau_n(k) = \frac{-\mathrm{d}\arg(X_n)}{\mathrm{d}\omega} = \Re\left(\frac{\tilde{X}_n(k)}{X_n(k)}\right) \qquad (3)$$

where $X_n(k)$ is the Fourier transform of $x_n(r)$ and $\tilde{X}_n(k)$ is the Fourier transform of $rx_n(r)$. If $x_n(r) = \delta(r - r_0)$, where $\delta(r)$ is a unit impulse function, it follows from (3) that $\tau_n(k) \equiv r_0 \forall k$. In the presence of noise, $\tau_n(k)$ becomes noisy, therefore an averaging procedure is performed over k. Different approaches are reviewed in [28]. The *Energy-Weighted Group Delay* is defined as

$$d(n) = \frac{\sum_{k=0}^{R-1} |X_n(k)|^2 \tau_n(k)}{\sum_{k=0}^{R-1} |X_n(k)|^2} - \frac{R-1}{2}. \qquad (4)$$

Manipulation yields the simplified expression

$$d(n) = \frac{\sum_{r=0}^{R-1} r x_n^2(r)}{\sum_{r=0}^{R-1} x_n^2(r)} - \frac{R-1}{2} \qquad (5)$$

which is an efficient time-domain formulation and can be viewed as a centre of gravity of $x_n(r)$, bounded in the range

$[-(R-1)/2, (R-1)/2]$. The location of the negative-going zero crossings of $d(n)$ give an accurate estimation of the location of a peak in a function.

It can be shown that the signal $d(n)$ does not always produce a negative-going zero crossing when an impulsive feature occurs in $e(n)$. In such cases, it has been observed that $d(n)$ consistently exhibits local minima followed by local maxima in the vicinity of the impulsive feature [14]. A *phase-slope projection* technique is therefore introduced to estimate the time of the impulsive feature by finding the midpoint between local maxima and minima where no zero crossing is produced, then projecting a line onto the time axis with negative unit slope.

2) Dynamic Programming: Erroneous GCI candidates are removed using known characteristics of voiced speech by minimising a cost function so as to select a subset of the GCI candidates which most likely correspond to true GCIs. The subset of candidates is selected according by minimising the following cost function

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \boldsymbol{\lambda}^T \mathbf{c}_{\Omega}(r), \tag{6}$$

where Ω is a subset with GCI candidates of size $|\Omega|$ selected to produce minimum cost, $\boldsymbol{\lambda} = [\lambda_A \ \lambda_P \ \lambda_J \ \lambda_F \ \lambda_S]^T = [0.8 \ 0.5 \ 0.4 \ 0.3 \ 0.1]^T$ is a vector of weighting factors, the choice of which is described in [14], and $\mathbf{c}(r) = [c_A(r) \ c_P(r) \ c_J(r) \ c_F(r) \ c_S(r)]^T$ is a vector of cost elements evaluated at the rth element of Ω. The cost vector elements are:

- *Speech waveform similarity*, $c_A(r)$, between neighbouring candidates, where candidates not correlated with the previous candidate are penalised.
- *Pitch deviation*, $c_P(r)$, between the current and the previous two candidates, where candidates with large deviation are penalised.
- *Projected candidate cost*, $c_J(r)$, for the candidates from the phase-slope projection, which often arise from erroneous peaks.
- *Normalised energy*, $c_F(r)$, which penalises candidates that do not correspond to high energy in the speech signal.
- *Ideal phase-slope function deviation*, $c_S(r)$, where candidates arising from zero-crossings with gradients close to unity are favoured.

C. The Zero Frequency Resonator-based technique

The Zero Frequency Resonator-based (ZFR) technique relies on the observation that the impulsive nature of the excitation at GCIs is reflected across all frequencies [20]. The GCI positions can be detected by confining the analysis around a single frequency. More precisely, the method focuses the analysis on the output of zero frequency resonators to guarantee that the influence of vocal-tract resonances is minimal and, consequently, that the output of the zero frequency resonators is mainly controlled by the excitation pulses. The zero frequency-filtered signal (denoted $y(n)$ here below) is obtained from the speech waveform $s(n)$ by the following operations [20]:

1) Remove from the speech signal the dc or low-frequency bias during recording:

$$x(n) = s(n) - s(n-1) \tag{7}$$

2) Pass this signal two times through an ideal zero-frequency resonator:

$$y_1(n) = x(n) + 2 \cdot y_1(n-1) + y_1(n-2) \tag{8}$$

$$y_2(n) = y_1(n) + 2 \cdot y_2(n-1) + y_2(n-2) \tag{9}$$

The two passages are necessary for minimizing the influence of the vocal tract resonances in $y_2(n)$.

3) As the resulting signal $y_2(n)$ is exponentially increasing or decreasing after this filtering, its trend is removed by a mean-substraction operation:

$$y(n) = y_2(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2(n+m) \tag{10}$$

where the window length $2N+1$ was reported in [20] to be not very critical, as long as it is in the range of about 1 to 2 times the average pitch period $\bar{T}_{0,mean}$ of the considered speaker. Accordingly, we used in this study a window whose length is $1.5 \cdot \bar{T}_{0,mean}$.

An illustration of the resulting zero frequency-filtered signal is displayed in Fig. 2(b) for our example. This signal is observed to possess two advantageous properties: 1) it oscillates at the local pitch period, 2) the positive zero-crossings of this signal correspond to the GCI positions. This is confirmed in Fig. 2(c), where a good agreement is noticed between the GCI locations identified by the ZFR technique and the actual discontinuities in the synchronized dEGG.

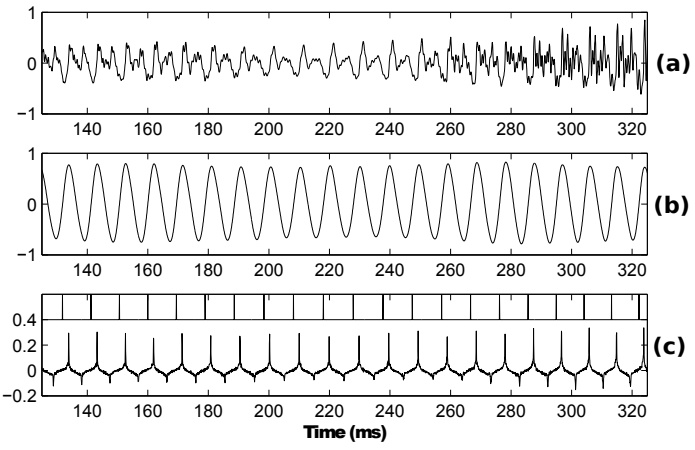

Fig. 2. Illustration of GCI detection using the Zero Frequency Resonator-based method on a segment of voiced speech. *(a) :* the speech signal, *(b) :* the zero frequency-filtered signal, *(c) :* the synchronized dEGG with the GCI positions located by the ZFR-based method.

D. The SEDREAMS algorithm

The Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) algorithm was recently proposed in [21] as a reliable and accurate method for locating both GCIs and GOIs from the speech waveform. Since the present study only focuses on GCIs, the determination of GOI locations by the SEDREAMS algorithm is omitted. The two steps involved in this method are: *i)* the determination of short intervals where GCIs are expected to occur and *ii)* the refinement of the GCI locations within these intervals. These two steps are described in the following subsections.

1) Determining intervals of presence using a mean-based signal: As highlighted by the ZFR technique [20], a discontinuity in the excitation is reflected over the whole spectral band, including the zero frequency. Inspired by this observation, the analysis is focused on a mean-based signal. Denoting the speech waveform as $s(n)$, the mean-based signal $y(n)$ is defined as:

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m) \qquad (11)$$

where $w(m)$ is a windowing function of length $2N+1$. While the choice of the window shape is not critical (a typical Blackman window is used in this study), it has been shown [21] that its length, which influences the time response of this filtering operation, may affect the reliability of the method.

A segment of voiced speech and its corresponding mean-based signal using an appropriate window length are illustrated in Figs. 3(a) and 3(b). Interestingly it is observed that the mean-based signal oscillates at the local pitch period. If the window is too short, it causes the appearance of spurious extrema in the mean-based signal, giving rise to false alarms. On the other hand, too large a window smooths it, leading to some possible misses. It has been observed in [21] that maximal reliability is obtained when the window length is between 1.5 and 2 times the average pitch period $\bar{T}_{0,mean}$ of the considered speaker. Accordingly, throughout the rest of this article a window whose length is $1.75 \cdot \bar{T}_{0,mean}$ is used for computing the mean-based signal of the SEDREAMS algorithm.

However the mean-based signal is not sufficient in itself for accurately locating GCIs. Indeed, consider Fig. 4 where, for five different speakers, the distributions of the actual GCI positions (extracted from synchronized EGG recordings) are displayed within a normalized cycle of the mean-based signal. It turns out that GCIs may occur at a non-constant relative position within the cycle. However, once minima and maxima of the mean-based signal are located, it is straightforward to derive short intervals of presence where GCIs are expected to occur. More precisely, as observed in Fig. 4, these intervals are defined as the timespan starting at the minimum of the mean-based signal, and whose length is 0.35 times the local pitch period (i.e the period between two consecutive minima). Such intervals are illustrated in Fig.3(c) for our example.

2) Refining GCI locations using the residual excitation: Intervals of presence obtained in the previous step give fuzzy short regions where a GCI should happen. The goal of the next step is to refine, for each of these intervals, the precise location of the GCI occuring inside it. The LP residual is therefore inspected, assuming that the largest discontinuity of this signal within a given interval corresponds to the GCI location.

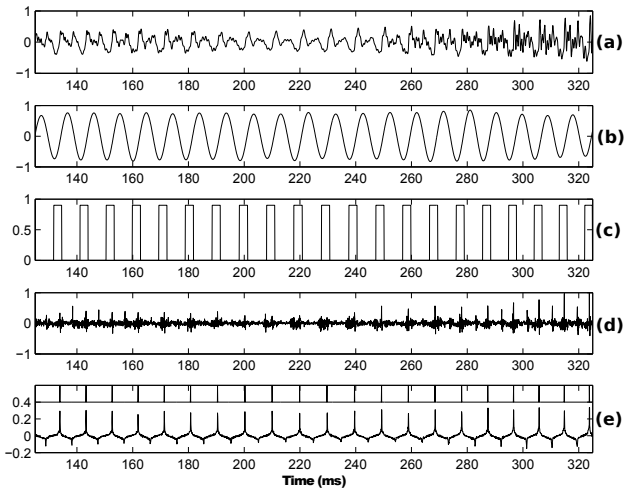

Fig. 3. Illustration of GCI detection using the SEDREAMS algorithm on a segment of voiced speech. *(a) :* the speech signal, *(b) :* the mean-based signal, *(c) :* intervals of presence derived from the mean-based signal, *(d) :* the LP residual signal, *(e) :* the synchronized dEGG with the GCI positions located by the SEDREAMS algorithm.

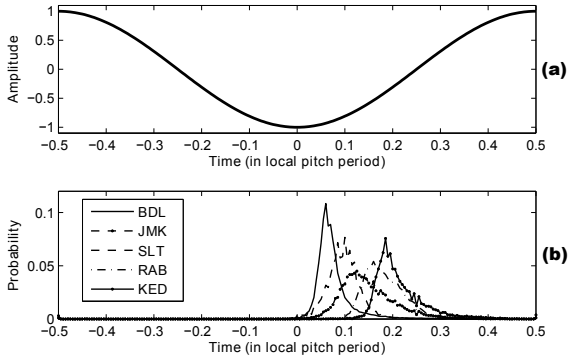

Fig. 4. Distributions, for five speakers, of the actual GCI positions (plot *(b)*) within a normalized cycle of the mean-based signal (plot *(a)*).

Figs. 3(d) and 3(e) show the LP residual and the time-aligned dEGG for our example. It is clearly noted that combining the intervals extracted from the mean-based signal with a peak picking method on the LP residue allows the accurate and unambiguous detection of GCIs (as indicated in Fig.3(e)).

It is worth noting that the advantage of using the mean-based signal is two-fold. First of all, since it oscillates at the local pitch period, this signal guarantees good performance in terms of reliability (i.e the risk of misses or false alarms is limited). Secondly, the intervals of presence that are derived from this signal imply that the GCI timing error is bounded by the depth of these intervals (i.e 0.35 times the local pitch period).

E. The YAGA algorithm

The Yet Another GCI Algorithm (YAGA) [15], like DYPSA, is an LP-based approach that employs N-best dynamic programming to find the best path through candidate GCIs. It differs in that candidates are estimated from an estimation of the voice source signal $u'(n)$, the time-derivative of glottal volume velocity, instead of a whitened LP residual, $e(n)$. The voice source signal is equivalent to the LP residual but without high-frequency preemphasis; in the case of the LP residual GCIs are manifest as impulsive features, whereas the voice source exhibits discontinuities at both GCIs and GOIs. For the purposes of this paper, only GCIs are considered.

Discontinuities are detected in $u'(n)$ by multiscale analysis [29] with 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 $u'(n)$, $1 \leq n \leq N$ at scale j is

$$\begin{aligned} d_j^s(n) &= W_{2^j} u'(n), \\ &= \sum_k g_j(k) a_{j-1}^s(n-k), \end{aligned} \tag{12}$$

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

$$a_j^s(n) = \sum_k h_j(k) a_{j-1}^s(n-k), \tag{13}$$

where $a_0^s(n) = u'(n)$ and $g_j(k)$, $h_j(k)$ are detail and approximation filters respectively that are upsampled by two on each iteration to effect a change of scale [29]. The multiscale product, $p(n)$, is formed by

$$p(n) = \prod_{j=1}^{j_1} d_j(n) = \prod_{j=1}^{j_1} W_{2^j} u'(n), \tag{14}$$

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 $u'(n)$ 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 [30]. The negative-going zero crossings of the group delay function of $p(n)$ are identified to locate these impulses, as described in II-B, forming a candidate set containing both GCIs and GOIs.

The GCIs are estimated from the candidate set by N-best dynamic programming. The set of cost functions is similar to that employed in DYPSA with two significant alterations. Firstly, waveform similarity is calculated the voice source signal $u'(n)$ instead of the speech signal $s(n)$; the absence of vocal tract resonances in this signal results in low similarity for those candidates not separated by one pitch period. Secondly, a measure of energy in the glottal closed phase assists the algorithm in finding only those candidates pertaining to the true GCIs.

III. Assessment of GCI Extraction Techniques

A. Speech Material

The evaluation of the GCI detection methods relies on ground-truth obtained from EGG recordings. The methods are compared on six large corpora containing contemporaneous EGG recordings whose description is summarized in Table I. The first three corpora come from the CMU ARCTIC databases [31]. They were collected at the Language Technologies Institute at Carnegie Mellon University with the goal of developing unit selection speech synthesizers. Each phonetically balanced dataset contains 1150 sentences uttered by a single speaker: BDL (US male), JMK (US male) and SLT (US female). The fourth corpus consists of a set of nonsense words containing all phone-phone transitions for English, uttered by the UK male speaker RAB. The fifth corpus is the KED Timit database and contains 453 utterances spoken by a US male speaker. These five first databases are freely available on the Festvox webpage [31]. The sixth corpus is the APLAWD dataset [32] which contains ten repetitions of five phonetically balanced English sentences spoken by each of five male and five female talkers. For each of these six corpora, the speech and EGG signals sampled at 16 kHz are considered.

Dataset	Number of speakers	Approximative duration
BDL	1	54 min.
JMK	1	55 min.
SLT	1	54 min.
RAB	1	29 min.
KED	1	20 min.
APLAWD	10	20 min.
Total	15	232 min.

TABLE I
DESCRIPTION OF THE DATABASES.

B. Objective Evaluation

The most common way to assess the performance of GCI detection techniques is to compare the estimates with the reference locations extracted from EGG signals (Section III-B1). Besides it is here proposed to evaluate also their efficiency on a specific application of speech processing: the causal-anticausal deconvolution (Section III-B2).

1) Comparison with Electroglottographic Signals: Electroglottography (EGG), also known as electrolaryngography, is a non-intrusive technique for measuring the impedance between the vocal folds. The EGG signal is obtained by passing a weak electrical current between a pair of electrodes placed in contact with the skin on both sides of the larynx. This measure is proportionate to the contact area of the vocal folds. As clearly seen in the explanatory figures of Section II, true positions of GCIs can then be easily detected by locating the greatest positive peaks in the differenced EGG signal. Note that, for the automatic assessment, EGG signals need to be time-aligned with speech signals by compensating the delay between the EGG and the microphone. This was done in this work by a manual verification for each database (inside which the delay is assumed to remain constant).

Performance of a GCI detection method can be evaluated by comparing the locations that are estimated with the synchronized reference positions derived from the EGG recording. For this, we here make use of the performance measure defined in

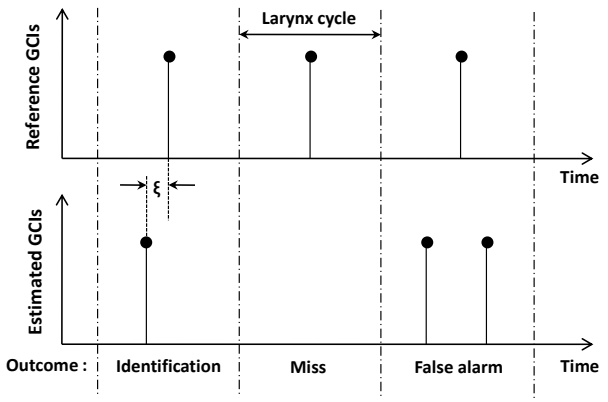

Fig. 5. Characterization of GCI estimates showing three larynx cycles with examples of each possible outcome from GCI estimation [14]. Identification accuracy is characterized by ξ.

[14], presented with the help of Fig. 5. The first three measures describe how *reliable* the algorithm is in identifying GCIs:

- the Identification Rate (IDR): the proportion of larynx cycles for which exactly one GCI is detected,
- the Miss Rate (MR): the proportion of larynx cycles for which no GCI is detected,
- and the False Alarm Rate (FAR): the proportion of larynx cycles for which more than one GCI is detected.

For each correct GCI detection (i.e respecting the IDR criterion), a timing error ξ is made with reference to the EGG-derived GCI position. When analyzing a given dataset with a particular method of GCI detection, ξ has a probability density comparable to the histograms of Fig. 8 (which will be detailed later in this paper). Such a distribution can be characterized by the following measures for quantifying the *accuracy* of the method [14]:

- the Identification Accuracy (IDA): the standard deviation of the distribution,
- the Accuracy to $\pm$ 0.25 ms: the proportion of detections for which the timing error is smaller than this bound.

2) A Speech Processing Application: the Causal-Anticausal Deconvolution: The causal-anticausal decomposition (also known as mixed-phase decomposition) is a non-parametric technique of source-tract deconvolution known to be highly sensitive to GCI location errors [9]. It can therefore be employed as a framework for assessing our methods of GCI extraction on a speech processing application. The principle of this decomposition relies on the mixed-phase model of speech [33], [9]. According to this model, voiced speech is composed of both minimum-phase (i.e causal) and maximum-phase (i.e anticausal) components. While the vocal tract response and the glottal *return phase* can be considered as minimum-phase signals, it has been shown [33] that the glottal *open phase* is a maximum-phase signal. The key idea of the causal-anticausal (or mixed-phase) decomposition is then to separate both minimum and maximum-phase components of speech, where the latter is only due to the glottal contribution. By isolating the anticausal component of speech, causal-anticausal separation allows to estimate the glottal open phase.

Two algorithms have been proposed in the literature for achieving the causal-anticausal separation: the Zeros of the Z-Transform (ZZT, [34]) method and the Complex Cepstrum-based Decomposition (CCD, [35]). It has been shown [35] that both algorithms are functionally equivalent and lead to a reliable estimation of the glottal flow. However the use of the CCD technique was recommended for its much higher computational speed compared to ZZT. Besides it was also shown in [35] that windowing is crucial and dramatically conditions the efficiency of the causal-anticausal decomposition. It is indeed essential that the window applied to the segment of voiced speech respects some constraints in order to exhibit correct mixed-phase properties. Among these constraints, the window should be synchronized on a GCI, and have an appropriate shape and length (proportional to the pitch period). If the windowing is such that the speech segment respects the properties of the mixed-phase model, a correct deconvolution is achieved and the anticausal component gives a reliable estimate of the glottal flow (i.e which corroborates the models of the glottal source, such as the LF model [36]), as illustrated in Fig. 6(a). On the contrary, if this is not the case (possibly due to the fact that the window is not perfectly synchronized with the GCI), the causal-anticausal decomposition fails, and the resulting anticausal component generally contains an irrelevant high-frequency noise (see Fig.6(b)).

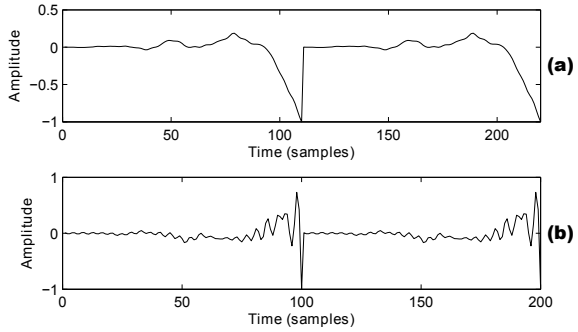

Fig. 6. Two cycles of the anticausal component isolated by mixed-phase decomposition *(*a): when the speech segment exhibits characteristics of the mixed-phase model, *(*b): when this is not the case.

As a simple (but accurate) criterion for deciding whether a frame has been correclty decomposed or not, the spectral center of gravity of the anticausal component is investigated. For a given dataset, this feature has a distribution as the one displayed in Fig. 7. A principal mode around 2 kHz clearly emerges and corresponds to the majority of frames for which a correct decomposition is carried out (as in Fig.6(a)). A second mode at higher frequencies is also observed. It is related to the frames where the causal-anticausal decomposition fails, leading to a maximum-phase signal containing an irrelevant high-frequency noise (as in Fig.6(b)). It can be noticed from this histogram that fixing a threshold at around 2.7 kHz optimally discriminate frames that are correctly and incorrectly decomposed.

In conclusion, it is expected that the use of good GCI estimates reduces the proportion of frames that are incorrectly decomposed using the causal-anticausal separation.

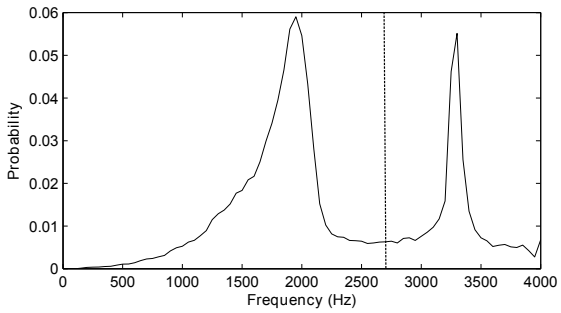

Fig. 7. Example of distribution for the spectral center of gravity of the maximum-phase component. Fixing a threshold around 2.7kHz makes a good separation between correctly and incorrectly decomposed frames.

IV. Experiments on Clean Speech Data

Based on the experimental protocol described in Section III, the performance of the five methods of GCI detection introduced in Section II is here compared on the original clean speech utterances.

A. Comparison with Electroglottographic Signals

Results obtained from the comparison with electroglottographic recordings are presented in Table II for the various databases.

In terms of *reliability* performance, SEDREAMS and YAGA algorithms generally give the highest identification rates. Amongs others, it turns out that SEDREAMS correctly identifies more than 98% of GCIs for any dataset. This is also true for YAGA, except on the RAB database where it reaches 95.70%. Although the performance of ZFR is below these two techniques for JMK, RAB and KED speakers, its results are rather similar on other datasets, obtaining even the best reliability scores on SLT and APLAWD. As for the DYPSA method, its performance remains behind SEDREAMS and YAGA, albeit it reaches IDRs comprised between 95.54% and 98.26%, except for the RAB speaker where the technique fails, leading to an important amount of false alarms (15.80%). Finally the HE-based approach is most of the time outperformed by all other methods. However it achieves on all databases identification rates, comprised between 91.74% and 97.04%.

In terms of *accuracy*, it is observed on all the databases, except for the RAB speaker, that YAGA leads the highest rates of frames for which the timing error is lower than 0.25 ms. The SEDREAMS algorithm gives almost comparable accuracy performance, just below the accuracy of YAGA. The DYPSA and HE algorithms, are outperformed by YAGA and SEDREAMS on all datasets. As it was the case for the reliability results, the accuracy of ZFR strongly depends on the considered speaker. It achieves very good results on the BDL and SLT speakers even though the overall accuracy is rather low especially for the KED corpus.

The accuracy performance is illustrated in Fig. 8 for the five measures. The distributions of the GCI identification error ξ is averaged over all datasets. The histograms for the SEDREAMS and YAGA methods are the sharpest and are highly similar. It is worth pointing out that some discrepancy is expected even if the GCI methods identify the acoustic events with high accuracy, since the delay between the speech signal, recorded by the microphone, and the EGG does not remain constant during recordings.

In conclusion from the results of Table II, the SEDREAMS and YAGA techniques, with highly similar performance, generally outperform other methods of GCI detection on clean speech, both in terms of reliability and accuracy. The ZFR method can also reach comparable (or even slightly better) results for some databases, but its performance is observed to be strongly sensitive to the considered speaker. In general, these three approaches are respectively followed by the DYPSA algorithm and the HE-based method.

B. Performance based on Causal-Anticausal Deconvolution

As introduced in Section III-B2, the Causal-Anticausal deconvolution is a well-suited approach for evaluating our techniques of GCI determination on a concrete application of speech processing. It was indeed emphasized that this method of glottal flow estimation is highly sensitive to GCI location errors. Besides we presented in Section III-B2 an objective spectral criterion for deciding whether the mixed-phase separation fails or not. It is here important to note that the constraint of precise GCI-synchronization is a necessary, but not sufficient, condition for having a correct deconvolution.

Figure 9 displays, for all databases and GCI estimation techniques, the proportion of speech frames that are incorrectly decomposed via mixed-phase separation (achieved in this work by the complex cepstrum-based algorithm [35]). It can be observed that for all datasets (except for SLT), SEDREAMS and YAGA outperform other approaches and lead again to almost the same results. They are closely followed by the DYPSA algorithm whose accuracy was also shown to be quite high in the previous section. The ZFR method turns out to be generally outperformed by these three latter techniques, but still gives the best results on the SLT voice. Finally, it is seen that the HE-based approach leads to the highest rates of incorrectly decomposed frames. Interestingly, these results achieved in the applicative context of the mixed-phase deconvolution corroborate the conclusions drawn from the comparison with EGG signals, especially regarding their accuracy to ± 0.25 ms (see Section IV-A). This means that the choice of an efficient technique of GCI estimation, as those compared in this work, may significantly improve the performance of applications of speech processing for which a pitch-synchronous analysis or synthesis is required.

V. Robustness of GCI Extraction Methods

In some speech processing applications, such as speech synthesis, utterances are recorded in well controlled conditions. For such high-quality speech signals, the performance of GCI estimation techniques was studied in Section IV. For many other types of speech processing systems however, there is no other choice than capturing the speech signal in a *real world environment*, where noise and/or reverberation may dramatically degrade its quality. The goal of this section is to

Database	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to ± 0.25ms (%)
BDL	HE	97.04	1.93	1.03	0.58	46.24
	DYPSA	95.54	2.12	2.34	0.42	83.74
	ZFR	97.97	1.05	**0.98**	0.30	80.93
	SEDREAMS	98.08	0.77	1.15	0.31	89.35
	YAGA	**98.43**	**0.39**	1.18	**0.29**	**90.31**
JMK	HE	93.01	3.94	3.05	0.90	38.66
	DYPSA	98.26	0.88	0.86	0.46	77.26
	ZFR	96.17	3.43	**0.4**	0.60	41.62
	SEDREAMS	**99.29**	**0.25**	0.46	0.42	80.78
	YAGA	99.13	0.27	0.60	**0.40**	**81.05**
SLT	HE	96.16	2.83	1.01	0.56	52.46
	DYPSA	97.18	1.41	1.41	0.44	72.17
	ZFR	**99.26**	0.15	**0.59**	**0.22**	83.70
	SEDREAMS	99.15	**0.12**	0.73	0.30	81.35
	YAGA	98.90	0.20	0.90	0.28	**86.18**
RAB	HE	92.08	2.55	5.37	0.78	38.67
	DYPSA	82.33	1.87	15.80	0.46	86.76
	ZFR	92.94	6.31	0.75	0.56	55.87
	SEDREAMS	**98.87**	**0.63**	**0.50**	**0.37**	**91.26**
	YAGA	95.70	0.47	3.83	0.49	89.77
KED	HE	94.73	1.75	3.52	0.56	65.81
	DYPSA	97.24	1.56	1.20	0.34	89.46
	ZFR	87.36	7.90	4.74	0.63	46.82
	SEDREAMS	**98.65**	0.67	**0.68**	**0.33**	94.65
	YAGA	98.21	**0.63**	1.16	0.34	**95.14**
APLAWD	HE	91.74	5.64	2.62	0.73	54.20
	DYPSA	96.12	2.24	1.64	0.59	77.82
	ZFR	**98.89**	0.59	0.52	0.55	57.87
	SEDREAMS	98.67	0.82	**0.51**	**0.45**	85.15
	YAGA	98.88	**0.52**	0.60	0.49	**85.51**

TABLE II
SUMMARY OF THE PERFORMANCE OF THE FIVE METHODS OF GCI ESTIMATION FOR THE SIX DATABASES.

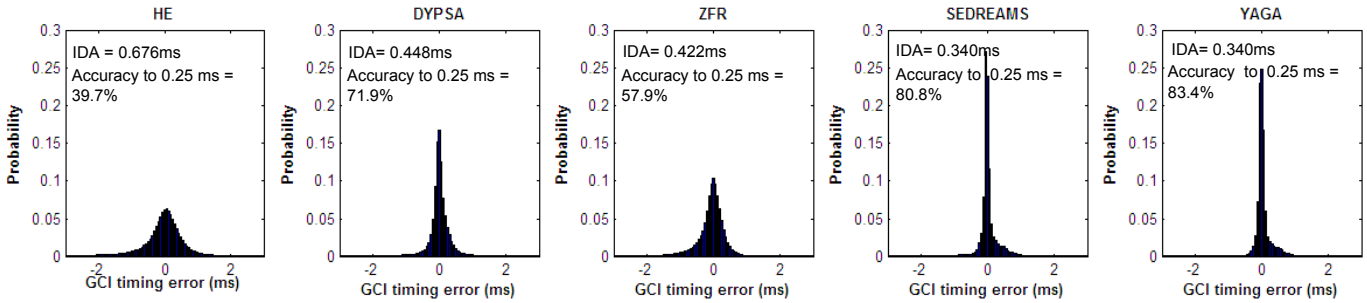

Fig. 8. Histograms of the GCI timing error averaged over all databases for the five compared techniques.

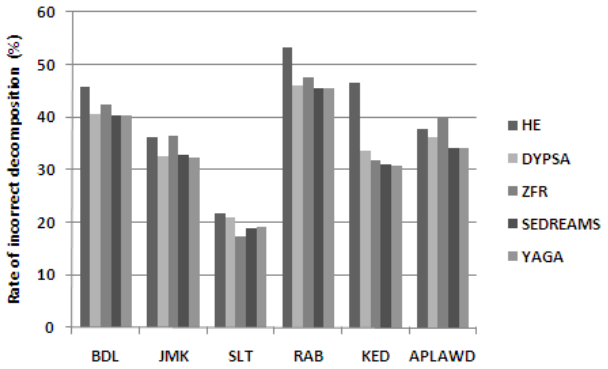

Fig. 9. Proportion of speech frames leading to an incorrect mixed-phase deconvolution using all GCI estimation techniques on all databases.

evaluate how GCI detection methods are affected by additive noise (Section V-A) and by reverberation (Section V-B). Note that results presented here below were averaged over the six databases.

A. Robustness to an Additive Noise

In a first experiment, noise was added to the original speech waveform at various Signal-to-Noise Ratio (SNR). Both a White Gaussian Noise (WGN) and a babble noise (also known as cocktail party noise) were considered. Results for these two noise types are exhibited in Figs. 10 and 11 according to the measures detailed in Section III-B1. It is observed that, for both noise types, the general trends remain unchanged. However it turns out that the degradation of reliability is more severe with the white noise, while the accuracy is more affected by the babble noise.

In terms of reliability, it is noticed that SEDREAMS and ZFR lead to the best robustness, since their performance is al-

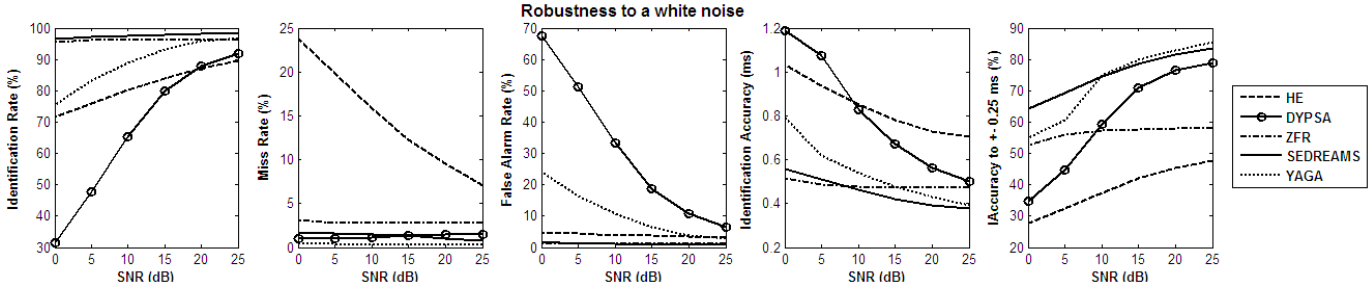

Fig. 10. Robustness of GCI estimation methods to an additive white noise, according to the five measures of performance.

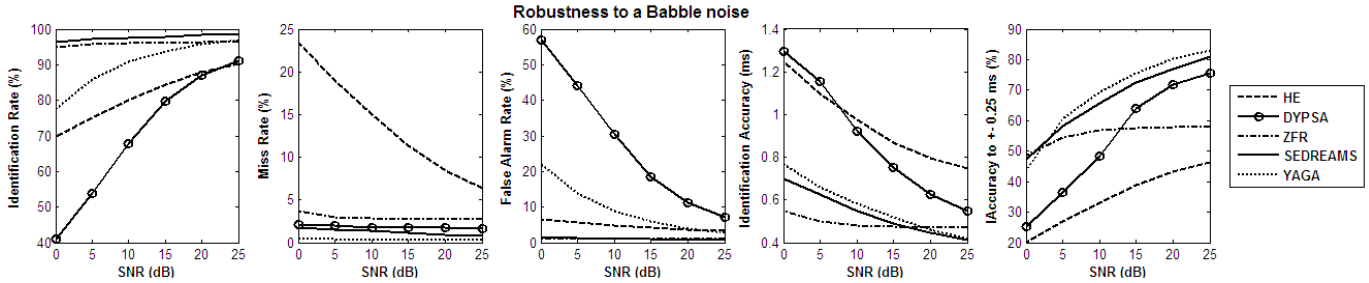

Fig. 11. Robustness of GCI estimation methods to an additive babble noise, according to the five measures of performance.

most unchanged up to 0dB of SNR. Secondly, the degradation for YAGA and HE is almost equivalent, while it is noticed that DYPSA is strongly affected by additive noise. Among others, it is observed that HE is characterized by an increasing missing rate as the noise level increases, while the degradation is reflected by an increasing number of false alarms for DYPSA, and for YAGA in a lesser extent. This latter observation is probably due to the difficulty of the dynamic programing process to deal with spurious GCI candidates caused by the additive noise.

Regarding the accuracy capabilities, the same conclusions almost hold. Nevertheless the sensitivity of SEDREAMS is this time comparable to that of YAGA and HE. Again, the ZFR algorithm is found to be the most robust technique, while DYPSA is the one presenting the strongest degradation and HE displays the worst identification accuracy.

B. Robustness to Reverberation

In many modern telecommunication applications, speech signals are obtained in enclosed spaces with the talker situated at a distance from the microphone. The received speech signal is distorted by reverberation, caused by reflected signals from walls and hard objects, diminishing intelligibility and perceived speech quality [37], [38]. It has been further observed that the performance of GCI identification algorithms is degraded when applied to reverberant signals [4].

The observation of reverberant speech at microphone m is

$$x_m(n) = h_m(n) * s(n), \quad m = 1, 2, \ldots, M, \qquad (15)$$

where $h_m(n)$ is the L-tap Room Impulse Response (RIR) of the acoustic channel between the source to the mth microphone. It has been shown that multiple time-aligned observations with a microphone array can be exploited for GCI estimation in reverberant environments [17]; in this paper we only consider the robustness of single-channel algorithms to the observation at channel $x_1(n)$. RIRs are characterised by the value T_{60}, defined as the time for the amplitude of the RIR to decay to -60dB of its initial value. A room measuring 3x4x5 m and T_{60} ranging $\{100, 200, \ldots, 500\}$ ms was simulated using the source-image method [39] and the simulated impulse responses convolved with the clean speech signals described in Section III.

The results in Figure 12 show that the performance of the algorithms monotonically reduces with increasing reverberation, with the most significant change in performance occurring between $T_{60} = 100$ and 200 ms. They also reveal that reverberation has a particularly detrimental effect upon identification rate of the LP-based approaches, namely HE, DYPSA and YAGA. This is consistent with previous studies which have shown that the RIR results in additional spurious peaks in the LP residual of similar amplitude to the voiced excitation [40], [41], generally increasing false alarm rate for DYPSA and YAGA but increasing miss rate for HE. Although spurious peaks result in increased false alarms, the identification accuracy of the hits is much less affected. The non-LP approaches generally exhibit better identification rates in reverberation, in particular SEDREAMS. The ZFR algorithm appears to be the least sensitive to reverberation while providing the best overall performance. However, the challenge of GCI detection from single-channel reverberant observations remains an ongoing research problem as no single algorithm consistently provides good results for all five measures.

VI. Conclusion

This paper gave a comparative evaluation of five of the most effective methods for automatically determining GCIs from

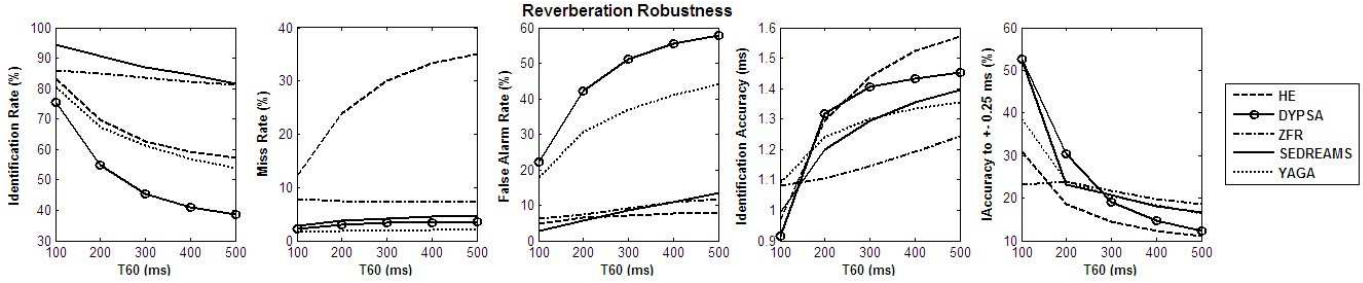

Fig. 12. Robustness of GCI estimation methods to reverberation, according to the five measures of performance.

the speech waveform: Hilbert Envelope-based detection (HE), the Zero Frequency Resonator-based method (ZFR), DYPSA, SEDREAMS and YAGA. The performance of these methods was assessed on six databases containing several male and female speakers, for a total amount of data of approximately four hours. In our first experiments on clean speech, the SEDREAMS and YAGA algorithms gave the best results, with a comparable performance. For *any* database, they reached an identification rate greater than 98% and more than 80% of GCIs were located with an accuracy of 0.25 ms. Although the ZFR technique can lead to a similar performance, its efficiency can also be rather low in some cases. In general, these three approaches were shown to respectively outperform DYPSA and HE. In a second experiment on clean speech, the impact of the performance of these five methods was studied on a concrete application of speech processing: the causal-anticausal deconvolution. Results showed that adopting a GCI detection with high performance could significantly improve the proportion of correctly deconvolved frames. In the last experiment, the robustness of the five techniques to additive noise, as well as to reverberation was investigated. The ZFR and SEDREAMS algorithms were shown to have the highest robustness, with an almost unchanged reliability. DYPSA was observed to be especially affected, which was reflected by a high rate of false alarms. Although the degradation of accuracy was relatively slow with the level of additive noise, it was noticed that reverberation dramatically affects the precision GCI detection methods.

ACKNOWLEDGMENT

Thomas Drugman is supported by the Belgian Fonds National de la Recherche Scientifique (FNRS).

REFERENCES

[1] J. C. Catford, *Fundamental Problems in Phonetics*. Indiana University Press, 1977.

[2] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

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

[4] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Cardiff, UK, 2007.

[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. 27, no. 4, pp. 350–355, Aug. 1979.

[6] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 21–29, 2001.

[7] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech Conference*, 2009.

[8] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice source waveform modelling," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[9] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *ISCA ITRW VOQUAL03*, 2003, pp. 21–24.

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

[11] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, 1988.

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

[13] 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.

[14] 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, 2007.

[15] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Detection of glottal opening and closing instants in voiced speech using the YAGA algorithm," *Submitted for peer review*, 2010.

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

[17] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Multichannel DYPSA for estimation of glottal closure instants in reverberant speech," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2007.

[18] 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, pp. 309–319, 1979.

[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] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, 2008.

[21] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech Conference*, 2009.

[22] A. Bouzid and N. Ellouze, "Glottal opening instant detection from speech signal," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Sep. 2004, pp. 729–732.

[23] V. N. Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Eurospeech*, Budapest, Sep. 1999, pp. 2805–2808.

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

[25] Y. M. 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.

[26] 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, 2007.
[27] 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)*, 1995, pp. 776–779.
[28] 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, 2006.
[29] 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.
[30] 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.
[31] [Online], "The festvox website," in http://festvox.org/.
[32] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Technical Report, 1987.
[33] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *ISCA ITRW VOQUAL03*, 2003, pp. 15–19.
[34] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Processing Letters*, vol. 12, 2005.
[35] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. Interspeech Conference*, 2009.
[36] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.
[37] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, vol. 21, pp. 577–580, 1949.
[38] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Frances, 2000.
[39] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
[40] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, 2001.
[41] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.