Circuits, Systems, and Signal Processing - Springer manuscript No.
(will be inserted by the editor)

Detection of the Glottal Closure Instants using Empirical Mode Decomposition

Rajib Sharma · S.R.M. Prasanna · Hugo Leonardo Rufiner · Gastón Schlotthauer

01 August 2017

Abstract This work explores the effectiveness of the *Intrinsic Mode Functions* (IMFs) of the speech signal, in estimating its *Glottal Closure Instants* (GCIs). The IMFs of the speech signal, which are its AM-FM or *oscillatory* components, are obtained from two similar non-linear and non-stationary signal analysis techniques - *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (ICEEMDAN), and *Modified Empirical Mode Decomposition* (MEMD). Both these techniques are advanced variants of the original technique - *Empirical Mode Decomposition* (EMD). MEMD is much faster than ICEEMDAN, whereas the latter curtails *mode-mixing* (a drawback of EMD) more effectively. It is observed that the partial summation of a certain subset of the IMFs results in a signal whose minima are aligned with the GCIs. Based on this observation, two different methods are devised for estimating the GCIs from the IMFs of ICEEMDAN and MEMD. The two methods are captioned *ICEEMDAN based GCIs Estimation* (IGE) and *MEMD based GCIs Estimation* (MGE). The results reveal that IGE and MGE provide consistent and reliable estimates of the GCIs, compared to the state-of-the-art methods, across different scenarios - clean, noisy, and telephone channel conditions.

Rajib Sharma (E-mail: s.rajib@iitg.ernet.in) · S.R.M. Prasanna (E-mail: prasanna@iitg.ernet.in)
Signal Informatics Laboratory, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, India.

Hugo Leonardo Rufiner (E-mail: lrufiner@sinc.unl.edu.ar)
Research Institute for Signals, Systems and Computational Intelligence – sinc(i), Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Santa Fe (3000), Argentina.

Gastón Schlotthauer (E-mail: gschlotthauer@conicet.gov.ar)
Laboratorio de Señales y Dinámicas no Lineales, CITER - CONICET, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde (3101), Entre Ríos, Argentina.

Keywords Glottal Closure Instants (GCIs) · Empirical Mode Decomposition (EMD) · Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN) · Modified Empirical Mode Decomposition (MEMD) · Intrinsic Mode Functions (IMFs)

1 Introduction

The *Glottal Closure Instants* (GCIs), or the *Epochs*, are the instants at which the speech production apparatus is *significantly excited* by an impulse-train like signal, generated by the quasi-periodic and abrupt closure of the vocal folds in the glottis, during the production of *voiced* speech [5, 50, 51, 71]. Analysis of the speech signal around its GCIs saves processing time and leads to more accurate analysis [71]. Accurate estimation of the GCIs helps in characterizing voice-quality features, and enables extraction of important acoustic-phonetic features such as glottal vibrations and formants [71]. As such, the information regarding the GCIs is utilized in many practical applications like prosodic speech modification [45], speech dereverberation [22], glottal flow estimation [67], speech synthesis [18, 60], data-driven voice source modeling [62] and causal-anticausal deconvolution of speech signals [7]. Henceforth, the detection of the GCIs has assumed considerable importance in Speech Processing [4, 15–17, 32, 46, 48, 49, 58, 59, 63].

Based on the popular *source-filter* theory of speech production [5, 50, 51], *Linear Prediction* (LP) analysis has been extensively used in the task of detecting the GCIs [4, 16, 48, 49, 58, 59, 63]. The ideal expected output of LP analysis is an LP filter with accurate estimation of the vocal tract resonances and the spectral slope of voiced speech, and an LP residual or error signal, which resembles a train of impulses separated by the time-varying pitch period of the speech signal. However, in practicality, the LP residual turns out to be a noisy signal, with relatively larger amplitudes in the vicinity of the GCIs. The noisy characteristics of the LP residual may be attributed to three main factors [4] :

(i) The inaccurate estimation of the coefficients of the poles, corresponding to the resonances of the vocal tract system, which makes the LP residual to have non-zero values at time-instants other than the GCIs.

(ii) The inaccurate estimation of the phase angles of the formants, which results in large bipolar swings in the LP residual, around the GCIs.

(iii) The presence of strong anti-resonances in the speech production system, which causes the large amplitudes in the LP residual to occur at time-instants other than the GCIs.

These differences between the ideal excitation signal and the LP residual reflect the mismatch between the true characteristics of the speech production system, and that modeled by the source-filter theory using LP analysis [5, 26, 27]. Nevertheless, due to lack of better alternatives, LP analysis,

and the LP residual, in particular, has remained the cornerstone in a multitude of efforts for estimating the GCIs. One of the first methods to utilize the LP residual in estimating the GCIs was the *Epoch Filter* (**EF**) [4]. In this method, the LP residual spectrum is whitened by multiplying with a hanning window. The large swings around the GCIs, in the whitened LP residual, are further reduced by taking its Hilbert envelope. Following this method, algorithms like the *Group Delay* (**GD**) [58], *Hilbert Envelope and Group Delay* (**HEGD**) [59], *Dynamic Programing and Phase Slope Algorithm* (**DYPSA**) [48], *Speech Event Detection using the Residual Excitation And a Mean-based Signal* (**SEDREAMS**) [16], *Yet Another GCI detection Algorithm* (**YAGA**) [63], and *Integrated Linear Prediction Residual using Plosion Index* (**ILPR-PI**) [49], all of which depend on LP analysis, have been developed.

The GD method evaluates the average slope of the phase spectrum, called the *phase slope function*, at each time-instant of the LP residual, to construct a *sinusoid-like* waveform, the positive-going zero-crossings of which coincides with the GCIs. The HEGD and the DYPSA methods try to refine the GD method by minimizing its computational cost, and the number of spurious GCIs estimated from it, respectively. The YAGA is same as the DYPSA algorithm but combines Wavelet Transform and LP analysis to derive a cleaner signal than the LP residual as a representation of the excitation source. The SEDREAMS algorithm uses a moving average filter to obtain a *sinusoid-like* waveform from the speech signal, where prospective regions of the GCIs are defined. Within these regions, the LP residual is used to estimate the GCIs. The ILPR-PI combines LP analysis with a non-linear operator, called the Dynamic Plosion Index, for estimating the GCIs. The *Zero Frequency Resonator* (**ZFR**) [46] is probably the first significant deviation from using LP analysis for estimating the GCIs. By using a cascade of two marginally stable 0-Hz resonators, on the pre-emphasized speech signal, the ZFR obtains an exponentially increasing/decreasing output. When the trend of such an output is removed in short segments of one to two pitch periods, the de-trended signal resembles a *sinusoidal* signal, whose positive-going zero-crossings coincide with the GCIs.

The above discussion reflects how useful a *sinusoidal representation of the source signal* is for the purpose of detecting the GCIs. In many of the popular methods, as discussed above, a sinusoid-like waveform is used directly or indirectly for estimating the GCIs. As most of these methods rely on LP analysis, they are prone to its limitations. The processing of the speech signal under the *arguable assumptions* [5, 26, 27] of short-time stationarity and linearity simplifies the analysis, but also induces errors. Such errors become significant in the case of speech signals subjected to noise, and hence the performances of the LP-analysis based techniques diminishes noticeably [49]. As an example, the *Identification Rate* (IR) of DYPSA is reported to be around 95 % for speech signals recorded in a clean noise-free environment, which falls to as low as 30 % when the same speech signals are corrupted by noise [49]. Similarly, when the speech signals are subjected to telephone channel conditions, the performances of the state-of-the-art methods are much worse. For example, the IR of

ZFR is reported to be $\gtrsim$ 95 % even when the speech signals are corrupted by noise, but drops to as low as 30 % for telephone-quality speech [49]. In other words, while most of the popular techniques perform excellently under clean and controlled data conditions, when the speech signal is subjected to external influences they are not as effective [17, 49]. There are significant variations in the performances of the techniques from one scenario to another [17, 49], as we will observe in Section 6 of this work. To summarize :

- The state-of-the-art methods for detecting the GCIs are much less effective for speech signals which are not recorded in a clean and controlled laboratory set-up. They are limited in adapting to changing conditions.
- Most of the popular methods utilize a sinusoid-like signal for detecting the GCIs.
- Most of the popular methods rely on LP analysis, which forces the assumptions of short-time stationarity and linearity.

The above three observations shape the objective of this work - *"to use some non-linear and non-stationary signal analysis method, to obtain sinusoid-like waveforms directly from the speech signal, and utilize them to develop a technique for detecting the GCIs which is robust to different scenarios"*. The objective also signifies the novelty of this work.

With the aforementined objective in mind, we investigate, in this work, the effectiveness of the various AM-FM or oscillatory components of the speech signal (utterance), as obtained from *Empirical Mode Decomposition* (**EMD**) [30,31,57], for the purpose of estimating the GCIs of its voiced regions. In particular, in this work, we study the utility of the AM-FM components, called *Intrinsic Mode Functions* (IMFs), obtained from two variants of EMD, which effectively curtail a drawback of EMD called *mode-mixing* [19,29–31,52,57,70]. The first variant, called the *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (**ICEEMDAN**) [13], is very effective in reducing *mode-mixing*, but at a large time-cost. The other variant, the *Modified Empirical Mode Decomposition* (**MEMD**) [55], reduces *mode-mixing* less effectively but provides a great time-advantage. As such, this work proposes two methods, but based on the same principle, which utilize the IMFs obtained from ICEEMDAN and MEMD respectively, for estimating the GCIs of the speech signal. The important contributions of this work may be summarized as :

- Experimentally demonstrate, using synthetic speech, the ability of a subset of IMFs in capturing epochal information. Experimentally verify the same on natural speech.
- Propose a framework/principle for estimating the GCIs of the speech signal from its IMFs.
- Propose two methods - one for the IMFs obtained from ICEEMDAN, and the other for the IMFs obtained from MEMD - for estimating the GCIs based on the proposed principle. The two methods are similar, though not the same.

- Evaluate the performances of the two proposed methods on two different databases - the CMU-Arctic [43], and the APLAWDW [3] ; and under different scenarios - clean, flat-band (White) noise, high-frequency (HFchannel) noise, low-frequency (Babble) noise, band-pass filtered telephone channel conditions, and ITU-standard telephone channel conditions. Show that the proposed methods are more consistent than the state-of-the-art methods across the databases and various scenarios.

The rest of the paper is organized as follows : Section 2 discusses the EMD, ICEEMDAN, and MEMD algorithms used for decomposing the speech signal. Section 3 discusses the principle used for detecting the GCIs, using the IMFs of ICEEMDAN and MEMD. Sections 4 and 5 describe the methods used for detecting the GCIs using the IMFs of ICEEMDAN and MEMD respectively. Section 6 presents the results and compares them to the standard algorithms. Section 7 concludes this work.

2 EMD vs. ICEEMDAN vs. MEMD

EMD is a non-linear and non-stationary data analysis technique, which is able to extract *meaningful* components of any time-series dataset, called its *Intrinsic Mode Functions* (IMFs), in the time-domain itself [31]. Such a decomposition has been of precious value to the varied fields of signal processing [6, 9–12, 23–25, 28, 30, 33–41, 47, 53, 54, 68]. The EMD algorithm is based on a core process, called *sifting*, which progressively breaks down the signal into its IMFs. The *sifting process* employs *cubic spline interpolation* iteratively on a signal, to segregate it into a high-frequency component (IMF), and a low-frequency component (*outer residue*). At first, the signal is broken down into its first IMF, and its first outer residue. The first outer residue is again segregated by applying the sifting process. This results in the second IMF and the second outer residue. The second outer residue is then segregated, and so on and so forth. Finally, an outer residue is obtained which cannot be decomposed further. This is called the *final residue*, and represents the trend of the signal. The sifting process is completely self sufficient in decomposing the signal, and does not utilize any *a priori* basis, or pre-knowledge about the characteristics of the signal. This makes EMD a truly data adaptive signal decomposition method. The IMFs generated resemble AM-FM or oscillatory signals, which have been observed to represent *useful latent information* embedded in the signal [30,31]. Theoretically, the IMFs are supposed to be locally symmetric. To generate an IMF from an outer residue, the sifting process employs multiple *sifting iterations* on it. There is no standard rule to determine the appropriate number of sifting iterations, and a number of *sifting criteria* have been proposed in the literature [57].

While the advantages of EMD has been well appreciated, it exhibits two phenomena which inhibit its true capabilities. These two phenomena are known as *end-effects* and *mode-mixing* [30, 31, 57]. For speech signals, which have silence-regions at its terminating ends, the end-effects are not significant [30,

31, 57]. However, mode-mixing, which is the presence of oscillations of disparate frequencies in the same IMF, or conversely, the presence of the same frequency oscillation in multiple IMFs, is a bottleneck. While many efforts have been made to reduce mode-mixing, the most promising results came through the infusion of noise to the signal. This development was termed *Ensemble Empirical Mode Decomposition* (EEMD) [70]. In the recent years, efforts to efficiently cancel out the noise infused into the signal has led to a number of EEMD variants [13,64,72]. It was observed that using white noise in pairs of opposite polarities substantially reduces the effect of the artificially added noise. This development was termed *Complementary Ensemble Empirical Mode Decomposition* (CEEMD) [72]. However, the problem that the number of IMFs produced could still be different for the different EMD processes of an EEMD or CEEMD decomposition still existed. To circumvent this problem, an algorithm was designed, which not only decomposes the signal but also the white noise realizations parallely with it. The IMFs obtained from the white noise realizations, which could be interpreted as correlated noise signals, are then fused with the outer residue signal, at the beginning of each sifting process. The signal IMFs are obtained progressively after averaging the results at each stage. This algorithm was termed *Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (CEEMDAN) [64]. However, it was observed that CEEMDAN sometimes produced some high-frequency spurious IMFs. To overcome this problem, the *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (ICEEMDAN) [13] was developed, which makes some subtle but effective modifications to the CEEMDAN algorithm.

The biggest drawback of EEMD and its variants is their time-efficiency - the algorithms need to process multiple copies of the signal independently, to finally extract its IMFs. Henceforth, with the aim of extracting IMFs with reduced *mode-mixing*, but with significantly lower time-cost, the *Modified Empirical Mode Decomposition* (MEMD) [55] has been recently proposed. The MEMD algorithm (denoted as M2-EMD (D2) in [55]) makes subtle changes to the *sifting* process, resulting in reduced mode-mixing.

Thus, given a digital speech signal, $s(n)$, EMD/ICEEMDAN/MEMD decomposes it into a small number of IMFs, given by,

$$s(n) = r_M(n) + \sum_{k=1}^{M} h_k(n) = \sum_{k=1}^{M+1} h_k(n) \tag{1}$$

In the above equation, $h_k(n)$ represents the k^{th} IMF of $s(n)$, and M the total number of IMFs generated. The signal $r_M(n) = h_{M+1}(n)$ represents the *final residue* or trend of the signal. Ideally, the decomposition is to be stopped when the *outer residue* takes the form of a trend, i.e., the number of extrema in $r_M(n)$ is two or less [29–31, 57]. Practically, however, the decomposition is stopped when a user-defined maximum number of IMFs, M, has been extracted. This is done in order to avoid unnecessary generation of low-frequency trend-like IMFs [57].

It has been observed with experiments on fractional Gaussian noise (fGn) that EMD (and hence ICEEMDAN and MEMD) behaves as a data dependent *dyadic filterbank* [20,21,30,69]. And this characteristic is well adhered to when the *number of sifting iterations*, N, is kept around ten [66, 70]. Henceforth, in this work, $N = 10$, is used in every *sifting process* in the EMD/MEMD algorithm. For the ICEEMDAN algorithm, however, N is determined by the *local-global stopping criterion*. Nevertheless, the maximum number of iterations per sifting process is not allowed to exceed 15, i.e., $N \leq 15$. Again, the initial ratio of the standard deviation of white noise to that of the signal is taken as $\epsilon_0 = 0.2$ [1, 2, 13, 52, 57]. Further, in this work, $L = 20$ white noise realizations are used for the ICEEMDAN algorithm.

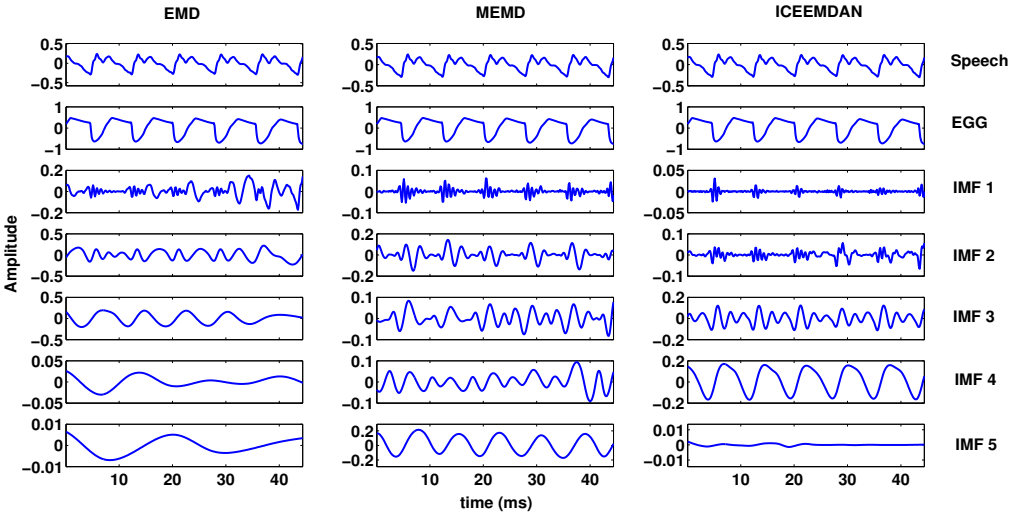

Fig. 1: **The first five IMFs, obtained from a speech signal, using EMD (left column), MEMD (middle column), and ICEEMDAN (right column). The second plot in each column shows the EGG signal corresponding to the speech signal.**

All experiments in this work are conducted on digital speech signals having *sampling frequency* (F_s) of 8 kHz. Assuming that the three algorithms exhibit an ideal filterbank nature, for an 8 kHz speech signal, the primary frequency reflected in IMF_k will have an upper-bound of $4000/(2^{k-1})$ Hz. Thus, beyond IMF_6, the IMFs represent very low-frequency trend-like signals, which are below the lower-limit of human pitch frequency. Therefore, we restrict the decomposition to ten components only ($M = 9$) for all the three algorithms, which will take care of any strong deviation from the ideal nature. In this work, for simplicity, we refer to the final residue as the last (tenth) IMF.

Thus, equation (1) reduces to,

$$s(n) = r_9(n) + \sum_{k=1}^{9} h_k(n) = \sum_{k=1}^{10} h_k(n) \tag{2}$$

Table 1: **Computational complexity of EMD, MEMD and ICEEMDAN. Time (in seconds) taken in decomposing a speech signal, of ∼4 s duration, into ten components, by EMD, MEMD, and ICEEMDAN.**

Method	EMD	MEMD	ICEEMDAN
Time (s)	0.82	0.90	62.76

Figure 1 shows the first five IMFs of a short segment of a speech signal taken from the CMU-Arctic database [43] after decomposition by EMD, MEMD, and ICEEMDAN. The presence of *mode-mixing* is easily observed in the case of EMD, particularly in IMF_1. MEMD exhibits substantially lesser *mode-mixing* in its IMFs, but the best decomposition is provided by ICEEMDAN. One may also observe the periodic nature of the glottal source, represented by IMF_3 in the case of EMD; IMF_5 in the case of MEMD; and IMF_4 in the case of ICEEMDAN [56]. This is reflected in the similarity of these IMFs with the *ElectroGlottoGraph* (EGG) signal, which is a measure of the periodic vibrations of the vocal folds, during the production of the speech signal. Figure 1 clearly demonstrates the superiority of MEMD and ICEEMDAN in representing the glottal source. Table 1 shows the time taken by each of the three methods in decomposing the speech signal. The algorithms are run in MATLAB, in desktop mode, in a machine with 8 GB RAM, using Intel quad-core i7 processor, of 2.9 GHz clock frequency. As the table indicates, MEMD provides the best trade-off between reducing *mode-mixing* and simultaneously maintaining a low time-cost. As such, in this work, only ICEEMDAN and MEMD are utilized in the experiments for extracting the GCIs of the speech signal.

3 Principle of estimating the GCIs

Let us consider a synthetic voiced speech signal, $s(n) = e_g(n) * h(n)$, constructed using the source-filter theory, at a sampling frequency of $F_s = 8$ kHz. The input excitation, $e_g(n)$, is modelled as a uniform train of impulses of unit strength, having *pitch* or *fundamental frequency* (F_0) equal to 100 Hz.

$$e_g(n) = -\sum_{m \in \mathbb{Z}} \delta(n - m\frac{F_s}{F_0}) \tag{3}$$

The impulse response of the vocal tract filter, $h(n)$, is modelled as a cascade of four resonators, simulating the four principal formants. The resonant peaks are considered at 700, 1200, 2400 and 3600 Hz, their bandwiths being 70, 140, 210

and 280 Hz respectively. The low-pass filter nature of the glottis, and the high-pass filter nature of the lips, are also included in $h(n)$. Figure 2(a) shows the speech signal so constructed, and Figure 2(b)-(h) the first seven IMFs derived from it using ICEEMDAN. IMFs 8-10 are not shown in the figure. It is evident from the figure that as the order of the IMF increases the frequency of the dominant oscillation manifested in it decreases. The mean frequency (F^m) of each IMF, which is a *measure of the dominant frequency reflected in the IMF*, is mentioned above the plot of the IMF. The mean frequency of IMF_k, denoted as F_k^m, is calculated from its power spectrum (squared magnitude spectrum), $S_k(f)$, as,

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)} \ , \ k = 1, 2, ..., 7 \tag{4}$$

In the above equation, f represents the analog frequencies corresponding to the discrete frequencies of the *Discrete Fourier Transform* (DFT) spectrum, from which the power spectrum is evaluated.

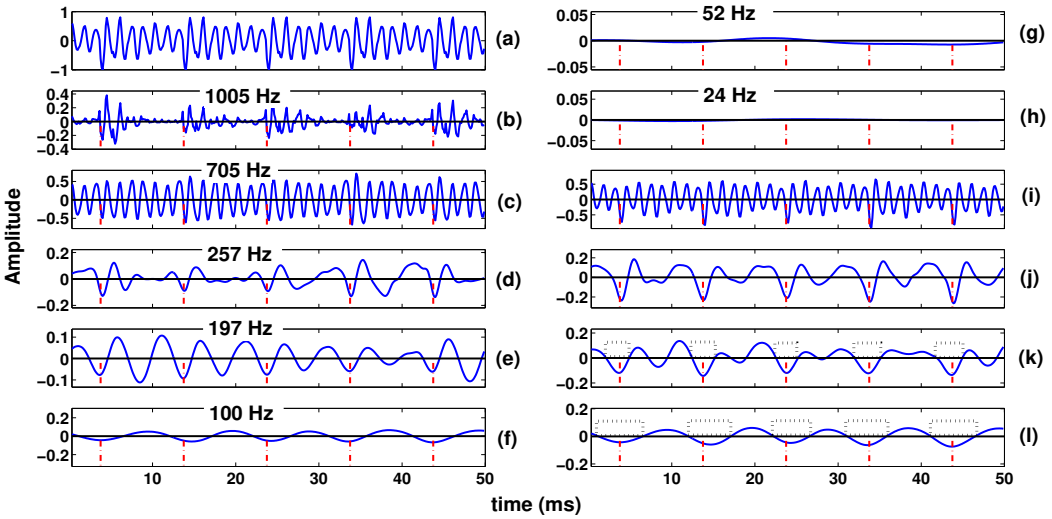

Fig. 2: **(a) A segment of a synthesized vowel-like speech signal $s(n)$ having $F_0 = 100$ Hz ; (b)-(h) are IMFs 1-7 respectively obtained from ICEEMDAN of $s(n)$. IMFs 8-10 are not shown ; (i) sum of IMFs 2-6 ; (j) sum of IMFs 3-6 ; (k) sum of IMFs 4-6 ; (l) sum of IMFs 5-6 ; Dashed vertical lines indicate the impulse excitation, $e_g(n)$. Dotted rectangles in (k) and (l) indicate regions of search for the GCIs.**

It can be seen from Figure 2 that as the F^m of the IMF decreases, it becomes more sinusoidal. The vertical dashed lines in Figure 2(b)-(l) indicate the input impulse-excitation, $e_g(n)$. It can be observed that the IMFs intersect

$e_g(n)$ at different points of their near-sinusoidal curves, i.e, the IMFs have a different phase-shift with respect to the impulse locations. Out of these seven IMFs, the higher-frequency IMFs, IMF_1 and IMF_2, seem to be aligned with the impulse excitation instants at some of their minima locations. However, there are other minima in the vicinity which could cause ambiguity. Similarly, IMF_4 also intersects $e_g(n)$ at some of its minima locations. However, it has stronger adjacent minima, which, as we will see later, might make the selection of the target minima difficult. IMF_3, Figure 2(d), as such, provides the best trade-off. IMF_3 intersects $e_g(n)$ at its prominent minima, with good accuracy, and with much lesser ambiguity with adjacent minima, compared to IMFs 1, 2 and 4. As such, IMF_3 may be described as an *oscillatory* or an *imperfect sinusoidal* signal, some of whose minima-locations coincide with the impulse locations of the input excitation. Thus, given an impulse location, $\{l_m^{ref} = m\frac{F_s}{F_0},\ m \in \mathbb{Z}\}$, IMF_3 may be represented as,

$$h_3(n) \approx a_3(n)\cos[2\pi\frac{f_3}{F_s}(n - l_m^{ref}) + \pi]\ , \tag{5}$$

where $a_3(n)$ is the time-varying amplitude envelope of $h_3(n)$, and f_3 the average frequency of the sinusoid over time. It is interesting to note that IMF_5, Figure 2(f), represents a near sinusoid with $F_5^m = F_0 = 100$ Hz, which shows the ability of ICEEMDAN to characterize glottal activity [56].

Figure 2(i)-(l) represent the sum of the IMFs $2-6$, $3-6$, $4-6$ and $5-6$, respectively. From a different perspective, Figure 2(i)-(l) also represent the sum of IMFs having different frequency content. Thus, Figure 2(l) represents the sum of the IMFs having mean frequency 50 Hz $\leq F^m \leq F_0 = 100$ Hz. Figure 2(k) represents the same for 50 Hz $\leq F^m \leq 2 \times F_0 = 200$ Hz, Figure 2(j) for 50 Hz $\leq F^m \leq 5 \times F_0 = 500$ Hz, and Figure 2(i) for 50 Hz $\leq F^m \leq 8 \times F_0 = 800$ Hz. As normal human speech does not have pitch frequency below 50 Hz, hence, the IMFs with F^m below 50 Hz, which are trend-like low amplitude signals, are left out of the summation. It can be seen that just like IMF_3, the sum of IMFs starting from IMF_3, Figure 2(j), manifests the impulse locations at its prominent minima, but with potentially less spurious minima in the neighbourhood. Thus, the signal represented by Figure 2(j) provides a better candidate for estimating the impulse locations of $e_g(n)$, than that by IMF_3, Figure 2(d). The challenge now is to identify the minima which coincide with the impulse train, amongst the multitude of minima in the auxiliary signal. For this purpose, we need to look at Figure 2(k),(l). It can be observed that in either of these two signals, the impulse locations lie within the regions defined by their positive and negative zero-crossings, as represented by dotted rectangles. Either of these signals, thus, represents a low-resolution signal for identifying the excitation instants. Of course, in the case of Figure 2(k), some spurious regions will also be detected. Such regions need to be eliminated by a post processing mechanism, described later. However, in principle, once such regions are identified, the signal represented by Figure 2(j) may be used for high-resolution analysis. The location of the minimum value of the signal in

Figure 2(j), within the identified regions, gives the instants of impulse excitation.

Having seen how ICEEMDAN can be used as a time-domain multi-resolution tool for synthetic speech, we apply it to a real/natural speech signal. Figure 3 shows the same plots as that of Figure 2, but the IMFs here are extracted from a randomly selected speech file (F_s = 8 kHz) from the CMU-Arctic database [43]. Figure 3(a) shows the speech signal. Figure 3(b) shows the difference EGG (dEGG) signal, corresponding to the speech signal. The dEGG signal is the first difference of the EGG signal, recorded simultaneously with the speech signal. A peak detection algorithm is used to detect the large negative peaks of the dEGG signal, as shown in Figure 3(b), which indicate the *instants of significant excitation of the vocal tract* [46]. These are taken as the true or reference GCIs, and indicated by the vertical dashed lines in Figure 3(b)-(n). The average distance between a pair of consecutive reference GCIs gives us the *reference time-period*, the inverse of which gives the *reference pitch frequency* (F_0^{ref}). For the given speech signal, F_0^{ref} = 116 Hz. Figure 3(c)-(j) represent the first eight IMFs of the speech signal. As in the synthetic speech case, the F^ms of the IMFs, listed above their plots, decrease monotonically with the IMF order. Correspondingly, the IMFs assume a more sinusoidal shape.

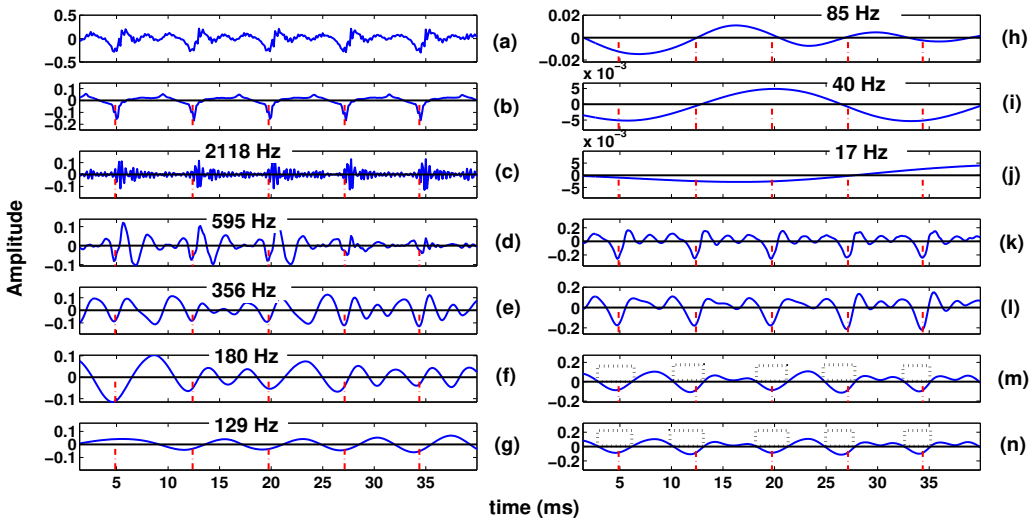

Fig. 3: **(a) A natural speech signal $s(n)$ having pitch frequency F_0^{ref} = 116 Hz ; (b) dEGG corresponding to $s(n)$. The negative peaks indicate the GCIs (c)-(j) are IMFs 1-8 respectively obtained from ICEEMDAN of $s(n)$. IMFs 9 and 10 are not shown ; (k) sum of IMFs 2-6 ; (l) sum of IMFs 3-6 ; (m) sum of IMFs 4-6 ; (n) sum of IMFs 5-6 ; Dashed vertical lines indicate the GCIs. Dotted rectangles in (k) and (l) indicate regions of search for the GCIs.**

It can be seen from Figure 3 that the GCIs correspond to different points in the near-sinusoidal IMFs. Amongst the IMFs, IMF_3, Figure 3(e), manifest the GCIs at some of its minima, with the least ambiguity - it seems the best candidate for GCIs estimation amongst the IMFs. Figure 3(k)-(n) represent the sum of IMFs $2-6$, $3-6$, $4-6$ and $5-6$, respectively. It may be observed that like IMF_3, the sum of IMFs 3-6, represented by Figure 3(l), also intersects the GCIs at some of its minima locations. It, however, has less spurious minima and is a better candidate for detecting the GCIs than IMF_3 itself. From a frequency domain perspective, Figure 3(l) represents the sum of IMFs having 50 Hz $\leq F^m \leq 5 \times F_0^{ref}$. Similarly, Figure 3(m) represents the sum of IMFs having 50 Hz $\leq F^m \leq 2 \times F_0^{ref}$, and Figure 3(n) the sum of IMFs having 50 Hz $\leq F^m \lesssim F_0^{ref}$, $F_m^5 = 129$ Hz $\approx F_0^{ref} = 116$ Hz. As in the synthetic case, either of the signals represented by Figure 3(m),(n) can be used for estimating the *regions of search* for the GCIs. Once such regions are identified, the signal in Figure 3(l) can be used to estimate the exact locations of the GCIs.

Expectedly, the preceding observations pertaining to the IMFs obtained from ICEEMDAN are also found to be true for the IMFs obtained from MEMD. Hence, figures equivalent to Figures 2 and 3, but for the IMFs obtained from MEMD are not included in this manuscript. However, while MEMD provides a significant time-advantage with respect to ICEEMDAN, it is comparatively less effective in reducing *mode-mixing*. As such, the sum of the lower-frequency IMFs are not very effective in providing us the *regions of search* for the GCIs, unlike in the case of ICEEMDAN. Therefore, we would require to devise an alternate methodology for finding the *regions of search* of the GCIs, in the case of MEMD.

Thus, in the case of both ICEEMDAN and MEMD, the partial sum of the IMFs, having their mean frequencies in the range $\{$ 50 Hz $\leq F^m \leq 5 \times F_0^{ref}$ $\}$, provides us with a *sinusoid-like* signal, some of whose minima locations manifest the GCIs. We may call this signal, resulting from the partial sum of the IMFs, as the *Source Enhanced and De-trended Speech* (SEDS) signal.

4 Procedure for ICEEMDAN based GCIs Estimation (IGE)

Based on the principle described in the previous section, the following procedure is employed to detect the GCIs from the speech signal.

- **(i)** Decompose $s(n)$ using ICEEMDAN. Construct the *de-trended speech* signal, $s_d(n)$.

$$s(n) = \sum_{k=1}^{10} h_k(n), \;\; s_d(n) = \Big\{ \sum_k h_k(n) | F_k^m > 50 \text{ Hz} \Big\} \tag{6}$$

- **(ii)** Estimate the average pitch frequency, F_0^{est}, from $s_d(n)$. In this work, the *Robust Algorithm for Pitch Tracking* (**RAPT**) [61] method is used to estimate the pitch frequencies of voiced frames of the de-trended speech

signal. Frames of 20 ms, with 50% overlap, are used. F_0^{est} represents the average pitch frequency over all the voiced frames.

- **(iii)** Sum up the components with 50 Hz $\leq F^m \leq 2 \times F_0^{est}$.

$$s_R(n) = \sum_k h_k(n) \; \forall \left\{ k \mid 50 \text{ Hz} \leq F_k^m \leq 2 \times F_0^{est} \right\} \tag{7}$$

- **(iv)** Detect the negative-going and positive-going zero-crossings of $s_R(n)$. The region from a given negative-going zero-crossing to the closest positive-going zero-crossing to its right, is a *region of search*. Let $\{R_r,\ r = 1, 2, ..., I_{ige}\}$ represent the regions of search, I_{ige} being the total number of regions.

- **(v)** Discard the first IMF, and sum up the rest of the components with 50 Hz $\leq F^m \leq 5 \times F_0^{est}$. This provides us with the SEDS signal.

$$s_e(n) = \sum_{k \geq 2} h_k(n) \; \forall \left\{ k \mid 50 \text{ Hz} \leq F_k^m \leq 5 \times F_0^{est} \right\} \tag{8}$$

- **(vi)** In any r^{th} region of search (R_r), find the location of minimum amplitude in $s_e(n)$. These locations provide the initial estimates of the GCIs, $\{l_r^i,\ r = 1, 2, ..., I_{ige}\}$.

Figure 4 shows the GCIs estimated using the above process from an arbitrary speech file. The dashed stem lines indicate the GCIs estimates obtained by the aforementioned process. However, as seen in Figure 4(d), along with the actual GCIs, some spurious GCIs are also estimated. The reason for such spurious estimates is the *imperfect sinusoidal nature* of the IMFs. The existence of false estimates is a familiar problem in many GCI estimation methods, and needs further refinement [48, 58, 59, 63]. To eradicate such false estimates, we apply the following simple methodology.

- **(i)** For the r^{th} estimated GCI, l_r^i, consider a region,

$$l_r^i - W/2 < L_r < l_r^i + W/2, \tag{9}$$

$$\text{where } W = \nu \frac{F_s}{F_0^{est}},\ \nu \in \mathbb{R}^+ \tag{10}$$

The size of the window, W, is determined by the factor ν, and the effect of this parameter is discussed in the results section.

- **(ii)** Within any p^{th} region, L_p, find the GCIs estimated within it, i.e, find the set

$$\{(l_r^i, v_r^i) \mid (l_r^i, v_r^i) \in L_p,\ r = 1, 2, ..., I_{ige}\}, \tag{11}$$

$$\text{where } v_r^i = s_e(l_r^i) \tag{12}$$

Compare the amplitudes, v_r^i, of this set. If the amplitude v_p^i is the minimum amplitude in this set, then l_p^i is considered as a legitimate GCI, otherwise it is rejected as a spurious one. May the *final estimated GCIs* be denoted by $\{l_r^{est},\ r = 1, 2, ..., F_{ige} \leq I_{ige}\}$.

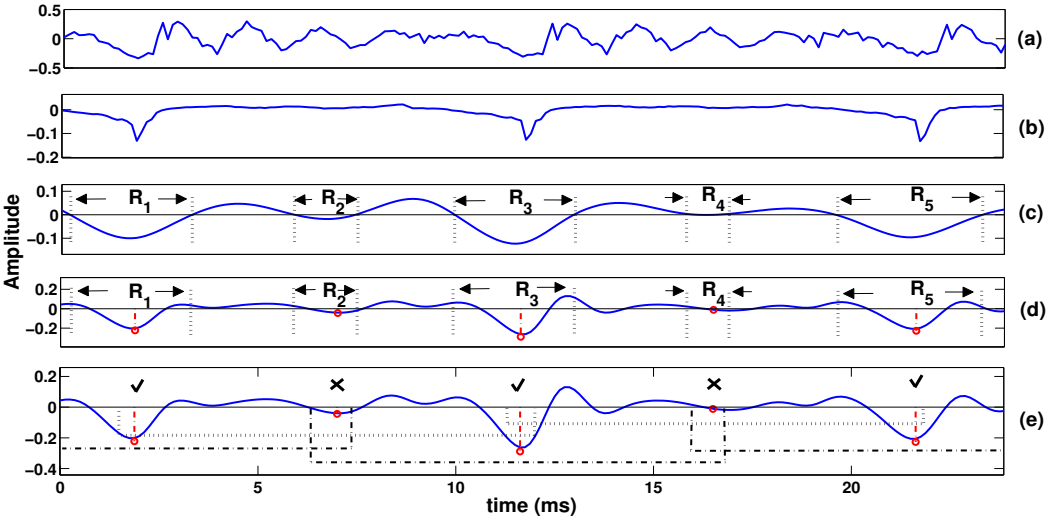

Fig. 4: **(a) Segment of a speech signal from the CMU-Arctic database ; (b) dEGG signal corresponding to the speech signal ; (c)** $s_R(n)$ **obtained from the IMFs of ICEEMDAN. The search regions are denoted by** $\{R_r, r = 1, 2, ..., 5\}$ **; (d)** $s_e(n)$**. The estimated GCIs within** $\{R_r, r = 1, 2, .., 5\}$ **are denoted by dashed stem lines ; (e)** $s_e(n)$**. Spurious and correct GCIs estimates are indicated by cross and ticks respectively. Dotted rectangles indicate the** L_r **regions of estimated spurious GCIs. Dash-dotted rectangles indicate the** L_r **regions of estimated correct GCIs.**

Figure 4(e) shows how the above process works. Let us consider the first (leftmost) GCI estimate, l_1^i. It has a region L_1 specified by the window W. There are two GCIs estimates within L_1. The amplitude, v_1^i, of $s_e(n)$ at l_1^i, is lesser than that of the other GCI estimated within the region. Hence, l_1^i is considered a correct GCI estimate. Now, let us consider the next GCI estimate, l_2^i. Clearly, it is a spurious one, and hence marked X. There are two other GCIs apart from l_2^i within the region L_2. Amongst these three GCIs locations, since the amplitude, v_2^i, of $s_e(n)$ at the spurious GCI location, l_2^i, is not the minimum, it is rejected. Next, consider the third GCI estimate, l_3^i, which is a correct estimate. There are two more GCIs estimated within the region L_3. In this case, the amplitude, v_3^i, of $s_e(n)$ at l_3^i, is the minimum amongst the amplitudes at all the three GCIs estimates within L_3. Hence, l_3^i is accepted as a correct estimate. The last two estimates are verified in the same fashion.

The entire process may be termed as *ICEEMDAN based GCIs Estimation* (**IGE**).

5 Procedure for MEMD based GCIs Estimation (MGE)

As MEMD is comparatively less effective than ICEEMDAN in reducing *mode-mixing*, the sum of lower-frequency IMFs are not very effective in providing us the *regions of search* for the GCIs. Therefore, we devise an alternate methodology for finding the *regions of search* of the GCIs, based on which the SEDS will provide the high-resolution estimates of the GCIs.

- **(i)** Decompose the speech signal $s(n)$ using MEMD. Construct the *de-trended speech* signal, $s_d(n)$.

$$s(n) = \sum_{k=1}^{10} h_k(n),\ s_d(n) = \Big\{ \sum_k h_k(n) | F_k^m > 50 \text{ Hz} \Big\} \tag{13}$$

- **(ii)** Estimate the average pitch frequency, F_0^{est}, from $s_d(n)$. In this work, the **RAPT** algorithm is used to estimate the pitch frequencies of voiced frames of the de-trended speech signal. Frames of 20 ms, with 50% overlap, are used. F_0^{est} represents the average pitch frequency over all the voiced frames.

- **(iii)** Let $s_h(n) = \Delta\Delta s_d(n)$, where Δ indicates a difference operation. Let n_d denote the minima locations of $s_h(n)$. Construct the envelope, $e_d(n)$, of $s_d(n)$, by using cubic spline interpolation, with the points of interpolation being $\{n_d, s_d(n_d)\}$.

- **(iv)** Detect the negative-going and positive-going zero-crossings of $e_d(n)$. The region from a given negative-going zero-crossing to the closest positive-going zero-crossing to its right, provides a prospective *initial region of search*. Let $\{R_r^i,\ r = 1, 2, ..., I_{mge}\}$ represent the initial regions, I_{mge} being the total number of regions.

- **(v)** In every $\{R_r^i,\ r = 1, 2, ..., I_{mge}\}$, find the point of minimum amplitude in $e_d(n)$, (n_r^i, v_r^i), where, $v_r^i = e_d(n_r^i)$.

- **(vi)** For a given $(n_r^i, v_r^i),\ r = 1, 2, ..., I_{mge}$, consider a window, L_r, given by

$$n_r^i - W/2 < L_r < n_r^i + W/2, \tag{14}$$

$$\text{where } W = \nu \frac{F_s}{F_0^{est}},\ \nu \in \mathbb{R}^+ \tag{15}$$

  The size of the window, W, is determined by the factor ν, and the effect of this parameter is discussed in the results section.

- **(vii)** Within any given p^{th} window, L_p, find the set $\{(n_r^i, v_r^i) \mid (n_r^i, v_r^i) \in L_p,\ r = 1, 2, ..., I_{mge}\}$. Compare the amplitudes, v_r^i, of this set. If v_p^i is the minimum amplitude in this set, then the region, R_p^i, is considered as a legitimate region of search, otherwise it is rejected as a spurious one. May the *final set of regions of search* be denoted by $\{R_r^f,\ r = 1, 2, ..., F_{mge} \leq I_{mge}\}$.

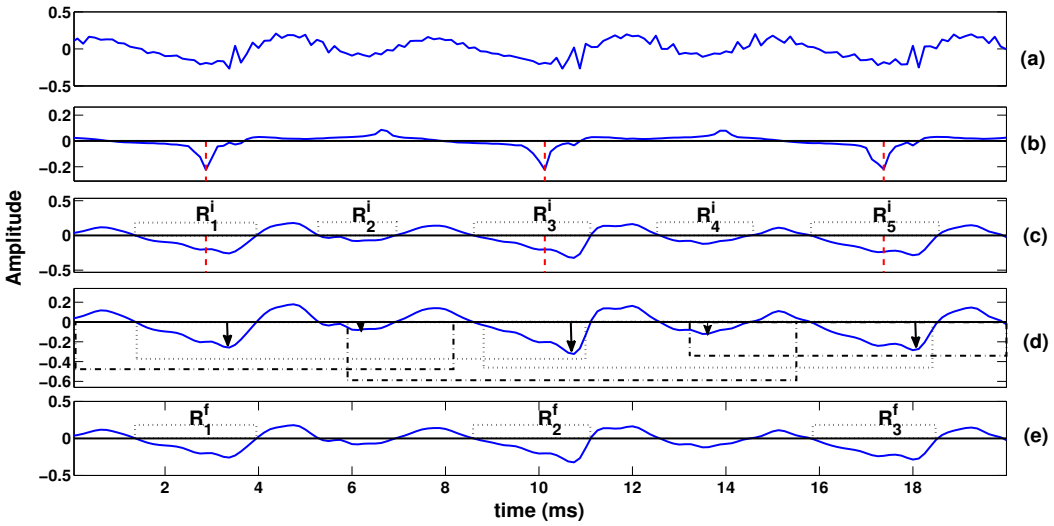

Fig. 5: **(a) Segment of a speech signal from the CMU-Arctic database ; (b) dEGG signal corresponding to the speech signal. The dashed vertical lines indicate the reference GCIs ; (c)** $e_d(n)$**. The dashed vertical lines indicate the reference GCIs. The dotted rectangles indicate the** *initial regions of search*, $\{R_r^i,\ r = 1, 2, ..., 5\}$ **; (d)** $e_d(n)$**. The vertical arrows indicate the** *points of minimimum amplitude*, $\{(n_r^i, v_r^i),\ r = 1, 2, ..., 5\}$**, corresponding to the initial regions. The dashed rectangles indicate the windows,** L_1^i, L_3^i**, and** L_5^i**. The dotted rectangles indicate the windows,** L_2^i**, and** L_4^i **; (e)** $e_d(n)$**. The dotted rectangles indicate the** *final regions of search*, $\{R_r^f,\ r = 1, 2, 3\}$**.**

Figure 5 demonstrates the working of the above process. As may be observed from Figure 5(c), $e_d(n)$, represents a symmetric smooth envelope of the given speech signal, Figure 5(a). There are five *initial regions of search* in $e_d(n)$, out of which three regions contain the GCIs (indicated by the vertical dashed lines), and the other two are spurious. The five *minimum value points* corresponding to these five regions are indicated by arrows in Figure 4(d). Centered around each *minimum value point*, a symmetric window of width W is considered. For the first (leftmost) point, (n_1^i, v_1^i), the window contains the points $\{(n_1^i, v_1^i),\ (n_2^i, v_2^i)\}$. As $v_1^i = \min\{v_1^i, v_2^i\}$, R_1^i is considered as a legitimate region. For the second point, (n_2^i, v_2^i), the window contains the points $\{(n_1^i, v_1^i),\ (n_2^i, v_2^i),\ (n_3^i, v_3^i)\}$. As $v_2^i \neq \min\{v_1^i, v_2^i, v_3^i\}$, R_2^i is considered as a spurious region. In the same manner, the legitimacy of the remaining regions, are evaluated. Finally, three legitimate regions remain, $\{R_1^f, R_2^f, R_3^f\} = \{R_1^i, R_3^i, R_5^i\}$.

Having derived the regions, $\{R_r^f,\ r = 1, 2, ..., F_{mge}\}$, where the GCIs are ought to be contained, the following procedure is applied :

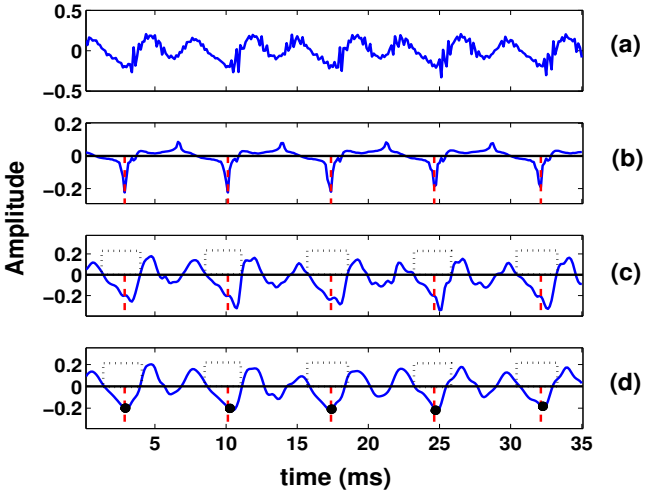

Fig. 6: **(a) Segment of a speech signal from the CMU-Arctic database ; (b) dEGG signal corresponding to the speech signal. The dashed vertical lines indicate the reference GCIs ; (c)** $e_d(n)$**. The dashed vertical lines indicate the reference GCIs. The dotted rectangles indicate the** *final regions of search* **; (d)** $s_e(n)$**. The dotted rectangles indicate the** *final regions of search***. The circles indicate the minimum amplitude points, within the** *final regions of search***, which give the estimates of the GCIs.**

- **(i)** Discard the first IMF, and sum up the rest of the components with 50 Hz $\leq F^m \leq 5 \times F_0^{est}$. This results in the SEDS signal, $s_e(n)$.

$$s_e(n) = \sum_{k \geq 2} h_k(n) \; \forall \left\{ k \mid 50 \text{ Hz} \leq F_k^m \leq 5 \times F_0^{est} \right\} \tag{16}$$

- **(ii)** In the regions of search, $\{R_r^f, \; r = 1, 2, ..., F_{mge}\}$, find the corresponding locations of minimum amplitude in $s_e(n)$. These locations, $\{l_r^{est}, \; r = 1, 2, ..., F_{mge}\}$, are the estimates of the GCIs.

Figure 6(d) shows the SEDS signal, $s_e(n)$, corresponding to the given speech signal. The *final regions of search*, as estimated from the speech envelope signal, $e_d(n)$, shown in Figure 6(c), provide the time intervals where the GCIs are contained. Then, within these intervals (shown by dotted rectangles), the points of minimum value (shown by circles) of the SEDS signal may be associated with the GCIs.

This complete process of estimating the GCIs is termed as *MEMD based GCIs Estimation* (**MGE**).

Table 2: **Description of the databases used. The CMU-Arctic database consists of five speakers, and the APLAWDW database consists of ten speakers.**

Database	CMU-Arctic (5 speakers)					APLAWDW (10 speakers)
Speakers	BDL (male)	JMK (male)	SLT (female)	EDX (male)	KDT (male)	All (5 male , 5 female)
Number of Utterances	∼1100	∼1100	∼1100	∼1900	∼500	∼50 per speaker
Average Duration	∼3 s	∼3 s	∼3 s	∼1 s	∼2 s	∼3 s

6 Results and Discussion

Two databases, the CMU-Arctic database [43], and the APLAWDW database [3], are considered for this work. The CMU-Arctic database consists of 5 speakers and the APLAWDW database consists of 10 speakers. The databases are described in Table 2. All utterances (from both the databases) considered in this work are of $F_s = 8$ kHz.

In order to evaluate the performances of the proposed IGE and MGE algorithms five objective metrics have been used [16, 17, 46, 48, 49, 59, 63]. They are described below :

- **Identification rate (IR)** : The percentage of glottal or larynx cycles in which only one GCI is estimated.
- **Miss rate (MR)** : The percentage of glottal cycles in which only no GCI is estimated.
- **False alarm rate (FAR)** : The percentage of glottal cycles in which more than one GCI is estimated.
- **Identification accuracy (IA)** : The standard deviation of the timing or identification error, ζ.
- **Accuracy to $\pm$ 0.25 ms (IA$'$)** : The percentage of GCIs which are within a timing error bound of $\pm$ 0.25 ms.

Figure 7 aids in understanding the metrics used for evaluation. Out of the five metrics, IR, MR, and FAR reflect the reliability of the estimation technique, whereas IA and IA$'$ represent the accuracy of the technique. Higher values of IR and IA$'$, and lower values of MR, FAR, and IA are desirable. Using these five metrics, the proposed algorithms are compared with five competitive state-of-the-art algorithms : DYPSA, YAGA, SEDREAMS, ZFR, and ILPR-PI.

The performances of the various algorithms are evaluated not only for clean speech, but speech degraded by White, Babble and HFchannel noise [65], with the *signal to noise ratio* (SNR) being varied from 0 to 20 dB, in steps of 5 dB. Apart from this, the performances for speech signals affected by telephone

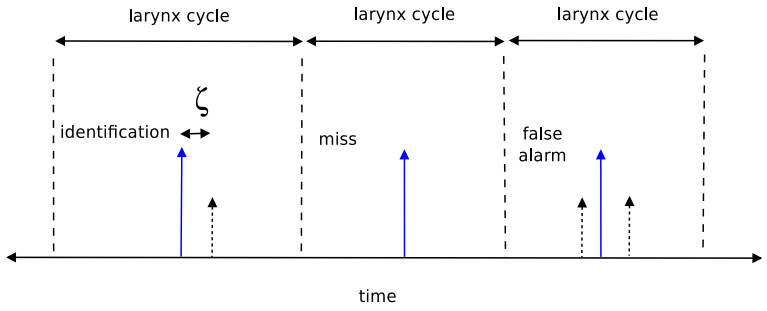

Fig. 7: **Characterization of the estimates of the GCIs showing three larynx or glottal cycles, with examples of each possible outcome from the estimates [16, 17, 46, 48, 49, 59, 63]. The solid arrows indicate the reference GCIs, obtained from the dEGG signal. The dotted arrows indicate the GCIs estimated. ζ is the** *identification error*.

channel effects are also evaluated. Two types of telephone-quality speech are considered in this work :

- **T-1** : The T-1 type of telephone-quality speech has been derived by band-pass filtering the speech signal between 300 - 3400 Hz, using the **VOICEBOX** toolbox [8]. The magnitude response of the filter is given by a raised cosine function in the range of 0-300 Hz and 3400-4000 Hz, and unity between 300 - 3400 Hz [8, 49].
- **T-2** : The T-2 type of telephone-quality speech has been derived by band-pass filtering the speech signal in accordance with ITU standards [42].

6.1 **Effect of ν in the GCIs estimation performance**

As discussed in the previous section, the existence of false alarms, i.e., false estimates of the GCIs, is almost inevitable in most of the GCIs estimation methods. In the case of IGE (MGE) too, such estimates (regions in the case of MGE) are obtained initially along with correct estimates (regions in the case of MGE). For eliminating such estimates (regions in the case of MGE), we employ a shifting window of size W, which is dependent on an adjustable parameter, ν. Hence, the effect of ν on the overall performance of the IGE and MGE algorithms needs to be studied, and its optimum range needs to be found out. It also needs to be investigated if the optimum range of ν varies for the two databases. For this purpose, the IR, MR, and FAR for the two methods are obtained for different values of $\nu \in [0.5, 2]$, for the two databases separately. Only clean speech signals are used for this study.

Figure 8 shows the plots of the three metrics for IGE, and Figure 9 shows the same for MGE. Separate plots are shown for the CMU-Arctic and APLAWDW databases. As can be observed from Figure 8, the IRs for both the databases obtain saturation somewhere between $\nu = 1.1$ to $\nu = 1.5$. The same is observed for MGE, as shown in Figure 9. In this range, for both the algorithms,

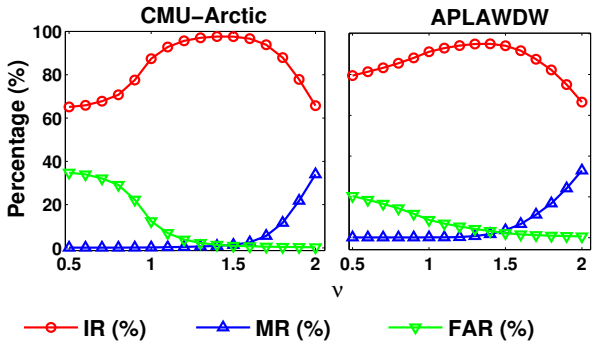

Fig. 8: **Effect of the window size parameter, ν, in estimating the GCIs using the IGE algorithm. Clean speech signals from the CMU-Arctic and APLAWDW databases are used.**

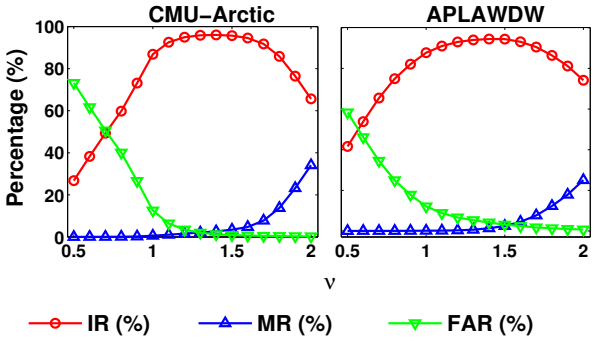

Fig. 9: **Effect of the window size parameter, ν, in estimating the GCIs using the MGE algorithm. Clean speech signals from the CMU-Arctic and APLAWDW databases are used.**

and both the databases, the FAR drops sharply, with little increase in MR. This means that, assuming F_0^{est} is a reasonably accurate estimate of the actual F_0, the window size should be slightly larger than the *pitch period* (the inverse of F_0), to eliminate the false estimates/regions, without adversely effecting the overall performance of the algorithm. Larger values of ν results in a large window, which not only eliminates false estimates/regions, but also the correct ones. Smaller values of ν, on the contrary, are not very efficient in eliminating the false estimates. It is also important to notice that the optimal range of ν is nearly the same for both the databases. Hence, the window size may be considered independent both in terms of the methodology employed to estimate the GCIs, and the change in dataset.

As such, using $\nu \in [1.1, 1.5]$ should give us results which is close to the best possible performance of the algorithm. Henceforth, $\nu = 1.4$ is used for estimating the GCIs by both IGE and MGE in this work, under all conditions.

6.2 **Performances under clean and telephone channel conditions**

Table 3: **Robustness of the GCIs estimation methods, according to the five performance metrics, for clean speech signals.**

Database	Metric	IGE	MGE	DYPSA	YAGA	SEDREAMS	ZFR	ILPR-PI
BDL (CMU)	IR(%)	97.59	97.63	94.96	97.82	97.51	97.12	**98.52**
	MR(%)	0.68	1.49	2.7	0.71	1.12	1.41	**0.62**
	FAR(%)	1.73	0.89	2.34	1.47	1.37	1.47	**0.86**
	IA(ms)	0.33	0.44	0.52	0.41	0.4	0.41	**0.31**
	IA'(%)	71.34	63.34	78.6	84.5	83.8	82.62	**87.3**
JMK (CMU)	IR(%)	**98.93**	98.9	97.5	98.76	98.56	95.83	98.7
	MR(%)	**0.35**	0.4	1.4	0.55	0.48	3.81	0.59
	FAR(%)	0.72	0.71	1.1	0.69	0.96	**0.36**	0.71
	IA(ms)	0.53	0.53	0.57	0.51	0.54	0.72	**0.35**
	IA'(%)	58.67	57.61	73.56	74.52	73.6	36.76	**83.7**
SLT (CMU)	IR(%)	**99.39**	99.28	97.15	98.15	98.45	98.96	98.83
	MR(%)	**0.1**	0.23	1.75	0.47	0.31	0.49	0.56
	FAR(%)	0.52	**0.49**	1.1	1.38	1.24	0.55	0.61
	IA(ms)	**0.23**	0.24	0.56	0.39	0.41	0.32	0.34
	IA'(%)	77.63	75.79	67.16	**81.23**	74.9	78.8	80.78
EDX (CMU)	IR(%)	96.74	94.19	80.03	95	98.16	91.5	**97.85**
	MR(%)	0.93	3.63	2.5	**0.9**	0.8	7.1	0.9
	FAR(%)	2.33	2.18	17.47	4.1	1.04	1.4	**1.25**
	IA(ms)	0.79	0.97	0.66	0.7	**0.53**	0.8	0.53
	IA'(%)	36.87	34.07	82.1	83.32	85.2	50.58	**85.6**
KDT (CMU)	IR(%)	94.49	91.7	95.1	96.51	96.9	84.67	**97.12**
	MR(%)	4.65	7.73	2.1	0.91	1.12	8.86	**0.21**
	FAR(%)	0.86	**0.56**	2.8	2.58	1.98	6.47	2.67
	IA(ms)	0.93	1.05	0.56	0.58	0.53	0.83	**0.37**
	IA'(%)	50.92	33.23	82.66	88.15	84.12	38.12	**90.15**
APLAWDW	IR(%)	95.05	94.46	94.5	**96.9**	96.8	96.75	95.5
	MR(%)	1.72	1.35	2.5	**1.1**	1.8	1.7	1.9
	FAR(%)	3.23	4.18	3	2	**1.4**	1.55	2.6
	IA(ms)	**0.39**	0.42	0.8	0.69	0.62	0.75	0.6
	IA'(%)	64.86	59.28	68.15	77.5	78.12	49.21	**80.15**

The performances of IGE, MGE, and the five state-of-the-art algorithms are first evaluated for clean speech signals of the the CMU-Arctic and APLAWDW databases. Table 3 shows the five performance metrics for the algorithms. The five speakers of the CMU-Arctic database are shown individually, for better analysis. The best performances are highlighted in bold-font. As can be

observed from the table, among the state-of-the art algorithms YAGA, SEDREAMS and ILPR-PI provide consistent performance across the different speakers of the CMU-Arctic database. DYPSA and ZFR, on the other hand, show some fluctuations in performance with respect to the change in speakers. The performance of DYPSA drops starkly for EDX, and that of ZFR drops for KDT. The performances of IGE are relatively invariant to speaker change, similar to that of YAGA, SEDREAMS and ILPR-PI. The performances of MGE are worse for the KDT and EDX speakers, compared to that of the other speakers, even though the variation in performance is not as alarming as that of DYPSA and ZFR. The low FAR values, in the case of both IGE and MGE, prove how well the methods for eliminating spurious GCIs work.

Overall, the performances of IGE and MGE are competitive with the state-of-the-art algorithms, for every speaker case. The performances of IGE and MGE for the APLAWDW database are similar to the state-of-the art algorithms. The only parameter, one may argue, in which the IGE and MGE lag behind with respect to the other algorithms is IA$'$. This is a direct result of the effect of mode-mixing, as it cannot be completely eliminated. Thus, while some GCIs estimates are extremely accurate, certain others are not exactly aligned with the reference GCIs. However, it must be noted that the IA$'$ values of IGE and MGE are still competitive with respect to the state-of-the-art algorithms.

Having observed the performances of IGE and MGE over clean speech signals, we now investigate whether they could perform credibly when the speech signals are affected by telephone channel conditions. Under telephone channel conditions, the low-frequency spectrum of the speech signal, which contains the glottal source information, is suppressed. This affects the waveform-shape of the speech signal, but the intelligibility of the speech signal is still intact. Thus, the performances of the algorithms for telephone-quality speech depicts their ability to extract critical information when the information source (the glottal characteristics) itself is significantly obscured or hidden. Further, the performances of the algorithms for telephone-quality speech is important as mobile communication is the cheapest and easiest means of acquiring speech data.

Table 4 shows the five performance metrics for the two types of telephone-quality speech, T-1 and T-2, for the CMU-Arctic and APLAWDW databases, for the seven algorithms. As the table shows, the performances of all the algorithms (particularly in terms of IR) degrade under telephone channel conditions. YAGA, SEDREAMS and ZFR, all of which performed credibly under clean speech conditions, show significant degradation for both types of telephone-quality speech. For T-1 type of telephone-quality speech, only IGE, MGE, DYPSA and ILPR-PI provide credible performances (IR $\gtrsim 80\%$). When the IR is itself low, the metrics related to accuracy (IA and IA$'$) become less relevant. In the case of T-2 type of telephone-quality speech, the performances of all the algorithms are worse than that of T-1 type. This is expected, as ITU telephone channel codecs suppress the lower-frequency spectrum of speech much more than simple band-pass filtering. For T-2 type, only IGE, MGE

Table 4: **Robustness of the GCIs estimation methods, according to the five performance metrics, for two types of telephone-quality speech (T-1 and T-2).**

Type	Database	Metric	IGE	MGE	DYPSA	YAGA	SEDREAMS	ZFR	ILPR-PI
T-1	CMU-Arctic	IR(%)	**91.1**	89.77	88.29	45.81	72.05	54.36	79.21
		MR(%)	5.39	7.14	0.8	0.9	0.12	**0.01**	1.05
		FAR(%)	3.5	**3.09**	10.91	53.29	27.83	45.63	19.74
		IA(ms)	1.03	1.1	0.4	1.28	0.34	**0.25**	1.07
		IA'(%)	33.76	32.98	81.35	27.26	**80.98**	75.68	42.62
	APLAWDW	IR(%)	86.04	84.59	**91.35**	75.33	70.67	72.26	85.52
		MR(%)	2.32	5.27	0.77	0.96	1.42	**0.13**	0.96
		FAR(%)	11.64	10.14	**7.88**	23.71	27.91	27.61	13.51
		IA(ms)	0.9	1.12	0.44	1.46	0.92	**0.76**	0.76
		IA'(%)	29.69	25.97	**79.79**	10.99	44.41	45	59.49
T-2	CMU-Arctic	IR(%)	85.36	**86.74**	80.62	36.3	18.23	17.65	65.82
		MR(%)	9.83	9.98	1.76	1.17	0.49	**0.08**	1.45
		FAR(%)	4.81	**3.28**	17.62	62.53	81.28	82.27	32.73
		IA(ms)	1.49	1.27	0.69	1.05	1.08	**0.58**	0.96
		IA'(%)	20.28	24.88	49.97	34.54	33.05	**55.46**	53.03
	APLAWDW	IR(%)	**85.22**	84.85	86.18	67.18	27.14	35.86	83.6
		MR(%)	6.68	8.55	2.22	**0.78**	2.43	0.62	1.89
		FAR(%)	8.1	**6.6**	11.6	32.04	70.43	63.51	14.51
		IA(ms)	1.13	1.08	0.62	0.74	1.57	1.69	**0.56**
		IA'(%)	17.64	19.2	48.34	53.3	15.55	16.83	**71.36**

and DYPSA provide credible performances, compared to the other algorithms. The performance of ILPR-PI degrades starkly for the CMU-Arctic database for T-2 type, compared to that of T-1 type.

Thus, out of the five state-of-the-art algorithms, only DYPSA, which did not perform as well as the other algorithms for clean speech, works credibly in the case of telephone-quality speech. Noticeably, IGE and MGE provide credible performance for telephone-quality speech, just like for clean speech.

6.3 **Performance under noisy conditions**

Having evaluated the performances of the algorithms on telephone-quality speech, we now proceed towards investigating their performances on speech signals corrupted by noise. Noisy speech signals represent the acquisition of speech data under practical or imperfect recording conditions. Evaluation of

the performances of the algorithms for noisy speech signals would depict the resistance of the algorithms to external sources (any other signal source not corresponding to the speaker) which try to mask or subdue the spectrum of the speech signal. Depending on both the level and type of noise, the addition of noise to the speech signal not only changes its waveform-shape but also its intelligibility. Therefore, three types of noise are considered [65] :

- **Babble :** Which masks the low-frequency speech spectrum.
- **White :** Which masks the entire speech spectrum.
- **HFchannel :** Which masks the high-frequency speech spectrum.

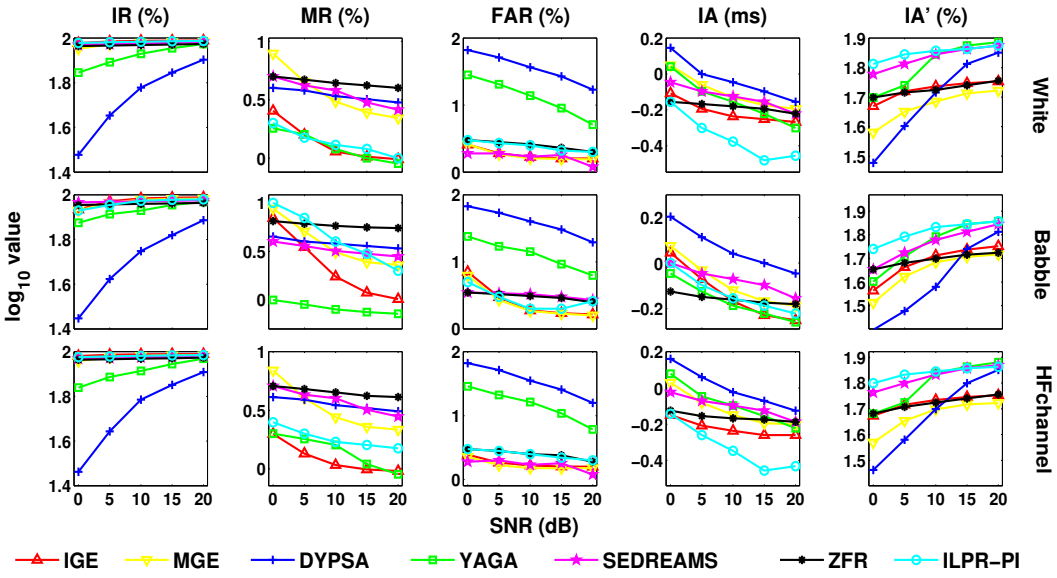

Fig. 10: **Robustness of the GCIs estimation methods, according to the five performance metrics, for speech signals corrupted by White, Babble, and HFchannel noise. The performance metrics are averaged over all the speech signals of the CMU-Arctic and APLAWDW databases combined. The SNR is varied from 0 dB to 20 dB for each type of noise.**

Figure 10 shows the performances of the seven algorithms for White, Babble and HFchannel noise, averaged over the two databases. For visual clarity, the plots are shown in logarithmic scale. First, let us consider the case of speech signals corrupted by White noise. As can be observed from the plots, all the algorithms perform similarly for all the metrics, particularly when the level of noise is low (SNR $\geq$ 15 dB). As the level of noise increases, the IRs of DYPSA and YAGA worsen substantially with respect to the other algorithms. ZFR, which performed poorly under telephone channel conditions, remarkably shows very little fluctuation with respect to the level of noise, for all the five metrics. The converse is true for DYPSA. This shows the limited adaptability

of the popular techniques with respect to changing conditions. For IGE, one may notice the consistently high values of IR (low MR and FAR) even at high noise levels. The same is true for MGE. Compared to IGE, however, the MR of MGE is higher at high noise levels. Again, as in the case of clean speech, the only metric that IGE and MGE might be considered lagging is IA′. As mentioned earlier, this is a result of mode-mixing. However, the values of IA′ for the two algorithms are still close to the state-of-the-art methods, particularly ZFR.

Next, let us consider the case of HFchannel noise. As the plots in Figure 10 show, the performances for HFchannel noise are similar to that of White noise for all the seven algorithms, and the five metrics. As such, the same observations made for White noise are applicable for HFchannel noise. In both these types of noise, IGE and MGE show limited fluctuations in their performances as the level of noise increases. This may be because the SEDS signal is constructed leaving out the first IMF (and other high-frequency IMFs). The high-frequency IMFs, thus, capture most of the White/HFchannel noise, making IGE and MGE immune to its influence.

Lastly, let us consider the case of Babble noise. As can be observed in Figure 10, the performances of the algorithms are worse for Babble noise, than that for White and HFchannel noise. DYPSA, which performed remarkably for telephone-quality speech, is also not immune to external low-frequency noise. Apart from DYPSA and YAGA, the rest of the algorithms almost converge to the same performance levels, and there is hardly any difference amongst their performances. The IRs of IGE and MGE are competitive with the best performing algorithms for different levels of noise. The MRs, FARs and IAs are similarly competitive. The IA′ values are now even closer to the best cases, and almost identical to that of ZFR.

Thus, considering various conditions - clean, telephone channel, and noise - both IGE and MGE perform not only competitively, but also consistently, with respect to the popular algorithms. Contrary to the state-of-the-art techniques, IGE and MGE may be deemed as “all weather techniques”.

6.4 **Computational complexity**

Having evaluated the performances of the algorithms, their time-costs also need to be considered for practical applicability. For this purpose, the same speech signal considered in Table 1 is again utilized. Table 5 lists the time taken by the seven methods in estimating the GCIs from the speech signal. The algorithms are run in MATLAB, in desktop mode, in a machine with 8 GB RAM, using Intel quad-core i7 processor, of 2.9 GHz clock frequency.

As is evident in Table 5, IGE is costly compared to the other algorithms. The efficiency of MGE is much better, and comparable to the state-of-the-art methods. The fact that the performance of MGE is not vastly different from that of IGE, while costing only a fraction of the time-cost of the IGE, makes it more suitable for practical applications. However, it is to be noted that

Table 5: **Computational complexity of the seven different methods for detecting the GCIs. Time (in seconds) taken by the seven algorithms for detecting the GCIs from a speech signal, of ∼4 s duration.**

Algorithm	IGE	MGE	DYPSA	YAGA	SEDREAMS	ZFR	ILPR-PI
Time (s)	63.17	1.62	0.76	1.19	0.74	2.28	0.54

the two proposed algorithms for estimating the GCIs from the IMFs do not account principally for their overall time-costs. It is the time taken to extract the IMFs that is the main reason (particularly for ICEEMDAN), as illustrated in Table 1. It is to be expected that with efficient programming, using parallel processing, the time-cost of extracting the IMFs using ICEEMDAN could be significantly reduced. Further, as the mathematical framework of EMD [14, 44] develops, better and efficient methods of curbing mode-mixing may be expected, and the IMFs could provide even more accurate estimates of the GCIs, with the same proposed methodologies. Thus, even though currently the large time-cost seems a bottleneck for practical applications, that may not be so in the near future, or even now with efficient programming.

7 Conclusion

This work focussed on detecting the GCIs of the speech signal using non-linear and non-stationary signal analysis, as an alternative to the source-filter theory or LP-analysis based methodologies of doing the same. The motivation behind using non-linear and non-stationary analysis comes from the fact that state-of-the-art techniques, mainly based on short-time LP analysis, provide inconsistent performances when the speech signal is subjected to external influences. With this motivation, this work investigated the capability of two non-linear and non-stationary signal analysis methods - ICEEMDAN and MEMD - for reliably extracting GCIs under varied conditions - clean, noisy and under telephone channel effects. It was observed that both ICEEMDAN and MEMD could extract sinusoid-like components, called the IMFs, from the speech signal, which could be utilized to detect the GCIs. Henceforth, two uncomplicated processes were developed, called IGE and MGE, for detecting the GCIs from the IMFs derived from ICEEMDAN and MEMD respectively. IGE and MGE are observed to provide comparable performances, under varied conditions, with the state-of-the-art algorithms. More importantly, IGE and MGE are more consistent with (adaptive to) changing conditions. The principal drawback of IGE is that it is time-costly. MGE, of course, is a much faster algorithm, but its performance is marginally inferior to that of IGE. For clean speech and telephone-quality speech, the difference between the two algorithms is marginal. On an average, IR of MGE is within 1 % of that of the IGE, whereas IA differs by < 0.2 ms between the two algorithms. Again, one may argue that the time-cost of IGE could be drastically reduced by parallel

programming and efficient coding. Nevertheless, as EMD improves as a non-linear and non-stationary data analysis method, the proposed principle and methodologies for detecting the GCIs could be expected to provide even more accurate estimates of the GCIs.

References

1. http://perso.ens-lyon.fr/patrick.flandrin/emd.html. URL `http://perso.ens-lyon.fr/patrick.flandrin/emd.html`
2. http://www.bioingenieria.edu.ar/grupos/ldnlys/index.htm. URL `http://www.bioingenieria.edu.ar/grupos/ldnlys/index.htm`
3. http://www.commsp.ee.ic.ac.uk/ sap/resources/aplawdw/. URL `http://www.commsp.ee.ic.ac.uk/~sap/resources/aplawdw/`
4. Ananthapadmanabha, T., Yegnanarayana, B.: Epoch extraction from linear prediction residual for identification of closed glottis interval. Acoustics, Speech and Signal Processing, IEEE Transactions on **27**(4), 309–319 (1979)
5. Benesty, J., Sondhi, M.M., Huang, Y.: Springer handbook of speech processing. Springer Science & Business Media (2008)
6. Bouchikhi, A., Boudraa, A.O.: Multicomponent am–fm signals analysis based on emd–b-splines esa. Signal Processing **92**(9), 2214–2228 (2012)
7. Bozkurt, B., Dutoit, T.: Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In: ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (2003)
8. Brookes, M.: Voicebox. Speech Processing Toolbox for Matlab, Department of Electrical & Electronic Engineering, Imperial College (2009)
9. Cexus, J.C., Boudraa, A.O.: Nonstationary signals analysis by teager-huang transform (tht). In: Signal Processing Conference, 2006 14th European, pp. 1–5. IEEE (2006)
10. Chatlani, N., Soraghan, J.J.: Emd-based filtering (emdf) of low-frequency noise for speech enhancement. Audio, Speech, and Language Processing, IEEE Transactions on **20**(4), 1158–1166 (2012)
11. Chen, K., Zhou, X.C., Fang, J.Q., Zheng, P.f., Wang, J.: Fault feature extraction and diagnosis of gearbox based on eemd and deep briefs network. International Journal of Rotating Machinery **2017** (2017). DOI https://doi.org/10.1155/2017/9602650
12. Chen, Y., Wu, C.t., Liu, H.l.: Emd self-adaptive selecting relevant modes algorithm for fbg spectrum signal. Optical Fiber Technology **36**, 63–67 (2017)
13. Colominas, M.A., Schlotthauer, G., Torres, M.E.: Improved complete ensemble emd: A suitable tool for biomedical signal processing. Biomedical Signal Processing and Control **14**, 19–29 (2014)
14. Colominas, M.A., Schlotthauer, G., Torres, M.E.: An unconstrained optimization approach to empirical mode decomposition. Digital Signal Processing **40**, 164–175 (2015)
15. Deepak, K., Prasanna, S.: Epoch extraction using zero band filtering from speech signal. Circuits, Systems, and Signal Processing **34**(7), 2309–2333 (2015)
16. Drugman, T., Dutoit, T.: Glottal closure and opening instant detection from speech signals. In: Interspeech - Tenth Annual Conference of the International Speech Communication Association, pp. 2891–2894 (2009)
17. Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T.: Detection of glottal closure instants from speech signals: a quantitative review. Audio, Speech, and Language Processing, IEEE Transactions on **20**(3), 994–1006 (2012)
18. Drugman, T., Wilfart, G., Dutoit, T.: A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In: Tenth Annual Conference of the International Speech Communication Association (2009)
19. Flandrin, P.: Some aspects of huang's empirical mode decomposition, from interpretation to applications. In: Int. Conf. Computat. Harmonic Anal. CHA, vol. 4 (2004)
20. Flandrin, P., Goncalves, P.: Empirical mode decompositions as data-driven wavelet-like expansions. International Journal of Wavelets, Multiresolution and Information Processing **2**(04), 477–496 (2004)

21. Flandrin, P., Rilling, G., Goncalves, P.: Empirical mode decomposition as a filter bank. Signal Processing Letters, IEEE **11**(2), 112–114 (2004)
22. Gaubitch, N.D., Naylor, P.A.: Spatiotemporal averagingmethod for enhancement of reverberant speech. In: Digital Signal Processing, 2007 15th International Conference on, pp. 607–610. IEEE (2007)
23. Guo, Y., Huang, S., Li, Y., Naik, G.R.: Edge effect elimination in single-mixture blind source separation. Circuits, Systems, and Signal Processing **32**(5), 2317–2334 (2013)
24. Guo, Y., Naik, G.R., Nguyen, H.: Single channel blind source separation based local mean decomposition for biomedical applications. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, pp. 6812–6815. IEEE (2013)
25. Hao, H., Wang, H., Rehman, N.: A joint framework for multivariate signal denoising using multivariate empirical mode decomposition. Signal Processing **135**, 263–273 (2017)
26. Hardcastle, W.J., Marchal, A.: Speech production and speech modelling. 55. Springer Science & Business Media (1990)
27. Holambe, R.S., Deshpande, M.S.: Advances in Non-Linear Modeling for Speech Processing. Springer Science & Business Media (2012)
28. Huang, H., Pan, J.: Speech pitch determination based on hilbert-huang transform. Signal Processing **86**(4), 792–803 (2006)
29. Huang, N.E.: Empirical mode decomposition and hilbert spectral analysis. Shock and Vibration, 69th„ Minneapolis, MN, United States (1998). URL `https://ntrs.nasa.gov/search.jsp?R=19990078602`
30. Huang, N.E., Shen, S.S.: Hilbert-Huang transform and its applications, vol. 5. World Scientific (2005)
31. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **454**(1971), 903–995 (1998)
32. Jain, P., Pachori, R.B.: Event-based method for instantaneous fundamental frequency estimation from voiced speech based on eigenvalue decomposition of the hankel matrix. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **22**(10), 1467–1482 (2014)
33. Khaldi, K., Alouane, M.T.H., Boudraa, A.O.: A new emd denoising approach dedicated to voiced speech signals. In: Signals, Circuits and Systems, 2008. SCS 2008. 2nd International Conference on, pp. 1–5. IEEE (2008)
34. Khaldi, K., Boudraa, A.: Audio watermarking via emd. Audio, Speech, and Language Processing, IEEE Transactions on **21**(3), 675–680 (2013)
35. Khaldi, K., Boudraa, A.O.: On signals compression by emd. Electronics letters **48**(21), 1329–1331 (2012)
36. Khaldi, K., Boudraa, A.O., Bouchikhi, A., Alouane, M.T.H.: Speech enhancement via emd. EURASIP Journal on Advances in Signal Processing **2008**(1), 873,204 (2008)
37. Khaldi, K., Boudraa, A.O., Komaty, A.: Speech enhancement using empirical mode decomposition and the teager–kaiser energy operator. The Journal of the Acoustical Society of America **135**(1), 451–459 (2014)
38. Khaldi, K., Boudraa, A.O., Torresani, B., Chonavel, T.: Hht-based audio coding. Signal, image and video processing **9**(1), 107–115 (2015)
39. Khaldi, K., Boudraa, A.O., Torresani, B., Chonavel, T., Turki, M.: Audio encoding using huang and hilbert transforms. In: Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on, pp. 1–5. IEEE (2010)
40. Khaldi, K., Boudraa, A.O., Turki, M.: Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement. IET Signal Processing **10**(1), 69–80 (2016)
41. Khaldi, K., Boudraa, A.O., Turki, M., Chonavel, T., Samaali, I.: Audio encoding based on the empirical mode decomposition. In: Signal Processing Conference, 2009 17th European, pp. 924–928. IEEE (2009)
42. King, S., Karaiskos, V.: The Blizzard Challenge 2009. Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK (2009). URL `http://www.festvox.org/blizzard/bc2009/summary_Blizzard2009.pdf`

43. Kominek, J., Black, A.W.: The cmu arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis (2004)
44. Lin, C.D., Anderson-Cook, C.M., Hamada, M.S., Moore, L.M., Sitter, R.R.: Using genetic algorithms to design experiments: A review. Quality and Reliability Engineering International **31**(2), 155–167 (2015). DOI 10.1002/qre.1591
45. Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech communication **9**(5-6), 453–467 (1990)
46. Murty, K.S.R., Yegnanarayana, B.: Epoch extraction from speech signals. Audio, Speech, and Language Processing, IEEE Transactions on **16**(8), 1602–1613 (2008)
47. Naik, G.R., Selvan, S.E., Nguyen, H.T.: Single-channel emg classification with ensemble-empirical-mode-decomposition-based ica for diagnosing neuromuscular disorders. IEEE Transactions on Neural Systems and Rehabilitation Engineering **24**(7), 734–743 (2016)
48. Naylor, P.A., Kounoudes, A., Gudnason, J., Brookes, M.: Estimation of glottal closure instants in voiced speech using the dypsa algorithm. Audio, Speech, and Language Processing, IEEE Transactions on **15**(1), 34–43 (2007)
49. Prathosh, A., Ananthapadmanabha, T., Ramakrishnan, A.: Epoch extraction based on integrated linear prediction residual using plosion index. Audio, Speech, and Language Processing, IEEE Transactions on **21**(12), 2471–2480 (2013)
50. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals, vol. 100. Prentice-hall Englewood Cliffs (1978)
51. Rabiner, L.R., Schafer, R.W.: Introduction to digital speech processing. Foundations and trends in signal processing **1**(1), 1–194 (2007)
52. Rilling, G., Flandrin, P., Goncalves, P., et al.: On empirical mode decomposition and its algorithms. In: IEEE-EURASIP workshop on nonlinear signal and image processing, vol. 3, pp. 8–11. NSIP-03, Grado (I) (2003)
53. Schlotthauer, G., Torres, M., Rufiner, H.: Voice fundamental frequency extraction algorithm based on ensemble empirical mode decomposition and entropies. In: World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany, pp. 984–987. Springer (2010)
54. Schlotthauer, G., Torres, M.E., Rufiner, H.L.: Pathological voice analysis and classification based on empirical mode decomposition. In: A. Esposito, N. Campbell, C. Vogel, A. Hussain, A. Nijholt (eds.) Development of multimodal interfaces: active listening and synchrony, pp. 364–381. Springer (2010)
55. Sharma, R., Prasanna, S.M.: A better decomposition of speech obtained using modified empirical mode decomposition. Digital Signal Processing **58**, 26 – 39 (2016). DOI http://dx.doi.org/10.1016/j.dsp.2016.07.012. URL `http://www.sciencedirect.com/science/article/pii/S1051200416300975`
56. Sharma, R., Prasanna, S.R.M.: Characterizing glottal activity from speech using empirical mode decomposition. In: National Conference on Communications 2015 (NCC-2015). Mumbai, India (2015)
57. Sharma, R., Vignolo, L., Schlotthauer, G., Colominas, M., Rufiner, H.L., Prasanna, S.: Empirical mode decomposition for adaptive am-fm analysis of speech: A review. Speech Communication **88**, 39 – 64 (2017). DOI http://dx.doi.org/10.1016/j.specom.2016.12.004. URL `//www.sciencedirect.com/science/article/pii/S0167639316302370`
58. Smits, R., Yegnanarayana, B.: Determination of instants of significant excitation in speech using group delay function. Speech and Audio Processing, IEEE Transactions on **3**(5), 325–333 (1995)
59. Sreenivasa Rao, K., Prasanna, S., Yegnanarayana, B.: Determination of instants of significant excitation in speech using hilbert envelope and group delay function. Signal Processing Letters, IEEE **14**(10), 762–765 (2007)
60. Stylianou, Y.: Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Transactions on speech and audio processing **9**(1), 21–29 (2001)
61. Talkin, D.: A robust algorithm for pitch tracking (rapt). Speech coding and synthesis **495**, 518 (1995)
62. Thomas, M.R., Gudnason, J., Naylor, P.A.: Data-driven voice soruce waveform modelling. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pp. 3965–3968. IEEE (2009)

63. Thomas, M.R., Gudnason, J., Naylor, P.A.: Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. Audio, Speech, and Language Processing, IEEE Transactions on **20**(1), 82–91 (2012)
64. Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P.: A complete ensemble empirical mode decomposition with adaptive noise. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4144–4147. IEEE (2011)
65. Varga, A., Steeneken, H.J.: Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech communication **12**(3), 247–251 (1993)
66. Wang, G., CHEN, X.Y., Qiao, F.L., Wu, Z., Huang, N.E.: On intrinsic mode function. Advances in Adaptive Data Analysis **2**(03), 277–293 (2010)
67. Wong, D., Markel, J., Gray, A.: Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transactions on Acoustics, Speech, and Signal Processing **27**(4), 350–355 (1979)
68. Wu, J.D., Tsai, Y.J.: Speaker identification system using empirical mode decomposition and an artificial neural network. Expert Systems with Applications **38**(5), 6112–6117 (2011)
69. Wu, Z., Huang, N.E.: A study of the characteristics of white noise using the empirical mode decomposition method. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **460**(2046), 1597–1611 (2004)
70. Wu, Z., Huang, N.E.: Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in adaptive data analysis **1**(01), 1–41 (2009)
71. Yegnanarayana, B., Gangashetty, S.V.: Epoch-based analysis of speech signals. Sadhana **36**(5), 651–697 (2011)
72. Yeh, J.R., Shieh, J.S., Huang, N.E.: Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. Advances in Adaptive Data Analysis **2**(02), 135–156 (2010)