Detection of Glottal Closure Instants based on the Microcanonical Multiscale Formalism

Vahid Khanagha, Khalid Daoudi, Hussein Yahia

Detection of Glottal Closure Instants based on the Microcanonical Multiscale Formalism

Vahid Khanagha, Khalid Daoudi, Hussein Yahia
INRIA Bordeaux Sud-Ouest (GEOSTAT team), 200 Avenue de la vielle tour, 33405 Talence, France.
Email: khanagha@umd.edu, {khalid.daoudi, hussein.yahia}@inria.fr
Webpage: http://geostat.bordeaux.inria.fr/

***Abstract*—This paper presents a novel algorithm for automatic detection of Glottal Closure Instants (GCI) from the speech signal. Our approach is based on a novel multiscale method that relies on precise estimation of a multiscale parameter at each time instant in the signal domain. This parameter quantifies the degree of signal singularity at each sample from a multi-scale point of view and thus its value can be used to classify signal samples accordingly. We use this property to develop a simple algorithm for detection of GCIs and we show that for the case of clean speech, our algorithm performs almost as well as a recent state-of-the-art method. Next, by performing a comprehensive comparison in presence of 14 different types of noises, we show that our method is more accurate (particularly for very low SNRs). Our method has lower computational times compared to others and does not rely on an estimate of pitch period or any critical choice of parameters.**

***Index Terms*—Detection of Glottal Closure Instant, non-linear speech analysis, multiscale signal processing.**

I. Introduction

According to the aerodynamic theory of voicing, during the production of a voiced sound, a stream of breath flows through the glottis and creates a push-pull effect on the vocal fold tissues that results in self-sustained oscillation of the vocal cords [1]. The push occurs during glottal opening when the glottis is convergent, whereas the pull occurs during glottal closing when the glottis is divergent. During glottal closure, the air flow is cut off until breath pressure pushes the folds apart and the flow starts up again, causing the cycles to repeat [1]. As such, the steady flow of air from the lower respiratory system is converted into a periodic train of flow pulses [2]. These glottal pulses form the actual excitation source of the vocal tract. However, to a first approximation, the significant excitations of the vocal tract systems can be considered to occur at discrete instants of time (epochs) [3]. There can be more than one epoch during a pitch period, but the major excitation usually coincides with the Glottal Closure Instants (GCIs) [4]. Indeed, once the vocal folds became sufficiently close, the Bernoulli force causes an abrupt closure, which in turn results in an abrupt excitation of the vocal tract system [5].

Precise detection of GCIs has found many applications in speech technology: closed phase Linear Prediction (LP) analysis [4], [6], [7], [8], pitch synchronous speech processing for converting the pitch and duration of speech [9], prosody modification [10], synthesis [11], [12], dereverbertion [13], casual-anticasual deconvolution [14], [15] and glottal flow estimation [16].

GCIs can be reliably detected using a simultaneously recorded Electro-Glotto-Graph (EGG) signal, which provides a non-invasive measurement of the vibratory motion of the vocal folds. Positive peaks of the differentiated EGG (dEGG) in each pitch period can be taken as GCI [17]. However, as an EGG device is not always available, there is a great interest in extracting them directly from the speech signal.

In this paper we present a novel GCI detection algorithm based on a novel multiscale formalism that we have been recently adapting for speech analysis [18], [19]. The formalism is called the Microcanonical Multiscale Formalism (MMF) and is centered on precise estimation of local quantities at each time instant in the signal domain that are called Singularity Exponents (SE). SEs provide a quantitative evaluation of singular behavior at different scales and as such, their values can be used to identify the geometrical structures (subset of signal samples) that share the same singular behavior (how the signal values vary w.r.t to their neighboring samples at different time resolutions, scales).

Of the highest interest is the subset of samples whose values of SEs indicates the existence of a consistent [highly] singular behavior across different scales. This set is called the Most Singular Manifold and is shown in some applications that it highlights the most informative subset of signal samples, in the sense that the whole image can be reconstructed using exclusively the

information carried by these samples [20]. For detection of GCIs from the speech signal, we argue that the impulsive excitation at GCIs produces consistent patterns of singularities at different scales of the speech waveform and hence, one can expect the MSM to give access to these time instants. We thus use the concept of MSM to develop a simple and efficient GCI detection algorithm.

The particularity of the algorithm is that it extracts GCIs directly from the speech signal without any limiting assumption (like the known polarity of the speech waveform) and does not require any information about the pitch period. We compare our algorithm with a recent method called SEDREAMS [21] that extracts GCIs by detection of discontinuities in residuals of LP analysis. SEDREAMS is recently shown to have the best of performances compared to several state-of-the-art methods, while being among the fastest methods [22]. We first show that in the case of clean speech, our method has very close performance to SEDREAMS and then, we perform an extensive comparison in the presence of fourteen different types of noises and we particularly report the results for very low values of SNR to compare the robustness of the two algorithms against these noises. The results show that our method is as reliable as SEDREAMS while it is more accurate in low SNR noisy situations.

This paper is organized as follows: section II provides a review on available GCI detection methods. In Section III we present the basic concepts of MMF and introduce the procedures for computation of SEs and formation of the MSM. In section IV, we introduce the details of our GCI detection algorithm. The experimental results are presented in section VI and finally in section VII, we draw our conclusion and perspectives.

II. REVIEW OF GCI DETECTION METHODS

Several algorithms have been developed to extract GCIs directly from the speech waveform that use different criteria for localization of these events. A common criterion is the occurrence of large values in the residual signal of Linear Prediction (LP) analysis [23], [4], [22]. LP residuals indeed show clear and strong peaks around GCIs, but they are vulnerable to noise (and reverberation [13], [8]). Moreover, it is known that LP residuals may contain peaks of random polarities [3]. There are however several other criteria that have been employed for GCI detection: the amount of frequency deviation of the zero-frequency filtered speech signal from the central frequency [3], Lines of Maximum Amplitudes (LoMA) of the wavelet transform [24], [25], discontinuities in an estimate of the voice source signal [5] and zero crossings of the slope function of phase spectrum (the unwrapped phase function of the short time Fourier transform) of LP residuals [26], [8] are some examples of these criteria.

All these criteria can be used to detect the GCIs. However, there is always the risk of false alarms or missed GCIs. The rate of false alarms and misses vary depending on the choice of algorithmic parameters like thresholds or the size of the analysis windows. The challenge is thus to design an algorithm that makes one and only one detection for each glottal cycle that is close enough to the actual GCI. For instance, the effect of the choice of window-size for using of zero crossings of phase spectrum slope is addressed in [8]: Ideally, the window should span exactly one pitch period. A large window covers more than one GCI and hence the zero crossing occurs in mid-way between two GCIs. A small window increases the number of false alarms, as spurious zero-crossings would occur in a window that does not contain any GCI. Consequently, for the DYPSA method [8], a moderately small window is employed to minimize the risk of having two GCIs in a single window. All the zero-crossings (each taken from one of these small analysis windows) are taken as candidates (plus some additional candidates taken from a phase-slope projection technique). The true GCIs are then found from the list of candidates, using an N-best Dynamic Programming. The cost function is defined in a way that its minimization maintains several desired properties in the final output sequence of the algorithm. For instance, based on the assumption of smooth variations of pitch over short segments, major pitch deviations are penalized heavily in the cost function (although the method does not require a supplemental pitch estimator). So in effect, for each pitch period, only one of the candidates is picked, which is the one providing the maximum consistency in terms of pitch-period variation. The same idea of DP is employed in YAGA to refine candidates which are taken by detection of discontinuities in an estimate of the voice source signal [5].

In [24] for the use of LoMA criterion, a coarse estimation of the fundamental frequency (F_0) is made to select the largest scale containing the F_0 and then, a local DP technique uses the pitch information to select only one LoMA per pitch period as the GCI. Finally, two heuristics are applied to reduce the errors corresponding to detection of more than one GCI per pitch period.

A simpler solution for controlling the GCI detection rate is employed in [3], [22] by the use of a smoothed mean-based signal that oscillates with the pitch period and assists in keeping the rate of detection to one detection per pitch period. The SEDREAMS method [22],

[21] uses the mean-based signal to localize a small interval in each pitch period that contains the GCI. This smoothing procedure provides robustness against noise but may compromise the accuracy. That is why in SEDREAMS [21] LP residuals are used for final localization of GCIs in each interval (which is found using the noise-robust mean-based signal). As for the use of LP residuals, SEDREAMs requires an estimation of their polarity to decide about the sign of peaks it takes as GCIs.

So in conclusion, the performance of a GCI detection method depends on two factors: the accuracy of the criterion used for localization of GCIs and the way of refining true GCIs from spurious candidates (especially in noisy scenarios) to control the number of detections per pitch period. Some algorithms use estimates of the pitch period so as to restrict their search-space to one pulse per period. This adds up to complexity of the algorithm while risking the accuracy. We take all these considerations into account to develop a simple and efficient algorithm that does not require any additional information (pitch or polarity) using a novel non-linear formalism that is introduced in the next section.

III. The Microcanonical Multiscale Formalism

Our GCI detection algorithm is based on a novel framework called the Microcanonical Multiscale Formalism (MMF) [27]. MMF allows the study of local geometrico-statistical properties of complex signals from a multiscale perspective. It is based on precise computation of local parameters called the Singularity Exponents (SE) at every time instant in the signal domain. SEs are local quantities that quantify the degree of regularity of the signal at each time instant. When correctly defined and estimated, these exponents alone can provide valuable information about local dynamics of complex signals and have recently proven their strength in many signal processing applications ranging from signal compression to inference and prediction in a quite diverse set of scientific disciplines such as satellite imaging [28], [29], [30], [31], adaptive optics [32], [33], computer graphics [34] and natural image processing [20], [35].

The singularity exponent $h(t)$ of a given d-dimensional signal $s(t)$ can be estimated by evaluating the power-law scaling behavior of a scale-dependent functional Γ_r over a set of fine scales r:

$$\Gamma_r\left(s(t)\right) \;=\; \alpha(t)\, r^{d+h(t)} + \mathrm{o}\left(r^{d+h(t)}\right) \qquad r \to 0 \quad (1)$$

where $\Gamma_r\left(\cdot\right)$ can be any multiscale functional complying with this power-law, whose choice is discussed in section III-A. The term $\alpha(t)$ is a factor that does not depend on the scale r and $\mathrm{o}\left(r^{d+h(t)}\right)$ means that for small scales the additive terms are negligible compared to the factor and thus $h(t)$ dominantly quantifies the multiscale behavior of the signal at the time instant t. Indeed, the value of $h(t)$ can be used to recognize geometrical superstructures (subsets of signal samples) inside the signal that share common multiscale properties. The most important among these sets is called the Most Singular Manifold (MSM), which is defined as the set of signal samples having the smallest SE values. Indeed for a given signal sample, the smaller the value of SE is, the higher unpredictability is at this time instant [27] (in the sense that the signal value at this point can not be predicted using its neighboring signal samples). It has been established that the critical transitions of the underlying physical system occur at these instants of time, and this fact has been used in many signal processing applications [30], [36]. It is shown that in case of natural images, the MSM contains the most informative subset of signal samples in the sense that the whole image can be reconstructed using only the information carried by the MSM [20]. MSM highlights a small subset of signal samples, from which the original image can be reconstructed by application of a reconstruction kernel. In this paper we will use the concept of MSM to access some of the important dynamical events in the speech signal (GCIs). We first discuss practical considerations in estimation of SEs for the case of speech signal and for the particular application of GCI detection.

A. The choice of $\Gamma_r\left(\cdot\right)$

An important aspect in the MMF is the choice of $\Gamma_r\left(\cdot\right)$ such that the inter-scale power-law correlations of the form presented in Eq. (1) are revealed. This multiscale measure can be simple linear increments (Hölder exponents), wavelet transform of the signal [37], the gradient-modulus measure or several other measures introduced in [27]. Depending on how any one of these $\Gamma_r\left(\cdot\right)$ cope with particularities of real world signals such as discretization, noise and long-range correlations, they may have their own benefits and disadvantages [37].

For the case of speech signals, we initially followed a similar path that is taken for natural images [27]. With the goal of achieving the most compact MSM from which the whole image can be reconstructed, a new $\Gamma_r\left(\cdot\right)$ was defined in [27] that is based on the local evaluation of a reconstruction kernel. In fact, $\Gamma_r\left(\cdot\right)$ is defined in a way that it penalizes predictability: the samples that

cannot be reconstructed from the information carried by their neighbors (i.e. they are less predictable) attain lower values of SE and thus they belong to the MSM. As such, the concept of MSM is linked to a local notion of predictability and, indeed, the resulting MSM is shown to be the most compact subset of image samples from which the whole image can be reconstructed. However, For the case of speech signals, the direct 1-D adaptation of the same procedure reduces the resulting $\Gamma_r(\cdot)$ to simple directional finite differences. In [38], we followed a similar path and searched for a $\Gamma_r(\cdot)$ that results in a relatively more compact MSM from which the whole speech signal can be reconstructed; we used a classical method for reconstruction of a given signal from a subset of its irregularly spaced samples (MSM in our case) and compared various definition of $\Gamma_r(\cdot)$ to find the one that results in a more compact MSM from which the signal can be reconstructed with good perceptual quality (evaluated using the PESQ measure of signal quality). As such, the multi-scale integral of the following scale-dependent functional was defined:

$$\Gamma_{r_i}(s(t)) = \mid 2s(t) - s(t - r_i) - s(t + r_i) \mid \quad (2)$$

We discussed in [38] that such definition reduces the effect of inter-sample correlations of the speech signal in estimation of SEs and we showed that it effectively results in a compact representation of the speech signal. On the other hand, the GCI detection application that we are considering in this paper allows us to provide an intuitive justification for this multiscale measure. Indeed, $\Gamma_r(\cdot)$ of Eq. (2) can be seen as a multiscale measure of the magnitude of an impulse. Assuming an ideal impulse of amplitude A at time t ($A\delta(t)$), this measure constantly returns the value of $2A$ at this time instant and for all different scales r_i. Considering the analogy between *scale* and *frequency*, this conforms with the frequency response of an impulse which is a discontinuity reflected over the whole spectral band (a fact that is exploited in [3] for GCI detection by zero-frequency filtering of speech). Note that for the GCI detection algorithm we look for a synergy of these measurements at different scales and hence, it is important that for different scales r the peak values of $\Gamma_r(\cdot)$ measurements of $A\delta(t)$ happen simultaneously at time n (or with exactly the same amount of delay) and the use of a directional (non-symmetric) measurement or a wavelet transform that has different group delays at different scales can not fulfill this requirement.

Of course we do not observe such ideal impulses at GCIs in the speech waveform and we only observe a filtered (smoothened) version of the hypothesized impulsive excitation. In other words, a portion of full-band spectral presence of excitation impulses are filtered out by the time varying vocal tract filter. However, since we evaluate several scales simultaneously, we can fairly expect the remaining spectral energy of the filtered impulses to be sufficient to exhibit a distinguishable singular behavior compared to other non-GCI signal samples where no impulse is present in the excitation.

B. Estimation of the singularity exponents

Once Γ_r is specified, $h(t)$ can be estimated from Eq. (1) by a $log - log$ regression of Γ_r versus the scales r. However, for the specific application of GCI detection we opted for the estimation method that is theoretically motivated in [36] by assuming the existence of a particular geometric multiplicative cascading behavior (implying that the measurements $\Gamma_r(\cdot)$ at different scales are directly related by a multiplicative factor that is independent of the scale r). The final estimation is performed as the sum of a set of *partial* SEs:

$$h(t) = \sum_{i=1}^{I} h_i(t) \quad (3)$$

where I is the number of scales that we use and $h_i(t)$ are the *partial* SEs which are computed by evaluation of Eq. (1) at each scale (assuming that $\alpha(t)$ is a factor that does not depend on scale):

$$h_i(t) = \frac{log(\Gamma_i(s(t)))}{log(r_i)} \quad (4)$$

This method of computation of SEs is particularly suited for GCI detection application. Indeed, the summation of partial singularities superimposes the amount of singular behaviors at different scales and as such, can reveal significant impulsive excitations that simultaneously appear over several scales. As mentioned earlier, we expect the excitation impulses at GCIs to produce *strong* local singularities at different scales of the speech waveform. This leads to simultaneous occurrence of highly negative partial SEs (Eq. (4)) at different scales around GCIs. The summation of these partial SEs (Eq. (3)) would thus result in significantly lower negative values and hence, those samples would belong to the MSM (defined as the subset of samples having the lowest values of SEs).

In practice, the finest accessible scales are dictated by the sampling frequency f_s of the discrete time signal $s[n] = s(\frac{n}{f_s})$. Consequently, Eq. (2) can only be evaluated at discrete instants of time $\frac{n}{f_s}$ with the scales r_i being

$$r_i = i/f_s\,,\; i \in [1 \cdots I] \quad (5)$$

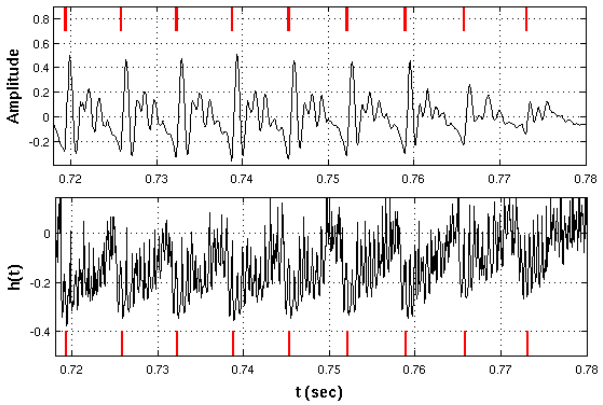

Fig. 1: **top:** A voiced segment of the speech signal "arctic_a0001" from CMU ARCTIC database [39] and **bottom:** the singularity exponents h[n]. Reference pitch marks are shown by red lines.

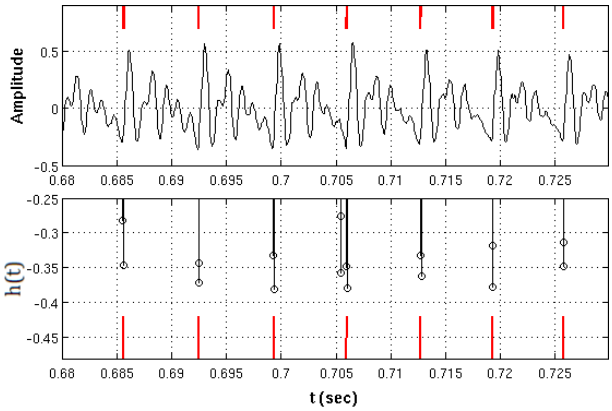

Fig. 2: **top:** A voiced segment of the speech signal "arctic_a0001" of the male speaker BLD from the "CMU ARCTIC" database [39]. **bottom:** MSM samples and their corresponding SE values. Reference pitch marks are shown by vertical red lines.

where I defines the largest scale that we use for computation of SEs. As such, the SEs are being estimated as a discrete time signal $h[n] = h(\frac{n}{f_s})$.

IV. The relevance of MSM to the GCIs

In this section we study the significance of the MSM w.r.t. to identification of GCIs. We start our study by a simple observation on the correspondence of the reference GCIs extracted from differentiated EGG signal with the negative peaks of $h[n]$. Fig. 1 shows a part of a voiced sound (top panel) and the corresponding SE values (bottom panel). The reference GCIs that are extracted from the differentiated EGG signal are also shown. It can be seen that $h[n]$ shows a sudden negative peak around GCIs.

Fig. 2 shows another example that confirms the intuition about the correspondence of the MSM with GCIs. The top panel shows another segment of a voiced sound along with its corresponding GCIs. The bottom panel shows the SE value of the samples belonging to the MSM. The MSM is formed as the 5% of samples having the lowest value of SE. It can be seen that MSM signal samples are indeed located around the reference GCI. Note also that around every single GCI, the MSM point with the lowest SE value is the closest one to the GCI mark. This example shows that MSM can indeed be considered as a criterion for localization of GCIs and can be used for development of an automatic GCI detection algorithm as presented in the following section.

V. A MSM-based GCI detection algorithm

The preliminary observations presented in section IV showed that MSM effectively points to GCIs. However, care must be taken for development of an automatic GCI detection algorithm so that the detection rate is maximized while false alarms are minimized. Indeed, practical formation of the MSM requires the specification of a threshold to be applied to singularity exponent values ($h[n]$). A global specification of the threshold would not be the best choice. Indeed, it may happen that a GCI point does attain a small $h[n]$ compared to its surrounding signal samples in one pitch period, but in a larger neighborhood, it may have higher value even compared to non-GCI samples. In practice, this may especially occur for the starting and ending segments of a voiced sound, where the energy of the signal is lower compared to the central segments. This case can be observed in Fig. 1 where, for the last two GCI marks, SE is actually smaller than the immediate neighbors but it is larger than many other non-GCI samples in a larger neighborhood that covers the whole segment. Hence, the application of a global threshold for the whole segment cannot refine GCIs from non-GCI samples. Another practical issue is that the presence of noise may cause the location of the samples having the smallest SE values to be shifted from the desired GCI locations.

To overcome these issues, we first mention that GCIs can be identified using two properties of SEs that can be observed in Fig. 1:

1) **correspondence of the MSM to GCIs**: in each

glottal cycle, $h[n]$ has the smallest value at the GCI. Particularly for the case of clean speech, the location of a*local* minimum usually coincides with the the GCI.

2) **regularity-drop**: there is a sudden fall in SE values right before each GCI. As such the local averages of SEs before and after each GCI are significantly different.

Our experiments showed that the first property indeed provides highly precise GCI localization in high-energy segments of clean speech, within a single pitch period. However as mentioned before, in a larger window, the GCIs at low-energy parts of speech may attain relatively higher values compared to the non-GCIs belonging to the high-energy parts. Also, the presence of noise may cause a GCI point to attain slightly higher values compared to its immediate neighbors. The second property on the other hand is not much affected by the local energy as the change in local averages (regularity-drop) is a relative quantity. Hence, this property seems more suitable for GCI detection in segments with lower energy. Also, even the presence of noise would not drastically affect this property as it is related to the average of SEs rather than their individual values. In order to make explicit and easy use of this property, we define the regularity-drop functional $\mathcal{D}_L[n]$ that measures the change in local average of SEs before and after any time instant n, on two windows of length T_L:

$$\mathcal{D}_L[n] = \sum_{k=n-T_L}^{n-1} h[k] - \sum_{k=n}^{n+T_L} h[k] \qquad (6)$$

Fig. 3 illustrates the resulting functional for a segment of voiced speech along with the reference GCIs. It can be seen that $\mathcal{D}_L[n]$ oscillates with the pitch period. In that sense, this is similar to the mean-based signal in [21] and can be similarly used to limit the search space for GCIs at each glottal cycle (to reduce false alarms). Indeed, the reason is that the regularity-drop of $h[n]$ occurs once per pitch period, and since it happens at GCI, $\mathcal{D}_L[n]$ displays a peak at this point. Also, each peak is preceded by a positive-going zero-crossing and is followed by a negative-going zero-cross. As these zero-crossings can be easily detected without ambiguity, they can be used for limiting the search space to positive half-periods of $\mathcal{D}_L[n]$ and consequently to reduce false alarms. Moreover, as the definition of $\mathcal{D}_L[n]$ involves time-domain averaging, it provides robustness against uncorrelated noise like white noise. Another advantage of the regularity-drop functional is due to its differential form that provides robustness against low-frequency correlated noises (such as car noise) that cause low-frequency shifts in SE values.

We use all these properties to develop an efficient GCI detection algorithm. We use the MSM as the primary criterion pointing to the GCI and we use $\mathcal{D}_L[n]$ to limit the search space while its peaks are also used as a secondary criterion for localization of GCIs. The final implementation is provided in Algorithm 1[1].

Algorithm 1 : GCI detection

1: Calculate $h[n]$ and $\mathcal{D}_L[n]$.
2: In $\mathcal{D}_L[n]$, for any positive-going zero-cross time instant n_{pos}, find the next negative-going zero-cross n_{neg}.
3: $n_{peak} \leftarrow \underset{n}{\operatorname{argmax}}\, \mathcal{D}_L[n], \quad n \in [n_{pos}, n_{neg}]$.
4: MSM formation: take n_1, n_2, n_3 having the lowest values of $h[n]$ in $n \in \{n_{pos}, n_{neg}\}$.
5: $n_{msm} \leftarrow \underset{n_i}{\operatorname{argmin}} |n_i - n_{peak}|$
6: $n_{gci} \leftarrow [(n_{peak} + n_{msm})/2]$

Note that in step 4 of the algorithm, we take three samples with the lowest value of singularity exponent so as to cope with noisy scenarios where $h[n]$ at GCI may be slightly higher than one or two of immediate neighbors. That is why the criterion of closeness to the peak of $\mathcal{D}_L[n]$ is used in step 5, to make the final decision. Using three out of eighty samples in each frame indeed conforms with the definition of the MSM: a small subset of samples having lowest SE values, while the size of the subset is chosen experimentally.

Indeed, $\mathcal{D}_L[n]$ is not simply used for constraining the detection to one detection per period, but rather, as its peak is expected to be located on GCI it is also contributing to an increase of accuracy.

VI. Experimental results

We compare our algorithm against a recent method called SEDREAMS [21], which is shown in [22] to provide the best of performances compared to several state-of-the-art methods, for both clean and noisy speech signals. We use the implementation that is made available on-line by its author [40]. We are basing our experimental protocol on [21] to compare the performance of our algorithm in clean and noisy situations. However, to make a more comprehensive evaluation, we use 14 different types of noises rather than four noises as in [21]. We test our algorithm on the CMU ARCTIC databases, which consist of 3 sets of 1150 phonetically balanced

[1]A Matlab implementation of this algorithm is made publicly available in http://geostat.bordeaux.inria.fr/

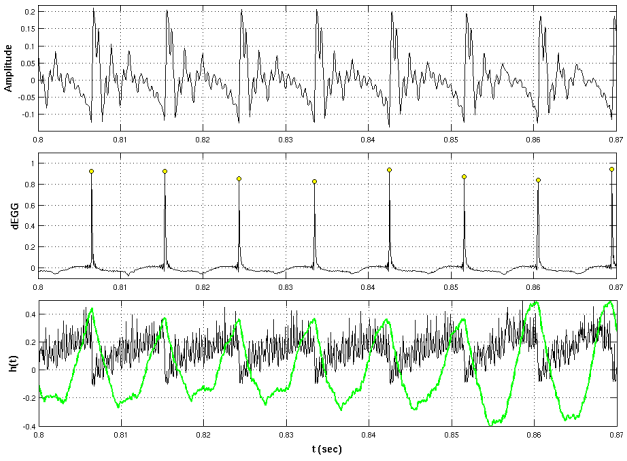

Fig. 3: **top:** A voiced segment of the speech signal taken from KED database. **middle:** the differenced EGG signal which serves for extraction of reference GCIs. The peaks are marked with yellow circles as the reference GCIs. **bottom:** singularity exponents are shown by black color and the regularity-drop functional $\mathcal{D}_L[n]$ is shown by green color.

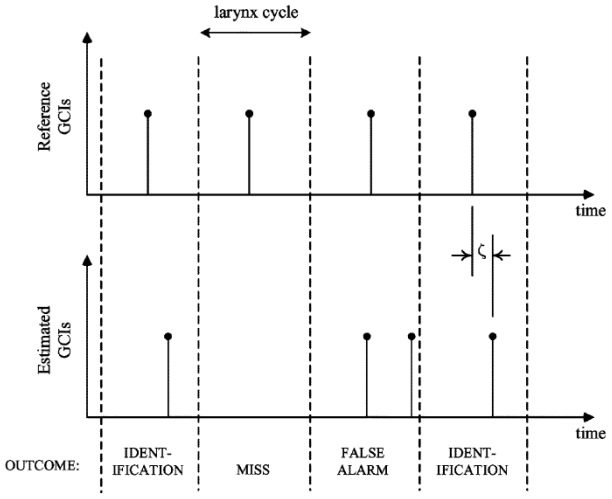

Fig. 4: Graphical representation of 4 different outcomes of the algorithm for any given larynx cycle. Identification accuracy is measured by ζ (the graphical representation is taken from [8]).

sentences, each uttered by a single speaker: BDL (US male), JMK (US male) and SLT (US female) [41]. We also test on the KED Timit database, which contains 453 utterances spoken by a US male speaker. All these freely available [41] datasets are chosen because they contain contemporaneous EGG recordings that can be used to extract reference GCIs. Note that the EGG signal is synchronized to the speech recordings such that SEDREAMS performance is maximized.

SE values are estimated using Eq. (3) with $I = 7$ and $\mathcal{D}_L[n]$ is computed with $T_L = 2.5ms$. Indeed, T_L must be smaller than the local pitch period for different speakers so as to avoid merging of two GCIs. On the other hand, the larger it is, the higher the robustness would be against uncorrelated noises like white noise. $T_L = 2.5ms$ is a reasonable choice that is smaller than the pitch periods of the speakers with pitch frequencies as high as 400 Hz. The outputs are only evaluated for the voiced parts of the speech signals. We use a separate Voiced/UnVoiced detector to compare and evaluate the different algorithms during the voiced parts of the speech signal.

A. Performance measures

We use the set of performance measures defined in [8] to evaluate the performance of our method. If the k-th reference GCI occurs at n_k, the corresponding larynx cycle can be defined as the range of samples $n \in (\frac{n_k+n_{k-1}}{2}, \frac{n_k+n_{k+1}}{2})$. Consequently, two sets of performance measures are defined using the graphical representation in Fig. 4. The first set includes three measures of the *reliability* of the algorithms:

- Hit Rate (HR): the percentage of larynx cycles for which exactly one GCI is detected.
- Miss Rate (MR): the percentage of larynx cycles for which no GCI is detected.
- False Alarm Rate (FAR): the percentage of larynx cycles for which more than one GCI is detected.

The second set defines two measures of the *accuracy* of the algorithms:

- Accuracy to ± 0.25 ms (A25 ms): the percentage of larynx cycles for which exactly one GCI is detected and the identification error ζ is within ± 0.25 ms.
- Identification Accuracy (IDA): the standard deviation of identification error ζ (the timing error between the reference GCIs and the detected GCIs in larynx cycles for which exactly one GCI has been detected).

B. Clean speech

Table I compares the performance of the two GCI detection methods for clean speech signals. Overall, it can be seen that SEDREAMS is more reliable, but the accuracy of the two methods are the same. Fig. 5 shows histograms of GCI detection timing error ζ for the two

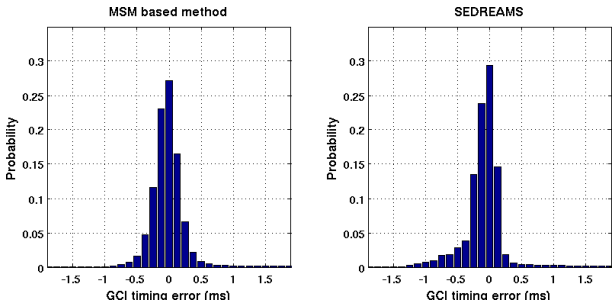

Fig. 5: Histogram of GCI detection timing error ζ. A reference GCI is considered to be correctly detected when exactly one detection has happened for the corresponding larynx cycle.

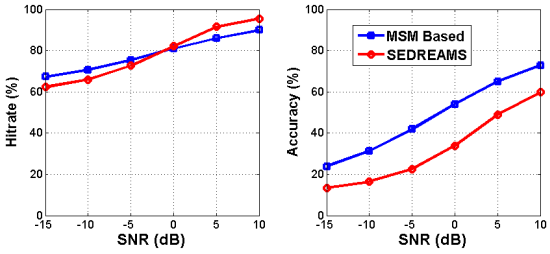

Fig. 7: Comparison of performances averaged over all the 14 types of noises taken from NOISEX database [42].

algorithms (over the whole four datasets). It can be seen that the distribution of timing error in identification of GCIs is almost the same.

C. Noisy speech

To assess the performance of our algorithm in more realistic scenarios, we evaluate its robustness against 14 different types of noises taken from the NOISEX-92 database [42]. Fig. 6 shows the results in the presence of different types of noises. To make the comparison easier, we only show two performance measures: the Hit Rate (HR) as a measure of reliability and the Accuracy to ± 0.25 ms as a measure of accuracy. It can be seen that in terms of reliability (Hit Rate), SEDREAMS overperforms in cases of white noise, Babble noise and destroyer engine noise. However, the MSM based method is more reliable in the presence of car interior noise, factory floor noise, Leopard military car noise and tank noise. For the remaining 7 types of noises, the reliability of the two methods is quite close, while SEDREAMS shows slightly better results, especially for higher SNRs. However, in terms of accuracy, the MSM based method is showing significantly higher performance for all the 14 types of noises. The higher accuracy of our method can be seen in Fig. 7, which shows the averaged results over all 14 types of noises.

SEDREAMS reliability can be explained by the adaptive control of the window length with a rough estimation of the average pitch period. This permits the algorithms to smoothen the signal as much as possible. That is why SEDREAMS shows much more reliable results in presence of an uncorrelated noise like white noise. We have however avoided assuming any prior knowledge about the average pitch period and instead chosen a reasonable fixed value of $2.5ms$ for T_L.

The more accurate result of our MSM-based algorithm compared to SEDREAMS can be explained by the difference between the regularity-drop function in our method and the mean-based signal used in SEDREAMS (both of them are used to constrain the number of detections in each pitch period to one). Apparently both of these functionals serve a similar goal to increase the *reliability* of the algorithms. However, the regularity-drop functional $\mathcal{D}_L[n]$ has two distinctive features that contribute not only to improve the *reliability* of our algorithm but also serve to improve the *accuracy* of it: first, its peak is located on the GCI and hence, it is a smooth (noise-robust) pointer to the GCI. The second difference is that $\mathcal{D}_L[n]$ is a relative quantity that results in its lower sensitivity to long-range correlations due to low-frequency noises like car-noise. It must be noted, however, that the high accuracy of the algorithm in localization of GCIs is mainly attributed to the high accuracy of SEs in localization of highly singular events in the signal domain

D. Analysis of algorithm parameters

In this section, the effect of change in algorithmic parameters is studied. Namely, we study the variation of window length (T_L) of the regularity drop functional ($\mathcal{D}_L[n]$) and the number of scales (I) that are used for computation of SEs using Eq. (3). The experiments are performed on 350 clean speech files randomly selected from the above databases. To study the case of noisy speech, for each selected clean speech file, we randomly select one of four different types of noises (white, babble, Volvo and factory) and add it to the clean file with the SNR of 0 dB.

Fig. 8 shows how the change in T_L affects the performance of the algorithm. As expected, the performance is significantly reduced for small window lengths but for $T_L > 2msec$ the algorithm almost performs in a consistent way for both noisy and clean scenarios (note that, especially for the noisy case there is an improvement

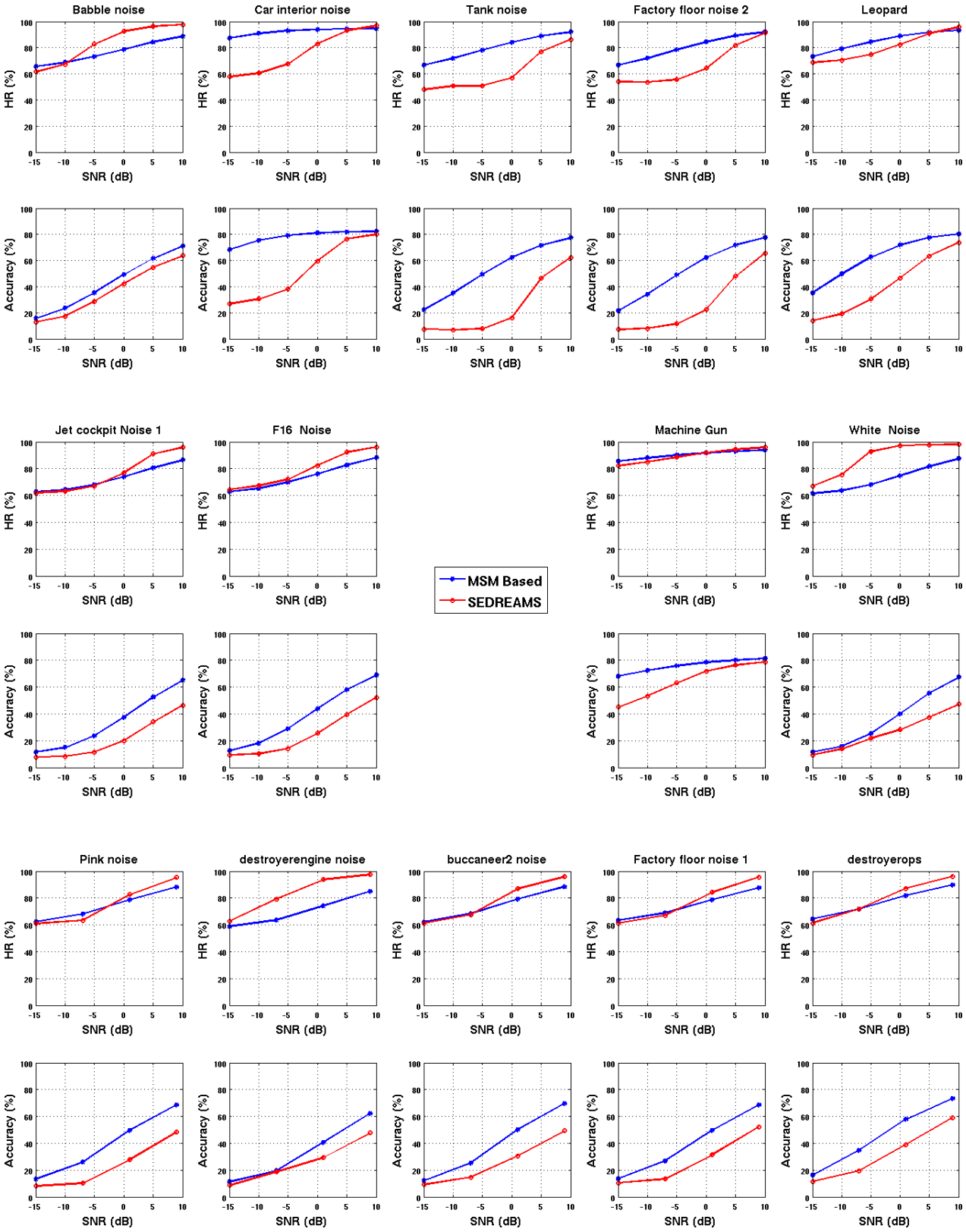

Fig. 6: Performance comparison in the presence of 14 different types of noises taken from the NOISEX database [42].

TABLE I: The comparative table of GCI detection performances for clean speech signals.

BDL dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25 ms (%)
MSM-based	94.7	2.7	2.5	0.54	79.5
SEDREAMS	97.4	0.85	1.7	0.38	85.43
JMK dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25 ms (%)
MSM-based	94.9	1.38	3.6	0.55	85.5
SEDREAMS	97.8	0.52	1.6	0.53	78.9
SLT dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25 ms (%)
MSM-based	94.1	4.42	1.4	0.39	80.91
SEDREAMS	98.3	0.02	1.6	0.31	80.25
KED dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25 ms (%)
MSM-based	97.4	1.07	1.5	0.39	96.24
SEDREAMS	98.8	0.05	1.14	0.34	94.33
Overall results for four speakers:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25 ms (%)
MSM-based	95.5	2.3	2.2	0.48	82.3
SEDREAMS	98.0	0.4	1.6	0.39	82.5

in False Alarm Rate for larger values of T_L but the improvement comes with the cost of increasing Miss Rates. The Hit Rate however, is almost constant in those regions, which means that the amount of useful GCI detections is almost the same). The results presented in sections VI-B and VI-E were obtained using the fixed value of $2.5ms$ which is smaller than the pitch period of most of ordinary speakers and as such, the two summations in Eq. (6) will only depend on SE values in a single pitch period.

Next, Fig. 9 shows the effect of change in number of scales (I) that are used in Eq. (3) to compute SEs. It is interesting to note how the performance is effectively improved by incorporation of more scales in computation of SEs (both for noisy and clean speech). At GCIs, where a glottal pulse appears across all scales, the co-existence of singular behavior at different scales (simultaneous occurrence of small *partial* SEs h_i of Eq. (4)) makes the final SE calculated by Eq. (3) to attain a very small value and hence, the sample will belong to the MSM and be detected as a GCI. However, at the non-GCI samples, even if for any reason a singular value appears at one of the scales (one small h_i) it will be canceled out by the larger h_i values at other scales, where the singularity does not appear (because there is no impulsive behavior to produce a consistent singular behavior across all scales). As such, the higher the number of scales are, the more amplification of true singular behavior will happen, while there will be more chance of canceling out the occasional

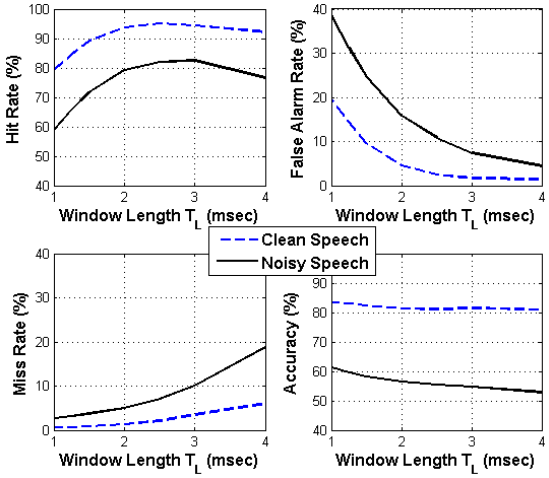

Fig. 8: The effect of variation of window length (T_L) of the regularity drop functional ($\mathcal{D}_L[n]$) on GCI detection performance (for $I = 7$).

appearance of small h_is at one of the scales for the non-GCI samples. The results presented in sections VI-B and VI-E, we use only the seven smallest scales ($I = 7$) to avoid excessive computations.

E. Computational complexity

We compare the computational complexity of our algorithm with that of SEDREAMS [22], which is shown

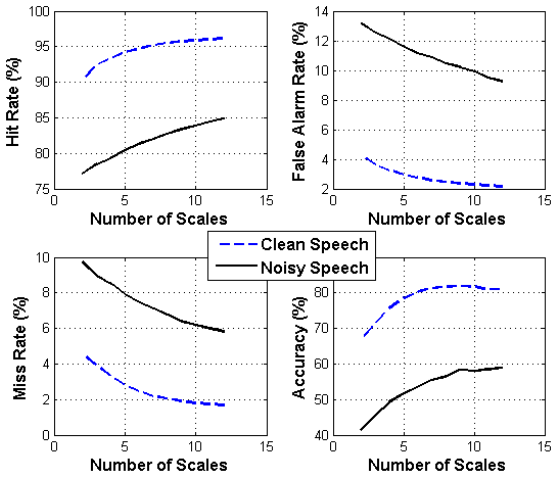

Fig. 9: The effect of variation of total number of scale used for computation of SEs, on GCI detection performance (for $T_L = 2.5msec$).

to be the most efficient algorithm compared to other state-of-the-art algorithms [22], [15]. The computational complexity of a GCI detection algorithm is indeed highly data-dependent and it is not easy to provide order of computation details as mentioned in [22], [15]. The computations of SEDREAMS can be divided into two major parts: computation of linear prediction residuals with computational complexity of $\mathcal{O}(N^2)$ (N being the total number of samples) and the formation of a so-called mean-based signal by averaging the windowed speech signal of length $1.75T_{0,mean}$, where $T_{0,mean}$ denotes the average pitch period of each speaker. As such for computation of the mean-based signal, the SEDREAMS requires $1.75T_{0,mean} \cdot F_s$ additions and the same number of multiplications (overall $3.5T_{0,mean}F_s$ operations) at each time instant.

Our MSM based method is similarly composed of two stages: the computation of singularity exponents with computational complexity of $\mathcal{O}(N)$ and the formation of the regularity-drop functional $D_L[n]$ which requires $2T_L \cdot Fs$ operations per sample. Considering that we are using $T_L = 2.5msec$ and assuming a speaker with $T_{0,mean} = 4msec$ we can see that the computation of $D_L[n]$ requires $5F_S/1000$ operations per sample whereas computation of the mean-based signal for SEDREAMS method requires $14Fs/1000$ operations per sample. If we also consider the lower computational complexity of the first stage ($\mathcal{O}(N)$ versus $\mathcal{O}(N^2)$), we can fairly conclude that our method is faster than SEDREAMS. We can also use an empirical metric called Relative Computation Time (RCT) as used in [22], [15] to compare the efficiency of these two algorithms in practice when applied to speakers with different average fundamental frequencies. The RCT is defined as:

$$RCT(\%) = 100.\frac{CPU\ time\ (s)}{Sound\ duration\ (s)} \qquad (7)$$

We compare the RCT of our mehtod with that of SEDREAMS. As for the SEDREAMS, we use the MATLAB codes that are made publicly available by its author in [40], where the original implementation is provided along with a faster implementation that uses a more efficient approach for computation of the mean-based signal. The computation times for these two different implementations of SEDREAMS (including the time for computation of residuals and detection of GCIs) are averaged over the whole database and reported in Table II, along with the RCT of our MSM based method. It can be seen that our method is almost 20 times faster than SEDREAMS. Also, if we compare with the fast implementation of SEDREAMS [22], [15], the MSM based method is 10 times faster. We underline however, that the results of sections VI-B and VI-E are obtained using the original implementation and not the fast one.

Finally, it is noteworthy that the overall processing delay of our GCI detection algorithm is about $3msec$ (a negligible delay for SE estimation plus $3msec$ for formation of $\mathcal{D}_L[n]$ and reaching of its negative-going zero-cross). This is much less than other methods which rely on dynamic programing techniques, the methods which use quantities like $T_{0,mean}$ whose computation require higher group delays or even from those who rely on the LP residual signal which is usually computed over a window of $25msec$ length. Together with the computational efficiency, this makes our method particularly suitable for real-time applications.

TABLE II: Comparison between the Relative Computation Time (RCT).

Method	RCT (%)
MSM-based	2.2
SEDREAMS [22]	43.8
fastSEDREAMS [22]	25.1

VII. Conclusion

In this paper we used a novel multiscale formalism called the MMF for development of a simple and efficient GCI detection algorithm. The MMF relies on precise estimation of local parameters called Singulrity Exponents (SE). We introduced the detailed procedure for estimation

of SEs for the case of speech signal and for the particular application of GCI detection. We showed that the subset of samples with lowest SE values (the MSM) indeed points towards the GCIs. We then used this property to develop an automatic GCI detection algorithm that compared to other methods, is more efficient and has less processing delay. The latter property makes the method appropriate for real-time implementations as it does not involve any type of batch processing. We showed that for clean speech signals our algorithm is almost as accurate and reliable as a recent stat-of-the-art method. But in presence of 14 different types of noises, and for very low SNRs, our method is more accurate. Moreover, this method does not rely on any model for speech production and does not require any estimate of the pitch period.

ACKNOWLEDGEMENT

This work were performed when the first author was a PhD student at INRIA Bordeaux-Sud Ouest, and was funded by the INRIA PhD Cordi-s program.

REFERENCES

[1] I. R. Titze, *The Myoelastic Aerodynamic Theory of Phonation*, The national center for voice & speech, 2007.

[2] B. H. Story, "An overview of the physiology, physics, and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23(4), pp. 195–206, 2002.

[3] K.S.R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1602–1613, 2008.

[4] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27 (4), pp. 309–319, 1979.

[5] M.R.P Thomas amd J. Gudnason and P.A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20 (1), pp. 82–97, 2012.

[6] E. N. Pinson, "Pitch synchronous time domain estimation of formant frequencies and bandwidths," *Journal of the Acoustical Society of America*, vol. 35 (8), pp. 1264–1273, 1963.

[7] K. Steiglitz and B. Dickinson, "The use of time-domain selection for improved linear prediction," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 25 (1), pp. 34–39, 1977.

[8] P. A. Naylor, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15 (1), pp. 34–43, 2007.

[9] T. Ewender and B. Pfister, "Accurate pitch marking for prosodic modification of speech segments," in *Proceedings of INTERSPEECH*, 2010.

[10] 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, 1990.

[11] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Interspeech conference*, 2010.

[12] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9 (1), pp. 21–29, 2001.

[13] N.D. Gaubitch and P.A. Naylor, "Spatiotemporal averagingmethod for enhancement of reverberant speech," in *15th International IEEE Conference on Digital Signal Processing*, 2007.

[14] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *ISCA Voice Quality: Functions, Analysis and Synthesis*, 2003.

[15] T. Drugman, *Advances in Glottal Analysis and its Applications*, Ph.D. thesis, University of Mons, 2011.

[16] D. Wong, J. Markel, and A. Jr. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350–355, aug 1979.

[17] A. Krishnamurthy and D. G. Childers, "Two channel speech analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34 (4), pp. 730–743, 1986.

[18] V. Khanagha, K. Daoudi, O. Pont, and Hussein Yahia, "A novel text-independent phonetic segmentation algorithm based on the microcanonical multiscale formalism," in *Proceedings of the INTERSPEECH*, 2010.

[19] Vahid Khanagha, Khalid Daoudi, Oriol Pont, and Hussein Yahia, "Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[20] A. Turiel and A. del Pozo, "Reconstructing images from their most singular fractal manifold," *IEEE Transactions on Image Processing*, vol. 11, pp. 345–350, 2002.

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

[22] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994 –1006, march 2012.

[23] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of Acoustic Society of America*, vol. 50 (2B), pp. 637–655, 1971.

[24] C. d'Alessandro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, pp. 601–622, 2011.

[25] N. Sturmel, C. d'Alessandro, and F. Rigaud, "Glottal closure instant detection using lines of maximum amplitudes (loma) of thewavelet transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, april 2009, pp. 4517–4520.

[26] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 3 (5), pp. 325–333, 1995.

[27] A. Turiel, H. Yahia, and C.J Pérez-Vicente., "Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, pp. 015501, 2008.

[28] J. Grazzini, A. Turiel, H. Yahia, and I. Herlin, "Edge-preserving smoothing of high-resolution images with a partial multifractal reconstruction scheme," in *International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2004.

[29] J. Grazzini, A. Turiel, and H. Yahia, "Multifractal Formalism for Remote Sensing: A Contribution to the Description and the Understanding of Meteorological Phenomena in Satellite Images," in *Complexus Mundi. Emergent Patterns in Nature*, Miroslav M. Novak, Ed., pp. 247–256. World Scientific Publishing Co. Pte. Ltd., 2006.

[30] H. Yahia, J. Sudre, C. Pottier, and V. Garçon, "Motion analysis in oceanographic satellite images using multiscale methods and the energy cascade," *Journal of Pattern Recognition*, 2010, to appear. doi:10.1016/j.patcog.2010.04.011.

[31] H. Yahia, J. Sudre, V. Garçon, and C. Pottier, "High-resolution

ocean dynamics from microcanonical formulations in non linear complex signal analysis," in *AGU FALL MEETING*, San Francisco, États-Unis, Dec. 2011, American Geophysical Union.
[32] S. K. Maji, O.l Pont, H. Yahia, and J. Sudre, "Inferring information across scales in acquired complex signals," in *European Conference on Complex Systems (ECCS)*, 2012.
[33] S. K. Maji, H. M. Yahia, O. Pont, J. Sudre, T. Fusco, and V. Michau, "Towards multiscale reconstruction of perturbated phase from hartmann-shack acquisitions," in *AHS*, 2012, pp. 77–84.
[34] Hicham Badri, "Computer graphics effects from the framework of reconstructible systems," M.S. thesis, Rabat faculty of science-INRIA Bordeaux Sud-Ouest, 2012.
[35] A. Turiel and N. Parga, "The multi-fractal structure of contrast changes in natural images: from sharp edges to textures," *Neural Computation*, vol. 12, pp. 763–793, 2000.
[36] O. Pont, A. Turiel, and C. J. Pérez-Vicente, "Description, modeling and forecasting of data with optimal wavelets," *Journal of Economic Interaction and Coordination*, vol. 4, no. 1, June 2009.
[37] A. Turiel, C.J. Pérez-Vicente, and J. Grazzini, "Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study," *Journal of Computational Physics, Volume 216, Issue 1, p. 362-390.*, vol. 216, pp. 362–390, 2006.
[38] V. Khanagha, K. Daoudi, O. Pont, H. Yahia, and A. Turiel, "Non-linear speech representation based on local predictability exponents," *Neurocomputing Journal*, Apr. 2013.
[39] CMU ARCTIC speech synthesis databases, ," [Online], http://festvox.org/cmu_arctic.
[40] T. drugman, "Gloat toolbox," [Online], http://tcts.fpms.ac.be/ drugman/.
[41] KED TIMIT database, ," [Online], http://festvox.org/.
[42] Noisex-92, ," [Online], www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

Khalid Daoudi received both the Master and the Ph.D. degrees in applied mathematics from University Paris 9 Dauphine, in 1993 and 1996, respectively. His Ph.D. dissertation was prepared at the Fractals Group of INRIA Rocquencourt, France. During 1997, he held a post-doctoral position at the Department of Mathematics, Ecole Polytechnique de Montral,Canada. From December 1997 to July 1999, he held a post-doctoral position at the Stochastic Systems Group (SSG) of the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge, USA. Since October 1999, he has a permanent position at INRIA Lorraine within the Speech Group. From October 2003 to February 2009, he was on leave at CNRS with the SAMOVA team of IRIT in Toulouse. Since March 2009, he is at INRIA Bordeaux with the GEOSTAT team. His research interests include Statistical Modeling and Estimation, Machine Learning, Multiscale Signal Processing and speech signal processing.

Hussein M. Yahia is the head of INRIA research team GEOSTAT (Geometry and Statistics in Acquisition Data, http://geostat.bordeaux.inria.fr). He specializes in nonlinear Signal Processing and the analysis of Complex Signals and Systems using advanced nonlinear Physics. He is the author and co-author of about 70 publications in international peer-reviewed journals and conferences. He has been managing international and European contracts and projects and has supervised 10 Ph.D. students.

Vahid Khanagha is currently a postdoctoral research associate at Institute for System Research of University of Maryland. His primary focus of research is front-end speech analysis techniques with emphasis on hardware efficient noise-robust parameter extraction methods. He received his Ph.D. degree in computer sciences from University of Bordeaux I in partnership with the French National research institute for computer sciences (INRIA). His Ph.D. dissertation was prepared at the GEOSTAT INRIA team in Bordeaux, France.