Phonetic segmentation of speech signal using local singularity analysis

Vahid Khanagha, Khalid Daoudi, Oriol Pont, Hussein Yahia

Phonetic segmentation of speech signal using local singularity analysis

Vahid Khanagha, Khalid Daoudi, Oriol Pont and Hussein Yahia

INRIA, GeoStat team
200 Avenue de la vielle tour, 33405 Talence cedex, France
Email: {vahid.khanagha, khalid.daoudi, oriol.pont, hussein.yahia}@inria.fr

Abstract

This paper presents the application of a radically novel approach, called the Microcanonical Multiscale Formalism (MMF) to speech analysis. MMF is based on precise estimation of local scaling parameters that describe the inter-scale correlations at each point in the signal domain and provides efficient means for studying *local* non-linear dynamics of complex signals. In this paper we introduce an efficient way for estimation of these parameters and then, we show that they convey relevant information about local dynamics of the speech signal that can be used for the task of phonetic segmentation. We thus develop a two-stage segmentation algorithm: for the first step, we introduce a new dynamic programming technique to efficiently generate an initial list of phoneme-boundary candidates and in the second step, we use hypothesis testing to refine the initial list of candidates. We present extensive experiments on the full TIMIT database. The results show that our algorithm is significantly more accurate than state-of-the-art ones.

Keywords: Nonlinear speech processing, Multiscale signal processing, phonetic segmentation of speech signals

1. Introduction

The existence of highly non-linear and turbulent phenomena in production process of the speech signal has been theoretically and experimentally established [1, 2]. However, most of the classical approaches in speech processing are based on linear techniques that may not adequately capture the complex dynamics of the speech signal. Some of the well-known examples of such non-linear phenomena include: the existence of turbulent sound source in production process of unvoiced sounds [1, 3, 4], the existence of a time spread and turbulent component for the excitation source of plosives [5] (that is idealized as an impulse in the linear framework) and the evidences regarding characterization of voiced sounds by highly complex air flows like jets and vortices [2].

There is a broad spectrum of non-linear speech analysis approaches [3, 6, 7, 8]. One common approach is to consider speech as a realization of a non-linear dynamical system and use the tools and methods in the study of such systems to catch the non-linear character of the speech signal. In these approaches, the non-linear aspects of the speech signal are associated to the turbulent nature of airflow in the vocal tract [1, 4, 2, 9], which in turn justifies the use of methods from the study of turbulent systems for speech analysis [6].

For instance in [10], it is assumed that the finite-dimensional embedded attractor of speech in the phase-space shares common features with its unknown non-linear production system. Consequently, topological features of the unknown dynamical system can be estimated from its embedded attractor. *Correlation dimension* and *general dimension* are two examples of such invariant quantities that are employed in [10] to define new features for speech analysis. Another characteristic of a dynamical system which remains intact by embedding procedure is Lyapanov Exponents (LE). LEs characterize the *global* degree of chaos or predictability in a dynamical system. They are shown in [11] to be useful features for phoneme classification. Also, a positive value of LE (which can be considered as an indicator for chaoticity) is observed for most of the phonemes.

These experiments show the potential of employing the tools and concepts coming from the study of chaotic, turbulent systems for speech analysis. However, most of these approaches suffer from practical and theoretical limitations. Apart from the rather involved mathematical procedures, the major limitation of most of these methods is their underlying assumption of stationarity, which is clearly not valid for a rapidly time-varying signal like speech. Consequently, these tools and methods can only be applied on isolated parts of the signal that are assumed stationary, like isolated phonemes. However, another issue appears when dealing with isolated phonemes: the length of a single phoneme is usually less than sufficient for precise estimation of parameters like LEs. This is why sustained vowels are used in [12] instead of naturally uttered ones. In another study [12] extensive experiments are conducted so as to choose the optimal dynamical model that can be used for relatively more precise estimation of LEs from very short signals like phonemes (to avoid the use of sustained vowels). Other than such practical issues, the fact that these approaches are based on the estimation of some *global* parameters, limits their usability for accessing many of the important *local* dynamical properties of the speech signal. These methods can generally recognize the existence of highly complex structures, but they can not provide local access to these structures.

In this paper, we use a radically novel framework coming from the study of highly disordered turbulent flows in statistical physics that does not suffer from the aforementioned limitations. This formalism offers new ways to evaluate quantitatively and locally the degree of predictability at every point in the signal domain. This is done trough the accurate estimation of local scaling parameters called Singularity Exponents (SE), which not only allows the identification of complexity inside signals but also allows to localize geometrically, inside the signal, where the complexity happens and how it organizes itself. In this paper we present a novel efficient method in estimation of SEs which is directly related to the degree of predictability around a point. We show how they convey meaningful information about some local dynamics of speech. Then, to demonstrate the usefulness of these exponents, we show how they can be easily used to develop an accurate and efficient phonetic segmentation algorithm.

Given that most of the recent developments in speech technology strongly rely on corpus-based methodologies that require the availability of precisely time-aligned labels of phonemes, phonetic segmentation is a very important task. The most precise segmentation method is to manually perform the task, which is extremely complex and expensive. Hence, *automatic* phonetic segmentation is of great importance and interest. The most frequent approach is to use an HMM-based phonetic recognizer supplied with the associated phonetic transcription and performing a forced alignment [13] to detect phoneme boundaries. Although such methods provide very good segmentation performances, they suffer from imposing linguistic constraints to the segmentation algorithm. This makes the algorithms restricted to the database used for training and hence, they can not be applied to different databases of different languages, contexts or accents. Moreover, it

is not always possible to access phonetic transcriptions to perform forced time-alignment. Such HMM based methods, or any other approach which relies on an externally supplied transcription of the sentence for determining phoneme boundaries are called Text-Dependent (TD) segmentation methods. Text Independent (TI) methods on the other hand rely exclusively on the acoustic information contained in the speech signal itself and do not require externally supplied transcriptions. TI methods can be used to simplify the manual labeling process for development of speech corpora. They may play a very important role in phoneme-based multi-lingual speech recognition systems [20].

The method we use in this paper belongs to the TI class of segmentation methods. We try to exploit our precisely estimated SEs to identify the phoneme boundaries. We employ a classical two-step methodology in change detection to identify the points in which local properties of SEs are changing. First, we use the measure we introduced in [15], which represents the instantaneous average of SEs to select a set of candidates for phonetic boundaries. To do so, we use the piece-wise-linear approximation of this measure applied to signal itself and its low-passed filtered version. In the second step, the initial list of candidates is refined using dynamic windowing and Log-Likelihood Ratio Test (LLRT). This two-step approach is similar to traditional segmentation methods that are based on a boundary pre-selection step followed by statistical tests to make the final decision [16, 17]. Some of the preliminary results of our algorithm are presented in [18]. In this paper, we describe the detailed procedure for SE estimation (a novel procedure that we have particularly adopted for phonetic segmentation problem (section 2.1)). Also, our novel dynamic-programming-based technique for piece-wise-linear approximation [which has crucial role in efficiency and text-independency of our segmentation algorithm] is fully presented in section 4.2. Moreover, we present more extensive experimental results (the *test* part of TIMIT database is included).

We evaluate our method over full TIMIT database including its *train* and *test* sets and we report all of the well-known performance measures such as detection, insertion and over-segmentation rates along with two global measures which simultaneously take these three into account. By doing so for several degrees of error tolerance, we provide results which are easy to interpret and to compare with. We compare our results with two state-of-the-art methods [19, 20] which have reported their results on these known datasets (*train* and *test* dataset respectively). [21] and [22] also report on the *train* dataset but the algorithms they propose are not fully unsupervised. The former assumes prior knowledge of the number of phonemes in the utterance, while the latter uses prior knowledge of all manual transcriptions to train a neural network (and hence the "text-independence" property of the system is compromised). Applied to the *train* set, our algorithm significantly outperforms [19]. Consequently, the application of the algorithm to the *test* dataset, without any tuning of parameters, shows comparable results w.r.t the best performance of [20] which is achieved after extensive tuning of parameters on the whole dataset.

The paper is structured as follows: in Section 2 we briefly introduce the basic concepts of the MMF and the detailed procedures for computation of SEs. In section 3 some examples about the informativeness of SEs are presented and the detailed algorithm for using them in the TI phonetic segmentation task is presented in section 4. The experimental results are presented in Section 5 and we draw our conclusion in Section 6.

2. The Microcanonical Multiscale Formalism

The approach we use in this paper for nonlinear speech analysis is based on the Microcanonical Multiscale Formalism (MMF) [23], which allows the study of local geometrico-statistical

properties of complex signals from a multi-scale perspective. The particularity of this formalism compared to its older counterparts (the canonical formulations [24]) is that it considers *local* quantities defined at each point of the signal domain, instead of global quantities like moments and structure functions. Central to the formalism is the computation of Singularity Exponents (SE) at each point in the signal domain. When correctly defined and estimated, these exponents alone can provide valuable information about local dynamics of complex signals and have been successfully used in many applications ranging from signal compression to inference and prediction [25, 26, 27, 28, 29]. SE values can be used to analyze the local scaling behavior of complex signals. It is shown [23] that for a given time-instant t, the smaller the value of SE is, the higher unpredictability is at this point. This property can be used to identify an important subset of signal samples that carry most of the information content of the signal: the Most Singular Manifold (MSM). MSM is defined as the subset of signal samples whose SE values are smaller than a threshold. It has been established that the critical transitions of the system occur at these points and this fact has been successfully used in many applications [23, 26].

The Singularity Exponent (SE), $h(t)$, for a signal $s(t)$, can be estimated by evaluation of the power-law scaling behavior of a multi-scale functional Γ_r over a set of fine scales r:

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

where $\Gamma_r\left(\cdot\right)$ can be any multi-scale functional complying with this power-law. The term $\mathrm{o}\left(r^{h(t)}\right)$ means that for small scales the additive terms are negligible and thus $h(t)$ dominantly quantifies the multi-scale behavior of the signal at time t. $\alpha(t)$ is a quantity that is independent of the scale and can be separated from $h(t)$.

An important aspect in the MMF is the definition $\Gamma_r(\cdot)$ such that the inter-scale power-law correlations in form of Eq. (1) are revealed. As suggested in [23], one reasonable choice is to integrate all the variations ($d\mu_{\mathcal{D}}$) of the signal on a ball B_r of radius r centered at t:

$$\Gamma_{\mu_r}\left(s(t)\right) = \int_{B_r} d\mu_{\mathcal{D}} \tag{2}$$

while the variations being defined as:

$$d\mu_{\mathcal{D}} = \mathcal{D}s\; dr \tag{3}$$

where $\mathcal{D}$ is an appropriate differential operator such as the norm of the gradient $\parallel \nabla s \parallel$, whose choice is motivated in [23] by the special power spectrum scaling of natural images. The use of the norm of the gradient helps to avoid the domination of the integral in Eq. (2) by the mean value of s, which may be far from zero due to the lack of stationarity. The resulting multi-scale measure is related to the typical characterization of intermittency in turbulence and describes the kinetic energy dissipation at scale r on B_r for a turbulent velocity field [24, 30].

2.1. Estimation of the singularity exponents

Ideally, if the power-law scaling behavior of Eq. (1) holds for a given measure $\Gamma_r(\cdot)$, the measurements at different scales are directly related. In this case, SEs can be estimated by evaluation of Eq. (1) using only one measurement at the finest scale: if one takes the logarithm of both sides of Eq. (1), $\alpha(t)$ appears as an additive term that would be negligible for an adequately small scale r_0 [31]. Now if we drop the constant shift d, one can have an estimate of the SE as:

$$h(t) = \frac{log\ \frac{\Gamma_{\mu_{r_0}}(s(t))}{\langle \Gamma_{\mu_{r_0}}(s(\cdot)) \rangle}}{log\ r_0} \tag{4}$$

where $\langle \Gamma_{\mu_{r_0}}(s(\cdot)) \rangle$ is the average of the functional over the whole domain of the signal and serves to diminish the additive factor due to $\alpha(t)$, as explained in [31]. There, it is shown that the additive factor due to $\alpha(t)$ is negligible provided that the finest accessible scale (r_0) is small enough.

By using the punctual estimation one preserves the finest accessible resolution in SE estimation. However, in practice we do not usually have access to the physical finest scale (due to discretization limitations) and hence the resulting estimate ($\Gamma_{\mu_{r_0}}$) might be unstable and perturbed by the additive term $\alpha(t)$. This perturbation can be corrected by incorporating the measurement at coarser scales. But this has the risk of compromising the resolution and the measurements at coarser sampling scales may simply combine the information carried by adjacent samples. For the application we consider in this paper, resolution is of very high importance. So we only add the measurement at the second finest scale ($\Gamma_{\mu_{r_1}}$) to regularize $\Gamma_{\mu_{r_0}}$ as:

$$\Gamma^{\kappa}_{r}(s(t)) = \kappa_t \Gamma_{\mu_{r_0}}(s(t)) \tag{5}$$

where the regularization term κ_t is defined as the quotient of the measurements at two scales:

$$\kappa_t = \sqrt{\frac{\Gamma_{\mu_{r_1}}(s(t))}{\Gamma_{\mu_{r_0}}(s(t))}} \tag{6}$$

where r_1 is the next coarser scale ($r_1 > r_0$). In the ideal case both measurements are completely correlated and $\kappa_t = 1$. But in practice, κ_t serves to compensate for the perturbations of a possibly noisy measurement made at the finest scale r_0. We thus replace $\Gamma_{\mu_{r_0}}$ in Eq. (4) with $\Gamma^{\kappa}_{r}(s(t))$ to compute $h(t)$. A summary of the overall procedure for computing the SE is provided in Alg. 1. Note that the procedure is written for a 1-D discrete time signal s[n] sampled from s(t) with the sampling frequency of fs. For such a discrete time signal, the finest accessible scale is the one provided by the sampling frequency. Consequently an estimate of the sum of variations at the finest scale (Eq. (2)) would simply be $\Gamma_{\mu_{r_0}}(s[n]) = |s[n] - s[n-1]|$ and at the second finest scale, we have two measurements to sum up as $\Gamma_{\mu_{r_1}}(s[n]) = |s[n] - s[n-1]| + |s[n+1] - s[n]|$.

Algorithm 1 The procedure for computation of the SEs for a discrete time signal

1: $\Gamma_{\mu_{r_0}}(s[n]) \leftarrow |s[n] - s[n-1]|$
2: $\Gamma_{\mu_{r_1}}(s[n]) \leftarrow |s[n] - s[n-1]| + |s[n+1] - s[n]|$
3: $\kappa_t[n] = \sqrt{\frac{\Gamma_{\mu_{r_1}}(s[n])}{\Gamma_{\mu_{r_0}}(s[n])}}$
4: $\Gamma^{\kappa}_{r}(s[n]) = \kappa_t[n] \Gamma_{\mu_{r_0}}(s[n])$
5: $h[n] = \frac{\log \frac{\Gamma^{\kappa}_{r}(s[n])}{\langle \Gamma_{\mu_{r_0}}(s[n]) \rangle}}{log\ r_0}$
6: ▷ $r_0 = \frac{1}{fs}$, $\langle \cdot \rangle$ denotes the average value over the whole sentence

3. A singularity analysis of phonemes

In this section we present some preliminary observations on how the temporal evolution of $h[n]$ conveys instructive information that can be used to detect phoneme boundaries. In Fig. 1–top a speech signal from the TIMIT database is shown along with the reference phoneme boundaries extracted from the manual transcription of TIMIT. Fig. 1–middle displays the time-evolution of

local histograms of $h[n]$. Each column in this figure shows a histogram of SE values computed on a window of 32ms length around the time t. The red color shows the highest probability and the blue color shows zero probability. One can easily observe that there is a change in the position of peak and in the variability of the distribution at phoneme boundaries.

The inter-phoneme variability of exponents can be further demonstrated using the concept of MSM. As mentioned in section 2, MSM is defined as the set of samples in the signal domain whose value of $h[n]$ is smaller than a threshold and it theoretically represents the most informative subset of samples in the signal domain. In our experiment, we set a global threshold h_{th} for the whole speech signal and in each window of length 32ms around t, we count the number of samples with $h[n] < h_{th}$. The result is shown in Fig. 1–bottom. It can be seen that the counter value is almost constant inside each phoneme (if we neglect the transition on phoneme boundaries where the sliding window partially covers both of the neighboring phonemes) and this constant value is different for adjacent phonemes.

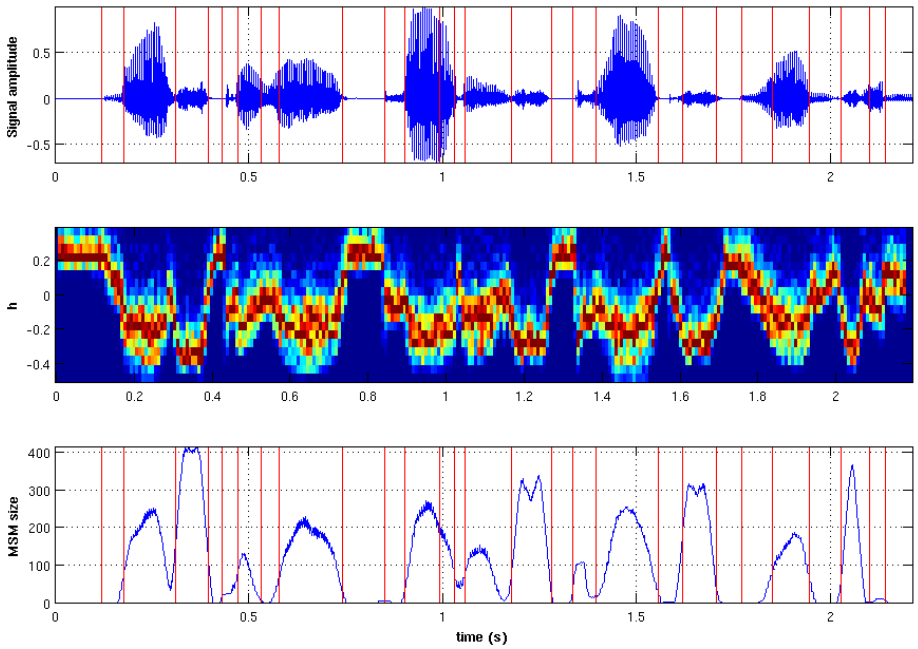

Figure 1: **(top)** A part of a speech signal from TIMIT database. Manually-positioned phoneme boundaries are marked with vertical red lines. **(middle)** The local histogram of singularity exponents in a 32ms window around each time instant. Red corresponds to maximum probability and dark blue corresponds to zero probability. **(bottom)** The number of points belonging to the MSM in a sliding window of 32ms length.

These observations suggest that SE do indeed convey valuable information about local dynamics of the speech, and in this very simple form can be readily used for phoneme segmentation. We next develop new tools to employ these properties for practical development of an automatic phonetic segmentation algorithm.

4. Application of singularity analysis to phonetic speech segmentation

We use a two steps procedure to employ the properties of SEs mentioned in section 3 for the task of automatic phonetic segmentation. We first define a measure that captures the inter-phoneme changes in the properties of SEs in a very simple form and use it to make a preliminary list of phoneme boundary candidates. Then we perform a dynamical windowing with a hypothesis testing to decide whether each candidate truly represents a phoneme transition.

4.1. The Accumulative function ACC

The simplest parameter that can quantify the changes in distribution of SEs between neighboring phonemes is their average. We consider $h[n]$ as a random variable whose average is changing between adjacent phonemes. We thus search for the locations of changes in local averages of $h[n]$ and consider them as the candidate phoneme boundaries. Local averages of this random variable can be estimated in a small sliding window. However, as SE estimations are available at the finest resolution and we are interested in preserving this resolution for the task of phonetic segmentation, we try to avoid windowing for estimation of averages. Instead, we use the running mean of $h[n]$ defined as:

$$ACC[n] \;=\; \sum_{k=1}^{n} (h[k] - \bar{h}) \tag{7}$$

where $\bar{h}$ is the global average of exponents (over the whole sentence) and N is the total number of speech samples. Indeed, inside the boundaries of each phoneme, the slope of $ACC[n]$ would be an estimate for the local average of the $h[n]$ minus the constant term $\bar{h}$. The resulting functional is plotted in Fig. 2. Just as we expected, this new functional reveals the changes in distribution in a more precise way. Indeed $ACC[n]$ is almost linear inside each phoneme, while there is a clear change in the slope at the phoneme boundaries. Extensive observations over different sentences confirm this behavior and show that these slope changes are even happening at the boundaries between extremely short phonemes, such as stops. Consequently, detection of the locations of changes in the slope of the ACC can lead to an automatic phonetic segmentation algorithm. A very simple solution is to fit a piece-wise-linear curve to the ACC and take the break-points as the candidates of change in distribution of SE, i.e. phoneme boundary candidates.

4.2. Piece-wise-linear approximation of ACC

For a speech signal of length N that contains K phoneme boundaries, we aim at fitting a piece-wise-linear curve to the corresponding $ACC[n]$ such that the breakpoints are located on the phoneme boundaries that we are searching for. If we denote the m-piece linear approximation of $ACC[n]$ ($n \in [1, N]$) by $LA_m(ACC, 1, N)$ and show the corresponding mean squared error of this approximation by $E_m^{1\to N}$ the following optimization problem can be used to reach our goal (conventional method):

- find $LA_m(ACC, 1, N)$ such that m is minimized and $E_m^{1\to N} < \epsilon$.

where ϵ is a reasonable threshold that is conventionally set by trial and error. Ideally, we wish that $m = K$ so that there are no false detections or deletion of boundaries. An intuitive solution for this problem is to recursively search through *all the points* in the domain of $ACC[n]$ to find the breakpoints (one after another) such that at each step the least possible value of approximation

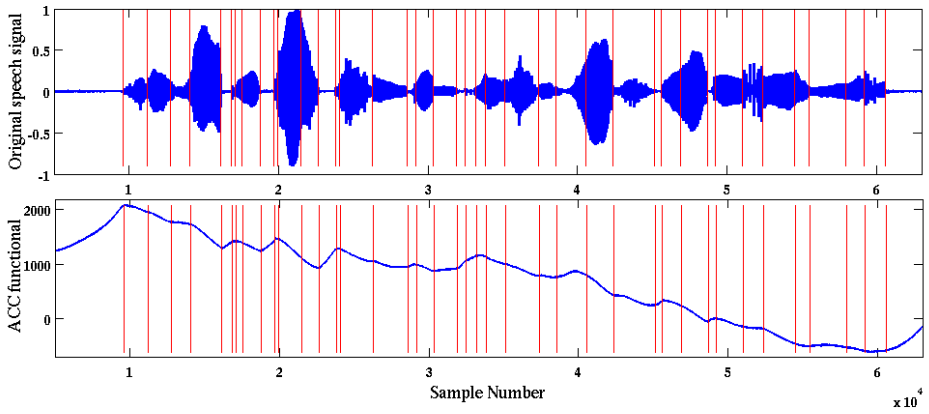

Figure 2: The speech signal "Masquerade parties tax one's imagination" from the TIMIT database and the *ACC* functional of Eq. (7).

error is achieved. This method is addressed in [32] for denoising of piece-wise constant functions and it is argued that when $K << N$, such a greedy search may result in a more efficient and more accurate approximation compared to other solvers. However, considering the large values of N and K in the particular problem of phonetic segmentation, the computational complexity of this method (which increases with K) is still unbearable when dealing with large databases of the speech signal. Also, since we don't know the number of phoneme boundaries (K), the stop condition of the algorithm (the value of ϵ) is vague. We thus develop a more efficient method whose computational load is independent of K.

Our optimization method is motivated by the dynamic programming approach of [33] where the above optimization problem is solved by recursive solution of easier sub-problems of the same type. In our method, instead of recursive search for K breakpoints over the *whole* signal each time, we perform a left-to-right search that passes through the signal only once: from left to right, each time a break-point is selected, the part of *ACC* before that breakpoint is removed from the search in the next iterations. More specifically, we start from $n = 1$ and search for the first point n_1 such that the error of the 1-piece approximation of $ACC[n]$ with a straight line from $ACC[1]$ to $ACC[c_i]$ becomes larger than a threshold ϵ. Clearly the location of n_1 is dependent on the choice of ϵ, which is not desirable because we are searching for specific fixed location of phoneme boundaries. We overcome this issue in the second step: we first opt for a relatively large threshold to ensure that noise-like variations are not mistaken as breakpoints. Once n_1 is found, we search for the breakpoint $c_1 \in [1, n_1]$ that once taken as a breakpoint, the 2-piece linear approximation [1] error in this segment is *minimized*. As such, the breakpoint c_1 is found through a *local minimization* procedure and not just a *thresholding* procedure and hence, we expect the overall algorithm to be less sensitive to the choice of the threshold ϵ. Once c_1 is fixed, we perform the exact same procedure, this time starting from c_1 to find the second breakpoint and repeat the procedure until the last sample is reached. This procedure is detailed in Alg. 2.

¢

[1] with the two lines that connects $ACC[1]$, $ACC[c_1]$ and $ACC[n_1]$

Algorithm 2 PLA procedure

1: $i \leftarrow 1; c_0 \leftarrow 1; k_1 \leftarrow 2;$
2: **while** $k_1 < N$ **do**
3: **if** $E_1^{c_{i-1} \rightarrow k_1} > \epsilon$ **then**;
4: $n_i \leftarrow k_1;$
5: **for** $k_2 \in (c_{i-1} \cdots n_i)$ **do**
6: $E_2^{c_{i-1} \rightarrow n_i}[k_2] \leftarrow E_1^{c_{i-1}+1 \rightarrow k_2} + E_1^{k_2+1 \rightarrow k_1};$
7: **end for**
8: $i \leftarrow i + 1;$
9: $c_i = \underset{k_2}{\text{argmin}}\, E_2^{c_{i-1} \rightarrow n_i}[k_2];$
10: $k_1 \leftarrow c_i;$
11: **end if**
12: $k_1 \leftarrow k_1 + 1;$
13: **end while**
14:
15: ▷ $E_1^{m_i \rightarrow m_j}$ is the MSE of approximating ACC by a straight line from $ACC[m_i]$ to $ACC[m_j]$.
16: ▷ At each iteration c_{i-1} denotes the previously taken phoneme candidate.

We emphasize that at each iteration, c_i is determined through a greedy minimization (Alg. 2 line 9) and not through a thresholding operation. This greedy minimization decreases the impact of the threshold ϵ (Alg. 2 line 3). We will later show in the experimental results that the algorithm is not so sensitive to the selection of ϵ. The insensitivity to the selection of algorithm parameters is an important property for text-independent phonetic segmentation, where no data is available to train the parameters.

4.3. The two-step segmentation algorithm

The piece-wise linear reconstruction of ACC as explained in section 4.2 is readily a working phonetic segmentation algorithm. However, by performing detailed analysis on the performance of this algorithm we found that it can not fully exploit the strength of SE for phone boundary detection. Motivated by some observations on the behavior of this simple algorithm at some particular phoneme transitions, we propose a two-step algorithm where we first pre-select candidate boundaries and then use statistical hypothesis test to make the final decision.

We first observed that some of the missed boundaries by the method of section 4.2 correspond to transitions between fricatives/stops and vowels. We also observed that transitions between speech and low energy segments (such as pauses and epenthetic silence) display strong and easy to detect changes in the slopes of ACC. Indeed, as shown in Fig. 1–middle, the SEs of low energy segments have high positive values, while they are mostly negative in active speech segments. Motivated by these observations and the fact that fricatives/stops are essentially high-band signals, we decided to compute ACC on a low-pass filtered version of the utterance. By doing so fricatives/stops-vowels transitions will be converted into silence-speech transitions which are easier to detect as shown in Fig. 3. It is known that most of the spectral energy of fricatives is located above 2000Hz and, for most stops, the active frequency bands start at 1800Hz [34]. We thus choose the cut-off frequency of the low-pass filter as 1800Hz.

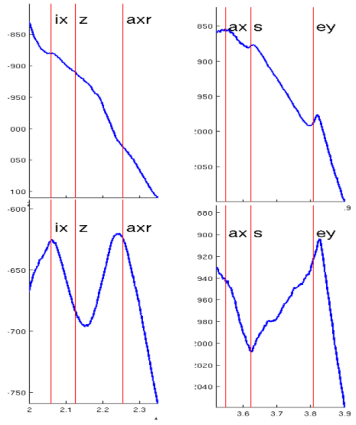

Figure 3: **(top)** Two examples of *ACC* for the original signal and **(bottom)** The *ACC* for the low-passed filtered signal. Phoneme boundaries are marked with vertical red lines.

The second and most important observation we made is related to the statistical distribution of SEs. We observed that some missed boundaries correspond to neighboring phonemes which although have a quite strong difference in their SE distribution, the change in their averages is not strong enough to be reflected as a change in the slope of *ACC* and thus it is not captured by the simple linear approximation procedure. It is then natural to consider a statistical hypothesis test over SE distributions in our segmentation algorithm in order to detect such boundaries.

Motivated by these observations, we use a similar approach as in [17] develop a new segmentation algorithm which consists of 2 steps. First, we apply Alg. 2 to the original signal and to its filtered version. We gather all the detected breakpoints and consider them as candidate boundaries. In the second step, we make the final decision on each candidate by performing a dynamic windowing over them followed by Log Likelihood Ratio Test (LLRT) over SE distributions of the *original* signal. We use a Gaussian hypothesis because our goal is to detect changes in mean and variance of SEs. More precisely, for each candidate c_i we consider the large window $Z = [c_{i-1}, c_{i+1}]$ and the two smaller windows $X = [c_{i-1}, c_i]$ and $Y = [c_i, c_{i+1}]$. We then compute LLR statistic to decide between the two hypotheses:

- H_0 : SE of Z are generated by a single Gaussian.
- H_1 : SE of Z are generated by two Gaussians on X and Y.

If H_1 is significantly likelier than H_0, we select c_i as a boundary. Otherwise, c_i is removed from the candidates list. We emphasize here that SE of the filtered signal are used only in the first step. The final decision is exclusively made upon the information conveyed by SE of the *original* signal. We also emphasize that this new algorithm is still simple and efficient as the original one.

5. Experimental results

Our evaluation is carried out on the *full* Train and Test sets of the TIMIT database which respectively contain 4620 and 1200 sentences uttered by 462 and 120 speakers. We first develop our algorithm on a small dataset of 30 sentences extracted from the Train set. First we test the

performance on the whole Train set and we compare the results with the state-of-the art methods. Then, to compare with [20], and to show that the promising result on Train dataset is not the result of over-trained parameters, we evaluate it on the Test dataset which is balanced for the phonological coverage.

5.1. Performance measures

The segmentation quality can be evaluated and analyzed using three "partial" scores: the *Hit Rate* (*HR*), the *False Alarm Rate* (*FA*) and the *Over Segmentation Rate* (*OS*). These three scores are defined in Table 1.

Table 1: *Partial performance measures. N_T is the total number of detected boundaries (correct and incorrect), N_H is the number of correctly detected boundaries and N_R is the total number of boundaries in the reference transcription*

Measure	Formula	Description
Hit Rate (*HR*)	$HR = \frac{N_H}{N_R}$	The percentage of reference boundaries that are correctly detected
Segmentation Rate (*OS*)	$OS = \frac{N_T - N_R}{N_R}$	*OS* shows how much more (or less) is total number of algorithm detections, compared to the total number of reference boundaries taken from the manual transcription.
False Alarm Rate (*FA*)	$FA = \frac{N_T - N_H}{N_T}$	The percentage of algorithm detections that are incorrect.

In order to assess the overall quality of a segmentation method, a global measure which simultaneously takes these scores in to account is required. A well known measure is the F_1-value:

$$F_1 = \frac{2 \times PCR \times HR}{PCR + HR} \tag{8}$$

where $PCR = 1 - FA$. Another global measure, called the *R*-value has been recently proposed in [35]. This measure makes more emphasize on over-segmentation by arguing that better hit rates might be achieved by simply adding random boundaries without any algorithmic improvement. This measure evaluates how close one is to the ideal segmentation $R = 1$:

$$\begin{gathered} R = 1 - \frac{|r_1| + |r_2|}{2} \\ r_1 = \sqrt{(1 - HR)^2 + OS^2},\ r_2 = \frac{HR - OS - 1}{\sqrt{2}} \end{gathered} \tag{9}$$

5.2. Results : Train dataset

Using different sizes of tolerance windows, we provide comparison of segmentation results for 3 methods. In the first one, we give the results reported in [19]. We mention here that [19]

report scores with 0ms, 10ms and 20ms tolerance windows. However, their approach is frame-based with a 10ms frame step size and they convert each manual boundary to the closest frame position. Thus, 5ms has to be added to their window size in order to make a fair comparison with our sample-based approach, which has the finest possible resolution. In the second one, we provide the results using our original SE-based algorithm summarized in section 4.2, we call it SE-ACC. In the third one, we present the results obtained using our new algorithm described in section 4.3, we call it SE-LLRT. Table 2 presents HR, FA and OS for the 3 methods. The first observation is that SE-LLRT outperforms SE-ACC for the 3 scores and all tolerance windows. In particular, a significant improvement is made in FA and OS. This shows that, as expected, some of the insertions introduced by the curve fitting procedure have been corrected by the LLRT. The second observation is that SE-LLRT yields higher accuracy than [19]. In particular, the smaller the tolerance window, the higher the relative improvement is. This shows that SE-LLRT is better suited for high precision detection of phoneme boundaries. To this regard, we can mention another interesting comparison with [22] which is also a sample-based segmentation method as ours. In [22], it is reported that 43.5% of their 86.8% detection output is located within the first bin of the cumulative histogram of distances from true boundaries. This corresponds to 43.5% $\times$ 86.8%=37.75% hit rate with 7.5ms tolerance. With SE-LLRT we obtain 44% hit rate which is significantly more accurate. More importantly, the algorithm in [22] is supervised (all manual transcriptions are used to train a neural network) while ours is fully unsupervised.

Table 2: *The comparative table of segmentation results. The scores are reported as percentages.*

tolerance	score	Dusan et al	SE-ACC-Train	SE-LLRT-Train	SE-LLRT-Test
5ms	HR	22.8	31.7	32.3	32.4
	FA	79.7	70.2	68.5	62.02
	OS	12.8	6.4	2.5	4.61
10ms	HR	-	52.8	53.6	53.16
	FA	-	50.4	47.7	49.18
	OS	-	6.4	2.5	4.61
15ms	HR	59.2	65.5	66.3	65.39
	FA	47.5	38.4	35.4	37.50
	OS	12.8	6.4	2.5	4.61
20ms	HR	-	72.4	72.5	71.7
	FA	-	31.94	29.2	31.5
	OS	-	6.42	2.5	4.61
25ms	HR	75.3	76.2	76.1	75.17
	FA	33.2	28.3	25.8	28.14
	OS	12.8	6.4	2.5	4.61
30ms	HR	-	78.8	80.5	77.41
	FA	-	26	23.7	26.00
	OS	-	6.4	2.5	4.61

Table 3 presents the performance of each of the 3 methods when evaluated using the global measures F_1 and R. The same observations we made above, still hold for the global performance evaluation. Indeed, SE-LLRT still outperforms SE-ACC for both F_1 and R. Moreover, about 6% (resp. 10%) improvement in R-value and 4% (resp. 10%) in F_1-value is achieved for 25ms of

tolerance (resp. 5ms and 15ms). This is a significant gain in accuracy that shows the strength of SEs in revealing the transitions fronts between phonemes.

Table 3: *The comparative table of global performance measures.*

tolerance	score	Dusan et al	SE-ACC-Train	SE-LLRT-Train	SE-LLRT-Test
5ms	R-value	0.29	0.39	0.41	0.41
	F_1-value	0.21	0.31	0.32	0.32
10ms	R-value	-	0.57	0.60	0.58
	F_1-value	-	0.51	0.53	0.52
15ms	R-value	0.60	0.68	0.70	0.69
	F_1-value	0.55	0.63	0.65	0.64
20ms	R-value	-	0.74	0.76	0.74
	F_1-value	-	0.70	0.72	0.70
25ms	R-value	0.73	0.77	0.79	0.77
	F_1-value	0.71	0.74	0.75	0.735
30ms	R-value	-	0.79	0.81	0.79
	F_1-value	-	0.76	0.77	0.76

5.3. Results : Test dataset

To the best of our knowledge, no detailed result is reported in the literature for the performance of a TI method on the TEST dataset. However, to facilitate later comparisons, we have provided the full segmentation results for this dataset, in the last columns of table 2 and 3. In [20] the result for Test dataset is reported only for 20ms tolerance, as shown in table 4. It can be seen that the hit-rate of [20] (8-Mel-bank coding scheme) is about 10% higher than that of ours, while its over-segmentation is about 13% worse. Hence a direct comparison in terms of hit-rate and over-segmentation is not meaningful. However, in terms of R-value the results are very close.

It should be mentioned that the results of [20] are attained by performing an extensive parameter-tuning procedure over the whole Test dataset. Quite the contrary, in our case we have applied our algorithm to the Test dataset without any parameter tuning. Our intention was to ensure that the promising performance of the algorithm is not the result of over-training of parameters. Even more, we emphasize an important feature of our algorithm which is its insensitivity to the threshold of the linear curve fitting. Recalling that this threshold is applied to the mean-squared approximation error of the *ACC*, an initial value for the threshold is selected as 0.001 times the maximum value of *ACC* (which is normalized to unity). Fig. 4 shows how *R*-value varied for different thresholds. We used a subset of 30 randomly selected sentences to

Table 4: *The segmentation results of [20], for three different coding schemes. Tolerance : 20ms.*

Coding scheme	(a,b,c)	HR	OS	FA	R-value	F-value
8-Mel-bank	(2,optimal value,5)	82	18.32	30.69	0.743	0.75
5-MFCC		76	12.31	31.33	0.736	0.72
Log Area Ratio		70	6.31	34.16	0.72	0.68

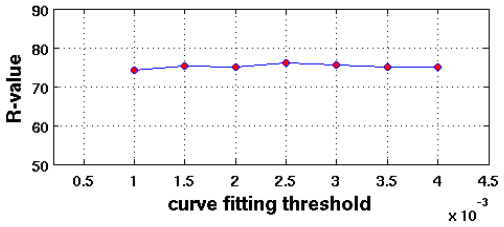

Figure 4: The sensitivity of the SE-LLRT to threshold.

compute these values. It can be seen that with about 400% change of the threshold value, the variance of changes in R-value is less than 0.5%. Thus, it seems fair to claim that our algorithm is threshold-independent: SE-ACC does not have any other parameter to tune and for the SE-LLRT, the threshold for LLRT is the only parameter to tune. This is a major advantage as most of the TI methods require accurate threshold tuning. For instance, the results of [20] are obtained by extensive optimization of three important parameters.

6. Conclusions

In this paper we used a novel framework in complex-signal analysis, called the MMF, to analyze the local dynamics of speech. We presented an efficient method for estimation of the singularity exponents which are the relevant features in this framework to identify critical transitions. Observing that temporal evolution of these exponents convey useful information about phoneme boundaries, we defined an accumulative functional upon these exponents which permits automatic detection of phone boundaries. This accumulative functional exhibits a piecewise-linear behavior with distinctive breakpoints at phoneme boundaries. We then developed a simple and efficient algorithm using dynamic programming to detect these break-points as the phoneme-boundary candidates (with special care for reducing computational complexity and also dependency on parameter tuning, which is necessary for text-independent phonetic segmentation). The final decision was made for each candidate, by performing a hypothesis test on the distribution of singularity exponents. We provided a detailed evaluation of our algorithm and compared it against several alternative approaches using the full *Train* and *Test* sets of TIMIT. The comparisons showed that our algorithm is more accurate than state-of-the-art ones. These encouraging results with a very simple analysis of singularity exponents confirm that nonlinear dynamics of the speech signal can indeed be geometrically accessed through the use of the MMF.

Acknowledgement

The first author is funded by the INRIA CORDIS doctoral program.

References

[1] J. F. Kaiser, Some observations on vocal tract operation from a fluid flow point of view, in: I. R. Titze, R. C. Scherer (Eds.), Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control, The Denver Center for the Performing Arts, 1983, pp. 358–386.

[2] H. M. Teager, S. M. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract, in: W. Hardcastle, A. Marchal (Eds.), Speech Production and Speech Modelling, NATO Advanced Study Institute Series D, 1989, pp. 241–261.

[3] G. Kubin, Speech coding and synthesis, Elsevier, 1995, Ch. 16: Nonlinear processing of speech, pp. 557–610.
[4] P. Maragos, A. Potamianos, Fractal dimensions of speech sounds: Computation and application to automatic speech recognition, Journal of Acoustic Society of America 105 (1999) 1925–1932.
[5] T. F. Quatieri, Discret-time speech signal processing: principles and practice, Prentice Hall PTR, 2002.
[6] M. Fandez-Zanuy, G. Kubin, W. B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, J. Schoentgen, Nonlinear speech processing: Overview and applications, Control and intelligent systems 30 (2002) 1–10.
[7] M. Little, Biomechanically informed nonlinear speech signal processing, Ph.D. thesis, Oxford University (2007).
[8] S. McLaughlin, P. Maragos, Nonlinear methods for speech analysis and synthesis, in Advances in Nonlinear Signal and Image Processing, Hindawi Publ. Corp., 2006.
[9] M. A. Little, P. McSharry, S. Roberts, D. Costello, I. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, BioMedical Engineering OnLine 6 (1) (2007) 23. doi:10.1186/1475-925X-6-23.
[10] V. Pitsikalis, P. Maragos, Analysis and classification of speech signals by generalized fractal dimension features, Speech Communication 51, Issue 12 (2009) 1206–1223.
[11] I. Kokkinos, P. Maragos, Nonlinear speech analysis using models for chaotic systems, IEEE Transactions on Speech and Audio Processing 13 (6) (2005) 1098–1109.
[12] M. Banbrook, S. McLaughlin, I. Mann, Speech characterization and synthesis by nonlinear methods, IEEE Transactions on Speech and Audio Processing 7 (Jan 1999) 1 – 17.
[13] D. Torre-Toledano, L. Hernandez-Gomez, L. Villarrubia-Grande, Automatic phonetic segmentation, IEEE Transactions on Speech and Audio Processing 11 (6) (2003) 617–625.
[14] B. Ata, J. Remde, A new model of lpc excitation for producing natural-sounding speech at low bit rates.
[15] V. Khanagha, K. Daoudi, O. Pont, H. Yahia, A novel text-independent phonetic segmentation algorithm based on the microcanonical multiscale formalism, in: INTERSPEECH, 2010, pp. 1393–1396.
[16] R. Andre-Obrecht, A new statistical approach for the automatic segmentation of continuous speech signals, IEEE Transactions on Acoustics, Speech and Signal Processing 36 (1) (1988) 29 –40. doi:10.1109/29.1486.
[17] G. Almpanidis, M. Kotti, C. Kotropoulos, Robust detection of phone boundaries using model selection criteria with few observations, IEEE Transactions on Audio, Speech, and Language Processing 17 (2009) 287–298.
[18] V. Khanagha, K. Daoudi, O. Pont, H. 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.
[19] S. Dusan, L. Rabiner, On the relation between maximum spectral transition positions and phone boundaries, Proceddings of INTERSPEECH/ICSLP 2006 (2006) 645–648.
[20] A. Esposito, G. Aversano, Text independent methods for speech segmentation, in: Nonlinear Speech Modeling and Applications, Vol. 3445 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2005, pp. 261–290.
[21] Y. Qiao, N. Shimomura, N. Minematsu, Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons, Proceedings of ICASSP2008 (2008) 885–888.
[22] Y. Lin, Y. Wang, Y.-F. Liao, Phone boundary detection using sample-based acoustic parameters, Proceedings of INTERSPEECH 2010.
[23] A. Turiel, H. Yahia, C. P. Vicente., Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis, Journal of Physics A: Mathematical and Theoretical 41 (2008) 015501.
[24] U. Frisch, Turbulence: The legacy of A.N. Kolmogorov, Cambridge University Press, 1995.
[25] J. Grazzini, A. Turiel, H. Y. 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.
[26] H. Yahia, J. Sudre, C. Pottier, V. Garçon, Motion analysis in oceanographic satellite images using multiscale methods and the energy cascade, Pattern Recognition 43 (10) (2010) 3591 – 3604. doi:http://dx.doi.org/10.1016/j.patcog.2010.04.011.
[27] S. K. Maji, O. Pont, H. Yahia, J. Sudre, Inferring information across scales in acquired complex signals, in: European Conference on Complex Systems (ECCS), 2012.
[28] S. K. Maji, H. M. Yahia, O. Pont, J. Sudre, T. Fusco, V. Michau, Towards multiscale reconstruction of perturbated phase from hartmann-shack acquisitions, in: AHS, 2012, pp. 77–84.
[29] H. Badri, Computer graphics effects from the framework of reconstructible systems, Master's thesis, Rabat faculty of science-INRIA Bordeaux Sud-Ouest (2012).
[30] A. Turiel, N. Parga, The multi-fractal structure of contrast changes in natural images: from sharp edges to textures, Neural Computation 12 (2000) 763–793.
[31] A. Turiel, C. Prez-Vicente, J. Grazzini, Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study, Journal of Computational Physics 216 (2006) 362–390.
[32] M. Little, N. Jones, Generalized methods and solvers for noise removal from piecewise constant signals: Parts i and ii, Proceedings of the Royal Society A, 2011,doi:10.1098/rspa.2010.0671.
[33] I. Ihm, B. Naylor, Scientific visualization of physical phenomena, Springer-Verlag New York, 1991, Ch. 32: Piece-

wise linear approximations of digitized space curves with applications.
[34] P. Ladefoged, K. Johnson, A Course in Phonetics, Chapter 8: Acoustic, Cengage Learning, 2010.
[35] O. J. Rasanen, U. K. Laine, T. Altosaar, An improved speech segmentation quality measure: the R-value, Proceedings of INTERSPEECH 2009.