Improving text-independent phonetic segmentation based on the Microcanonical Multiscale Formalism

Vahid Khanagha, Daoudi Khalid, Oriol Pont, Hussein Yahia

IMPROVING TEXT-INDEPENDENT PHONETIC SEGMENTATION BASED ON THE MICROCANONICAL MULTISCALE FORMALISM

Vahid Khanagha, Khalid Daoudi, Oriol Pont and Hussein Yahia

INRIA Bordeaux Sud-Ouest (GEOSTAT team)
351 Cours de la Libération, BAT. A29, 33405 Talence, France
email: vahid.khanagha@inria.fr, http://geostat.bordeaux.inria.fr/

ABSTRACT

In an earlier work, we proposed a novel phonetic segmentation method based on speech analysis under the Microcanonical Multiscale Formalism (MMF). The latter relies on the computation of local geometrical parameters, singularity exponents (SE). We showed that SE convey valuable information about the local dynamics of speech that can readily and simply used to detect phoneme boundaries. By performing error analysis of our original algorithm, in this paper we propose a 2-steps technique which better exploits SE to improve the segmentation accuracy. In the first step, we detect the boundaries of the original signal and of a low-pass filtred version, and we consider the union of all detected boundaries as candidates. In the second step, we use a hypothesis test over the local SE distribution of the original signal to select the final boundaries. We carry out a detailed evaluation and comparison over the full training set of the TIMIT database which could be useful to other researchers for comparison purposes. The results show that the new algorithm not only outperforms the original one, but also is significantly much more accurate than state-of-the-art ones.

Index Terms— phonetic segmentation, non-linear speech processing, multiscale signal processing, complex signals and systems.

1. INTRODUCTION

In an earlier work [1], we proposed a radically new approach for text-independent (TI) phonetic segmentation based on the Microcanonical Multiscale Formalism (MMF). Using methods from the statistical physics field, MMF provides accurate analysis of the nonlinear dynamics of complex signals. It relies on the estimation of local geometrical parameters, the *singularity exponents* (SE), which quantify the degree of predictability at each point of the signal. When correctly defined and estimated, these exponents can provide valuable information about the local dynamics of complex signals and has been successfully used in many applications ranging from signal representation to inference and prediction [2, 3, 4]. We showed that MMF is valid for speech signal and that a simple analysis of SE yields a fast and relatively accurate TI phonetic segmentation algorithm. Experiments on TIMIT showed that, while our approach is simple and conceptually novel, it still achieves better segmentation accuracy than state-of-the-art methods. Moreover our algorithm is almost parameter/threshold free, which is a major advantage w.r.t. traditional algorithms and a desirable property for text and language independent applications.

By performing error analysis of our original MMF-based algorithm, in this paper we propose a technique which better shows the strength of SE and further improves the segmentation accuracy. This technique is a 2-step procedure. In the first step, we detect the phoneme boundary candidates on the given signal and on a low-pass filtered version using our MMF-based algorithm. We then consider the set of all detected boundaries (those of the signal and those of its filtered version) as candidate boundaries. In the second step, we use dynamic windowing and Log-Likelihood Ratio Test (LLRT) to decide which are the correct boundaries. This 2-step structure is similar to that of traditional segmentation methods where there is a boundary pre-selection followed by statistical tests to make the final decision [5].

Another objective of this paper is to address a common difficulty in comparing TI segmentation methods which is the diversity of evaluation datasets and also incoherencies in performance measures. Indeed, while the literature is rich in phonetic segmentation methods, it is relatively poor in material for performance comparison. Most of the papers report either on undefined subsets of known databases or on personal/unaccessible databases. Moreover, it is often difficult to analyze the reported accuracy scores because of the diversity of measures used. This makes it difficult to make fair comparisons between different segmentation algorithms. To the best of our knowledge, the only TI and unsupervised segmentation algorithms that report on known and accessible database are [6] and [7]. The latter reports on the full training set of TIMIT and has the advantage of providing different scores with different sizes of tolerance windows. [8] and [9] also report on the same dataset but the algorithms they propose are not fully unsupervised. The former assumes prior knowledge of the the number of phonemes in the utterance, while the latter uses prior knowledge of all manual transcriptions to train a neural network. We will thus report our results on the full training set of TIMIT and compare them to [7]. Using different sizes of tolereance windows, we provide detection, insertion and oversegmentation rates and also 2 different global performance measures, F_1 and R-value. By doing so, we attempt to provide results which are easy to interpret and compare with.

The paper is structured as follows. In Section 2 we briefly introduce MMF and our previous work on its application to phonetic segmentation. In Section 3 we present our new MMF based segmentation algorithm. The experimental results are presented in Section 4 and we draw our conclusion in Section 5.

2. OVERVIEW ON PREVIOUS WORK

In [1] the very first steps in applying MMF for catching non-linear dynamics of speech signal were taken. In this section, we briefly introduce MMF and our previous work on its application to the phonetic segmentation of speech signal.

2.1. Microcanonical Multiscale Formalism

MMF is based on the computation of the local scaling exponents of a given signal, whose distribution is the key quantity defining intermittent dynamics of the signal. These exponents are a useful tool for the study of geometrical properties of signals, and have been used in a wide variety of applications ranging from signal compression to inference and prediction [3, 4]. These exponents are associated to the evaluation of a local power-law scaling behavior at each point in the signal domain. The validity of MMF for a given signal $s(t)$ relies on the existence of such relationship for at least one scale-dependent functional Γ_r, for each time instance t and for small scales r:

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

where $h(t)$ is the so-called *singularity exponent* (SE). Turiel *et al.* [10] proposed a method for accurate estimation of SE by choosing Γ_r to be the gradient-modulus measure:

$$\Gamma_r\left(s(t)\right) \;:=\; \frac{1}{\Lambda(B_r)} \int_{B_r(t)} \mathrm{d}\tau \, |s'(\tau)| \tag{2}$$

It is shown that, the exponent associated to the corresponding power law characterizes the information content and the dynamical transitions of the signal in terms of the scale [11, 12]. Practical implementation to avoid noise and discretization artifacts consists in using a continuous wavelet transform $\mathbb{T}_\Psi\left[|s'|\right](r,t) \;\propto\; r^{h(t)}$. We use the Lorentzian wavelet because it provides an accurate estimation for small exponents which are the most informative ones [13]. In fact, for a given point, the smaller the value of SE is, the higher predictability is in the neighborhood of this point [10]. 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 [2, 4, 14].

2.2. Application of MMF to the phonetic segmentation

The validity of MMF and the availability of precise estimates of SE for speech signal, was proved by an extensive evaluation on TIMIT database. We then showed how SE convey valuable information about dynamical transitions of the speech signal. Indeed, since different phonemes should have different geometrical and statistical properties, we expected the corresponding SE to have different behavior inside the boundaries of each phoneme. This was verified by evaluating the time evolution of the distribution of SE. This led us to the development of an automatic segmentation algorithm, by exploiting the easiest interpretation of the changes in distributions, which is the change in averages. In other words, we expect that different phonemes have different averages of exponents compared to their neighboring phonemes. We proposed to use the primitive of the SEs function over time as an estimator of the instantaneous average:

$$ACC(t) \;=\; \int_{t_0}^{t} \mathrm{d}\tau \; h(\tau) \tag{3}$$

The resulting functional, with a detrending to enhance the presentation, is plotted in Figure 1a. As expected, this new functional revealed the changes in distribution in a more precise way. Indeed, inside each phoneme the functional ACC is almost linear. Moreover, there is a clear change in the slope at the phoneme boundaries. Extensive observations over different sentences confirm this behavior, and thus the strength of the proposed functional, Eq. (3).

In order to develop an automatic segmentation algorithm, a very simple solution was employed to fit a piecewise linear curve to ACC and identify the breaking points. To do so, we performed a left-to-right search to find the hypothesized boundaries as the points where the mean squared error of the linear fit is below a certain threshold. Finally we reject the nodes located in the silence as the points where the average power in a 30ms window is less than -30dB.

3. IMPROVING THE MMF-BASED SEGMENTATION

In this section we present a technique which better exploits the strength of singularity exponents in order to improve the accuracy of the phonetic segmentation algorithm described above. This technique is motivated by some observations about the behavior of the latter algorithm at some particular phoneme transitions. This leads us to propose a two-steps algorithm where we first pre-select candidate boundaries and then use statistical hypothesis test to make the final decision.

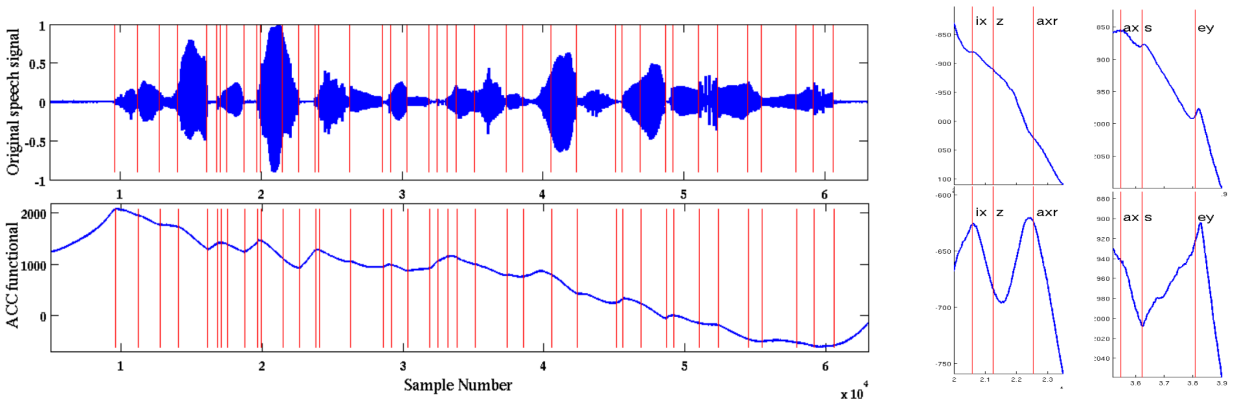
By performing error analysis of the MMF-based algorithm, we first observed that some of the misssed boundaries correspond to transitions between fricatives/stops and vowels. We also observed that transitions between speech and low energy segments (such as pauses and epenthetic silence) display strong and easy to detect changes in the slopes of ACC. Indeed, the singularity exponents 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 propose 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 much easier to detect as shown in Figure 1b. 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. We thus choose the cutoff frequency of the low-pass filter as 1800Hz.

The second and most important observation we made is related to the statistical distribution of singularity exponents. We observed that some missed boundaries correspond to neighboring phonemes which have a quite distinctive difference in their SE distribution. However, the change in their averages is not strong enough to be translated as a change in the slope of ACC, and thus it is not captured by the simple curve-fitting procedure. It is then natural to think about including a statistical hypothesis test over SE distributions in our segmentation algorithm in order to detect such boundaries.

Motivated by these observations, we develop new segmentation algorithm which consists in 2 steps. First, we use our original MMF-based algorithm to detect the boundaries of the original and filtered signal. We gather all the detected boundaries and consider them as candidate boundaries. In the second step, we make the final decision by performing a dynamic windowing over these candidates followed by Log Likelihood Ratio Test (LLRT) over SE distributions of the *original* signal. We use a Gaussian hypothesis because our purpose is to detect changes in mean and variance. 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 hypothesis:

- 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 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.



(a) (b)

Fig. 1: (a) A speech signal from TIMIT and its ACC functional. (b) **TOP:** Two examples of ACC functional for the original signal where the change in the slopes are not clear, **BOTTOM:** The ACC functional for the low-passed filtered signal. The changes in slopes are clearer. Phoneme boundaries are marked with vertical red lines.

4. EXPERIMENTAL RESULTS

Our evaluation is carried out on the *full* Train set of the TIMIT database which contains 4620 sentences uttered by 462 speakers. This set contains a wide variabilitiy of speakers and is balanced for dialectical coverage which is desirable for the evaluation of TI segmentation methods.

4.1. Performance measures

The segmentation quality can be evaluated and analyzed using three "partial" scores: the *Hit Rate* (HR) which is the rate of correctly detected boundaries; the *False Alarm Rate* (FA) which is the rate of erroneously detected boundaries and the *Over Segmentation Rate* (OS). These three scores are defined as:

$$HR = \frac{N_H}{N_R},\ OS = \frac{N_T - N_R}{N_R},\ FA = \frac{N_T - N_H}{N_T} \tag{4}$$

where, N_T is the total number of detected boundaries, N_H is the number of correctly detected boundaries and N_R is the total number of boundaries in the reference transcription. In order to 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:

$$F_1 = \frac{2 \times PCR \times HR}{PCR + HR} \tag{5}$$

where $PCR = 1 - FA$ is the *precession rate*. Another global measure, called the R-value, which is supposed to be more accurate than F_1 has been recently proposed in [15]. 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 $R = 1$:

$$r_1 = \sqrt{(1 - HR)^2 + OS^2},\ r_2 = \frac{HR - OS - 1}{\sqrt{2}} \tag{6}$$

$$R = 1 - \frac{|r_1| + |r_2|}{2} \tag{7}$$

4.2. Results

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 [7]. We mention here that [7] 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 MMF-based algorithm [1] summarized in section 2.2, we call it MMF-ACC. In the third one, we present the results obtained using our new algorithm described in section 3, we call it MMF-LLRT. Table 1 presents HR, FA and OS for the 3 methods. The first observation is that MMF-LLRT outperforms MMF-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 has been corrected by the LLRT. The second observation is that MMF-LLRT yields considerably much higher accuracy than [7]. In particular, the smaller tolerance window is, the higher relative improvement is. This shows that MMF-LLRT is better suited for high precision detection of phoneme boundaries. To this regard, we can mention another interesting comparison with [9] which is also a sample-based segmentation method as ours. In [9], 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 MMF-LLRT we obtain 44% hit rate which is significantly more accurate. More importantly, the algorithm in [9] is supervised (all manual transcriptions are used to train a neural network) while ours is fully unsupervised.

Table 2 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, MMF-LLRT still outperforms MMF-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 singularity exponents in revealing the transitions fronts between phonemes.

Finally we emphasize an important feature of our algorithm

Table 1: *The comparative table of segmentation results ("-" means not available). The scores are reported as percentages.*

tolerance	score	Dusan et al [7]	MMF-ACC
5ms	HR	22.8	31.7
	FA	79.7	70.2
	OS	12.8	6.4
10ms	HR	-	52.8
	FA	-	50.4
	OS	-	6.4
15ms	HR	59.2	65.5
	FA	47.5	38.4
	OS	12.8	6.4
20ms	HR	-	72.4
	FA	-	31.94
	OS	-	6.42

Table 2: *The comparative table of global performance measures.*

tolerance	score	Dusan et al [7]	MMF-ACC	MMF-LLRT
5ms	R-value	0.29	0.39	0.41
	F_1-value	0.21	0.31	0.32
10ms	R-value	-	0.57	0.60
	F_1-value	-	0.51	0.53
15ms	R-value	0.60	0.68	0.70
	F_1-value	0.55	0.63	0.65
20ms	R-value	-	0.74	0.76
	F_1-value	-	0.70	0.72
25ms	R-value	0.73	0.77	0.79
	F_1-value	0.71	0.74	0.75
30ms	R-value	-	0.79	0.81
	F_1-value	-	0.76	0.77

which is its insensitivity to the threshold of the linear curve fitting. Figure 2 displays the R-value for different thresholds. we used a subset of 30 randomly selected sentences to compute these values. One can see that with about 400% change in the value of threshold, that variance of changes in R-value is less than 0.5%. Thus, we can fairly consider that our algorithm is threshold-free. This is a major advantage as most of the TI methods require accurate threshold tuning.

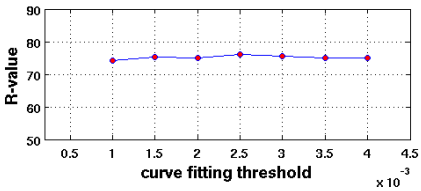


Fig. 2: The sensitivity of the MMF-LLRT to threshold.

5. CONCLUSIONS

By performing error analysis of our original MMF-based phonetic segmentation algorithm, we presented a new technique which confirms the strength of singularity exponents in detecting phoneme boundaries, and which further improves the segmentation accuracy. We provided a detailed evaluation and comparison using the full training set of TIMIT that could be useful to other researchers for comparison purposes. The results show that the new algorithm not only outperforms the original one, but also is significantly much more accurate than state-of-the-art ones. We are still at the beginning of exploration of speech analysis from the MMF perspective. The encouraging results we obtained so far suggest that the MMF has indeed big potential in speech processing and should be further investigated. This will be the purpose of future communications.

6. REFERENCES

[1] V. Khanagha, K. Daoudi, O. Pont, and Hussein Yahia, "A novel text-independent phonetic segmentation algorithm based on the microcanonical multiscale formalism," *Proceedings of INTERSPEECH2010*, 2010.

[2] 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.

[3] O. Pont, A. Turiel, and C.J. Perez-Vicente, "Empirical evidences of a common multifractal signature in economic, biological and physical systems," *Physica A*, vol. 388, no. 10, pp. 2025–2035, May 2009.

[4] 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, pp. 39–54, June 2009.

[5] G. Almpanidis, M. Kotti, and C. Kotropoulos, "Robust detection of phone boundaries using model selection criteria with few observations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 287–298, 2009.

[6] A. Esposito and G. Aversano, "Text independent methods for speech segmentation," in *Nonlinear Speech modelling. G. Chollet et al Eds.*, 2004, pp. 261–290.

[7] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," *Proceddings of INTERSPEECH/ICSLP 2006*, pp. 645–648, 2006.

[8] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," *Proceedings of ICASSP2008*, pp. 885–888, 2008.

[9] Y. Lin, Y. Wang, and Yuan-Fu Liao, "Phone boundary detection using sample-based acoustic parameters," *Proceedings of INTERSPEECH2010*, 2010.

[10] A. Turiel, H. Yahia, and C. P. Vicente., "Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis," *J. Phys. A, Math. Theor.*, vol. 41, pp. 015501, 2008.

[11] U. Frisch, *Turbulence: The legacy of A.N. Kolmogorov*, Cambridge Univ. Press, 1995.

[12] 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.

[13] A. Turiel and C. Pérez-Vicente, "Multifractal measures: definition, description, synthesis and analysis. a detailed study," in *Proceedings of the "Journées d'étude sur les méthodes pour les signaux complexes en traitement d'image"*, J.-P. Nadal, A. Turiel, and H. Yahia, Eds., Rocquencourt, 2004, pp. 41–57, INRIA.

[14] O. Pont, A. Turiel, and C. Pérez-Vicente, "On optimal wavelet bases for the realization of microcanonical cascade processes," *Int. J. Wavelets Multi.*, vol. 9, 2011.

[15] Okko J. Rasanen, Unto K. Laine, and Toomas Altosaar, "An improved speech segmentation quality measure: the R-value," *Proceedings of INTERSPEECH2009*, 2009.