APPLICATION OF THE MICROCANONICAL MULTISCALE FORMALISM TO SEGMENTATION OF SPEECH SIGNALS

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
phone: + (33) 0524574109, fax: + (33) 0524574124, email: vahid.khanagha@inria.fr
web: geostat.bordeaux.inria.fr

ABSTRACT

In this paper we use a novel framework, the Microcanonical Multiscale Formalism (MMF), to analyze speech signals. The MMF is based on the computation of geometrical and local parameters – the singularity exponents – which allow nonlinear analysis of their complex dynamics and, particularly, characterize their intermittent signature. We define an accumulative measure on these exponents which has the nice property of producing clear and distinctive changes at phoneme boundaries. We present preliminary experiments on the TIMIT database, which show that singular exponents convey indeed valuable information about the local dynamics of speech. They also show that the measure we define has a good potential to provide a new and powerful method for text-independent phonetic segmentation.

1. INTRODUCTION

It is theoretically and experimentally established that turbulence and high nonlinear phenomena are present in the speech production process [8, 2, 9, 10]. However, the traditional approach to speech processing is based on linear techniques which basically rely on the source-filter model. The linear approach cannot adequately take into account or capture the complex dynamics of speech. For this reason, nonlinear speech processing has gained a significant attention during the last years.

In this paper we analyze the nonlinear dynamics of speech using concepts and methods from the framework of turbulent systems. Our approach is based on the Microcanonical Multiscale Formalism (MMF) which is a novel framework to study the geometric-statistical properties of complex signals from a multiscale perspective [12, 17]. The MMF has proved to be a valuable approach to model and analyze empirical complex and turbulent systems. This is particularly true for scale-invariant systems, i.e., systems that have corresponding statistical properties at different scales [11].

The MMF is an extension of its more standard Canonical counterpart [5, 1]. The particularity of MMF is that it is based on geometrical and local parameters, rather than relying on statistical averages – such as structure functions or partition functions – as it is the case in the canonical framework [11]. Hence, MMF makes it possible to locally study the dynamics of complex signals.

In this paper we show that speech signals lie in the domain of applicability of MMF. We then use the local parameters computed by the MMF, called singularity exponents [18], and show how they convey meaningful information for the identification of phoneme boundaries.

Speech segmentation has many potential applications in speech technology, from speech synthesis to Automatic Speech Recognition (ASR). Segmentation could be the first stage of an ASR systems, but the lack of satisfying segmentation algorithms has led to a reversed approach: a wide class of segmentation methods are the adapted versions of HMM-based phonetic recognizers [15]. This class of segmentation methods are known as text-dependent methods since they rely on an externally supplied database of target vocabulary and its manual transcriptions. On the other hand, there exists a class of text-independent segmentation methods which are based on the identification of variations in feature-based distances [3]. Text independent methods are not limited to a specific corpus and they rely on either some model-based feature vectors or some raw spectral measures.

In this paper we exploit the behavioural changes in distribution of singularity exponents over time, through the use of an accumulative measure. We present preliminary experiments that show that this measure can be readily used to detect phoneme boundaries.

The paper is structured as follows. In Section 2 we introduce the basic concepts of MMF, the algorithm for singularity exponents estimation, and then present the validation procedure for a given signal to be evaluated under MMF. In Section 3 we show that speech is an appropriate candidate for such formalism. In Section 4 we discuss the use of singularity exponents for segmentation of speech signals. Finally, in Section 5 we draw our conclusions.

2. MICROCANONICAL MULTISCALE FORMALISM

In this section we give a brief overview on the basics of MMF. A more extensive review of the theory and tools can be found in [17].

MMF is based on the computation of the local scaling exponents of a given signal, whose distribution is the key quantity defining its intermittent dynamics. 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 [16, 13].

Before applying MMF to a given signal, the first step is to study its validity for that signal. The validity of

 ISSN 2076-1465

MMF for a signal relies on the existence of a local power-law scaling behaviour at each point in the signal domain [17]. Formally, for at least one scale-dependent functional Γ_r, the following relation must hold for any time t and for small scales r:

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

where $h(t)$ is the so-called singularity exponent [17]. The multiplicative factor $\alpha(t)$ depends on the chosen functional Γ_r, but for some systems such as scale-invariant ones, the exponent $h(t)$ is independent of it. However it is beyond the scope of this paper to study whether speech has scale-invariance properties. The term $\mathrm{o}\left(r^{h(t)}\right)$ means that for very small scales the additive terms are negligible compared to the factor and thus $h(t)$ dominantly quantifies the degree of "regularity" of $s(t)$ at each time instance.

If the functional is chosen as the linear increment, $\Gamma_r\left(s(t)\right) = s(t+r) - s(t)$, the resulting exponents are Hölder exponents and they characterize causal power-law correlations. When empirical data are analyzed, it is often difficult to obtain good estimation of Hölder exponents from linear increments. Discretization, noise and long-range correlations hinder the practical calculation of these exponents from Eq. (1).

There is an alternative and more robust definition for the functional Γ_r in Eq. (1), which is defined from the typical characterization of intermittence in turbulence: the gradient-modulus measure. This measure on a ball of radius r for a turbulent velocity field describes the kinetic energy dissipation at scale r. Therefore, it is a quantity linked to the transfer of energy from one scale to another. Thus, the exponent associated to the power law in terms of the scale characterizes the information content and the dynamical transitions of the signal [5, 18]. The functional is defined as the r-radius gradient-modulus measure divided by the volume of the r-radius ball:

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

where s' is the derivative of s, B_r is the r-radius ball and Λ means the Lebesgue measure on the real line. Practical implementation to avoid noise and discretization artifacts consists in using a wavelet support for the ball $B_r(t)$.

Finally, we mention the importance of a particular set of points that convey most of informations about the nonlinear dynamics of signal: the most singular component. In fact, for a given point, the smaller value of singularity exponent, the higher predictability implied at this point [16]. It is has been established that the critical transitions of the system occur at these most singular points, and this fact has been successfully used in many applications such as edge detection or data reconstruction [17].

2.1 Estimation of singularity exponents

The method used in this paper to estimate singularity exponents is the continuous-wavelet-transform equivalent of Eq. (2). This has the general advantage of coping with the particularities of real-world data such as discretization, acquisition noise and long-range correlations. Also, wavelet transform vanishes polynomial contributions in the additive term $\mathrm{o}\left(r^{h(t)}\right)$ which are a common obstacle for the precise estimation of singularity exponents. Overall, we examine the following power-law relationship for each time instance:

$$\mathbb{T}_\Psi\left[|s'|\right](r,t) \;\propto\; r^{h(t)} \qquad (3)$$

where $\mathbb{T}_\Psi\left[x\right](r,t) := \left(\Psi_r * x\right)(t)$ stands for the continuous wavelet transform, $\Psi_r(t) := r^{-1}\,\Psi(r/t)$ and Ψ is a wave-like function called mother wavelet.

It is appropriate to mention another advantage of using the continuous wavelet transform for this estimation: the possibility of computing the transform over a set of non-integer scales in discretized signals. The scale variable r in Eq. (3) could be assigned any non-integer value, providing a smooth interpolation scheme for the discrete-time signal.

2.2 MMF validation

It is easy to see that taking the logarithm of both sides of Eq. (3) reveals a linear relationship between the logarithm of the wavelet transform and the logarithm of the scale. So it is possible to estimate the singularity exponent $h(t)$ at each time t by performing a linear regression of the wavelet transform vs. the scale in a log-log plot. Therefore, Eq. (3) and consequently Eq. (1) are verified for a given signal if we attain acceptable correlation coefficients for such linear regression. When this occurs, the MMF is valid for the signal.

3. SPEECH SIGNALS IN THE MICROCANONICAL FRAMEWORK

In this section we study the validity of MMF for speech signals. In order to estimate the singularity exponents, we use a slight modification of Eq. (3). Indeed, in our experiment with speech signals, we observed that better correlation coefficients are obtained when we take the logarithm of both sides of Eq. (3) and divide with $\log(r)$, so that we have the linear relationship:

$$\frac{\log \mathbb{T}_\Psi\left[s'\right](r,t)}{\log r} \;=\; \alpha_\Psi(t)\,\frac{1}{\log r} + h(t) \qquad (4)$$

Then, by performing the linear regression, the singularity exponent $h(t)$ is estimated as the bias of this linear relationship.

The wavelet we use is the Lorentzian wavelet. This wavelet defines an accurate estimation for smaller exponents, at the expense of a saturation of all exponents ≥ 1 [19]. This is desirable because small exponents are the most informative ones and, in the presented case, retrieved exponents are far from the saturation and so it does not appear as an actual limitation. To perform the regression we chose 10 scales which are log-uniformly spaced between 1 and 100 samples (which correspond to the interval from 62.5 μs to 6.25 ms).

We first check the existence of the power-law scaling Eq. (3) on phonemes, as they are the basic acoustic speech units. All our experiments are carried out on the TIMIT database [6]. We use the transcriptions provided by TIMIT to construct a test database of 3000

phonemes: for each phoneme family (vowels, fricatives, stops, semi-vowels and glides, affricates and nasals) we take 500 different instances of a representative phoneme. The estimation of singularity exponents using Eq. (4) is performed on the 500 instances and the resulting average correlation coefficients are reported in Table 1.

Next, we performed the same procedure over whole sentences. 500 different speech signals with an approximate length of 3–5 seconds were used for this experiment. We obtained an average correlation coefficient of 0.96. The small loss compared to the average of values in Table 1 (which is 0.98) is explained by the presence of long segments of silence when we process whole sentences.

Overall, these experiments show that excellent correlation coefficients are obtained using our estimation procedure. First, this suggests that we achieve very precise estimation of the singularity exponents. Second, it suggests that the MMF is valid for speech signals. We can thus proceed now to study how these exponents convey useful information about the speech dynamics.

4. APPLICATION OF MMF FOR SPEECH SEGMENTATION

Speech is a non-stationary signal which is formed by concatenation of small acoustic units called phonemes. The automatic detection of boundaries between phonemes is a challenging task and is still an open problem which has many applications in speech technology. In Section 3 we demonstrated the validity of MMF for speech signals. Here, we present our observations on the instructive information of singularity exponents about the variable temporal dynamics of speech signals.

Since different phonemes are basically different signals with different frequency content and statistical properties, we expect the corresponding singularity exponents to have different behaviour inside the boundaries of each phoneme. In order to demonstrate these changes, we provide a graphical presentation in Figure 1, showing the changes in distribution of singularity exponents conditioned on the time, $\rho(h|t)$.

In Figure 1–top, the original speech signal is shown and the phoneme boundaries are represented by vertical red lines. These boundaries are extracted from the manual transcription of TIMIT database. Figure 1–middle displays the time evolution of the conditional distribution of singularity exponents. In the vertical axes we show the rank of the singularity exponents in bins of 5 percentiles. Then, at each time instance t, we take a 30 ms window centred around t and we accumulate the exponents to the globally computed bins. As we want to represent conditional probability each row is norm-∞ normalized. It is remarkable that there is a change in the position of maxima and in the variabilities of h distribution. Moreover, the distribution alternates from uni-modal to multi-modal, with uni-modal cases centred at the middle of the global range and multi-modal cases typically with two modes: one at each extreme of the range.

However, although these changes in distribution behaviour are visually apparent they would be extremely difficult to detect numerically and automatically. Hence, it would be appropriate to define a new measure to exploit these distributive changes. With this purpose in mind, notice that the easiest interpretation of the changes in distributions of Figure 1–middle is the change in averages. In other words, we expect that different phonemes have different averages of singularity exponents compared to their neighbouring phonemes. In order to check this, we use the primitive of the singularity-exponent function over time as an estimator of the instantaneous average. Formally, we define the new functional as:

$$ACC(t) = \int_{t_0}^{t} \mathrm{d}\tau \; h(\tau) \tag{5}$$

The resulting functional is plotted in Figure 1–bottom for the same speech signal as before. To enhance presentation of the values of resulting time varying function in an observable window, we detrend it. Just as we expected, this new functional reveals the changes in distribution in a more precise way. Indeed, inside each phoneme the functional ACC is almost linear (if we neglect the small scale fluctuations). Moreover, there is a clear change in the slope at the phoneme boundaries. These slope changes are even able to identify the boundaries between extremely short phonemes, such as stops. Extensive observations over different sentences confirms this behaviour, and thus the strength of the proposed functional, Eq. (5).

These experiments suggest that the singularity exponents computed in the MMF convey indeed meaningful information about the critical transitions in speech signals. They also suggest that we can readily use these exponents to develop a new, robust method for phonetic segmentation.

An accurate evaluation of such method requires the implementation of a numerical algorithm for unsupervised identification of breaking points in noisy piecewise linear curves. This is the purpose of our ongoing research. At the time of writing of this paper, we are studying the use of Free Knot B-spline algorithms [14] to achieve this goal.

The closest methods to ours that we found in the literature are the ones in [7] and [4]. In the former, the analysis of trajectory of the Variance Fractal Dimension (VFD) is used for phonetic segmentation. In the latter, the authors propose a fractal based approach which uses the transitions of the envelope of the local fractal dimension to determine the boundaries between words and phonemes. Performing an extensive comparison between these methods and ours is beyond the scope of this paper, particularly since these methods – like ours, for the moment – visually differentiate the phonemes without giving an automatic segmentation procedure. However, in our approach the latter could be solved by a simple piece-wise linear approximation. Thus we can fairly say that our approach is much easier to incorporate in an automatic segmentation algorithm than the measures given in [7] and [4].

5. CONCLUSIONS

In this paper we first showed that MMF is a valid framework for the study of speech signals. From this perspective, we then analyzed the local properties of speech sig-

Phoneme type	Vowel	Fricative	Stop	Semi Vowel & Glide	Affricate	Nasal
Phoneme	/aa/	/dh/	/b/	/el/	/ch/	/en/
Average Correlation Coef.	0.97	0.99	0.99	0.99	0.99	0.99

Table 1: The average correlation coefficients of the linear regression Eq. (4) for a representative phoneme of each 6 different families.

nals through the singularity exponents computed in the MMF. We showed that these exponents are interestingly informative about the speech dynamics. Finally we proposed a geometrical quantifying measure that produces clear and distinctive changes at phoneme boundaries, and thus can be used for automatic text-independent phonetic segmentation. We emphasize that the complete application of the MMF framework for speech signal requires more accurate justifications to cope with all the particularities of speech signals. Still, the study presented in this paper reveals the informativeness of the singularity exponents, without any extra manipulations.

REFERENCES

[1] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy. *Ondelettes, multifractales et turbulence*. Diderot Editeur, Paris, France, 1995.

[2] A. Barney, C. Shadle, and P. Davies. Fluid flow in a dynamical mechanical model of the vocal folds and tract: part 1 & 2. *J. Acoust. Soc. Amer.*, 105(1):444–466, Nov. 1999.

[3] A. Esposito and G. Aversano. Text independent methods for speech segmentation. In *Summer School on Neural Networks 2004*, pages 261–290, 2004.

[4] P. C. Fantinato, R. C. Guido, S.-H. Chen, B. L. S. Santos, L. S. Vieira, S. B. J, L. C. Rodrigues, F. Sanchez, J. Escola, L. M. Souza, C. D. Maciel, P. R. Scalassara, and J. Pereira. A fractal-based approach for speech segmentation. In *Tenth IEEE International Symposium on Multimedia*, pages 551–555, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

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

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus. Technical report, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[7] W. Kinsner and W. Grieder. Speech segmentation using multifractal measures and amplification of signal features. In *Cognitive Informatics, 7th IEEE International Conference on*, Oct. 2008.

[8] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6):1098–1109, Jan. 2005.

[9] A. Kumar and S. Mullick. Nonlinear dynamical analysis of speech. *J. Acoust. Soc. Amer.*, 100(1):615–629, 1996.

[10] M. Little. *Biomechanically Informed Nonlinear Speech Signal Processing*. PhD thesis, Oxford University, 2007.

[11] O. Pont, A. Turiel, and C. Perez-Vicente. Empirical evidences of a common multifractal signature in economic, biological and physical systems. *Physica A*, 388(10):2025–2035, May 2009.

[12] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Application of the microcanonical multifractal formalism to monofractal systems. *Physical Review E*, 74:061110–061123, 2006.

[13] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Description, modeling and forecasting of data with optimal wavelets. *Journal of Economic Interaction and Coordination*, 4(1):39–54, June 2009.

[14] H. Schwetlick and T. Schutze. Least squares approximation by splines with free knots. *BIT Numerical Mathematics*, 35(3):361–384, Sep. 1995.

[15] D. Torre-Toledano, L. Hernandez-Gomez, and L. Villarrubia-Grande. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625, 2003.

[16] A. Turiel and A. del Pozo. Reconstructing images from their most singular fractal manifold. *IEEE Trans. on Im. Proc.*, 11:345–350, 2002.

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

[18] A. Turiel and N. Parga. The multi-fractal structure of contrast changes innatural images: from sharp edges to textures. *Neural Computation*, 12:763–793, 2000.

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

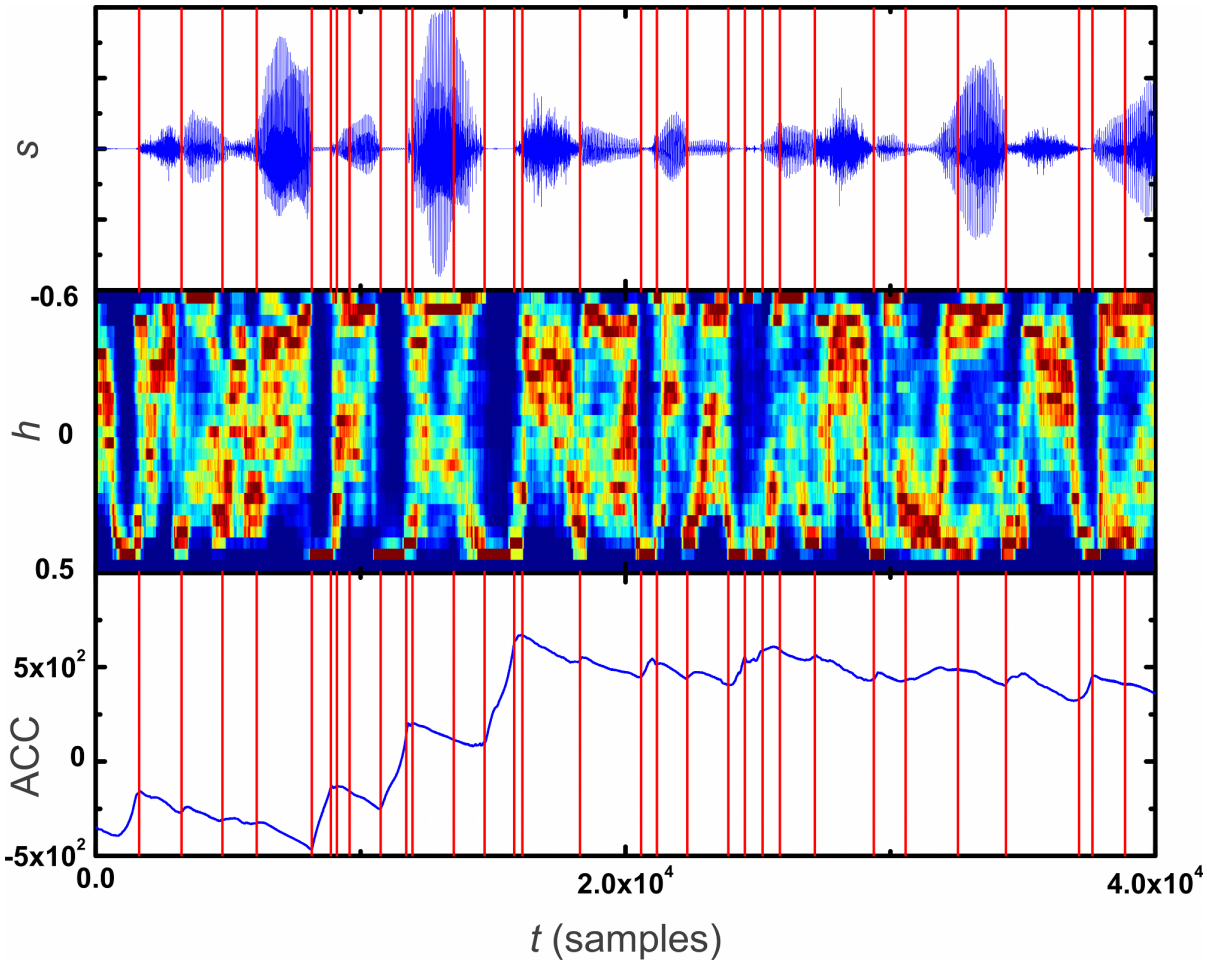

Figure 1: **TOP:** A normalized speech signal from TIMIT database. The signal was sampled at 16 kHz. Manually-positioned phoneme boundaries are marked with vertical red lines. **MIDDLE:** Joint histogram of the distribution of singularity exponents (vertical axis) conditioned to the time window (horizontal axis). Red corresponds to maximum probability and dark blue corresponds to zero probability. The horizontal axis is divided in 30 ms bins. The vertical axis is divided in global 5-percentile bins, so that it is proportional to the global rank of the singularity exponents, not to their value. This avoids low-probability distortions. **BOTTOM:** Proposed functional for the identification of phoneme boundaries. It is remarkable that most phoneme boundaries co-localize with strong changes in slope. To enhance presentation, the *ACC* functional presented in Eq. (5) has been globally detrended (for the whole sentence, not the presented portion).