# Non-linear speech representation based on local predictability exponents

V. Khanagha, K. Daoudi, O. Pont, H. Yahia, A. Turiel*

*INRIA, GeoStat team, http://geostat.bordeaux.inria.fr*
*200 Avenue de la Viell Tour, 33405 Talence cedex, France*
*Email: {vahid.khanagha, khalid.daoudi, oriol.pont, hussein.yahia}@inria.fr*

**ICM-CSIC, Physical Oceanography Department, Barcelona, Spain*
*Email: {turiel}@icm.csic.es*

## Abstract

Looking for new perspectives to analyze non-linear dynamics of speech, this paper presents a novel approach based on a microcanonical multiscale formulation which allows the geometric and statistical description of multiscale properties of the complex dynamics. Speech is a complex system whose dynamics can be, to some extent, geometrically and statistically accessed by the computation of Local Predictability Exponents (LPEs) unlocking the determination of the most informative subset (Most Singular Manifold or MSM), leading to associated compact representation and reconstruction. But the complex intertwining of different dynamics in speech (added to purely turbulent descriptions) suggests the definition of appropriate multiscale functionals that might influence the evaluation of LPEs, hence leading to more compact MSM. Consequently, by using the classical and generic Sauer/Allebach algorithm for signal reconstruction from irregularly spaced samples, we show that speech reconstruction of good quality can be achieved using MSM of low cardinality. Moreover, in order to further show the potential of the new methodology, we develop a simple and efficient waveform coder which achieves almost the same level of perceptual quality as a standard coder, while having a lower bit-rate.

*Keywords:* nonlinear speech processing, multiscale signal processing, complex signals and system

## 1. Introduction

The existence of highly non-linear and turbulent phenomena in the 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 which basically rely on the source-filter model. These linear approaches cannot adequately take into account or capture the complex dynamics of speech (despite their undeniable importance). For instance, it has been shown that the Gaussian linear prediction analysis, which is a ubiquitous technique in current speech technologies, cannot be used to extract all the dynamical structure of real speech time series (for all simple vowels of US English and for both male and female speakers) [3]. Among numerous methods proposed for non-linear speech processing [4], a wide class considers speech as a non-linear dynamical system and attempts to use the available tools and methods in the study of such systems to catch non-linear features of the signal. For example in [5], Minkowski-Bouligand dimensions of speech (related to the amount of turbulence in a speech sound) is used for phoneme classification. Also in [6], Lyapunov Exponents (associated to the degree of chaos or predictability in a dynamical system) are used to model and analyze speech by relating its non-linear character to the concept of predictability. However, apart from the practical issues in the estimation of such quantities for speech signal (because of its time-varying dynamics), they can only provide a global characterization of the system's dynamics.

Recently, we have been conducting research in non-linear speech analysis by using principles and analogies coming from the study of complex systems in statistical physics. Our strategy is also to relate the non-linear character of speech signal to the concept of predictability [7] but in a *local* manner. To do so, we use a microcanonical approach which relies on the precise evaluation of local geometrical scaling parameters, that we call Local Predictability Exponents (LPEs) whose values are naturally associated to criticality. Critical transitions in the system are described by the values of LPEs, the latter being computed in the speech signal domain itself. This property can be used for geometric localization of a subset of points, called the Most Singular Manifold (MSM) composed of the points with the lowest values of LPEs. It has been shown that some signals can be reconstructed using only the information contained in the MSM [8] and hence MSM can be considered as the most informative subset.

The computation of LPEs relies on the choice of some multiscale functionals operating on the signal. In a purely turbulent signal, with no more regular dynamics superimposed, different multiscale functionals should lead to the same LPEs [9]. But the physical process of speech production indicates the existence of different dynamics added to the purely turbulent ones. How these added dynamics influence compact representations, correlations and the determination of most informative subsets open fascinating areas of research. In [10], a new definition for this multi-scale functional is discussed for the case of 2-D natural images, which is based on a recently patented algorithm [11]. The definition is based on the connection of LPEs with the concept of image reconstruction from the MSM, as it uses a discretized combinatorial mask which makes a local evaluation of a 2-D reconstruction kernel and hence, the highest quality of reconstruction from the MSM is yielded. We presented in [12], preliminary experiments about how by a direct 1D adaptation of the method in [11], MSM can provide a relatively parsimonious representation of speech. In this paper we introduce an alternative multiscale functional for the particular case of the speech signal, which leads to a parsimonious MSM which permits speech reconstruction of better quality. As the MSM corresponds to a subset of irregularly spaced speech samples, in this paper we use a classical method for the interpolation of irregularly spaced samples, the Sauer-Allebach algorithm [13], to reconstruct the speech signal from its MSM. We show that by using this simple and generic algorithm, and even by violating its conditions, good quality speech reconstruction can still be achieved from a MSM of low cardinality. This demonstrates that we do indeed capture a highly informative subset with the new multiscale functional. We then use the representation given by irregularly spaced MSM samples to build a *waveform* speech coder. Irregular sampling for waveform speech coding has been used in [14, 15]. They both take local extrema of the speech signal as irregularly spaced samples (hence, contrary to the MSM, the size of their sampling is constant and is not controllable) and reconstruction is done either through sinusoidal interpolation [14], local polynomial interpolation or cubic spline interpolation [15]. In this paper, In order to show the potential of our methodology in *waveform* speech coding, we use a very simple quantization and encoding of the MSM to develop a fully operational waveform coder and we compare it to the standard G.726 ADPCM [16] waveform coder. We show that our coder achieves almost the same level of perceptual quality as G.726, while having a lower bit-rate.

This paper is organized as follows: section 2 provides a brief review of the basic concepts of the microcanonical formalism we use including the general definition of LPEs. In section 3 we introduce a new multiscale functional for the computation of LPEs and we present the reconstruction algorithm. In section 4 the experimental results are presented and finally, in section 5, we draw our conclusions.

## 2. The microcanonical approach to complexity

The approach we use for the non-linear analysis of speech is a novel framework to study the geometrico-statistical properties of complex signals from a multiscale perspective [9]. It can be seen as an extension of previous approaches [17] for the analysis of turbulent data, in the sense that it considers quantities defined at each point of the signal's domain, instead of averages used in canonical formulations (moments and structure functions) [18]. Central to the approach is the computation of LPEs at every point in a signal's domain which unlocks the relations between geometry and statistics in a complex signal. When correctly defined and estimated, these exponents alone can provide valuable information about the local dynamics of complex signals and have been successfully used in many applications ranging from signal compression to inference and prediction [8]. The scaling exponent $h(t)$ for any given $d$-dimensional signal $s(t)$, can be estimated by the evaluation of the limiting power-law scaling behavior of a multiscale functional $\Gamma_r$ over a set of fine scales $r$:

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

where $\Gamma_r\left(\cdot\right)$ can be any multi-scale functional complying with this power-law. Many choices are possible for $\Gamma_r\left(\cdot\right)$: it can be the gradient-based measure introduced in [9] or the one we will later introduce in section 3.1. The term $\mathrm{o}\left(r^{d+h(t)}\right)$ means that the additive terms are negligible and thus $h(t)$ dominantly quantifies the multiscale behavior of the signal at each time instant $t$. Once $\Gamma_r\left(\cdot\right)$ is selected, $h(t)$ can be estimated by multiscale evaluation of Eq. 1. To do so, we use a computationally efficient method proposed in [19] as will be explained in section 3.1.

After precise estimation of $h(t)$, the signal can be hierarchically decomposed into the subsets sharing similar multiscale behavior. In fact, the geometrical distribution of such points can be defined as the level sets of the form $\mathcal{F}_h = \{t \in \Omega \mid h(t) = h\}$ ($\Omega$ is the domain of $t$). This arrangement is naturally related to the notion of information content of a large class of signals and there is a particular set among these level sets, which carries most of the information content of the signal, in the statistical sense. This set is called the Most Singular Manifold and comprises the points having the smallest LPEs, which provide indications in the acquired signal (a pressure i.e. an intensive physical variable) of most critical transitions of the associated dynamics [9]. These are the points where sharp and sudden local variations take place and hence they have the lowest possible predictability from their neighboring points. The formal definition of MSM reads:

$$\mathcal{F}_\infty = \{t \in \Omega \mid h(t) = h_\infty\}, \qquad h_\infty = min(h(t)) \tag{2}$$

In practice, once the signal is discretized, $h_\infty$ should be defined within a certain quantization range and hence MSM is formed as a set of points where $h(t)$ is below a certain threshold. It has been shown that [9] for many real world signals, the whole signal can be reconstructed using only the information carried by the MSM. For example, a reconstruction operator is defined for natural images in [9] which allows very accurate reconstruction of the whole image when applied to its gradient information over the MSM. The reconstruction quality can be further improved, using the $\Gamma_r\left(s(t)\right)$ measure defined in [10] which makes a local evaluation of the reconstruction operator to penalize the unpredictability. We used direct 1D adaptation of the latter 2D approach, to the case of speech signal and we showed that it is possible to form a compact MSM from which the whole speech signal is reconstructible with a reasonable perceptual quality [12]. In this paper we proceed to further study the particularities of the speech signal being analyzed within this microcanonical formalism.

## 3. LPEs and inter-sample dependencies

The raw 1-D adaptation of the methodology proposed in [10] to the speech signal and the associated reconstruction of the speech signal from its MSM is described in [12]. However, it is known that speech sampled at the Nyquist rate or faster exhibits significant correlation between successive samples [20]. Indeed, inter-sample dependencies imply a certain degree of predictability of each sample from its direct neighboring ones. This predictability, certainly introduces some redundancy to any form of speech representation which does not take these correlations into account. Instead, one can remove the predictable portion of any given sample (from its previous samples) and concentrate on the true information contribution of each sample. This strategy, for instance is central to the Differential Pulse Code Modulation (DPCM) which encodes the difference between samples and their predictions, and thus leads to superior compression performance compared to the simple PCM [20].

### *3.1. A new multiscale functional for the estimation of LPEs*

Our goal is to define a multiscale functional $\Gamma_r(s(t))$ for the specific case of speech, which takes into account these inter-sample dependencies. The simplest solution to remove the samples redundancy is to work on the differences between successive samples (as in the basic form of DPCM [21]). But the inter-sample correlations are also scale-dependent (within specific limits imposed by the system) because of the wide variation of scales in which these dependencies appear. Clearly any single-scale measure will fail in providing a balanced evaluation of predictability about the smoother parts of the speech signal (like voiced parts) and the noise-like behavior of other parts (like unvoiced speech). Thus, we must provide a functional for cross-scales inter-samples dependencies. Taking these considerations into account, we define the following multiscale functional operating at the scale $r$ on the signal $s(t)$:

$$\Gamma_r(s(t)) \;=\; |s(t) - \frac{s(t-r)}{2} - \frac{s(t+r)}{2}| \tag{3}$$

This $\Gamma_r(s(t))$ must be evaluated in accordance with the power-law of Eq. (1) over a set of scales. To do so, we use the theoretical developments in [19] which associates the power-law scaling of Eq. (1) to the existence of an underlying cascade process. In such processes, energy or information is transferred between scale levels of the signal. This way, the MSM actually corresponds to the set of points where information concentrates as it transfers across scales and, in that sense, it is a *least predictable/reconstructible manifold.* The cascade variable of this process must follow an infinitely divisible distribution; a property which permits a simple estimation of the desired scaling exponents, as the sum of a set of *transitional* exponents [19]:

$$h(t) = \sum_{i=1}^{k} h_{r_i}(t) \tag{4}$$

where $h_{r_i}(t)$ are the *transitional* exponents, which can be computed by direct evaluation of Eq. (1) at each scale, using the proposed measure in Eq. (3):

$$h_{r_i}(t) = \frac{log(\Gamma_{r_i}(s(t)))}{log(r_i/f_s)} \tag{5}$$

where $f_s$ is the sampling frequency of the signal. The LPEs computed according to Eq. (4) are then being used to form the MSM as explained in section 2.

### *3.2. Speech reconstruction from the MSM*

For the reconstruction of signal from the MSM, the Shanon's interpolation principle can not be applied since the samples in the MSM are irregularly spaced. Instead, we make use of a classical method in the interpolation of irregularly spaced samples, called the Sauer-Allebach algorithm [13]. The latter guaranties perfect reconstruction of any band-limited discrete signal $s[k]$ of length $N$ from irregularly spaced samples $s[k_i], k_i = 1..p < N$, with the condition that the maximal gap between samples must be less than the Nyquist rate. The algorithm consists in performing the following recursion:

$$\begin{gathered} s_0[k] = P_{B_\omega}(\mathcal{A}s[k_i]) \\ s_{n+1}[k] = s_n[k] + \lambda P_{B_\omega}(\mathcal{A}e_n[k_i]), \quad e_n[k_i] = s[k_i] - s_n[k_i] \end{gathered} \tag{6}$$

where $\mathcal{A}$ is an approximation operator (linear interpolation for instance) and $P_{B_\omega}$ is a low-pass filter matched to the original $B_\omega$-band limited signal. It is proven that $s_{n+1}$ converges to the original signal with a geometric rate [22]. However, in case of reconstruction of speech signal from its MSM, there are some problems in the direct application of the Sauer-Allebach algorithm. Indeed, the MSM structure generally violates the maximal gap condition. Also, the time-varying nature of speech does not permit the determination of a unique $B_\omega$ to be used for the whole signal. In fact, the unvoiced parts may have effective frequency content up to $8kHz$ while for the voiced parts there is no considerable activity for frequencies higher than $4kHz$. So the application of the above algorithm is not straightforward. We thus use a simple frame-based implementation of the interpolation algorithm to account for the time-varying nature of speech. In non-overlapping frames of length 20 ms, we first make a simple voiced/unvoiced decision: if the average distance

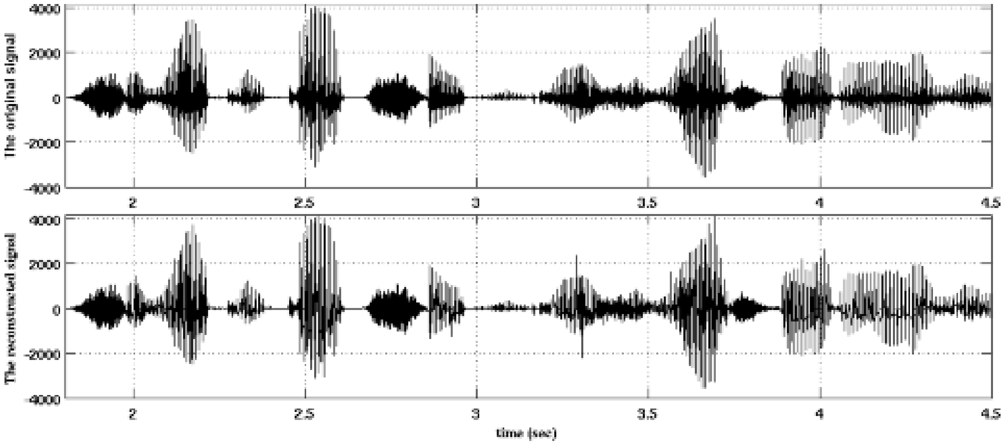

Figure 1: Waveforms of [top] the original signal and [bottom] the reconstructed signal from an MSM containing 14.7% of the points.

of successive MSM points is more than a specific value (depending on the sampling frequency), we consider the frame as a voiced frame and we set $B_\omega = 4kHz$. Otherwise we put $B_\omega = 8kHz$ to maintain the quality of unvoiced frames. Of course, since we are violating the maximal gap condition of the Sauer-Allebach reconstruction algorithm, we can not expect perfect reconstruction of the original signal. However, as experimental results will show, a good quality of reconstruction is still achievable.

## 4. Experimental results

Our experiments are carried out on about 2.5 hours of speech signal, composed of 3000 utterances from speakers of different sexes, accentes and ages, which are randomly taken from the TIMIT database. The overal Voice Activity Factor (VAF) of the dataset is 0.91.

In our application we compute the $h_{r_i}(t)$ for the four finest scales ($k = 4$) and use them to compute the exponent using Eq. (4). Then, we can use these LPEs to form the MSM as explained in section 2. Practically, one can sort all the samples of a given signal according to their value of LPE and take as many points as necessary to achieve a desired level of reconstruction in terms of Mean Squared Error (MSE). However, considering the time-varying nature of speech, a global formation of MSM would be problematic. To account for such time-varying characteristics, we perform a *local* formation of the MSM in non-overlapping frames of length 20 ms. In each frame, we

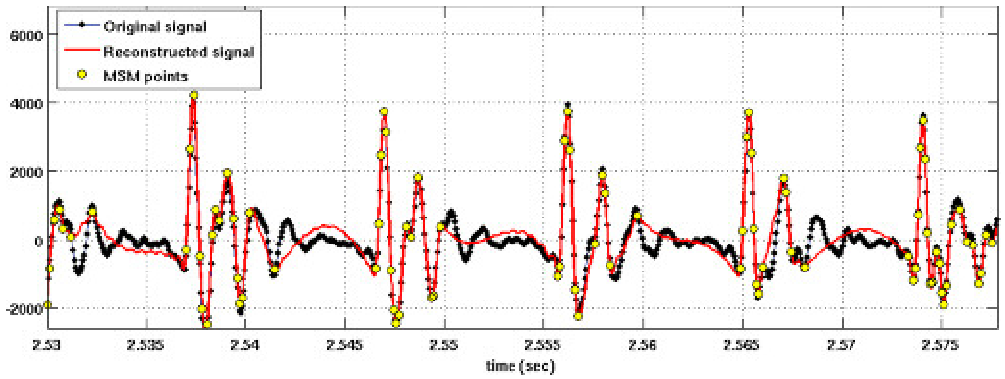

Figure 2: Waveforms of the original signal and the reconstructed signal. Samples belonging to MSM are marked with yellow circles.

sort the samples according to their value of $h(t)$ and we take as many samples as necessary to achieve the desired level of MSE from the *local* reconstruction. We then use the ensemble of these points to form the final MSM. Finally, we use the method explained in section 3.2 to reconstruct the whole signal from the MSM.

Figure 1-top shows a segment of speech signal from the dataset and the reconstructed waveform, from a MSM containing 14.7% of samples, is shown in Figure 1-bottom. It can be seen that the overall shape of the speech signal with its diverse dynamical regimes is preserved. There exist few occasional peaks (such as the one around $t = 3.25$ sec), which are duo to rare events of very large distances between consecutive MSM points (much more than Nyquist limit). However, the frequency of occurrence of these events is too low, such that they have no significant effect on the perceived quality. Figure 2 shows a zoom on the composition of the MSM and the corresponding waveform reconstruction.

We compare the quality of reconstrction for three different sampling methods: the MSM formed using our proposed measure in Eq. (3), the MSM formed by the measure of [11] and the subset of equally spaced samples. The same reconstruction method is used for all these three sampling methods as proposed in section 3.2. We emphasize that the reconstruction we use in this paper is different from the simple interpolation we use in [12]. The most accurate method for evaluating speech quality is believed to be the subjective listening test [23]. Our informal subjective listening tests confirm that the perceptual quality of the reconstructed signal from an MSM containing 14.7%

of the points is quite natural. Also, the quality is clearly better when compared to [11] (with the same size of MSM). In addition, to confirm that such quality can not be achieved by the application of Sauer-Allebach reconstruction algorithm to any subset of points, the results of the reconstruction from the MSM are compared with that of the subset of equally spaced samples (of the same size). In this case, our method gives much better quality.

In order to have an objective evaluation of the relative quality of different reconstructions, we use two objective measure of perceptual quality: the Perceptual Evaluation of Speech Quality (PESQ) as recommended by International Telecommunication Union (ITU) [16] and also a composite objective measure (denoted by Csig) of speech distortion [23]. The latter measure is a combination of several basic perceptual measures, and has shown high correlation with subjective listening test which is believed to be the most accurate method for evaluating speech quality. Both of these objective measures provide a number in the range of 1 (the worst quality) to 5 (the best quality) which is shown to have very high correlation with the average score that would be given by a panel of listeners about the quality of the processed speech signal.

Figure 3 shows a comparison of the resulting PESQ, CMOS and SNR versus the size of the subset of samples for all these sampling methods that clearly confirms our subjective informal evaluations. For instance, to achieve the PESQ equals to 3, the proposed method requires 16% of points in the MSM, while 27% of samples are required with the method of [11]. In case of the subset of equally spaced samples, such quality is not achieved with less than 33% of the samples.

Now that we have seen the reconstructibility of the speech signal from the MSM, we show how it can be used to build a waveform coder. We emphasize however, that our goal is not to achieve the best coding performances, but rather to demonstrate the potential of the proposed methodology. In fact, the required data for the reconstruction of the signal are the MSM signal amplitudes along with their relative position on the time axis ($\Delta t$). In order to use fewer bits to encode the MSM signal amplitudes, we encode the average gradient values. Indeed, the dynamic range of these values is lower compared to the original signal amplitudes and hence they can be quantized with fewer bits. We then use a 6-bit non-uniform quantizer [20] matched to the histogram of average gradient values of a very small development set ($20sec$ of speech, uttered by 3 men and 3 women). We observe that the distance between 95% of MSM points is less than 16 samples ($f_s = 16kHz$). Therefore,

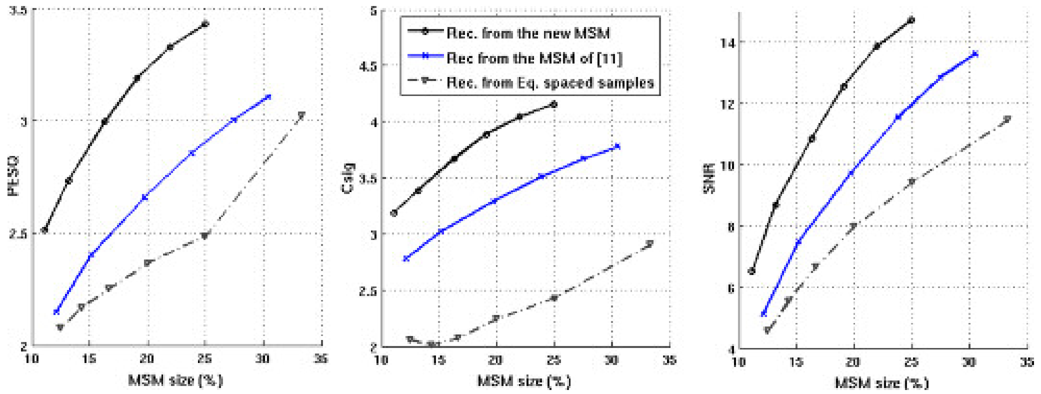

Figure 3: Comparison of reconstruction quality for three different sampling methods: the MSM formed using our measure in Eq. (3), the MSM formed by the measure of [11] and the subset of equally spaced samples. The same reconstruction method is used for all these three sampling methods as proposed in section 3.2.

we dedicate 4-bits to encode the $\Delta t$ information of successive MSM samples. If, as a result of pauses in the speech signal, the distance is more than 16 samples, a whole 10 bit word would be dedicated to transfer the $\Delta t$ information. Consequently, there will be about 5% data overhead for the distant MSM samples.

By application of this coder to the full 2.5 hour dataset (converted to $VAF = 0.5$ as classically done), we achieved an avarge perceptual score of 3.4 with a total bit-rate of 16.8 kbps. This is a very good coding performance if we compare it with the performance of the G.726 standard of ITU which uses Adaptive DPCM coding and achieves the average PESQ of 3.5 with the bit-rate of 24 kbps [16]. Moreover, we achieved this coding performance using a very simple quantization and we measure the quality on the raw output of the reconstruction without any enhancement. Indeed, the distortions in the reconstructed signal which origins from occasional loss of some intermediate information, could be corrected with further post-processing.

## 5. Conclusion

In the continuation of our research [12, 24] in the study of the non-linear characteristics of speech, we showed in this paper that a very compact representation of speech signal is possible to achieve by the study of geometrical scaling exponents which relate non-linearity to the multiscale structure of complex signals. In a complex system whose dynamics is entirely described by, for instance Fully Developed Turbulence (FDT), LPEs can be computed indifferently by various multiscale measures (speed increments, energy dissipation etc.) all giving rise to the same MSMs and reconstruction properties. The voice production mechanism tends to indicate that speech involves the superposition of different dynamics on top of a purely chaotic one. We showed that such approach results in speech reconstruction of good quality using MSM of low cardinality. Consequently, to demonstrate the compactness of this representation, we proposed a simple coding of the MSM which yields similar perceptual quality as that of G.726 ADPCM coder, with considerably lower bit rate. We emphasize however that our goal was not to achieve the best coding performances, but rather to continue demonstrating the potential of the radically novel perspective we are adopting for non-linear speech analysis.

## Acknowledgment

This work was 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.

[3] M. Little, P. E. McSharry, I. Moroz, S. Roberts, Testing the assumptions of linear prediction analysis in normal vowels, Journal of the Acoustical Society of America 119 (January 2006) 549–558. doi:10.1121/1.2141266.

[4] G. Kubin, Speech coding and synthesis, Elsevier, 1995, Ch. Chapter 16: Nonlinear processing of speech, pp. 557–610.

[5] 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.

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

[7] G. Boffetta, M. Cencini, M. Falcioni, A. Vulpiani, Predictability: a way to characterize complexity, Physics Reports 356 (6) (2002) 367–474, doi:10.1016/S0370-1573(01)00025-4.

[8] A. Turiel, A. del Pozo, Reconstructing images from their most singular fractal manifold, IEEE Transactions on Image Processing 11 (2002) 345–350.

[9] 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.

[10] O. Pont, A. Turiel, H. Yahia, An optimized algorithm for the evaluation of local singularity exponents in digital signals, in: 14th International Workshop on Combinatorial Image Analysis, 2011.

[11] A. Turiel, Method and system for the singularity analysis of digital signals, patent registered under number pct/es2008/070195 (2008).

[12] V. Khanagha, H. Yahia, K. Daoudi, O. Pont, A. Turiel, Reconstruction of speech signals from their unpredictable points manifold, in: NOLISP, Lecture Notes in Computer Science, Springer, 2012.

[13] K. Sauer, J. Allebach, Iterative reconstruction of bandlimited images from nonuniformly spaced samples, IEEE Transactions on Circuits and Systems 34 Issue:12 (Dec 1987) 1497 – 1506.

[14] J. Rheem, B. Kim, S. Ann, A nonuniform sampling method of speech signal and its application to speech coding, Signal Processing 41 (1) (1995) 43–48. doi:10.1016/0165-1684(94)00089-I.

[15] P. Ghosh, T. Sreenivas, Waveform reconstruction from non-uniform samples with application to speech coding, in: IEEE-Eurasip Nonlinear Signal and Image Processing (NSIP) 2005, Abstracts., 2005, p. 35.

[16] G726 recommendation: 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation, international telecommunication union (1990).

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

[18] U. Frisch, Turbulence: The legacy of A.N. Kolmogorov, Cambridge University Press, 1995.

[19] O. Pont, A. Turiel, C. Pérez-Vicente, On optimal wavelet bases for the realization of microcanonical cascade processes, International Journal of Wavelets, Multiresolution and Information Processing 9 (2011) 35–61.

[20] W. C. CHU, Speech coding algorithms: foundation and evolution of standardized coders, Wiley-Interscience, 2003.

[21] J. G. Proakis, D. K. Manolakis, Digital Signal Processing: Principles, Algorithms and Applications, Prentice Hall, 1995.

[22] H. G. Feichtinger, K. Grochenig, Theory and practice of irregular sampling, CRC-Press, 1993, Ch. Wavelets Mathematics and Applications.

[23] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on Audio Speech Language Processing 16 (2008) 229 – 238.

[24] 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.