Reconstruction of Speech Signals from their Unpredictable Points Manifold

Vahid Khanagha, Hussein Yahia, Khalid Daoudi, Oriol Pont, Antonio Turiel

Reconstruction of Speech Signals from Their Unpredictable Points Manifold

Vahid Khanagha[1], Hussein Yahia[1], Khalid Daoudi[1], Oriol Pont[1], and Antonio Turiel[2]

[1] INRIA Bordeaux Sud-Ouest (GEOSTAT team), 351 Cours de la Liberation, 33405 Talence cedex, France
[2] Institut de Ciències del MAR, ICM - CSIC, Passeig de la barceloneta 37-49, 08003 Barcelona, Spain
http://geostat.bordeaux.inria.fr
{vahid.khanagha,hussein.yahia,khalid.daoudi, oriol.pont}@Inria.fr,turiel@icm.csic.es

Abstract. This paper shows that a microcanonical approach to complexity, such as the Microcanonical Multiscale Formalism, provides new insights to analyze non-linear dynamics of speech, specifically in relation to the problem of speech samples classification according to their information content. Central to the approach is the precise computation of Local Predictability Exponents (LPEs) according to a procedure based on the evaluation of the degree of reconstructibility around a given point. We show that LPEs are key quantities related to predictability in the framework of reconstructible systems: it is possible to reconstruct the whole speech signal by applying a reconstruction kernel to a small subset of points selected according to their LPE value. This provides a strong indication of the importance of the Unpredictable Points Manifold (UPM), already demonstrated for other types of complex signals. Experiments show that a UPM containing around 12% of the points provides very good perceptual reconstruction quality.

Keywords: non-linear speech processing, multi-scale methods, complex signals and systems, predictability, compact representation.

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 [5, 15]. However, most of the classical approach in speech processing are based on linear techniques which basically rely on the source-filter model. These linear approachs can not 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) [9]. For this reason, non-linear speech processing has recently gained a significant attention, seeking for alternatives of these dominant linear methods [8]. The use of Lyapunov Exponents (associated to the degree of chaos or predictability in a dynamical system) [7] or fractal dimensions of speech (Minkowski-Bouligand dimensions, related to the amount of turbulence in a speech sound)[11] for phoneme classification are successful examples of such methods considering speech as a realization of complex system in a physical sense.

Recent advances in the analysis of complex signals and systems have shown that a microcanonical approach associated to a precise evaluation of Local Predictability Exponents (LPEs) can be derived from analogies with Statistical Physics. These methods can be used to get access to non-linear characteristics of the speech signal and to relate them with the geometrical localization of highly unpredictable points [2] in the signal domain. The framework called Microcanonical Multiscale Formalism (MMF) can be used to compute these LPEs [18]. In an earlier work [6] we have presented the potential of these exponents in the identification of transition fronts in the speech signal and we used them to develop a powerful phonetic segmentation algorithm. In this paper, by showing how the evaluation of predictability is embedded in the estimation procedure of LPEs, we study how they truly quantify the information content at any given point. We show how the LPEs can be used to determine a proper subset in the signal domain made of points which are the most informative (i.e. less predictable). This is proved, in the framework of reconstructible systems, by successful reconstruction of the whole signal from that proper subset, called for that reason the Unpredictable Points Manifold (UPM) [14], which turns to be identical with the Most Singular Manifold previously defined in [18].

We use an objective measure of perceptual quality to evaluate the reconstructibility of speech signal from the UPM. We show that quite natural reconstruction can be achieved by applying the reconstruction formula to a UPM which contains only around 12% of samples. This implies the possibility of the future development of an efficient speech compression algorithm. Moreover, this significant redundancy reduction, together with the previous promising phonetic segmentation results [6], add to demonstrate the high potential of LPEs in the analysis of speech signal.

This paper is organized as follows: section 2 provides a brief review of basic concepts of the MMF. In section 3 we present the detailed procedure for the estimation of LPEs through a local reconstruction procedure. In section 4 the experimental results are presented and finally, in section 5 we draw our conclusions.

2 Microcanonical Multiscale Formalism

Microcanonical Multiscale Formalism (MMF) is a microcanonical approach to the geometric-statistical properties of complex signals and systems [18]. It can be seen as an extension of previous approaches [1] 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) [3, 13]. MMF is based on the computation of a singularity exponent at every point in a signal's domain; we call the singularity exponents Local Predictability Exponents or LPEs. LPEs unlock the relations between geometry and statistics in a complex signal, and have been used in a wide variety of applications ranging from signal compression to inference and prediction [16, 12]. LPEs are defined by the evaluation of the limiting behavior of a multiscale functional Γ_r at each point t and scale r:

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

If $\Gamma_r\left(s(t)\right) = |s(t+r) - s(t)|$, then evaluation of LPE $h(t)$ results in Hölder exponents which are widely used in fractal analysis. However it is often difficult to obtain good estimation of Hölder exponents because of the sensitivity and instability of linear increments. Another choice for Γ_r is introduced in [18] is a measure operating on the derivative of the signal s':

$$\Gamma_r\left(s(t)\right) := \frac{1}{2r}\int_{t-r}^{t+r} |s'(\tau)|\mathrm{d}\tau \tag{2}$$

In equation 1 the singularity exponent $h(t)$ is called a LPE at t. It can be evaluated through $\log - \log$ regression of wavelet projections [10, 17] but the resolution capability of a wavelet depends on the number of its zero-crossings, which is increased in higher order wavelet but is minimum for positive wavelets. So, the introduction of gradient measures improves the spatial resolution of LPE estimation. However it is still possible to achieve better precision in the estimation of LPEs, particularly while attempting to mitigate the phenomenon of oscillations in a wavelet decomposition or while avoiding the problem of determining a wavelet adapted to the nature of the signal. This leads to the estimation of LPEs as presented in the next section.

3 Evaluating LPEs through local reconstruction

In [16] is presented, in the case of 2D signals, a formal reconstruction of a signal from partial information about its gradient. This reconstruction is properly defined for signals having at each point t a value $h(t)$ defined by equation 1 (in the case of the functional in Eq. (2)). The singularity exponents $h(t)$ define a hierarchy of sets having a multiscale structure closely related of the cascading properties of some random variables associated to the macroscopic description of the system under study. Among these sets, the Most Singular Manifold $\mathcal{F}_\infty$ (MSM, defined by the points in the signal's domain having the minimal singularity exponent at a given threshold) maximizes the statistical information in the signal. Consequently, there is a universal operator that reconstructs the signal from its gradient restricted to the MSM. The set $\mathcal{F}_\infty$ defines a current of the *essential gradient* defined by:

$$\nabla_{\mathcal{F}_\infty} s(t) = s'(t)\delta_{\mathcal{F}_\infty}(t) \tag{3}$$

where $\delta_{\mathcal{F}_\infty}$ is the distribution associated to the continuum of the MSM. According to these notations, the reconstruction formula reads

$$s(t) = (g \cdot \nabla_{\mathcal{F}_\infty} s)(t) \tag{4}$$

where g is a universal reconstruction kernel defined in Fourier space by:

$$g(\hat{\mathbf{k}}) = i\mathbf{k}/|\mathbf{k}|^2, i = \sqrt{-1} \tag{5}$$

This reconstruction kernel g acts as an inverse derivative operator in Fourier space and there is at least one set $\mathcal{F}_\infty$ for which the reconstruction formula is trivial: if we take as $\mathcal{F}_\infty$ the whole signal domain, then the reconstruction formula takes the form $\nabla_{\mathcal{F}_\infty} s(t) = s'(t)$. If $\mathcal{F}$ is any set for which the reconstruction is perfect, then, according to [16, 14], the decision to include or not a point t in $\mathcal{F}$ is local around t. By definition, the Unpredictable Point Manifold (UPM) $\mathcal{F}_{upm}$, the collection of all unpredictable points, is the smallest set for which perfect reconstruction is achieved. The basic conjecture of the framework of reconstructible systems is that $\mathcal{F}_{upm} = \mathcal{F}_\infty$ i.e. UPM = MSM. Assuming this, we will therefore define a quantity associated with the local degree of predictability at each point. This quantity is a special vectorial measure defined by a wavelet projection of the gradient which penalizes unpredictability.

Given the point t in the domain of discrete signal $s(t)$, the simplest neighborhood associated to the predictability at time t, can be formed by the three points (p_0, p_1, p_2) where $p_0 = t$, $p_1 = t + 1$ and $p_2 = t - 1$. In order to avoid standard harmonics of the form $(e^{2ik\pi/n})_k$ which are dependent to the size n, we first mention that the simplest Nyquist frequency in the two directions of a given point t, is $2\pi/3$. Consequently we introduce the complex number $\omega = e^{2i\pi/3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$, $\bar{\omega} = \omega^2$ along with the matricial operator:

$$\mathcal{M} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & \omega & \bar{\omega} \\ 1 & \bar{\omega} & \omega \end{bmatrix} \tag{6}$$

Now we define a derivation operator d_x in Fourier space, which is naturally associated to a half-pixel difference between a given point and its immediate neighbors:

$$d_x = \mathcal{M}^{-1}\hat{d}_x\mathcal{M} \tag{7}$$

where $\hat{d_x} = (0, i\sqrt{3}, -i\sqrt{3})$ and its action on a vector is being done componentwise. Indeed, we multiply $\mathcal{M}$ by a vector and then we apply $\hat{d_x}$ and multiply the resulting vector by $\mathcal{M}^{-1}$. The local reconstruction operator is then defined as:

$$\mathcal{R} = \mathcal{M}^{-1}\hat{\mathcal{R}}\mathcal{M} \tag{8}$$

where $\hat{\mathcal{R}} = (0, -i\sqrt{3}, i\sqrt{3})$. We use these gradient and reconstruction operators to define a UPM measure of local correlation as follows. Given a point

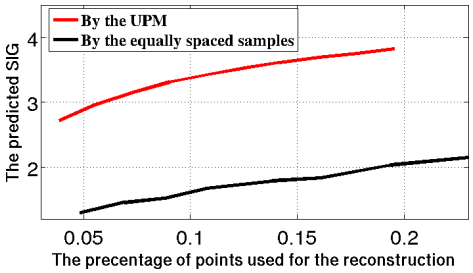
t_0 and a scale r_0, we define the neighborhood of t_0 as (p_0, p_1, p_2) and the associated signal values of this neighborhood as (s_0, s_1, s_2). Given the average $\bar{s} = \frac{1}{3}(s_0 + s_1 + s_2)$ we form the conveniently detrended vector (u_0, u_1, u_2) as $u_0 = p_0 + \bar{s}$, $u_1 = p_0 - \bar{s}$, $u_2 = p_0 - \bar{s}$. We apply d_x to the vector (u_0, u_1, u_2) to obtain (g_0, g_1, g_2) whose we save its first element as $A = g_0$. The local reconstruction operator is then applied to (g_0, g_1, g_2) in order to deduce the reconstructed signal (q_0, q_1, q_2). Once again, we apply d_x to the latter to obtain (ρ_0, ρ_1, ρ_2). The UPM measure of local correlation is then defined as:

$$\mathcal{T}_{\Psi_{lcsm}} \Gamma_{r_0}(t_0) = |A - \rho_0| \tag{9}$$

The exponents $h(t)$, thus called Local Predictability Exponents (LPEs), are then obtained using a point-wise evaluation of Eq. (1) with this UPM measure.

4 Experimental results

The practical formation of UPM includes the determination of a threshold h_θ, such that the application of the reconstruction formula to the points in which their LPE is inferior to this threshold (i.e. $< h(t) < h_\theta$), provides negligible reconstruction error. However, in the case of speech signal, a global determination of h_θ can be problematic. Indeed, speech is a non-stationary signal which can be regarded as the concatenation of small units (phonemes) which essentially possess diverse statistical characteristics. Hence a globally selected h_θ might lead in perfect reconstruction of some phonemes (where the global h_θ has formed a dense UPM around the neighborhood), while providing poor reconstruction of some other parts (where a less dense UPM is formed).



Score	Description
5	Very natural, no degradation
4	Fairly natural, little degradation
3	Somewhat natural, somewhat degraded
2	Fairly unnatural, fairly degraded
1	Very unnatural, very degraded

Fig. 1: **Left-** The perceptual quality of the reconstructed signals resulted from the application of reconstruction formula (Eq. (4)) to the UPM and to the subset of equally spaced samples. **Right-** the description of the SIG scale.

To preserve the relative reconstruction quality for the whole signal, we define a global UPM density and we locally select h_θ such that this density is preserved in each neighborhood. For instance, to form a UPM containing 10% of the points of the signal, we use a sliding window (without overlapping) and for

each window, we take 10% of the points in the neighborhood having the least values of LPE (i.e. they are less predictable). The window length of $20msecs$ is employed so that the statistical variations of speech signal in each window might be neglgible. We take $10secs$ of speech signal from TIMIT database to assess the reconstructibility of speech signal from their UPM. The sample speech signal used for this experiment has a high voice activity factor($= 0.9$). Hence, the resulting redundancy reduction could not be related to a simple voice activity detection.

To evaluate the quality of the reconstructed speech signals, a composite objective measure of speech distortion[4] is used. This composite 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. This measure provides a predicted rating of speech distortion according to the SIG scale. Figure 1 shows the description of SIG scale, along with the resulting reconstruction scores for different UPM densities. In order to demonstrate how UPM truly corresponds to the most informative subset, we have compared the results of reconstruction from UPM with that of reconstruction from another subset which has the same size but is formed as the collection of equally spaced samples. It can be seen that the reconstruction scores obtained by the UPM is significantly superior. Moreover, quite natural reconstruction is achieved with only around 12% of the points in the UPM. In this case, the standard error of reconstruction (as defined in [18]) is equal to 27dB, which confirms the quality of the *waveform* reconstruction, besides the *perceptual* quality evaluated by the aforementioned composite measure.

5 Conclusion

By precise estimation of the LPEs according to the evaluation of the degree of local reconstructibility, we showed that it is possible to recognize the most informative subset inside the speech signal, called the UPM. We successfully used UPM to reconstruct the whole speech signal, with enough naturality. Indeed, a SIG score of 3.5 is achieved by a reconstruction from 12% of speech samples. Following our successful LPE-based phonetic segmentation [6], such significant redundancy reduction with a simple use of LPEs gives more ground to the relevance of these parameters in the analysis of the complex dynamics of speech signal, while leaving the door open for particularities of the speech signal w.r.t. the MMF, yet to be discovered in future studies.

Acknowledgment

This work was funded by the INRIA CORDIS doctoral program.

References

1. Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J., Muzy, J.F.: Ondelettes, multifractales et turbulence. Diderot Editeur, Paris, France (1995)

2. Boffetta, G., Cencini, M., Falcioni, M., Vulpiani, A.: Predictability: a way to characterize complexity. Physics Reports 356(6), 367–474 (2002), doi:10.1016/S0370-1573(01)00025-4
3. Frisch, U.: Turbulence: The legacy of A.N. Kolmogorov. Cambridge Univ. Press (1995)
4. Hu, Y., Loizou, P.C.: Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Language Processing 16, 229 – 238 (2008)
5. Kaiser, J.F.: Some observations on vocal tract operation from a fluid flow point of view. In: Titze, I.R., Scherer, R.C. (eds.) Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control, pp. 358–386. The Denver Center for the Performing Arts (1983)
6. Khanagha, V., Daoudi, K., Pont, O., Yahia, H.: Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2010)
7. Kokkinos, I., Maragos, P.: Nonlinear speech analysis using models for chaotic systems. IEEE Transactions on Speech and Audio Processing 13(6), 1098–1109 (Jan 2005)
8. Kubin, G.: Nonlinear processing of speech. Chapter 16 on Speech coding and synthesis. Elsevier (1995)
9. Little, M., McSharry, P.E., Moroz, I., Roberts, S.: Testing the assumptions of linear prediction analysis in normal vowels. Journal of the Acoustical Society of America 119, 549–558 (January 2006)
10. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press (1999)
11. Maragos, P., Potamianos, A.: Fractal dimensions of speech sounds: Computation and application to automatic speech recognition. Journal of Acoustic Society of America 105, 1925–1932 (March 1999)
12. Pont, O., Turiel, A., Pérez-Vicente, C.J.: Description, modeling and forecasting of data with optimal wavelets. Journal of Economic Interaction and Coordination 4(1), 39–54 (June 2009)
13. Pont, O., Turiel, A., Perez-Vicente, C.: Empirical evidences of a common multifractal signature in economic, biological and physical systems. Physica A 388(10), 2025–2035 (May 2009)
14. Pont, O., Turiel, A., Yahia, H.: An optimized algorithm for the evaluation of local singularity exponents in digital signals. In: 14th International Workshop on Combinatorial Image Analysis (2011)
15. Teager, H.M., Teager, S.M.: Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle, W., Marchal, A. (eds.) Speech Production and Speech Modelling. NATO Advanced Study Institute Series D (1989)
16. Turiel, A., del Pozo, A.: Reconstructing images from their most singular fractal manifold. IEEE Trans. on Im. Proc. 11, 345–350 (2002)
17. Turiel, A., Pérez-Vicente, C., Grazzini, J.: Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study. Journal of Computational Physics, Volume 216, Issue 1, p. 362-390. 216, 362–390 (2006)
18. Turiel, A., Yahia, H., Vicente., C.P.: Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis. J. Phys. A, Math. Theor. 41, 015501 (2008)