Efficient multipulse approximation of speech excitation using the most singular manifold

Vahid Khanagha, Daoudi Khalid

Efficient multipulse approximation of speech excitation using the most singular manifold

Vahid Khanagha and Khalid Daoudi

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

Abstract

We propose a novel approach to find the locations of the multipulse sequence that approximates the speech source excitation. This approach is based on the notion of Most Singular Manifold (MSM) which is associated to the set of less predictable events. The MSM is formed by identifying (directly from the speech waveform) multiscale singularities which may correspond to significant impulsive excitations of the vocal tract. This identification is done through a multiscale measure of local predictability and the estimation of its associated singularity exponents. Once the pulse locations are found using the MSM, their amplitudes are computed using the second stage of the classical MultiPulse Excitation (MPE) coder. The multipulse sequence is then fed to the classical LPC synthesizer to reconstruct speech. The resulting MSM-based algorithm is shown to be significantly more efficient than MPE. We evaluate our algorithm using 1 hour of speech from the TIMIT database and compare its performances to MPE and a recent approach based on compressed sensing (CS). The results show that our algorithm yields similar perceptual quality as MPE and outperforms the CS method when the number of pulses is low.

Index Terms: Multipulse speech coding, source excitation approximation, multiscale signal processing, singularity exponents.

1. Introduction

Multipulse source coding has been widely used and studied within the framework of Linear Predictive Coding (LPC). It consists in finding a sparse representation of the excitation source (or residual) which yields a source-filter reconstruction with high perceptual quality. The MultiPulse Excitation (MPE) method [1, 2] is the first and one of the most popular techniques to achieve this goal. MPE provides a sparse excitation sequence through an iterative Analysis-by-Synthesis procedure to find the position and amplitudes of the excitation one at a time [1], and then re-optimizing the amplitudes once the locations for all of the pulses are found [2]. The well known Code Excited Linear Prediction (CELP) is essentially a multipulse coder which uses vector quantization to search in a codebook of excitation signals to determine the excitation sequence.

Most of multipulse coding methods depend on the choice on the vocal tract model, such as the all-pole filter parameters in LPC. That is, a predictor (typically an autoregressive model) is first learned to model the vocal tract transfer function, then the sparse residual approximation is obtained using this predictor. In this paper we adopt a different strategy, we argue that the *locations* of the pulse excitations can be retrieved *directly* from the observed speech waveform and *independently* of the predictor. The principle behind this strategy is the following: in voiced speech, impulsive excitation takes places around Glottal Closure Instances (GCI), i.e., when air flow through the glottis is blocked by closure of the vocal folds. This produces multiscale singularities on the observed waveform and, hence, prediction cannot be correctly performed around these instances. We thus define a measure of predictability and argue that the sparse multipulse excitation should be located within the set of less predictable points. This measure is not specific to voiced speech, but is rather geometric and local. We thus assume that it should also work for unvoiced speech.

To define this measure of predictability we follow the same philosophy as in our recent work [3, 4] on speech analysis using the Microcanonical Multiscale Formalism (MMF). In that work, we showed that singularity exponents (a notion central to MMF) of speech signals permit to develop an accurate and efficient algorithm for unsupervised phonetic segmentation which outperforms state-of-the-art techniques. In this paper, we go further and analyze another notion central to MMF, the Most Singular Manifold (MSM). We show that MSM provides a good approximation to the locations of the sparse multipulse excitations and compare it to the standard MPE method [2] and a recent Compressed Sensing (CS) based approach [5]. The results show that our MSM approach yields similar performances than MPE while it is much faster. They also show that that our approach outperforms the CS method, which has roughly the same computational as the MPE, when number of pulses per speech frame is low.

The paper is organized as follows. In section 2 we recall the basic concepts of MMF, define the measure of local predictability and describe how to form the MSM. In section 3 we present the MSM-based algorithm to approximate the multipulse source excitation. In section 4 the experimental results are presented. Finally, in section 5, we draw our conclusion and perspectives.

2. Most Singular Manifold of speech signals

We have been recently developing a novel framework for nonlinear analysis of speech signals based on the Microcanonical Multiscale Formalism (MMF) [6]. The latter allow the study of the local geometrico-statistical properties of complex signals from a multiscale perspective. It can be seen as an extension of previous approaches 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) [7]. Central to the formalism is the computation of Singularity Exponents (SE) 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, 9]. The singularity exponent $h(t)$ for any given d-dimensional signal $s(t)$, can be estimated by the evaluation of the limiting power-law scaling behaviour of a multiscale functional Γ_r over a set of fine scales r:

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

where $\Gamma_r\left(s(t)\right)$ can be any multiscale functional complying with this power-law like the gradient-based measure introduced in [6]. The term $\mathrm{o}\left(r^{d+h(t)}\right)$ means that for small scales the additive terms are negligible compared to the factor and thus $h(t)$ dominantly quantifies the multiscale behaviour of the signal at the time instant t.

In MMF, a particular set of interest is the level set called the *Most Singular Manifold* (MSM) which comprises the points having the smallest SE, and which provides indications in the acquired signal (a pressure i.e. an intensive physical variable) of most critical transitions of the associated dynamics [6]. These are the points where sharp and sudden local variations take place and hence they have the lowest possible predictability from their neighbouring points. The formal definition of MSM reads:

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

In practice, once the signal is discretized, h_∞ 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, 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 [8] 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 Γ_r measure defined in [10] which makes a local evaluation of the reconstruction operator to penalize unpredictability.

For the application we consider in this paper, we have tried different measures Γ_r and different estimation methods of their associated SE. The Γ_r we tried are however all gradient-modulus based as suggested in [6]. We found that the measure which yields the best performances is the following:

$$\Gamma_r(s(t)) \;=\; \int_0^r |\nabla_\tau s(t)| d\tau \quad (3)$$

where

$$\nabla_\tau s(t) \;=\; |2s(t) - s(t-\tau) - s(t+\tau)| \quad (4)$$

Once Γ_r is specified, there are many ways to estimate the SE $h(t)$ (log-log regression, adaptation of the 2D method in [10] to 1D,...). We found that the estimation method which yields the best performances is the one used and theoretically motivated in [9, 11]. With this method, the MSM actually corresponds to the set of points where energy concentrates as it transfers across scales and, in that sense, it is a *least predictable/reconstructible manifold*. Under some hypothesis, this leads to a simple estimation method of the SE, as the sum of a set of *transitional* exponents:

$$h(t) = \sum_{i=1}^{j} h_{r_i}(t) \quad (5)$$

where $h_{r_i}(t)$ are the *transitional* exponents, which can be computed as:

$$h_{r_i}(t) = \frac{log(\Gamma_{r_i}(s(t)))}{log(r_i/f_s)} \quad (6)$$

where f_s is the sampling frequency of the signal. The SE computed according to Eq. (5) are then being used to form the MSM as explained above. We will come back in the next section on this method of SE estimation to provide a practical explanation of its good performances.

3. MSM multipulse approximation of source excitation

It is well known that significant impulsive excitations are reflected over the whole speech spectral band [12]. Consequently, excitation impulses would produce "strong" local singularities at different scales of the waveform. This legitimate the use of the multiscale power-law of Eq. (1) to identify and quantify these singularities. This also makes it natural to expect the co-existence of negative transitional SE (Eq. (6)) at different scales around these singularities. Their summation (Eq. (5)) would thus result in lower negative values. Such singularities would then belong to the MSM and considered as unpredictable. Fig.1 shows an example which confirms this intuition. In Fig.1(top) a segment of voiced sound is shown, along with its corresponding pitch marks given by the Electro-Glottal Graph (EGG) data [13]. The pitch marks are taken around the Glottal Closure Instants (GCI) when the most significant impulsive excitation of vocal cords happens. In Fig.1(bottom), the MSM points of this segment are shown with their value of SE. The MSM is formed as the 5% of samples having lowest value of SE. It can be seen that MSM points are indeed located around the reference GCI. Note also that, around every single GCI, the MSM point with the lowest SE value is the closest one to the GCI mark. This example shows that our MSM can indeed identify the location where significant impulsive excitations occur. We emphasize however that our purpose is not to develop a GCI identification method, we did not investigate (yet) this matter and it is beyond the scope of this paper. We recall that our purpose in this paper is to define a notion of predictability which allows a multipulse approximation of the excitation using the set of less predictable points, the MSM.

Once the MSM is formed, we have to estimate the corresponding pulse amplitudes and then feed it to a filter which models the vocal tract in order to reconstruct the speech signal. To do so we use the second stage of the standard MultiPulse Excitation (MPE) coder [2]. We recall that the latter applies an Analysis-by-Synthesis scheme in two stages. In the first one, pulses are added iteratively one at a time by minimizing mean squared error (mse) between the original and reconstructed signal. The computation required in this first stage is K searches of order N, where K is the number of desired pulses and N is the number of signal samples. In the second stage, once the locations of all pulses are found, their amplitudes are jointly re-optimized such that the mse is minimized [2].

In our MSM based algorithm, we replace the first stage of MPE, i.e., the iterative search to find the pulse locations, by the

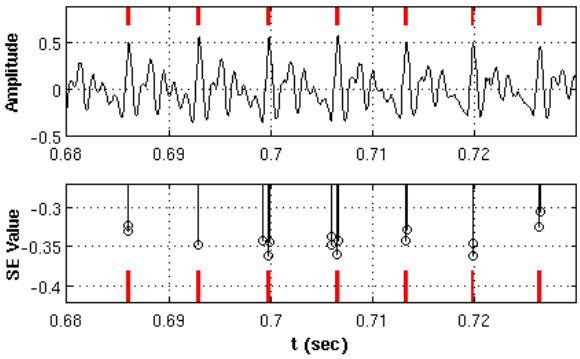

Figure 1: (top) A voiced segment of the speech signal "arctic_a0001" of the male speaker BLD from the publicly available "CMU ARCTIC" database [13]. The reference pitch marks are represented by vertical red lines, (bottom) MSM samples and their corresponding SE values.

following procedure. We form the MSM by taking $2K$ samples having the lowest SE values. Then, assuming that the pulses are located on the MSM grid, we find their amplitudes using the same joint optimization as in the second stage of the MPE. Finally, we choose the K pulses with the highest amplitude as the excitation sequence. Clearly, our approach is computationally more efficient than MPE since the whole first stage of the classical MPE (K searches of order N) is replaced by a simple sort of SE values to form the MSM.

4. Experimental results

We have tried to follow the same experimental protocol as in [5]. That is, we evaluate our method using about 1 hour of clean speech signal randomly chosen from TIMIT database (re-sampled to 8kHz) uttered by speakers of different genders, accents and ages which provides enough diversity in the characteristics of the analyzed signals. 10 prediction coefficients are computed for frames of 20ms (N=160) and the search for impulses are separately performed in each of two subframes of 10 ms, following the procedure explained in [2]. We use the four finest scales to estimate SE, $j = 4$ in Eq. (5). No long-term pitch prediction is performed. All the results we show are without quantization and are almost the same with different randomizations. We compare the performance of our method with the classical MPE [2] and the CS-based method [5].

Before starting this comparison, we first show an example of the reconstruction quality using the MSM method. Fig. 2 shows a stationary voiced sound (a), the MSM excitation sequence with 7 (resp. 14) pulses per 20 ms (b) (resp. (d)), and its reconstruction (c) (resp. (e)). This example shows clearly that our method can indeed yield good reconstruction quality of voiced speech even when using only few pulses.

Fig. 3 shows the average normalized reconstruction error ($\bar{e}_N = \frac{\|s-\hat{s}\|_2}{\|s\|_2}$) of the MSM and MPE methods for different number K of pulses per 20ms of speech. This results shows that our method achieves a satisfactory performance but is still less accurate than MPE in terms of mean square error (mse). This can be explained by the fact that MPE finds and adds iteratively one pulse to the previous ones so as to minimize the mse. In others words, MPE focuses on mse minimization which is not the focus of the MSM method as it provides pulse locations

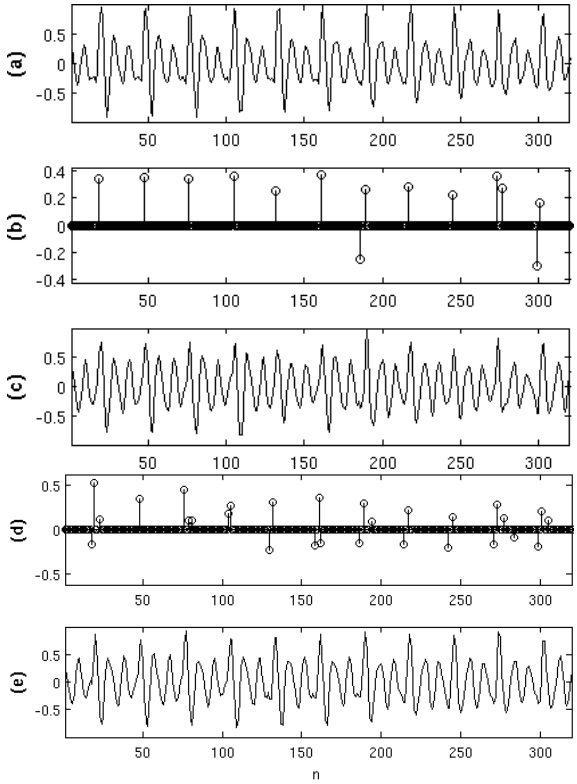

Figure 2: (a) 40 ms segment of stationary voiced speech, (b) the MSM excitation sequence using 7 pulses per 20 ms, (c) the reconstructed signal, (d) the MSM excitation sequence using 14 pulses per 20 ms, and (e) the reconstructed signal.

candidates using a multiscale geometrical approach. This definitely penalizes the MSM method in terms of mse, but it makes it much more efficient than MPE. Still, our method surprisingly outperforms the CS one [5] in terms of mse when $K < 10$ (see Fig.1 in [5]). For instance, for K=8, we achieve $\bar{e}_N = 0.55$ while CS method gives $\bar{e}_N \approx 0.68$. For $K = 10$, which is the typical operating point of a multipulse coder, our method and the CS method perform almost the same. Meanwhile, the CS method has a computational complexity which is roughly the same as MPE, so our method is also much more efficient than the CS one.

The computational processing times are compared in Table 1, in terms of the average empirical Relative Computation Time:

$$RCT(\%) = 100.\frac{CPU\ time\ (s)}{Sound\ duration\ (s)}$$

On the other hand, mse is not the best way to assess the perceptual quality of reconstructed speech. First, our informal subjective listening test showed that the perceptual quality of our method is indeed very close to that of MPE. Especially for K=20, both methods provide almost the same perceptual quality. Second, we evaluated the perceptual quality of reconstructed speech from MSM and MPE using the composite measure of speech quality CMOS [14]. This measure is a combination of PESQ, Cepsterum distance measure, LLR and Itakura-Saito distance. It provides a score of perceptual quality in the range of 1 (the worst quality) to 5 (the best quality). The results are shown in Table 1. This comparison confirms our in-

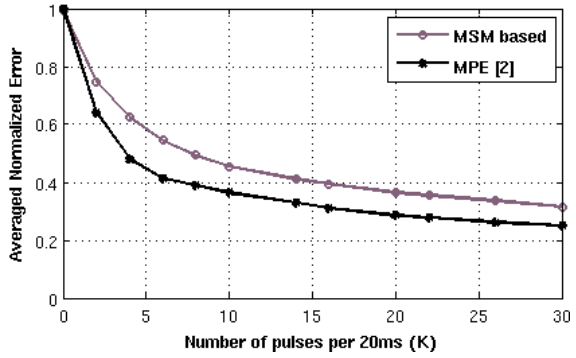

Figure 3: The average normalized reconstruction error, averaged over 1 hour of speech signals from TIMIT database.

formal listening tests. Indeed, the perceptual quality (measured in terms of CMOS) of the MSM and MPE methods are roughly the same.

In summary, all these results suggest that the MSM method achieves similar perceptual quality of reconstruction as MPE [2], with much higher computational efficiency. They also suggest that our method outperforms the recent CS-based method [5] when $K < 10$ in terms of both mean squared error and efficiency.

Table 1: *Comparison between the average perceptual quality of reconstruction and the average Relative Computation Time.*

Method	K	CMOS	RCT (%)
MSM	10	4.0	9.6
	20	4.2	21.7
MPE [2]	10	4.1	71.9
	20	4.2	143.7

5. Conclusion and perspectives

Following our recent research on the use of the Microcanonical Multiscale Formalism (MMF) for speech analysis, we introduced in this paper the concept of MSM and its relation to unpredictability. We defined a multiscale measure of local predictability and provided an estimation algorithm of its associated singularity exponents. We first showed that the resulting MSM can indeed identify (directly from the waveform) singularities which correspond to significant impulsive excitations (GCI for instance). We then used the MSM to efficiently determine the locations of the multipulse sequence, their amplitudes are then found using the second stage of MPE. We showed that the resulting algorithm is significantly more efficient than MPE. The experimental results showed that the MSM algorithm achieves similar perceptual quality as MPE and outperforms the recent CS method in terms of mse when $K < 10$. These encouraging results suggest (again) that the MMF has indeed a promising potential in speech processing and should be further investigated. Many perspectives can be drawn from the presented work. For instance, as our approach is independent of the predictor, we can investigate the use of a sparse predictor such as in the CS formulation, instead of the LPC minimum variance predictor. Another interesting and challenging problem is to explore the automatic identification of GCI. This would open the gap for all the GCI related applications, in particular closed-phase LPC would be the readiest application of our approach. This will be the purpose of future communications.

Acknowledgement

The authors would like to thank O. Pont and H. Yahia for helpful discussions. The first author is funded by the INRIA CORDIS doctoral program.

6. References

[1] B. Atal and J. Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1982.

[2] S. Singhal and B. Atal, "Amplitude optimization and pitch prediction in multipulse coders," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 317 – 327, 1989.

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

[4] ——, "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.

[5] D. Giacobello, M. Christensen, M. Murthi, S. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction," *IEEE Signal Processing Letters*, vol. 17, 2010.

[6] A. Turiel, H. Yahia, and C. P. Vicente., "Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, p. 015501, 2008.

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

[8] A. Turiel and A. del Pozo, "Reconstructing images from their most singular fractal manifold," *IEEE Transactions on Image Processing*, vol. 11, pp. 345–350, 2002.

[9] 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, June 2009.

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

[12] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1602–1613, 2008.

[13] "Cmu arctic speech synthesis databases," [Online], http://festvox.org/cmu_arctic.

[14] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio Speech Language Processing*, vol. 16, pp. 229 – 238, 2008.