ATTENUATED EMBEDDING ESTIMATORS FOR SPEECH SIGNALS

D. Napoletani[1]*, C.A. Berenstein*[2]*, T. Sauer*[3]*, D.C. Struppa*[3] *and D. Walnut*[3]

[1] School of Computational Sciences
George Mason University, Fairfax VA 22030
dnapolet@gmu.edu

[2] Institute for Systems Research
University of Maryland, College Park MD 20742

[3] Department of Mathematical Sciences
George Mason University, Fairfax VA 22030

ABSTRACT

In an earlier paper we utilized techniques from the theory of non-linear dynamical systems to define a notion of embedding threshold estimators. These estimators were based on the analysis of delay-coordinates embeddings of sets of coefficients of the measured signal in some chosen frame.
One of the motivations behind the method was the desire of building an estimator that would perform equally well regardless of the type of white noise contamination. In this paper we explicitely write the structure of one particular algorithm, the attenuated embedding threshold estimator, and we show the performance of the algorithm when different types of white noise are added to speech time series.

1. INTRODUCTION

In this paper we explore the performance of a method of denoising that is designed to be efficient for a variety of white noise contaminations, while keeping a fixed choice of parameters of the algorithm itself. The method is based on a loose distinction between the geometry of delay-coordinates embeddings of, respectively, deterministic time series and non-deterministic ones. The techniques we employ are taken from the theory of non-linear dynamical systems as described for example in [5]. Let $F[n]$, $n = 1, ..., N$, be a discrete signal of length N, and let $X[n] = F[n] + W[n]$, $n = 1, ..., N$, be a contaminated measurement of $F[n]$, where $W[n]$ are realizations of a white noise process W. Any discrete periodic signal $X[n]$, $n \in \mathbb{Z}$ with period N can be represented in a discrete windowed Fourier frame. The atoms in this frame are of the form

$$g_{m,l}[n] = g[n-m]exp(-\frac{i2\pi ln}{N}), \;\; n \in \mathbb{Z}. \qquad (1)$$

We choose the window g to be a symmetric N-periodic function of norm 1 and support q. Specifically we can choose g to be the characteristic function of the $[0, 1]$ interval; we realize that this may not be the most robust choice in many cases, but we have purposedly selected this function to avoid excessive smoothing which was found to adversely affect our algorithm. Under the previous conditions x can be completely reconstructed from the inner products $\mathcal{F}X[m, l] =< X, g_{m,l} >$, i.e.,

$$X = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{l=0}^{N-1}\mathcal{F}X[m,l]\tilde{g}_{m,l} \qquad (2)$$

where

$$\tilde{g}_{m,l}[n] = g[n-m]\exp(\frac{i2\pi ln}{N}), \;\; n \in \mathbb{Z} \qquad (3)$$

We denote the collection $\{< X, g_{m,l} >\}$ by $\mathcal{F}X$. For finite discrete signals of length N the reconstruction has boundary errors. However, the region affected by such boundary effects is limited by the size q of the support of g and we can therefore have perfect reconstruction if we first extend X suitably at the boundaries of its support and then compute the inner products $\mathcal{F}X$. More details can be found in [4] and references therein. We can now define a collection $\mathcal{C}_p$ of double-indexed paths by choosing an integer p and setting

$$\gamma_{\bar{m},\bar{l}} = \{g_{m,l} \text{ such that } l = \bar{l}, \bar{m} \leq m \leq \bar{m} + p\}, \qquad (4)$$

that is, paths that are oriented in the time direction. The choice of p is very important as different structure in speech signals (our main case study) is evident at different time scales. Let $I = I(\gamma_{\bar{m},\bar{l}}) = I(\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}})$ be a function defined for each path $\gamma_{\bar{m},\bar{l}} \in \mathcal{C}_p$. We defined in [3] *a semi-local thresholding estimator* in the window Fourier frame as follows:

$$\tilde{F} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{l=0}^{N-1} d_{I,T}(\mathcal{F}X[m,l])\tilde{g}_{m,l} \qquad (5)$$

where $d_{I,T}(\mathcal{F}X[m, l]) = \mathcal{F}X[m, l]$ if $I(\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}}) \geq T$ for some $\gamma_{\bar{m},\bar{l}}$ containing (m, l), and $d_{I,T}(\mathcal{F}X[m, l]) = 0$ if $I(\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}}) < T$ for all $\gamma_{\bar{m},\bar{l}}$ containing (m, l).

The 'semilocal' quality of $\tilde{F}$ is evident from the fact that all coefficients in several $\mathcal{F}X_\gamma$ are used to decide the action of the thresholding on each coefficient. Windowed Fourier transform followed by local thresholding is a well known strategy, but its success is crucially determined by the specific choice of threshold functions utilized; we propose here a novel use of embedding

techniques from non-linear dynamical systems theory to choose the specific shape of I.

We first recall a fundamental result about reconstruction of the state space realization of a dynamical system from its time series measurements. Suppose S is a dynamical system, with state space $\mathbb{R}^k$ and let $h : \mathbb{R}^k \to \mathbb{R}$ be a measurement, i.e., a continuous function of the state variables. Define moreover a function F of the state variables X as

$$F(X) = [h(X), h(S_{-\tau}(X)), ..., h(S_{-(d-1)\tau}(X))] \quad (6)$$

where by $S_{-i\tau}(X)$ we denote the state of the system with initial condition X at $i\tau$ time units earlier.

We say that $A \subset \mathbb{R}^k$ is an invariant set with respect to S if $X \in A$ implies $S_t(X) \in A$ for all t. Then the following theorem is true (see [5] and [1]):

Theorem: *Let A be an m-dimensional submanifold of $\mathbb{R}^k$ which is invariant under the dynamical system S. If $d > 2m$, then for generic measuring functions h and generic delays τ, the function F defined in (6) is one-to-one on A.*

Given this background, let again $X = F + W$ be contaminated measurements of a signal F and assume that the time series $\mathcal{F}F_{\gamma_{\bar{m},\bar{l}}}$. are generated approximately by some unknown dynamical system with $d_A > 0$ for a subset of paths in $\mathcal{C}_p$. Choose some relatively large embedding dimension ($d > 3$) and find the largest and smallest singular values σ_1 and σ_d for the embeddings of $\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}}$. Set $I_{svd}(\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}}) = \frac{\sigma_1}{\sigma_d}$, In [3] we show that, for a white noise processes W, it is very unlikely that $I_{svd}(\mathcal{F}W_{\gamma_{\bar{m},\bar{l}}}) > T$, for some T that depends on the type of windowed Fourier frame and on the choice of paths. Therefore we can use the embedding index I_{svd} as a way to identify paths in $\mathcal{C}_p$ that are likely generated by non random sources.

2. THE BASIC ALGORITHM

In this section we develop a possible algorithm based on these ideas. The notion of semilocal estimator is slightly expanded to improve the actual performance of the estimator itself. To this extent, define neighborhoods for each atom in the windowed Fourier frame, i.e.:

$$\mathcal{O}(g_{m,l}) = \{g_{m',l'} \text{ s.t. } |l' - l| \leq 1, |m' - m| \leq 1\}, \quad (7)$$

Such neighborhoods are used in the algorithm as a way to make a decision on the value of the coefficients in a *two dimensional* neighborhood of $\mathcal{F}X_\gamma$ based on the the analysis of the *one dimensional* time series $\mathcal{F}X_\gamma$ itself.

(A)*Set $\tilde{F} = 0$.*

(B)*Given X, choose $q > 0$ and expand X in a windowed Fourier frame with window size q.*

(C)*Choose sampling intervals $S_{\bar{l}}$ for time coordinate and $S_{\bar{m}}$ for the frequency coordinate. Choose the path length p. Build a collection of paths $\mathcal{C}_p$ as in Eq. 4.*

(D)*Choose embedding dimension d and delay τ along the path. Compute the index $I_{svd}(\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}})$ for each $\mathcal{F}X_{\gamma_{\bar{m},\bar{l}}} \in \mathcal{C}_p$.*

(E)*Choose attenuation coefficient α. Set $\mathcal{F}Y[m,l] = \alpha\mathcal{F}X[m,l]$ if $I_{svd}(\mathcal{F}X_\gamma) \geq T$ for some γ containing $g_{m',l'}$, $g_{m',l'} \in \mathcal{O}(g_{m,l})$, otherwise set $\mathcal{F}Y[m,l] = 0$ if $I_{svd}(\mathcal{F}X_\gamma) < T$ for all γ containing $g_{m',l'}$, $g_{m',l'} \in \mathcal{O}(g_{m,l})$.*

(F)*Let Y be the inversion of $\mathcal{F}Y$. Set $\tilde{F} = \tilde{F} + Y$ and $X = X - Y$.*

(G) *Choose a paramenter $\epsilon > 0$, if $|Y| > \epsilon$ go to step (B).*

Note that the details of the implementation (A)-(G) are in line with the general strategy of matching pursuit. The window length q in step (B) could change from one iteration to the next to 'extract' possible structure belonging to the underlining signal at several different scales. In the experiments performed in the following section we generally alternate between two window sizes q_1 and q_2.

The attenuation introduced in (E) has some additional ad hoc parameters in the definition of the neighborhoods in Eq.7 and in the choice of the attenuation parameter α. By the double process of increasing the number of nonzero coefficients chosen at each step and decreasing their contribution we are allowing more information to be taken at each iteration of the projection pursuit algorithm, but in a slow learning framework that in principle (and in practice as we found out) should increase the sharpness of the distinct features of the estimate, on this issue see the discussion in [2] chapter 10.

Remark: The algorithm we described in (A)-(G) requires the choice of several parameters: we choose a dictionary of analysis $\mathcal{D}$, a collection of discrete paths C_p, the embedding parameters τ (time delay) and d (embedding dimension), and the learning parameters T (threshold level), α (attenuation coefficient) and ϵ. The choice of $\mathcal{D}$ and C_p are dependent on the type of signals we analyze. Since we analyze speech signals, we look at windowed Fourier frames; the algorithm is not very sensitive to the choice of the length q of the window, while the use of several windows is found to be always beneficial. The choice of C_p is also dependent on the type of signals analyzed, speech signals have specific frequencies that change in time, so a set of paths parallel to the time axis was natural in this case. The algorithm has three parameters associated with C_p, the time and frequency sampling rates $\bar{l}$ and $\bar{m}$ and the length p of the paths. Essentially we want to set these parameters so that the number of paths that have index $I_{svd} > T$ is sizeable for a training set of speech signals and marginal for white noise series. Our experience is that such choice is possible and non unique. Finally the choice of α and ϵ is completely practical in nature, ideally we want α and ϵ as close to zero as possible, but, to avoid making the algorithm unreasonably slow, we must set values that are found to give good quality reconstructions on some training set of speech signals while they require a number of iterations of the algorithm that is compatible with the compution and time requirements of the specific problem.

Note:*The algorithm described in this paper and some variations are being patented, with provisional patent application number 60/562,534 filed on April 16, 2004.*

3. DENOISING

In this section we explore the quality of the attenuated embedding threshold in the context of the windowed Fourier frame and with the class of paths $\mathcal{C}_p$ defined in Eq.4. We apply the algorithm to a speech signal from the TIMIT database contaminated by different types of white noise with several intensity levels. We show that the attenuated embedding threshold estimator performs well for all white noise contaminations we consider.

The delay along the paths is chosen as $\tau = 4$, the length of the paths is $p = 2^8$, the embedding dimension $d = 4$. With these parameters, $T \approx 28$ ensures that, for most white noise time series W, $\mathcal{F}W_{\gamma_{\bar{m},\bar{l}}}$, $I_{svd}(\mathcal{F}W_{\gamma_{\bar{m},\bar{l}}}) > T$. The sampling interval of the paths in the frequency direction is $S_{\bar{m}} = 1$ and along the time di-

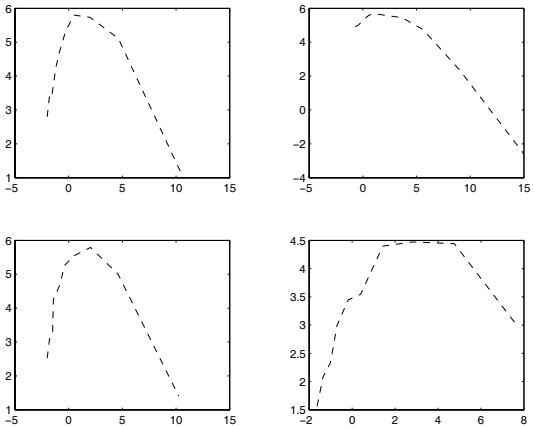

Fig. 1. Scaled SNR gain in decibel of the attenuated embedding estimates plotted against the scaled SNR of the corresponding measurements. From top left in clockwise order we consider the case of: a)Gaussian white noise; b) uniform noise; c)discrete bimodal distribution; d) asymmetrical superposition of Gaussians.

rection is $S_{\bar{l}} = p/8$ We select $\alpha = 0.1$, as small values of α seem to work best. The algorithm is applied to short consecutive speech segments to reduce the computational cost of computing the windowed Fourier transform on very long time series, therefore, to keep the running time uniformly constant for all such segments, we decided to iterate the algorithm (A)-(F) a fixed number of times (say 6) instead of choosing a parameter ϵ in (G). The window size q in (B) alternates between $q_1 = 100$ and $q_2 = 25$. It is moreover important to note that the attenuated embedding threshold is able to extract only a small fraction of the total energy of the signal f, therefore the Signal-to-Noise Ratio (SNR) computations are done on scaled measurements X, estimates $\tilde{F}$, and signals F set to be all of norm 1, we call such estimations *scaled* SNR.

In Figure 1 we show the gain in decibel of the scaled SNR of the reconstructions (with the attenuated embedding threshold estimator) plotted against the corresponding scaled SNR of the measurements. From top left in clockwise direction we have: a) measurements with Gaussian white noise contaminations of decreasing variance; b) measurements contaminated by uniform white noise; c) then we test the case of discrete bimodal white noise contaminations with values chosen with equal probability in the set $\{-Q, Q\}$ and several choices of Q; d) finally we have the case of white noise contaminations with probability distributions g_k that are the scaled superposition of two Gaussians g_1 and g_2 with repectively standard deviations $\sigma_1 = 100$ and $\sigma_2 = 300k$ and with means $m_1 = 200k$ and $m_2 = -100k$, $k = 1, ..., 10$. The g_k are then shifted to have mean 0. We build these superpositions to test the algorithm in an asymmetric and non central case. Note that the overall shape of the scaled SNR gain is similar for all distributions (notwithstanding that the discrete plots do not have exactly the same domain). The maximum gain seems to happen for measurements with scaled SNR around 2 decibel. Note that in the case of uniform noise the right tail of the SNR gain takes negative values; this is due to the attenuation effect of the estimator that is pronunced for the high intensity speech features, but it is by no means indicative of worse perceptual quality with respect to the measurements. The reconstructions for the case of uniform and

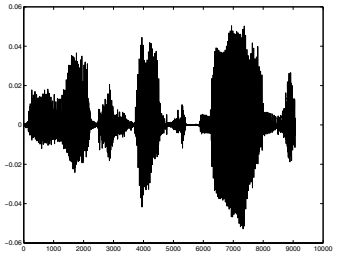

Fig. 2. Original speech signal scaled to have norm 1.

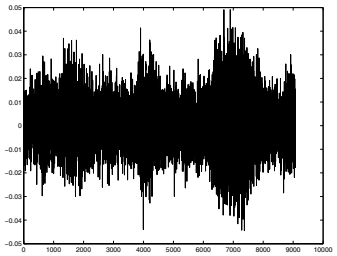

Fig. 3. Noisy measurement of the signal in Figure 2 with Gaussian white noise and scaled SNR of about 2db.

Gaussian noise show better perceptual quality than for the two bimodal distributions, a fact worth of further analysis. In Figure 2 we show the original speech signal, Figure 3 shows the measurement in the presence of Gaussian noise corresponding to the 'peak' of the SNR gain curve (measurement SNR ≈ 2), Figure 4 shows the corresponding reconstruction with attenuated embedding threshold estimator. Similarly Figure 5 shows the measurement with uniform noise corresponding to the 'peak' of the uniform noise SNR gain curve (measurement SNR ≈ 1.2), while Figure 6 shows the corresponding reconstruction. In both cases the perceptual quality is better than the noisy measurements, which is not the case, for example, for hard threshold estimators in wavelet bases (when the SNR of the measurements is so low, see [3] for more details).

We stress again that none of the parameters of the algorithm were changed as we went from Gaussian white noise contaminations to more general white noise processes. Data files for the signal, measurement and reconstructions used to compute the quantities in all the figures are available upon request for direct evaluation of the perceptual quality.

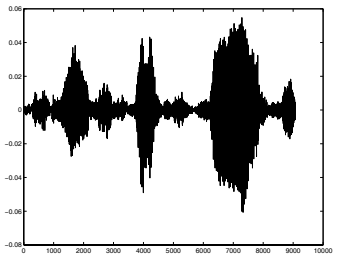

Fig. 4. Attenuated embedding estimate from the measurement in Figure 3, scaled to have norm 1.

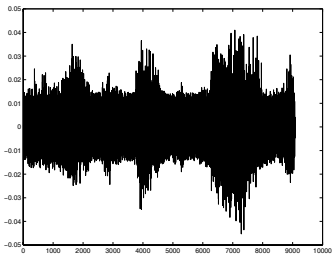

Fig. 5. Noisy measurement of the signal in Figure 2 with uniform white noise and scaled SNR of about 1.2db.

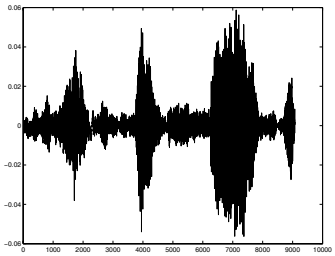

Fig. 6. Attenuated embedding estimate from the measurement in Figure 5, scaled to have norm 1.

4. REFERENCES

[1] K. T. Alligood, T. D. Sauer, J. A. Yorke, *Chaos. An introduction to Dynamical Systems*, Springer, 1996.

[2] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[3] D. Napoletani, C.A. Berenstein, T. Sauer, D. C. Struppa, D. Walnut, A Threshold Estimator for Speech Signals Based on State Space Embedding Reconstructions, submitted, 2004.

[4] T. Strohmer, Numerical Algorithms for Discrete Gabor Expansions, in *Gabor Analysis and Algorithms. Theory and Applications*, H. G. Feichtinger, T. Strohmer editors. Birkhauser, 1998.

[5] T. Sauer, J. A. Yorke, M. Casdagli, Embedology, *Journal of Statistical Physics*,**65** (1991), 579-616.