Delay-coordinates embeddings as a data mining tool for denoising speech signals

D. Napoletani[a)]
Center for Applied Proteomics and Molecular Medicine, George Mason University, Manassas, Virginia 20110

D. C. Struppa
Department of Mathematics and Computer Science, Chapman University, Orange, California 92866

T. Sauer
Department of Mathematical Sciences, George Mason University, Fairfax, Virginia 22030

C. A. Berenstein
Institute for Systems Research, University of Maryland, College Park, Maryland 20742

D. Walnut
Department of Mathematical Sciences, George Mason University, Fairfax, Virginia 22030

(Received 20 March 2006; accepted 11 October 2006; published online 8 November 2006)

In this paper, we utilize techniques from the theory of nonlinear dynamical systems to define a notion of embedding estimators. More specifically, we use delay-coordinates embeddings of sets of coefficients of the measured signal (in some chosen frame) as a data mining tool to separate structures that are likely to be generated by signals belonging to some predetermined data set. We implement the embedding estimator in a windowed Fourier frame, and we apply it to speech signals heavily corrupted by white noise. Our experimental work suggests that, after training on the data sets of interest, these estimators perform well for a variety of white noise processes and noise intensity levels. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2384909]

In this paper, we introduce a denoising algorithm that is designed to be efficient for a variety of white noise contaminations and noise intensities. We expand a measurement X in a windowed Fourier frame and then we use delay-coordinates embeddings as a data mining tool to extract coefficients likely to be generated by some underlying signal. We assume such signal belongs to a predetermined data set of interest (speech signals in our test case). The algorithm needs to be trained on the class of data we want to denoise and on a collection of white noise distributions with kurtosis up to some maximum allowed kurtosis K. Once the parameters of the algorithm are found, we verify that they do not need to be changed as we change noise variance, or noise probability density function as long as its kurtosis is less than K.

I. INTRODUCTION

A. Overview of the method

In this work, we analyze time series of the type $X(n)=F(n)+W(n)$, $n=1,\ldots,N$, where F is some sampled speech signal, and each $W(n)$ is a realization of some white noise process. Recall that, for a time series $X(n)$, $n=1,\ldots,N$, an embedding dimension d, and a time delay τ (in this context both taken to be positive integers), the *delay-coordinates embedding* $\mathbf{X}(n)=[X(n),X(n-\tau),\ldots,X(n-(d-1)\tau)]$, $n=(d-1)\tau+1,\ldots,N$, gives a faithful description of the underlining finite dimensional dynamics (if any) as long as the embedding dimension d is big enough (at least twice the dimension of the dynamical system[1,2]). We refer to Ref. 3 for a review of this and other nonlinear time-series techniques.

There are several powerful uses of nonlinear time-series techniques: among the better known are chaos control[4,5] and chaos synchronization with its applications to communication theory.[6,7] There has been a lot of emphasis on how to cope with noise and imperfect modeling (see, for example, Refs. 8 and 9), a fundamental problem when dealing with high-dimensional systems. In particular, for delay-coordinates embeddings, the appropriate choice of delay τ and embedding dimension d is a difficult practical issue, especially for high-dimensional systems.

This paper suggests a way to deal with this issue for a specific class of estimation problems and a specific class of signals, namely signals that have a sparse, localized representation in time frequency domain such as speech signals. Indeed, unlike what happens in most applications of delay-coordinates embeddings, we do not need in our work a precise reconstruction of the underlying dynamics. Instead, we measure the "squeezing" of the local dynamics along the principal direction, measured by the *embedding index* $\mathcal{I}(X)=\sigma(1)/\sigma(d)$, i.e., *the quotient of the largest and smallest singular values of the embedding image* $\mathbf{X}$ *of* X. This point of view reflects the focus of this particular application on data mining as compared to system identification.

Singular value decompositions have been used effectively to recover dynamics from noisy measurements glo-

[a)]Electronic mail: dnapolet@gmu.edu

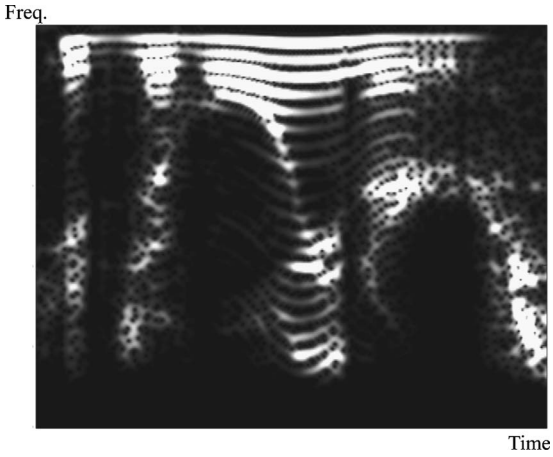

FIG. 1. We show part of a windowed Fourier representation of a speech signal; horizontal axis is the time coordinate, vertical axis is the frequency coordinate. Lighter areas correspond to coefficients with larger absolute value.

bally in Ref. 10 and locally in Ref. 11. In this paper, we look at singular value decompositions locally in a time-frequency representation, and we iteratively extract interesting dynamics in different such representations. The iterative procedure reduces the emphasis on reconstruction of a *specific and unique dynamics*, and closely connects the delay coordinate denoising methods to sparsity denoising techniques more frequently seen in signal processing and wavelet literature.

To localize the delay-coordinates embedding method in the time-frequency domain, we compute the embedding index for time series supported by line segments belonging to a collection C_p of horizontal line segments γ of length p in a time-frequency windowed Fourier frame expansion such as in Fig. 1 (see Ref. 12, Chap. 5 and Sec. I B in this paper for more on Fourier frames). We choose this shape for elements in C_p because speech signals are well represented in windowed Fourier frames and so we expect them to have sparse and very localized representations in the time-frequency domain. Moreover, we observed in practice that speech signals exhibit a nontrivial, but simple, dynamics along the time axis in the regions of the time-frequency domain where their representation is mostly concentrated. By sparse representation we mean that only a small fraction of the coefficients will have large absolute values. Note that white noise does not have such sparse representation in windowed Fourier frames (for a review of the idea of sparsity and its applications in signal processing, see Ref. 13, Chap. 11).

Given a choice of embedding dimension d and time delay τ, for each time series X defined on a line of coefficients in C_p we compute the embedding index $\mathcal{I}(X)$ and define $Q_X(t)$, the cumulative index density function that for each t gives the proportion of lines with index above t. We compute $Q_X(t)$ for 10 speech signals S_i and for several white noise time series W_j with bounded kurtosis. We will show that $Q_X(t)$ is very similar for all chosen probability density functions. Moreover, the decay of $Q_X(t)$ is much faster for noise than for the set of speech signals. If we call $Q_S(t)$ the average cumulative density function for our training speech series and $Q_W(t)$ the average cumulative density function for the white noise processes, we can define an index threshold T as the smallest positive value of t for which $Q_S(T) > 10Q_W(T)$.

We use the notion of index threshold to build a recursive algorithm that, given a measurement $X=F+W$, extracts lines of coefficients in the windowed Fourier frame to estimate F. The embedding threshold is used to extract from a measurement X a fraction of lines that are unlikely to be the result of white noise processes. The algorithm begins by setting the initial estimate $\tilde{F}=0$. For $k=1,\ldots,K$, the core steps to be repeated recursively are as follows:

A. Given a measurement X, compute its windowed Fourier frame coefficients.
B. Extract all coefficient lines in C_p with embedding index above the set threshold level T.
C. Generate a partial estimate F_k by using only the coefficients from the extracted lines.
D. Attenuate the partial estimate by some small coefficient $\alpha > 0$, i.e., set $F_k = \alpha F_k$. Put $X = X - F_k$, $\tilde{F} = \tilde{F} + F_k$.

One iteration of A–D essentially does the following: with steps A and B, we take the coefficients of the measurement X in the time-frequency domain that appear to exhibit significant dynamics locally along the time axis. With step C, we use only these significant coefficients to generate a partial estimate. Since we do not want to disrupt too much the dynamics of the windowed Fourier coefficients of X that we do not extract, in step D we reduce the norm of the partial estimate, and then we subtract it from the measurement. This partial estimate is also added to whatever current estimate we already have.

Since the repeated application of the loop A–D generates attenuated estimates, to evaluate the performance of the method we compute the signal-to-noise ratios (SNR) on measurements, speech signals, and estimates all scaled to have norm 1. Note that we need to train several parameters on the data set of interest. A proper selection of such parameters is shown to be possible for speech signals. After choosing the parameters, the algorithm is robust with respect to noise level and noise type. (The algorithm described in this paper is being patented, with provisional patent application number 60/562,534 filed on April 16, 2004.) Our work shows that delay-coordinates embeddings are an effective tool in denoising speech signals, but the method could be applied in principle to other types of data sets with a suitable choice of frames and with a different construction of the collection C_p.

Threshold estimators in time frequency and time scale representations have been used effectively to estimate signals from noisy measurements when it is possible to assume a sparse representation of the signals of interest. We mention here the seminal work of Ref. 14 on the optimality of wavelet threshold denoising for piecewise regular signals corrupted by the addition of Gaussian white noise. Several techniques have been developed to deal with the non-Gaussian case, some of the most successful are the Efromovich-Pinsker (EP) estimator (see Ref. 15) and the block threshold estimators of Cai and collaborators (see Ref. 16 and the more recent Ref. 17). In these methods, the variance of the white

process needs to be estimated from the data; moreover, since the threshold is designed to evaluate intensities (or relative intensities) of the coefficients in blocks of multiwavelets, low intensity details may be filtered out, as is the case for simpler denoising methods.

The method we describe in this paper does not require knowledge of the noise intensity level (thanks to the use of *quotients* of singular values of embedding images), and it is remarkably robust to changes in the type of noise distribution. For the same reason, fine low-intensity details can in principle be preserved, since quotients of singular values of embeddings are not dependent on intensity levels of the time series. This strength is achieved at a price: the inner parameters of the algorithm need to be adjusted to the data. This is true to some extent for the EP and block thresholding algorithms as well, but the number and type of parameters that need to be trained in our approach are increased by the need to choose a reasonably "good" delay-coordinates embedding suitable for the data we would like to denoise. In Sec. III, we briefly suggest possible ways to make the training on the data fully automatic, but it is yet to be seen at this stage which data sets are amenable to the analysis we propose.

In Sec. I B, we briefly review windowed Fourier frames and we formally define the collection of paths C_p. Moreover, we review the results on delay-coordinates embeddings that are used in our method. In Sec. II, we explore the properties of the cumulative density functions attached to the embedding index. In Sec. III, we define the attenuated embedding estimators that are the core of algorithm A–D and we give a technical version of the algorithm itself. In Sec. IV, we apply our method to several speech signals contaminated by four types of white noise, and we discuss the quality of the estimates.

B. Fourier frames and delay-coordinates embeddings

We now formally define some of the objects and techniques introduced in the previous subsection. Let $F[n]$, $n=1,\ldots,N$, be a discrete signal of length N, and let $X[n]=F[n]+W[n]$, $n=1,\ldots,N$, be a contaminated measurement of $F[n]$, where $W[n]$ are realizations of a white noise process W. For a given discrete orthonormal basis $B=\{g_m\}$ of the N-dimensional space of discrete signals, we can write $X=\sum_{m=0}^{N-1}X_B[m]g_m$, where $X_B[m]=\langle X,g_m\rangle$ is the inner product of X and g_m.

Note that any discrete periodic signal $X[n]$, $n\in\mathbb{Z}$ with period N can be represented in a discrete windowed Fourier frame. This frame is a localization of the Fourier transform that allows us to look at changes of frequency of signals in short neighborhoods by using a masking window g; more specifically, the elements of the windowed Fourier frame are of the form

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

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 for many applications, but we have deliberately selected this function to avoid excessive smoothing, which we found to adversely affect our algorithm. Under the previous conditions, the signal X can be completely reconstructed from the inner products $\mathcal{F}X[m,l]=\langle X,g_{m,l}\rangle$, i.e.,

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

where

$$\tilde{g}_{m,l}[n]=g[n-m]\exp\left(\frac{i2\pi ln}{N}\right),\quad n\in\mathbb{Z}. \tag{3}$$

We denote the collection $\{\langle X,g_{m,l}\rangle\}$ by $\mathcal{F}X$. For finite discrete signals of length N, this 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 Refs. 12 and 18.

We now 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

$$F(X)=[h(X),h(S_{-\tau}(X)),\ldots,h(S_{-(d-1)\tau}(X))], \tag{4}$$

where by $S_{-j\tau}(X)$ we denote the state of the system with initial condition X at $j\tau$ 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 Refs. 1, 2, and 19):

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.*

Keeping in mind that generally the most significant information about g is the knowledge of the attractive invariant subsets, we can say that delay maps allow a faithful description of the underlying finite dimensional dynamics, if any. The previous theorem can be extended to invariant sets A that are not topological manifolds; in that case more sophisticated notions of dimension are used (see Ref. 2).

Let now the measuring function h be the identity function and assume from now on that τ is an integer delay so that $F(W[n])=[W[n],W[n-\tau],\ldots,W[n-(d-1)\tau]]$. Note that if the delay-coordinate procedure is applied to the time series $W[n]$, $n=1,\ldots,N$, for W an uncorrelated random process, then for any embedding dimension d the state space will be filled according to a spherically symmetric probability distribution. In other words, the expected value of the embedding index $\mathcal{I}(W)=\frac{\sigma_1}{\sigma_d}$ is 1 regardless of the choice of embedding parameters. Indeed, since W is a white noise process, each coordinate of $F(W[n])$ is a realization of some random variable with some given probability density function g, therefore $F(W)$ is a realization of a multivariate random variable of dimension d and symmetric probability distribution.

Since speech signals exhibit smooth and regular dynamics locally, we expect their embedding index computed on elements of C_p to be far from 1, i.e., not to fill the state space uniformly. This observation motivates our attempt to separate (locally in time frequency space) structure belonging to speech signals from background noise in the windowed Fourier frame, by using the embedding index. Note, however, that, even when X is a pure white noise process, the windowed Fourier frame will enforce a certain degree of smoothness along each line in C_p since consecutive points in each line segment are inner products of frame atoms with partially overlapping segments of X. So there will be some correlation in the coefficients along elements of C_p even when X is an uncorrelated time series, and the embedding index calculated along lines in C_p may be much larger than 1 even for such pure noise processes. We need, therefore, to carefully analyze in the next section the value distribution of the embedding index $\mathcal{I}$ to understand whether it behaves differently for uncorrelated random processes and for signals in a database of speech signals.

II. SEPARATING SPEECH SIGNALS AND RANDOM PROCESSES

Our interest in this paper is in estimating speech signals from noisy measurements, and much of the structure of speech signals in the time frequency domain is contained in localized "ridges" that are oriented in the time direction (as we can see from Fig. 1). Given this localization property of speech signals, we expect the collection C_p of double-indexed horizontal lines

$$\gamma_{\bar{m},\bar{l}} = \{g_{m,l} \text{ such that } l = \bar{l}, \bar{m} \leq m \leq \bar{m} + p\}, \tag{5}$$

where p is some positive integer, to be relatively sensitive to local time changes of such ridges, since each element in C_p is a short line segment in the time frequency domain oriented in the time direction. For a given time series X and choice of parameters (q,p,τ,d) (recall q is the length of the window in the windowed Fourier frame), we compute the collection of embedding indexes $\mathcal{I}(\mathcal{F}X) = \{\mathcal{I}(\mathcal{F}X_\gamma), \gamma \in C_p\}$. Define now formally the *index cumulative function* as

$$Q_X(t) = \frac{\#\{\gamma \text{ such that } \mathcal{I}(\mathcal{F}X_\gamma) > t\}}{\#\{\gamma\}}, \tag{6}$$

i.e., for a given t, $Q_X(t)$ is the fraction of paths that have index above t. A simple property of Q_X will be crucial in the following discussion.

Lemma 1: *If X is a white noise process and $X' = aX$ is another random process that component by component is a rescaling of X by a positive number a, then the expected function Q_X and $Q_{X'}$ are equal.*

Proof: Each set of embedding points generated by one specific path γ is, coordinate by coordinate, a linear combination of some set of points in the original time series. Therefore, if $X' = aX$, $\mathcal{F}X'_\gamma = a\mathcal{F}X_\gamma$, but the quotient of singular values of a set of points is not affected by rescaling of all coordinates, therefore the distributions of $\mathcal{I}(\mathcal{F}X)$ and $\mathcal{I}(\mathcal{F}X')$ are equal. Since $Q_{X'}$ and Q_X are defined in terms of $\mathcal{I}$, they are equal as well.

Remark 1: The choice of p in C_p is very important in practice. The speech signals that we consider are sampled at a sampling frequency of about 8100 Hz. We choose support of the window $q=64$ and length of the paths $p=2^8$. This assures that the length of each path is at least of the same order of magnitude as the duration of stationary vocal emissions. Given this length p for γ, we need to have embedding dimension d and time delay τ so that $d\tau \ll p$, so that for each path we will generate a sufficiently large number of points in the embedding image. Because of these restrictions, we set $d=4$ and $\tau=4$, which give $d\tau = 2^4 \ll p = 2^8$; we generate in this way 240 points for each path. We heuristically adjusted the embedding parameters d and τ and the length p of the paths so that the qualitative behavior of speech signals and white noise processes was as distinct as possible. See the discussion in Sec. III, remark 4 for a possible way to make the choice of parameters automatic. We now expand some uncorrelated zero mean random processes of length $N=2^{11}$ on the windowed Fourier frame with the parameters $q=64$, $p=2^8$, $d=4$, and $\tau=8$.

Noise distributions: The specific random processes we use are time series with each point a realization of a random variable with the following:

(1) Gaussian probability density function (pdf).
(2) Uniform probability density function.
(3) Tukey probability density function, that is, a sum of two normal distributions with uneven weight: each point of the time series is a realization of the random variable $W = RN_1 + (1-R)4N_2/\sqrt{r+16(1-r)}$, where N_1 and N_2 are Gaussian random variables, and R is a Bernoulli random variable with $P(R=1)=0.9$ and $r=P(R=1)$. This is an interesting example of heavy-tail distribution (used in Ref. 15 as well).
(4) Discrete bimodal pdf with values in $\{-V,V\}$ for some positive V.

All probability density functions are set to have mean zero and variance 1, since by Lemma 1 we know Q_* will not be affected by changes of the variance. The Tukey pdf has heavy tail and is therefore a good example of highly non-Gaussian distribution. The kurtosis is 3 for the pdf in (1), about 1.8 for the pdf in (2), about 13 for the pdf in (3), and about 1.2 for the pdf in (4).

In Fig. 2(a), we plot $Q_X(t)$ for the white noise processes generated with pdf's in (1)–(4), averaged over ten repetitions for each random distribution. Note that the qualitative behavior of Q_X is very similar for all chosen distributions; in particular, they all exhibit a very fast decay for larger values of t. The maximum L^2 distance between any two Q_X in the interval $[0,40]$ is ≈ 0.54 (or some 6% of the average L^2 norm of the Q_X); we found that even for distributions with kurtosis up to 50, the maximum distance was less that 0.8 (about 8.5% of the average L^2 norm of Q_X), irrespective of the specific pdf, moreover most of the error is concentrated in regions of high intensity of the derivative and it does not affect much the behavior of the right tail of the curves Q_X.

Remark 2: To speed up the computation, we sampled the indexes $(\bar{m},\bar{l})$ of the lines in (5); specifically we selected

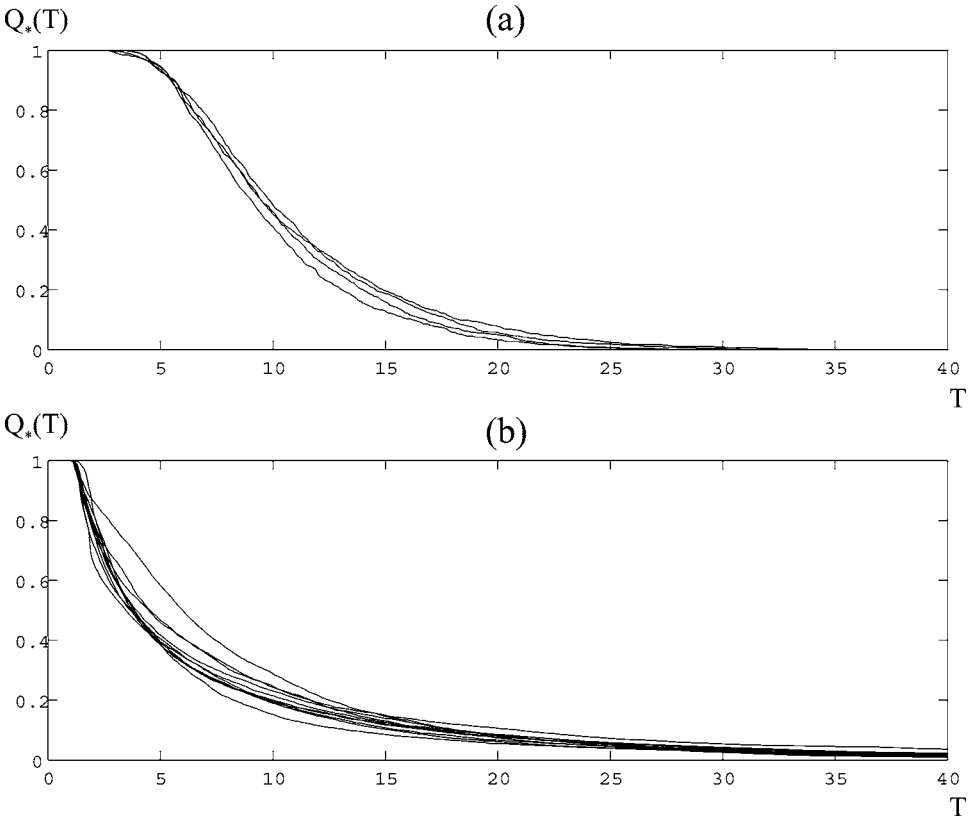

FIG. 2. From top to bottom, this figure shows Q_*, as defined in Eq. (6) for (a) uncorrelated random processes (1) to (4) defined in the text; (b) ten randomly selected segments of speech signal from the TIMIT database.

a sampling length of $S_{\bar{m}}=1$ for the frequency index $\bar{m}$ and a sampling length of $S_{\bar{l}}=p$ for the time index.

The implication of our computations is that moderately heavy tail distributions will not exhibit a significantly different behavior for Q_X with respect to the Gaussian distribution, supporting our claim that Q_X is robust with respect to the choice of white noise distribution. For each probability density function, the shape of Q_X is affected by the correlation introduced by the length of q (the window support of the windowed Fourier frame): if $\tau<q$, some coordinates in each embedding point will be correlated and this will cause the decay of Q_X to be slower when τ is smaller. Compute now Q_X (with the same choice of parameters) for a collection of 10 randomly selected segments of speech signals of length 2^{11}. The rate of decay of the functions Q_X is significantly different, and the tail of the functions is still considerably thick by the time the rate of decay of Q_X for most random processes is almost zero [see Fig. 2(b)]. To have a significantly larger fraction of paths retained for speech signals rather than noise, we select the threshold T as follows:

Determination of threshold (DT): *Given a choice of* (q,p,τ,d), *a collection of training speech time series* $\{S_j\}$, *and a selection of white noise processes* $\{W_i\}$, *choose* T_0 *to be the smallest t so that the mean of* $Q_{S_j}(T_0)$ *is one order of magnitude (10 times) larger than the mean of* $Q_{W_i}(T_0)$.

This heuristic rule gives, for the parameters in this section, $T_0 \approx 28.2$. Rule (DT) gives us as experimental way to determine a threshold $T=T_0$ for the index $\mathcal{I}$ that removes most of the time frequency structure of some *predetermined* noise distributions, while it preserves a larger fraction of the time frequency structure of speech signals. Since, moreover, "reasonable" distributions exhibited a Q_X similar to the one of Gaussian distributions, we can in practice train the threshold only on Gaussian noise and be assured that it will be a meaningful value for a larger class of distributions.

Note that even though very low energy paths could have in principle a high embedding index, the energy concentration in paths that have very high index tends to be large for speech signals. To see that, for a given signal X, let

$$E_X(t) = \frac{\Sigma\{|\mathcal{F}X_\gamma|_2 \text{ such that } \mathcal{I}(\mathcal{F}X_\gamma) > t\}}{\Sigma|\mathcal{F}X_\gamma|_2} \tag{7}$$

be the fraction of the total energy contained in paths with index above t. We can see in Fig. 3 that the amount of energy contained in paths with high index value is significantly larger for speech signals than for noise distributions. More precisely, the fraction of the total energy of the paths carried by paths with $\mathcal{I}>T_0$ is on average 0.005 for the noise distributions and 0.15 for the speech signals, or an increase by a factor of 30. It seems, therefore, that the embedding index $\mathcal{I}$, with our specific choice of parameters, is quite effective in separating a subset of lines that are likely to be generated by speech signals. Note, moreover, that similar results can be obtained by perturbing p, τ, and d, which suggests an intrinsic robustness of the separation with respect to the parameters.

The ability of the embedding index to detect significant lines of coefficients in C_p is mostly attributed to the very nice properties of speech signals as they are well represented in windowed Fourier frames and so we expect them to have sparse and very localized representations in the time fre-

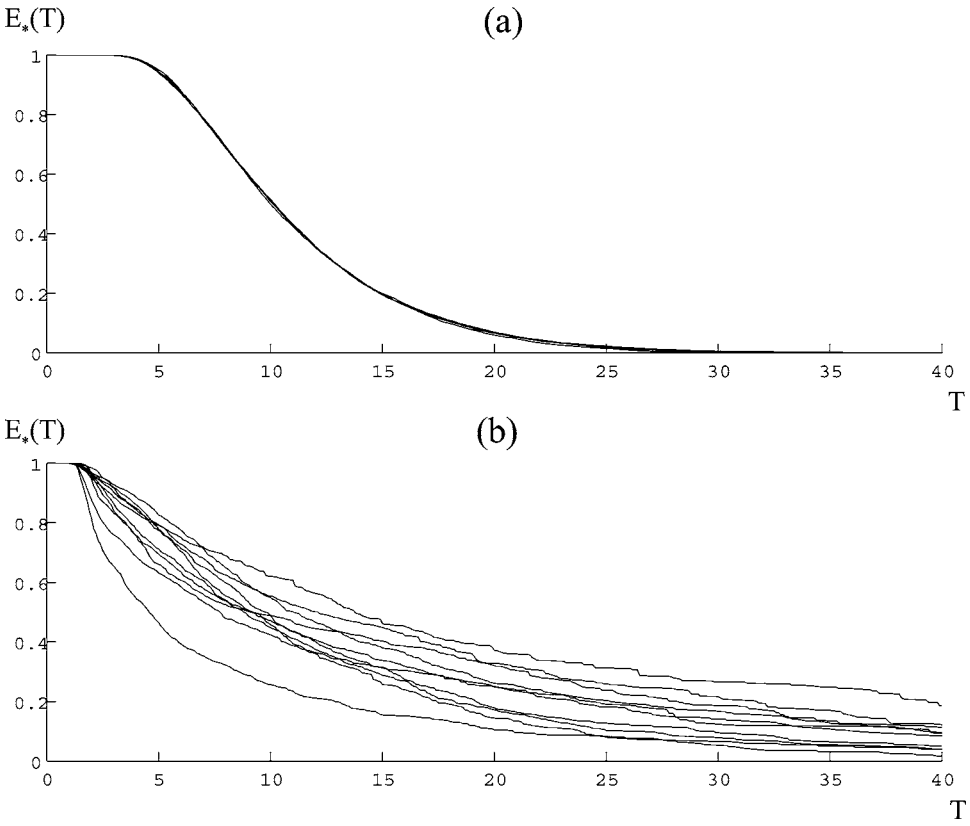

FIG. 3. From top to bottom, this figure shows E_*, as defined in Eq. (7) for (a) the uncorrelated random processes (1) to (4) defined in the text; (b) the segments of speech signals as in Fig. 1(b).

quency domain, while white noise does not have such sparse representation. Moreover, the absolute value of the coefficients of speech signals will smoothly, but irregularly, oscillate along the time axis in those regions of the time frequency domain where the signals are well represented. Therefore, speech signals appear to exhibit a nontrivial, but relatively simple, dynamics locally along the time axis in the regions of the time frequency domain where their representation is mostly concentrated.

There is a large literature on possible ways to distinguish deterministic dynamical systems from random behavior (see, for example, the articles collected in Ref. 20). Generally, the identification of the "best" τ and d that allow a faithful representation of the invariant subset is considered very important in practical applications of delay-coordinates embeddings (as discussed, for example, in Ref. 21), as it allows to make transparent the properties of the invariant set itself; specifically, we want to deduce from the data themselves the dimension m of the invariant set (if any) so that we can choose a d that is large enough for the theorem to apply. Moreover, the size of τ has to be large enough to resolve the image far from the diagonal, but small enough to avoid decorrelation of the delay coordinates point.[22]

The algorithm A–D outlined in Sec. I uses the structure of the embedding in such a way that the identification of the most suitable τ and d is not crucial (as it would be if we were interested in extracting the exact dynamics underlying the time series). While we do need to train the algorithm on the available data to find the most suitable τ, d, and embedding threshold T [by analyzing the cumulative density functions $Q_*(t)$ of the embedding index], such analysis does not require a full resolution of the underlying dynamics and is conceptually quite simple.

The use of embedding techniques in the context of computational harmonic analysis frees us from the need to use embedding techniques to effectively model the signals. Instead, we use time delay embeddings as *data mining tools*. We specifically use the term "data mining" to highlight the necessity of adjusting the parameters of the algorithm according to predetermined training sets of signals of interest and of relevant white noise distributions. The goal becomes simply to "separate" to some extent the behavior in the time frequency domain of the embedding index of the two training sets of interest by a suitable choice of parameters.

Remark 3: Note that if the dimension of the invariant set A is $d_A=0$, then for any white noise process W, $X+W$ has spherically symmetric embedding image and $\frac{\sigma_1}{\sigma_d}\approx 1$ for any embedding dimension d as in the case of pure white noise. This means that an estimator based on $\mathcal{I}$ is not able to estimate noisy constant time series on a given path γ. This restriction can be eased by allowing information on the distance of the center of the embedding image to be included in the definition of the embedding threshold estimator. In this paper, for simplicity we assumed $d_A>0$ for all lines in C_p. This seems to be sufficient since, as we already pointed out, speech signals generally have nontrivial dynamics locally along lines of C_p in the regions of the time frequency domain where they are mostly concentrated.

III. ATTENUATED EMBEDDING ESTIMATORS

In this section, we refine the denoising algorithm A–D based on the embedding index. Essentially we use the em-

bedding index to select those coefficients in the window Fourier frame that are likely to carry significant information on the signal, and we recursively extract them to generate partial estimates of the signal itself. Following the structure of threshold estimators (see Refs. 14 and 12, Chap. 10), we define *an embedding estimator* in the windowed Fourier frame as follows:

$$\tilde{F} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{l=0}^{N-1} d_{\mathcal{I},T}(\mathcal{F}X[m,l])\tilde{g}_{m,l}, \tag{8}$$

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

Essentially the embedding estimator in (8) is a modification of the reconstruction formula in (2) where we use only coefficients contained in lines with large embedding index. This estimator is the core of our algorithm A–D, but it is slightly modified to improve the actual performance by attenuating each partial estimate in algorithm A–D by the coefficient α. To implement a relatively fast version of the algorithm, we define tubular neighborhoods for each atom in the windowed Fourier frame,

$$\mathcal{O}(g_{m,l}) = \{g_{m',l'} \text{ s.t. } |l'-l| \leq 1, |m'-m| \leq 1\}, \tag{9}$$

and we make a decision on the value of the coefficients in the *two-dimensional* neighborhood of the line of coefficients $\mathcal{F}X_\gamma$ based on the embedding index of the *one-dimensional* line of coefficients $\mathcal{F}X_\gamma$.

The following is a technical version of the method A–D described in the Introduction. We assume that embedding dimension d and time delay τ have already been set. Let K be some large integer, set values for the windows q_k, $k=1,\ldots,K$, and use (DT) to find the thresholds $T(q_k) = T(d,\tau,q_k)$. We denote by $\tilde{F}$ the partial current estimate of the signal and by R the current residue. Set initially $\tilde{F}=0$ and $R=X$, where X is the given measurement. Choose a positive small attenuation coefficient α. For $k=1,\ldots,K$, repeat $\tilde{A}$–$\tilde{D}$.

- $\tilde{A}$ Given R, expand R in a windowed Fourier frame with window size q_k.
- $\tilde{B}$ Set $\mathcal{F}Y[m,l] = \mathcal{F}R[m,l]$ if $\mathcal{I}(\mathcal{F}R_\gamma) \geq T(q_k)$ 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 $\mathcal{I}(\mathcal{F}R_\gamma) < T(q_k)$ for all γ containing $g_{m',l'}$, $g_{m',l'} \in \mathcal{O}(g_{m,l})$.
- $\tilde{C}$ Let Y be the inverse image of $\mathcal{F}Y$.
- $\tilde{D}$ Set $Y = \alpha Y$. Set $\tilde{F} = \tilde{F} + Y$ and $R = R - Y$.

Note that the details of the implementation $\tilde{A}$-$\tilde{D}$ are in line with the general strategy of greedy regression.[22] The window length q in step $\tilde{A}$ can change from one iteration to the next to "extract" possible structure belonging to the underlining signal at several different scales. The "thickening" of the lines in Eq. (9) increases the number of nonzero coefficients chosen in $\tilde{B}$, while the attenuation performed in $\tilde{D}$ decreases their contribution. In this way, we are allowing more information to be taken at each iteration of algorithm, but in a slow learning fashion that in practice increases the sharpness of the estimate since it does not disrupt too much the dynamics of the coefficients of the measurements at each iteration. On the general issue of attenuated learning processes, see the discussion in Ref. 23, Chap. 10. *Note that the attenuation of coefficients leads to improved results only when it is part of a recursive algorithm, otherwise it gives only a rescaled version of the estimate.*

Choice of parameters: One drawback of the algorithm $\tilde{A}$–$\tilde{D}$ is the need to choose several parameters. We choose the window q for the windowed Fourier frames, the length p of lines in C_p, the embedding parameters τ (time delay) and d (embedding dimension), and the learning parameters T (threshold level), α (attenuation coefficient), and K (number of iterations).

We stress that all such choices are context-dependent, and are the price to pay to have an estimator that is relatively intensity-independent and applicable to wide classes of noise distributions. We give here some general indications on how to select these parameters. First of all, the algorithm is not very sensitive to the choice of the length q of the window in the Fourier frame, while the use of several windows is found to be always beneficial. Let us explore now the relation of parameters associated with C_p, embedding parameters τ and d, and threshold T. Recall that for the collection C_p we have as parameters the time and frequency sampling rates $\bar{l}$ and $\bar{m}$ and the length p of the paths. The frequency sampling rates $\bar{l}$ and $\bar{m}$ are necessary only to speed up the algorithm; ideally we would like a dense sampling. The same considerations apply to the "thickening" of the lines in (9). We basically try to speed up the algorithm by collecting more data at each iteration. The truly essential parameters are the path length p, the embedding parameters, and the threshold T. Essentially we want to set these parameters so that the number of paths that have index $\mathcal{I} > T$ is sizeable for a training set of speech signals and marginal for the white noise time series of interest. Our experience is that such a choice is possible and robust; in the previous section, we gave a simple rule to find the threshold T in step (DT), given a choice of (q,p,τ,d). The choice of α and K is completely practical in nature: we probably want α as close to zero as possible and K as large 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 computing and time requirements of the specific problem.

Remark 4: We did not give an automatic algorithm to select the paths' length and embedding parameters: a possibility in this direction is to build a learning algorithm to find all of T, line length p, and embedding parameters. More specifically, let $\bar{Q}_S(x)$ be the mean of the functions $Q_{S_i}(x)$ for a training set of speech signals S_i and $\bar{Q}_W(x)$ be the mean of the functions $Q_{W_i}(x)$ for a set of white noise time series W_i. We can split the process so that we first find d, τ, and p such that the distance of the functions $\bar{Q}_W(x)$ and $\bar{Q}_S(x)$ is maximum in the L^2 norm. After finding these parameters, we can find a value of T according to rule (DT).

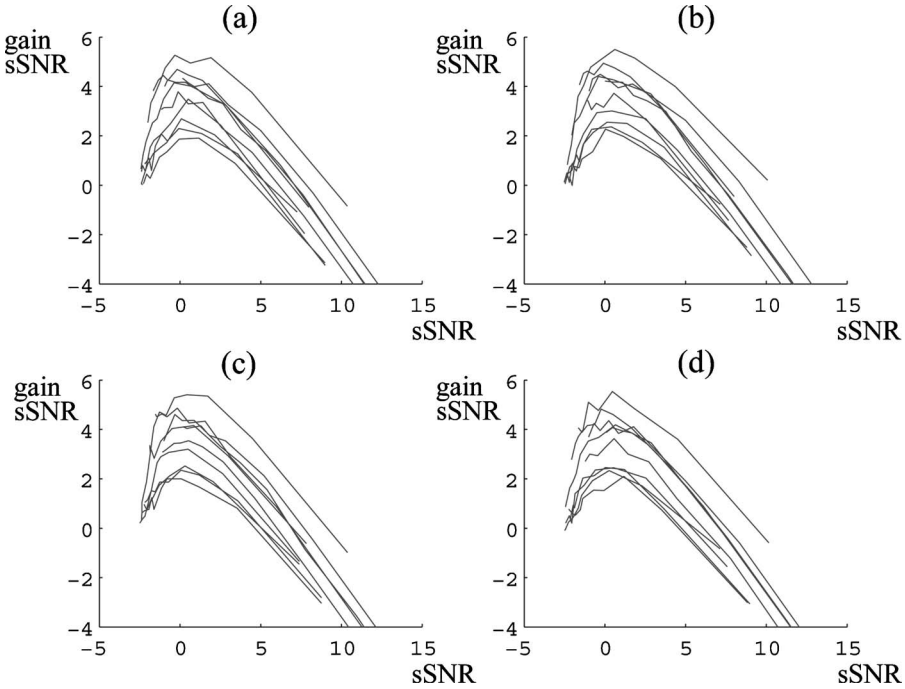

FIG. 4. Scaled SNR gain in decibel of the attenuated embedding estimates plotted against the scaled SNR of the corresponding measurements. From top left in clockwise order we consider the case of (a) Gaussian white noise; (b) uniform noise; (c) Tukey white noise; (d) discrete bimodal distribution.

IV. RESULTS AND DISCUSSION

In this section, we explore the quality of the attenuated embedding threshold as implemented in the windowed Fourier frame and with our class of lines $\mathcal{C}_p$. We apply the algorithm to 10 speech signals 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 embedding dimension is chosen to be $d=4$, the delay along the paths $\tau=4$, the length of the paths is $p=2^8$, and we set the window length of the windowed Fourier transform to be $q_k=25$ for k even and $q_k=100$ for k odd (to detect both features with good time localization and those with good frequency localization). For these parameters and for the speech signals that we used as a training set, we have $T \approx 26.8$ when $q_k=100$ and $T \approx 27.4$ when $q_k=25$ using the procedure (DT) of Sec. II. The sampling interval of the paths in the frequency direction is $S_{\vec{m}}=3$ and along the time direction is $S_{\vec{l}}=p/2$ We select $\alpha=0.1$ and $K=6$. 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.

We note that the attenuated embedding threshold is able to extract only a small fraction of the total energy of the signal f, exactly because of the attenuation process, therefore the signal-to-noise ratio (SNR) computations are done on scaled measurements X, estimates $\widetilde{F}$, and signals F set to be all of norm 1. We call such estimations *scaled* SNR or sSNR, and we explicitly write, for a given signal F and estimation Z,

$$\mathrm{sSNR}(Z) = 10 \log_{10} \frac{1}{E(|F/|F| - Z/|Z||)}.$$

We compute $\mathrm{sSNR}(X)$ and $\mathrm{sSNR}(\widetilde{F})$ by approximating the expected values $E(|F/|F| - X/|X||)$ and $E(|F/|F| - \widetilde{F}/|\widetilde{F}||)$ with an average over several realizations for each white noise contamination.

In Fig. 4, we show the gains of the scaled SNR of the reconstructions (with the attenuated embedding threshold estimator) plotted against the corresponding scaled SNR of the measurements. Each curve corresponds to one of 10 speech signals of approximately one second used to test the algorithm. From the top left in the clockwise direction, we have measurements contaminated by random processes with pdf's (1) to (4) as defined in Sec. II and with several choices of variance. Note that the overall shape of the sSNR 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 sSNR around 1 decibel. Note that the right tail of the sSNR gains takes often negative values; this is due to the attenuation effect of the estimator that is pronounced for the high-intensity speech features, but it is not necessarily indicative of worse perceptual quality with respect to the measurements. Some of the figures in the following will clarify this point.

For Gaussian white noise, we compared our algorithm to the block thresholding algorithm described in Ref. 16. We used the Matlab code implemented in Ref. 24, made available at `www.jstatsoft.org/v06/i06/codes/` as a part of their comparison of denoising methods. As the block thresholding estimator was implemented in a symmlet wavelet basis that is not well adapted to the structure of speech signals, a more compelling comparison would require the development of an embedding threshold estimator in a wavelet basis. In Fig. 5, we show the sSNR gain for all tested speech signals using the block threshold estimator (right plot) and attenuated embedding estimator (left plot). Note how, for the embedding threshold estimator, there is a higher maximum sSNR gain and how the sSNR improvement is more uniform for all tested speech signals. Consider that, for very low sSNR of the measurements, the degradation of the

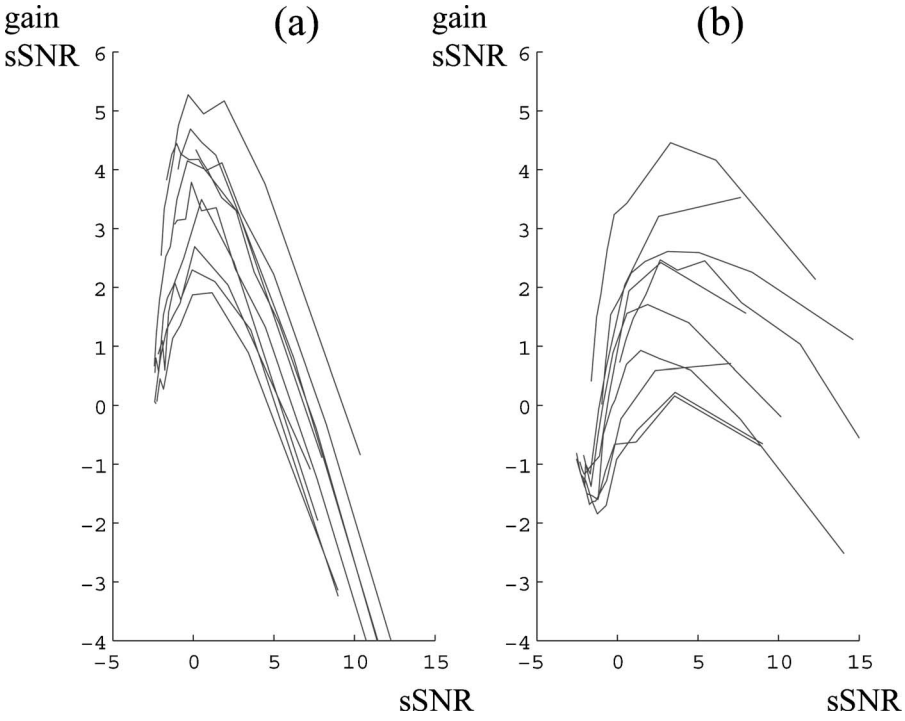

FIG. 5. sSNR gain for the estimates of ten speech signals and Gaussian additive noise using (a) the embedding threshold estimator; (b) the block thresholding estimator.

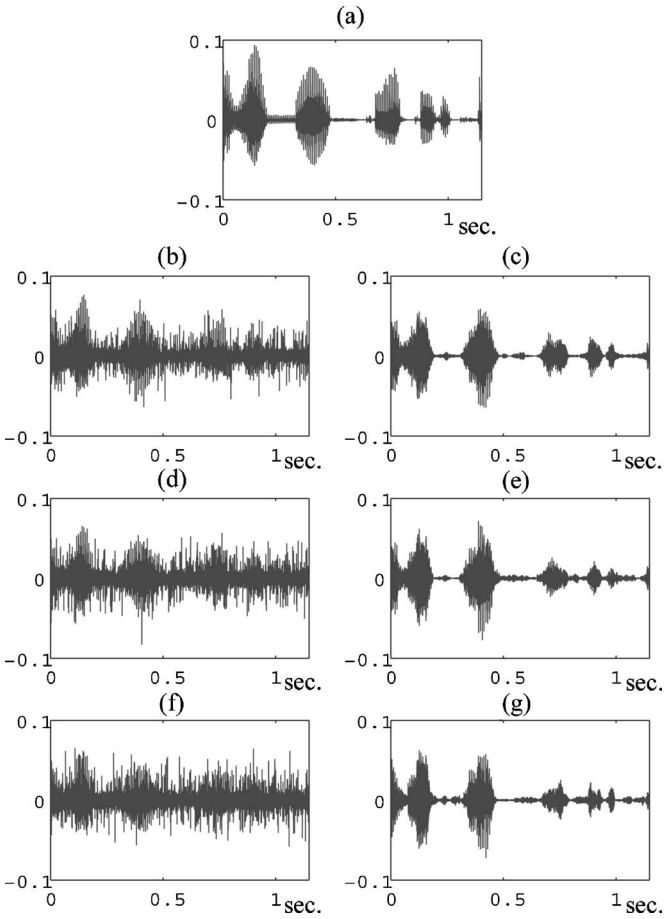

FIG. 6. The subplot (a) on the top shows a speech signal scaled to have norm 1. Subplots (b), (d), and (f) show noisy scaled measurement of the signal with Tukey white noise and decreasing scaled SNR (sSNR), respectively, equal to 4.6 , 2.6 , and 1.3 db. Subplots (c), (e), and (g) show the corresponding scaled attenuated embedding estimate with sSNR equal to, respectively, 5.8 , 5.3 , and 4.6 db.

block thresholding estimates is mostly due to a loss of low-intensity details. We believe the performance of the embedding threshold estimators is particularly compelling exactly in these cases; most speech structures are detected even in very noisy measurements with our method.

In Fig. 6, we show, on the left, top to bottom, three measurements of decreasing sSNR with added Tukey white noise and on the right the corresponding embedding threshold estimates. In particular, subplots (f) and (g) show measurement and estimate for measurement sSNR ≈ 1 (i.e., corresponding to the "peak" of the sSNR gain curve). Note how the shape and position of all speech envelopes is detected in this extremely high noise case; this is even more striking considering the highly non-Gaussian nature of the added Tukey noise (with kurtosis equal to 13). In all cases, the perceptual quality of the estimates appears to be better than that of the noisy measurements. It must be noted, though, that the estimates for bimodal and uniform noise were not intelligible at the peak of the sSNR gain curve (just as the measurements were not).

We stress that even though the threshold T was found using only Gaussian white noise as the training distribution, none of the parameters of the algorithm were changed as we went from Gaussian white noise contaminations to more general white noise processes, yet the sSNR gain was similar, 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.

[1]F. Takens, *Detecting Strange Attractors in Turbulence*, Lecture Notes in Mathematics Vol. 898 (Springer-Verlag, Berlin, 1981).
[2]T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," J. Stat. Phys. **65**, 579 (1991).
[3]H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 2003).

[4]E. Ott, C. Grebogi, and J. A. Yorke, "Controlling chaos," Phys. Rev. Lett. **64**, 1196 (1990).
[5]F. J. Romeiras, C. Grebogi, E. Ott, and W. P. Dayawansa, "Controlling chaotic dynamical systems," Physica D **58**, 165 (1992).
[6]L. M. Pecora and T. L. Carroll, "Synchronization in chaotic systems," Phys. Rev. Lett. **64**, 821 (1990).
[7]T. L. Carroll and L. M. Pecora, "Using multiple attractor chaotic systems for communication," Chaos **9**, 445 (1999).
[8]K. Judd and L. A. Smith, "Indistinguishable states II: The imperfect model scenario," Physica D **196**, 224 (2004).
[9]L. A. Smith, "Disentangling uncertainty and error: On the predictability of nonlinear systems," in *Nonlinear Dynamics and Statistics*, edited by A. I. Mees (Birkhauser, Boston, 2000), pp. 31–64.
[10]D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," Physica D **20**, 217 (1986).
[11]T. Sauer, "A noise reduction method for signals from nonlinear systems," Physica D **58**, 193 (1992).
[12]S. Mallat, *A Wavelet Tour of Signal Processing* (Academic Press, San Diego, 1998).
[13]S. Jaffard, Y. Meyer, and R. Ryan, *Wavelets: Tools for Science and Technology* (SIAM, Philadelphia, 2001).
[14]D. Donoho and I. Johnstone, "Minimax estimation via wavelet shrinkage," Ann. Stat. **26**, 879 (1998).
[15]S. Efromovich, J. Lakey, M. C. Pereyra, and N. Tymes, "Data-driven and optimal denoising of a signal and recovery of its derivative using multiwavelets," IEEE Trans. Signal Process. **52**, 628 (2004).
[16]T. Cai and B. W. Silverman, "Incorporating information on neighboring coefficients into wavelet estimation," Sankhya, Ser. B **63**, 127 (2001).
[17]T. Cai and M. Low, "Nonparametric function estimation overshrinking neighborhoods: Superefficiency and adaptation," Ann. Stat. **33**, 184 (2005).
[18]T. Strohmer, "Numerical algorithms for discrete Gabor expansions," in *Gabor Analysis and Algorithms. Theory and Applications*, edited by H. G. Feichtinger and T. Strohmer (Birkhauser, Boston, 1998).
[19]K. T. Alligood, T. D. Sauer, and J. A. Yorke, *Chaos. An Introduction to Dynamical Systems* (Springer, New York, 1996).
[20]*Nonlinear Dynamics and Statistics*, edited by A. Mees (Birkhauser, Boston, 2001).
[21]E. Bradley, "Time-series analysis," in *An Introduction to Intelligent Data Analysis*, edited by M. Berthold and D. Hand (Springer Verlag, Berlin, 1999).
[22]G. Davis, S. Mallat, and M. Avelaneda, "Adaptive greedy approximations," Constructive Approx. **13**, 57 (1997).
[23]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).
[24]A. Antoniadis, J. Bigot, and T. Sapatinas, Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study, 2001, available http://www.jstatsoft.org/v06/i06/