Denoising Human Speech Signals Using Chaoslike Features

Rainer Hegger, Holger Kantz, and Lorenzo Matassini

Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 38, D 01187 Dresden, Germany

(Received 15 April 1999; revised manuscript received 22 October 1999)

A local projective noise reduction scheme, originally developed for low-dimensional stationary deterministic chaotic signals, is successfully applied to human speech. This is possible by exploiting properties of the speech signal which resemble structure exhibited by deterministic dynamical systems. In high-dimensional embedding spaces, the strong inherent nonstationarity is resolved as a sequence of many different dynamical regimes of moderate complexity.

PACS numbers: 87.19.Dd, 05.40.Ca, 05.45.Jn, 43.60.+d

Nonlinear time series analysis comprises a set of techniques for the analysis, manipulation, and understanding of aperiodic signals relying on the hypothesis of deterministic chaos [1,2]. This means that the signal reflects the complex dynamics of a purely deterministic (often few-degree-of-freedom) system. Since such signals represent a very limited class with most probably no relevance for field measurements, it is of considerable interest to explore how far these concepts can be applied more generally to aperiodic signals with nondeterministic origin and strong nonstationarity.

Noise reduction for human speech is of high technological relevance. Background noises of all kinds are inconvenient in telecommunication, deteriorate automatic speech recognition, and cause serious difficulties for the users of electronic hearing aids. A well-established remedy is filtering of the signal in the frequency domain, in the simplest case by bandpass filters. Because of the aperiodic nature of voice signals, expressed by broadband power spectra, simple filters distort the signal more than they reduce noise, such that more refined methods are required. As a consequence of the high technical relevance of this field, very sophisticated combinations of filtering techniques using Fourier transforms, wavelet transforms, and other methods are in use in the specialized field of speech processing. Examples of state-of-the-art results are Refs. [3,4], and a modern spectral subtraction scheme which became a kind of standard was introduced in [5].

Models of the sound generating mechanism of humans indicate that stationary articulated human voice such as the vowel "aaaaaa" (German pronunciation) have a low-dimensional origin. In [6], a two-mass model of vocal-fold vibrations is analyzed with methods from nonlinear dynamics; it is shown that a sufficiently large tension imbalance of the left and right vocal fold induces bifurcations to subharmonic regimes, toroidal oscillations, and chaos. The reconstruction of attractors and the estimation of their properties indicate low dimensionality of the attractors generating the signal. Furthermore, it was shown by means of empirical orthogonal functions that normal phonation is well represented by only two eigenmodes. The simulation of disordered voice has shown that the three strongest modes contain 90% of the variance [7]. In contrast to stationary voice signals, the concatenation of different phonemes to full words or sentences does not represent a low-dimensional system, since there are frequent and arbitrary switches between different kinds of dynamical behavior. Because of the transition regions from the preceding phoneme and to the successive phoneme, a phoneme inside a word even differs from the same phoneme when it is isolated and elongated.

In this Letter we report on the successful application of a noise reduction scheme which was originally developed using exclusively properties of low-dimensional stationary deterministic dynamical systems. The success is possible by the identification and exploration of quasideterministic structure in the voice signal. Hereby, the reconstruction of an embedding space allows one to efficiently cope with the problem of nonstationarity. The different phonemes are identified implicitly. As argued before and as it will be proven below, the dynamics inside the single phonemes is very close to low-dimensional deterministic. The application of this raw concept yields already typical gains in the signal-to-noise ratio which are comparable to the most recent results published in the signal processing literature (e.g., [3,4]). The results can be improved by postprocessing the denoised signal with other filters which rely on properties different from those exploited by our method to distinguish between signal and noise (e.g., spectral properties).

To get an impression of how the method works, assume that one has to eliminate noise from music stored on an old-fashioned long playing record, induced by scratches on the black disc. The task becomes almost trivial if one has several samples of this LP. When playing them synchronously, the signal part of the different tracks is identical, whereas the noise part is independent. Already simple averaging will enhance the signal. In deterministic chaotic signals, this redundancy is stored in the past: Since determinism means that similar initial conditions will behave in a similar way (at least for short periods), one solely has to look for almost repetitions of the present signal in the past. Based on this idea, several approaches for noise

 © 2000 The American Physical Society

reduction for deterministic chaotic data have been developed [8]. One of them [9] will be sketched in the following.

Let a dynamical system be given by the map $\mathbf{F}$: $\Gamma \to \Gamma$ in a state space $\Gamma \subset \mathbf{R}^d$. The equation of motion thus reads $\mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n)$. Not knowing $\mathbf{F}$, one can determine it in linear approximation from a long time series $\{\mathbf{x}_k\}$, $k = 1, \ldots, N$, by determining a set of neighboring points $\mathcal{U}_n$ of $\mathbf{x}_n$ and minimizing

$$s_n^2 = \sum_{k:\mathbf{x}_k \in \mathcal{U}_n} (\mathbf{A}_n \mathbf{x}_k + \mathbf{b}_n - \mathbf{x}_{k+1})^2, \qquad (1)$$

the one-step prediction error, with respect to $\mathbf{A}_n$ and $\mathbf{b}_n$ [10,11]. The implicit relation $\mathbf{A}_n \mathbf{x}_k + \mathbf{b}_n - \mathbf{x}_{k+1} = 0$ expresses that data are confined to a hyperplane in the extended phase space. When the signal $\mathbf{x}_k$ is superimposed by random noise, $\mathbf{y}_k = \mathbf{x}_k + \eta_k$, the set $\mathcal{U}_n$ will no longer be embedded in a manifold whose tangent space is the hyperplane defined by $\mathbf{A}_n$ and $\mathbf{b}_n$, but will form a cloud scattered around it. Reducing noise now means to project the noisy $\mathbf{y}_n$ onto this hyperplane. If not $\mathbf{x}_k$ but only a scalar observable s_k is measured, one can reconstruct a phase space by the Takens time delay embedding method [12,13], combining successive elements of the time series $\{s_k\}$ to vectors in $\mathbf{R}^m$, $\mathbf{s}_n = [s_n, s_{n-\tau}, \ldots, s_{n-(m-1)\tau}]$. Here, the embedding dimension m and the lag τ has to be chosen suitably.

The noise reduction scheme outlined above is called local projective noise reduction, as illustrated in Fig. 1, and can be formalized by a minimization problem. The conceptual steps in the numerical algorithm are the following: (i) For every delay vector $\mathbf{s}_n$, all neighbors in the delay embedding space are collected (i.e., $\mathcal{U}_n$ is formed). (ii) The covariance matrix $C_{ij} = \sum_{\mathcal{U}_n} (\hat{\mathbf{s}}_k)_i (\hat{\mathbf{s}}_k)_j$ is computed [$\hat{\mathbf{s}}_k$ means that the mean value (on $\mathcal{U}_n$) has been subtracted], and its singular values are determined. (iii) The vectors corresponding to the largest singular values are supposed to represent the directions spanning the hyperplane defined above by $\mathbf{A}_n$ and $\mathbf{b}_n$. (iv) To reduce noise, $\mathbf{s}_n$ is projected onto these dominant directions. The method is iterated a few times for convergence. The choice of the parameters entering the algorithm (m, τ, diameter of $\mathcal{U}_n$, number of singular vectors to project on) is the crucial problem and has to respect the particular properties of the signal and of the noise.

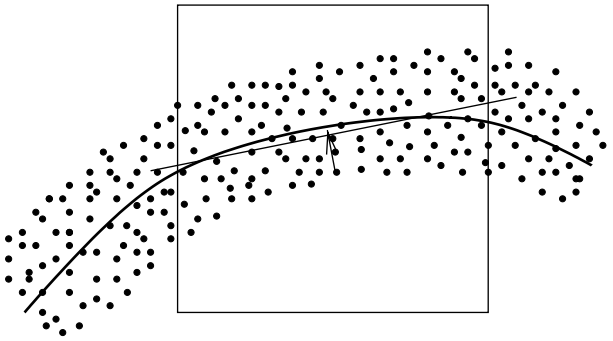

FIG. 1. Schematic representation of the noise reduction method.

We use a recording lasting 3 sec of the vowel "a" for a demonstration of the method for an aperiodic stationary signal. Is is contaminated numerically by 10% white noise, and afterwards filtered by the noise reduction algorithm. The power spectrum before adding noise, after adding noise, and after noise reduction is shown in Fig. 2. Obviously, we were able to restore significant parts of the

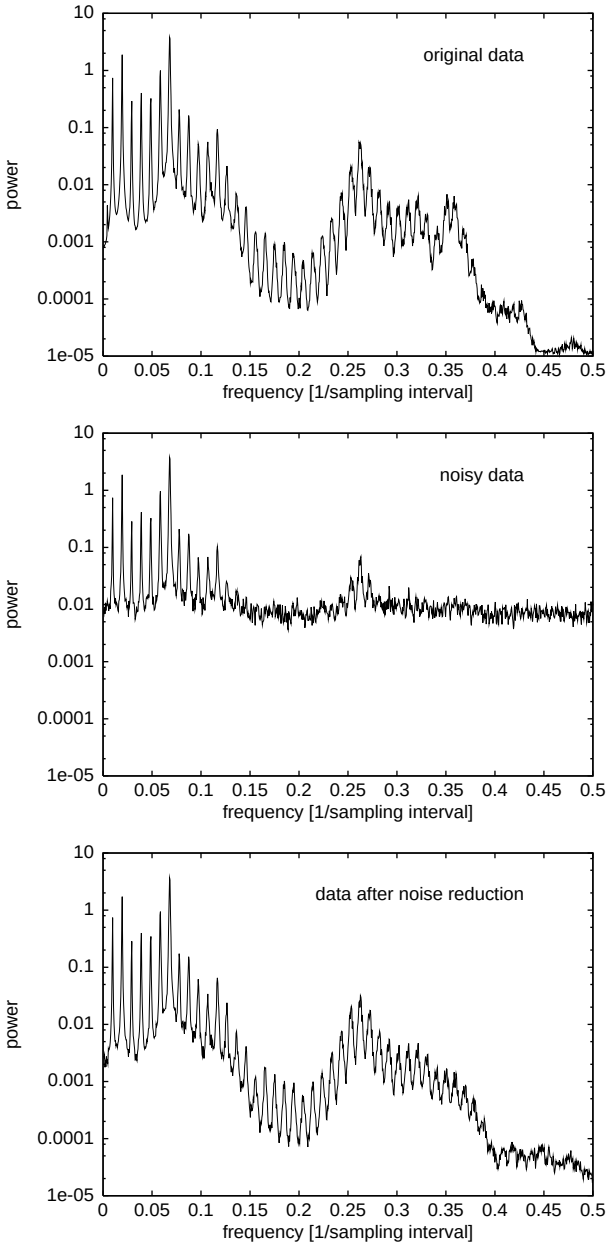

FIG. 2. The power spectrum of 3 sec of the vowel "a": original recording, after adding noise numerically, and after nonlinear noise reduction. Most of the structure of the original spectrum below the noise level could be reconstructed.

spectrum which are well below the noise level and thus invisible for any bandpass filter.

The voice signals studied in the following are taken from a language course on CD ROM with a sampling rate of 22.050 kHz and a precision of 12 bit. The data were converted to real numbers and numerically contaminated by different types and amplitudes of noise and subjected to the noise reduction algorithm. As a measure of performance, we use the gain in dB, given by

$$\text{gain} = 10 \log_{10} \frac{\sum (y_k - s_k)^2}{\sum (\hat{y}_k - s_k)^2}, \qquad (2)$$

where s_k is the clean signal, y_k the noisy signal, and $\hat{y}_k$ the signal after noise reduction. Additionally, we reconvert noisy and denoised signals into the wav-audio format and inspect the results accoustically.

The crucial aspect for the application of the nonlinear noise reduction algorithm is the choice of the correct embedding parameters and the dimension of the linear subspace onto which we project in order to capture the structure we want to preserve. The surprising issue of this paper is that structure in embedding space does in fact exist in human voice signals. Human voice forms an aperiodic and highly nonstationary signal. In Fig. 5 we show the trace of the Italian words "buon giorno." They are composed of subunits, called phonemes, which can be considered as different types of dynamics. Careful investigation of time and length scales, including the use of recurrence plots [14,15] (Fig. 3), shows that the sound wave characterizing a single phoneme (duration between 50 and 150 ms) has a characteristic profile on about 5–10 ms. A kind of phase angle on this highly nontrivial oscillation will then identify the instantaneous amplitude. Thus our time delay embedding space should allow us to identify the actual phoneme and the phase inside the phoneme. Both are achieved by 20- to 30-dimensional delay vectors with a time lag of 3 to 5, covering a time interval of about 8 ms. In this reconstructed state space neighbors represent very similar wave forms and carry thus the redundancy needed to reconstruct the original signal. The neighborhood sizes for the local linear reconstruction of the dynamical constraints are chosen to guarantee about 5–20 neighboring states. Although this number is undesirably small, it cannot easily be increased. A single phoneme does not offer more than ≈20 almost repetitions of a given wave. Searching for neighbors in other words (presumably in identical phonemes) introduces large numerical effort, requires longer sentences, and, most importantly, does not improve the situation much, since changed amplitudes and dilatation or compression of identical phonemes in different words destroy the similarity. Thus all results presented here were gained from intraphoneme neighbors. Figure 3 is a section of a recurrence plot proving these claims.

The projection is done onto subspaces of dimension of 3–7. The success of this strict reduction of dimensionality

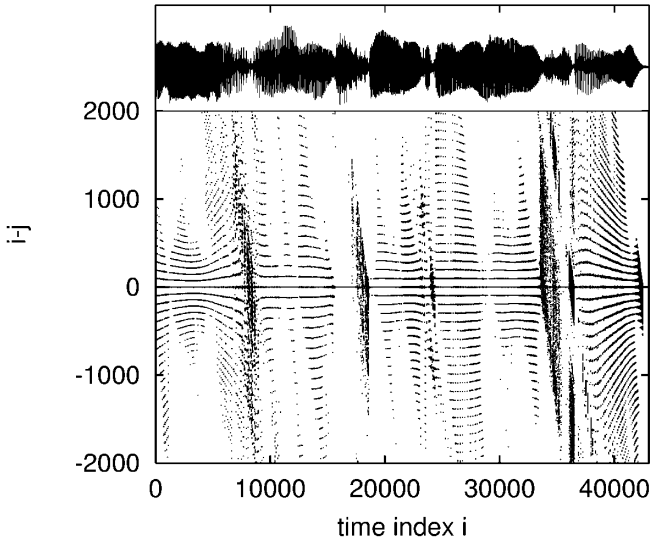

FIG. 3. Main panel: Section of a recurrence plot: In the plane of indices i, j a dot is printed, whenever the delay vectors fulfill $|\mathbf{s}_i - \mathbf{s}_j| < \epsilon$. It proves that our delay vectors really represent meaningful states, where the line structure shows the approximate periodicity inside the phonemes and the number of intraphoneme neighbors. There are almost no dots for $|i - j| > 2000$, reflecting the lack of interphoneme similarities (for this particular ϵ). Upper panel: The speech signal underlying the recurrence plot.

implies that the wave dynamics inside a given phoneme represents only a few degrees of freedom, once it has been identified in the high-dimensional space. This result is in agreement with the study presented in [16] and explains also why as few as 20 neighboring points are sufficient for the algorithm: Only the subspace onto which the projection is done has to be identified, all the remaining directions are irrelevant. Thus only the large singular values and the corresponding singular vectors of the covariance matrix C_{ij} are needed.

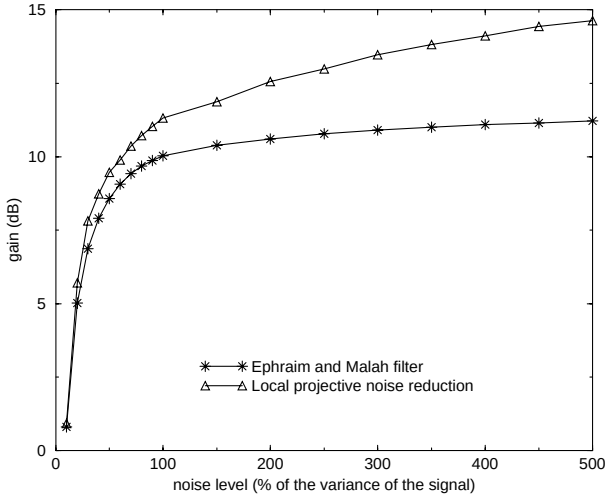

FIG. 4. The gain of the noise reduction scheme as a function of the noise level. A comparison with the Ephraim and Malah filter is shown.

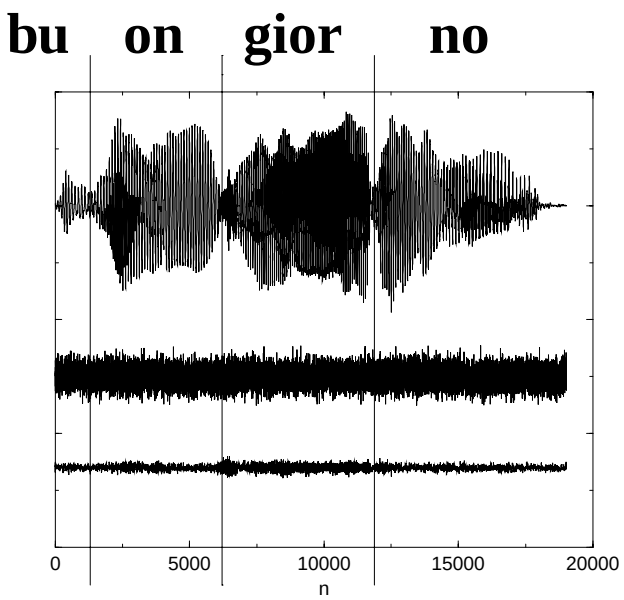

FIG. 5. The clean signal of "buon giorno," noise added to the signal, and the remaining noise after noise reduction (same scale, different offsets). The amplitude of the remaining noise varies systematically, i.e., the success of the noise reduction depends partly on the signal.

The high dimension of the embedding space helps to identify neighbors also for rather high noise levels: Usually, neighborhoods merge if all data are contaminated by large amounts of noise. Here, due to the fact that the signal is rather sparsely filling the 20- to 30-dimensional space, this is not a problem, and we find reasonable results (Fig. 4). For noise levels bigger than 150% the computation of the singular values of the covariance matrix gives unreliable results and the filtering is performed via an averaging between the identified neighbors rather than a projection into the approximated submanifold. However, this may be seen as a degenerate projection onto a zero-dimensional space and is fully contained in the general algorithm. An ideal filter would leave the unperturbed signal unaffected. Because of the violation of the basic assumptions about stationarity and determinism, this is not true for the voice signal, but only below 5% of noise do these distortions become larger than the gain due to noise elimination.

To summarize, nonlinear projective noise reduction is a method which originally was designed to treat deterministic chaotic data with low dimension and constant system parameters. The essential ingredients are the ideas of embedding of scalar data, which in the mathematical framework only makes sense for deterministic stationary signals, and of manifolds representing dynamical constraints. This work demonstrates that the representation in a state space and the confinement to manifolds are approximately present locally for human voice signals, although these signals are nondeterministic and strongly nonstationary. The high embedding dimension here helps to resolve the nonstationarity, since different dynamical regimes (different phonemes) occupy distinct regions in this space and are thus naturally separated. Quantitative comparison with current methods shows the validity of the idea and the power of the proposed filtering scheme. A similar scheme has been used for almost stationary weakly nondeterministic signals with the aim of signal separation as a kind of nonlinear three-way filter [17].

[1] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer, New York, 1996).
[2] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, England, 1997).
[3] T. Gülzow, A. Engelsberg, and U. Heute, Signal Process. **64**, 5–19 (1998).
[4] I. Y. Soon, S. N. Koh, and C. K. Yeo, Speech Commun. **24**, 249–257 (1998).
[5] Y. Ephraim and D. Malah, IEEE Trans. Acoust. Speech Signal Process. **32**, 1109–1121 (1984).
[6] I. Steinecke and H. Herzel, J. Acoust. Soc. Am. **97**, 1874–1884 (1995).
[7] D. A. Berry, I. R. Titze, H. Herzel, and K. Krischer, J. Acoust. Soc. Am. **95**, 3595–3604 (1994).
[8] E. J. Kostelich and T. Schreiber, Phys. Rev. E **48**, 1752 (1993).
[9] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber, Chaos **3**, 127 (1993).
[10] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).
[11] J. P. Crutchfield and B. S. McNamara, Complex Systems **1**, 417 (1987).
[12] F. Takens, Detecting Strange Attractors in Turbulence, Lecture Notes in Mathematics (Springer, New York, 1981), Vol. 898.
[13] T. Sauer, J. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[14] J. P. Eckmann, S. O. Kamphorst, and D. Ruelle, Europhys. Lett. **4**, 973 (1987).
[15] M. Casdagli, Physica (Amsterdam) **D108**, 12–44 (1997).
[16] H. Herzel, D. Berry, I. Titze, and I. Steinecke, Chaos **5**, 30–34 (1995).
[17] T. Schreiber and D. Kaplan, Phys. Rev. E **53**, 4326 (1996).