Nonlinear Analysis of Speech Signals: Generalized Dimensions and Lyapunov Exponents

Vassilis Pitsikalis, Iasonas Kokkinos and Petros Maragos

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.

Email:[vpitsik,jkokkin,maragos]@cs.ntua.gr

Abstract

In this paper, we explore modern methods and algorithms from fractal/chaotic systems theory for modeling speech signals in a multidimensional phase space and extracting characteristic invariant measures like *generalized fractal dimensions* and *Lyapunov exponents*. Such measures can capture valuable information for the characterisation of the multidimensional phase space - which is closer to the true dynamics - since they are sensitive to the frequency with which the attractor visits different regions and the rate of exponential divergence of nearby orbits, respectively. Further we examine the classification capability of related nonlinear features over broad phoneme classes. The results of these preliminary experiments indicate that the information carried by these novel nonlinear feature sets is important and useful.

1. Introduction

There are multiple theoretical and experimental evidences, e.g. in [1, 2, 3], for the existence of rich nonlinear structure in speech signals. Motivated by these evidences, our on-going speech research focuses on the detection of nonlinear phenomena in speech dynamics like *turbulence* and extraction of related information as acoustic signal features for ASR. In a previous work [4], the *short-time fractal dimsension* of speech sounds was measured as a feature to approximately quantify the degree of turbulence in them and used to improve phoneme recognition. Moving a step further, instead of the quantification in the *scalar* phase space, in this paper we extend our work in [5, 6] on using concepts from chaos to model the nonlinear dynamics in speech of the chaotic type and then compute characteristic *invariant* measures. Some previous work in similar directions can be found in [7, 8].

A speech signal segment can be thought of as a 1D projection of a vector function applied to the unknown *multidimensional* speech production system. It is possible that this projection is responsible for a loss of information. By a reverse procedure a multidimensional phase space is reconstructed satisfying the major requirement to be diffeomorphic to the original phase space, so that determinism and differential information of the dynamical system are preserved [9]. According to the *embedding* theorem [10], the reconstructed space can be formed by samples of the original signal delayed by multiples of a constant time delay and defines a motion in a reconstructed multidimensional space that has many common aspects with the original phase space (e.g. fractal dimensions and Lyapunov exponents). Thus, by studying the constructible dynamical system we can uncover useful information about the original unknown dynamical system provided that the unfolding of the dynamics is successful. The parameters that need to be set (i.e. embedding dimension and time delay) can be determined respectively by use of a non-linear correlation measure (i.e. the average mutual information of the signal), and a measure that quantifies how much the manifolds in the phase space are folded, due to projection [10]. In the unfolded phase space one should measure invariant quantities of the attractor that would be conserved from the original phase space.

The analysis via nonlinear models aims to capture these invariant measures of the speech production system's dynamics. In this way, we shall gain insight on the phenomena which take effect by quantifying various characteristics of them. Measures satisfying our requirements are the fractal dimensions and the Lyapunov exponents. The former correspond to the number of active degrees of freedom and the underlying complexity (geometrically and/or probabilistically). Moreover the analysis with generalized fractal dimensions provides a measure which has the potential to detect inhomogeneity of a set, in which case the set is called a multifractal. In this case the description of the set with a class of generalized dimensions is indispensable. On the contrary, if the set is homogeneous then any fractal dimension out of the class of generalized dimensions can work as a representative, but even in this case the knowledge that the fractal dimension remains constant is useful. The Lyapunov exponents quantify the sensitivity to the initial conditions and the rate of exponential divergence of nearby orbits on the attractor. Thus, they are complementary to the dimensions as they carry information closely related to the dynamics of the system.

In this work the effort was placed firstly on the modeling and the analysis of speech signals with methods that reflect the above two directions described above and are computationally efficient. As a preliminary ASR application, we have also experimented on broad class phoneme classification with promising results. Section 2 of this paper summarizes the basic concepts for the analysis with generalized dimensions and a preliminary classification of phonemes. The other invariant measure, i.e. the analysis with Lyapunov exponents, is presented in Section 3 followed by classification experiments.

2. Generalized Dimensions

As an attempt towards a more detailed characterization of complexity and 'strangeness' in phoneme attractors, we have explored the direction of generalized dimensions (e.g. Renyi hierarchy). The description of a phase space via one and only number (e.g. D_1 or D_2), might be too restricting to represent the amount of information possibly residing in it, as far as the underlying probability density distribution is concerned, since it might be more populated in certain regions than others. Although fractal dimensions of the probabilistic type (such as information or correlation dimension) do take under consideration the variable "visitability" of the attractor in different regions, they still are a global weighted average.

A measure among others [11] that can be applied for the extension of the analysis, is the generalized dimension function which defines an infinite class of dimensions, introduced in [12] (where an extensive analysis can be found). In brief, this is accomplished by an analysis of the generic moments of nearest neighbors' distances among randomly chosen points on the attractor. More precisely, for a reference point x in the attractor X and a predefined number of points n, if $\delta(n)$ is its nearest's neighbor distance among the $n-1$ others, and $P(\delta, n)$ is the probability distribution of δ, then the generic moment of order γ of these distances is

$$\langle \delta^\gamma \rangle \equiv M_\gamma(n) = \int_0^\infty \delta^\gamma P(\delta, n) d\delta.$$

Since $\langle \delta^\gamma \rangle$ is argued [12] to depend on n as $\sim n^{-\frac{\gamma}{D(\gamma)}}$, the dimension function is defined as: $D(\gamma) = -\lim_{n \to \infty} \frac{\gamma \ln n}{\ln M_\gamma(n)}$ where γ is the parameter that suppresses or enhances different δ scales of distances. Since for increasing γ the larger distances are more weighted and vice versa, $D(\gamma)$ is a monotonic nondrecreasing function of γ. Among the infinite number of dimensions, one can find the Renyi class of dimensions D_q for $q \geq 0$ with which the corespondance is given by the formula $D[\gamma = (1-q)D_q] = D_q$. Geometricaly the D_q's are the intersection of the graph of the $D = D(\gamma)$ function with a series of straight lines with slope $\frac{1}{1-q}$ (e.g. D_0 is the point that $\gamma = D(\gamma)$; D_1 is the intersection with $\gamma = 0$). If $D(\gamma)$ does not vary for different γ values yields, then the set is homogeneous with constant fractal dimension ($D_0 = \ldots = D_q, q \geq 0$).

The integral equation of $\langle \delta^\gamma \rangle$ can be rewritten as a sum for a discrete signal of finite length N: $M_\gamma(n) = \frac{1}{N} \sum_{i=1}^{N} \delta_i^\gamma(n) P(\delta_i, n)$ where i is an index for the points of the data set. The second term in this sum i.e. the probability density function $P(\delta, n)$ can be computed for an arbitrary scale δ_j as the difference of volume estimates based on the resolution of the successive scales [13]. Let $\{y(k): k = 1, \ldots, M\}$ be a set of uniform random numbers of the same dimensionality as the data set X, and let us define the membership function $f_{\delta_j}(k) = 1$ if $\text{dist}(y(k), X) \leq \delta_j$ and 0 otherwise where $\text{dist}(y(k), X) = \inf_{x \in X} \|y(k) - x\|$. Then the Monte Carlo volume estimate of a δ_j-cover of the set X is: $A(\delta_j) \equiv \frac{1}{M} \sum_{k=1}^{M} f_{\delta_j}(k)$. Given the above, $P(\delta_j, N) \approx A(\delta_j) - A(\delta_{j+1})$ is the probability that some point has a nearest neighbor at distance $\delta \in (\delta_{j+1}, \delta_j]$.

When arbitrary signals are involved, an infinite (or very large for implementation reasons) amount of data is considered to be available (the number of points used in [12] or [13] are of the order of 10^6). Unfortunately this is not the case for speech signals (due to non-stationarity), especially if there has to be some physical interpretation of the state that the speech production system was, while generating a certain phoneme. As far as the direct computation of a generalized dimension by estimation of a multidimensional histogram (as the mass function) is concerned, the more usual limitations are observed due to the insufficient statistics of the data in the multidimensional bins, as they tend to get sparser. The random nature of the approach described above, makes it appealing for this experimental application on speech signals.

Figure 2 shows the dimension functions $D = D(\gamma)$ for different, arbitrary selected, phonemes (from the TIMIT database). It can be clearly seen that in some cases $D(\gamma)$ is varying in the range of γ values that has been computed, which might be different in each case, because of, among others, the phoneme length, speaker and allophone dependency. Such dependence

of the dimension function on γ indicates non-homogeneity of the set. However, there have been observed cases in which the dimension function is not monotonic and/or nondecreasing (see Fig. 2 (a)), or cases that same phonemes uttered either by the same speaker or not, had totally different profile of dimension function. To explore the existence of any classification capability of the measures described above, certain simple characteristic features have been selected such as: the mean value of the generalized dimension function, the coefficients of a 1st or 2nd order polynomial fit to the generalized dimension function. In Fig. 2 the 2D PCA projection of this feature vector is plotted for the broad classes of stops, unvoiced fricatives and vowels or unvoiced fricatives and vowels (in all cases the phonemes are of one speaker whose TIMIT-identity is mentioned). In general we have observed some basic clustering (greater for vowels, less for fricatives, and even less for stops), although there is some diffusion among the classes. Stops tend to mix more than the other classes so they have been omitted in the last two plots. Further in order to quantify our observations, we have used GMM (HTK, 16 mixtures) for, speaker independent, isolated broad class phoneme classification, yielding 83% correct rate for vowels(V), 75% for fricatives(F) and 70% for stops(S) with an overall correct rate of 78% (out of 32616 test phonemes). In the same task the 12 cepstrum coefficients alone (extracted framewise and then mapped by averaging on one feature vector per phoneme, without any deltas, so as to compare approximately under the same terms) scored respectively 67%(V), 48%(F), 88%(S) with overall correct rate of 67%. Concluding, these preliminary experiments are promising because they provide an efficient way to uncover some types of nonlinear information with good potential for categorization of broad phoneme classes.
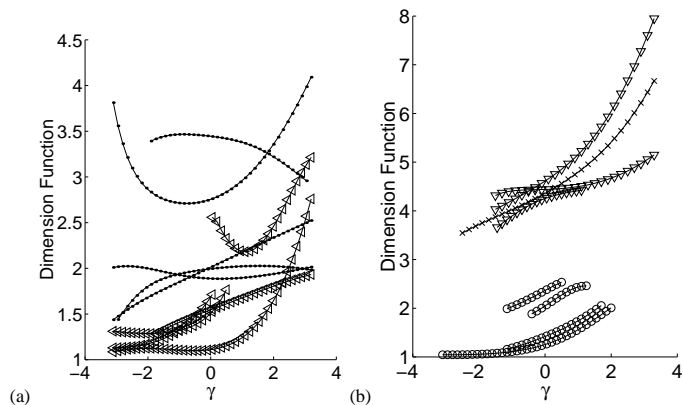

Figure 1: Generalized dimension for:(a) vowels /iy/ (·) and stops /b/ (◁) uttered by the same speaker (mrws1); (b) fricatives /v/ (○), /z/ (▽), and /f/ (×) uttered by mixed speakers.

3. Lyapunov Exponents

Modeling and Prediction on the Reconstructed Attractor: The task of predicting a chaotic signal that has been produced by a system whose dynamics $X_{n+1} = F[X_n]$ are described by a function F, can be formulated as finding a function $\hat{F}$ that approximates F in an optimal way. Only a time series of output observations are given, which can be used to reconstruct the system's attractor, where prediction is done. Numerous techniques have been proposed for the purpose of prediction [14]; these models include approximations based on global or local polyno-
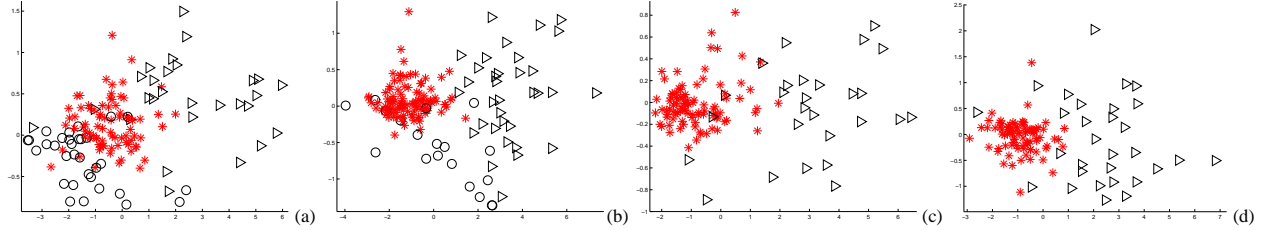
Figure 2: The 2 principal axes after PCA projection of D_γ related features for broad phoneme classes: unvoiced fricatives ($\triangleright$), vowles ($*$) and stops ($\circ$) for: (a) male speaker (mjsw0),(b) female speaker (fdac1). Unvoiced fricatives ($\triangleright$) and vowels ($*$) for: (c) male speaker (mdab0) and (d) male speaker (mrjo0).

mials as well as approximations inspired from machine learning such as radial basis function networks, fuzzy-logic systems and support vector machines. Our focus has been on facilitating the application of the methods of chaotic signal analysis even when a short time series is available (ca. 500 samples), which is our case in speech. Amongst the models that we tested, the one that strikes the best balance between model complexity, training time and ability to capture the system dynamics seems to be the Takagi-Sugeno-Kang (TSK) [15] model with linear local approximations (TSK-1).

TSK models [15] constitute the Fuzzy-Logic approach to function approximation. In brief, the idea is to partition the input space of the function in M fuzzy sets, i.e. sets having no crisp borders, to fit a local model to each subdivision of the input space and to express the approximation of the function as a weighted average of local models: $F(X) = \frac{\sum_{i=1}^{M} \mu_i(X) l_i(X)}{\sum_{i=1}^{M} \mu_i(X)}$ where μ_i measures the degree of membership of X in the $i-th$ fuzzy set and $l_i(X)$ is the local model of the system dynamics for the $i-th$ fuzzy set. If constants are used as local models i.e. $l_i(X) = B_i$, the model is called a TSK-0 model. If we use a linear expression as a local model i.e. $l_i(X) = A_i X + B_i$, the model is called TSK-1.

The number of membership functions and their spreads (if they are Gaussians) have to be determined, as well as A_i, B_i. If the centers and the spreads are known, the optimal A_i and B_i can be calculated by the normal equations and if A_i and B_i are considered stable, the centers and the σ's can be learned using a gradient descent algorithm. We used a variant of the popular AN-FIS system [16] for Fuzzy Logic based function approximation, starting with a relatively large number of membership functions and using the SVD-QR [17] technique for elimination of unnecessary membership functions; for fine-tuning a robust variant of the back-propagation algorithm , namely Resilient Propagation (RPROP) [18], has been used. For details on Fuzzy-Logic based function approximation, we refer the interested reader to the above citations.

Lyapunov exponents (LEs): LEs can be used to characterize a dynamical system, since they are independent of a particular coordinate system and embedding dimension. Divergence of nearby orbits results in a positive LE and convergence of orbits results in a negative LE. For a conservative system the sum of LEs has to be negative, so that the orbits are bounded, while a chaotic system has at least one positive LE. LEs can be calculated as [19]: assume an initial state X_0 which is perturbed by ΔX to a new one X_0'. The values of the their orbits will differ by

$$|X_k' - X_k|^2 = \Delta^T X J^T(X_0) \cdot J^T(X_k) \cdot J(X_k) \cdot J(X_0) \Delta X$$

$k = 1, 2, 3 \ldots, J(X_n)$ is the Jacobian of F at X_n and $|\cdot|$

is the euclidian norm of a vector. We can estimate J by using the predictor which approximates F. The quantity $J(X_k) \cdots J(X_0) J^T(X_0) \cdots J^T(X_k)$ when $k \rightarrow \infty$ converges to the Oseledec matrix OSL of F. The logarithm of the eigenvalues of the Oseledec matrix are equal to the LEs of the system whose dynamics are described by F. Since we usually do not have that long a time-series, we use an approximation of OSL which involves only the first k matrixes, from which we calculate the so called *local* Lyapunov exponents.

A problem that arises when calculating the eigenvalues of the Oseledec matrix is its ill-conditioned nature which causes numerical inaccuracies. The recursive QR decomposition technique has been proposed, which breaks the problem into smaller ones: the matrix OSL can be viewed as the product of $2m$ matrixes, $A_{2m} \cdot A_{2m-1} \cdots A_1$ each of which can be expressed as $A_j Q_{j-1} = Q_j R_j \ \forall j, \ Q_0 = I$ where Q_j, R_j result from the QR-decomposition of A_j. Q is an orthogonal matrix and R is upper diagonal with decreasing diagonal elements. Thus, we can simplify the diagonalization of OSL as follows [19]: $A_{2m} A_{2M-1} \cdots A_1 = Q_{2m} R_{2m} R_{2m-1} \cdots R_1$. Since Q_{2m} is orthogonal the eigenvalues of the last expression shall be equal to the eigenvalues of the product of the $R_{1 \ldots 2m}$ matrices, so their eigenvalues shall equal the elements of their diagonal. Subsequently, the i-th LE can be expressed as $\lambda_i = \sum_{j=1}^{2k} \log(d_{ji})$ where d_{ji} is the i-th element of the diagonal of R_j. Another problem we may encounter is due to the fact that the embedding dimension is not necessarily the intrinsic dimension of the system, but can be a larger one, which guarantees the unfolding of the attractor. As a by-product of the embedding process, more LEs than the true ones are calculated and they are called *spurious* exponents. One can resolve this problem by reversing the order of the data and calculating once more the LEs of the system. The true ones should flip sign, since convergence of nearby orbits now becomes divergence and vice-versa. The spurious ones, however, are an artifact of the embedding process and should stay negative, since they only represent how the rest of the dimensions should collapse to the attractor of the system, independently of the nature of the system. This method was proposed in [19] and works well with clean and long data sets; One of our main criteria for choosing a certain predictor was how well this can be done with short and noisy data sets.

Applications to speech signals: We want to extract some meaningful features (LEs) from the speech signal that could be used for speech analysis. *Vowels* have small positive exponents (usually only one) and 1-2 LEs negative and very close to zero. *Unvoiced fricatives* give no validated exponents; in particular all the direct and inverse exponents are negative (hence no exponents are validated). This is a consequence of the highly noisy nature of unvoiced fricatives. Validated exponents of
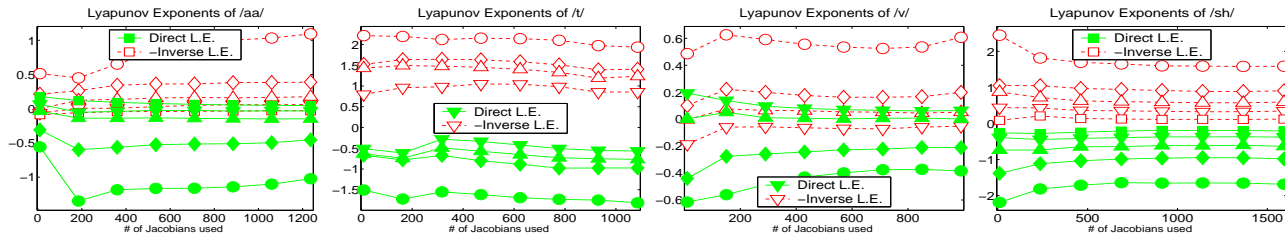
Figure 3: Direct and inverse Lyapunov exponents of a vowel, an unvoiced stop sound, a voiced and an unvoiced fricative.
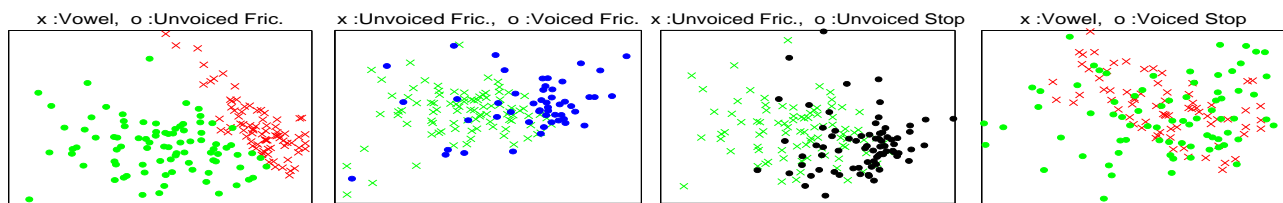

Figure 4: One excellent, one good, one average and one bad class separation using only LEs (The 2 principal components of the L.E.data are plotted).

voiced fricatives are usually higher than those of vowels which is somehow expected, since fricatives are less predictable than vowels. However, it happens usually that a strong noise component in the signal causes none of its exponents to be validated. *Stop sounds* were found to be vaguely separated in two clusters: namely unvoiced/voiced . For the first group it was impossible to find any validated exponents, while the exponents of the second group were validated some of the times. However, really short and non-stationary time series correspond to stop sounds and it is therefore not safe to draw any conclusions about the dynamics of the system based on the LEs; they still can be useful, however for classification. The fact that no LEs are validated may prove to be useful information since this distinguishes stop sounds and unvoiced fricatives from vowels and voiced fricatives. Separation of the phoneme classes is possible in some cases by using the first three LEs of phonemes, as can be seen in Fig.3. The separation is almost perfect for vowels/unvoiced fricatives, but it is not always that successful.

Classification Experiments: In order to somehow quantify the usefulness of LEs we used a directed acyclic graph (DAG) classifier that uses a K-NN classifier to distinguish between every pair of classes, giving a vote to the winning class for every comparison. The class that gets the most votes is considered to be the class that the phoneme belongs to. This architecture allows us to use the LEs only for the comparison between those classes for which they may have a positive contribution e.g. for the comparison between vowels and unvoiced fricatives. Using only the first three LEs computed with the TSK-1 model we achieved a 62% l-o-o (leave one out) correct classification rate for phonemes that have more than 400 samples and 55% when no restriction on the phoneme size was used, which is still much larger than the percentage corresponding to random choice of one class (20%). This testifies that LEs can serve as a useful feature for speech analysis. Using 12 Mel-Frequency Cepstrum Coefficients (MFCC) for the same purpose yielded a 78% correct classification. By combining the MFCC components with the first 3 LEs for the cases they were found to be useful we got an 81% correct classification rate.

4. References

[1] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modelling*,W.J. Hardcastle & Marchal, Eds., NATO ASI Series D, vol. 55,1989.

[2] J. F. Kaiser, "Some Observations on Vocal Tract Operation from a Fluid Flow Point of View", in *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, I. R. Titze and R. C. Scherer (Eds.), Denver Center for Performing Arts, Denver, CO, pp. 358–386, 1983.

[3] T. J. Thomas, "A finite element model of fluid flow in the vocal tract", *Comput. Speech & Language*, 1:131-151, 1986.

[4] P. Maragos and A. Potamianos, "Fractal Dimensions of Speech Sounds: Computation and Application to Automatic Speech Recognition", *J. Acoust. Soc. Amer.*, 105 (3), pp.1925–1932, March 1999.

[5] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models", *Proc. IEEE ICASSP'02*, Orlando, pp. 533–536, 2002.

[6] P. Maragos, A. Dimakis and I. Kokkinos, "Some Advances In Nonlinear Speech Modeling Using Modulations, Fractals, and Chaos" in *Proc. Int'l Conf. on Digital Signal Processing (DSP-2002)*, Greece, July 2002.

[7] M. Banbrook, S. McLaughlin, and I. Mann, "Speech Characterization and Synthesis by Nonlinear Methods", *IEEE Transactions on Speech and Audio Processing*, vol. 7, 1999.

[8] G. Kubin, "Synthesis and Coding of Continuous Speech with the Nonlinear Oscillator Model", *Proc. IEEE ICASSP'96*, pp. 267–270, 1996.

[9] T. Sauer, J.A. Yorke and M. Casdagli, "Embedology", *J. Stat. Physics*, vol.65, Nos. 3/4, 1991.

[10] H. D.I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, 1996.

[11] H.G.E Hentschel and I. Procaccia, "The Infinite Number of Generalized Dimensions of Fractals and Strange Attractors", Physica 8D, pp. 435–444,1983.

[12] R. Badii and A. Politi, "Statistical description of chaotic attractors: the dimension function", *J. Stat. Phys.*, 40, pp.725–750, 1984.

[13] F. Hunt, and F. Sullivan, "Efficient algorithms for computing fractal dimensions", in *Synergetics*, G. Mayer-Kress, editor, Springer Series, 32, Springer-Verlag, New York, pp.74-81, 1986.

[14] M. Casdagli and S. Eubank (eds.), *Nonlinear modeling and forecasting*, Proc. vol. in Santa Fe Inst. Studies in the Sciences of Complexity, 1992.

[15] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control", *IEEE Trans. Systems, Man and Cybenetics*, vol.15, no.1, 1985.

[16] J. Jang "ANFIS:Adaptive Network-Based Fuzzy Inference System", *IEEE Trans. Systems, Man & Cybenetics*, vol. 23, pp. 665–685, 1993.

[17] G. Golub and C. Van Loan, *Matrix Computations*, The John Hopkins University Press, 1989.

[18] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm", in *Proc. IEEE Int'l Conf. on Neural Networks*, San Francisco, CA, pp. 586–591, March, 1993.

[19] J.-P. Eckmann, K. Oliffson, D. Ruelle and S. Ciliberto, "Lyapunov exponents from time series", *Phys. Rev. A*, 34 (6), Dec. 1986.