SOME ADVANCES IN NONLINEAR SPEECH MODELING USING MODULATIONS, FRACTALS, AND CHAOS

Petros Maragos, Alexander G. Dimakis, and Iasonas Kokkinos

Dept. of E.C.E., National Technical University of Athens, Zografou, Athens 15773, Greece.
E-mail: maragos@cs.ntua.gr, adim@softlab.ntua.gr, jkokkin@cs.ntua.gr.

ABSTRACT

In this paper we briefly summarize our on-going work on modeling nonlinear structures in speech signals, caused by modulation and turbulence phenomena, using the theories of modulation, fractals, and chaos as well as suitable nonlinear signal analysis methods. Further, we focus on two advances: i) AM-FM modeling of fricative sounds with random modulation signals of the 1/f-noise type and ii) improved methods for speech analysis and prediction on reconstructed multidimensional attractors.

1. INTRODUCTION

For several decades the traditional approach to speech modeling has been the linear (source-filter) model where the true nonlinear physics of speech production is approximated via the standard assumptions of linear acoustics and 1D plane wave propagation of the sound in the vocal tract. This approximation leads to the well-known linear prediction model for the vocal tract where the speech formant resonances are identified with the poles of the vocal tract transfer function. The linear model has been applied to speech coding, synthesis and recognition with limited success; to build successful applications, deviations from the linear model are often modeled as second-order effects or error terms. However, there is strong theoretical and experimental evidence [1, 2, 3, 4, 5] for the existence of important nonlinear aerodynamic phenomena during the speech production that cannot be accounted for by the linear model. In our work we view the linear model only as a first-order approximation to the true speech acoustics which also contain second-order and nonlinear structure. The investigation of speech nonlinearities can proceed in at least two directions: (i) numerical simulations of the nonlinear differential (Navier-Stokes) equations [6] governing the 3-D dynamics of the speech airflow in the vocal tract, as e.g., in [3, 7], and (ii) development of nonlinear signal processing systems suitable to detect such phenomena and extract related information. In our research we focus on the second approach, which is computationally much simpler, i.e., to develop models and extract related acoustic signal features describing two types of nonlinear phenomena in speech, *modulations* and *turbulence*. Turbulence can be explored both from the geometric aspect, which brings us to *fractals* [16], and from the nonlinear dynamics aspect, which leads us to *chaos* [23, 22].

Thus, in our on-going work we explore models suitable to extract information about the modulation, fractal and chaotic structure of speech signals and use it for applications such as recognition and synthesis. The purpose of this paper is doublefold: First, we briefly summarize the main concepts, models and algorithms that we have used or developed in the three above nonlinear methodologies for speech analysis. Second, we focus on two advances: i) an AM-FM model for fricative sounds using random processes with 1/f spectrum for the instantaneous nonlinear phase fluctuation, and ii) some improved techniques for nonlinear speech analysis and prediction on reconstructed multidimensional attractors.

2. SPEECH MODULATIONS

By ‘speech resonances’ we shall loosely refer to the oscillator systems formed by local vocal tract cavities emphasizing certain frequencies and de-emphasizing others. Although the linear model assumes that each speech resonance signal is a damped cosine with constant frequency within 10-30 ms and exponentially decaying amplitude, there is much experimental and theoretical evidence for the existence of *amplitude modulation (AM) and frequency modulation (FM)* in speech resonance signals, which make the amplitude and frequency of the resonance vary instantaneously within a pitch period. First, due to the *airflow separation* [1, 6], the air jet flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses. This can cause modulations of the air pressure and velocity fields. Also, during speech production *vortices* can easily be generated and propagate along the vocal tract [6, 3], while acting as modulators of the energy of the jet. Motivated by this evidence, in [8, 9] we proposed to *model each speech resonance with an AM-FM signal*

$$x(t) = a(t)\cos[\phi(t)] = a(t)\cos[2\pi \int_0^t f(\tau)d\tau] \qquad (1)$$

and the total speech signal as a superposition of such AM-FM signals, $\sum_k a_k(t)\cos[\phi_k(t)]$, one for each formant. Here $a(t)$ is the instantaneous amplitude signal and $f(t)$ is the instantaneous cyclic frequency representing the time-varying formant signal. The short-time formant frequency average $f_c = (1/T)\int_0^T f(t)dt$, where T is in the order of a pitch period, is viewed as the carrier frequency of the AM-FM signal. The classical linear model of speech

This research work was supported by the Greek Secretariat for Research and Technology and by the European Union under the program ΕΠΕΤ-98 with Grant # 98ΓΤ26. It was also supported by the basic research program ARCHIMEDES of the NTUA Institute of Communication and Computer Systems.

Proceedings of Int’l Conference DSP-2002, July 1-3, 2002, Santorini, Greece.

views a formant frequency as constant, i.e., equal to f_c, over a short time (10-30 ms) frame. However, the AM-FM model can both yield the average f_c and provide additional information about the formant's instantaneous frequency deviation $f(t) - f_c$ and its amplitude intensity $|a(t)|$.

For demodulating a single resonance signal, in [9] we used the nonlinear Teager-Kaiser energy-tracking operator $\Psi[x(t)] \triangleq [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$, where $\dot{x} = dx/dt$, to develop the following nonlinear algorithm

$$\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx 2\pi f(t) \quad , \quad \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)| \qquad (2)$$

This is the *energy separation algorithm (ESA)* and provides AM-FM demodulation by tracking the physical energy implicit in the source producing the observed acoustic resonance signal and separating it into its amplitude and frequency components. It yields very good estimates of the instantaneous frequency signal $f(t) \geq 0$ and of the amplitude envelope $|a(t)|$ of an AM-FM signal, assuming that $a(t), f(t)$ do not vary too fast (small bandwidths) or too greatly compared with the carrier frequency f_c.

There is also a *discrete* version of the ESA, called *DESA* [9], which is obtained by using a discrete energy operator on discrete-time nonstationary sinusoids. The DESA is a novel and very promising approach to demodulating speech resonances for many reasons: (i) It yields very *small errors* for AM-FM demodulation. (ii) It has an extremely *low computational complexity*. (iii) It has an excellent time resolution, almost *instantaneous*; i.e., operates on a 5-sample moving window. Extensive experiments on speech demodulation using the DESA in [9, 12, 13] indicate that these amplitude/frequency modulations *exist* in real speech resonances and are necessary for its *naturalness*. The main disadvantage of the DESA is a moderate sensitivity to noise. This can be reduced by first interpolating the discrete-time signal with *smoothing splines* to create a continuous-time signal, then applying the continuous-time ESA (2), and finally sampling the information-bearing signals to obtain estimates of the instantaneous amplitude and frequency of the original discrete signal $x[n]$. This whole approach is called the *Spline-ESA* and is developed in [10].

The ESAs are efficient demodulation algorithms only when they are used on narrowband AM-FM signals [11]. This constraint makes the use of *filterbanks* inevitable for wideband signals like speech. Thus, each short-time segment (analysis frame) of a speech signal is simultaneously filtered by all the bandpass filters of the filterbank, and then each filter output is demodulated using the ESA. In our on-going research [12, 13, 14, 15] we have been using filterbanks with Gabor bandpass filters whose center frequencies are spaced either linearly or on a mel-frequency scale. See Fig. 1.

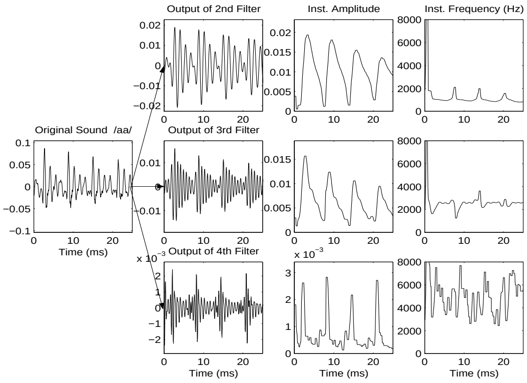


Fig. 1. Demodulating a speech phoneme using a Gabor filterbank and the Spline-ESA. From [15].

Random Modulations and 1/f Noises

While the instant frequency signals produced by demodulating resonances of speech vowels have a quasiperiodic structure, those of fricatives look random. Since fricative and stop sounds contain turbulence, motivated by Kolmogorov's multiscale model of turbulence, we are proposing a random modulation model for resonances of fricatives and stops where the instant phase modulation signal is a random process from the 1/f-noise family. Specifically, we are modeling each such speech resonance $R(t)$ as

$$R(t) = a(t)\cos(2\pi f_c t + p(t)), \quad E[|P(\omega)|^2] \propto \frac{\sigma^2}{|\omega|^\gamma}$$

where $p(t)$ is a random nonlinear phase signal, $P(\omega)$ is its power spectral density (PSD), and E[.] denotes expectation. The PSD, measured either by a sample periodogram $|P(\omega)|^2$ or empirically via filterbanks, is assumed to obey a $1/\omega^\gamma$ power law; such processes are called "1/f noises". A popular fractal model for a subclass of 1/f noises are the fractional Brownian motions (FBMs) [17]. The method we used to solve the inverse problem, i.e. that of extracting the phase modulation $p(t)$ from the speech resonance, is summarized below in four steps: (1) Isolate the resonance by bandpass filtering the speech signal. We used a Gabor filter due to its minimal duration-bandwidth product. (2) Use the ESA to estimate the AM and FM signals, $a(t)$ and $f(t)$. (3) Median filter the FM signal for reducing some extreme spikes as discussed in [9]. (4) Estimate the phase modulation signal $p(t)$ by integrating the instant frequency: $\hat{p}(t) = 2\pi \int_0^t (f(\tau) - f_c)d\tau$, where f_c is the short-time average of $f(t)$.

To test the efficiency of this method we created artificial resonace signals with 1/f phase modulation signal and compared the initial $p(t)$ with its estimate $\hat{p}(t)$ via the above procedure. The original 1/f phase modulation was created by filtering white noise; however, any known method for 1/f noise synthesis can be used. As seen in Figs. 2(d),(e) the reconstructed phase modulation $\hat{p}(t)$ is a low pass version of the original $p(t)$. This is due to the Gabor filtering and the inherent limit to the amount of information that can be carried by phase modulation.

Next we present strong experimental evidence that certain classes of speech signals have resonances that can be effectively modeled as phase modulated 1/f signals. In order to test the validity of the model we demonstrate log-axes plots of the estimated power spectrum of $\hat{p}(t)$ that clearly follow a spectral $1/\text{f}^\gamma$ power law. All the power spectra were estimated by using Welch's averaged modified periodogram method. Another test we used was the variance of the wavelet coefficients. Following [18], if $\psi_n^m(t)$ is an Rth-order regular wavelet basis (R depends on γ), then the process constructed via the expansion $p(t) = \sum_m \sum_n x_n^m \psi_n^m(t)$ is "nearly 1/f" when the wavelet coefficients have variances $var x_n^m = \sigma^2 2^{-\gamma m}$. We have experimentally found that many real speech phase modulation

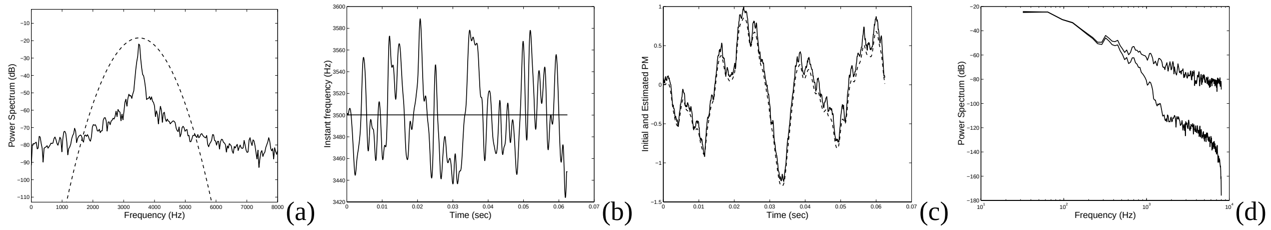


Fig. 2. (a) Artificial Resonance with 1/f phase $p(t)$. b) Instant Frequency. c) $p(t)$ and $\hat{p}(t)$. d) PSD of $p(t)$ and $\hat{p}(t)$.

signals seem to obey this law.

After estimating the phase modulation signal, the problem of estimating its spectral exponent γ naturally follows. Many methods have been proposed for this estimation problem. They include, among others, least squares estimation of the slope of log-axes plots of sample periodograms, methods based on wavelets, and maximum likelihood (ML) schemes. For a detailed review see [19]. The ML estimators are considered the most suitable because they are able to cope with measurement noise. These methods are based on the well-known FBM model [17]. Unfortunately, FBM is not suitable for processes where $\gamma > 3$ because the theory does not directly accommodate such cases. Further, the fact that the signal $\hat{p}(t)$ (from which γ will be estimated) is a *low-pass filtered* 1/f process creates difficulties for any estimator based on an exact 1/f model. In our experimental study we used the simple method of using least-squares estimate on a log-axes plot of Welch's periodogram using only the part of the power spectrum not affected by low pass filtering. The spectral exponents estimated were roughly in the range $\gamma \in (2.5, 4)$ demonstrating that FBM is not suitable to model such high correlated 1/f processes also known as "black noises". The wavelet EM approximation algorithm proposed in [18], an interesting approach not based on FBM, was recently shown in [19] to provide satisfactory estimation only when $0 < \gamma < 1$. Therefore, the only method that worked relatively well was to use a least squares fit on the frequencies of the periodogram not affected by the low pass filtering. This method only provides a rough estimate of γ exponent and is sensitive on measurement noise.

Figure 3 demonstrates the application of the above described 1/f-phase modulation model to an unvoiced and a voiced fricative. We have also performed numerous other similar experiments on real speech signals (from the TIMIT database), by following the same procedure: A strong speech resonance is located, possibly by using the iterative ESA method. Then the ESA is used to extract the phase modulation. (The phase modulations were also estimated via the Hilbert transform to make sure that the ESA does not introduce any artifacts.) The estimated phase is assumed to be a low passed version of a $1/\mathrm{f}^{\gamma}$ random process and the γ exponent is estimated from the slope of the power spectrum (as well as from the variance of wavelet coefficients). In all these experiments our conjecture that the phase modulation of random speech resonances has a $1/\mathrm{f}^{\gamma}$ spectrum has always been verified.

Our on-going work in this area includes better estimation algorithms and a statistical study relating estimated exponents with types of sounds.

3. SPEECH TURBULENCE

Conservation of momentum in the air flow during speech production yields the Navier-Stokes equation [6]

$$\rho(\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \nabla \vec{u}) = -\nabla p + \mu \nabla^2 \vec{u} \qquad (3)$$

where ρ is the air density, p is the air pressure, $\vec{u}$ is the (vector) air particle velocity, and μ is the air viscosity coefficient. It is assumed that flow compressibility is negligible and hence $\nabla \cdot \vec{u} = 0$. An important parameter characterizing the type of flow is the Reynolds number Re=$\rho UL/\mu$, where U is a velocity scale for $\vec{u}$ and L is a typical length scale, e.g., the tract diameter. For the air we have very low μ and hence high Re. This causes the inertia forces to have a much larger order of magnitude than the viscous forces $\mu \nabla^2 \vec{u}$. A *vortex* is a region of similar (or constant) vorticity $\vec{\omega}$, where $\vec{\omega} = \nabla \times \vec{u}$. Vortices in the air flow have been experimentally found above the glottis in [1, 3] and theoretically predicted using simple geometries in [2, 1, 4]. There are several mechanisms for the creation of vortices: 1) velocity gradients in boundary layers, 2) separation of flow, which can easily happen at cavity inlets due to adverse pressure gradients (see [1] for experimental evidence), and 3) curved geometry of tract boundaries, where due to the dominant inertia forces the flow follows the curvature and develops rotational components. After a vortex has been created, it can propagate downstream as governed by the vorticity equation [6]

$$\frac{\partial \vec{\omega}}{\partial t} + \vec{u} \cdot \nabla \vec{\omega} = \vec{\omega} \cdot \nabla \vec{u} + \nu \nabla^2 \vec{\omega} \quad , \quad \nu = \mu/\rho \qquad (4)$$

The term $\vec{\omega} \cdot \nabla \vec{u}$ causes vortex twisting and stretching, whereas $\nu \nabla^2 \vec{\omega}$ produces diffusion of vorticity. As Re increases (e.g., in fricative sounds or during loud speech), all these phenomena may lead to instabilities and eventually result in *turbulent flow*, which is a 'state of continuous instability' [6] characterized by broad-spectrum rapidly-varying (in space and time) velocity and vorticity. Many speech sounds, especially fricatives and stops, contain various amounts of turbulence. In the linear speech model this has been dealt with by having a white noise source exciting the vocal tract filter.

Modern theories that attempt to explain turbulence [6] predict the existence of eddies (vortices with a characteristic size λ) at multiple scales. According to the energy cascade theory, energy produced by eddies with large size λ is transferred hierarchically to the small-size eddies which actually dissipate this energy due to viscosity. A related result is the Kolmogorov law

$$E(k, r) \propto r^{2/3} k^{-5/3} \qquad (5)$$

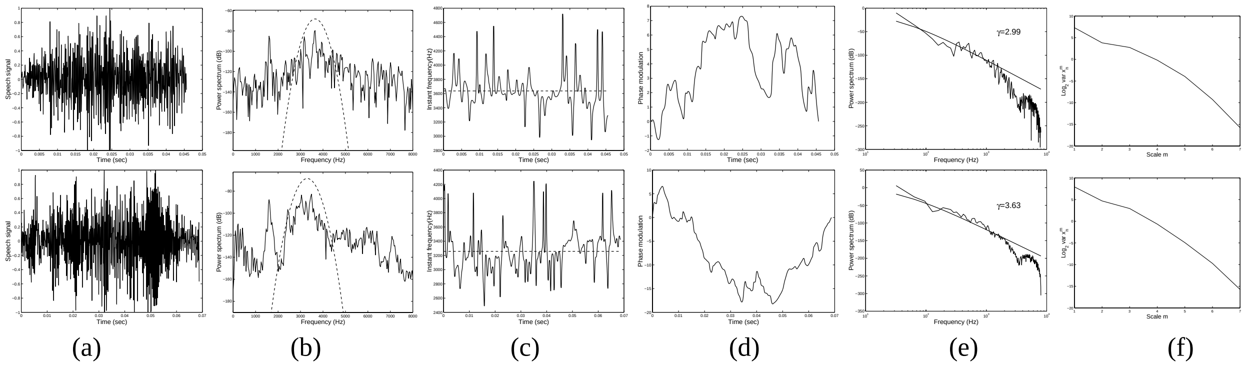


Fig. 3. Experiments with phoneme /s/ (top row) and /z/ (bottom row). Columns: a) Speech signal $s(t)$. b) PSD of $s(t)$ and Gabor filter. c) Instant Frequency. d) Phase modulation $\hat{p}(t)$. e) PSD of $\hat{p}(t)$ and estimated slope. f) Variance of the wavelet coefficients.

where $k = 2\pi/\lambda$ is the wavenumber in a finite nonzero range, r is the energy dissipation rate, and $E(k, r)$ is the velocity wavenumber spectrum, i.e., Fourier transform of spatial correlations. This multiscale structure of turbulence can in some cases be quantified by *fractals*. Mandelbrot [16] and others have conjectured that several geometrical aspects of turbulence (e.g., shapes of turbulent spots, boundaries of some vortex types found in turbulent flows, shape of particle paths) are fractal in nature. We may also attempt to understand aspects of turbulence as cases of chaos. Specifically, chaotic dynamical systems converge to attractors whose sets in phase space or related time-series signals can be modeled by fractals; references can be found in [23]. Now there are several mechanisms in high-Re speech flows that can be viewed as routes to chaos; e.g., vortices twist, stretch, and fold [6, 16]. This process of twisting, stretching, and folding has been to found in low-order nonlinear dynamical systems to give rise to chaos and fractal attractors.

3.1. Speech Analysis using Fractals

Motivated by Mandelbrot's conjecture that fractals can model multiscale structures in turbulence, in [20] we used the *short-time fractal dimension* of speech sounds as a feature to approximately quantify the degree of turbulence in them. Although this may be a somewhat simplistic analogy, we have found in previous work [20, 21] the short-time fractal dimension of speech to be a feature useful for speech sound classification into phonetic classes, segmentation, and recognition. An efficient algorithm developed in [20] to measure it consists of using multiscale morphological filters that create geometrical covers around the graph of the speech signal, whose fractal dimension D can then be found by

$$D = \lim_{s\to 0} \frac{\log[\text{Area of dilated graph by disks of radius } s/s^2]}{\log(1/s)} \tag{6}$$

D is between 1 and 2 for speech signals; the larger D is, the larger the amount of geometrical fragmentation of the signal graph. In practice, real-world signals do not have the same structure over all scales; hence D is computed by least-squares fitting a line to the log-log data of (6) over a small scale window that can move along the s axis and thus create a profile of local *multiscale fractal dimensions* $D(s, t)$ at each time location t of the short speech analysis frame. The function $D(s, t)$ is called a *fractogram*. The fractal dimension at the smallest scale $(s = 1)$ can provide some discrimination among various classes of sounds such as vowels (very low D), unvoiced fricatives (very high D), and voiced fricatives (medium D). At higher scales, the fractogram multiscale fractal dimension profile can also offer additional information that helps in discriminating among speech sounds. See Fig. 4.

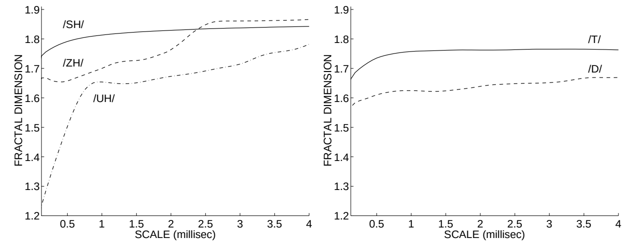


Fig. 4. Multiscale fractal dimension of phonemes /sh/, /zh/, /uh/ and /t/, /d/, averaged over 200 instances [21].

Related to the Kolmogorov 5/3-law (5) is the fact that the variance between particle velocities at two spatial locations X and $X + \Delta X$ varies $\propto |\Delta X|^{2/3}$. By linking this to similar scaling laws in FBMs, it was concluded in [20] that speech turbulence leads to fractal dimension of $D = 5/3$, which was often approximately observed during experiments with fricatives.

3.2. Speech Attractor Analysis using Chaotic Models

Attempting to explore the link between turbulence and chaos, we have used concepts and methods from chaotic systems to model and analyze nonlinear dynamics in speech signals. Most of the techniques we used can be found in [22]. Some preliminary efforts in our work to apply these advanced techniques to speech signals are discussed in [25]. Previous work on using chaotic systems to model speech includes [26, 27, 28].

Embedding and Attractor Reconstruction. We assume that (in discrete time n) the speech production system can be viewed as a nonlinear (but finite dimensional due to dissipativity) dynamical system $Z_{n+1} = F(Z_n)$ where the phase space of Z_n is multidimensional. A speech signal segment $s(n)$, $n = 1, ..., N$, can be considered as a 1D projection of a vector function applied to the unknown dynamic variables Z_n. According to the *embedding* theorem

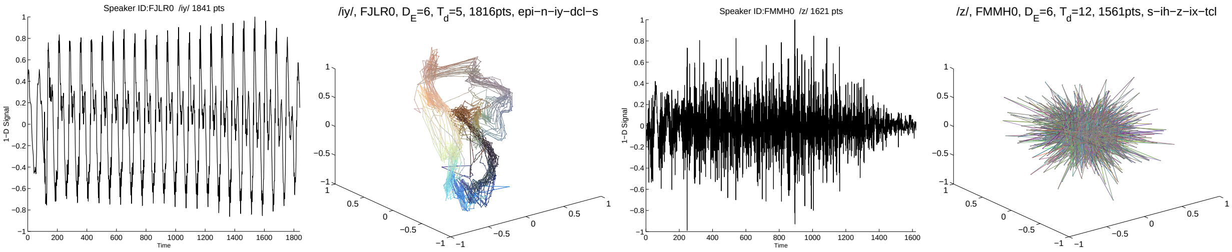


Fig. 5. Waveforms for phonemes /iy/ and /z/ and attractors of embedded signals. From [25].

[22], the vector $X_n = [s(n), s(n+T_d), s(n+2T_d), \ldots, s(n+(D_e-1)T_d]$ formed by samples of the original signal delayed by multiples of a constant time delay T_d defines a motion in a reconstructed D_e-dimensional space that has many common aspects with the original phase space of Z_n. Specifically, many quantities of the original dynamical system (e.g. generalized fractal dimensions and Lyapunov exponents) in the original phase-space Z_n are conserved in the reconstructed space traced by X_n. Thus, by studying the constructible dynamical system $X_n \rightarrow X_{n+1}$ we can uncover useful information about the original unknown dynamical system $Z_n \rightarrow Z_{n+1}$ provided that the unfolding of the dynamics is successful, e.g. the embedding dimension D_e is large enough. However, the embedding theorem does not specify a method to determine the required parameters (T_d, D_e) but only sets constraints on their values. Hence, procedures to estimate good values of these parameters are essential. T_d is related to the correlation or mutual information among speech samples. As in [22, 25] we choose T_d as the first minimum location of a function $I(T)$ equal to the *average mutual information* between speech samples that are T positions apart. Due to the projection, samples of the 1D signal are not necessarily in their relative positions because of the true dynamics of the multidimensional system (true neighbors). As in [22, 25], we find the embedding dimension D_e by increasing its value until the percentage of false neighbors goes to zero (or minimized in the existence of noise). After choosing T_d and D_e, the task of embedding the speech signal in a multidimensional phase space and reconstructing its *attractor* is completed. See Fig. 5.

Dimensions. In the unfolded state-space one can measure invariant quantities of the attractor, which if chaotic would be characterized by sensitive dependence on initial conditions, dense periodic points and mixing [23], and fractal-type dimensions of geometrical (e.g. box-counting dimension) and/or probabilistic (e.g. information dimension) character. The dimension of the attractor except from being a measure of complexity, corresponds to the number of active degrees of freedom of the system. A useful generalized dimension of probabilistic type is the *correlation dimension* [24, 23]

$$D_C = \lim_{r\to 0} \lim_{N\to\infty} \log C(N,r)/\log r$$

where C is the correlation sum, equal for each scale r to the number of point pairs with distances less than r normalized by the number of pairs.

Modeling and Prediction on Reconstructed Attractor

The task of predicting a chaotic signal that has been produced by a system whose dynamics 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 $s(n)$ 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, ranging from local linear models to complex neural networks [29]. Various models (e.g. RBF networks, zeroth and first order TSK models [32], local polynomials) have been tested and found improper when applied to a short data set (ca. 500 samples), which is our case in speech; hence, only those that gave satisfying results will be presented. The performance of predictors has been evaluated on the data they have been trained with, giving a measure of how well a model can learn the data it has been given. However, this may be misleading when we want to extract some useful features about the system dynamics.

Lyapunov Exponent (LEs). A chaotic system is characterized by extreme sensitivity on initial conditions and rapid divergence of nearby orbits. LEs measure the exponential rate of divergence of orbits on phase-space and 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 follows [34]: assume an initial state X_0 which is slightly 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,$A_{2m} \cdot A_{2m-1} \cdot \cdot A_1$ 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 [34]:

$$A_{2m}A_{2M-1}\cdots A_1 = Q_{2m}R_{2m}R_{2m-1}\cdots R_1.$$

Since Q_{2m} is orthogonal the eigenvalues of the last expression shall be equal to the eigenvalues of the product of the $R_{1\cdots 2m}$ matrixes, 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 [34] and works well with clean and long data sets using local polynomials for the identification of the system dynamics. One of our main criteria for choosing a certain predictor was how well this can be done with short and relatively noisy data sets.

Speech Dynamic Models on Attractor. The model that was found to be optimal for the purpose of prediction with a relatively small number of parameters was a rather simple one, which is an extension of the well known linear (AR) model. Instead of assuming that the next value of the state-space vector can be expressed as a linear combination of the previous values of the signal we can use an expression that uses higher order terms, i.e. a *global polynomial* instead of a global linear model. The parameters of a global polynomial that fits the data in an optimal way can be calculated using the family Φ of *orthonormal multivariate polynomials*:

$$\phi_k(X = [x_1, ..., x_{De}]) = \sum_{m=1}^{k} A_m^k \prod_{i=1}^{D_e} x_i^{e_i(m)} \qquad (7)$$

where $\{e_1(m), \cdots, e_{D_e}(m)\} = \mathrm{I}(m)$, and I is a one-to-one correspondence with the property $m_2 > m_1 \Longrightarrow e_i(m_2) \geq e_i(m_1)$, $i = 1 \ldots D_e$. The coefficients A_m^k for the polynomials belonging to family Φ can be derived using a rather sophisticated (but fast) method, so we assume they have already been calculated; for a complete presentation see [30]. We can then express F over the basis Φ as

$$F(X) = \sum_{i=1}^{\infty} C_i \cdot \phi_i(X), \quad F \simeq \hat{F}(X) = \sum_{i=1}^{k} C_i \cdot \phi_i(X), \qquad (8)$$

where $C_i = \sum_{n=1}^{N} F(X_n) \cdot \phi_i(X_n)$. So the approximation $\hat{F}$ is the is the most accurate expansion of F over Φ using only k terms. This model is quite efficient for the purpose of representing a signal where a rather crude approximation of the dynamics is achieved using a very small number of parameters. However, it is rather inadequate when our goal is to capture the dynamics of a system to calculate its LEs. Having tested numerous models, we finally decided to make use of *Support Vector Machines (SVMs)* for regression [33].

SVMs are based on novel ideas from the field of neural networks and have proven to give excellent results when applied to chaotic signals [31]. What distinguishes them from other models is that they aim to minimize the generalization error of the predictor rather than its training error, so a fairly accurate model of the system dynamics that is not biased in favor of the training data can be produced. Training an SVM for regression, whose output is $y = W^T X + b$, can be expressed as a quadratic programming problem:

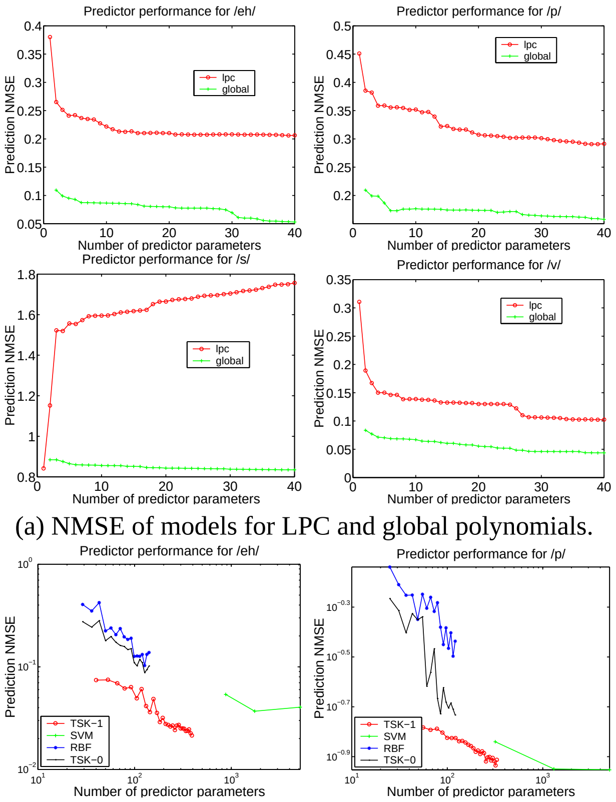
$$\begin{array}{ll} \text{minimize} & \frac{1}{2}W^T W + \kappa \sum_{i=1}^{L}(\xi_i + \xi_i^*) \\ \text{subject to} & \left\{ \begin{array}{ll} y_i - W^T X_i - b & \leq \epsilon + \xi_i \\ W^T X_i + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{array} \right. \end{array}$$

where L is the training set length. The term $W^T W$ penalizes model complexity while the second term tries to minimize prediction error. The first constraint penalizes positive prediction errors e larger than ϵ by $\xi = e - \epsilon$, and the second does the same thing for negative errors i.e. an ϵ-insensitive error function is used which results in a robust to noise and outliers predictor. From the nature of the optimization problem a sparse set of data points can be used to approximate the function F and those X_i that determine the value of W are called *Support Vectors*. The expression of the approximating function is

$$F(X) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) X^T X_i + b$$

where α, α^* are Lagrange multipliers from the dual optimization problem and only dot products are used. Any kernel that satisfies the Mercer conditions [33] can be used instead, such as odd B-splines, the Gaussian kernel and polynomial kernels. SVMs using Gaussian kernels have proven to give the best results amongst other techniques that have been tested on short time-series for the purpose of capturing the system's dynamics and extracting the LEs. One of their most important features is that they make it possible to validate a LE based on the fact that LEs flip sign, even when using only few and relatively noisy data.

Applications to speech signals. Our interest in applying the methods from chaos to the speech signal is twofold: we wish to predict the speech signal using a predictor with a relatively small number of parameters and we want to extract some meaningful features (LEs) from the speech signal that could be used for speech analysis. For every phoneme class the global polynomials model can achieve lower MSE than the LPC model using the same number of parameters, even when only linear terms are included in the expression for the global polynomials (Fig. 6). This verifies that predicting the speech signal on the reconstructed phase-space is more efficient. Specifically, in the context of prediction of chaotic signals, the LPC model assumes $T_d = 1$ for any signal and D_e equal to the number of parameters and tries to fit a global linear model to the data. On the contrary, using global polynomials, a fixed D_e is assumed and T_d is calculated using a principled way, so that any additional complexity of the predictor results in a more accurate reconstruction of the system dynamics rather than just increasing D_e. The results are quite impressive especially for vowels, where the LPC model is supposed to be at its best. Applying some of the other models that we have tested we managed to have a much lower MSE but at the



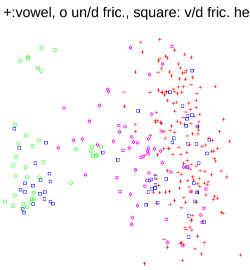
(a) NMSE of models for LPC and global polynomials.

(b) NMSE vs. parameters for TSK-$\{0,1\}$, RBFs, SVMs

Fig. 6. Prediction of speech signals on their attractor.

cost of larger predictors (fig.6). Our model of choice for speech representation is therefore a global polynomial.

We calculated the LEs for different phonemes to see whether some meaningful results can be obtained from their values. After extensive testing, we decided to use an SVM with error tolerance $\epsilon = 0.01$ and Gaussian kernels, whose spread was set to $\sigma = 0.8R$, where R is the divergence of data from the mean of the attractor. In brief, the results obtained are: (1) Vowels have small positive exponents (usually only one) and 2-3 negative. (2) Stop sounds have no validated exponents, i.e. no LEs flip sign when the data are presented with the inverse direction. This is a property characteristic of random or non-stationary signals: the methods applied to chaotic signals then break down and cannot yield meaningful results. (3) For voiced fricatives it is possible to find some validated positive exponents while for unvoiced fricatives the same problem arose as with stop sounds. This is a consequence of the highly noisy nature of unvoiced fricatives that causes the methods of chaotic analysis to break down. The fact that no LEs are validated may still be used as information since this distinguishes stop sounds /unvoiced fricatives from vowels/ voiced fricatives.



A vague separation of the phoneme classes can be accomplished using the first three LEs of phonemes, as can be seen in the right figure. After Principal Components Analysis (PCA) we have projected the data in two dimensions, where the four major classes can be separated, up to a certain degree. Having selected the most robust algorithms for feature extraction our future work shall concentrate on incorporating the new features in the speech recognition process. Some primary results are encouraging, since by enhancing the original feature vector (12 mel-cepstrum coefficients) with the two larger LEs we achieved a 10% decrease, on average, in cross-validation error, when classification in the four main phoneme classes was attempted. We used K-NN classifiers where K ranged from 1 to 50 and the best performing classifier was selected.

4. NONLINEAR FEATURES AND ASR

Although there have been some preliminary efforts to apply fractal and modulation ideas to speech vocoders [35, 13], so far we have mainly applied these models to automatic speech recognition (ASR). In our work [21, 15, 25] we have been developing improved acoustic features for ASR by augmenting the 'standard' feature vector (mel-frequency cepstrum coefficients-MFCC) and its time derivatives with information from the modulation[1] and turbulence structure of speech. Thus, as short-time acoustic representations of speech we use *hybrid feature vectors* that contain information both from the the linear model (smoothed cepstrum) which represents a first-order approximation to the true speech acoustics, as well as from the speech modulations and the chaotic dynamics, which contain information from the second-order non-linear speech acoustics. We have used these hybrid feature vectors as input to hidden Markov model (HMM)–based speech recognizers.

1) In [21] combining the 'standard' (MFCC) with *fractal features* consisting of samples of the multiscale fractal dimension were applied to recognizing the highly confusable e-set (spoken letters: b, c, d, g, p, t, v, z) of ISOLET database and yielded up to 18% reduction in the word error rate over using the 'standard' features alone.

2) In [15] *modulation features* were extracted from each frame as FM percents (bandwidth/mean of instantaneous frequency) at the outputs of a Gabor filterbank. These FM features were used to augment the standard (MFCC) feature vector. The hybrid features showed a significant improvement yielding a word recognition error rate reduction over the TIMIT database that approached 40% for a medium number of mixture components.

3) In [15] after the embedding of speech signal in an unfolded state space, *chaotic features* were computed consisting of the mean and standard deviation of the scale-varying correlation dimension and its integral. Augmenting the standard features (MFCC) with these chaotic features gave better word recognition results over the TIMIT database (error reduction of 29%). Combining the MFCC with both the modulation and the chaotic features further increased the error reduction to 42%.

Some current directions of our on-going research include: experimentation with more sophisticated chaotic and fractal features; better integration of chaotic features with modulation features; apply the nonlinear features for speech recognition in noisy environments and for large vocabulary speech recognition.

[1] Some preliminary work on using Teager energy features (that indirectly contain pre-modulation information) in speaker and speech recognition include [36, 37, 38, 14].

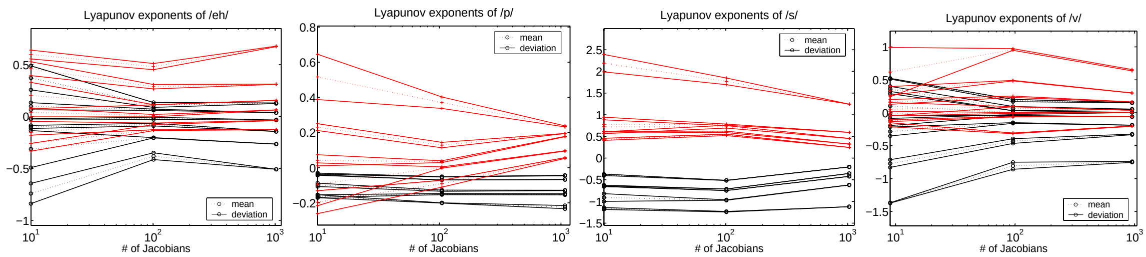


Fig. 7. Direct and inverse Lyapunov exponents of 4 different phoneme classes.

5. 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 and A. Marchal, Eds., NATO Adv. Study Inst. Series D, vol. 55, France, July 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 & R. C. Scherer (Eds.), Denver Center for Performing Arts, Denver, 1983.

[3] T. J. Thomas, “A finite element model of fluid flow in the vocal tract”, *Comput. Speech & Language*, vol. 1, pp. 131–151, 1986.

[4] R. S. McGowan, “An Aeroacoustics Approach to Phonation,” J. Acoust. Soc. Am., 83 (2), pp. 696-704, 1988.

[5] A. Barney, C.H. Shadle and P.O.A.L. Davies, “Fluid Flow in a Dynamical Mechanical Model of the Vocal Folds and Tract: Part 1 & 2”, *J. Acoust. Soc. Amer.*, 105 (1): 444-466, Jan. 1999.

[6] D. J. Tritton, *Physical Fluid Dynamics*, 2nd edition, Oxford Univ. Press, New York, 1988.

[7] G. Richard, D. Sinder, H. Duncan, Q. Lin, J. Flanagan, S. Levinson, M. Krane, S. Slimon, D. Davis, “Numerical Simulation of Fluid Flow in the Vocal Tract”, *Proc. EuroSpeech*, 1995.

[8] P. Maragos, T. F. Quatieri, and J. F. Kaiser, “Speech Nonlinearities, Modulations, and Energy Operators”, in *Proc. ICASSP-91*, Toronto, pp. 421–424, May 1991.

[9] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis”, *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[10] D. Dimitriadis and P. Maragos, “An Improved Energy Demodulation Algorithm Using Splines”, *Proc. ICASSP*, Salt Lake, 2001.

[11] A. C. Bovik, P. Maragos, and T.F. Quatieri, “AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators”, *IEEE Trans. Signal Processing*, vol. 41, Dec. 1993.

[12] A. Potamianos and P. Maragos, “Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation”, *J. Acoust. Soc. Amer.*, 99 (6), pp.3795–3806, June 1996.

[13] A. Potamianos and P. Maragos, “Speech Processing Applications Using an AM–FM Modulation Model”, *Speech Communication*, vol.28, pp.195-209, 1999.

[14] A. Potamianos and P. Maragos, “Time-Frequency Distributions for Automatic Speech Recognition”, *IEEE Trans. Speech and Audio Processing*, vol.9, pp.196-200, Mar. 2001.

[15] D. Dimitriadis, P. Maragos, and A. Potamianos, "Modulation Features for Speech Recognition”, *Proc. ICASSP-2002*, Orlando.

[16] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, NY, 1982.

[17] B.B. Mandelbrot and J.W. Van Ness, “Fractional Brownian motion, fractional noises and applications”, *SIAM Review*, vol.10(4), p.422–437, 1968

[18] G. Wornell, *Signal Processing with Fractals: A Wavelet-Based Approach*, Prentice Hall, 1995.

[19] B. Ninness, “Estimation of 1/f Noise”, *IEEE Trans. Info. Theory*, vol.IT-44, pp.32-46, Jan. 1998.

[20] P. Maragos, “Fractal Aspects of Speech Signals: Dimension and Interpolation”, *Proc. ICASSP–91*, Toronto, May 1991.

[21] 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.

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

[23] H.O. Peitgen, H. Jurgens and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer Verlag, Berlin 1992.

[24] P. Grassberger and I. Procaccia, “Measuring the Strangeness of Strange Attractors”, Physica 9D, pp. 189-208, 1983.

[25] V. Pitsikalis and P. Maragos, "Speech Analysis and Feature Extraction Using Chaotic Models”, *Proc. ICASSP-2002*, Orlando.

[26] T. F. Quatieri and E. M. Hofstetter, “Short-Time Signal Representation by Nonlinear Difference Equations”, *Proc. IEEE ICASSP’90*, Albuquerque, NM, pp. , April 1990.

[27] H. P. Bernhard and G. Kubin, “Speech Production and Chaos”, *XIIth Int. Congress Phonetic Sciences*, Aix-en-Provence, Aug. 1991.

[28] G. Kubin, “Synthesis and Coding of Continuous Speech with the Nonlinear Oscillator Model”, *Proc. ICASSP*, Atlanta, 1996.

[29] M. Casdagli and S. Eubank (eds.), *Nonlinear modeling and forecasting*, Proceedings volume in the Santa Fe Institute Studies in the Sciences of Complexity, 1992.

[30] G. Gouesbet and C. Letellier, “Global vector field reconstruction by using a multivariate polynomial L_2 approximation on nets”, *Physical Review E*, 49, 1994.

[31] S. Munkherjee, E. Osuna and F. Girosi, “Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines”, *Proc. IEEE NNSP’97*, Amelia Island, FL, Sep. 1997.

[32] 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.

[33] V. Vapnik, S.E Golowich and A.Smola, “Support vector method for function approximation, regression estimation, and signal processing.” In *Advances in Neural Information Processing Systems 8*, The MIT press, 1996.

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

[35] P. Maragos and K. L. Young, “Fractal Excitation Signals For CELP Speech Coders”, *Proc. ICASSP*, Albuquerque, NM, April 1990.

[36] T. F. Quatieri, C. R. Jankowski and D. A. Reynolds, “Energy Onset Times for Speaker Identification”, *IEEE Signal Process. Lett.*, vol.1(11), pp.160-162, Nov. 1994.

[37] H.Tolba and D. O’Shaughnessy, “Automatic speech recognition based on cepstral coefficients and a mel-based discrete energy operator”, in *Proc. ICASSP*, Seattle, May 1998.

[38] G. Zhou, J. Hansen and J. F. Kaiser, “Linear and nonlinear speech feature analysis for stress classification”, in *Proc. Int. Conf. Speech & Lang. Process.*, Sydney, Dec. 1998.