SPEECH ANALYSIS AND FEATURE EXTRACTION USING CHAOTIC MODELS

Vassilis Pitsikalis and Petros Maragos

Dept. of Electrical & Computer Engineering, National Technical University of Athens,
Zografou, Athens 15773, Greece. E-mail: [vpitsik,maragos]@cs.ntua.gr

ABSTRACT

Nonlinear systems based on chaos theory can model various aspects of the nonlinear dynamic phenomena occuring during speech production. In this paper, we explore modern methods and algorithms from chaotic systems theory for modeling speech signals in a multidimensional phase space and for extracting nonlinear acoustic features. Further, we integrate these chaotic-type features with the standard linear ones (based on cepstrum) to develop a generalized hybrid set of short-time acoustic features for speech signals and demonstrate its efficacy by showing significant improvements in HMM-based word recognition.

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 are 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 [12, 13]; to build successful applications deviations from the linear model are often modeled as second-order effects or error terms. There is indeed strong theoretical and experimental evidence [15, 5, 19, 17] for the existence of important nonlinear aerodynamic phenomena during the speech production that cannot be accounted for by the linear model. The investigation of speech nonlinearities can proceed in at least two directions: (i) numerical simulations of the nonlinear differential (Navier-Stokes) equations governing the 3D dynamics of the speech airflow in the vocal tract, 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 nonlinear phenomena in speech like *turbulence*.

To be physically meaningful mathematical representations and derived features of speech signals should be derived based on important aspects of the physics of speech production, such as the acoustic dynamics of 3D speech airflow, geometry of vocal tract, and nonstationarity of speech. The nowadays "standard" speech features used in automatic speech recognition (ASR) are based on short-time smoothed cepstra stemming from the linear model. This representation ignores the nonlinear aspects of speech. Adding new robust nonlinear information is quite promising to lead to improved performances and robustness. In this paper, we also develop robust nonlinear features based on chaotic models for speech production and apply these features to increase the recognition performance of ASR systems whose pattern classification part is based on Hidden Markov Models (HMM). Our motivation for this part of our research work includes the following: (1) By using concepts from fractals [7] to quantify the geometrical roughness of speech waveforms, one of the authors was able to extract fractal features from speech signals and use them to improve phonemic recognition [9]. (2) Fractals can quantify the geometry of speech turbulence. A fuller account of the nonlinear dynamics can be obtained by using chaotic models for general time-series as in [1].

Section 2 of this paper summarizes the basic concepts and algorithms for analyzing speech signals with chaotic models. In Section 3 we describe how to extract short-time feature vectors from speech signals that contain chaotic dynamics information, integrate these nonlinear speech features with the standard linear ones (cepstrum), and develop a generalized set of acoustic features for improving HMM-based phonemic recognition.

2. SPEECH ANALYSIS USING CHAOTIC MODELS

It has been shown experimentally and predicted theoretically that many speech sounds contain various amounts of *turbulence* [8]. Specifically, due to *airflow separation* [15, 19], 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. *Vortices* can easily be generated along the vocal tract [19, 17] and then propagate while twisting, stretching and diffusion occurs. Such phenomena are encountered in many speech sounds and lead to turbulent flow; especially fricatives, plosives and vowels uttered with some speaker-dependent aspiration, 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. It has been conjectured that geometrical structures in turbulence can be modeled using fractals [7, 8], while its dynamics can be modeled using the theory of chaos. In a previous work [9], one of the authors measured the *short-time fractal dimsension* of speech sounds as a feature to approximately quantify the degree of turbulence (based on its multiscale structure) in them and used it to improve phoneme recognition. Moving a step further, instead of the above quantification in the scalar phase space, we shall use, in this paper, concepts from chaos [1] to model the nonlinear dynamics in speech of the chaotic type, as an attempt to penetrate into its 'hidden' aspects. Previous work on using chaotic systems to model

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.

speech can be found in [11, 18, 2, 6].

We assume that (in discrete time n) the speech production system can be viewed as a nonlinear but finite dimensional (due to dissipativity [16]) dynamical system $X(n) \rightarrow F[X(n)] = X(n+1)$. A speech signal segment $s(n)$, $n = 1, ..., N$, can be considered as a 1D projection of a vector function applied to the unknown *multidimensional* dynamic variables $X(n)$. It is possible that the complexity or randomness observed in the scalar signal could be due to loss of information during the projection. It is questionable whether there exists a reverse procedure by which a phase space of $Y = Y(n)$ is reconstructed - using information provided by the scalar signal - satisfying the major requirement to be diffeomorphic to the original phase space, so that determinism and differential information of the dynamical system are preserved [14].

According to the *embedding* theorem [1], the vector

$$Y(n) = [s(n), s(n+T_D), \ldots, s(n+(D_E-1)T_D] \quad (1)$$

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 $X(n)$. Particularly, many quantities of the original dynamical system (e.g. generalized fractal dimensions and Lyapunov exponents) in the original phase space $X(n)$ are conserved in the reconstructed space traced by $Y(n)$. The fact that the multidimensional phase space can be fully reconstructed is intuitively justified as there is no disconnected subset of variables of the nonlinear system, nor one can be created by a smooth transformation. Thus, by studying the constructible dynamical system $Y(n) \rightarrow Y(n+1)$ we can uncover useful information about the original unknown dynamical system $X(n) \rightarrow X(n+1)$ 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 For example, D_E must be greater than twice the box-counting dimension of the attractor set and T_D may have any value except from $p\Delta t$, where $p = 1, 2$ and Δt corresponds to periods of possible periodic orbits of the system. Hence, procedures to estimate the values of these parameters are essential.

The time delay corresponds to the constant time difference between the neighboring elements of each reconstructed vector. The smaller T_D gets, the more will the successive elements be correlated, as not enough time will have elapsed for the system to generate sufficient amounts of information and all connected variables affect the observed one. As a consequence the reconstructed vectors will populate along the separatrix of the multidimensional phase space. On the contrary, the greater T_D gets, the more random will the successive elements be and any preexisting 'order' will be lost. Thus it is necessary to compromise between these two conflicting arguments. To achieve this, the following measure of nonlinear correlation introduced by Fraser & Swinney is used for dealing with chaotic data $s(n)$ [1]:

$$I(T) = \sum_{n=1}^{N-T} P(s(n), s(n+T)) \cdot \log_2 \left[\frac{P(s(n), s(n+T))}{P(s(n)) \cdot P(s(n+T))} \right] \quad (2)$$

where $P(\cdot)$ denotes probability. Each log term in the above sum is the mutual information for a pair of observed values $s(n), s(n+T)$ which are apart from each other by a delay T. If these values are independent, their mutual information is zero, as their joint probability factorizes to the product of the two probabilities. Thus, $I(T)$ is the *average mutual information* between pairs of samples of the signal segment that are T positions apart. Then, the 'optimum' time delay T_D is selected as the smallest T at which the average mutual information assumes a minimum value:

$$T_D = \min\{\arg \min_{T \geq 0} \ I(T)\} \quad (3)$$

The next step is to select the dimension D_E of the reconstructed vectors. As a consequence of the projection, points of the 1D signal are not necessarily in their relative positions because of the true dynamics of the multidimensional system (true neighbors); manifolds are folded and different distinct orbits of the dynamics are intersecting. A true vs. false neighbor criterion is formed by comparing the distance between two points S_n, S_j embedded in successive increasing dimensions. If their distance $d_D(S_n, S_j)$ in dimension D is significantly different than their distance $d_{D+1}(S_n, S_j)$ in dimension $D+1$, then they are considered to be a pair of *false neighbors*. Equivalently, if $R^D(S_n, S_j) = \frac{d_{D+1}(S_n, S_j) - d_D(S_n, S_j)}{d_D(S_n, S_j)}$ exceeds a threshold (usually in the range $[10, 15]$), then the two points are false neighbors, under the assumption that any distance difference is not greater than some second order multiple of the attractor diameter $R_A = \frac{1}{N} \sum_{n=1}^{N} \|s(n) - \bar{s}\|$. The dimension D at which the percentage of false neighbors goes to zero (or minimized in the existence of noise) is chosen as the embedding dimension D_E.

In the unfolded phase space one can measure invariant quantities of the attractor, which if chaotic would be characterized [10] by dense periodic points and mixing, such as fractal 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. The *correlation dimension* [4, 10] (belonging to a greater set of generalized dimensions of probabilistic type) is defined as

$$D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C(N, r)}{\log r}, \quad (4)$$

where C is the correlation sum, i.e. for each scale r the number of points with distances less than r normalized to the number of pairs of points:

$$C(N, r) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \theta(r - \|X_i - X_j\|) \quad (5)$$

where θ is the Heavyside unit-step function. For small 'enough' scales and for N large 'enough' $C(r)$ is proportional to $\chi(r) r^{D_C}$, where $\chi(r)$ stands for the lacunarity of the set [7].

Figure 1 shows the waveforms of four speech phonemes, their attractors and local-scale correlation dimension measurements. The shape[1] differences or similarities (complex rough spikes for fricatives, smooth flow/cycles for vowels) in the attractors are consistent with the corresponding physics for each phoneme.

3. CHAOTIC FEATURES AND SPEECH RECOGNITION

The analysis described in Section 2 has been applied to a large number of phonemes. Experimental observations of the dynamics in

[1] The visualization of the multidimensional attractors has been done by showing the first three elements of each vector in 3D space and the last three as RGB color components.

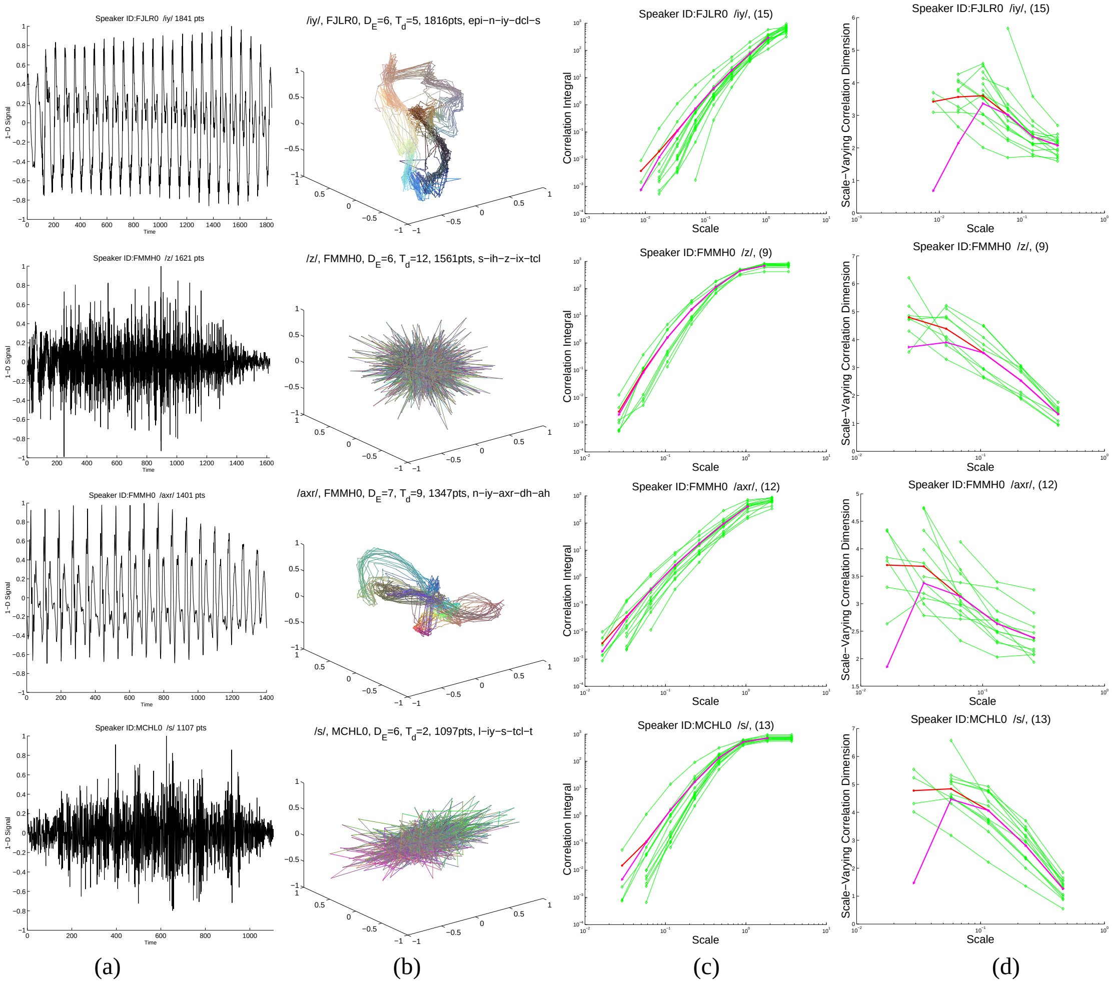

Fig. 1. (a) Speech Waveforms, (b) Attractors of Embedded Signals, (c) Correlation Sums, (d) Scale-Varying Correlation Dimensions. 1st row (top): vowel /iy/, 2nd row: voiced fricative /z/, 3rd row: vowel /axr/, 4th row (bottom): unvoiced fricative /s/ . (In (c) and (d) thick lines show average curves.)

the reconstructed phase space have shown the formation of general patterns among phonemes of the same type, both from a qualitative and a quantitative point of view (i.e., the attractors' topology and the scale-varying correlation dimensions, respectively). Less well-formed patterns were observed in the case of phonemes of the same class (e.g. fricatives, plosives, vowels). Further, even for the same phoneme uttered by the same speaker, there were some cases of variabilities depending on neighboring phonemes (allophones). Motivated by similar classifications of fractal speech characteristics in a previous work [9], we attempted to extract *features* related to chaotic dynamics and apply them to an automatic speech recognition (ASR) system based on hidden Markov models (HMM)[2].

The feature vectors used in speech recognition are typically computed over a 20-30 ms window and are updated every 5-10 ms. The 'standard' feature set consists of the mean square amplitude (usually called 'energy') the first twelve *mel-frequency cepstrum coefficients (MFCC)* and their first and second time derivatives. We shall augment the 'standard' feature vector and thus create a *hybrid feature vector* by incorporating information from the nonlinear structure of speech of the chaotic type as additional features. Thus, as short-time acoustic representations of speech we use feature vectors that contain information both from the smoothed cepstrum of the linear model, which represents a first-order approximation to the true speech acoustics, as well as from the chaotic dynamics, which contain information from the second-order nonlinear speech acoustics. The input feature vectors are split into two different data streams (MFCC and chaotic) belonging to independent probability 'streams' with independent probability distributions. The TIMIT[3] database was used for the recognition experiments.

Through an automated procedure, each speech analysis frame (25-ms frames, updated every 10 ms) has been embedded in a multidimensional phase space using the appropriate parameters (T_D, D_E). The physical justification of embedding only a frame instead of a whole phoneme is that the reconstructed space in this

[2] The HTK [20] HMM-recognition system was used.

[3] The TIMIT database consists of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of US. All speech signals in TIMIT are sampled at 16 kHz. The training set consists of 3696 sentences and the test set of 1344 sentences.

occasion belongs to the short-time phase space of the dynamic system during the time period it produced the current frame. Next, we computed a feature vector that was related to the correlation sum and the scale-varying correlation dimension and hence carried information about the chaotic dynamics of each frame. Specifically, we selected a set of four chaotic features: (1) the mean of the correlation sum C, (2) the standard deviation of C, (3) the mean of the scale-varying correlation dimension D_C, and (4) the standard deviation of D_C. This feature set also included the first and second time derivatives of these four features.

The recognition results (see Table 1) of the hybrid feature set were quite promising, even though our preliminary first application of chaotic features used the *fewest* and simplest possible such features. The relative word error rate reduction of 18% and 29% (with 8 and 16 mixtures respectively) over using only the standard features is possibly due to the detection of nonlinear phenomena which remain “hidden” in the 1D dynamics. Unfolding the signal to the original phase space enables the observation of the true dynamics of the system; furthermore a broad variety of new measurements can be performed on the unfolded attractor that can yield fractal and/or chaotic features adding considerable information even in a four-component feature vector.

	Word Percent Correct	
# Gaussian Mixtures	MFCC	MFCC+Chaotic
8	73.95	78.61
16	78.76	85.01

Table 1. Recognition Results

In [3] we have also used this chaotic feature vector in combination with other nonlinear features of the modulation type. This yielded a relative error rate reduction by 42%, which outperformed experiments in which only one type of feature set was used.

4. CONCLUSIONS

In this paper we have described how to apply modern concepts and algorithms from chaotic systems to analyzing speech signals in order to create a multidimensional model that exploits nonlinear dynamic information and extract related novel acoustic features of chaotic type. Further we have developed a hybrid feature set for speech recognition that includes both the standard linear features as well as the chaotic features and applied this new feature set to HMM-based word recognition. Our experimental results, have shown a significant improvement in recognition over the TIMIT database. Clearly, information provided by the new (nonlinear) features deals with different aspects of the speech dynamics and therefore is valuable for the recognition process.

In our on-going speech research, we are also working to enhance the nonlinear speech analysis described herein in various directions such as: exploring more sophisticated chaotic features, such as generalized dimensions and Lyapunov exponents which also contain dynamical information; extracting chaotic features in noisy environments; integration of chaotic features with other nonlinear features; application of chaotic features to large vocabulary speech recognition problems. Further results will be presented in a forthcoming paper.

5. REFERENCES

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

[2] H. P. Bernhard and G. Kubin, “Speech Production and Chaos”, *XIIth Intern. Congress of Phonetic Sciences*, Aix-en-Provence, August 1991.

[3] D. Dimitriadis, P. Maragos, V. Pitsikalis and A. Potamianos, "Modulation and Chaotic Acoustic Features for Speech Recognition”, *J. Control and Intelligent Systems*, 2002.

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

[5] 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.

[6] G. Kubin, “Synthesis and Coding of Continuous Speech with the Nonlinear Oscillator Model”, *Proc. IEEE ICASSP’96*, pp. 267–270, 1996.

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

[8] P. Maragos, “Fractal Aspects of Speech Signals: Dimension and Interpolation”, *Proc. IEEE ICASSP’91*, Toronto, pp. 417-420, May 1991.

[9] 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.

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

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

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[13] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

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

[15] 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 ASI Series D, vol. 55, 1989.

[16] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, Applied Mathematical Sciences, vol.68, 1993.

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

[18] N. Tishby, *Proc. IEEE ICASSP’90*, pp. 365–368, 1990.

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

[20] S. Young, *The HTK Book*, Cambridge Research Lab: Entropics, Cambridge, England, 1995.