ISCA Archive
http://www.isca-speech.org/archive

ITRW on Non-Linear Speech
Processing (NOLISP 03)
Le Croisic, France
May 20-23, 2003

Some Advances on Speech Analysis using Generalized Dimensions

Vassilis Pitsikalis and Petros Maragos

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.
Email:[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 extracting characteristic invariant measures such as the generalized fractal dimensions. Such measures can capture valuable information for the characterisation of the multidimensional phase space since they are sensitive on the frequency that the attractor visits its different regions. Further, we integrate some of 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 slight improvements in HMM-based phoneme recognition without the use of any language model.

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. There is indeed strong theoretical and experimental evidence [22, 10, 25, 23] for the existence of important nonlinear aerodynamic phenomena during the speech production that cannot be accounted for by the linear model. Thus, in our research we focus on the development of nonlinear signal processing systems suitable to detect such phenomena and extract related information as acoustic signal features describing these nonlinear phenomena in speech like *turbulence*.

To be physically meaningful mathematical representations and extracted 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. However, linear acoustics has been used for the past 50 years to model speech signals both during generation by the human vocal tract as well as during perception by the human auditory system, ignoring the true nonlinear physics of speech production and hearing. The analysis via nonlinear models aims to capture the invariant measures of the speech production system's dynamics. By this way, we shall gain insight on the phenomena which take effect by quantifying various characteristics of them. The analysis with generalized fractal dimensions [6, 12, 16] provides a measure which has the potential to detect inhomogeneity or nonuniformity 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 indespensable. On the contrary, if the set is uniform then any fractal dimension out of the class of generalized dimensions can work as a representative.

The nowadays "standard" speech features used in automatic speech recognition (ASR) are based on short-time smoothed cepstra stemming from the linear model [19, 20]. 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 a simple set of 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).

Section 2 of this paper summarizes the basic concepts for analyzing speech signals with chaotic models ([17]), extracting short-time feature vectors that contain relative information and after integrating them with the standard linear ones (cepstrum), develop a generalized set of acoustic features for improving HMM-based phonemic recognition. Section 3 presents the analysis using concepts as the generalized dimensions and a preliminary application to speech phonemes.

2. Speech Analysis using Chaotic Models and ASR Aplication

It has been shown experimentally and predicted theoretically that many speech sounds contain various amounts of *turbulence* [14]. Several phenomena (airflow separation[22, 25], instability, generation of vortices[25, 23]) encountered in many speech sounds lead to turbulent flow; especially for fricatives, plosives and vowels uttered with some speaker-dependent aspiration. It has been conjectured that geometrical structures in turbulence can be modeled using fractals [13, 14], while its dynamics can be modeled using the theory of chaos. In a previous work [15], one of the authors measured the *short-time fractal dimsension* of speech sounds as a feature to approximately quantify the degree of turbulence in them and used it to improve phoneme recognition. Moving a step further, instead of the quantification in the *scalar* phase space, we extend our work [17] by using, concepts from chaos [1] to model the nonlinear dynamics in speech of the chaotic type and then compute characteristic invariant measures. Previous work can be found in [18, 24, 5, 11].

A speech signal segment can be thought of as a 1D projection of a vector function applied to the unknown *multi-dimensional* 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 - 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 [21]. According to the *embedding* theorem [1], the reconstructed space is 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 are the embedding dimension and the time delay; they are determined respectively by use of a nonlinear 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. In the unfolded phase space one can measure invariant quantities of the attractor like fractal dimensions (corresponding to the number of active degrees of freedom and the underlying complexity) and Lyapunov exponents. The scale varying *correlation dimension* [7, 16], as an easy to compute measure, has been evaluated [17].

We then attempted to extract *features* related to chaotic dynamics and apply them to an automatic speech recognition (ASR) system based on hidden Markov models (HMM) (The HTK [26] HMM-recognition system was used) . It is noted that no language model was used, so as to explore only phonetic contribution of the features, without any statistical information about the language. The analysis described above has been applied to each speech analysis frame (30-ms frames, updated every 10 ms). The physical justification of embedding only a frame instead of a whole phoneme is that the reconstructed space in this occasion belongs to the short-time phase space of the dynamic system during the time period it produced the current frame. We compute a feature vector that is related to the scale-varying correlation dimension and hence carried information about the chaotic dynamics of each frame. Specifically, the following set of elements has been selected: (1) the mean of scale-varying correlation dimension at smaller scales, (2) its standard deviation, (3) the mean of scale-varying correlation dimension at larger scales, and (4) its standard deviation. The feature set is extended with the first and second time derivatives and then merged with the standard MFCC's in 2 different probability streams for speech recognition over the TIMIT database.

The recognition result of the hybrid feature set (accuracy for hybrid feature set 53.91% , Baseline accuracy 53.41%) were slightly promising, even though our preliminary first application of chaotic features used the *fewest* and simplest possible such features. The relative pnone error rate reduction of 1.07% (with 16 mixtures) 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 will be performed on the unfolded attractor that can yield fractal and/or chaotic features adding considerable information even in a four-component feature vector.

3. Speech Analysis using Generalized Dimensions

As an attempt towards the *finer* characterization (complexity and strangeness) of phoneme attractors, the direction of generalized dimensions ([16, 8]) (eg Renyi hierarchi) was followen. The representation and description of a phase space via one and only number (e.g. D_1, D_2), might be too restricting to stand alone for the ammount 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 ([16]) that can be applied for the extension of the analysis, is the generalized dimension function which defines an infinite class of dimensions, introduced in [2, 3] (where an extensive analysis can be found). In brief, this is accomplished by an analysis of the generic momments 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 ammong the $n-1$ others, and $P(\delta, n)$ is the probability distribution of δ, then the generic momment of order γ of these distances is

$$\langle \delta^\gamma \rangle \equiv M_\gamma(n) = \int_0^\infty \delta^\gamma P(\delta, n) d\delta. \qquad (1)$$

Since $\langle\delta^{\gamma}\rangle$ is argued [2] to depend on n as $\sim n^{-\frac{\gamma}{D(\gamma)}}$ the dimension function is defined as:

$$D(\gamma) = -\lim_{n\to\infty}\frac{\gamma\ln n}{\ln M_{\gamma}(n)} \qquad (2)$$

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 $\gamma = D(\gamma)$, and D_1 is the intersection with $\gamma = 0$) . The case that $D(\gamma)$ does not vary for different γ values yields that the set is uniform with constant fractal dimension ($D_0 = D_1 = D_2 = \ldots = D_q, q \geq 0$).

The integral Equation 1 can be rewriten 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 index i stands for all the reference points of the 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 [9]. Let $Y = \{y(k) : k = 1, \ldots, M\}$ be a set of uniform random numbers of the same dimensionality as the data set X, and $f_{\delta_j}(k) = 1$ if $\text{dist}(y(k), X) \leq \delta_j$ and 0 otherwise; given that $\text{dist}(y(k), X) = \inf\|y(k) - x\|$ for $x \in X$. Then the Monte Carlo volume estimate of a δ_j-cover of the set X is: $A(\delta_j) \equiv \frac{1}{M}\sum_{i=1}^{M} f_{\delta_j}(k)$. 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 implimentation reasons) ammount of data is considered to be available (the number of points used in [3] or [9] are of the order of $10^5, 10^6$). 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. The more usual limitations are observed in the computation of higher dimensional histograms due to the insufficient statistics of the data in the multi-dimensional bins. The random nature of the approach described above, makes it appealing for an experimental application on speech signals.

In Figures 3(a,b) the dimension functions $D = D(\gamma)$ are presented for different, – arbitrary selected – phonemes (extracted 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 dependency and allophone dependency. Such dependence of the dimension function on γ indicates non uniformity of the set. Though, there have been observed cases in which the dimension function is not monotonic and/or nondecreasing (see Figure 3(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. Further in order to quantify our observations, we have used Gaussian Mixture Models (HTK toolkit,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.

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 measure invariant quantities like generalized fractal dimensions i.e. an infinite number of dimensions, which carry much more information compared to a fractal dimension that consists of a single number. In our on-going speech research, we are working to enhance the nonlinear speech analysis described herein, in various directions such as: exploring different ways for computing the generalized dimensions and apply them to a broad class of phonemes; computation of Lyapunov exponents which also contain dynamical information; analysis of features as far as their dependencies upon various attributes is concerned; extracting chaotic features in noisy environments; 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] R. Badii and A. Politi, “Hausdorff Dimension and Uniformity Factor of Strange Attractors”, *J. Stat. Phys.*, Phys. Rev. Lett. 52, pp.1661–1664, 1984.

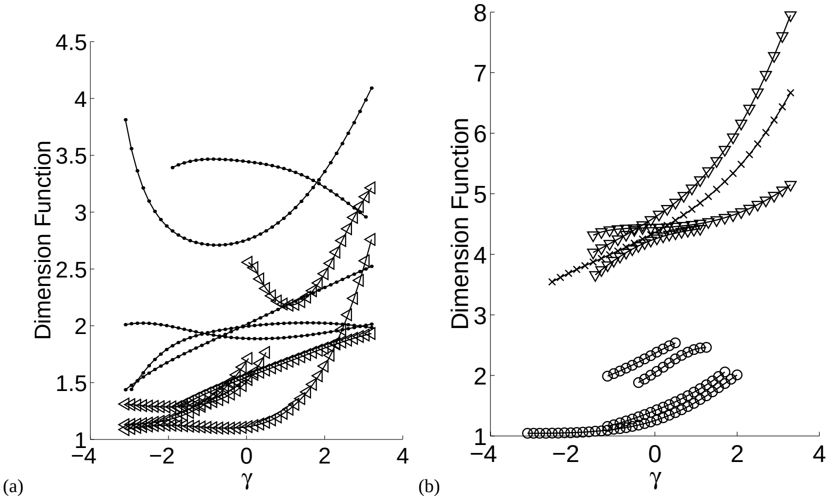

Figure 1: Generalized dimension for:(a) vowels /iy/ (·) and stops /b/ (◁) uttered by the same speaker (mrws1); (b) fricatives /v/ (○), /z/ (▽), and /f/ (×) uttered by mixed speakers.

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

[4] R. Badii and A. Politi, "Intrinsic oscillations in measuring the fractal dimension", *Phys. Lett.*, 104A, pp.303–305, 1984.

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

[6] K. Falconer, "Fractal Geometry: Mathematical Foundations and Applications", John Wiley and Sons, New York, 1990.

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

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

[9] 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.

[10] 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.

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

[12] J. L. McCauley, "Chaos, Dynamics and Fractals", Cambridge University Press, 1993.

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

[14] P. Maragos, "Fractal Aspects of Speech Signals: Dimension and Interpolation", *Proc. IEEE ICASSP'91*, Toronto, pp.417-420, May 1991.

[15] 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.

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

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

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

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

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

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

[22] 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.

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

[24] N. Tishby, "A dynamical systems approach to speech processing", *Proc. IEEE ICASSP'90*, pp. 365–368, 1990.

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

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