Nonlinear Dynamical Invariants for Speech Recognition

S. Prasad, S. Srinivasan, M. Pannuri, G. Lazarou and J. Picone.

Center for Advanced Vehicular Systems
Mississippi State University, MS, USA
{prasad, srinivas, pannuri, glaz, picone}@cavs.msstate.edu

Abstract

There is growing interest in modeling nonlinear behavior in the speech signal, particularly for applications such as speech recognition. Conventional tools for analyzing speech data use information from the power spectral density of the time series, and hence are restricted to the first two moments of the data. These moments do not provide a sufficient representation of a signal with strong nonlinear properties. In this paper, we investigate the use of features, known as invariants, that measure the nonlinearity in a signal. We analyze three popular measures: Lyapunov exponents, Kolmogorov entropy and correlation dimension. These measures quantify the presence (and extent) of chaos in the underlying system that generated the observable. We show that these invariants can discriminate between broad phonetic classes on a simple database consisting of sustained vowels using the Kullback-Leibler divergence measure. These features show promise in improving the robustness of speech recognition systems in noisy environments.

1. Introduction

Speech recognition systems today still exploit the linear acoustics model of speech production, and rely on traditional measures of the spectrum based on Fourier transforms. Though applications of machine learning to speech recognition have made great strides in recent years, high performance speech recognition systems are still sensitive to mismatches in training and evaluation conditions, or dramatic changes in the acoustic environments in which they operate. We refer to this as the robustness problem – can a speech recognition system achieve high performance on noisy data that has not been observed during training? Our goal in this work is to produce new features for speech recognition that do not rely on traditional measures of the first and second order moments of the signal.

Dynamical systems can be represented by state-space models, where the states of the system evolve in accordance with a deterministic evolution function, and the measurement function maps the states to the observables. The path traced by the system's states as they evolve over time is referred to as a *trajectory*. An *attractor* is defined as the set of points in the state space that are accumulated in the limit as $t \to \infty$. *Invariants* of a system's attractor are measures that quantify the topological or geometrical properties of the attractor, and are invariant under smooth transformations of the space. These smooth transformations include coordinate transformations such as Phase Space Reconstruction of the observed time series [1].

These invariants are a natural choice for characterizing the system that generated the observable. These measures have been previously studied in the context of analysis and synthesis research [1][2] and more recently in the context of speech recognition [3]. In this paper, we review algorithms to extract these invariants using a pilot database consisting of elongated pronunciations of a small set of phones, and study discriminability in a feature space comprised of these invariants.

Lyapunov exponents [4] associated with a trajectory provide a measure of the average rates of convergence and divergence of nearby trajectories. Fractal dimension [5] is a measure that quantifies the number of degrees of freedom and the extent of self-similarity in the attractor's structure. Kolmogorov entropy [5] defined over a state-space, measures the rate of information loss or gain over the trajectory. These measures search for a signature of chaos in the observed time series. Since these measures quantify the structure of the underlying nonlinear dynamical system, they are prime candidates for feature extraction of a signal with strong nonlinearities.

Our long-term goal in this research is to model phones using dynamical systems, where the state-space that generated the observed acoustic sound corresponds to a unique configuration of the articulators and the driving process. Since each configuration corresponds to a unique attractor in the phase space, it is expected that the invariants extracted from different phones will mirror differences in the corresponding attractors.

Lyapunov exponents [3] have been employed as features in a phonetic recognition system, and studied in combination with conventional cepstral features. In this paper, we extend the analysis to three standard invariants of a dynamical system. The motivation behind studying such invariants from a signal processing perspective is to capture the relevant nonlinear dynamical information from the time series – something that is ignored in conventional spectral-based analysis.

The outline of this paper is as follows. In Section 2 we review phase-space reconstruction techniques, which are the starting point for most nonlinear dynamical system analysis. We provide a brief review of the algorithms we employed for the extraction of three dynamical invariants from a time series. In Section 3, we describe the experimental setup. We also explain the choice of various parameters involved in the estimation of the invariants from speech data. In Section 4, we present the results of extracting the invariants from speech data, and quantify the discriminability of these invariants across phonetic attractors using the Kullback-Leibler divergence measure.

2. Nonlinear Dynamical Invariants

To characterize the structure of the underlying strange attractor from an observed time series, it is necessary to reconstruct a phase space from the time series. This reconstructed phase

space captures the structure of the original system's attractor (the true state-space that generated the observable). The process of reconstructing the system's attractor is commonly referred to as embedding.

The simplest method to embed scalar data is the method of delays. In this method, the pseudo phase-space is reconstructed from a scalar time series, by using delayed copies of the original time series as components of the RPS. It involves sliding a window of length m through the data to form a series of vectors, stacked row-wise in the matrix. Each row of this matrix is a point in the reconstructed phase-space. Letting $\{x_i\}$ represent the time series, the reconstructed phase space (RPS) is represented as:

$$X = \begin{pmatrix} x_0 & x_\tau & \cdots & x_{(m-1)\tau} \\ x_1 & x_{1+\tau} & \cdots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(m-1)\tau} \\ \vdots & & & \vdots \end{pmatrix}, \quad (1)$$

where m is the embedding dimension and τ is the embedding delay.

Taken's theorem [4] provides a suitable value for the embedding dimension, m. The first minima of the auto-mutual information versus delay plot of the time series is a safe choice for embedding delay [4].

2.1. Lyapunov Exponents

The analysis of separation in time of two trajectories with infinitely close initial points is measured by Lyapunov exponents [4]. For a system whose evolution function is defined by a function *f*, we need to analyze

$$\Delta x(t) \approx \Delta x(0) \frac{d}{dx}(f^N)x(0) \ . \quad (2)$$

To quantify this separation, we assume that the rate of growth (or decay) of the separation between the trajectories is exponential in time. Hence we define the exponents, λ_i as

$$\lambda_i = \lim_{n\to\infty} \frac{1}{n} \ln(\mathrm{eig_i} \prod_{p=0}^{n} J(p)) \ , \quad (3)$$

where, J is the Jacobian of the system as the point p moves around the attractor. These exponents are invariant characteristics of the system and are called Lyapunov exponents, and are calculating by applying (3) to points on the reconstructed attractor. The exponents read from a reconstructed attractor measure the rate of separation of nearby trajectories averaged over the entire attractor.

2.2. Fractal Dimension

Fractals are objects which are self-similar at various resolutions. Self-similarity in a geometrical structure is a strong signature of a fractal object. Correlation dimension [5] is a popular choice for numerically estimating the fractal dimension of the attractor. The power-law relation between the correlation integral of an attractor and the neighborhood radius of the analysis hyper-sphere can be used to provide an estimate of the fractal dimension:

$$D = \lim_{N\to\infty} \lim_{\varepsilon\to 0} \frac{\partial \ln C(\varepsilon)}{\partial \ln \varepsilon} \ , \quad (4)$$

where $C(\varepsilon)$, the correlation integral is defined as:

$$C(\varepsilon) = \frac{2}{N*(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(\varepsilon - \left\| \vec{x}_i - \vec{x}_j \right\|) \ , (5)$$

where $\vec{x}$ is a point on the attractor (which has N such points). The correlation integral is essentially a measure of the number of points within a neighborhood of radius ε, averaged over the entire attractor. To avoid temporal correlations in the time series from producing an underestimated dimension, we use Theiler's correction for estimating the correlation integral [5].

2.3. Kolmogorov-Sinai Entropy

Entropy is a well known measure used to quantify the amount of disorder in a system. It has also been associated with the amount of information stored in general probability distributions.

Numerically, the Kolmogorov entropy can be estimated as the second order Renyi entropy (K_2) and can be related to the correlation integral of the reconstructed attractor [5] as:

$$C_d(\varepsilon) \sim \lim_{\substack{\varepsilon\to 0 \\ d\to\infty}} \varepsilon^D \exp(-\tau\, d\, K_2) \ , \quad (6)$$

where D is the fractal dimension of the system's attractor, d is the embedding dimension and τ is the time-delay used for attractor reconstruction. This leads to the relation

$$K_2 \sim \frac{1}{\tau} \lim_{\substack{\varepsilon\to 0 \\ d\to\infty}} \ln \frac{C_d(\varepsilon)}{C_{d+1}(\varepsilon)} \ . \quad (7)$$

In a practical situation, the values of ε and d are restricted by the resolution of the attractor and the length of the time series.

3. Experimental Setup

In this work, we attempt to extract various nonlinear dynamical invariants of the underlying attractor from the observed acoustic utterances. We collected artificially elongated pronunciations of several vowels and consonants from 4 male and 3 female speakers. Each speaker produced sustained sounds (4 seconds long) for three vowels (/aa/, /ae/, /eh/), two nasals (/m/, /n/) and three fricatives (/f/, /sh/, /z/). The data was sampled at 22,050 Hz. For this preliminary study, we wanted to avoid artifacts introduced by coarticulation.

The acoustic data from each phoneme is embedded into a reconstructed phase space using time delay embedding with a delay of 10 samples. This delay was selected as the first local minimum of the auto-mutual information vs. delay curve averaged across all phones.

The choice of an embedding dimension of 5 was made after observing the plots of the Lyapunov spectra vs. embedding dimension over a range of embedding dimensions, and noting

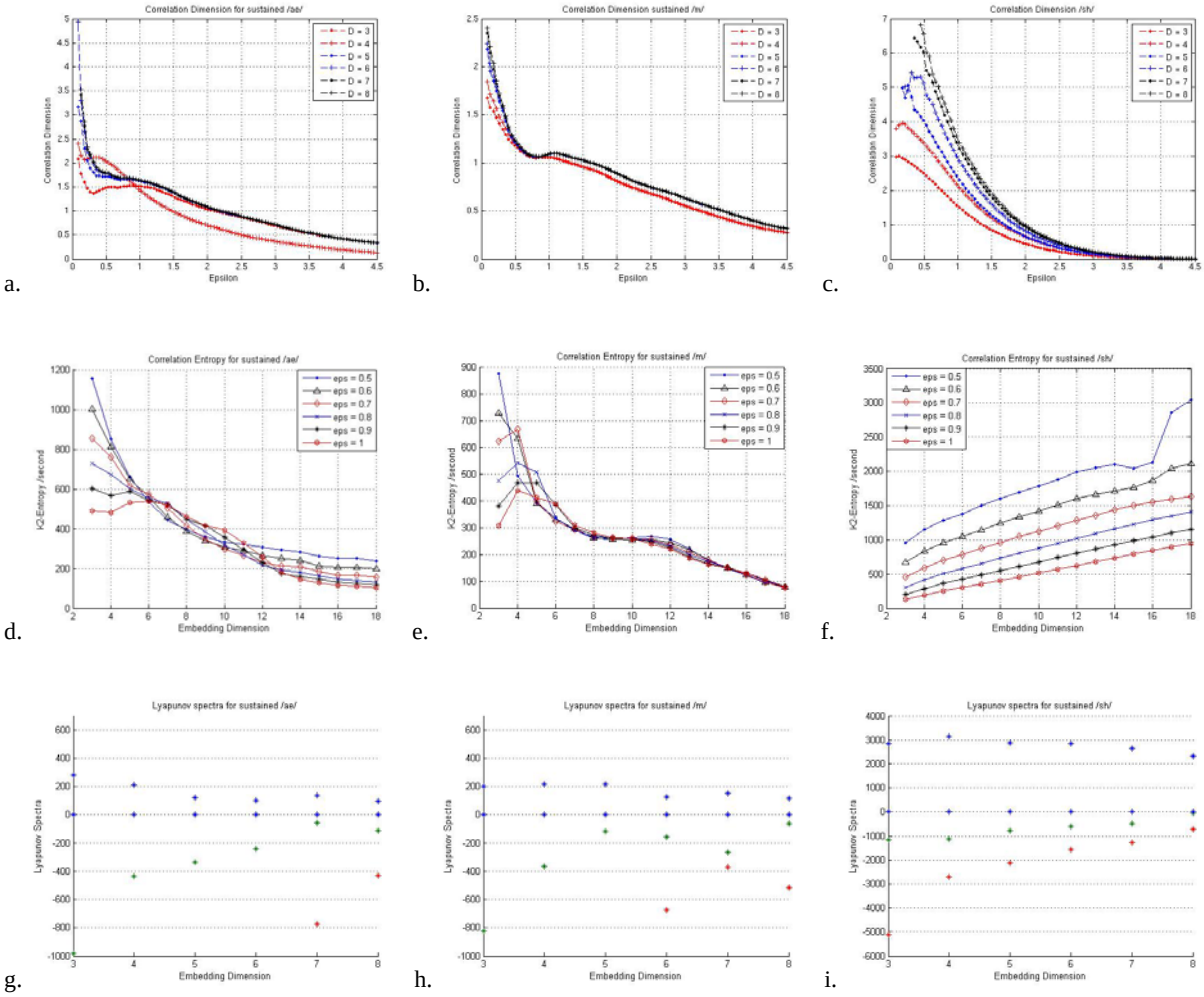

Figure 1 *Correlation Dimension (a through c), Kolmogorov Entropy (d through f), and Lyapunov Spectra (g through i) estimates for a vowel, a nasal and a fricative.*

that the estimates of the Lyapunov spectra converge at an embedding dimension of 5 for most phones, as shown in Figure 1.

To estimate the Lyapunov spectra from speech data, we used the algorithm described in [4]. We experimentally found the optimal (by varying the parameters and choosing the value at which we obtain convergence of the largest Lyapunov exponent) number of nearest neighbors to be 30, the evolution step size to be 5, and the number of sub-groups of neighbors as 15. A more detailed explanation of these parameters can be found in [2].

For estimates of Kolmogorov entropy, an embedding dimension of 15 was used. It is clear from (2) that for reliable entropy estimates, a high embedding dimension must be used.

As a measure of discrimination information between two statistical models representing dynamical information, we chose the Kullback-Leibler divergence measure [6]. We measured invariants for each phoneme using a sliding window, and built an accumulated statistical model over each such utterance. The discrimination information between a pair of models $p_i(\vec{x})$ and $p_j(\vec{x})$ is given by:

$$J(i,j) = \int_{\vec{x}} p_i(\vec{x}) \ln \frac{p_i(\vec{x})}{p_j(\vec{x})} d\vec{x} + \int_{\vec{x}} p_j(\vec{x}) \ln \frac{p_j(\vec{x})}{p_i(\vec{x})} d\vec{x} \ . \quad (8)$$

$J(i, j)$ provides a symmetric divergence measure between two populations i and j, from an information theoretic perspective. We use J as the metric for quantifying the amount of discrimination information across dynamical invariants extracted from different broad phonetic classes.

4. Results

Figure 1 shows the three dynamical invariants extracted from various phones using a variety of analysis parameters. For these experiments, we chose a window size of 1,500 samples. For the set of plots (a) through (c) in Figure 1, we vary the value of the neighborhood radius (epsilon) and study the variation in estimated fractal dimension with this parameter. We observe a

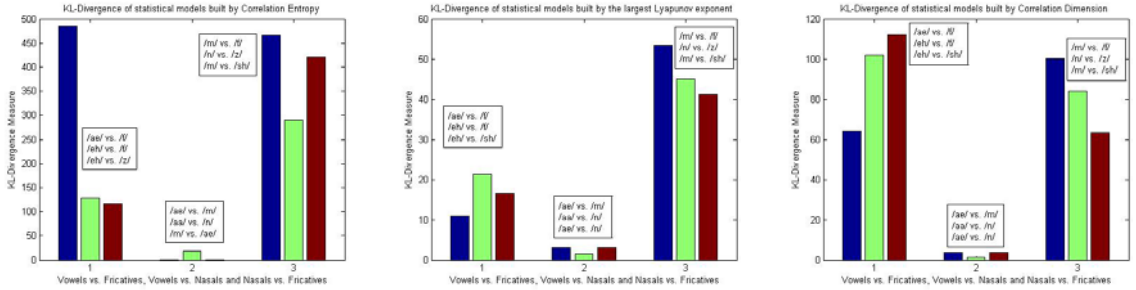

Figure 2 *KL Divergence Measure across various phonemes, using the three dynamical invariants.*

clear scaling region (where the dimension estimate is unaffected by variations in the neighborhood radius) for vowels and nasals (at epsilon ~ 0.75). Such a scaling region is not present in dimension estimates from fricatives. Also note that the estimate of fractal dimension for vowels and nasals is not sensitive to variations in embedding dimension from 5 through 8. However, the dimension estimate for fricatives increases consistently with an increase in the embedding dimension.

A similar trend is observed for plots (d) through (f), representing the Kolmogorov entropy estimates as a function of the embedding dimension. Once again, vowels and nasals have entropy estimates that stabilize at an embedding dimension of approximately 15. The entropy estimates for fricatives increase consistently with the embedding dimension. This behavior, along with the variation in dimension estimates with embedding dimension, reaffirms the conventional belief that unvoiced fricatives can be modeled using the combination of a noisy source and linear constant coefficient digital filter. If a time series were generated from an IID stochastic process, an increase in the embedding dimension adds to the randomness in the reconstructed phase space of this series, and hence leads to consistently increasing estimates of fractal dimension and attractor entropy. In [3], estimates of Lyapunov exponents could not be validated for fricatives, which is consistent with our observations using fractal dimension and Kolmogorov entropy estimates.

Plots (g) through (i) depict the Lyapunov spectra as a function of various embedding dimensions. Note that the positive exponent converges to a stable value at an embedding dimension of 5. Another technique for estimating the appropriate embedding dimension from a time series is the method of false nearest neighbors [5].

Figure 2 depicts the KL-divergence measure between phone models formed using the nonlinear dynamical invariants as features. Equation 3 has a closed form expression for normal distributions with different mean vectors and covariance matrices, which is what we used for estimating these divergence measures. We used a sliding window of length 36 ms to extract the invariants. The plots in this figure indicate the separation between statistical models generated using correlation entropy, Lyapunov exponents and correlation dimension extracted from utterances of all seven speakers. Note that the discrimination information of these features is high between vowels and fricatives and nasals and fricatives. The separation between nasals and vowel sounds is small.

5. Conclusions and Future Work

In this paper, we presented algorithms for extraction of three nonlinear dynamical invariants from speech data. We demonstrated the between-class separation of these invariants across 8 phonetic sounds. The results show promise in the potential use of these invariants for speech recognition applications.

In future work, we plan to study the speaker variability of these invariants, hoping that variations in the vocal tract response across speakers will result in different attractor structures, which will be captured by such invariants. We also plan to perform a pre-filtering of the analysis window before forming the reconstructed phase space for a more robust extraction.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0414450. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. References

[1] Kumar, A. and Mullick, S.K., "Nonlinear Dynamical Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 615-629, July 1996.

[2] Banbrook, M., *Nonlinear Analysis of Speech From a Synthesis Perspective*, Ph.D. Thesis, The University of Edinburgh, Edinburgh, UK, 1996.

[3] Kokkinos, I.; Maragos, P., "Nonlinear Speech Analysis using Models for Chaotic Systems," *IEEE Transactions on Speech and Audio Processing*, pp. 1098- 1109, Nov. 2005.

[4] Eckmann, J.P. and Ruelle, D., "Ergodic Theory of Chaos and Strange Attractors," *Reviews of Modern Physics*, vol. 57, pp. 617-656, July 1985.

[5] Kantz, H. and Schreiber T., *Nonlinear Time Series Analysis*, Cambridge University Press, UK, 2003.

[6] Campbell, J. P., "Speaker Recognition: A Tutorial," *Proceedings of IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.