Continuous Speech Recognition Using Nonlinear Dynamic Invariants

Daniel May[1]*, Tao Ma*[1]*, Sundar Srinivasan*[1]*, Georgios Lazarou*[2] *and Joseph Picone*[1]

[1] Department of Electrical and Computer Eng., Mississippi State University, MS State, MS, USA
[2] New York City Transit Authority, New York, New York, USA
{dom5, tm334, ss754, glaz, picone}@ece.msstate.edu

Abstract

In this paper, we combine traditional MFCCs with nonlinear dynamic invariants in an effort to produce a more robust feature vector for continuous speech recognition. This new feature vector exploits the underlying nonlinear dynamic properties that traditional linear techniques fail to capture. We performed a set of phoneme classification experiments using these new features and saw a maximum relative improvement of 10.3% for certain phoneme types. Evaluations of the Aurora-4 continuous speech recognition corpus show a maximum relative increase of 11.1% for the clean evaluation set. However, an average relative decrease of 7.6% was observed for the data sets containing noise.

Index Terms: nonlinear systems, nonlinear features, speech recognition

1. Introduction

For the past several decades, acoustic modeling for speech recognition has been based on the source-filter model and one-dimensional wave propagation in the vocal tract. The signal processing techniques that parameterize acoustic speech data into features operate primarily in the signal's frequency domain. This approach models the vocal tract as a linear filter and captures the lower-order characteristics of the speech production process. Recent theoretical and experimental evidence has suggested the existence of nonlinear characteristics in different types of speech and that that these characteristics contain significant information about speech production. While the traditional linear representation of speech has shown to be a reasonable means of acoustic modeling, it fails to capture this higher-order information of the acoustic dynamic system [1][2].

Dynamic systems can be represented by phase 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 $t \rightarrow \infty$. *Invariants* of a system's attractor are measures that quantify the topological or geometrical properties of the attractor and do not change under smooth transformations of the space. These smooth transformations include coordinate transformations such as phase space reconstruction of the observed time series [3].

Dynamic 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 [3][4] and more recently in the context of speech recognition [5]. Our previous work involves a thorough analysis of these invariants and their ability to discriminate between different types of speech signals [6]. Using a small database of elongated pronunciations of phones, we measured the between-class separation in a feature space comprised of these invariants and found that they were capable of discriminating between sustained phones.

In this paper, we continue our analysis of three standard dynamic invariants: Lyapunov exponents, fractal dimension, and Kolmogorov entropy. Lyapunov exponents [7] associated with a trajectory provide a measure of the average rates of convergence and divergence of nearby trajectories. Fractal dimension [8] is a measure that quantifies the number of degrees of freedom and the extent of self-similarity in the attractor's structure. Kolmogorov entropy [8] 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 dynamic system, they are prime candidates for feature extraction of a signal with strong nonlinearities. The motivation behind studying such invariants from a signal processing perspective is to capture the relevant nonlinear dynamic information from the time series – something that is ignored in conventional spectral-based analysis.

Recent work has shown that the combination of fractal dimension with Mel-frequency cepstral coefficients (MFCCs) improves recognition performance for speech contaminated with noise [9]. This provides sufficient motivation for an investigation into additional dynamic invariants. We combine the three invariants mentioned above with the traditional MFCCs to create a new feature vector that exploits both the linear acoustic model and the nonlinear dynamic information of the signal. We use this new feature vector to evaluate the Aurora-4 large vocabulary evaluation corpus and compare the recognition accuracy to a system using only MFCCs. 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 dynamic system analysis. We provide a brief review of the algorithms we employed for the extraction of three dynamic invariants from a time series. In Section 3, we describe the preliminary signal classification experiments that demonstrate the ability of these invariants to model acoustics better than traditional MFCCs by themselves. Finally, in Section 4 we present continuous speech recognition results of the Aurora-4 corpus using different combinations of MFCCs and invariants.

2. Nonlinear Dynamic Invariants

Nonlinear systems can best be represented by their phase space which defines every possible state of the system. The dimensions of the phase space correspond to the system's dynamic variables, and each point in the space corresponds to a unique state of the system. 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:

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

where m is the embedding dimension and τ is the embedding delay. Taken's theorem [7] 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 [7].

2.1. Lyapunov Exponents

The analysis of separation in time of two trajectories with infinitely close initial points is measured by Lyapunov exponents [7]. For a system whose evolution function is defined by a function *f*, we need to analyze

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

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

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

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 and quantify the level of chaos present in the attractor. Attractors corresponding to chaotic systems will generally have high Lyapunov exponents while the exponents from more stable, periodic systems will have lower exponents. Through experimentation, it was found that an embedding dimension of 5 since the Lyapunov spectra converge at 5 over a range of embedding dimensions. A more detailed explanation of this and other parameter values can be found in [6].

2.2. Fractal Dimension

Some geometrical objects have a characteristic called self-similarity. An object is characterized as self-similar if a close-up examination of the object reveals that it is composed of smaller versions of itself. Self-similarity in a geometrical structure can be quantified and defines the degree to which it occupies a space. This value is called fractal dimension.

Correlation dimension [8] is a popular choice for numerically estimating the fractal dimension of an 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:

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

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

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

where $\vec{x}$ is one of *N* points on the attractor. 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 [8].

2.3. Kolmogorov 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 [8] as:

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

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

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

In practice, the values of ε and *d* are restricted by the resolution of the attractor and the length of the time series. We found that an embedding dimension 15 gives consistent estimations of Kolmogorov entropy [6].

3. Phoneme Classification Experiments

In this work, we combine the traditional 39 dimensional MFCC feature vector (consisting of 12 MFCCs, absolute energy, and their first and second derivatives) with nonlinear dynamic invariants and evaluate this combination on the Wall Street Journal derived Aurora-4 large vocabulary evaluation corpus. This corpus represents a well-established LVCSR benchmark and constitutes a balanced trade-off between computational resources and complexity. Also, the limited 5,000 word vocabulary makes this corpus conducive to acoustic modeling research. The subset of the corpus used for our experiments is divided into a training set and seven evaluation sets. The training set consists of 7,138 utterances from 83 speakers totaling 14 hours of speech. The evaluation sets consist of one clean set, and six sets consisting of various levels of digitally-added noise. Each evaluation set consists of 330 utterances from 8 different speakers. All utterances are sampled at 16 kHz.

In an effort to determine whether or not the combination of these invariants with MFCCs is able to better model continuous speech, we perform a set of preliminary phoneme classification experiments. Using automatic, time-aligned phonetic transcriptions of the clean corpus data, we match segments of the continuous speech to 40 phonemes. For each of the feature combinations, a 16-mixture GMM is estimated for every phoneme. Using the same data, we then classify each of the signal frames as one of the 40 phonemes. Table 1 summarizes the relative difference in classification accuracy between the baseline MFCC feature vector and the MFCC/Invariant combination feature vector. Figure 1 illustrates relative improvements for several individual phonemes.

In Table 1, we see that the average relative classification accuracy increases significantly for affricates and stops, with the most dramatic increase for affricates using the correlation dimension invariant where we get an increase of 10.3%. Stops show a fairly consistent increase for all three invariants. The use of the correlation entropy invariant resulted in an improvement for all phoneme types except for fricatives. Many of the phoneme types saw little or no improvements, and although some suffered a decrease in accuracy, these decreases are minimal.

Figure 1 illustrates some of the results seen in Table 1 by showing the relative classification improvement for several individual phones. The relative improvements for affricates and stops are high for each of the invariants while the nasal phonemes saw little or no improvements. These results are encouraging. The accuracy improvements in these low-level phoneme recognition experiments suggest that we will likely see accuracy increases in continuous speech recognition experiments.

4. Speech Recognition Experiments

Our preliminary experiments provide strong support that the addition of these nonlinear invariants the standard MFCC feature vector will improve the accuracy of speech recognition tasks. We next present two sets of continuous speech recognition experiments, each using acoustic models trained from the clean training set mentioned in the previous section. The first set evaluates the noise-free test set using each of the different MFCC/invariant feature vector combinations. The results of these experiments are outlined in Table 2. The purpose of these experiments is to determine whether these new feature vectors will improve recognition performance for an evaluation set with environmental conditions that match those of the training set. The second set of experiments evaluates seven different test sets, each with varying levels and types of additive noise that would be encountered in the following environments: an airport, random babble, a vehicle, a restaurant, the street, and on a train. The results of these experiments are outlined in Table 3. The purpose of this second set is to determine whether or not these nonlinear invariants improve the robustness of the acoustic models to noise conditions that are unseen in the training data.

Table 1. *Average relative phoneme classification improvements using MFCC/Invariant combination*

	Correlation Dimension	Lyapunov Exponent	Correlation Entropy
Affricates	10.3%	2.9%	3.9%
Stops	3.6%	4.5%	4.2%
Fricatives	-2.2%	-0.6%	-1.1%
Nasals	-1.5%	1.9%	0.2%
Glides	-0.7%	-0.1%	0.2%
Vowels	0.4%	0.4%	1.1%

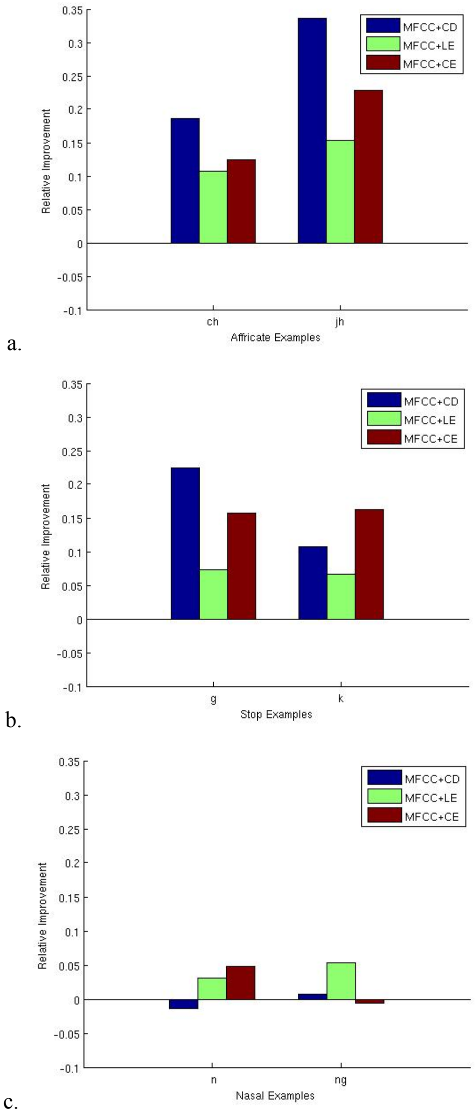

Figure 1: *Relative improvements for several phonemes.* (*Affricates (a), Stops (b), Glides (c)*)

All experiments use the ISIP prototype system developed at Mississippi State University. This open-source speech

Table 2. *Continuous Speech Recognition Results for Clean Evaluation Data (no additive noise) and the Relative Improvement vs. the Baseline MFCCs*

	WER (%)	Improvement (%)
Baseline	13.5	--
Correlation Dimension (CD)	12.2	9.6
Lyapunov Exponent (LE)	12.5	7.4
Correlation Entropy (CE)	12.0	11.1
All Invariants	12.8	5.2

Table 3. *Continuous Speech Recognition Results for Noisy Evaluation Data*

	WER (%)					
	Airport	Babble	Car	Restaurant	Street	Train
Baseline	53.0	55.9	57.3	53.4	61.5	66.1
CD	57.1	59.1	65.8	55.7	66.3	69.6
LE	56.8	60.8	60.5	58.0	66.7	69.0
CE	52.8	56.8	58.8	52.7	63.1	65.7
All	58.6	63.3	72.5	60.6	70.8	72.5

recognition system uses HMMs to model acoustics and a trigram backoff language model. The models trained for these experiments are cross-word context dependent HMMs with underlying 4-mixture Gaussians.

The recognition results for the clean test set are very encouraging. Each of the MFCC/invariant feature combinations results in a significant recognition performance increases over the baseline MFCC experiments. Correlation entropy results in the largest relative improvement of 11.1%. This reflects the results in Section 3 where we saw a relatively consistent improvement in phoneme accuracy for correlation entropy. While combining all three of the invariants results in an improvement over the baseline, this improvement is not as significant as each of the invariants by themselves. This seems to suggest that the new features contribute a certain level of overlapping information.

The recognition results for the noisy test sets are less encouraging as each experiment resulted in a performance decrease compared the baseline. These results contradict our theory that the addition of invariants would result in a feature vector that is more robust to noisy conditions unseen in the training set. We are currently doing further research to understand this discrepancy, and are focused on a closer examination of our invariant computations. We are also more closely examining some filtering methods which may enhance the algorithms' robustness to noise.

5. Conclusions and Future Work

In this paper, we presented the technique of combining nonlinear dynamic invariants to traditional MFCCs to create a feature vector that is able to simultaneously model the linear acoustics and the nonlinear dynamic information of a speech signal. We saw that some of these invariants are able to improve classification of certain phonemes within continuous speech. We also found that these invariants are able to improve the recognition accuracy of continuous speech recognition tasks when the evaluation data is not contaminated with noise. However, when evaluation data is contaminated with noise, our experiments indicate an increase in WER. We are still investigating the cause of this performance decrease and experimenting with some various filtering methods which will attempt to remove the adverse noise effects from the attractor.

In future work, we hope to develop a method for directly modeling the attractor and use this model to replace traditional HMMs for continuous speech recognition

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

[2] Lindgren, A. C., Johnson, M. T. and Povinelli, R. J., "Speech Recognition Using Reconstructed Phase Space Features", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 60—63, 2003.

[3] 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.

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

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

[6] Prasad, S., Srinivasan, S., Pannuri, M., Lazarou, G., Picone, J., "Nonlinear Dynamical Invariants for Speech Recognition," *Proceedings of the International Conference on Spoken Language Processing*, pp. 2518-2521, Pittsburgh, Pennsylvania, USA, Sept. 2006.

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

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

[9] Pitsikalis, V. and Maragos, P., "Filtered Dynamics and Fractal Dimensions for Noisy Speech Recognition," *IEEE Signal Processing Letters*, vol. 13, no. 11, pp. 711-714, Nov. 2006.