CONTINUOUS SPEECH RECOGNITION USING
NONLINEAR DYNAMIC INVARIANTS

Abstract

In this paper, we combine the traditional MFCC feature vector with nonlinear dynamic invariants in an effort to produce a more robust feature vector for continuous speech recognition. This new feature vector exploits the linear acoustic properties of the speech signal as well as the underlying nonlinear dynamic properties that traditional linear techniques fail to capture. We perform a set of preliminary experiments which use these new feature vectors to classify segments of continuous speech as different phonemes, and the results of these experiments provide motivation for larger scale speech recognition experiments. Continuous speech recognition experiments on the Aurora-4 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.

# 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. Recent theoretical and experimental evidence has found the existence of nonlinear characteristics in speech and suggests 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 the higher-order information of the acoustic dynamic system (Maragos, Dimakis and Kokkinos, 2002; Lindgren, Johnson and Povinelli, 2003).

Dynamic systems can be represented by state-space models, where the states of the system evolve in accordance with a deterministic evolution function. The path traced by the system’s states as they evolve over time is referred to as a *trajectory*. An *attractor* is the set of points in the state space that are accumulated as *t*→∞. *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 such as phase space reconstruction of the observed time series (Kumar and Mullick, 1996).

Dynamic invariants are a natural choice for characterizing the system that generated the observables. These measures have been previously studied in the context of analysis and synthesis research (Kumar and Mullick, 1996; Banbrook, 1996), and more recently in the context of speech recognition (Kokkinos and Maragos, 2005). Current research involves a thorough analysis of these invariants and their ability to discriminate between different types of speech signals (Prasad, *et al*., 2006).

In this paper, we continue the analysis of three standard dynamic invariants: Lyapunov exponents, fractal dimension, and Kolmogorov entropy. Lyapunov exponents (Eckmann and Ruelle, 1985) associated with a trajectory provide a measure of the average rates of convergence and divergence of nearby trajectories. Fractal dimension (Kantz and Schreiber, 2003) is a measures the extent of self-similarity in the attractor’s structure, and Kolmogorov entropy (Kantz and Schreiber, 2003) measures the rate of information loss or gain over the trajectory. These measures search for a signature of chaos in the observed time series. The motivation behind studying such invariants from a signal processing perspective is to capture the relevant nonlinear dynamic information from the time series which 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 accuracy (Pitsikalis and Maragos, 2006). This provides sufficient motivation for an investigation into additional dynamic invariants. We combine the three invariants mentioned above with the traditional MFCCs to form a new feature vector that exploits both the linear acoustic model and the nonlinear dynamic information of the signal. We evaluate performance of this new feature vector on 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, which is the starting point for computing invariants, and provide a brief overview of the algorithms employed for the extraction of 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 alone. Finally, in Section 4 we present continuous speech recognition results of the Aurora-4 corpus using combinations of MFCCs and invariants.

# Nonlinear Dynamic Invariants

To characterize the structure of the underlying attractor of 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 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:

  ,              (1)

where *m* is the embedding dimension andis the embedding delay.

Taken’s theorem (Eckmann and Ruelle, 1985) provides a suitable method for estimating the value of the embedding dimension,. The first minima of the auto-mutual information versus delay plot of the time series is a safe choice for embedding delay (Eckmann and Ruelle, 1985).

## Lyapunov Exponents

The analysis of separation in time of two trajectories with infinitely close initial points is measured by Lyapunov exponents (Eckmann and Ruelle, 1985). For a system with an evolution function defined by *f*, we need to analyze

 . (2)

To quantify this separation, we assume that the rate of separation between the trajectories is exponential in time. Hence we define the exponents,as

 , (3)

where, J is the Jacobian of the system as the point *p* moves around the attractor. These exponents are referred to as Lyapunov exponents and are calculating by applying  to points on the reconstructed attractor, and averaging them over the entire attractor.

## Fractal Dimension

Fractals are geometrical structures which are self-similar at various resolutions. Correlation dimension (Kantz and Schreiber, 2003) 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:

 ,                        (4)

where, the correlation integral is defined as:

,  (5)

where 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, averaged over the entire attractor.

## 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 () and can be related to the correlation integral of the reconstructed attractor (Kantz and Schreiber, 2003) as:

 ,               (6)

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

 .                        (7)

In practice, the values ofand are restricted by the resolution of the attractor and the length of the time series.

# Phoneme Classification Experiments

In this work, we combine the traditional MFCC feature vector with nonlinear dynamic invariants and evaluate this combination on the Wall Street Journal derived Aurora-4 corpus. This corpus represents a well-established LVCSR benchmark and constitutes a balanced trade-off between computational resources and complexity. The subset of the corpus used for our experiments is divided into a training set and seven evaluation sets. The evaluation sets consist of one clean set, and six sets consisting of various levels of digitally‑added noise.

To determine whether the combination of these invariants with MFCCs is able to better model continuous speech, we perform a set of preliminary phoneme classification experiments. Using 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 phonemes. The results showed a relative improvement of 1.65% for fractal dimension, 1.5% for Lyapunov exponents, and 1.42% for entropy.

The accuracy improvements in these low-level phoneme recognition experiments suggest that we will likely see performance increases in continuous speech recognition experiments.

# Speech Recognition Experiments

The experiments outlined in Section 3 provide motivation for running larger-scale continuous speech recognition experiments. We next present two sets of 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 new MFCC/invariant feature vector combinations. The results of these experiments are outlined in . The purpose of these experiments is to determine whether the 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. The results of these experiments are outlined in . The purpose of this second set is to determine whether these nonlinear invariants improve the robustness of the acoustic models to noise conditions that are unseen in the training data.

|  |  |  |
| --- | --- | --- |
|  | **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 1: Continuous Speech Recognition Results for Clean Evaluation Data (no additive noise) and the Relative Improvement vs. the Baseline MFCCs

|  |  |
| --- | --- |
|  | **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 |

Table 2: Continuous Speech Recognition Results for Noisy Evaluation Data

All experiments use the ISIP prototype system developed at Mississippi State University. This open-source speech recognition system uses HMMs to model acoustics and a trigram 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%. 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.

# Conclusions and Future Work

In this paper, we presented a technique for combining nonlinear dynamic invariants with 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 observed 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 degradation.

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. We are also investigating ways to extend these techniques to the language modeling problem in speech recognition.

# 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.

# References

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

Eckmann, Jean-Pierre and Ruelle, David, “Ergodic Theory of Chaos and Strange Attractors,” *Reviews of Modern Physics*, vol. 57, pp. 617‑656, July 1985.

Kantz, Holger and Schreiber, Thomas, *Nonlinear Time Series Analysis*, Cambridge University Press, UK, 2003.

Kokkinos, Iasonas; Maragos, Petros, “Nonlinear Speech Analysis using Models for Chaotic Systems,” *IEEE Transactions on Speech and Audio Processing*, pp. 1098- 1109, Nov. 2005.

Kumar, Arun 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.

Lindgren, Andrew C.; Johnson, Michael T. and Povinelli, Richard J., “Speech Recognition Using Reconstructed Phase Space Features”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 60—63, 2003.

Maragos, Petros; Dimakis. Alex G. and Kokkinos, Iasonas, “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.

Pitsikalis, Vassilis and Maragos, Petros, “Filtered Dynamics and Fractal Dimensions for Noisy Speech Recognition,” *IEEE Signal Processing Letters*, vol. 13, no. 11, pp. 711-714, Nov. 2006.

Prasad, Saurabh; Srinivasan, Sundar; Pannuri, Madhulika; Lazarou, Georgios; Picone, Joseph, “Nonlinear Dynamical Invariants for Speech Recognition,” *Proceedings of the International Conference on Spoken Language Processing*, pp. 2518-2521, Pittsburgh, Pennsylvania, USA, Sept. 2006.