Nonlinear Statistical Modeling of Speech

S. Srinivasana, T. Maa, D. Maya, G. Lazaroub and J. Piconea

aDepartment of Electrical and Computer Eng., Mississippi State University, MS, USA

bNew York City Transit Authority, New York, New York, USA

 {ss754, tm334, dom5}@ece.msstate.edu, {georgios.lazarou, joseph.picone}@gmail.com

**Abstract.** Contemporary approaches to speech and speaker recognition decompose the problem into four components: feature extraction, acoustic modeling, language modeling and search. Statistical signal processing is an integral part of each of these components, and Bayes Rule is used to merge these components into a single optimal choice. Acoustic models typically use hidden Markov models based on Gaussian mixture models for state output probabilities. This popular approach suffers from an inherent assumption of linearity in speech signal dynamics. Language models often employ a variety of maximum entropy techniques, but can employ many of the same statistical techniques used for acoustic models.

In this paper, we focus on introducing nonlinear statistical models to the feature extraction and acoustic modeling problems as a first step towards speech and speaker recognition systems based on notions of chaos and strange attractors. Our goal in this work is to improve the generalization and robustness properties of a speech recognition system. Three nonlinear invariants are proposed for feature extraction: Lyapunov exponents, correlation fractal dimension, and correlation entropy. We demonstrate an 11% relative improvement on speech recorded under noise-free conditions, but show a comparable degradation occurs for mismatched training conditions on noisy speech. We conjecture that the degradation is due to difficulties in estimating invariants reliably from noisy data.

 To circumvent these problems, we introduce two dynamic models to the acoustic modeling problem: (1) a linear dynamic model (LDM) that uses a state space-like formulation to explicitly model the evolution of hidden states using an autoregressive process, and (2) a data-dependent mixture of autoregressive (MixAR) models. Results show that LDM and MixAR models can achieve comparable performance with HMM systems while using significantly fewer parameters. Currently we are developing Bayesian parameter estimation and discriminative training algorithms for these new models to improve noise robustness.

Keywords: nonlinear statistical models, chaos, machine learning, speech recognition.

# Introduction

Statistical or machine-learning techniques, such as Hidden Markov models (HMMs) and Gaussian mixture models (GMMs), have dominated the signal processing and pattern recognition literature for the past 25 years. However, such approaches are prone to overfitting and have problems with generalization. For example, delivering high performance on previously unseen noise conditions remains an elusive goal. A lack of robustness to previously unseen conditions is a major impediment to the success of human language technology in many important application spaces. In this paper, we summarize our attempts to advance state of the art by applying principles of nonlinear statistical modeling to acoustic modeling in speech recognition and speaker verification.

A typical pattern recognition approach to speech recognition based on a Bayesian model is shown in FIGURE 1. The acoustic front-end and the acoustic model, which are the focus of this research, are used to compute the maximum likelihood contribution of the overall posterior probability. Popular approaches for both of these suffer from an inherent assumption of linearity in speech signal dynamics. The language model is used to compute prior probabilities, often employing a variety of maximum entropy techniques but is not the subject of this research. However, the techniques described in this paper can also be applied to language modeling research, something we hope to investigate in future research. The search component ties all these models together by searching a very large space for the most probable sequence of words (or symbols), and is also not the subject of this paper. Statistical signal processing is an integral part of each of these components, and Bayes Rule is used to merge these components into a single optimal choice.

Based on notions of chaos and strange attractors, we focus on introducing nonlinear statistical models to the feature extraction and acoustic modeling problems as a first step towards speech and speaker recognition systems. Three nonlinear invariants are proposed for feature extraction: Lyapunov exponents, correlation fractal dimension, and correlation entropy. We demonstrate an 11% relative improvement on speech recorded under noise-free conditions, but show a comparable degradation occurs for mismatched training conditions on noisy speech. We conjecture that the degradation is due to difficulties in estimating invariants reliably from noisy data. To circumvent these problems, we introduce two dynamic models to the acoustic modeling problem: (1) a linear dynamic model (LDM) that uses a state space-like formulation to explicitly model the evolution of hidden states using an autoregressive process, and (2) a data-dependent mixture of autoregressive (MixAR) models.

FIGURE . An overview of a typical pattern recognition approach to speech recognition. In this work, we have focused on the acoustic front-end and acoustic modeling.

# Nonlinear Dynamic Invariants as Speech Features

Nonlinear systems can best be represented by their phase space which defines every possible 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. In this work, we use embedding to achieve this [1,2].

Phase spaces can be characterized by nonlinear dynamic invariants. Three nonlinear dynamic invariants are considered [3,4,5,6] – Lyapunov Exponents measuring the rates of convergence and divergence of nearby trajectories, Correlation Fractal Dimension quantifying number of degrees of freedom and the extent of self-similarity in the attractor’s structure, and Correlation entropy measuring the rate of information loss or gain over the trajectory.

## 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.

|  |
| --- |
| **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% |

In an effort to determine whether or not the combination of these invariants with MFCCs is able to better model continuous speech, we performed a set of preliminary phoneme classification experiments. Using automatic, time-aligned phonetic transcriptions of the clean corpus data, we matched segments of 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. summarizes the relative difference in classification accuracy between the baseline MFCC feature vector and the MFCC/Invariant combination feature vector.

In , 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.

According to the results seen in we can see 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.

## Speech Recognition Experiments

Our preliminary experiments provided strong support that the addition of these nonlinear invariants the standard MFCC feature vector will improve the accuracy of speech recognition tasks. We next conducted 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 . The purpose of these experiments was 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.

|  |
| --- |
| **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 |

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 . 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 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 |

All experiments use the ISIP prototype system [7] developed at Mississippi State University. This open-source speech 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 resulted in a significant increase in recognition performance over the baseline MFCC experiments. Correlation entropy resulted in the largest relative improvement of 11.1%. While combining all three of the invariants resulted in an improvement over the baseline, this improvement was 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 were 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 conjecture that the degradation is due to difficulties in estimating invariants reliably from noisy data.

|  |
| --- |
| **TABLE 4.** Classification (% accuracy) results for the Aurora-4 large vocabulary corpus (the relative improvements are shown in parentheses) |
| **Model** | **Clean Dataset** | **Noisy Dataset** |
| HMM (4-mix) | 46.9(-) | 36.8(-) |
| LDM | 49.2 (4.9%) | 39.2 (6.5%) |

# Linear Dynamic Modeling for Speech Classification

In a Linear Dynamic Model (LDM), systems are described as underlying states and observables combined by a measurement equation [8]. Suppose *yt* is a *p*-dimensional observation vector and *xt* is a *q*-dimensional internal state vector, the LDM formulation is based on a state-space model as follows:

  (1)

where ***F*** is the state evolution matrix and ***H*** is the observation transformation matrix. The variables *εt* and *ηt* are assumed to be uncorrelated white Gaussian noise with covariance matrices ***C*** and ***D***, respectively, which drive the linear stochastic system. The sequences of observations *yt* and the underlying states *xt* are finite and dimensionless following multivariate Gaussian distributions for all *t*.

A phone classification experiment was conducted to evaluate LDMs on the Aurora‑4 large vocabulary corpus with 5K words dictionary. An HMM system was used to generate alignments at the phone level. Each phone instance is treated as one segment. A total of 40 LDM phone models were estimated. Each model was trained using the segmental features derived from 13-dimensional frame-level feature vectors comprised of 12 cepstral coefficients and absolute energy.

 summarizes the results of the Aurora-4 phoneme classification experiments. We can see that the LDM classifiers achieved superior performance to baseline HMM classifiers with a classification accuracy of 49.2% for the clean evaluation data and 39.2% for the noisy evaluation data. This represents a 4.9% relative and a 6.5% relative increase in performance over a comparable HMM system with 3-state models. LDM appears to offer improved generalization over the HMM baseline system across different channel conditions, which makes LDM a more robust speech recognition technique.

# Mixture of Autoregressive (MixAR) models for Speaker Recognition

A mixture autoregressive process (MixAR) of order *p* with *m* components, *X*={*x*[*n*]}, is defined as [9]:

  ()

where, *εi* is a zero-mean Gaussian random process with a variance of *σj2*, “w.p.” denotes “with probability” and the gating weights, *Wi* sum to 1. The linear prediction coefficients, {*ai*}, represent the dynamic model, where *ai,0* are the component means, while {*wi*, *gi*} are called gating coefficients. It is apparent that an *m*-mixture MixAR process is the weighted sum of *m* Gaussian autoregressive processes, with the time-dependent weights depending on previous data and the gating coefficients.

We applied the MixAR model to the 1-speaker detection task in the 2001 NIST SRE Corpus. Only the development database was used. All 60 speakers were used for training and all 78 utterances were used for evaluation. Each training utterance was about 2 minutes long, while the test utterances were of varying length not exceeding 60 seconds. Static (13 MFCCs), delta (26 MFCCs) and delta-delta (39 MFCCs) features were extracted.

MixAR and GMM performance was then evaluated as a function of the number of mixtures. The EER results are shown in . Also indicated in parenthesis is the number of parameters for each case. From this table it is clear that MixAR can achieve about the same performance using almost 4x fewer parameters than GMM. This reduction in the number of parameters points to the efficiency of MixAR in capturing the dynamic information. Moreover, even when considering the best case scenario for GMM with a large number of parameters (8 mixtures with static as well as velocity and acceleration coefficients), there is a 10.6% relative reduction in EER with MixAR. This is a strong indication that there is some amount of nonlinear evolution information in speech features that GMM model cannot capture using linear derivatives alone and MixAR can effectively employ this information for achieving better speaker recognition.

|  |
| --- |
| **TABLE 5.** Speaker recognition EER with NIST for MixAR and GMM as a function of #mix. (the numbers of parameters are shown in parentheses) |
| **# mix.** | **MixAR Static+∆+∆∆** | **MixAR Static** |
| 2 | 23.1(216) | 24.1(120) |
| 4 | 21.7(432) | 19.2(240) |
| 8 | 20.5(864) | 19.1(480) |
| 16 | 20.5(1728) | 19.2(960) |

# Conclusions

In this paper, we explored three statistical pathways to improve the performance of current automatic speech and speaker recognition systems by using nonlinear dynamic information in speech. We saw that nonlinear invariant features are able to improve classification of certain phonemes within continuous speech as well as to improve speech recognition performance for clean continuous speech. However this method was not robust to noise primarily because of difficulties in obtaining reliable estimates of invariants. We explored LDM to model the acoustics of speech and found that it improves classification accuracy especially in noisy conditions. Finally, we applied a mixture autoregressive model for speaker recognition and obtained improved performance with 4x fewer parameters compared to that with traditional GMM models.

In future work, we hope to develop Bayesian parameter estimation and discriminative training algorithms for these new models to improve noise robustness.

# 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

1. P. Maragos, A.G. Dimakis and I. Kokkinos, “Some Advances in Nonlinear Speech Modeling Using Modulations, Fractals, and Chaos,” in *Proc. Int. Conf. on Digital Signal Processing (DSP-2002)*, Santorini, Greece, July 2002.
2. A.C. Lindgren, M.T. Johnson, and R.J. Povinelli, “Speech Recognition Using Reconstructed Phase Space Features,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 60-63, April 2003.
3. A. Kumar, and S.K. Mullick, “Nonlinear Dynamical Analysis of Speech,” *Journal of the Acoustical Society of America*, vol.  100, no.  1, pp.  615-629, July 1996.
4. I. Kokkinos and P. Maragos, “Nonlinear Speech Analysis using Models for Chaotic Systems,” *IEEE Transactions on Speech and Audio Processing*, pp. 1098-1109, November 2005.
5. J.P. Eckmann and D. Ruelle, “Ergodic Theory of Chaos and Strange Attractors,” *Reviews of Modern Physics*, vol. 57, pp. 617-656, July 1985.
6. D. May, Nonlinear Dynamic Invariants For Continuous Speech Recognition, M.S. Thesis, Dept. of Elect. and Comp. Eng., Mississippi State University, May 2008.
7. ISIP prototype system, *http://www.isip.piconepress.com/projects/speech/software/*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, August 2008.
8. J. Frankel and S. King, “Speech Recognition Using Linear Dynamic Models,” *IEEE Trans. on Speech and Audio Proc.*, vol. 15, no. 1, pp. 246-256, January 2007.
9. C.S. Wong and W.K. Li, “On a Mixture Autoregressive Model,” *Journal of the Royal Statistical Society*, vol. 62, no. 1, pp. 95-115, February 2000.