09/30/07 — 08/31/08: RESEARCH ACTIVITIES

The primary goal of this project is to develop novel nonlinear modeling techniques for speech and speaker recognition systems. In previous years of this project, we have explored various classical nonlinear models with limited success. This year we focused on two approaches, a nonlinear mixture autoregressive model and a linear dynamic model. We also completed experiments on adding nonliear features to the acoustic feature vector.

An overarching challenge in this project has been dealing with the nonstationarity of the speech signal. We have explored several models that, if given ample amounts of data, appear capable of modeling nonlinar behavior. The parameter estimates of these models converge over time scales of seconds. However, the speech signal varies on the order of 10 to 30 msec. Some phonemes crucial to good recognition and verification performance last only a few tens of milliseconds, on the order of a few frames of data. We have yet to find success at estimating parameters of these nonlinear models over such short durations of speech. In such cases, one only has a few hundred samples with which to compute the model parameters. Worse, coarticulation phenomena often impinge on these samples, making it very difficult to estimate phonemes with short duration. This is one of the great challenges of speech processing – estimating model parameters from a signal that is evolving on very short time scales. We even attempted to oversample the signal in an effort to increase the amount of data, but even this approach suffers from some well-known drawbacks.

One of the reasons we shifted the focus of the original proposal from recognition to verification was to provide a longer time epoch over which parameters could be estimated. Unfortunately, though in verification you have access to a long speech utterance, the fact that the signal model is continually evolving during this time makes it difficult to separate out speaker variations from phoneme variations. Nonlinear models attempting to directly estimate long-term suprasegmental parameters related to the speaker's identity also did not perform well.

Therefore, in the final year of this project we focused on approaches that blend well with the traditional, piecewise linear approximation that has been so successful in speech recognition. We also attempted to leverage as much of the hidden Markov model infrastructure as possible. In this report, we summarize findings of the three approaches mentioned above. One M.S. thesis is available that describes the feature extraction work in more detail; two PhD theses are under development that describe the other work in greater detail. All are available from our web site while under construction in an effort to disseminate the project information as quickly as possible.

A. Continuous Speech Recognition with Nonlinear Dynamic Invariants

This year we completed our work on nonlinear dyanmic invariants, culminating with the publication of an M.S. thesis [1] and submissions to several conferences including INTERSPEECH 2008 [2][4]. We have previously reported our general approach was to combine traditional MFCCs with nonlinear dynamic invariants in an effort to produce a more robust feature vector for continuous speech recognition.

We continued our analysis of three standard dynamic invariants: Lyapunov exponents, fractal dimension, and Kolmogorov entropy. Lyapunov exponents associated with a trajectory provide a measure of the average rates of convergence and divergence of nearby trajectories. Fractal dimension is a measure that quantifies the number of degrees of freedom and the extent of self-similarity in the attractor's structure. Kolmogorov entropy 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.

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 evaluated these features on two sets of continuous speech recognition experiments involving the Aurora 4 task (a variant of the WAJ 5K task).

The recognition results for the clean test set were very encouraging. Each of the MFCC/invariant feature combinations resulted in a significant recognition performance increases over the baseline MFCC experiments. Correlation entropy resulted in the largest relative improvement of 11.1%. The recognition results for the noisy test sets were less encouraging as each experiment resulted in a performance decrease (higher error rate) compared to 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 also examined more closely some state-space filtering methods which we hoped would enhance the algorithms' robustness to noise. These also did not result in any improvements. This work is described in greater detail in [1].

B. MixAR Modeling of Speech Signals for Speaker Recognition

Gaussian mixture models are a very successful method for modeling the output distribution of a state in a hidden Markov model (HMM). However, this approach is limited by the assumption that the dynamics of speech features are linear and can be modeled with static features and their derivatives. In this work, a nonlinear mixture autoregressive model is used to model state output distributions (MAR-HMM). This model can handle both static and dynamic features.

The MAR-HMM model we have developed is a generalized version of [5] that has been extended to handle vector observations, so that we can operate on the speech feature vector stream rather than speech samples. One property of MAR that is of particular relevance is the ability of MAR to model nonlinearity in time series. Though the individual component AR processes are linear, the probabilistic mixing of these AR processes constitutes a nonlinear model. In a GMM, the distribution remains invariant to the past samples due to the static nature of the model. For MAR, the conditional distribution given past data varies with time. This model is capable of modeling both the conditional means and variances. Thus, MAR can model time series that evolve nonlinearly.

To better understand the efficacy of the MAR-HMM model, we evaluated its performance on two simple pattern recognition tasks. The first task represents data with known nonlinearities. The MAR, with just 2 components and 8 parameters, achieved 100% classification accuracy using only static features. The GMM approach using only static features was unable to do much better than a random guess strategy since the two classes have similar static marginal distribution. With the inclusion of delta coefficients, the GMM performance increases significantly, but even in this case it achieved only 85% accuracy with 28 parameters. Though delta features capture some amount of dynamic information in the features, it is still only a linear approximation, and we cannot capture their nonlinear evolution with just GMMs.

The second task was a simple phone classification task. To test the efficacy of MAR-HMMs in speech modeling, we made 16 kHz recordings of three distinct phones – "aa" (vowel), "m" (nasal), and "sh" (sibilant). For each phone and for each speaker, 35 recordings were made to serve as training database, while another 15 were reserved for testing. Silence was removed so that we could focus on the ability of the approach to model speech.

We evaluated the performances of 2-, 4-, 8-, and 16 mixture GMM-HMM and MAR-HMM with the 13 dimensional static MFCC features. For an equal number of parameters, MAR outperformed GMM significantly. For instance, MAR-HMM achieved a phone classification accuracy of 94.4% with only 320 parameters while a GMM system using 432 parameters could only achieve a 93.3%.

To determine whether MAR is more effective at exploiting dynamics than what GMM can achieve using dynamic features, we also performed another experiment with 39-dimensional features containing both static as well as velocity and acceleration coefficients. In this case, the results were not conclusive. While MAR-HMM showed an accuracy rate of 97.8% with 472 parameters, GMM-HMM attained only 96.7% accuracy with 632 parameters.

Unfortunately, the performance of MAR-HMM saturated with an increase in the number of parameters. For example, MAR-HMM at 1888 parameters achieved only 98.9% accuracy while GMM-HMM achieved 100% with 1264 parameters. We suspect that this could be due to the fact that our parameter estimation and likelihood computation procedures assume that the features are independent. It is well-known that the static MFCC features are uncorrelated (at least, theoretically), but obviously the delta features are correlated with the static ones. While this should also cause problems for GMM, the problem is more acute for MAR because in this case, unlike GMMs, we employ the past history explicitly.

This work is described in detail in a paper that will be presented at INTERSPEECH'2008 [6].

C. Linear Dynamic Models for Speech Recognition

Last year we proposed Linear Dynamic Models (LDMs) as an alternative to hidden Markov models (HMMs) for robust speech recognition in noisy environments. HMMs in speech recognition typically utilize a diagonal covariance matrix assumption in which correlations between feature vectors for adjacent frames are ignored. LDMs use a state space-like formulation that explicitly models the evolution of hidden states using an autoregressive process. This smoothed trajectory model allows the system to better track speech dynamics in noisy environments.

This year we refined this model and completed its evaluation on a complex speech recognition task – the Aurora 4 task. This task is attractive because it contains speech degraded by a variety of noise condictions (digitally) and it removes the language model problem by using a closed-set 5K language model. An HMM baseline system was used to generate alignments at the phone level. Each phone instance was treated as one segment. A total of 40 LDM phone models, one classifier per model, were used to cover the pronunciations. Each classifier was trained using the segmental features derived from 13-dimensional frame-level feature vectors comprised of 12 cepstral coefficients and absolute energy. The full training set has as many as 30k training examples per classifier. Each phone-level classifier was trained as a one-vs-all classifier. The classifiers were used to predict the probability of an acoustic segment.

The HMM system achieves up to 46.9% and 36.8% accuracy for the clean evaluation data and noisy evaluation data respectively. The LDM classifiers achieved superior performance to the 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.

We are currently developing a HMM/LDM hybrid decoder architecture to model the frame correlation using LDMs as well as utilizing HMMs techniques for phone segment alignment. Preliminary experiments will be presented on the Alphadigits (AD) and Resource Management (RM) speech corpora.

This HMM/LDM hybrid decoder architecture will be a good evaluation of LDMs on continuous speech recognition tasks, and can be compared to other hybrid decoders we have developed that utilize other nonlinear statistical models (e.g., support vector machines and relevance vector machines).

D. Additional Comments

Progress in the final year of this project was slowed somewhat because the two PhD graduate students working on the project had opportunities to do internships. Mr. Sundar Srinivasan spent Summer'2008 at Motorola's research lab working on rapid adaptation techniques in speech recognition. Mr. Tao Mao spent the summer with Intel's speech research group working on methods for speeding up speech recognition on Intel processors. He will continue there through December 2008. These students were identified by these employers based on their work on this project. For each of them, it was their first chance to do an internship during their graduate studies, so the PIs felt it was worthwhile to let them pursue these opportunities. Each had a very productive summer.

As a result, we applied and received an extension of the project through December 2008. This will allow Mr. Srinivasan to present his work at INTERSPEECH, and to complete some experiments on larger-scale speech processing tasks. He is expected to complete his disseration by August 2009.

Mr. Tao will complete his graduate studies as a graudate teaching assistant in the ECE Department at MS State. This will allow him to complete his dissertation on the work previously described. He is expected to graduate by December 2009.

The final report for this project will be delivered by the end of the 2008 calendar year, when we expect to have expended all of the funds.

E. References

[1] D. May, *Nonlinear Dynamic Invariants For Continuous Speech Recognition*, M.S. Thesis, Department of Electrical and Computer Engineering, Mississippi State University, May 2008.

[2] D. May, T. Ma, S. Srinivasan, G. Lazarou and J. Picone, "Continuous Speech Recognition Using Nonlinear Dynamic Invariants," submitted to INTERSPEECH, Brisbane, Australia, September 2008.

[3] D. May, S. Srinivasan, T. Ma, G. Lazarou and J. Picone, "Continuous Speech Recognition Using Nonlinear Dynamic Invariants," submitted to the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, USA, June 2008.

[4] D. May, S. Srinivasan, T. Ma and J. Picone, "Continuous Speech Recognition Using Nonlinear Dynamical Invariants," submitted to the International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, Nevada, USA, March 2008.

[5] 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.

[6] S. Srinivasan, T. Ma, D. May, G. Lazarou and J. Picone, "Nonlinear Mixture Autoregressive Hidden Markov Models For Speech Recognition," to be presented at INTERSPEECH, Brisbane, Australia, September 2008.

[7] T. Ma, S. Srinivasan, D. May, G. Lazarou and J. Picone, "Robust Speech Recognition Using Linear Dynamic Models," submitted to INTERSPEECH, Brisbane, Australia, September 2008.