Nonlinear Mixture Autoregressive Hidden Markov Models For Speech Recognition

Sundar Srinivasan[1], Tao Ma[1], Daniel May[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
{ss754, tm334, dom5, glaz, picone}@ece.msstate.edu

Abstract

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 paper, a nonlinear mixture autoregressive model is used to model state output distributions (MAR-HMM). Estimation of model parameters is extended to handle vector features. MAR-HMMs are shown to provide superior performance to comparable Gaussian mixture model-based HMMs (GMM-HMM) with lower complexity on two pilot classification tasks.

Index Terms: mixture autoregressive models, speech recognition, nonlinear statistical modeling

1. Introduction

The majority of speech recognition systems employ hidden Markov models (HMMs) in which each state is assumed to emit observations modeled by a Gaussian mixture distribution [1]. We refer to this traditional approach as GMM-HMM. These systems typically employ a standard feature vector containing absolute spectral information (e.g., mel frequency-scaled cepstral coefficients and energy) and the first and second derivatives of these features. This popular approach to feature extraction is often referred to as MFCCs.

In spite of the popularity of this model, several deficiencies have been documented. One particular drawback of GMM-HMMs that is the focus of this work is the assumption of conditional time-independence of speech features, i.e., for any state, the probability of observing a feature is assumed to be independent of previous frame features. One popular approach to overcoming this deficiency is to include derivative information in the feature vector, as is done in the standard MFCC feature vector. These derivatives capture information about the dynamics of the speech signal. It is well known that the use of dynamic information in MFCCs significantly enhances recognition performance.

However, the derivatives of the cepstral features are only a linear approximation of the actual dynamics of the static features. Recent work suggests that speech signals have nonlinearities that could contain relevant information for speech recognition [2][3]. A common approach to exploit the nonlinear speech information is to explicitly quantify the degree of nonlinearity using nonlinear invariants such as Lyapunov exponents and fractal dimensions, and to include these with MFCCs as features to be modeled by GMM-HMM. One drawback of this approach arises from the difficulty of estimating these invariant features reliably for short speech signal frames [3].

A more serious shortcoming in our view is that these invariants are only crude quantifiers of nonlinearity and cannot be used to model the nonlinear evolution itself -- two signals can have different nonlinear dynamics, yet have the same "amount" of nonlinearity as quantified by the invariants. For example, two periodic signals could be quite different from each other, yet both would be assigned a maximal Lyapunov exponent value of zero. Perhaps because of this difficulty, the method of using invariants as features has not improved performance of speech recognition systems significantly. At best, they have been shown to be of use for broad phone class recognition [3][4].

In this work, we apply a nonlinear mixture autoregressive hidden Markov model (MAR-HMM) to capture the nonlinear dynamics in speech MFCCs [5] as shown in Figure 1. In addition to capturing the dynamic information directly from the features, it can also simultaneously model the static information. One of the earliest applications of autoregressive-HMMs (AR-HMMs) considered an autoregressive filter to model state observations in a 5-state HMM for speaker verification [6]. A more recent investigation of AR-HMMs [7] used a switching autoregressive process to capture signal correlations during state transitions. Results on speech recognition showed that at best their model was only comparable to cepstral-based GMM-HMM recognizer.

A more sophisticated model introduced in [5] considers a mixture of autoregressive filters for the observation model. The MAR-HMM model we develop here is a generalized version of that model, and has also been extended to handle vector observations, so that we can operate on the speech feature vector stream rather than speech samples. The latter is more common in the nonlinear dynamics literature.

This paper is organized as follows. In Section 2 we introduce MAR and note some relevant properties of this model. We also briefly discuss the problem of parameter estimation using the Expectation Maximization (EM) algorithm in the HMM framework. An extension of the scalar MAR to the vector case, and also a discussion of a strategy to initialize the EM iterations are included. We show how MAR-HMM is a generalization of these other models. In Section 3 we present some preliminary results on the application of MAR-HMM to two simple classification tasks, one involving a phone classification experiment. We conclude the paper in Section 4 with a discussion of on-going research.

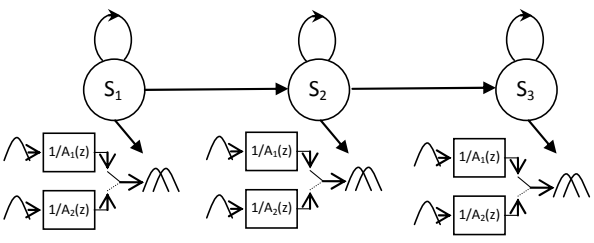

Figure 1: *An overview of the MAR-HMM approach.*

2. Mixture Autoregressive HMMs

A mixture autoregressive process (MAR) of order p with m components, $X=\{x[n]\}$, is defined as [5]:

$$x[n]=\begin{cases} a_{1,0}+\sum_{i=1}^{p} a_{1,i}x[n-i]+\varepsilon_1[n] & \text{w.p. } w_1 \\ a_{2,0}+\sum_{i=1}^{p} a_{2,i}x[n-i]+\varepsilon_2[n] & \text{w.p. } w_2 \\ \vdots & \vdots \\ a_{m,0}+\sum_{i=1}^{p} a_{m,i}x[n-i]+\varepsilon_m[n] & \text{w.p. } w_m \end{cases} \quad (1)$$

where, ε_i is a zero-mean Gaussian random process with a variance of σ_j^2, "w.p. w_i" denotes "with probability w_i" and the weights, w_i sum to 1. For i = 1, …, m, $\{a_i\}$ are the linear predictor coefficients, and $a_{i,0}$ are the component means. It is apparent that an m-mixture MAR process is the weighted sum of m Gaussian autoregressive processes.

One convenient way of viewing this model is as a process in which each data sample at any one point in time is generated from one of the component AR mixture process chosen randomly according to its weight w_i. It is easy to draw parallels between MAR and GMM models. In particular, MAR can be viewed as a generalization of GMM that models each component as a sum of the output of an autoregressive filter with a specified mean. It should be noted that MAR with all the component orders set to zero reduces to a GMM. This similarity between the two makes it straightforward to replace GMM with MAR in HMMs.

One property of MAR that is of particular relevance here 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. This property becomes important in speech processing in the light of recent work on nonlinear processing of speech [2][3]. Some other properties of MAR including conditions required for the process to be stationary are derived in [5].

2.1. Parameter Estimation Using EM

Similar to GMM training, maximum likelihood estimates for MAR parameters can be calculated using the Expectation Maximization (EM) algorithm [9]. Given the order, p, the parameter set for each of the m components of a MAR model consists of p+1 predictor coefficients (including the mean), the error variance, and mixing weight:

$$\theta_l = \{a_{l,0}, a_{l,1}, \ldots, a_{l,p}, \sigma_l, w_l\} \quad l = 1, \ldots, m \quad (2)$$

To estimate these parameters, we first need an initial guess for these parameters and then we iterate with EM to successively refine the estimates. An initialization strategy that we found to work reasonably well was to first train a GMM with the same number of mixtures and then set each component of the MAR to have the same mean, variance, and weight as the GMM model. Without loss of generality, we initialize the predictor coefficients of MAR to zero.

These initial parameters can be then refined recursively using an E-step [8]:

$$\gamma_l[n] = \frac{w_l\, p_l(x[n]\,|\,\theta)}{\sum_{k=1}^{m} w_k\, p_k(x[n]\,|\,\theta)} \quad (3)$$

where

$$p_l(x[n]\,|\,\theta) \propto \frac{1}{\sigma_l} e^{\frac{-1}{2\sigma_l}(x[n]-a_{l,0}-\sum_{i=1}^{m} a_{l,i}\,x[n-i])^2} . \quad (4)$$

is the probability that sample was generated from component l at time instant n. The corresponding M-step is given by:

$$\hat{w}_l = \frac{\sum_{n=p+1}^{N}\gamma_l[n]}{N-p} \quad (5)$$

$$\hat{A}_l = R_l^{-1}\, r_l \quad (6)$$

$$\hat{\sigma}_l^2 = \frac{\sum_{n=p+1}^{N}\gamma_l[n](x[n]-\hat{a}_{l,0}-\sum_{i=1}^{m}\hat{a}_{l,i}\,x[n-i])^2}{\sum_{n=p+1}^{N}\gamma_l[n]} \quad (7)$$

where

$$R_l = \sum_{n=p+1}^{N}\gamma_l[n]\,X_{n-1}\,X_{n-1}^T \quad (8)$$

$$r_l = \sum_{n=p+1}^{N}\gamma_l[n]\,X_{n-1}\,x[n] \quad (9)$$

and

$$X_{n-1} = [1\; x[n-1]\; x[n-2] \ldots x[n-p]]^T . \quad (10)$$

Note that the solution for the predictor coefficients is a weighted form of the standard covariance method for estimating linear prediction coefficients.

Also observe that in these estimation equations, the updates for variances require updated values of the predictor coefficients and the mean for that component. A straightforward implementation would thus require all data to be buffered so that the predictor coefficients and the mean could be computed in a first pass, and then another pass through all the data can be made to compute the variance using these updates. This is impractical for very large data sets.

An alternative strategy is to identify and accumulate the required statistics on a sample basis and then update the parameters only at the end of data. Fortunately in this case, the required statistics for the variance update turn out to be only R_l and r_l,which we already compute while updating the prediction coefficients and the mean, and also weighted variance, which is easily computed.

2.2. Vector Extension of the MAR Model

MAR has previously only been applied to a scalar time series. We can extend this to vector time series, but to make this extension more tractable, we will assume that the component dimensions of the vector are uncorrelated. This is equivalent to the assumption of diagonal covariance that is commonly made for GMM-HMMs. In this case, each MAR component is a weighted mixture of vector autoregressive processes, but due to the independence assumption, every dimension still has an individual set of predictor coefficients and variance. However there is only a single weight for each component. Thus, only the update equation for w_l needs to be modified.

For a D-dimensional vector, the update equation for w_l becomes:

$$\hat{w}_l = \frac{\sum_{d=1}^{D}\sum_{n=p+1}^{N}\gamma_{l,d}[n]}{\sum_{k=1}^{m}\sum_{d=1}^{D}\sum_{n=p+1}^{N}\gamma_{k,d}[n]}. \quad (11)$$

where $\gamma_{l,d}$ and $\gamma_{k,d}$ are similar to $\gamma_l[n]$ in (3) except that these are now computed for each element in the D-dimensional vector. In the context of HMMs, the $\gamma_{k,d}$ values would be incorporated along with the language model probabilities in the forward-backward recursions during training and in the likelihood computations during decoding as well. Since the formulation of MAR parallels that of GMM, the higher levels in the typical HMM hierarchy for speech recognition are unchanged.

2.3. Relationship to Other AR-HMM Models

One of the first applications of autoregressive HMMs in speech processing assumed an autoregressive (AR) model for each state, so that the short term correlations in the speech signal and known linguistic properties of sound combinations could be modeled by the state transitions [6]. This model was effectively a single-mixture component MAR-HMM.

The next step was the introduction of mixture autoregressive HMMs in [6]. This work applied a weighted mixture of AR filters to model observations at each state. While this appears to be very similar to the MAR-HMM developed in this paper, this approach had two major shortcomings. The model in [6] assumed that all AR components had the same variance, and that each was zero mean. This is equivalent to constraining the MAR model to have zero means and equal variance. In this respect the MAR-HMM considered in our work is more general than the autoregressive models previously applied to speech.

A variant of the original AR-HMM, using switching autoregressive process was considered in [7]. In this approach, the signal correlations during HMM state transitions were also modeled by the switching process. However, this model again was restricted to a single component AR, and thus it too is equivalent to a single-component MAR. Moreover, these variants of AR-HMMs considered only scalar speech time series as observations. Our extensions to vector time series are crucial to application of these models to speech recognition.

3. Pilot Experiments

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 second task is a simple phone classification task.

3.1. Two-Class Problem

The MAR-HMM approach, like GMM-HMMs, can perform classification using a maximum likelihood approach. The log likelihood of data given a set of MAR-HMM model parameters is used to score each model and the class with the maximum score is chosen. A two-class classification problem was designed where data are randomly generated randomly according to the following MAR parameters:

$$\Theta_1 : \begin{cases} p = 1, m = 2, w_1 = 0.4, w_2 = 0.6, \\ a_{1,0} = -1, a_{1,1} = 0.2, a_{2,0} = 1, a_{2,0} = 0.2,. \\ \sigma_1 = 0.25, \sigma_2 = 0.2 \end{cases} \quad (12)$$

$$\Theta_2 : \begin{cases} p = 0, m = 4, \\ w_1 = 0.2, w_2 = 0.2, w_3 = 0.3, w_4 = 0.3 \\ a_{1,0} = -1.03, a_{2,0} = -0.86, \\ a_{3,0} = 1.13, a_{4,0} = 0.98 \\ \sigma_1 = 0.3263, \sigma_2 = 0.2906, \\ \sigma_3 = 0.2598, \sigma_4 = 0.2894 \end{cases} . \quad (13)$$

where Θ_1 and Θ_2 correspond to classes 1 and 2, respectively.

For this example we chose the parameters for class 2 such that the marginal distribution is about the same as that of the first class, but it lacked the dependence on past samples unlike class 1. Hence the data for class 2 follows only a GMM distribution. This was done to demonstrate a case where GMM would be unable to achieve good classification due to its ability to capture the dynamics in the model.

The results of these experiments, along with the number of parameters for each model, are shown in Table 1. In addition to listing accuracy, the numbers of parameters for each model are shown. Since in this case we knew that the distribution can have a maximum of 4 modes, we use only 2- and 4-mixture models. It can be observed that MAR, with just 2 components and 8 parameters can achieve 100% classification accuracy using only static features. The GMM approach using only static features is unable to do much better than a random guess strategy since the two classes have similar static marginal distribution. This demonstrates MAR-HMM's ability to learn dynamic information.

With the inclusion of delta coefficients, the GMM performance increases significantly, but even in this case it achieves 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. From the above, it is clear that at least some dynamic information is better modeled using MAR-HMM.

3.2. Phone Classification Experiments

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.

From these sound files, the static feature database comprising of 13 MFCC coefficients (including energy) was extracted for each frame 10 ms in duration using a window duration of 25 ms. In addition, we also created a database containing 39-dimensional features by concatenating Δ and $\Delta\Delta$ MFCC features.

First we evaluated the performances of 2-, 4-, 8-, and 16-mixture GMM-HMM and MAR-HMM with the 13-dimensional static MFCC features. The results are shown in Table 2. From this, it can be seen that for equal number of parameters MAR outperforms GMM significantly. For

Table 1: *Classification (% accuracy) results for synthetic data.*

# mixts.	GMM Static only	MAR Static only	GMM static+ Δ+$\Delta\Delta$	MAR static+ Δ+$\Delta\Delta$
2	47.5 (6)	100.0 (8)	82.5 (14)	100.0 (20)
4	52.5 (12)	100.0 (16)	85.0 (28)	100.0 (40)

Table 2: *Sustained phone classification (% accuracy) results with MAR and GMM using 13 MFCC features (the numbers of parameters are shown in parentheses).*

# mixtures	GMM	MAR
2	77.8 (54)	83.3 (80)
4	86.7 (108)	90.0 (160)
8	91.1 (216)	94.4 (320)
16	93.3 (432)	95.6 (640)

instance, MAR-HMM achieves a phone classification accuracy of 94.4% with only 320 parameters while a GMM system using 432 parameters could achieve only 93.3%. This clearly shows that MAR can exploit the dynamical information to better model the evolution of MFCCs for phones that GMM is unable to model.

To determine whether MAR is more effective at exploiting dynamics than what GMM can achieve using dynamic features, we also perform another experiment with 39-dimensional features containing both static as well as velocity and acceleration coefficients. The results are tabulated in Table 3. In this case, the results are not conclusive. While MAR-HMM shows an accuracy rate of 97.8% with 472 parameters, GMM-HMM attains only 96.7% accuracy with 632 parameters.

However, the performance of MAR-HMM saturates with an increase in the number of parameters. For example, MAR-HMM at 1888 parameters achieves only 98.9% accuracy while GMM-HMM achieves 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 acutely so for MAR because in this case, unlike GMMs, we employ the past history explicitly.

4. Summary

In this paper, we presented a novel nonlinear mixture autoregressive HMM (MAR-HMM) for speech recognition. We described techniques for estimating the parameters of this model based on the EM algorithm. We also compared this model to other autoregressive HMM models used in speech recognition. We presented preliminary results on two small pattern recognition tasks. We showed that MAR-HMM can outperform GMM-HMM if the patterns have a nonlinear evolution function. We applied this model to a phone classification experiment and showed that MAR-HMMs can exploit the dynamic information in speech to achieve significantly higher phone classification accuracy.

When combining dynamic features, the results for MAR-HMM were mixed. MAR-HMM delivered better performance for smaller numbers of mixtures, resulting in a reduction in the number of parameters. However, GMMs performed better for larger numbers of mixtures. We conjecture that this is due to correlation between static and delta features that invalidates the assumption of independence both during parameter estimation and likelihood computation.

We are in the process of applying MAR-HMM to the large vocabulary continuous speech recognition problem using the Aurora [10] corpus. An overarching goal of our work on nonlinear statistical modeling has been the belief that nonlinear systems can be more robust on problems involving previously unseen channel conditions. The Aurora task is an excellent opportunity to evaluate this hypothesis because it contains a number of mismatched evaluation conditions.

Table 3: *Sustained phone classification (% accuracy) results with MAR and GMM using static+Δ+ΔΔ MFCC features (the numbers of parameters are shown in parentheses).*

# mixtures	GMM	MAR
2	92.2 (158)	94.4 (236)
4	94.4 (316)	97.8 (472)
8	96.7 (632)	97.8 (944)
16	100 (1264)	98.9 (1888)

One of the problems we found with the MAR-HMM approach is that it is difficult to learn parameters reliably from continuous speech in an unsupervised manner. Hence, we are currently building a GMM-HMM/MAR-HMM hybrid system, where a first pass with GMM-HMM would be used to provide phone boundaries after which MAR-HMM would be used to process each phone segment. We also plan to use phone-dependent decorrelation of features as a pre-processing step to improve the reliability of estimates both for MAR-HMMs and GMM-HMMs.

5. 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.

6. References

[1] Huang, X., Acero, A. and Hon, H., *Spoken Language Processing: A guide to Theory, Algorithm, and System Development*, Prentice-Hall, 2001.

[2] 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.

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

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

[5] Wong, C. S. and Li, W. K., "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society*, vol. 62, no. 1, pp. 95-115, February 2000.

[6] Juang, B. H., and Rabiner, L. R., "Mixture Autoregressive Hidden Markov Models for Speech Signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 6, pp. 1404-1413, December 1985.

[7] Ephraim, Y. and Roberts, W. J. J., "Revisiting Autoregressive Hidden Markov Modeling of Speech Signals," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 166-169, February 2005.

[8] Poritz, A. B., "Linear Predictive Hidden Markov Models," *Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech*, Princeton, New Jersey, USA, pp. 88-142, October 1980.

[9] Dempster, A., Laird, N., and Rubin, D., "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1-38, February 1977.

[10] Parihar, N. and Picone, J., "An Analysis of the Aurora Large Vocabulary Evaluation," *Proceedings of the European Conference on Speech Communication and Technology*, pp. 337-340, Geneva, Switzerland, September 2003.