A Nonlinear Mixture Autoregressive Model for Speaker Recognition

S. Srinivasan1, T. Ma1, D. May1, G. Lazarou2 and J. Picone1

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

Abstract

Gaussian mixture models (GMMs) are a very successful method for modeling the distribution of speaker features. In this approach, the dynamics of the speech spectrum are typically encapsulated in the feature vector through the use of derivatives. This model 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 (MixAR) is used to model speaker features. Experiments show that MixAR performs better than a GMM when the signal contains strong evidence of nonlinear behavior. On the 2001 NIST Speaker Recognition Evaluation Corpus, MixAR is shown to lower the equal error rate by 10.6% relative and uses significantly fewer parameters than GMM.

**Index Terms**: mixture autoregressive models, speaker verification, nonlinear statistical modeling

# Introduction

The majority of speaker recognition systems employ a Gaussian mixture model (GMM) to capture modalities in the distribution of features for a speaker . 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 MFCC features.

Despite the widespread popularity of this model, several well-known deficiencies exist [2]. One particular drawback of a GMM 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 speaker recognition performance [3].

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 [4][5]. A common approach to exploit nonlinearities in speech is to explicitly quantify the degree of nonlinearity using nonlinear invariants such as Lyapunov exponents, and to concatenate these with the MFCC features [5]. However, this approach has produced only modest success on limited tasks, and has not fundamentally improved the robustness of the technology in extremely harsh or mismatched operating environments [5].

In this work, we apply a nonlinear mixture autoregressive model, known as MixAR [7], to capture the nonlinear dynamics in the speech feature vector, as shown in . Since the MixAR model captures static as well as dynamic information it can be used to capture all relevant information using only static features. Our goal is to reduce the number of parameters necessary to obtain the same or better performance as compared to a GMM. In addition, we hope that the MixAR model would capture information in nonlinear dynamics of speech features that GMM cannot model, and improve overall speaker recognition performance.



Figure : *An overview of the MixAR approach.*

Previous work [8][9] on mixture autoregressive modeling for speech has been in the context of hidden Markov models for speech recognition. 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 [8]. A more recent investigation of AR-HMMs [9] 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 an MFCC-based HMM using a GMM observation model. Another model considered speech features as a GMM white noise process filtered through an autoregressive signal [10].

A more sophisticated model introduced in [11] considers a mixture of autoregressive filters (MAR) for the observation model. Our earlier work [6] considered this model for phone classification. MixAR [7] is a generalization of MAR, where the mixture weights are allowed to be time-varying and data-dependent. All earlier applications of MixAR and related models [12] have been in the context of time-series prediction. In this work, we apply the MixAR model to feature vectors in a speaker recognition task.

This paper is organized as follows. In Section  we describe the MixAR model and note some relevant properties. We also briefly discuss the problem of parameter estimation and optimization using the Expectation Maximization (EM) algorithm. In Section 3 we present preliminary results on the application of MixAR to a 2-way classification task involving synthetic speech‑like data in which the amount of nonlinearity is under parametric control. In Section 4 we present our results on speaker verification tasks with 2001 NIST Speaker Recognition Evaluation (SRE) Corpus [13]. We conclude the paper in Section 5 with a summary and discussion of ongoing research.

# Mixture Autoregressive Models

A mixture autoregressive process (MixAR) of order *p*with *m* components, *X*={*x*[*n*]}, is defined as [7]:

|  |  |  |
| --- | --- | --- |
|  |  |  |

where, εi is a zero-mean Gaussian random process with a variance of σj2, “w.p.” denotes “with probability” and the gating weights, *W*i sum to 1. The linear prediction coefficients, {*a*i}, represent the dynamic model, where *a*i,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.

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 processes chosen randomly according to its weight *W*i. It is easy to find parallels between the MixAR and GMM models. In particular, MixAR 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, and with mixture weights determined by a gating system similar to a mixture of experts. It should be noted that with the component orders and *gi* set to zero, MixAR, reduces to the familiar GMM. This similarity between the two makes it straightforward to replace GMM with MixAR for speaker recognition.

One property of MixAR that is of particular relevance here is the ability of MixAR 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. Even when the mixture weights are fixed, the model reduces to MAR, which is still nonlinear. The addition of a gating system layer for weight generation increases the flexibility of the model even further, allowing us to model distributions as a function of past data.

In a GMM, the distribution remains invariant to the past samples due to the static nature of the model. For MixAR, the conditional distribution given past data varies with time. This model is capable of modeling both the conditional means and variances. Thus, MixAR 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 [4][5].

Some other properties of MixAR, including a mathematically rigorous proof of the ability of MixARs to arbitrarily closely model stochastic processes are derived in . Note that in the original formulation, both the gate and prediction orders were constrained to be equal. In this paper, we restrict our use of MixAR order to one to avoid difficulties during parameter estimation.

## Parameter Estimation Using EM

Similar to GMM training, maximum likelihood estimates for MixAR prediction and variance parameters can be calculated using the Expectation Maximization (EM) algorithm [13]. 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:

|  |  |  |
| --- | --- | --- |
|  |  |  |

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 MixAR to have the same mean, variance, and weight as the GMM model. We initialize the predictor coefficients and the data-dependency gating coefficients, {*Ai*} of MixAR to zero.

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

|  |  |  |
| --- | --- | --- |
|  |  |  |

where

|  |  |  |
| --- | --- | --- |
|  |  |  |

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

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

where

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

and

|  |  |  |
| --- | --- | --- |
|  |  |  |

Refer to comments on estimation of predictor coefficients and variances for MAR in [6] for further details.

However, a complication arises with respect to the estimation of gating coefficients. There is no closed-form solution for these, and hence a Newton gradient-ascent approach must be used:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

where *Q* denotes the likelihood of the MixAR model for the training data. β and Δ are design parameters to be chosen empirically. In our experiments, we found that fixing Δ = 0.01 and running 10 iterations each with β = 0.9, β = 0.5, and β = 0.2 in succession provided a smooth and reasonably quick convergence.

Table : *Classification Error Rate (%) with 12 speech MFCC-like synthetic features for GMM and MixAR Number of parameters in each case is in paranthesis. (\*: For this case, GMM performed better with only static features, and this value is stated)*.

|  |  |  |
| --- | --- | --- |
| *α* | GMM-8mix. Static+∆ | MixAR-4-mix. Static |
| 0.0\* | 1.5 (288) | 1.5 (240) |
| 0.25 | 3.25 (576) | 3.5 (240) |
| 0.50 | 10.25 (576) | 6.25 (240) |
| 0.75 | 24.75 (576) | 9.75 (240) |
| 1.0 | 26.75 (576) | 13.75 (240) |

# Pilot Experiments

To better understand the efficacy of the MixAR model, we evaluated its performance on two pattern classification tasks. The first task represents generic data with known nonlinearities. The second task is a simple classification task with data for the two classes synthesized from models trained on true speaker data.

## Two-Way Classification with Synthetic Data

A simple 2-way classification experiment was designed to study the performance of MixAR and GMM. Two-dimensional data for the first class was generated using a linear dynamic system:

|  |  |  |
| --- | --- | --- |
|  |  |  |

Data for the second class was generated using the simple nonlinear equation:

|  |  |  |
| --- | --- | --- |
|  |  |  |

In both cases, ***E*** denotes an uncorrelated 2-D Gaussian (normal) random variable with a zero mean and unit variance.

For each class, the training data consisted of a sequence of 10,000 vectors, and evaluation data consisted of 100 segments of 200 feature vectors each (the log-likelihood of the entire segment was used to assign a segment to a class). The classification error results are stated in Table 1. Clearly, when using only static features, MixAR does much better than GMM if nonlinearities are present. The use of dynamic features enhances GMM performance considerably but still falls far short of MixAR's performance.

## Two-way Classification with Speech-like Data

In order to evaluate how well MixAR does as compared to GMM for speech-like signals, two speakers from the 2001 NIST SRE Corpus [13] were selected. A 3-state HMM with 4 Gaussian mixtures per state and a MixAR model with 4 mixtures were trained over 12 static MFCC coefficients for each speaker. For each class (speaker), two speech-like signals of 40,000 vectors were generated – a linear speech-like signal (***X1***) was synthesized from the HMM model, and a nonlinear speech-like signal (***X2***) was generated from the MixAR model. To simulate a range of signals with varying degrees of nonlinearity, the two signals were mixed with a mixing coefficient alpha:

|  |  |  |
| --- | --- | --- |
|  |  |  |

The first 20,000 vectors from each ***Xα*** were used as a training set while the remaining vectors were split into 200 segments of 100 vectors each for evaluation. The results are shown in Table 2.

From the table we can see that when the amount of nonlinearity is insignificant, GMM performs as well as MixAR. However, as the amount of nonlinearity in the signal increases, MixAR performs significantly better with just static features as compared to GMM with static+∆ features. This clearly demonstrates the superiority of MixAR when dynamics in the data are nonlinear.

# Speaker Verification Experiments

Next we applied the MixAR model to the 1-speaker detection task in the 2001 NIST SRE Corpus [13]. 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.

First we evaluated performance with and without delta features and energy for a fixed number of mixtures. The results are tabulated in Table 3. For GMM, substantial improvement is obtained using the delta features and marginal improvements were obtained using delta-delta features. For MixAR, the use of any delta features provides no measurable improvements. This clearly indicates that MixAR can extract all necessary information from only the static features.

Table 1: *Classification (% error) results for synthetic data (the numbers of parameters are shown in parentheses).*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # mix. | GMM Static | MixAR Static | GMM Static+∆ | MixAR Static+∆ |
| 0 | 36.0(12) | 6.5(20) | 10.0(24) | 5.5(40) |
| 4 | 35.5(24) | 6.0(40) | 11.5(48) | 4.5(80) |

Table 3: *Speaker recognition EER with NIST for different feature combinations (number of features in parenthesis).*

|  |  |  |
| --- | --- | --- |
| Features | GMM-16-mix. | MixAR-8-mix. |
| Static(12) | 22.1 | 19.1 |
| Static+E(13) | 33.1 | 41.1 |
| Static+Δ(24) | 20.6 | 20.4 |
| Static+Δ+ΔΔ(36) | 20.5 | 20.5 |

MixAR and GMM performance was then evaluated as a function of the number of mixtures. The detection error trade-off (DET) curves are shown in Figure 2. The EER results are shown in Table 4. 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.

# Summary

In this paper, we presented a novel nonlinear mixture autoregressive model (MixAR) for speaker recognition. We described techniques for estimating the parameters of this model based on the EM algorithm and ways to resolve some of the algorithmic issues in this procedure. We presented preliminary results on two small pattern recognition tasks. We showed that MixAR can outperform GMM if the patterns have a distinct nonlinear evolution function. We applied this model to a speaker verification task and showed that MixAR can exploit the dynamic information in speech to achieve appreciably better verification performance with significantly fewer parameters compared to a GMM.

The experiments with MixAR on speaker recognition have yielded very encouraging results. 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. We plan to next extend this work to a large vocabulary speaker independent speech recognition task.

Figure 2: *Speaker recognition DET curves with NIST.*

# 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

Table 4: *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) |

1. D. A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication*, vol. 17, no. 1-2, pp. 91-108, August 1995.
2. X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice-Hall, 2001.
3. M. Nosratighods, E. Ambikairajah, and J. Epps, “Speaker verification using a novel set of dynamic features,” *Proceedings of The 18th International Conference on Pattern Recognitio*n, pp. 266-269, Hong Kong, China, September 2006.
4. I. Kokkinos, and P. Maragos, “Nonlinear Speech Analysis using Models for Chaotic Systems,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1098-1109, November 2005.
5. D. May, *Nonlinear Dynamic Invariants For Continuous Speech Recognition*, M.S. Thesis, Department of Elect. and Comp. Eng., Mississippi State University, May 2008.
6. S. Srinivasan, T. Ma, D. May, G. Lazarou and J. Picone, "Nonlinear Mixture Autoregressive Hidden Markov Models For Speech Recognition," *Proceedings of the International Conference on Spoken Language Processing*, pp. 960-963, Brisbane, Australia, September 2008.
7. M. Zeevi, R. Meir, and R. Adler, “Nonlinear models for time series using mixtures of autoregressive models”,

Unpublished Technical Report, 1999, http://ie.technion.ac.il/~radler/mixar.pdf

1. B. H. Juang, and L. R. Rabiner, “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.
2. Y. Ephraim, and W. J. Roberts, “Revisiting Autoregressive Hidden Markov Modeling of Speech Signals,” *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 166-169, February 2005.
3. M. E. Ayadi, *Autoregressive models for text independent speaker identification in noisy environments*, PhD Thesis, Department of Elect. and Comp. Eng., University of Waterloo, September 2008.
4. 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.
5. A. X. Carvalho and M. A Tanner, “Modeling nonlinearities with mixtures-of-experts of time series models,” *International Journal of Mathematics and Mathematical Sciences*, vol. 2006, no. 9, pp. 1-22, May 2006.
6. National Institute of Standards and Technology,

“The 2001 NIST Speaker Recognition Evaluation,” http://www.nist.gov/speech/tests/spk/2001, 2001.

1. A. Dempster, N. Laird, and D. Rubin, “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.