[[1]](#footnote-1)

Robust Speaker Verification Using a Nonlinear Mixture Autoregressive Model (MixAR)

Sundararajan Srinivasan1, Tao Ma1, Georgios Lazarou2, and Joseph Picone3

*Abstract*—Gaussian Mixture Modeling (GMM) has been the most popular approach in speaker verification for over two decades. The inefficiencies of this model for nonlinear dynamics are well-documented. In this work, we applied a nonlinear mixture autoregressive model (MixAR) to the problem of speaker verification. Initial experiments with synthetic data showed that the performance of MixAR exceeded that for a GMM while using fewer parameters, particularly in noisy environments. Experiments were conducted using three kinds of noise (white, car, and babble) artificially added to the clean TIMIT data. MixAR using static features and fewer than half of the parameters of a comparable GMM consistently outperformed GMM. Experiments on NTIMIT verified that MixAR gives comparable performance to a GMM and yet uses half the parameters.

*Index Terms*—mixture autoregressive, robust speaker verification, Gaussian mixture.

EDICS: SPE-RECO

# INTRODUCTION

T

he goal in speaker verification is to accept or reject the identity claim made by a speaker. This is widely used in a variety of applications ranging from secured access and surveillance to multimodal verification. A challenge for statistical modeling in speaker verification is to accurately and efficiently represent the probability distribution of speaker features so that even similar sounding speakers can be distinguished. The majority of speaker recognition systems today utilize Gaussian Mixture Models (GMMs) either entirely or as part of a hybrid model [1].

There are two well-known drawbacks of the GMM model. The first involves redundancy – there is obviously statistical dependence between absolute, static, and acceleration feature coefficients, but building GMMs over the complete concatenated vector does not take this redundancy into account. Hence, we tend to use more parameters than might be necessary. The second more serious drawback, which is the focus of this work, is the implicit assumption of linearity in the MFCC dynamics. The derivatives of the cepstral features are only a linear approximation of the actual dynamics of the static features. However, a survey of studies on the subject shows that the speech signal contains significant nonlinear information, and using only derivative features to represent speech MFCC dynamics with GMM modeling is tantamount to discarding any nonlinear information present in the signal .

An obvious fix to this problem is to add features that can represent the nonlinear dynamic information. However, adding nonlinear invariants as features has not improved the robustness of speech and speaker recognition technologies in harsh or mismatched environments. Three reasons can be attributed to this failure. First, it is difficult to estimate invariants reliably from speech, resulting in parameter estimation algorithms that need to be extensively tuned. Second, these estimation algorithms typically require an acoustic event to have a long duration . This gravely undermines the applicability of invariant features for a short-term stationary signal like speech. Even if it was somehow possible to estimate the invariants accurately, there is a third and more fundamental problem that invariants only quantify the degree of nonlinearity and do not characterize the nature of the dynamics completely.

The primary goal of this work is to approach the information representation problem at the modeling level using a nonlinear mixture autoregressive model (MixAR) , thereby accounting for the nonlinear dynamics of speech in the base model and minimizing the dimensionality of the feature space. To our knowledge, this is the first attempt at using this model for speaker verification. Previous work on mixture autoregressive modeling for speech has been in the context of hidden Markov models for speech recognition . A more recent investigation of AR-HMMs  used a switching autoregressive process to capture signal correlations during state transitions. Results on speech recognition showed that at best the 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 for speaker identification .

(a)



(b)



Fig. 1. An overview of the GMM (a) and MixAR (b) approaches.

A more sophisticated model introduced in  considers a mixture of autoregressive filters (MAR) for the observation model. Our earlier work  considered this model for phone classification. MixAR is a generalization of MAR, where the mixture weights are allowed to be time-varying and data-dependent. In this work, we apply the MixAR model to feature vectors in a speaker recognition task.

The rest of the paper is organized as follows: Section II defines the MixAR model and explains some of the relevant properties. Results of experiments using synthetic data are included in Section III and our speaker verification experiments with real speech data are in Section IV. Finally, in Section V we present our conclusions and discuss future work.

# MIXAR MODEL Overview

## Definitions and Properties

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

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

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* and *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 insightful 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 (see Fig. 1). One property of MixAR that is of particular relevance here is the ability of MixAR to model nonlinear time series [7]. 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.

Several other properties of MixAR, including a mathematically rigorous proof of the asymptotic performance of a MixAR model for stochastic processes are derived in . The problem of parameter estimation is also discussed in . Note that in the original formulation, both the gate and prediction orders were constrained to be equal. In this paper, we restrict ourselves to MixAR models of order one to avoid difficulties during parameter estimation. We used the ISIP public domain speech recognition software  to implement the MixAR model as well as integrate it into an existing speaker verification system.

# Synthetic Data Experiments

## 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  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 (e.g., a 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 α:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

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 of the classification experiments with this data are shown in Table I. 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. These results validated the basic model and provided motivation to do further testing on more realistic data.

## Speaker Verification Experiments with Synthetic Data

Since our goal is to study speaker verification, we usedvthe development database in the 1-speaker detection task of the 2001 NIST SRE Corpus  for our next set of experiments. This development database is small enough to make it manageable and yet large enough to provide a reliable estimate of the performance. All 60 speakers in the training set were used. Each training utterance was about 2 minutes long. Static (13 MFCCs), delta (26 MFCCs) and delta-delta (39 MFCCs) features were extracted.

Two types of clean data were synthesized. For the first type, a 10-state HMM with 4-Gaussians per state was trained for each utterance for each MFCC. For the second type, a 32-mixture MixAR model of prediction order 1 was trained for each utterance and for each MFCC. For each of the models trained, new training data of about 30,000 frames per speaker and evaluation data of 20 utterances with about 200 frames for each utterance per speaker were generated according to that model.

Similarly, two types of noisy data were generated. First, the clean training utterances from the development data were corrupted with car noise to achieve an SNR of 5 dB. This approach followed a methodology previously used to generate the AURORA I database . The remainder of the steps to yield the two types of noisy data were the same as those for the clean case. The goal of creating data in this way was to simulate 4 different test conditions: clean/linear, clean/nonlinear, noisy/linear and noisy/nonlinear.

TABLE I

Classification Error Rate (%) with synthetic data

|  |  |  |
| --- | --- | --- |
| *α (nonlinearity)* | GMM-8mix. Static+∆ | MixAR-4-mix.  (Static only) |
| 0.00\* | 1.50 (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) |
| 1.00 | 26.75 (576) | 13.75 (240) |

\* For this case, GMM performed better with only static features.

Using the synthesized training data, both GMMs and prediction order-1 MixARs were trained for each speaker under each condition. Then the corresponding synthesized evaluation data were used for evaluating speaker verification performance. For the clean case, there was little difference in performance between GMM and MixAR. For noisy evaluation data at 5 dB SNR, there was not much variation in performance between GMM and MixAR for HMM-generated data.

However, for the data generated from the nonlinear MixAR model and with the addition of noise, the MixAR model showed a significant improvement in performance using far fewer parameters. This is evident from the DET plot in Fig. 2. These results provide support to the hypothesis that when there are significant nonlinearities in the signal, using this information makes the nonlinear model much more robust to the presence of noise.



Fig. 2. Speaker Verification DET curves for MixAR-generated nonlinear data with 5B car noise.

# Speaker Verification Experiments

## Evaluation under Noisy Conditions

To evaluate the robustness of MixAR compared to GMM on unseen noise conditions, several noise conditions were simulated with the TIMIT database  by adding synthesized noise from three different noise sources: white, car, and babble. Three SNR levels were used: 10, 5 and 0 dB (in addition to the clean set). The core test partition of the database containing 168 speakers was used. The three types of noise sources were chosen to represent commonly occurring types of noise. The matrix of experimental results is shown in Table II. From this table, it is clear that while unseen noise conditions degrades performance for both models, MixAR performs relatively better than GMM and also uses 2.5 times fewer parameters.

## Evaluation under Channel Variations

Channel variation is another problem that degrades the performance of speech processing systems. NTIMIT is a database that was created by transmitting TIMIT utterances over different telephone channels . We studied speaker verification performance on NTIMIT using the core test set of 168 speakers by splitting the data for each speaker into 8 utterances for training and the remaining 2 utterances for evaluation. The DET performance curves for the 8-mixture MixAR using only static MFCCs (with 480 parameters) and for the 16-mixture GMM (with 1168 parameters) using both static and delta features is shown in Fig. 3. The corresponding EERs are shown in Table III. From this it is clear that MixAR using 2.5 times fewer parameters achieves the same or higher level of performance as a GMM.

TABLE II

Speaker Verification Performance (EER)   
For different noise conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GMM\* (1168) | SNR (dB) | Car Noise | White Noise | Babble Noise |
| Clean | 2.4 | | |
| 10 dB | 19.7 | 48.7 | 40.6 |
| 5 dB | 31.2 | 50.0 | 44.7 |
| 0 dB | 39.3 | 49.8 | 48.2 |
| MixAR (480) | Clean | 1.8 | | |
| 10 dB | 13.7 | 47.0 | 36.9 |
| 5 dB | 23.2 | 47.6 | 42.8 |
| 0dB | 33.9 | 48.5 | 47.6 |

\* Number of parameters in each case is in parenthesis.

# Conclusion

In this work, we applied a nonlinear mixture autoregressive model to several speaker verification tasks. Our experiments with synthetic as well as real speech data show that MixAR model outperforms GMM especially under unseen noisy conditions. Moreover, in all cases we tested, MixAR did not require delta features and used 2.5x fewer parameters to achieve comparable or better performance as that of GMM. The dynamic modeling capability of MixAR is effective at capturing and exploiting speech dynamics. Future work will focus on deriving an adaptation framework for MixAR to demonstrate that it is more effective than conventional adaptation approaches on unseen channel conditions. We are also integrating the MixAR approach into a large vocabulary speech recognition system.



Fig. 3. Speaker Verification DET curves for GMM and MixAR models on TIMIT and NTIMIT databases.

# References

TABLE III

Speaker Verification Performance (EER) with NTIMIT & TIMIT

|  |  |  |
| --- | --- | --- |
| Database | GMM (1168)  (Static +∆ MFCCs) | MixAR (480)  (Static MFCCs Only) |
| TIMIT | 2.4 | 1.8 |
| NTIMIT | 21.0 | 20.9 |

Number of parameters in each case is in parenthesis.

1. X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice-Hall, 2001.
2. D. A. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models”, *Speech Communication*, vol. 17, no. 1-2, pp. 91-108, 1995.
3. 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.
4. D. May, *Nonlinear Dynamic Invariants For Continuous Speech Recognition*, M.S. Thesis, Department of Electrical and Computer Engineering, Mississippi State University, USA, May 2008.
5. A. Petry, D. Augusto, and C. Barone, “Speaker Identification using Nonlinear Dynamical Features,” *Chaos, Solitons & Fractals*, vol. 13, Issue 2, pp. 221-231, February 2002.
6. M. Banbrook, G. Ushaw, and S. McLaughlin, “How to Extract Lyapunov Exponents from Short and Noisy Time Series,” *IEEE Transactions on Signal Processing*, vol. 45, no. 5, pp. 1378-1382, May 1997.
7. M. Zeevi, R. Meir, and R. Adler, “Nonlinear Models for Time Series using Mixtures of Autoregressive Models”, *Technical Report*, Technion University, Israel, available online at: *http://ie.technion.ac.il/~radler/mixar.pdf*, October 2000.
8. 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.
9. 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.
10. M. E. Ayadi, “Autoregressive models for text independent speaker identification in noisy environments,” PhD. Thesis, Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, September 2008.
11. 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.
12. 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.
13. J. Picone, “ISIP Foundation Classes,” *http://www.isip.piconepress.com/projects/speech /software/downloads/production/*, Department of Electrical and Computer Engineering, Temple University, Philadelphia, Pennsylvania, USA, January 2010.
14. National Institute of Standards and Technology, “The 2001 NIST Speaker Recognition Evaluation,” *http://www.nist.gov/speech/tests/spk/2001*, 2001.
15. “TIMIT Speech Corpus,” *http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1*, Linguistic Data Consortium, Philadelphia, 1993.
16. D. Pearce and H. Hirsch, The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions, *ICSLP 2000*, pp. 29-32, October 2000.
17. “NTIMIT Speech Corpus,” *http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S2*, Linguistic Data Consortium, Philadelphia, 1993.

1. Manuscript received May 12, 2010. 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.

   1. S. Srinivasan (corresponding author) and T. Ma are with Mississippi State University, MS, USA (email: [ss754@msstate.edu](mailto:ss754@msstate.edu), tm334@msstate.edu).

   2. G. Lazarou is with New York Transit Authority, New York City, NY, USA (email: georgios.lazarou@gmail.com).

   3. J. Picone is with Temple University, Philadelphia, PA, USA (email: joseph.picone@gmail.com). [↑](#footnote-ref-1)