SPEAKER VERIFICATION: MINIMIZING THE CHANNEL EFFECTS USING AUTOASSOCIATIVE NEURAL NETWORK MODELS

S.P.Kishore and B.Yegnanarayana

Department of Computer Science and Engineering
Indian Institute of Technology, Madras, 600 036, India
E-mail: {kishore@fant., yegna@}iitm.ernet.in

ABSTRACT

The characteristics of telephone channel and handset have significant effect on the performance of speaker verification systems. The channel/handset mismatch between the training and testing data degrades the performance of speaker verification systems. In this paper, we show that the Autoassociative Neural Network (AANN) models can be used to minimize the effects of channel characteristics on the performance of text-independent speaker verification system. This paper also compares two approaches to represent the background model for AANN based speaker verification system.

1. INTRODUCTION

Speaker verification is to verify the speaker from his/her voice. It has been shown that the telephone channel and bandwidth have significant effect on the performance of speaker verification systems [1] [2]. From the results of National Institute of Standards and Technology (NIST) speaker verification evaluations [3], we can observe that the channel mismatch between the training and testing data further degrades the performance of speaker verification systems.

To compensate for the effects of channel/handset characteristics on the performance of speaker verification system, either robust features are extracted or a robust model is built to represent the speaker-specific features. Most of the current speaker recognition systems use Gaussian Mixture Models (GMM) to model the speaker-specific features [4]. Recently, Ikbal [5] [6] has proposed the autoassociative neural network models to capture the speaker-specific features. The IIT Madras speaker verification system in NIST-99 speaker verification evaluation [3] was based on AANN models. The speech corpus used in NIST-99 speaker verification evaluation was conversational database collected over the telephone handsets. The verification conditions were close to the practical situations like channel mismatch and handset mismatch.

The purpose of this paper is to show that the AANN models can be used to minimize the effects of channel characteristics on the performance of speaker verification system. The effects of background normalization on AANN based speaker verification system are also studied. All the studies reported in this paper are made on 230 speakers (male subset) of NIST-99 database. The performance of speaker verification was evaluated on 1448 male test utterances of NIST-99 database with 11 claimants for each test utterance. The paper is organized as follows: Section 2 describes the autoassocaitive neural network and the estimation of data distribution by the AANN model. Section 3 describes the AANN models for speaker verification. Section 4 describes two approaches for background model representation and their effects on the performance of AANN based speaker verification system. In section 5, we discuss the capability of AANN models to minimize the channel effects.

2. AUTOASSOCIATIVE NEURAL NETWORKS

An autoassociative neural network is a feedforward neural network [7] as shown in Fig.1. It consists of an input layer, an output layer and one or more hidden layers. The input and output layers have the same number of processing units. The number of processing units in one of the hidden layer is less than the number of processing units in the input layer. This layer is called the *dimension compression hidden layer*, as this layer causes the input vectors to go through a dimension

We would like to thank Prof. Hynek Hermansky of OGI, Portland, USA, for supporting this work at IIT Madras.

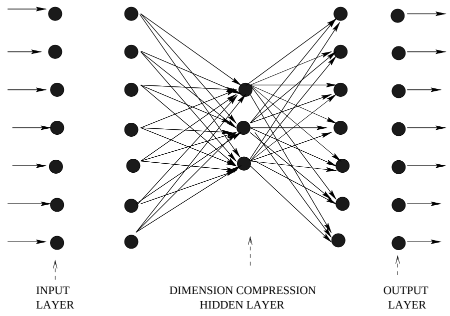

Figure 1: An Autoassociative Neural Network.

compression process. The processing units of the input layer and the output layer are linear, whereas the units in the hidden layer are nonlinear. During training of the network, the target vectors are same as the input vectors. To realize the input vectors at the output layer, the network projects an M-dimensional vector in the input space R^M onto a vector in the subspace R^N, and then maps it back onto the M-dimensional space, where N<M. The network performs *nonlinear principal component analysis* for projecting the input vectors onto the subspace R^N. The subspace R^N is the space spanned by the first N principal components derived from the training data. The value of N is determined by the number of units in the *dimension compression layer*. The mapping of the subspace R^N back onto the M-dimensional space R^M determines the way in which the subspace R^N is embedded in the original space R^M. It has been shown that the AANN trained with a data set will capture the subspace and the hypersurface (in the space spanned by the vectors of the input layer) along the surface of maximum variance of the data [6] [5]. In other words, the AANN can be used to capture the distribution of the given data set.

3. AANN MODELS FOR SPEAKER VERIFICATION

The property of the AANN model to capture the distribution of the given data set is used in building the speaker models for 230 speakers (male subset) of NIST-99 database. Each speaker model is built by training an AANN with the feature vectors extracted from the utterance of a particular speaker. The total duration of the utterances by each speaker is approximately 2 minutes. The speaker model may be an independent model or adapted from the background model. Section 4 discuss the background normalization in detail.

The feature vectors used in this study are 19-dimensional, mean removed , weighted cepstral coefficients [8]. These cepstral coefficients are obtained from the 16 predictor coefficients using a frame size of 27.5ms and a frame shift of 13.75ms. The silence frames are removed using an amplitude threshold. The network is trained using backpropagation learning algorithm in pattern mode. During the testing phase, the feature vectors extracted from the test utterances are given to the claimant model and background model. The score obtained by the model is the error (Euclidean distance between the desired and actual output) averaged over all the feature vectors. The claimant model score is normalized with the score from the background model as described in section 4. EER is used as the measure of performance of the speaker verification system. Results obtained by the speaker verification system using AANN models are discussed in section 4 and 5, for background normalization and channel effects, respectively.

4. BACKGROUND NORMALIZATION

The use of background normalization improves the performance of speaker verification systems [4] [2]. We considered two approaches of background model representation to obtain the normalized scores. One is Universal Background Model (UBM) [4] and the other is Individual Background Models (IBM).

4.1. Universal Background Model

The UBM is built by training an AANN with the feature vectors (see Section 3) extracted from the utterances of 500 speakers (250 male and 250 female). This set of 500 speakers belongs to NIST-98 database. The UBM generated using the pooled feature vectors of all the speakers is called as pooled background model. The duration of each background speaker utterance is approximately 2min. We considered 200 feature vectors per speaker to generate the pooled background model.

The speaker models for the 230 speakers (male subset) of NIST-99 database are adapted from the UBM by training the model with all the feature vectors of a particular speaker. In the testing phase, the claimant model score (S_{cm}) and the background model score (S_{bg}) are obtained as described in Section 3. The normalized score (S_n) of the claimant model is obtained using $S_n = S_{bg} - S_{cm}$ [2] [6]. The performance of speaker verification system using pooled background model is as shown in Table 1.

Table 1: Results of single speaker detection using pooled background model

Environment Between Training and Testing	Equal Error Rate (EER)
Matched (same channel)	12.8%
Channel Mismatch	28.0%
Handset Mismatch	43.2%
All Cases	26.0%

4.2. Individual Background Models

The disadvantage of the UBM lies in choosing the parameters (for e.g., number of speakers, number of epochs, number of feature vectors/speaker) to generate the model. It has been shown that the performance of speaker verification system is sensitive to these parameters [2]. To circumvent this problem we generated 92 speaker models using NIST-98 database, with the same parametric specifications used to generate the speaker models of NIST-99 database. This set of 92 speakers are male speakers. The utterances of 46 speakers are collected over electret handset and the utterances of remaining 46 speakers are collected over carbon-button. These models are called as Individual Background Models.

The speaker models for 230 speakers (male subset) of NIST-99 database are generated independent of the background models. In the testing phase, the feature vectors of the test utterance are given to the claimant model and to all the Individual Background Models. The position (R) of the claimant model score in the ascending list of the scores from Individual Background Models is converted into normalized score (S_n) using $S_n = N/R+1$, where N denotes the number of Individual Background Models. The performance of speaker verification system using Individual Background Models is as shown in Table 2.

Table 2: Results of single speaker detection using Individual Background Models

Environment Between Training and Testing	Equal Error Rate (EER)
Matched (same channel)	9.8%
Channel Mismatch	27.5%
Handset Mismatch	45.0%
All Cases	26.0%

5. MINIMIZING THE CHANNEL EFFECTS USING AANN MODELS

In this section, we discuss the effectiveness of AANN models to minimize the channel effects on the performance of speaker verification system. Sambur [9] used a method called orthogonal linear prediction, based on the extraction of orthogonal parameters. These orthogonal parameters are the eigenvectors obtained from the covariance matrix of the linear predictor coefficients. Reflection coefficients and log area coefficients were also studied. He analyzed the eigenvalues of the orthogonal parameters and proposed that “the first few most significant parameters would be indicative of the linguistic content of the utterance.., the least significant parameters would be indicative of the telephone media, and the remaining parameters would be indicative of the speaker”.

As discussed in section 2, the AANN model projects the input vectors onto the first N principal components to realize them at the output layer. The value of N is determined by the number of units in the dimension compression hidden layer of the AANN model. In other words, the number of nodes in the dimension compression hidden layer determines the number of least varying components ignored. The results shown in Table 1 and Table 2 are obtained by ignoring the five least varying components derived from the training data. The structure of the AANN model used is $19L14N32N22N19L$, where L denotes linear units and N denotes nonlinear units. The integer value denotes the number of units in that particular layer. The performance of speaker verification system obtained by changing the structure of the AANN model to $19L10N32N22N19L$ is as shown in Table 3. The elimination of nine least varying components has significantly improved the performance of speaker verification system. The normalization procedure of Individual Background Models is used to obtain these results.

Table 3: Results of single speaker detection obtained by eliminating the 9 least varying components

Environment Between Training and Testing	Equal Error Rate (EER)
Matched (same channel)	8.2%
Channel Mismatch	20.8%
Handset Mismatch	41.0%
All Cases	23.5%

From Table 2 and Table 3, we can observe that the relative reduction in the EER for matched, channel mismatch and handset mismatch conditions are 16.32%, 24.36%, 8.88% respectively. The reduction in the EER for handset mismatch conditions is comparatively low. It can be attributed to the fact that the first few most varying components do contain significant information about the telephone handset [10].

The normalization procedure of Individual Background Models compares the claimant model score with the scores of all the Individual Background Models. Instead, if we restrict our comparison to the Individual Background Models, whose utterances are collected over the same handset type of the claimant [11], the performance of speaker verification system improves as shown in Table 4. The results obtained by changing the structure of the AANN model to $19L8N32N22N19L$ is as shown in Table 5. Table 6 shows the performance of speaker verification system using the AANN model with the structure $19L6N32N22N19L$. The rise of EER in Table 6 (last row), indicates that the further reduction in the number of nodes of the dimension compression hidden layer may lead to loss of speaker-specific features.

Table 4: Results of single speaker detection obtained by eliminating the 9 least varying components and using handset dependent background

Environment Between Training and Testing	Equal Error Rate (EER)
Matched (same channel)	8.8%
Channel Mismatch	22.8%
Handset Mismatch	33.8%
All Cases	20.0%

Table 5: Results of single speaker detection obtained by eliminating the 11 least varying components and using handset dependent background

Environment Between Training and Testing	Equal Error Rate (EER)
Matched (same channel)	7.9%
Channel Mismatch	20.3%
Handset Mismatch	33.0%
All Cases	19.0%

Table 6: Results of single speaker detection obtained by eliminating the 13 least varying components and using handset dependent background

Environment Between Training and Testing	Equal Error Rate (EER)
Matched (same channel)	8.3%
Channel Mismatch	20.0%
Handset Mismatch	31.2%
All Cases	19.1%

6. CONCLUSION

This paper compared the two approaches (UBM and IBM) to represent the background model for AANN based speaker verification system. When the environment between the training and testing data was not considered these two approaches gave a similar performance. For the matched conditions between the training and testing data, the approach of Individual Background Models gave a relative reduction of 23.4% in the EER. The effects of channel characteristics on the performance of speaker verification system was minimized by reducing the number of nodes in the dimension compression hidden layer of the AANN model. We also observed that this technique had little impact on handset characteristics.

7. REFERENCES

[1] D. A. Reynolds and *et al.*, "The effects of telephone transmission degradations on speaker recognition performance," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 329–332, 1995.

[2] Hemant Misra, *Development of a Mapping Feature for Speaker Recognition*. MS dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, May 1999.

[3] "Speaker recognition workshop," *Proc. NIST 1999,University of Maryland, USA*, Jun 3-4 1999.

[4] D. A. Reynolds, "Comparision of background normalisation methods for text–independent speaker verification," in *Eurospeech*, (Greece), pp. 963–966, 1997.

[5] M. S. Ikbal, H. Misra, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *Int. Joint Conf. on Neural Networks*, (Washington, USA), July 1999.

[6] M. Shajith Ikbal, *Autoassociative neural network models for speaker verification*. MS dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, May 1999.

[7] B. Yegnanarayana, *Artificial Neural Networks*. New Delhi: Prentice-Hall of India, 1999.

[8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, pp. 254–272, Apr. 1981.

[9] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 283–289, Aug. 1976.

[10] S.P.Kishore and B.Yegnanarayana, "Identification of handset type using autoassociative neural network models," in *Int. Conference on Advances of Pattern Recognition*, (ISI, Calcutta, INDIA), Dec. 1999.

[11] L. P. Heck and M. Weintraub, "Handset–dependent background models for robust text–independent speaker recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Munich, Germany), April 1997.