Identification of Handset Type using Autoassociative Neural Networks

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 telephone handset mismatch between training and testing data degrades the performance of speaker recognition systems. To compensate for the handset mismatch, the handset type of the test utterance should be known. In real applications the handset type of the test utterance is not known *a priori*. The handset type should be identified from the speech signal itself. In this paper, we study the effectiveness of an autoassociative neural network (AANN) model to capture the distribution of a given data set, and to identify the handset type from the speech signal. The handset type identification of 85% was achieved on NIST-99 speaker evaluation database. When the procedure was applied for gender identification we obtained a performance of 93% on the NIST-99 database.

1 Introduction

The characteristics of telephone channel and handset have significant effect on the performance of speaker recognition systems. The channel or handset mismatch between training and testing data degrades the performance of speaker recognition systems. The channel mismatch may be compensated to some extent by mean subtraction of the parameter vectors, assuming that the distortion caused by the telephone channel on the parameters is linear [1]. However, the distortion caused by the handset is nonlinear, and is difficult to handle [2].

*corresponding author

The distribution of the data set undergoes affine transformation due to the distortion introduced by the handset [3]. It has been observed from the NIST speaker evaluations that the Equal Error Rate (EER) for the mismatched handset condition is 3 to 4 times higher than the EER for matched handset condition and telephone channel [4]. To compensate for the effect of handset mismatch, normalization techniques like h_norm [2], handset mapper [5] and selection of suitable background speakers [6] have been proposed.

To implement these normalization techniques, the handset type of the test utterance should be known. In this paper, an approach based on autoassociative neural network (AANN) model is proposed to identify the handset type from the speech signal. We show that similar approach can be adopted for the identification of gender from speech data.

The paper is organized as follows: Section 2 describes the autoassociative neural network and the estimation of data distribution by AANN. Section 3 describes the AANN models to perform the handset identification task. In Section 4, the testing procedure and the experimental results are discussed.

2 Autoassociative Neural Networks

An autoassociative neural network is a feedforward neural network [7][8] as shown in Fig.1. It consists of an input layer, an

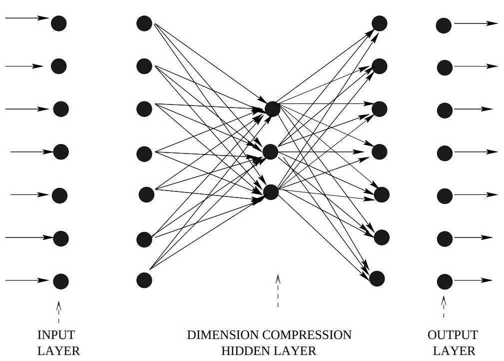

Figure 1: An autoassociative Neural Network.

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 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 output layer) along the surface of maximum variance of the data [9] [10]. In other words, the AANN can be used to capture the distribution of the given data set. We explore this feature of AANN to identify the handset type from the speech signal as discussed in Sections 3 and 4.

3 AANN Models for Handset Types

The basic idea is to use an AANN to capture the distribution of feature vectors extracted from the utterances collected over the handset of a particular type. The different handset types used for this study are electret and carbon–button. Utterances of 58 speakers collected over the electret handset are used to train one AANN. The utterances of another 58 speakers collected over carbon–button handset are used to train a second AANN. Each set of 58 speakers consists of 29 male and 29 female speakers. The speaker evaluation database of NIST–98 is used for this training phase.

The features used in this study are 19–dimensional weighted cepstral coefficients [1]. These cepstral coefficients are obtained from the 16 linear prediction coefficients computed using a frame size of 20ms and a frameshift of 10ms. The silence frames are removed by using an amplitude threshold.

The structure of the AANN is $19L - 6N - 22N - 19L$, where L denotes linear units and N denotes nonlinear units. The integer value denotes the number of units in that particular layer. The activation function of the nonlinear unit is the *hyperbolic tangent* function. The network is trained using backpropagation learning algorithm in pattern mode [7][8].

4 Performance of AANN Models for Identification of Handset Type

Feature vectors are extracted from a given test utterance. Each feature vector is given as input to both the AANNs and the error at the output layer is obtained. The error is the Euclidean distance between the actual output and the desired output. The error for all frames in the test utterance is averaged for each AANN. A decision on the type of handset is based on the network that gives less error.

Handset type identification test was performed on 1448 test utterances of NIST-99 speaker evaluation database. The performance of the handset type identification is given in Table 1. The overall performance of the handset type identification is 84.46%. An AANN trained with the utterances of 58 speakers(29 male and 29 female) may capture the gross characteristics of the speakers and the linguistic information. The good performance of handset type identification shows that the effect of handset characteristics is significant on the distribution of feature vectors.

Table 1 : Results of Handset Type Identification

Handset Type	No. of Test Utterances	No. of Utterances for which Handset Type is identified correctly (Percentage in Parentheses)
Carbon-button	591	479(81.04%)
Electret	857	744 (86.81%)

By training seperate AANNs with speech data for male and female speakers, one can study the effectiveness of these models for gender identification. The performance of two AANNs trained for gender identification is as shown in Table 2. It is interesting to note that the distribution of features captured by the networks do possess the discriminability of the gender.

Table 2 : Results of Gender Identification

Gender	No. of Test Utterances	No. of Utterances for which Gender is identified correctly (Percentage in Parentheses)
Male	1448	1331(91.91%)
Female	1972	1834(93.00%)

In this paper, an AANN–based approach was proposed to identify the handset type automatically from the speech signal. The results also supports the conjecture that AANNs do capture the distribution of feature vectors of the training data set, and the effect of handset characteristics is significant on the distribution of the feature vectors.

Acknowledgement

We would like to thank Hemant, Shajith Ikbal and Mathew for many useful discussions. We would also like to thank Prof. Hynek Hermansky of OGI, USA for supporting this study.

References

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

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

[3] R. J. Mammone, X. Zhang, and R. P. Ramachandran, “Robust speaker recognition: A feature–based approach,” *IEEE Signal Processing Magazine*, vol. 13, pp. 58–71, Sept. 1996.

[4] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on switch board corpus," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 113–116, 1996.

[5] T. Quatieri, D. A. Reynolds, and G. Leary, "Magnitude-only estimation of handset nonlinearity with application to speaker recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 745–748, 1998.

[6] 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.

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

[8] S. Haykin, *Neural networks: A comprehensive foundation*. New Jersey: Prentice-Hall International, 1999.

[9] 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.

[10] 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.