SOURCE AND SYSTEM FEATURES FOR SPEAKER RECOGNITION USING AANN MODELS

B. Yegnanarayana, K. Sharat Reddy and S. P. Kishore

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai-600036, INDIA
email: {yegna,sharat,kishore}@speech.iitm.ernet.in

ABSTRACT

In this paper we study the effectiveness of the features extracted from the source and system components of speech production process for the purpose of speaker recognition. The source and system components are derived using linear prediction (LP) analysis of short segments of speech. The source component is the LP residual derived from the signal, and the system component is a set of weighted linear prediction cepstral coefficients. The features are captured implicitly by a feedforward autoassociative neural network (AANN). Two separate speaker models are derived by training two AANN models using feature vectors corresponding to source and system components. A speaker recognition system for 20 speakers is built and tested using both the models to evaluate the performance of source and system features. The study demonstrates the complementary nature of the two components.

1. INTRODUCTION

Speaker recognition involves speaker identification or speaker verification based on his/her voice in the form of speech. Speech signal carries information about speech message, speaker and also the channel/environment of recording. For speaker recognition, speech data from a speaker is collected and is used to develop a model for capturing the speaker specific information. For text-independent speaker recognition the speech data is usually of about one minute duration. The model for each speaker is either a statistical model like a Gaussian Mixture Model or Hidden Markov Model, or a neural network model like feedforward autoassociative network [1] [2] [3]. After developing separate models for each speaker, recognition involves determining the probability that a given test utterance (usually of 30 sec or more) belong to one of the models. The speaker for the model that gives maximum probability for the given test utterance is marked as the speaker of the test utterance. In speaker identification the speaker characteristics of the test utterance is matched with a finite set of speaker models, i.e., the population is fixed. In speaker verification, on the other hand, the claim of the given speaker is to be verified by determining the extent of match between the speaker characteristics of test utterance with the speaker model. If the match exceeds a preset threshold, the claim is accepted. Otherwise, the claim is rejected.

For speaker recognition studies, the speech signal is processed to extract features suitable for the task. Usually the features represent short-time (10-30 ms) spectral information of the speech signal. Statistical distributions of the spectral features are used to build speaker models. It is interesting to note that human beings recognize speakers mostly from the source characteristics such as glottal vibrations, and prosodic features such as intonation and duration. Since it is difficult to reliably extract and build a speaker model using this information, not much effort has gone in developing speaker recognition system using the source and suprasegmental information. This paper is an attempt to derive a speaker-specific model using predominantly the source characteristics of the speech production, and use this model for speaker recognition. Performance of this system is compared with the performance of the system using the conventional spectral-based methods. For both types of systems, an autoassociative neural network model is used to represent the speaker characteristics [2] [4].

This paper is organized as follows: In Section 2, we discuss the source and system features used in this study. In Section 3, the speaker recognition system using the AANN model is described. The performance of the recognition systems based on source and system features is evaluated, and the results are discussed in Section 4. Section 5 gives a summary of the paper.

2. FEATURES FOR SPEAKER RECOGNITION

Speech data for each speaker is represented in the form of feature vectors, each vector is derived from a short segment (10-30 ms) of data. Both the source and system features are derived using linear prediction analysis [5]. An 8^{th} order linear prediction analysis is performed for every frame (20 ms) on preemphasised (differenced) speech. 19 linearly weighted cepstral coefficients are derived from the 8 LPCs to represent the short-time spectral envelope [6] [7]. The 19-dimensional weighted cepstral coefficient feature vector is used to represent the vocal tract system characteristics. Linear prediction residual is used to derive the source characteristics. The hypothesis is that the source characteristics of the speaker may be present in the higher (> 2) order correlations among the samples, which are difficult to extract. Note that the second order correlations are nearly zero in the LP residual, as these correlation components are extracted by the LP analysis to represent the vocal tract system characteristics.

Fig.1 shows a segment of speech, the LP residual and the LP spectrum. The residual signal and the LP spectrum can be viewed as decomposition of the signal into approximate source and system components. It is clear that while the LP spectral information can be represented in the form of parameters like weighted LP cepstral coefficients, it is not obvious how to represent the information present in the LP residual samples. Since spectral envelope of the residual is nearly flat, attempts to represent the spectral information from the residual may not bring out significant source features [8]. Hence we are exploring nonlinear models like neural networks to extract this source information from the LP residual.

In order to extract the unknown higher order correlations among samples, a five layer autoassociative neural network model is used with nonlinear units in the three hidden layers and linear units in the input and output layers [9]. The number of units in the input and output layers are equal, and correspond to the block size of the residual samples used to train the model. The number of units in the middle hidden layer is less than the size of the block, and hence is called a compression layer. Details of the network structure and the speaker recognition system based on this structure is described in the next section.

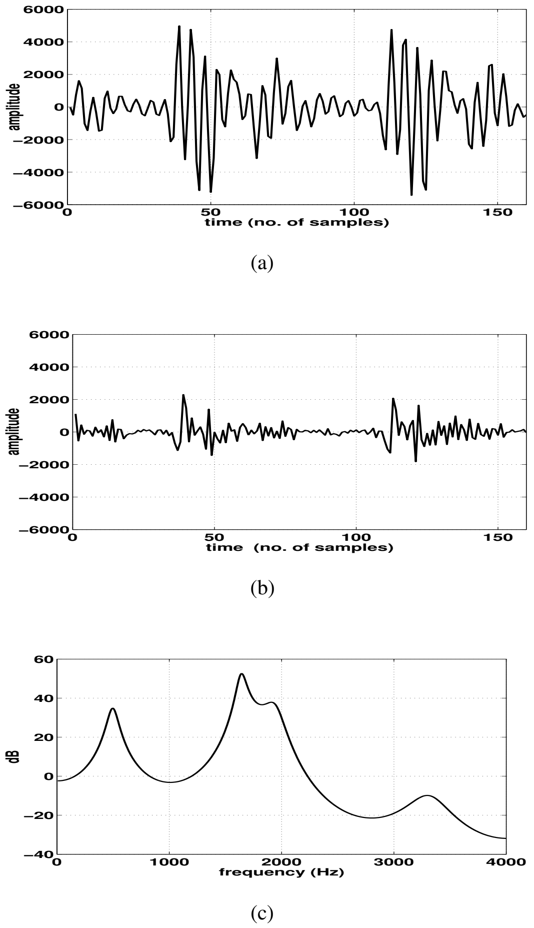

Fig. 1. (a) Speech segment. (b) Residual of the speech segment obtained from 8^{th} order linear prediction. (c) 8^{th} order LP spectrum.

3. AANN-BASED SPEAKER RECOGNITION SYSTEM

Separate AANN models are developed for speaker recognition systems based on vocal tract system characteristics and source characteristics. The model using the 19-dimensional weighted LP cepstral feature vector captures the distribution of the vocal tract system feature vectors of a given speaker [10]. The distributions are usually different for different speakers. Thus, each AANN model trained with one speaker's data captures the distribution of that speaker. The AANN model for the vocal tract system features is shown in Fig.2. We call this as Model 1.

The model is trained with feature vectors derived from one minute of speaker data. The feature vectors are computed for every 27.5 ms frames separated by 13.75 ms. After removing the low energy and silence frames, the total number of frames per speaker are approximately 6000. The model is trained using backpropagation learning algorithm for 60 epochs [11]. Each feature vector is normalized to unit magnitude before giving as input to the model. One model is created for each speaker.

To capture the source characteristics from the LP residual, a block size of 40 samples is used. The corresponding AANN model is shown in Fig.3. We call this as Model 2.

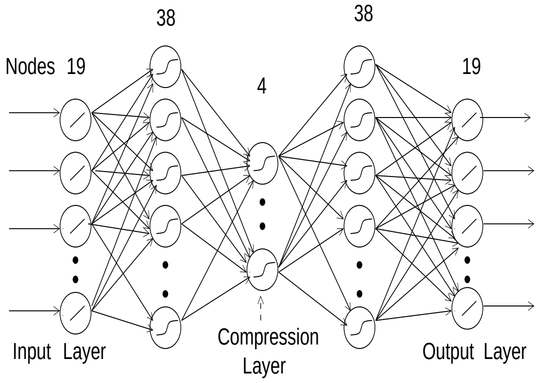

Fig. 2. Model 1: AANN model for system features

The structure of the model is based on some preliminary experimentation. It must be emphasized that the structure is not optimized for this study.

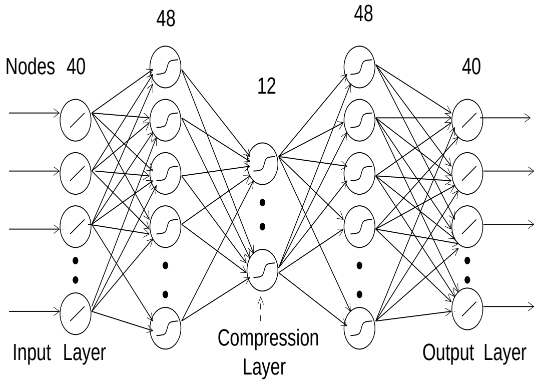

Fig. 3. Model 2: AANN model for source features

The model is trained with blocks of residual samples taken with one sample shift. The number of blocks used for training is nearly equal to the number of samples in the speech data for the speaker, after removing silence and low energy frames. The model is trained for 60 epochs. Each block is normalized to unit magnitude before giving as input to the model. One model is created for each speaker.

For testing, a test utterance of 60 sec duration for each speaker is used. The system feature vectors and the LP residual are extracted using 8^{th} order LP analysis. The 19-dimension feature vector (normalized) for Model 1 and a block of 40 samples (normalized) of LP residual for Model 2 are given as inputs. The output of each model is compared with its input to compute the squared error for each frame or block. The error (E_i) for the i^{th} frame or block is transformed into a confidence value by using $c_i = exp(-\lambda E_i)$, where the constant $\lambda = 1$ throughout this study. This confidence value will be larger for smaller values of error, i.e., frames or blocks matching with the corresponding models. The c_i value will be lower for large error value, thus giving less emphasis to frames or blocks not matching with their respective models. A given test utterance is compared with each of the 20 speaker models to obtain the average confidence value $c = \sum_i c_i$ for each model. The average confidence value is used to compare the performance of the speaker recognition systems based on source and system characteristics. The performance evaluation is discussed in the next section.

4. PERFORMANCE EVALUATION OF SPEAKER RECOGNITION SYSTEM

We have used data for 20 speakers from NIST 99 development data for evaluation and testing [4]. The performance of the recognition system using source and system features is given in Table 1 for two independent sets of speakers data, where each set consists of 20 speakers. From Table 1, it is evident that both features seem to give good performance. The results also indicate the complementary nature of the source and system components for speaker recognition. We can use this feature to combine the results of both the models. A simple way of combining is to add the scores from the source and system models, and then rank the test speaker according to the new scores. In Table 1, these results are shown as the rank of the combined model. The results show that the overall ranking of the speakers have improved significantly.

It is interesting to note that the source features are derived from the LP residual, which does not have any spectral information. Still it is giving recognition performance nearly as well as the system features. One may be able to exploit this feature to reduce the effects of channel and handset mismatch between training and testing. Also the selection of blocks of residual based on signal characteristics may help to reduce the effects of noise [12]. If the test signal is noisy, fewer blocks can be used to compute the confidence value.

5. SUMMARY AND DISCUSSION

We have presented a method for developing a speaker recognition system using source features of speech production. LP residual is used to represent the source features. The speaker information present in the source features is captured using an autoassociative neural network model. Since the residual does not contain any significant spectral information, this method may provide robustness against channel and handset effects, which are known to degrade the performance of a speaker recognition system.

Table 1. Performance of Speaker Recognition using source and system features. The table shows the rank of the speaker obtained by matching with 20 speakers.

	Speaker No.→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Set I	Rank of Model 1 (system features)	1	1	1	1	1	2	1	1	1	8	1	1	1	1	1	1	1	1	1	1
	Rank of Model 2 (source features)	2	1	1	1	1	1	4	1	1	1	1	2	1	1	1	13	1	1	1	1
	Rank of Combined Model	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1

	Speaker No.→	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Set II	Rank of Model 1 (system features)	1	1	1	4	1	1	1	1	1	1	2	1	1	1	1	1	5	1	1	1
	Rank of Model 2 (source features)	1	1	1	1	1	10	1	1	1	1	10	1	1	1	1	3	2	2	1	1
	Rank of Combined Model	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1	1	2	1	1	1

In this study the structure of the autoassociative neural network model is based on some preliminary experimentation. A systematic study is needed to determine a suitable structure of the model and also the order of the LP analysis and the block size of the residual. It is likely that such an experimentally optimized model may perform better than the one described in this paper. Since the two models are based on independent features, it is possible to suitably combine the results of both to obtain significantly better performance compared to individual models.

6. REFERENCES

[1] D. A. Reynolds, "Speaker identification and verification using gaussian mixture models," *Speech Comm.*, vol. 17, pp. 91–108, Aug. 1995.

[2] S. P. Kishore and B. Yegnanarayana, "Speaker verification: Minimizing the channel effects using autoassociative neural network models," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Istanbul, 2000, pp. 1101–1104.

[3] M. Shajith Ikbal, Hemant Misra, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *Int. Joint Conf. on Neural Networks*, Washington, USA, 1999.

[4] NIST, "Speaker recognition workshop notebook," *Proc. NIST 2000 Speaker Recognition Workshop, University of Maryland, USA*, Jun 26-27 2000.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[7] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[8] Phillippe Thevenaz and Heing Hugli, "Usefulness of lpc-residue in text-independent speaker verification," *Speech Comm.*, vol. 17, pp. 145–57, 1995.

[9] Mark A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE*, vol. 37, no. 2, pp. 233–243, Feb. 1991.

[10] S. P. Kishore and B. Yegnanarayana, "Speaker verification using autoassociative neural network models," *IEEE Trans. Acoust., Speech, Signal Processing (communicated)*.

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

[12] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech, Audio Processing*, vol. 8, pp. 267–281, 2000.