Asian Journal of Information Technology 13 (10): 595-598, 2014
ISSN: 1682-3915

Acoustic Feature Extraction Methods LPC, LPCC and RASTA-PLP in Speaker Recognition

R. Visalakshi and P. Dhanalakshmi
Department of Computer Science and Engineering,
Annamalai University, Chidambaram, India

Abstract: In this study, researchers have analyzed the performance of a Speaker recognition system based on features extracted from the speech recorded using close speaking and a throat microphone in clean and noisy environment. In general, clean speech performs better for Speaker recognition system. Speaker recognition in noisy environment, using a transducer held at the throat results in a signal that is clean even in noisy. The proposed techniques of Speaker Recognition (SR) are done in two ways such as acoustic feature extraction and classification. Initially, the speech signal is given to various acoustic features that include Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC) and Relative Spectral-Perceptual Linear Predictive (RASTA-PLP). Second the extracted features are given to Auto Associative Neural Networks (AANN). The experimental results show that the performance of RASTA-PLP with AANN Model gives an accuracy of 95% in clean and 89% in noisy environments using throat microphone.

Key words: Autoassociative Neural Network (AANN), Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Relative Spectral Perceptual Linear-Prediction (RASTA-PLP), Speaker Recognition (SR), Throat Mmicrophone (TM), close speaking microphone

INTRODUCTION

The throat microphone is a transducer that is placed in contact with the skin surrounding the larynx near the vocal folds. It converts the vibrations that picks up into equivalent speech signals. Typically, the throat, speech is a low amplitude signal and its speech is of high quality. In a noisy environment, the intelligibility of close speaking microphone speech is affected as the microphone picks up not only the voice but also the background noise. But the intelligibility of the throat microphone signal is nearly the same as that of the signal obtained in a noise-free environment. Hence, the throat microphone is a preferred choice for use in speech applications even in adverse conditions (Shahina *et al*., 2004).

Applications such as military field, music industry, cockpit, firefighters, soldiers, airplane, motorcycle, factory (Erzin, 2009) or street crowd environment, whisper speech and analyzing the performance of speech impaired people. Speaker recognition is a task of person identification using speech as the biometric authentication (Atal, 1976). As speech interaction with computers becomes more pervasive in activities such as financial services and information retrieval from speech databases, the utility of automatically recognizing a speaker based solely on vocal characteristics increases. Given a speech sample, speaker recognition is concerned with extracting clues to the identity of the person who was the source of that utterance (Kinnunen and Li, 2010). Speaker recognition is divided into two specific tasks: verification and identification (Reynolds, 2002). In speaker verification the goal is to determine from a voice sample if a person is whom he or she claims. In speaker identification the goal is to determine which one of a group of known voices best matches the input voice sample. In either case the speech can be constrained to a known phrase (text-dependent) or totally unconstrained (text-independent) (Balakrishnan, 2005). In most of the applications, voice is used to confirm the identity claim of a speaker. Speaker recognition system may be viewed as working on four stages, namely features, extraction, modeling and testing.

Outline of the study: In order to distinguish the two categories of speech data from clean and noisy environment, the features are extracted using Linear Prediction Analysis (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP). The auto Associative Neural Network (AANN) is used to capture the distribution of the acoustic feature vectors in the feature

Corresponding Author: R. Visalakshi, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India

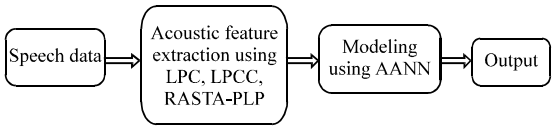

Fig. 1: Block diagram of speaker recognition

space. Back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. Experimental results show that the accuracy of AANN with Relative Spectral Perceptual linear prediction features can provide a better result. Figure 1 illustrates the block diagram of speaker recognition.

ACOUSTIC FEATURE EXTRACTION TECHNIQUES

In this proposed method feature extraction based on the LPC, LPCC and RASTA-PLP are used for speaker recognition.

Preprocessing: To extract the features from the speech signal, the signal must be pre-processed and divided into successive windows or analysis frames. Throughout this research, sampling rate of 8 kHz, 16 bits monophonic, Pulse Code Modulation (PCM) format in wave speech is adopted (Rabiner and Juang, 2003). Speech signal which is recorded using a close-speaking microphone from the collected speaker database speech data is pre-processed before extracting features. This involves the detection of begin and end points of the utterance in the speech waveform, pre-emphasis and windowing of the frame. The process of pre-emphasis provides high frequency emphasis and windowing reduces the effect of discontinuity at the ends of each frame of speech. The speech samples in each frame are pre-processed using a different operator to emphasize the high frequency components.

Linear prediction analysis: For acoustic feature extraction, the difference speech signal is divided into frames of 20 msec with a shift of 10 msec. A pth order LP analysis is used to capture the properties of the signal spectrum.

In the LP analysis of speech each sample is predicted as the linear weighted sum of the past p samples where p represents the order of prediction (Dhanalakshmi, 2010). If s(n) is the present sample then it is predicted by the past p samples as:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \quad (1)$$

RASTA-PLP: RASTA-PLP is an extension to PLP. PLP was originally proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information (Hermansky *et al*., 1991). RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel, e.g., from a telephone line. The RASTA-PLP algorithm proceeds as follows (Hermansky, 1990):

- Compute the critical-band-spectrum (as in the PLP) and take its logarithm
- Transform spectral amplitude through compressing static nonlinear transformation
- Filter the time trajectory of each transformed spectral component
- Transform the filtered speaker representation through expanding static non-linear transformation
- As in conventional PLP, multiply by the equal loudness curve and raise to the power 0.33 to simulate the power law of hearing
- Compute an all-pole model of the resulting spectrum

AANN MODEL FOR SPEAKER RECOGNITION

Autoassociative neural network models are feed forward neural networks performing an identity mapping of the input space and are used to capture the distribution of the input data (Yegnanarayana and Kishore, 2002). The distribution capturing ability of the AANN Model is described in this study. Let us consider the five layer AANN Model shown in Fig. 2 which has three hidden layers. This layer is called the dimension compression hidden layer as this layer causes the input vectors to go through a dimension compression process (Yegnanarayana *et al*., 2002). In this network, the second and fourth layers have more units than the input layer. The third layer has fewer units than the first or fifth. The processing units in the first and third hidden layer are nonlinear and the units in the second compression/hidden layer can be linear or non-linear. The activation functions in the second, third and fourth layer are non-linear. The structure of the AANN Model used in the study is 14L 38N 4N 38N 14L for LPC, 19L 38N 4N 38N 19L for LPCC and 13L 26N 4N 26N 13L for RASTA-PLP, for capturing the distribution of acoustic features where L denotes a linear unit and N denotes a non linear unit. The integer

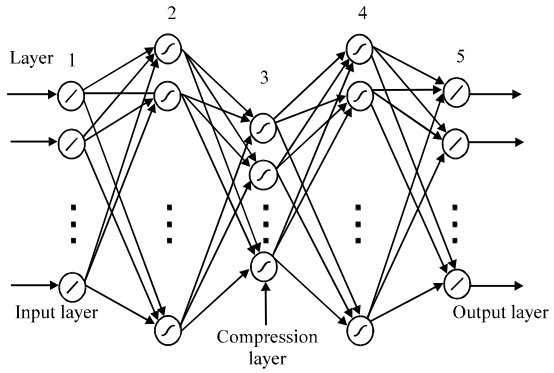

Fig. 2: Autoassociative neural network

value indicates the number of units used in that layer. The non-linear units use tanh(s) as the activation function where s is the activation value of the unit. A Back Propagation Learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector (Yegnanarayana, 1999; Haykin, 2001). During training the target vectors are same as the input vectors. The AANN trained with a data set will capture the subspace and the hypersurface along the surface of maximum variance of the data (Haykin, 2001).

EXPERIMENTAL RESULTS

Datasets: The evaluation of the proposed Speaker recognition system is performed by using a speech database which consists of the following contents: recording was done in the laboratory under clean and noisy environment. Speech from volunteers was acquired using the close-speaking and throat microphones for 2-5 sec. The data obtained from each of the 400 speakers is used to train a speaker model. The recordings for training and testing the speaker models were carried out in separate sessions.

Recognizing speaker using AANN: The five layer Auto Associative Neural Network Model as described in study is used to capture the distribution of the acoustic feature vectors. The structure of the AANN Model used in the study is 14L 38N 4N 38N 14L for LPC, 19L 38N 4N 38N 19L for LPCC, 13L 26N 4N 26N 13L for RASTA-PLP for capturing the distribution of the acoustic features of a class where L denotes a linear unit and N denotes a nonlinear unit. The non-linear units use tanh(s) as the activation function where s is the activation value of the unit. The Back Propagation Learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. The speech signals

Table 1: Speaker recognition performance using AANN

Categories	LPC (%)	LPCC (%)	RASTA-PLP (%)
Clean	89	90	95
Noisy	82	84	89

are recorded for 2-5 sec at 8000 samples/sec and divided into frames of 20 msec with a shift of 10 msec. A 14th order LP analysis is used to capture the properties of the signal spectrum as described in study. The recursive relation (Yegnanarayana, 1999) between the predictor coefficients and cepstral coefficients is used to convert the 14 LP coefficients into 19 cepstral coefficients. The LP coefficients for each frame is linearly weighted to form the LPCC. The distribution of the 14 dimensional LPC feature vectors, 19 dimensional LPCC feature vectors and 13 dimensional RASTA-PLP feature vectors in the feature space are captured using an AANN Model. The performance of acoustic features such as LPC, LPCC and RASTA-PLP for AANN in clean and noisy environment is given in Table 1.

CONCLUSION

In this study, researchers have proposed a Speaker recognition system using AANN Model. LPC, LPCC and RASTA-PLP features are extracted from the voice signal or speech data that can later be used to represent each speaker. Experimental results show that then Characteristics of the speech data are collected from clean and noisy environment using close speaking and throat microphone. The degradation in performance is due to the change in speaker characteristics. Researchers used AANN Models to train the noisy speech which ensures that the voice characteristics of the speaker is similar in both the training and testing stage. By comparing the results of the various acoustic features, the performance of RASTA-PLP with AANN model gives an accuracy of 95% in clean and 89% in noisy environments using throat microphone.

REFERENCES

Atal, B.S., 1976. Automatic recognition of speakers from their voices. Proc. IEEE, 64: 460-475.

Balakrishnan, N., 2005. Improved text-independent speaker recognition using gaussian mixture probabilities. M.S Thesis, Carnegie Mellon University.

Dhanalakshmi, P., 2010. Classification of audio for retrieval applications. Ph.D. Thesis, AU, Chidambaram.

Erzin, E., 2009. Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. Audio Speech Language Process. Trans., 17: 1316-1324.

Haykin, S., 2001. Neural Networks a Comprehensive Foundation. Pearson Education, Asia.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87: 1738-1752.

Hermansky, H., N. Morgan, A. Bayou and P. Kohn, 1991. RASTA-PLP speech analysis. Technical Report (TR-91-069), International Computer Science Institute, Berkeley, CA.

Kinnunen, T. and H. Li, 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun., 52: 12-40.

Rabiner, L. and B. Juang, 2003. Fundamentals of Speech Recognition. Pearson Education, Singapore.

Reynolds, D.A., 2002. An overview of automatic speaker recognition technology. IEEE Acoustics Speech Signal Proc., 4: 4072-4075.

Shahina, M., B. Yegnanarayana and M.R. Kesheorey, 2004. Throat microphone signal for speaker recognition. Department of Computer Science and Engeenring, Indian Institute of Technology Madras.

Yegnanarayana, B. and S. Kishore, 2002. AANN: An alternative to GMM for pattern recognition. Neural Networks, 15: 459-469.

Yegnanarayana, B., 1999. Artificial Neural Networks. Prentice Hall of India, New Delhi.

Yegnanarayana, B., S. Gangashetty and S. Palanivel, 2002. Autoassociative Neural Network Models for Pattern Recognition Tasks in Speech and Image. In: Soft Computing Approach to Pattern Recognition and Image Processing, Ghosh, A. and S.K. Pal (Eds.). World Scientific Publishing Co. Pvt. Ltd., Singapore, pp: 283-305.