



Throat Microphone for Speaker Recognition Using AANN

R. Visalakshi*, P.Dhanalakshmi

Department of Computer Science Engineering
Annamalai University, India

Abstract— In this paper, we have analyzed the performance of speaker recognition system based on features extracted from the speech recorded using throat microphone in clean and noisy environment. In general, clean speech performs better for speaker recognition system. Speaker recognition in noisy environment, using transducer held at the throat results in a signal that is clean even in noisy. This speaker recognition system is also beneficial to visually impaired person. The system recognizes the speakers from acoustic features of mel-frequency cepstral coefficients (MFCC). AANN is one of the modeling techniques used to capture the feature. The auto associative neural network (AANN) is used to capture the distribution of the acoustic feature vectors in the feature space. This model captures the distribution of the acoustic feature of a class, and the backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. The experimental results show that, the performance of AANN using MFCC gives an accuracy of 94.93% in clean and noisy environment.

Keywords— Autoassociative neural network(AANN); Mel-frequency cepstral coefficients(MFCC); Speaker Recognition(SR); Throat microphone; Visually impaired.

I. 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 [13].

Applications such as military field, music industry, cockpit, fire fighters, soldiers, air-plane, motorcycle, factory[7], 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 [1]. 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 [8]. Speaker recognition is divided into two specific tasks: verification and identification [6]. 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) [11]. In most of the applications voice is used to confirm the identity claim of a speaker. Speaker recognition system may be viewed as working in four stages namely Analysis, Features Extraction, Modeling and Testing as shown in Fig.1

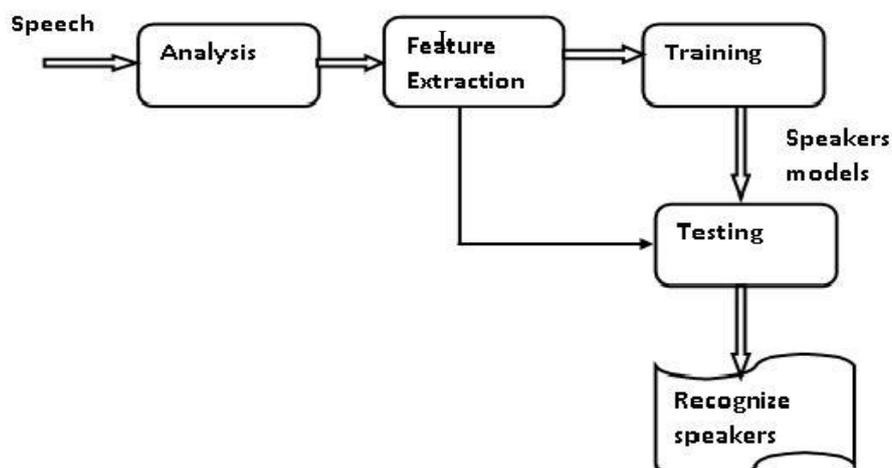


Fig. 1 Stages in the Development of Speaker Recognition System

Speech data contains different types of information that conveys speaker identity. These include speaker specific information due to the vocal tract, excitation source and behavioural traits. The speech signal is produced from the vocal tract system. The physical structure and dimension of vocal tract as well as the excitation source are unique for each speaker. This uniqueness is embedded into the speech signal during speech production and can be used for speaker recognition. To obtain good representation of these speaker characteristics, speech data needs to be analysed. The speech analysis stage deals with the selection of suitable frame size and frame shift for segmenting the speech signal for further analysis and feature extraction [12].

A. Outline of the work

In order to distinguish the two categories of speech data from clean and noisy environment, the features are extracted using Mel frequency Cepstral Coefficients (MFCC). The auto associative neural network (AANN) is used to capture the distribution of the acoustic feature vectors in the feature 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 Mel-frequency cepstral features can provide a better result.

II. Acoustic Feature Extraction Techniques

In this proposed method feature extraction based on MFCC are used for speaker recognition

A. Pre-processing

To extract the features from the speech signal, the signal must be pre-processed and divided into successive windows or analysis frames. Throughout this work, sampling rate of 8 kHz, 16 bit monophonic, pulse code modulation (PCM) format in wave speech is adopted [9]. 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 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 preprocessed using a difference operator to emphasize the high frequency components.

B. Mel frequency cepstral coefficients

The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly affects the recognition performance [16]. A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved more efficient. The calculation of the MFCC includes the following steps.

C Mel-frequency wrapping

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency f , measured in Hz, a subjective pitch is measured on a scale called the mel scale. The melfrequency scale is linear frequency spacing below 1KHz and a logarithmic spacing above 1kHz. As a reference point, the pitch of a 1 KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels.

Our approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant band width and spacing on a mel frequency scale.

D Cepstrum

The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for a given frame analysis, because the mel spectrum coefficients (and their logarithm) are real numbers. Hence, we can convert them into the time domain using the discrete cosine transform (DCT). In this final step, log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^K (\log S_k) \cos \left[n(k - 0.5) \frac{\pi}{K} \right], \tag{1}$$

Where, $n=1, 2, \dots, L$ Denoting the output of the filter bank by S_k ($k = 1, 2, \dots, K$), complete process for the calculation of MFCC is shown in

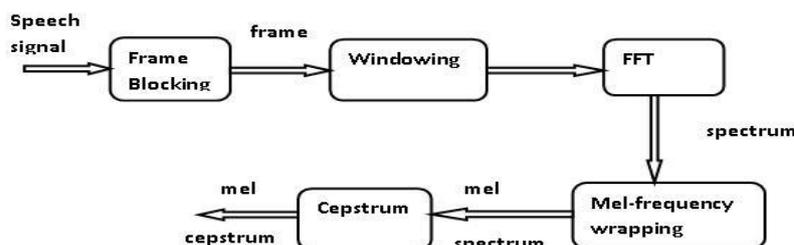


Fig. 2 Block Diagram of Mel Cepstral Coefficients

III MODELING TECHNIQUES FOR SPEAKER RECOGNITION

A. Autoassociative Neural Network Models

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [3]. The distribution capturing ability of the AANN model is described in this section. Let us consider the five layer AANN model shown in Fig. 3, 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 [5]. 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 nonlinear.

The activation functions at the second, third and fourth layer are nonlinear. The structure of the AANN model used in our study is 39L 78N 4N 78N 39L for MFCC, for capturing the distribution of acoustic features, where L denotes a linear unit, and N denotes a non linear unit. The integer 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. Back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector [2][14][15].

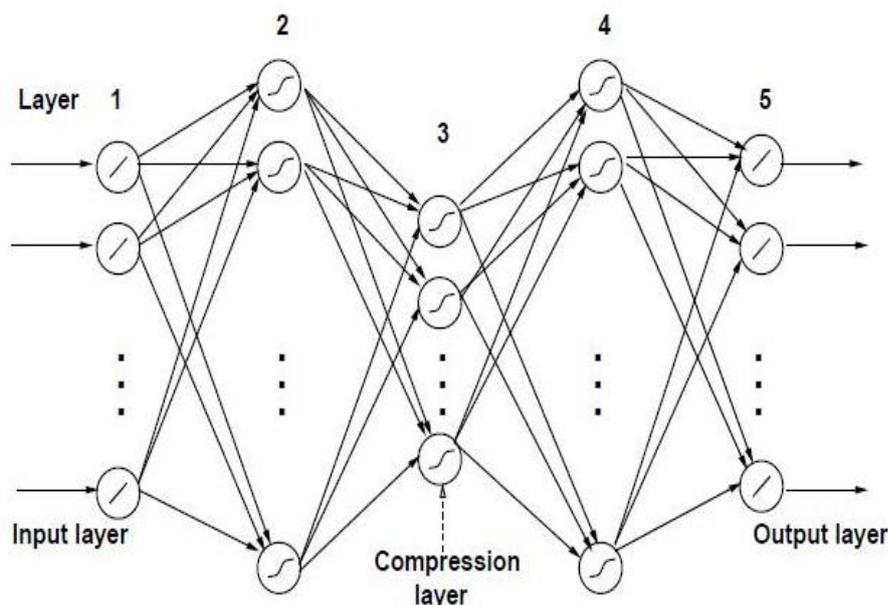


Fig. 3 Autoassociative Neural Network Models.

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 [10].

IV. Experimental Results

A. 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 throat speaking microphones for 2 secs to 5 secs. Text-independent speech was used in this study. The data obtained from each of the 25 speakers is used to train a speaker model. Each test utterance was of 2 secs to 5 secs duration. The recordings for training and testing the speaker models were carried out in separate sessions. About 225 test utterances obtained from the 25 speakers under clean and noisy conditions were used in this study.

B. Modeling using AANN

The five layer autoassociative neural network model as described in Section III is used to capture the distribution of the acoustic feature vectors. The structure of the AANN model used in our study is 39L 78N 4N 78N 39L for MFCC, for capturing the distribution of the acoustic features of a class, where L denotes a linear unit, and N denotes a nonlinear unit. The nonlinear units use tanh(s) as the activation function, where s is the activation value of the unit. The backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. The speech signals are recorded for 2 sec to 5 sec at 8000 samples per second and divided into frames of 20 msec, with a shift of 10 msec. The distribution of the 39 dimensional MFCC feature vectors in the feature space is captured using an AANN model. Separate AANN models are used to capture the distribution of feature vectors of each class. The distribution is usually different for different speaker [4].

The acoustic feature vectors are given as input to the AANN model and the network is trained for 300 epochs. One epoch of training is a single presentation of all the training vectors to the network. The AANN training process creates separate model for each category. For testing, the speech signal is recorded for 2 sec to 5 sec. For evaluating the performance of the system, the feature vector is given as input to each of the model. The output of the model is compared

with the input to compute the normalized squared error. The normalized squared error (E) for the feature vector y is given by, $E = \frac{\|y - o\|^2}{\|y\|}$, where o is the output vector given by the model. The error (E) is transformed into a confidence score (C) using $C = \exp(-E)$. The average confidence score is calculated for each model. The class is decided based on the highest confidence score. The performance of the system is evaluated, and the method achieves about 94.93%.

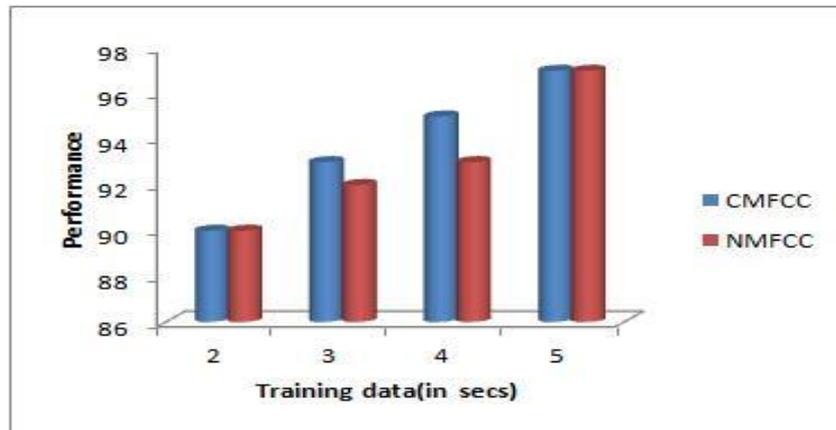


Fig. 4 Performance of AANN for Speaker Recognition

V. CONCLUSIONS

In this paper, we have proposed a speaker recognition system using AANN. MFCC features are extracted from the voice signal that can later be used to represent each speaker. Experimental results show that, the characteristics of the speech data are collected from clean and noisy environment using throat microphone. The performance of the system using close-speaking microphone data degrades as the background noise increase, whereas the performance of the system using throat microphone data is likely to be unaffected by the background noise. The proposed system is a generalized work which can be useful for visually impaired people and also in areas which require high security like defence, aircraft and military. The speaker information present in the source features is captured using AANN model. The Experimental results of AANN model using MFCC performs better in clean and noisy environment. In future, throat microphone can be used to analyse the performance of speech impaired people. Various acoustic features can be analysed and the performance of different pattern recognition techniques can be studied.

REFERENCES

- [1] Atal B (1976) Automatic recognition of speakers from their voices. In: Proc. IEEE, pp 460–475
- [2] BYegnanarayana (1999) Artificial neural networks, Prentice Hall of India, NewDelhi
- [3] BYegnanarayana, Kishore S (2002) AANN: an alternative to GMM for pattern recognition. Neural Networks 15:459–469
- [4] BYegnanarayana, Reddy KS, Kishore SP (2001) Source and system features for speaker recognition using aann models. In: IEEE Int. Conf. Acoust., Speech and Signal Processing, Salt Lake City, Utah, USA, pp 409–412
- [5] BYegnanarayana, SV-Gangashetty, SPalanivel (2002) Autoassociative neural network models for pattern recognition tasks in speech and image. In: Soft Computing Approach to Pattern Recognition and Image Processing, World Scientific publishing Co. Pte. Ltd, Singapore, pp 283–305
- [6] DAReynolds (2002) An overview of automatic speaker recognition technology. ICASSP 2002
- [7] Erzin E (2009) Improved throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. IEEE 17(7):1558–7916
- [8] KTomi, Li H (2010) An overview of text independent speaker recognition: From features to supervectors. ScienceDirect, Speech Communication 2010
- [9] LRabiner, Juang B (2003) Fundamentals of Speech Recognition, Pearson Education, Singapore
- [10] MSIkbal, Misra H, Yegnanarayana B (1999) Analysis of autoassociative mapping and neural networks. In: Int. Joint Conf.on Neural Networks, Washington, USA
- [11] NBalakrishnan (2005) Improved Text-Independent Speaker Recognition using Gaussian Mixture Probabilities, M.S thesis, Carnegie Mellon university
- [12] PRupali, Kulkarni M (2010) Analysis of FFSR, VFSR, MFSR techniques for feature extraction in speaker recognition: a review. IJCSI 7(4):1694–0814
- [13] Shahina M B Yegnanarayana (2004) Throat microphone signal for speaker recognition. In: Proc. Int. Conf., Spoken Language Processing
- [14] SHaykin (2001a) Neural Networks: A Comprehensive Foundation, Pearson Education, Singapore
- [15] SHaykin (2001b) Neural networks a comprehensive foundation, Pearson Education, Asia
- [16] TVibha (2010) Mfcc and its applications in speaker recognition. In: International Journal on Emerging Technologies