This full-text paper was peer-reviewed and accepted to be presented at the IEEE WiSPNET 2017 conference.

Speaker Recognition and Verification Using Artificial Neural Network

Neha Chauhan[1] and Mahesh Chandra[2]
Department of Electronics and communication, Birla Institute of Technology, Mesra, Ranchi-835215, India
Email: [1]nutanneha@gmail.com [2]shrotriya@bitmesra.ac.in

***Abstract*—Speaker recognition is a biometric technique which uses individual voice samples for recognition purpose. Speaker recognition is mainly divided into speaker identification and speaker verification. In this paper, a comparative study is made between various combinations of features for speaker identification. Mel frequency Cepstral Coefficient (MFCC) features are combined with spectral centroid and spectral subtraction and tested for improvement in efficiency. Feed forward artificial neural network is used as a classifier. System was tested for 30 speakers. For speaker identification, an average identification rate of 65.3% is achieved when MFCC is combined with centroid features and an identification rate of 60% is achieved when MFCC is combined with spectral subtraction. For speaker verification, an average verification rate of 65.7% is achieved when MFCC is combined with spectral subtraction and a verification rate of 75.3% is achieved when MFCC is used along with centroid.**

***Index Terms*—MFCC, Artificial neural network, Spectral features.**

I. Introduction

Speaker Identification is a process of identifying one particular speaker from the set of different speakers. Biometric recognition through voice is a rising area of research and can be used along with other security checks for user's authentication. During speaker verification one speaker voice is matched with one sample and accordingly speaker is accepted or rejected [1].

The main aim of speaker identification is to identify a speaker from the set of different speakers on the basis of his/her speech features [1]. Speaker identification system uses voice samples to verify individual identity and used in application like voice dialling, voice operated access control systems. Speaker recognition system is mainly divided into two parts-Text dependent and Text independent. During text dependent speaker is not free to say anything, his speech depends on text speaker speech and can be used in 'PIN' type application [1], [2]. While in text independent system speaker is free to say anything with no constraints on what speaker should speak. MFCC are widely used as feature extraction technique and can be obtained by taking Fourier transform of windowed speech signal [3]. Every person voice has different speech quality due to different physiological behavioral characteristics which includes size of the vocal tract, pitch. Other variations in speech signal amongst various individuals are speaking habits and accent [4]. Fig. 1 shows steps for speaker recognition system [1], [2].

Paper is divided into four sections-Database preparation, feature extraction, classification, result and conclusion.

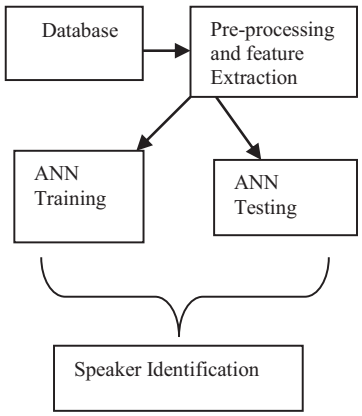

Fig. 1. Automatic speaker recognition system.

II. Database Preparation

Experiment was carried out using voice samples of different speakers. Jharkhand districts and mandies, database of different speakers are recorded on line through mobiles. Each speaker recorded his/her voice samples 32 times be uttering district and mandi name of Jharkhand state. Database of total 30 speakers were recorded at sampling frequency 8 kHz.

III. Feature Extraction

Feature Extraction is a process of extracting useful information of speech samples and discarding unwanted information like noise [5]. Feature extraction techniques are mainly classified as-Spectral feature and temporal feature. Spectral features are obtained by converting time based signal to frequency based like centroid and MFCC. Temporal feature are time based like zero crossing rate [6], [7].

Spectral subtraction method is used to remove noise from the speech samples which increases the efficiency of the system. Spectral subtraction reduces the spectral component which changes more rapidly than speech [8].

A. Spectral Centroid

Spectral centroid define center of gravity of magnitude spectrum of short time Fourier transform and locate the position of formant for each suband. Spectral features improves the performance of MFCC as it calculate information related to vocal source e.g. voiced and unvoiced excitation. Centroid gives single value which represents frequency domain characteristic of a speech signal [6].

978-1-5090-4442-9/17/$31.00

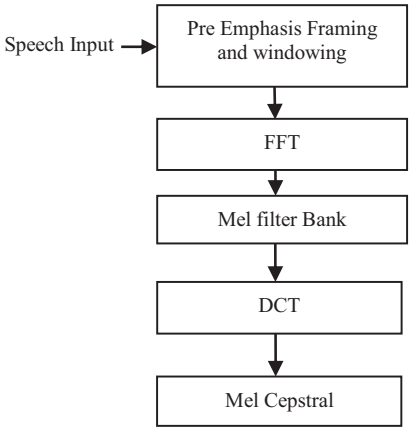

Fig. 2. Block diagram of MFCC.

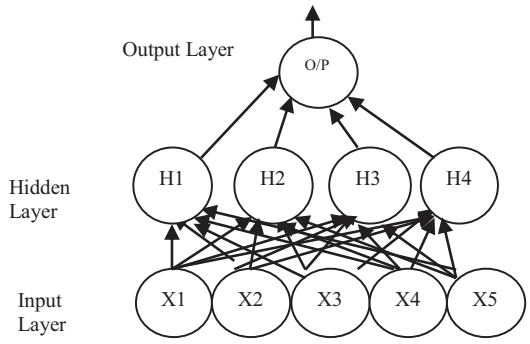

Fig. 3. Artificial neural network.

TABLE I
AVERAGE SPEAKER IDENTIFICATION EFFICIENCY OF DIFFERENT NUMBERS OF SPEAKERS.

Number of speakers	% Accuracy using MFCC	% Accuracy using MFCC + SS	% Accuracy using MFCC + Centroid
5	95	95	96
10	84.7	85.3	85.6
20	75.2	77.3	78.3
30	50	60	65.3

B. Mel Frequency Cepstral Coefficient

MFCC is one of the most popular feature extraction technique used to extract the important feature of speech signal discard all the unwanted information. MFCC is prone to noise which reduces system efficiency and therefore when MFCC values are combined with spectral features increases the efficiency of system. During MFCC speech signals are first divided into small frames consisting of arbitrary number of samples. Overlapping of the frame is done to preserve the smallest unit of sound which are phonemes. Hamming window is applied to each frame for the smooth transition. Filter coefficient of hamming window is calculated using formula (n) is window function

$$W(n) = 0.54 - 0.46\cos(2\pi n/N - 1), \quad 0 \le n \le N - 1$$
$$= \text{otherwise} \tag{1}$$

where N is total number of sample and n is current sample.

Fast Fourier transform of each frame is calculated which speed up the process. Logarithmic Mel scale is applied to FFT frame which is linear up to 1 kHz and logarithmic at greater frequencies. The relationship between frequency of speech and Mel scale can be established as:

$$\text{Frequency (Mel Scaled)} = [2695\log(1 + f(\text{Hz}))/700]. \tag{2}$$

Last step is to calculate discrete cosine transform which de-correlates the features and arranges them in descending order of information, they contain about speech signal. Hence first 13 coefficients are used as MFCC features for creating model [7]. MFCC, spectral centroid and spectral subtraction values are used to train neural network and efficiency is calculated. MFCC steps are shown in Fig. 2.

IV. CLASSIFICATION

Classification is the process of identifying unknown speaker by matching his/her feature with existing database using classifier. In this paper artificial neural network is used as a classifier. Artificial Neural network works in the same way like human brain and consist of different neurons which is used to carry message from one layer to other. Artificial Neural Network mainly consists of three layers-Input layer, hidden layer and output layer. The network has varying neurons input n, which receive input of different sets features. The number of hidden layer varies from 1 to 4 and neurons in each hidden layer varies from 10 to 60 [7] (see Fig. 3).

Multilayer Feed forward Neural Network with back propagation is used to build the model. Features of 30 speakers are taken as an input. Classification is mainly done in three stages which are Training, testing and validation. During training, model of the system is created, validation is set to minimize over fitting of data and testing checks the accuracy of the system that is build during training [1], [2]. Also during testing phase, it is checked whether the speaker is an authorized user or not and this process of detecting authorized speaker is called as speaker verification. 70% of the total samples are taken for training and remaining are taken for validation and testing.

V. RESULT AND DISCUSSION

A. Speaker Identification Result with Different Feature Extraction Techniques

Table I shows the efficiency of different number of speakers with different feature extraction techniques. From above table we observe that highest efficiency is obtained when MFCC features are combined with spectral centroid. Following chart shows the comparison in efficiency between different feature extraction techniques (see Fig. 4).

B. Speaker Verification Result

For speaker verification threshold value is set for the output matrix and speakers voice sample above threshold are considered as an authorised user and remaining speakers with value

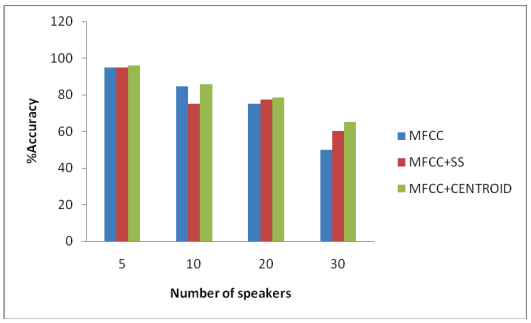

Fig. 4. Comparison chart between different feature extraction.

TABLE II
AVERAGE SPEAKER VERIFICATION EFFICIENCY OF 30 SPEAKERS.

Feature extraction technique	Average verification efficiency (%)
MFCC	50
MFCC + SS	65.7
MFCC + Centroid	75.3

less than threshold are rejected and considered as unauthorized user. Depending on the level of security required, the threshold can be adjusted. Threshold value set should not be too high or very small because very low value sometimes can select false user and very high value may reject true user [11], [12]. Table II shows the verification result of 30 speakers with MFCC, combination of MFCC with centroid value and MFCC with spectral subtraction method.

C. Speaker Verification Result of 30 Speakers

Results of speaker recognition and verification shows that efficiency of system increases when MFCC is combined with other spectral features. For speaker identification system, efficiency increases from 50% to 65.3% and 50% to 60% when MFCC is combined with centroid and spectral subtraction respectively. For speaker verification efficiency increases from 50% to 75.3% using MFCC with centroid features over single MFCC features. In case of spectral subtraction verification efficiency increases from 50% to 65.7%. Combination of MFCC and centroid gives best result in both cases of speaker identification and speaker verification system (see Fig. 5).

VI. CONCLUSION

In this paper, performance evaluation of various combinations of features is successfully studied for speaker identification and verification.

Initially, the system was tested with only MFCC features and a low efficiency of 50% was found for both speaker identification and verification. To improve the efficiency of the system, we experimented on adding more information to our training matrix by appending features of centroid and spectral subtraction. This resulted in improvement of efficiency to 10%–15%. Result was tested for 30 speakers. For speaker identification an average efficiency of 60% is achieved when MFCC is combined with spectral subtraction and an average

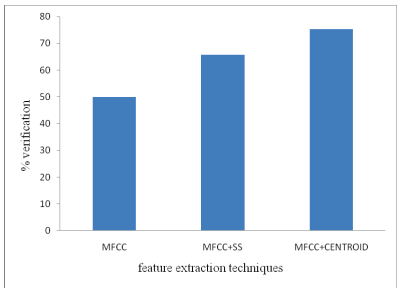

Fig. 5. Speaker verification system comparison chart.

efficiency of 65.3% is achieved when MFCC features are combined with centroid. For speaker verification an average verification efficiency of 65.7% is achieved when MFCC is combined with spectral subtraction and verification efficiency of 75.3% is achieved when MFCC is combined with centroid features.

REFERENCES

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Proc. J Acoustic Soc. Am.*, vol. 55, no. 6, June 1974.
[2] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, Sep. 1997.
[3] M. S. Sinith, Anoop Salim, K. Gowri Sankar, K. V. Sandeep Narayanan, and Vishnu Soman, "A novel method for text-independent speaker identification using MFCC and GMM," in *Proc. IEEE, Audio Language and Image Processing (ICALIP)*, Shanghai, Nov. 2010.
[4] Puebla Cholula, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," *Proc. IEEE*, 978-1-4577-1326-2, Feb. 2012.
[5] Jayant M. Nayak, "Speaker verification: a tutorial," *Proc. IEEE*, vol. 28, no. 1, 0163-6804, Jan. 1990.
[6] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech and Language Processing.*, vol. 15, no. 5, pp. 1711–1723, July 2007.
[7] Noraziahtulhidayu kamarudin, S. A. R. Al-Haddad Shaiful Jahari Hashim, Mohammad Ali Nematollahi, and Abd Rauf Bin Hassan, "Feature extraction using spectral centroid and mel frequency cepstral coefficient for quranic accent automatic identification," *Research and Development (Scored), IEEE Student Conference on*, 16–17 Dec. 2014.
[8] M. Sahidullah and G. Saha, "Design, analysis and experimentalevaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, pp. 543–565, 2012.
[9] WU Zunjing and CAO Zhigang, "Improved MFCC-based feature for robust speaker I identification," *Tsinghua Science and Technology*, vol. 10, no. 2, pp. 158–161, Apr. 2005.
[10] Brett R. Wildermoth and Kuldip K. Paliwal, "Use of voicing and pitch information for speaker recognition," in *8TH Australian International Conference Speech Science and Technology*, 2000.
[11] Ahmad R. Abu-El-Quran and Rafik A. Goubra, "Pitch-based feature extraction for audio classification," in *Haptic, Audio and Visual Environments and Their Applications, 2003. HAVE 2003. Proceedings. The 2nd IEEE International Workshop on*, 21–21 Sep. 2003.
[12] N. Mirghafori and L. P. Heck, "An adaptive speaker verification system with speaker dependent a priori decision thresholds," in *Proc. ICSLP'02*, pp. 589–592, 2002.
[13] B. Anuradha Shafee, "Speaker identification and Spoken word recognition in noisy background using artificial neural networks," in *ICEEOT*, 2016.
[14] R. Togneri and D. Pullella, "An overview of speaker identification: accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, Second Quarter 2011, 2011, pp. 23–61.
[15] Abhilasha Sukhwal and Mahendra Kumar, "Comparitive study of different classifiers based speaker recognition system using modified MFCC for noisy environment," in *International Conference on Green Computing and Internet of Things*, 2016.