SPEAKER IDENTIFICATION IN A MULTI-SPEAKER ENVIRONMENT

Manthan Thakker
Information Technology
Thadomal Shahani Engineering College
Mumbai, India
Email ID: manthanthakker40@gmail.com

Prachi Ved
Information Technology
Thadomal Shahani Engineering College
Mumbai, India
Email ID:prachiv2612@gmail.com

Shivangi Vyas
Information Technology
Thadomal Shahani Engineering College
Mumbai, India
Email ID: shivangivyas.1812@gmail.com

Shanthi Therese S.
Information Technology
Thadomal Shahani Engineering College
Mumbai, India
Email ID:shanthitherese123@gmail.com

Abstract. Human beings are capable of performing unfathomable tasks. A human being is able to focus on a single person's voice in an environment of simultaneous conversations. We have tried to emulate this particular skill through an artificial intelligence system. Our system identifies an audio file as a single or multi-speaker file as the first step and then recognizes the speaker(s). Our approach towards the desired solution was to first conduct pre-processing of the audio (input) file where it is subjected to reduction and silence removal, framing, windowing and DCT calculation, all of which is used to extract its features. Mel Frequency Cepstral Coefficients (MFCC) technique was used for feature extraction. The extracted features are then used to train the system via neural networks using the Error Back Propagation Training Algorithm (EBPTA). One of the many applications of our model is in biometric systems such as telephone banking, authentication and surveillance.

Keywords: Speaker identification, neural network, Multi-Speaker, Mel Frequency Cepstral Coefficients (MFCC).

1 INTRODUCTION

Speaker recognition is defined as identifying a person based on his/her voice characteristics. This is useful in applications for authentication to identify authorized users i.e., enable access control using voice of an individual. Most of the times there are scenarios where multiple speakers speak simultaneously [2], [7]. Single speaker identification systems fail to handle such audio signals. Therefore, there it is essential to make the speaker recognition systems to handle multi-speaker audio files and classify them [2], [7].

2 REVIEW OF LITERATURE

The paper that we have chosen as the foundation of our project is a technical paper [1] written by Wei-Ho Tsai and Shih-Jie Liao from the National Taipei University of Technology. The paper highlights the issue of identifying separate speakers in a multi-speaker environment.

The paper introduces 'Single Speaker Identification' which has seen a lot of development and success and goes on to explain the problem of multiple speakers and their identification in a conversation.

The important applications of multi-speaker identification are also listed, which include the likes of suspect identification in police work and automated minuting of meetings. The paper further explains two approaches to solving this problem.

1. A two stage process where the signal is first tested to identify whether it contains speech from a single speaker or from multiple speakers.
2. The second approach is a single stage process that carries out the single speaker and multi-speaker identification in parallel.

3 SYSTEM DESIGN

3.1 Mel frequency cepstrum coefficients

This method was implemented as a feature extraction technique [3], [5]. To pre-emphasize speech signal, a high pass filter is implemented in this process. As speech is a non-stationary signal, which means the statistical properties of such speech is not constant all the time, we assume that the signal is made stationary by using a window of frame size 25ms and frame shift of 10ms [5]. We then apply the MFCC algorithm to determine 20 coefficients for the data set.

2.

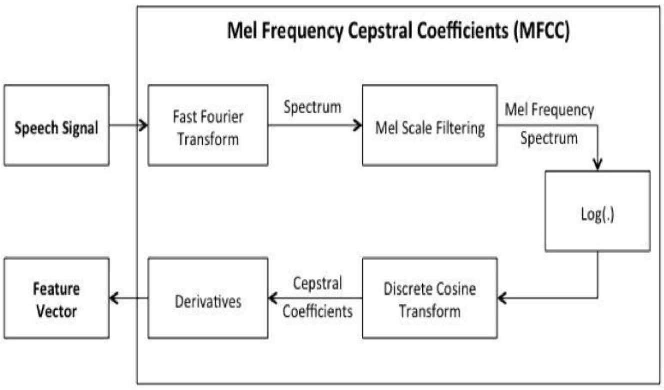

Fig. 3.1. MFCC process

3.2 Normalization

In order to make the neural network training more efficient, we have normalized the input features [6]. In this system, the values are normalized in the range of 0 to 1.

Therefore, the normalization technique used is:

1. Select the maximum value from the input data set.

2. Divide all data set values by the maximum value to get the normalized matrix.

Suppose X is the input matrix and x is the normalized matrix then

n=max(X);

x=X/n;

3.3 Neural networks

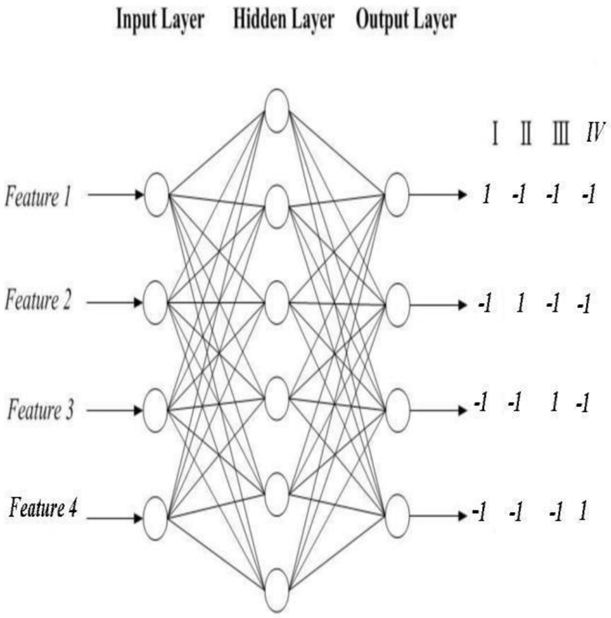

Fig. 3.2. Back Propagation algorithm methodology

We have used neural networks for training and testing the data set.

1. The neural network consists of one input layer, one hidden layer and one output layer of neurons [4], [6]. The input neurons correspond to the features extracted (MFCC) per frame. An input matrix consisting of all features is given as input to the neural network. There are nine neurons in our output layer for nine speakers to be recognized, i.e. one neuron for one speaker.
2. The basic neural network having four outputs is as shown in figure 3.2. If the identified speaker is speaker 1, the first output neuron gives an output of 1 and the rest output neurons give an output of -1. Similarly, for second, third and fourth speaker, output neurons 2, 3 and 4 are fired and they give an output of 1 respectively.

3. The number of hidden neurons depends upon the number of hyperplanes required to correctly classify the input set into individual speakers in n-dimensional space [6] (In our case 20 dimensional space).

3.4 Tools used

1. Audacity :

   This tool was used to digitally mix the audio files for multi-speaker recognition. It was also used to pre-process the audio files before using the audio files for training.

2. Text2speech.org

   This site was used to generate audio files which were used for testing and training. The data set consisted of 10 audio files per speaker i.e., each speaker speaking the digits 0-9 and some words[9]. There are 25 recordings which serves as multi-speaker files [8].

3. Matlab

   This software was used to acquire MFCC and to train and test the system using Error Back Propagation Training Algorithm (EBPTA).

4 RESULT ANALYSIS

Table I. Result obtained

Sr. No.	Training Set (No. of recordings)	Test Set (No. of recordings)	Accuracy (%)
1.Single Speaker	9	110	95.45% (105 recordings identified correctly)
2.Multi-Speaker	--(same training set as single speaker)	25	88% (22 recordings identified correctly)

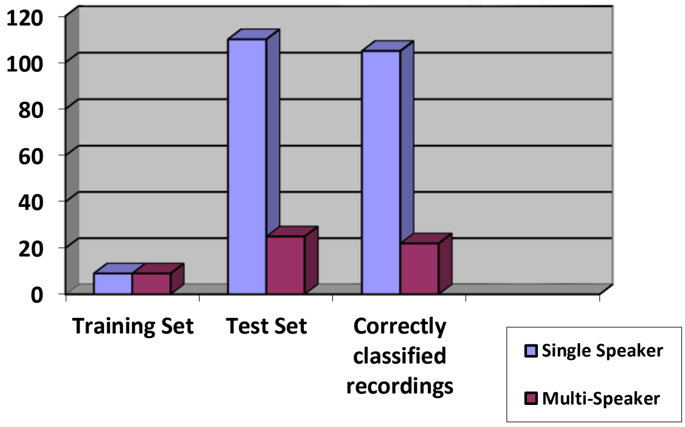

Fig. 4.1 Result Bar Chart

5 CONCLUSION

We were successful in identifying all the speakers using the Mel Frequency Cepstral Coefficients technique for feature extraction and Error Back Propagation Training Algorithm (EBPTA) for feature matching. To make our system more robust and adaptable to real life application, our system also identifies speakers in a multi-speaker environment. As the algorithm used is neural network that basically tries to mimic the working of a human brain, it is always adaptable to learning with new datasets. The multispeaker environment detection and learning capability of our system are the novel and user friendly features of our propesd system.

6 FUTURE WORK

1. Speaker identification using large data sets:

To make an application for identifying speakers in real-time, it is necessary to use cluster of computers for training to utilize parallel computing in neural networks. Various technologies such as ‘MapReduce’ [10] could be used for large datasets for training.

2. Speech Diarization:

As the system proposed by us gives us frame by frame classification, one can easily perform speech diarization i.e. identifying who speaks when.

3. Speech Isolation:

Once the frames are identified, one can also isolate the speech in multi-speaker environment so as to understand what each individual said.

4. Speech Recognition (speech to text conversion):

Similar architecture could be used to develop a system wherein speech could be accurately determined by the system which means identifying the letters, words and number being spoken.

7 ACKNOWLEDGMENT

Our special thanks to Mr. Arun Kulkarni, our Head of Department (Information Technology) for his cooperation and unconditional support. As our teacher he provided us with his useful insights and extended a helping hand whenever it was required.

We are highly indebted to the faculty of Thadomal Shahani Engineering College for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in making the project.

REFERENCES

[1] Wei-Ho Tsai and Shih-Jie Liao, “Speaker Identification in Overlapping Speech”, Journal Of Information Science and Engineering paper published in 2010.

[2] Barry Arons, “A Review of The Cocktail Party Effect”, MIT Media Lab.

[3] Amit Sahoo and Ashish Panda, “Study of Speaker Recognition Systems”,National Institute of Technology,Rourkela, 2011.

[4] PPS Subhashini, Dr. M.Satya Sairam ,Dr. D Srinivasarao,”Speaker Identification with Back Propagation Neural Network Algorithm”, International Journal of Engineering Trends and Technology paper published in 2014.

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi,“Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Volume 2, Issue 3, Journal of computing,, March2010.

[6] Noor Khaled, Saad Najiam Al Saad,” Neural Network Based Speaker Identification System Using Features Selection”, Department of Computer Science, Al-Mustansiriyah University, Baghdad,Iraqi.

[7] M.K. Alisdairi, ”Speaker Isolation in a “Cocktail-Party” Setting”, Term Project Report ,Columbia University,2002.

[8] Ms. Asharani V R, Mrs. Anitha G , Dr. Mohamed Rafi,” Speakers Determination and Isolation from Multispeaker Speech Signal”, Volume 4 Issue 4, International Journal of Computer Science and Mobile Computing,,April 2015

[9] Douglas A. Reynolds,” Automatic Speaker Recognition Using Gaussian Mixture Speaker Models”, Volume B, Number 2, The Lincoln Laboratory Journal,1995.

[10] Changlong Li, Xuehai Zhou, “Implementation of Artificial Neural Networks in MapReduce Optimization “ University of Science and Technology of China, 2 Texas Tech University,US,2014.