International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

An MFCC Based Speaker Recognition using ANN with Improved Recognition Rate

Sundeep Sivan[1], Gopakumar C.[2]
[1]M.Tech. Scholar, Department of Electronics, College of Engineering Chengannur, Kerala, INDIA
[2]Associate Professor, Department of Electronics, College of Engineering Karunagapally, Kerala, INDIA

Abstract: *Speaker recognition is used to recognize persons from their voice. It has many applications such as in security systems, database access services, forensics etc. In most of the today's literatures for improvement of speaker recognition system are limited to the study of feature matching techniques. This paper deals with a text dependent speaker recognition system using neural network and also proposing a method to improve the accuracy of recognition by changing the number of Mel Frequency Cepstral coefficients (MFCC) used in training stage. Voice Activity Detection (VAD) is used as a preprocessing step to further improve the accuracy*.

Keywords: *Speaker Recognition; MFCC; VAD; Raspberry Pi; Neural Network; Speech Processing*

I. Introduction

The speech signal conveys many levels of information to the listener. At the primary levels speech conveys a message. Also speech conveys information about the gender, emotion and the identity of the speaker. The goal of the speaker recognition is to extract and characterize the information about a speaker identity from speech signal. There are two types of speaker recognition tasks such as Text-dependent speaker recognition and text-independent speaker recognition. In text-dependent speaker recognition, recognition phrases are fixed, whereas in the case of text-independent speaker recognition, the recognition phrases are not fixed.

Speaker recognition consists of mainly two steps. They are feature extraction and classification stage [1]. In feature extraction stage, features are extracted from the speech signal. There are many methods for feature extraction. Most widely used methods are MFCC (Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coding), LPCC (Linear Predictive Cepstral Coefficient) etc. [2]. Also there are many methods for feature matching. They are HMM (Hidden Markov Model), DTW (Dynamic Time Warping), ANN (Artificial Neural Network) etc. In this paper we describe about a text-dependent speaker recognition using Artificial Neural Network and proposes a method to improve the accuracy of the speaker recognition system. Also hardware implementation of the speaker recognition system is done.

In most of the today's literatures for improvement of accuracy of speaker recognition system are limited to the study of the feature matching stage. The main disadvantage of this method is that it is very complex. To avoid this complexity here we propose a method which improves the accuracy of the system by changing the parameters in feature extraction stage. Voice Activity Detection is used as a preprocessing stage to further improve the accuracy.

II. Literature Survey

Speech is produced when air is pushed through the trachea and vocal folds. There are two traits that determine speaker-specific characteristics: physical and learned traits. Learned traits are speaking-rate, timing patterns, pitch usage and dialect. Physical traits are formed by the vocal tract which includes size and shape of laryngeal pharynx, oral pharynx, oral cavity, nasal cavity and nasal pharynx. It is common for speaker verification systems to mainly depend on characteristics derived from the vocal tract. Although the high-level cues (learned traits) are more robust and are not much affected by noise and channel mismatch, here we limit scope in the low-level cues (physical traits) because they are easy to be automatically extracted and suitable for our purpose.

A) Classification of Speaker Recognition Systems

1. *Text-dependent vs. Text-Independent*

   In the text-dependent speaker recognition system a fixed word is always used for training and testing. In this scenario a two leveled security system is obtained. Firstly, the voice has to be produced by an authorized user and secondly, the user has to provide the proper password. In text-independent speaker recognition system there is no specific word for training and testing.

2. *Speaker Verification vs. Speaker Identification*

   In a speaker verification system, the user will first provide his identity to the user and then the system will check if this identity is correct by analysing the speaker voice. In a speaker identification system, user does not need to provide his identity to the system. Instead the system will check which speaker model will matches the input. And makes a decision based on that.

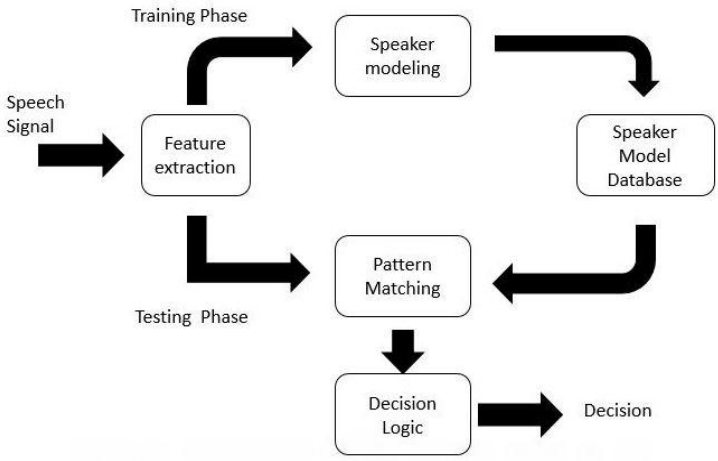
3. Open-set vs. Closed-set

A closed-set recognizer will consider that the user who is trying to access the system belongs to a group of known users. However, an open-set system will consider the possibility that the user who is attempting to enter in the system can be unknown.

III. Methodology

In speaker recognition systems, the two main operations performed are feature extraction and feature classification [1]. The feature extraction is a data reduction process that captures speaker specific properties. There are various methods for feature extraction of speech signals. They are linear prediction coefficients (LPCs), the Mel-Frequency Cepstral Coefficients (MFCCs) and the Linear Prediction Cepstral Coefficients (LPCCs). Classification step consists of two phases; speaker modelling and speaker matching. In the speaker modelling step, the speaker is enrolled to the system using features extracted from the training data. When a sample of data from some unknown speaker arrives, pattern matching techniques are used to compare the features from the input speech sample to a model corresponding to a known speaker. The combination of a speaker model and a matching technique is called a classifier. A general block diagram of a speaker recognition system is shown below.

Figure 1. **Speaker Recognition System**



Speaker recognition system can be divided into two steps. Feature extraction and Classification. The classification module has two components: pattern matching and decision. The feature extraction module extracts a set of features from the speech signal that represent some speaker-specific information. The pattern matching module is responsible for comparing the estimated features to the Speaker models. In this work we are using MFCC (Mel Frequency Cepstral Coefficients) as features and ANN (Artificial Neural Network) as feature matching technique.

A. Voice Activity Detection (VAD)

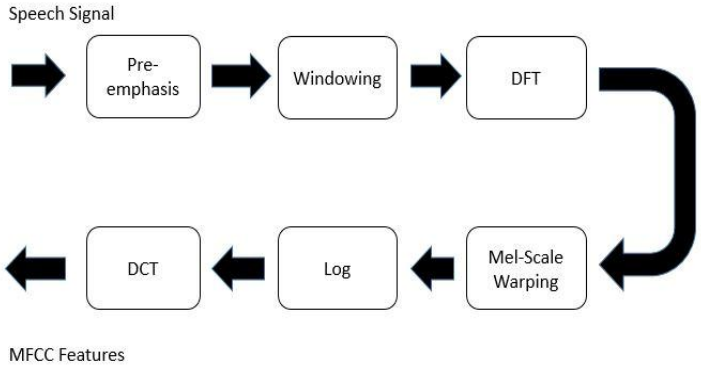
Voice activity detection is a technique to detect the presence or absence of human speech in a waveform. The challenge of voice activity detection is to find an efficient algorithm to discriminate speech from non-speech in a sound file. A human being can easily recognize human speech from other sound. But from a computer's perspective it is not an easy task. In the case of a computer it is solved by finding features in human speech that makes it possible to identify speech in a sound file [7].

Here we use three different features per frame. The first feature is the widely used short-term energy (E) [6]. Energy is the most common feature for speech/silence detection. When a person is speaking, he excites energy with his vocal cords. Therefore energy is most dominant feature to detect speech. By summing the squared absolute value of each sample in a frame, the energy feature is calculated. The most dominant frequency component of the speech frame spectrum can be another very useful feature in discriminating between speech and silence frames [7]. This feature is simply computed by finding the frequency corresponding to the maximum value of the spectrum magnitude. Zero-crossing rate (ZCR) is another feature used for Voice Activity Detection. Zero crossing rate of noise is larger than that of speech. If the sign is shifted between two adjacent samples, we have a zero-crossing. By summing up every zero-crossing in a frame, the zero-crossing rate is calculated. In each frame Energy, Most Dominant frequency and Zero Crossing Rate are find out and compare it with a predefined threshold values, and if any 2 of them satisfies the condition, then that frame is marked as speech frame otherwise it is discarded.

B. MFCC

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. [3]. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. Block diagram of MFCC is shown below.

Figure 2. MFCC Block Diagram



The First step is to pre-emphasis the signal. Here the signal is sent through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequencies. The next step is framing and windowing. Hamming window is used for this purpose. The windowed signal is then undergone Discrete Fourier Transform (DFT). The output of the DFT module is then warped on mel scale. The fifth step involves taking logarithm. And in the last step, Discrete Cosine Transform (DCT) is done for calculating MFCC.

C. Artificial Neural Network

A neural network is a machine learning approach that is designed to model the way in which how human brains performs a particular task. A neural network consists of a massive interconnection of nodes, called neurons. A neural network consists of an input layer, an output layer and one or more hidden layers. Here we are using a feed forward Neural Network for pattern matching with 2 hidden layers.

D. Database

A database which consists of word "hello" spoken by 6 persons (3 males and 3 females) 70 times is created in a studio quality .Sampling rate is fixed at 16Khz and stored it in .wav format.

IV. Results

In this section speaker experiments are carried out in different conditions. Voice Activity Detection is used to find the start and end point of a word. As a preprocessing step, voice activity detection is implemented here.

Figure 3. Voice Activity Detection Output

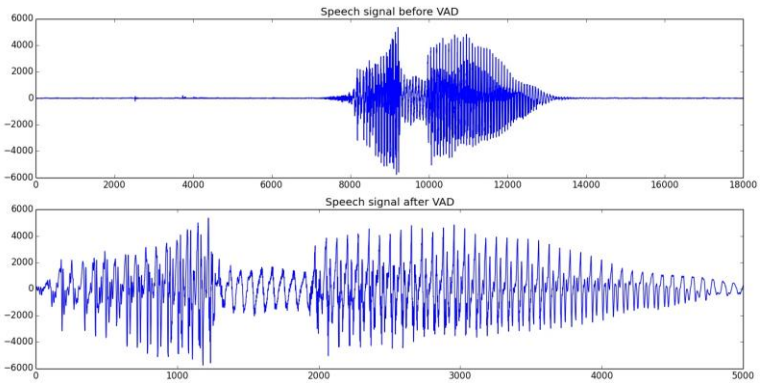


Figure 3 shows the output of Voice Activity Detection. From the figure it is clear that Voice activity detection successfully eliminated silent part from speech signal.

In the training phase of automatic speaker recognition, a database is first created.3 males and 3 females are used to generate this database, each one uttered the word "hello" 70 times. From this database 30 samples

from each speaker is used to calculate MFCC. And this MFCC is used to train the neural network. In the testing phase 40 samples from each speaker is used.

A. Effect of Number of MFCC Coeffcients

24 filter MFCC is used for experiment. The effect of number of MFCC coefficients in speaker recognition task is analyzed. The result is summarized in the following tables.

No. of MFCC coefficients = 14			
Speaker	**No. of Attempts**	**Correctly Recognized**	**Wrongly Recognized**
S1	40	28	12
S2	40	24	16
S3	40	30	10
S4	40	27	13
S5	40	29	11
S6	40	31	9
Total	240	169	71

No. of MFCC coefficients = 16			
Speaker	**No. of Attempts**	**Correctly Recognized**	**Wrongly Recognized**
S1	40	31	9
S2	40	30	10
S3	40	32	8
S4	40	29	11
S5	40	31	9
S6	40	29	11
Total	240	182	58

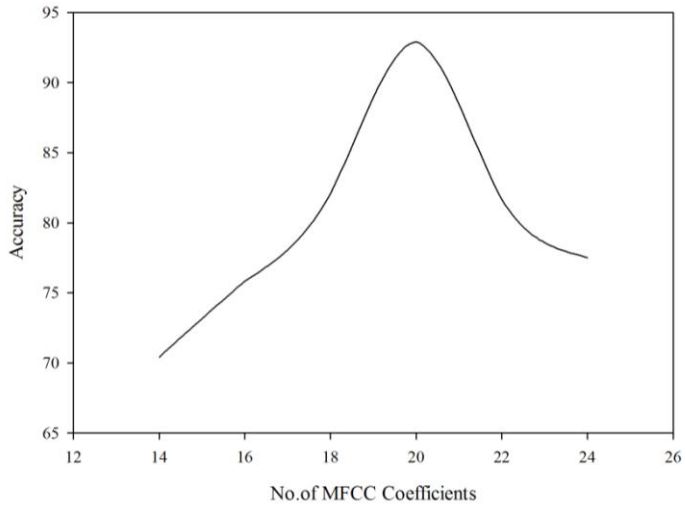
No. of MFCC coefficients = 18			
Speaker	**No. of Attempts**	**Correctly Recognized**	**Wrongly Recognized**
S1	40	32	8
S2	40	33	7
S3	40	34	6
S4	40	31	9
S5	40	35	5
S6	40	32	8
Total	240	197	43

No. of MFCC coefficients = 20			
Speaker	**No. of Attempts**	**Correctly Recognized**	**Wrongly Recognized**
S1	40	35	5
S2	40	37	3
S3	40	38	2
S4	40	36	4
S5	40	38	2
S6	40	39	1
Total	240	223	17

No. of MFCC coefficients = 22			
Speaker	**No. of Attempts**	**Correctly Recognized**	**Wrongly Recognized**
S1	40	34	6
S2	40	33	7
S3	40	33	7
S4	40	34	6
S5	40	30	10
S6	40	32	8
Total	240	169	44

No. of MFCC coefficients = 24			
Speaker	**No. of Attempts**	**Correctly Recognized**	**Wrongly Recognized**
S1	40	33	7
S2	40	32	8
S3	40	30	10
S4	40	31	9
S5	40	29	11
S6	40	31	9
Total	240	169	54

Figure 4. MFCC Coefficients vs



From the above graph it is clear that as the number of MFCC Coefficients increases, accuracy also increases. But after a particular point accuracy starts decreasing. So accuracy of the speaker recognition system can be improved by increasing the number of MFCC filter coefficients up to a particular point. So, for optimum performance of the speaker recognition system not less than 4/5th of total number of MFCC coefficients should be taken.

B. Hardware Implementation

The advantage of hardware implementation is that, computation time will be less. We successfully implemented speaker recognition system in Raspberry Pi Also as a future work we would like to make our speaker recognition system real time.

Figure 5. Hardware Implementation of Speaker Recognition System

V. Conclusion

An efficient method to improve the accuracy of text-dependent speaker recognition system is discussed. Present methods of improvement of accuracy of speaker recognition system mainly depend upon changing parameters in the speaker matching stage. The proposed method improves the accuracy of speaker recognition system by changing the number of MFCC coefficients used for training with VAD as a preprocessing stage. Using the proposed method maximum accuracy of 92.91% is achieved. Speaker recognition system is implemented in Raspberry Pi.

References

[1] Wael Al-Sawalmeh, Khaled Daqrouq,Abdel-Rahman Al-Qawasmi, “Use of Wavelets in Speaker Feature Tracking Identification System Using Neural Network”, *WSEAS Transactions on Signal Processing*,May 2009.

[2] Joseph Keshet, SamyBengio, "Automatic Speech and Speaker Recognition",*WILEY*.

[3] John R. Deller Jr., John H.L.Hansen, John G. Proakis, “Discrete Time Processing of speech signal”, *IEEE press classic*.

[4] HomayoonBeigi,”Fundamentals of Speaker Recognition”, *Springer*, 2011.

[5] LindasalwaMuda, MumtajBegam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", *Journal of Computing*, Volume 2, Issue 3, March 2010, ISSN 2151-9617.

[6] S. Gökhun Tanyer and Hamza Özer, "Voice Activity Detection in Nonstationary Noise", *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 4, JULY 2000.

[7] M. H. Moattar and M. M. Homayounpour, “A simple but efficient real-time voice activity detection Algorithm”, *17th European Signal Processing Conference (EUSIPCO 2009)*, pp.2549-2553, August 24-28, 2009.

[8] James A. Freeman David M. Skapura, "Neural Networks Algorithms, Applications, and Programming Techniques",*Addison-Wesley*.