IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATIC SPEAKER RECOGNITION OF SINGLE DISTANT AND MULTIPLE DISTANT MICROPHONE SIGNALS USING BIC

[1]Umaisa Hassan [2]Sukhvinder Kaur

[1]Student,[2] AP & head of the department

Department of electronics and communication engineering, SDDIET, Barwala, India

Abstract: This paper introduces an idea of Speaker Recognition (SR) system. Speech is considered as most reliable form of communication between humans, which gave the idea for the production of human machine interface technology. One of the important aspects of human machine speech interface technology is speaker recognition which is used in Biometric system. SR is the process of automatically perceiving the oneness of a speaker on the basis of discrete information extracted from their speech signals. It tends to be extensively grouped into two sections: Speaker Identification and Speaker Verification. Also there are two modes to record the database for SR: Single Distant Microphone mode (SDM) and Multiple Distant Microphone mode (MDM). In this paper features of both SDM and MDM signals have been extracted using MFCC and proposed algorithm. For feature matching Bayesian Information Criterion (BIC) has been used, finally the performance has been accessed using ROC (Receiver operating Characteristics) and DET (Detection Error Trade-Off).

Index terms: **Single Distant Microphone (SDM) signal, Multiple Distant Microphone (MDM) Bayesian information criterion (BIC), Mel Frequency Cepstral Coefficient (MFCC), ROC (Receiver operating Characteristics) and DET (Detection Error Trade-Off).**

INTRODUCTION

This chapter includes basic introduction to understand what Speaker recognition system is , how it works, what are its components. Also the areas where it is put to use or can be used in future is discussed here.

In our regular day to day existences there are numerous types of correspondence, for example: non-verbal communication, literary language, pictorial language and discourse, and so forth. Anyway among these structures discourse is constantly viewed as the most impressive structure in light of its rich measurements character. Aside from the discourse text (words), the rich measurements likewise allude as the sex, demeanour, feeling, wellbeing circumstance and character of a speaker. Such data is significant for a viable correspondence[1]. The Speech is one of the most important tools for communication between humans. Therefore manufacturing of ASR is need for human being all the time. Speech recognition made it feasible for machine to understand human languages[2]. From most recent twenty years increasingly more consideration has been paid on speaker Recognition field. Speaker Recognition includes two stages: speaker identification and speaker Verification and is the cycle of consequently perceiving who is talking based on singular data removed from discourse signals. This method makes it conceivable to utilize the speaker's voice to check their character and control admittance to administrations, for example, voice dialing, banking by phone, phone shopping, data set admittance administrations, data administrations, voice message, security control for classified data territories, and far off admittance to PC's [3,4]. Speaker verification (*SV*) is defined as the process of determining whether the speaker identity is who the person claims to be. There are a various terms that have same meaning as that of speaker verification, such as voice verification, voice authentication, speaker/talker authentication, talker verification. A one-to-one comparison (also called binary decision) between the features of an input voice and those of the claimed voice that is registered in the data base is done in speaker verification.

In this paper, section 1 includes introduction of the system, section 2 describes proposed speaker recognition system, section 3 includes results and discussion having ROC and DET curves. Finally conclusion and future scope finds its place in last section.

1. SPEAKER RECOGNITION SYSTEM

In every speaker recognition system there are two main processes; feature extraction and feature matching. Feature extraction is the process that takes an adequate amount of data from the speech signal that can later be used to address each speaker. Feature matching involves the certifiable framework to identify the unknown speaker by contrasting highlighted features from his or her voice input with the ones from a database of known speakers[5].

Speaker recognition in humans is not merely based on pitch of the speech signal. In fact it is a complex combination of many different features. It is important to keep in mind that a machine based recognition system need not to be same as that of human's method. Fig. 1 depicts basic block diagram for Speaker recognition System.

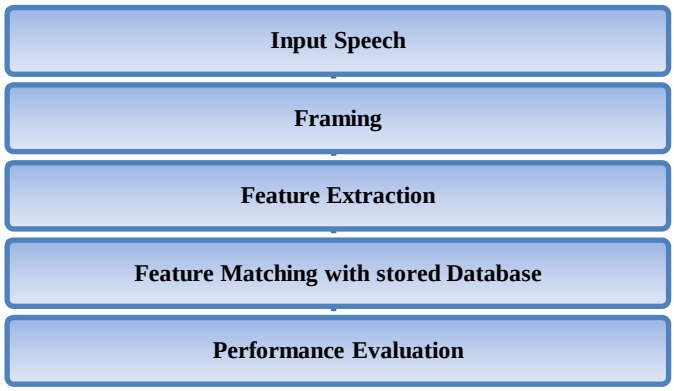

fig. 1. basic block diagram of speaker recognition system

Speaker recognition process includes two steps; speaker identification and speaker verification.

- Speaker identification is a measure of extracting the features associated with the speaker. A lot of feature extraction tools are available these days. The tools used in this paper are DWT, MFCC and NEO. Here test speaker's features are matched with the data base features.
- Speaker verification is a measure of accepting or dismissing the speaker identity by matching algorithm based on their features. Matching can be done using Distance Metric algorithms, BIC is used in this paper so that error is reduced and efficiency is increased.

Fig. 2 gives the methodology that has been used in this paper. It comprises of different blocks. The left hand side of flow chart is for single distant microphone with 16 speakers whose features are extracted using DWT, MFCC and proposed algorithm. On the right hand side are the multi distant 5 microphones, same extraction algorithms have been applied as that of SDM. For matching process BIC algorithm has been used and performance is evaluated using ROC and DET.

2.1 INPUT/DATABASE USED

Database is a collection of speech signals from different speakers. There are total 16 different speakers and two types of data bases. In this paper standard PDA data has been used.

- One database used is Single distant microphone (SDM) containing 16 speech signals.
- Second one is multiple distant microphone (MDM) containing 5 distant microphones and a total of 80 signals are present.

2.2 FRAMING

Since our ears cannot respond to very fast change of speech data continuously thus it is beneficial to frame the data into small portions called frames. Framing is done because frames are easy to handle and work upon

2.3 FEATURE EXTRACTION

Feature extraction or frontend processing focuses on converting the speech waveform to some type of parametric representation for additional investigation and processing, so as to create the speaker discriminative highlights. The speech signal characteristics are fixed for brief periods (between 5 and 100 mSec). However, these characteristics start to change to reflect speaker specific information for longer periods (1/5 sec or more)[6]. Therefore, the speech signal is a gradually time varying signal and, to characterize the speech signal, we use the most common technique: short term spectral analysis[7]. This technique carries the speech in frequency domain with portions of speech through a succession of examinations.

This is critically most important block of recognition system. In this paper Mel Frequency Cepstral Coefficient (MFCC) and Discrete Wavelet Transform (DWT) Combined with Non linear Energy Operator (NEO) which is the proposed algorithm is used.

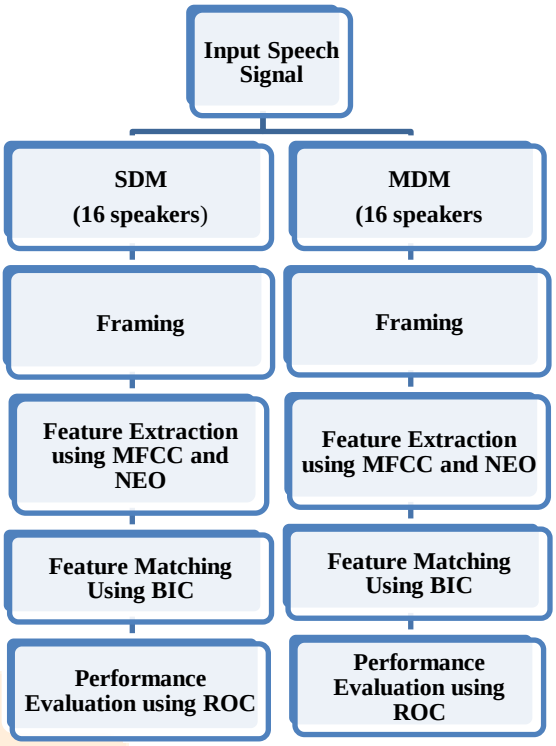

fig.2. design of speaker recognition proposed model with methodology

Mel Frequency Cepstral Coefficients:

MFCC"s depends on the known dissimilarity of the human ears critical bandwidths with frequency; filters separated linearly at low frequencies and logarithmically at high frequencies are used to capture the phonetically important characteristics of speech. It is an approach based on hearing behavior that cannot acknowledge frequencies over 1KHz [8]. The signal is communicated in the MEL scale, which is linear frequency dividing under 1000 Hz and a logarithmic dispersing above 1000 Hz[8,9]. Psychophysical examines have indicated that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. MFCC highlights depend on the known variation of the human ears basic data transmissions with recurrence.

Proposed Algorithm:

The features of the speech signal are first derived using Discrete Wavelet Transform. On the obtained features Non Linear Energy Operator (NEO) is applied.
The nonlinear energy administrators are equipped for accessing the speech signal energy as these administrators are basically energy tracking administrators. The nonlinear differential energy administrators like Teager Kaiser Energy Operator (TEO) can recognize formant AMFM modulations by accessing the product of their time varying amplitude and frequency. The Teager Energy Operator is exceptionally a very high resolution energy operator[10,11]. Rather than experiencing the definite numerical examination of the calculations we show the recreation after effects of those calculations.
Definition of TEO: The nonlinear Teager operator can be defined as in [10]

$$\Psi[x'(t)] = [x(t)]^2 - [x(t)x''(t)] \qquad \text{(I)}$$

And the discrete version of the operator can be defined

$$\Psi[x(n)] = x^2[n] - x[n-1]x[n+1] \qquad \text{(II)}$$

2.4 FEATURE MATCHING

For feature matching Bayesian Information Criterion (BIC) is used. The information stream is a Gaussian cycle in the Cepstral space. We present a maximum likelihood approach to deal with identity turns of a Gaussian cycle the choice of a turn is based on the Bayesian Information criterion (BIC), a model choice standard in the insight writing. BIC is a probability based basis penalized by the model unpredictability. This report focuses on the distance measure between test speaker feature and the database features. We have stored frames of 16 different speakers in 16 different cells in case on Single distant microphone and frames of 16 different speakers with 5 microphones in different 80 locations. Each cell contains the feature extracted from a particular speaker. These segments should be as long as possible, because BIC works best for speaker segments of long length (>3sec). Let us consider two audio segments (i, j) of parameterized acoustic vectors of $X_i = \{x_1, x_2, \ldots x_N\}$ and $x_j = \{x_1, x_2, \ldots x_k\}$of lengths

Bayesian Information Criterion (BIC) is one of the most popular techniques for detecting speaker change point in an audio recording. It's the statistical measure used in statistical hypothesis testing. Let's say the model trained on segment X1 and X2 is M1 and M2 respectively. Then BIC for each segments are,

$$A(x) = \log(p(x_i|M_1)) - \log(p(x|M_{UBM})) \quad \text{(III)}$$

$$B(x) = \log(p(y_i|M_2)) - \log(p(x|M_{UBM})) \quad \text{(IV)}$$

where, $X = \{x_1, x_2, \ldots, x_N; y_1, y_2, \ldots, y_M\}$, x_i ; y_i are the feature vectors, M1 is the model estimated using feature vectors of speaker S1, M2 is the model of speaker S2, UBM is the universal background model. The distance between speaker S1 and S2 is then computed, a smaller value of Td indicates that two speakers are more similar to each other[5].
Analysis is done with distance measuring metrics (BIC) by calculating their score file and comparing it with reference file.

2. RESULTS AND DISCUSSION

2.1 Database used

Database is a collection of speech signals from different speakers. There are total 16 different speakers and two types of data bases.

- One database used is Single distant microphone (SDM) containing 16 speech signals
- Second one is multiple distant microphone (MDM) containing 5 distant microphones and a total of 80 signals are present.

All the speeches have the same language and the language is English. Standard PDA data has been used, which is also helpful to determine the minimum length for speaker recognition system.

a) Performance Evaluation

The performance of both the algorithms is compared using Direct Error Trade OFF (DET) curve and Receiver Operator Characteristics (ROC)

b) Experimental Results

DET curve shows how miss rate and False are related with each other. Graph of DET curves using MFCC and NEO are shown below

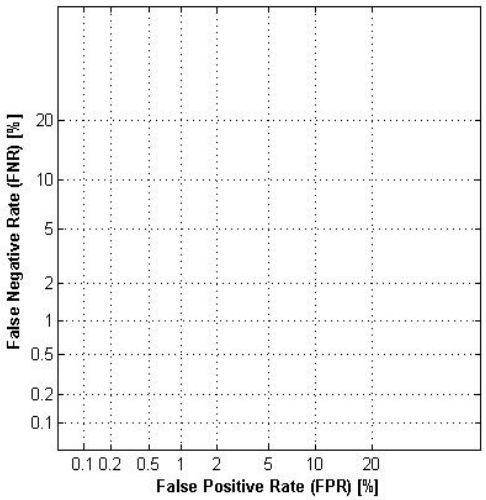

Fig.3.1. DET Curve of SDM Signal Using MFCC

Fig. 3.1 DET curve shows how False Positive Rate and False Negative Rate are related to each other. It shows DET curve of SDM signal using MFCC features. The error rate is 55.7678% providing an efficiency of 44.2322%.

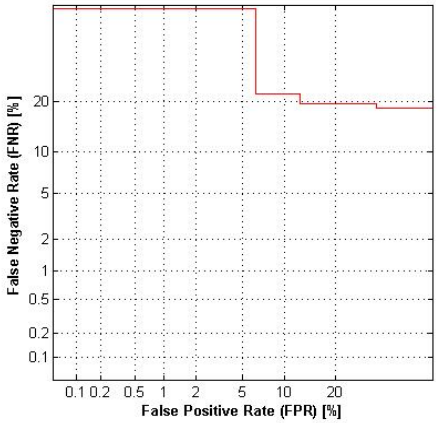

Fig.3.2. DET Curve of SDM Signal Using NEO

Fig. 3.2 shows the DET curve of SDM signal whose features are sought using NEO. It shows relationship between FPR and FNR. The error rate is 18.3825% providing an efficiency of 81.6175%. It is evident from graph that NEO provides better results for SDM signal than MFCC.

Fig. 3.3 shows relationship between FPR and FNR of MDM signal using MFCC. The error rate is 36.4772% providing an efficiency of 63.5228%.

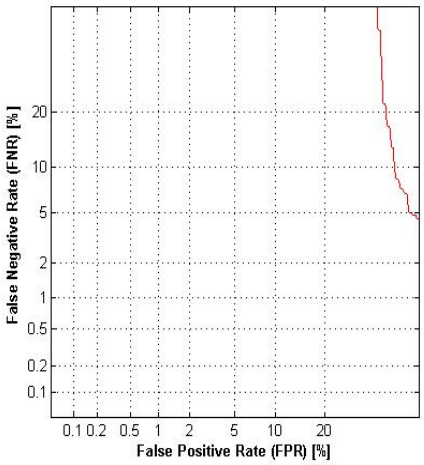

Fig. 3.3. DET Curve of MDM Signal Using MFCC

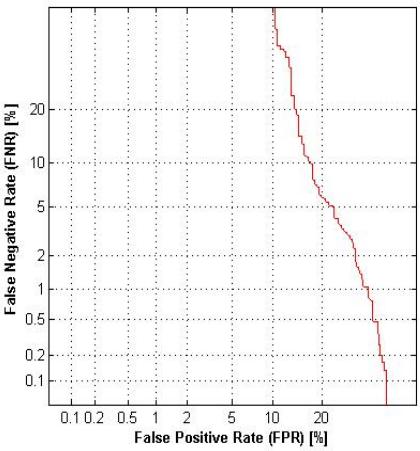

Fig.3.4. DET Curve of SDM Signal Using NEO

Fig. 3.4 shows relationship between FPR and FNR of MDM signal using NEO. The error rate is 14.7309% providing an efficiency of 85.7309%. From all the DET curves above it is clear that the proposed algorithm presents greater efficiency.

Results using ROC Curves:

A Receiver Operating Characteristic (ROC) bend is a graphical plot that shows the indicative capacity of a paired classifier framework as its segregation limit is fluctuated. It is made by plotting the True Positive Rate (TPR) and false positive rate (FPR). Fig.3.5 shows trade-off between FPR and TPR of SDM signal using MFCC with error rate is 55.7678%.

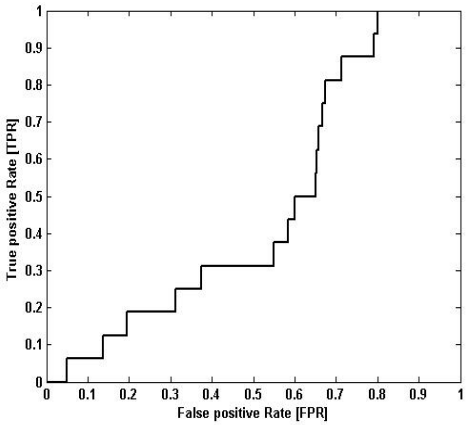

Fig.3.5 ROC Curve of SDM Signal Using MFCC

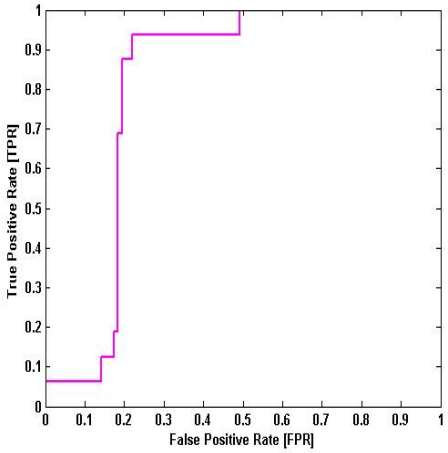

Fig.3.6 ROC Curve of SDM Signal Using NEO

Fig. 3.6 shows relation between FPR and TPR of SDM signal using NEO with error rate is 18.3825% .

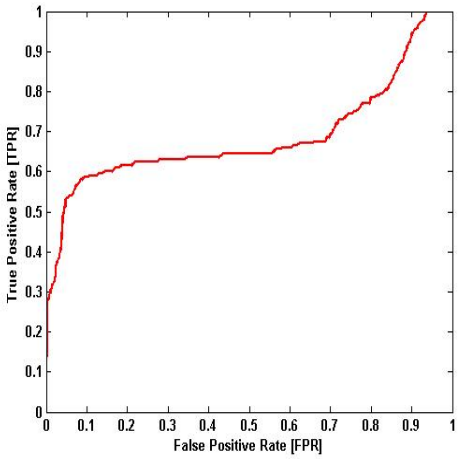

Fig.3.7 ROC Curve Of MDM Signal Using MFCC

Fig. 3.7 shows relation between FPR and TPR of MDM signal using MFCC with error rate is 36.4772% .

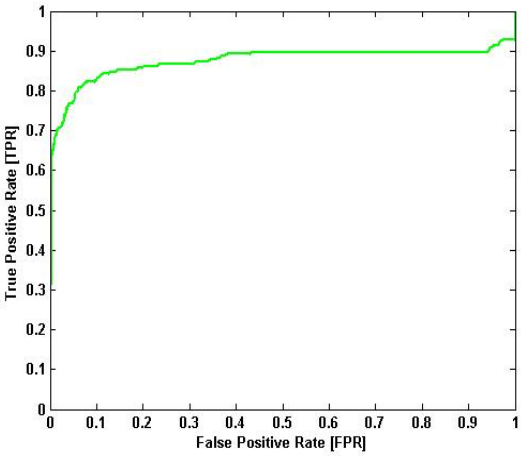

Fig.3.8 ROC Curve of MDM Signal Using NEO

Fig. 3.8 shows relation between FPR and TPR of MDM signal using NEO with error rate is14.7309% . It is clearly visible proposed algorithm provides better results than MFCC.

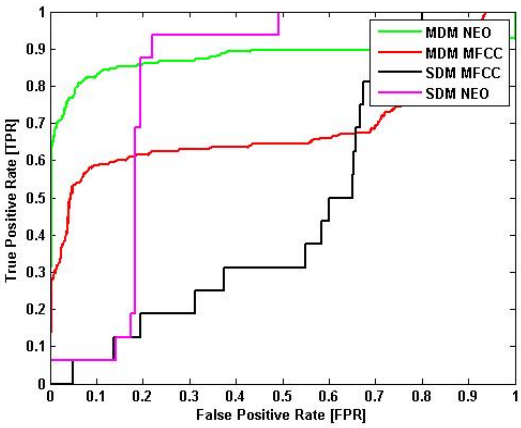

Fig. 3.9 the ROC Curve of Both SDM and MDM Signals Using MFCC and NEO

Fig. 3.9 shows all the four ROC curves of SDM and MDM signals using both MFCC and NEO. Colour coding has been used to distinguish them from one another. It is highly evident from fig. 4.9 that the proposed algorithm provides better results in both SDM and MDM signals.

3. CONCLUSION

Speaker acknowledgment framework is exceptionally famous in voice confirmation for personality and access control to administrations. Speaker ID and check are essential pieces of a speaker acknowledgment framework. In this paper another methodology for speaker recognition has been utilized. For speaker recognition we have used freely available PDA data that contains two types of speech signals recorded using Single distant Microphone (SDM) and multiple distant Microphone (MDM).
In this system firstly we frame the speech signals and then these signals are squeezed using DWT for noise reduction and more acceptable sampling frequency. Further more features of compressed signal are extracted with the help of Mel frequency Cepstral coefficients (MFCC) and non linear energy operator (NEO). These features are further used for identification and verification of speaker's voice. The distance metric incorporated is Bayesian Information Criteria (BIC). At the end results are evaluated with detection error trade of curve (DET) and receiver operator characteristics (ROC) curve. We have compared output of four distances using two different features, MFCC and proposed algorithm. The best result is shown when features are sought using DWT and then on the result NEO is used, which is the proposed algorithm, it is more efficient. It counts minimum false alarm when compared with others. The error percentage in MDM signals using proposed algorithm is 14.7309 which is the least of all.

4. FUTURE SCOPE

- After doing all the hard work and getting results that are accurate up to a large extent we can still say there is scope of further work to be done in speaker recognition system, since it is a vast subject of application today.
- For future, work must be done to develop algorithm which can give better results even with smaller length segments of speech in multi-distant micro-phone signals.
- In future, text dependent systems can be developed and used in practical applications.

The current techniques of feature extraction perform fairly well, but their execution deteriorates in noisy environment. So in future, work can be done in such a way that noise will least affect the performance of an ASR.

REFRENCES

[1] Ling Fen, "Speaker Recognition", KGS lengby, 2004.

[2] Shipra j. Arora, Rishi Pal Singh , "Automatic Speech Recognition :A Review",international journal of computer applications,vol.60,no. 9 ,pp.34, 2012.

[3] Zan Win Aung, A Robust Speaker Identification System , IJTSRD,VOL.2,NO.5, PP. 2057, 2018

[4] R. A. Cole and colleagues, Survey of the State of the Art in Human Language Technology, National Science Foundation European Commission, 1996. http://cslu.cse.ogi.edu/HLTsurvey/ch1node47.htm l

[5] Sukhvinder Kaur, J. S. Sohal, Monica, "Evaluation of Speaker Recognition System Using Different Distance Metrics", International Journal of Scientific Research in Computer Science, Engineering and Information Technology Vol. 2 , no. 5 ,pp. 541-546, 2017.

[6] D. S. Rodr´ıguez, "Text-independent speaker identification," Master's Thesis, AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY KRAKOW, 2008

[7] M. R. Hasan, M. Jamil, M. Rabbani, and M. Rahman, "Speaker identification using mel frequency cepstral coefficients variations", vol. 1, p. 4, 2004.

[8] Varun Shrama, "A Review on Speaker recognition approaches and challenges", International Journal of Engineering Research & Technology (IJERT) Vol. 2, no.5, 2013

[9] Rosé Ramón Calvo de Lara, "A Method of Automatic Speaker Recognition Using Cepstral Features and Vectorial Quantization", M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 146 – 153, 2005. Springer- Verlag Berlin Heidelberg 2005.

[10] Subradeep Pal, "Speech Signal Processing: Non Linear Energy Operator Centric Review", International Journal of Electronic Engineering Research,Volume 4, Number 3 pp. 205-221, 2012.

[11] L.G. Ceballos, J.H.L. Hansen and J.F. Kaiser, "Vocal Fold Pathology Assignment using AM Autocorrelation Analysis of the Teager

[12] EnergyOperator",