Speaker Recognition using Gaussian Mixture Model

H.AaliyaAmreen[#1], K.KhadarNawas[#2],

[#1]*PG Student,* [#2]*Assistant Professor*

[#1]*Department of Computer Science and Engineering, B.S.Abdur Rahman University, Vandalur-48.*
[#2]*Department of Computer Science and Engineering, VIT University, Chennai.*

Abstract- *Speaker recognition is a term which is most popular in biometric recognition technique that tends to identify and verify a speaker from his/her speech data. Speaker recognition system uses mechanism to recognize the speaker by using the speaker's speech signal. It is mainly useful in applications where security is the main and important one. Generally, speech information are recorded though the air microphone and these speech information collected from various speakers are used as input for the speaker recognition system as they are prone to environmental background noise, the performance is enhanced by integrating an additional speech signal collected through a throat microphone along with speech signal collected from standard air microphone. The resulting signal is very similar to normal speech, and is not affected by environmental background noise. This paper is mainly focused on extraction of the Mel frequency Cepstral Coefficients (MFCC) feature from an air speech signal and throat speech signal to built Gaussian Mixture Model(GMM) based closed-set text independent speaker recognition systems and to depict the result based on identification.*

Keywords- *Speaker Recognition, GMM, MFCC, Throat Microphone*

I.INTRODUCTION

Speaker recognition is a biometric procedure that uses an individual's speech for recognition purposes. The speaker recognition process specified by both the physical structure of an individual's vocal tract and the behavioural characteristics of the individual[1]. Speaker recognition technique uses the speaker's voice to verify their identity and provides services such as voice dialling, database access services,[2] information services and voice mail. Speech is a complicated signal produced as a result which provides different levels of medium such as semantic, linguistic and acoustic.[4] Besides, there are speaker related differences which as a result specify a combination of anatomical differences that inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these signals are taken into account and used to discriminate between speakers. Speaker recognition can be classified into different categories such as Open Set vs. Close Set, Identification vs Verification and Text Dependent vs Text independent speaker recognition. Text dependent uses a constrained mode and Text independent uses an unconstrained mode[sy using text dependent speech, the individual utters either a fixed password or prompted phrase that is programmed into the system and this type of system can improve performance especially with cooperative users. A text independent system has no knowledge of the presenter's phrasing and is much more flexible in situations where the individual submits the sample which may be unaware of the collection or unwilling to cooperate, that presents a more difficult challenge[6]. The Fig.1. shows how the speaker recognition is classified based on the trained speakers in the system. An open set system can have any number of speakers[9] that are trained and registered, but in closed set system it can have only a fixed number of users. Identification is the task of determining an unknown speaker's identity, but verification is the process of rejecting the identity claim of a speaker.

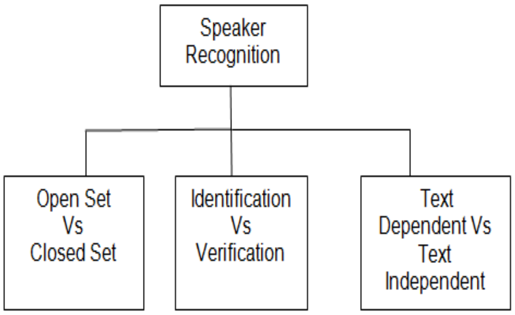

Speaker Recognition
Open Set Vs Closed Set
Identification Vs Verification
Text Dependent Vs Text Independent

Fig.1. Classification of Speakers

.

A. SPEECH SENSORS

It is a best opportunity for speech sensors to provide an multimodal speech processing with subject to automatic speech recognition and speaker recognition systems. Speech sensors provide functional measurement for glottal excitation with vocal tract articulator movements that leads to acoustic disturbances and can supplement the acoustic speech waveform. It provides a brief explanation of how the standard air phone and throat phone were built for creation of features for the recorded sample and also it explains how a Speaker recognition system is built, based on the decision of the output by computing Test Speech and Reference Speech is specified and speaker recognition systems. System sensors provide functional measurement for glottal excitation with vocal tract articulator movements that are relatively immune to acoustic disturbances . It provides a brief explanation of how the standard air phone and throat phone were built for creation of features for the recorded sample and also it explains how a Speaker recognition system is built, based on the decision of the output by computing Test Speech and Reference Speech is specified in Fig.2.

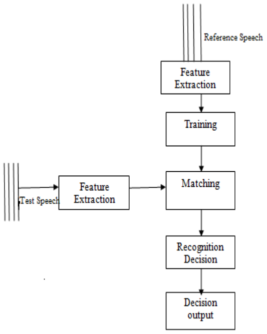


Fig.2. Speaker recognition systems

B .FEATURES FOR SPEAKER RECOGNITION

The speech signal is a form that can be represented by a sequence of feature vectors without the loss of generality. These features are used for speaker dependent and also for speaker independent recognition systems that rely on real life systems.

C.EXTRACTION METHODS

Various methods available for feature extraction are
1)Mel-Frequency Cepstral Coefficients (MFCC),
2) Real Cepstral Coefficients (RCC),
3)Linear Prediction Coding(LPC),
4)Linear Predictive Cepstral Coefficients (LPCC)
5) Perceptual Linear Predictive Cepstral Coefficients (PLPC).

1) Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is one of the most popular technique and commonly used in most of the applications of speech signal for feature extraction.[5] It is based on the human peripheral auditory method. According to human perception the frequency contents of sounds does not follow a linear scale. It is mainly used in speaker or speech recognition systems.

2)Real Cepstral Coefficients(RCC)

RCCs are signals that are transformed from the time domain to the frequency domain by applying a Fast Fourier Transform (FFT) to each frame. The results of this logarthim specify inverse Fast Fourier transform (IFFT)[7]and then it is applied to get the real Cepstrum of the signal,.

3)Linear Prediction Coding(LPC)

This technique analysis the speech signal by estimating the formants. LPC is a form that removes the effects of formants and calculates the intensity and frequency of the remaining buzz from the speech signal[10].

4) Linear Predictive Cepstral Coefficients (LPCC)

It is also a technique with widely used extracted features from speech signal. In this process it specify LPC parameters that can be effectively used to energy and frequency spectrum[12].

5) Perceptual Linear Predictive Cepstral Coefficients (PLPC).

This technique is based on the magnitude spectrum of the speech analysis window. Other techniques such as MFCC and LPC are cepstral techniques while this PLPCC is a temporal technique[11].

II.ARCHITECTURAL DIAGRAM

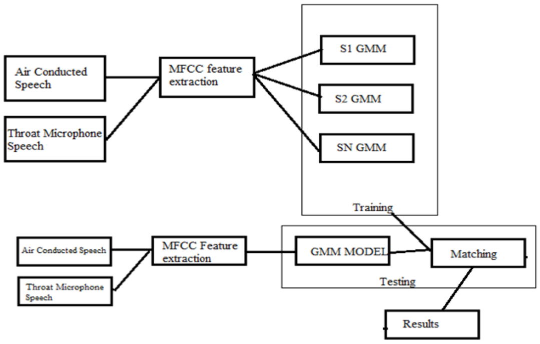


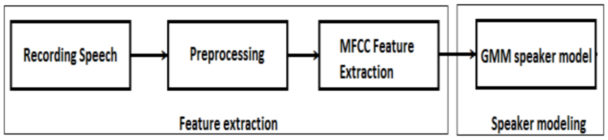
Fig.3. Combination of Air phone and Throat Phone

As such of all other pattern recognition systems this speaker recognition systems also has two phases such as training and testing. Training is the process which specifies familiarity to the system with the voice characteristics of the speakers who are all registered[5]. Testing is the process with actual recognition task. The above block diagram of Fig.3. specifies that in training phase the speech signals from two kinds of phones are depicted and features of those signals are extracted with the help of MFCC and then all the speech signals are modelled using GMM[11]. In testing phase again the speech signals and features for those signals are gathered and extracted accordingly and then the next step of modelling happens then in this step a new formation such as matching the models, and at last results are produced by specifying he is the recognized speaker.

III. DESIGNING SPEAKER RECOGNITION SYSTEM

The speaker recognition system is designed and implemented based on two phases or modules in which one module is based on feature extraction and modelling of datasets and the other module is based on testing and training of datasets.

A.EXTRACTION AND MODELLING



1)Recording Speech

A standard air conduction microphone is used to record the speech. The recorded speech is stored as '.wav' format. Various speeches from different speaker are recorded to train the speaker recognition system. The recorded speech is not clean and contains background noise due to microphone recording. The collected speech requires pre processing to make it suitable for the feature extraction to which sample features are generated based on two techniques of pre processing to make the recorded speech to work properly without occurring any error rate to do this speech data has to be normalized .

2) Pre-Processing

It is the process of removing the unwanted or channel error disturbances to make the sample to process without error rate, During pre-processing it helps in lowering of high frequency energies that are not useful for creating GMM model.

3) Frame Blocking

In this process continuous speech signal are taken which is divided into frames of some N samples, with adjacent frames being separated by some M samples with the value M less than that of N. [11]Considering the first frame with N samples and second frame with M samples overlapping takes place by N - M samples . This step continues until all the speech is accumulated for using one or more frames. For example values of M and N are said to be specified as N = 256 and M = 128 respectively. The N's value is taken as 256 it specifies that speech signal is assumed to be periodic.

4) Windowing

It is a process of taking a small subset of a larger data set, for processing and analysis. A new approach, is that a rectangular window, involves the data set before and after the window by truncating, but not modifying the contents of the window at all.[10] The next step is to minimize the signal discontinuities of each frame at the beginning and end . The below Fig.4. shows how Hamming Window is related in terms of co-efficient and gain based on how frames are segregated.

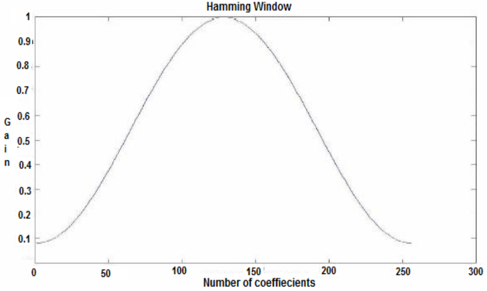
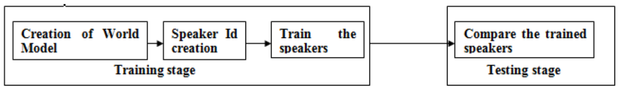


Fig.4. Hamming Window

5) Mel-Frequency Cepstral coefficient(Mfcc)

MFCC Calculation is the second feature extraction method to be used, for which it is based on the known variation critical bandwidths from human ear's with frequency. Filters are spaced linearly at low frequencies and logarithmically at high frequencies that has been used to capture the phonetically important characteristics of speech.[6] This is expressed in mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

B.TRAINING AND TESTING



1) Creation Of World Model

In this step the various speakers samples are collected and it is led to the processing by combining it in one model known as world model.

2)Speaker Id Generation

This process specifies that after the World model gets created to identify the corresponding speaker ,id's are generated.

3) Train The Speakers

After the id's are generated, then it is easier to train speakers for identification

4) Compare The Trained Speakers

After the speakers have been trained, they have led to the process of comparison by comparing the speech sample of trained speaker with the new uttered speech it is depicted that he/she is an authorized/not authorized speakers.

IV. GAUSSIAN MIXTURE MODEL

Definition of GMM specifies that it is the density function with probability parameters that are represented as a weighted sum of Gaussian component densities,[4] It is also a form of parametric model for probability distribution with continuous measurements in biometric system. And the estimation is done for training data using the iterative Expectation-Maximization (EM) algorithm or Maximum Posteriori (MAP) from a well-trained prior model. It is also a form of non-parametric methods for speaker identification .Feature vectors for this GMM are provided in d-dimensional feature space for clustering as they are related to Gaussian distribution,[2] which specifies that each corresponding cluster can be seen as Gaussian probability distribution with features belonging to the clusters of probability values. Below Fig.5. depicts the GMM model with its corresponding feature space. The usage of Gaussian mixture density for speaker identification is motivated by two facts. They are:

- Individual Gaussian classes are represented as the set of acoustic classes. These acoustic classes formulate vocal tract information.
- Gaussian mixture densities provide free approximation to distribute all feature vectors in multi-dimensional feature space.

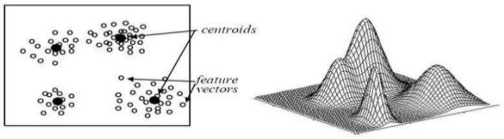


Fig.5. GMM model

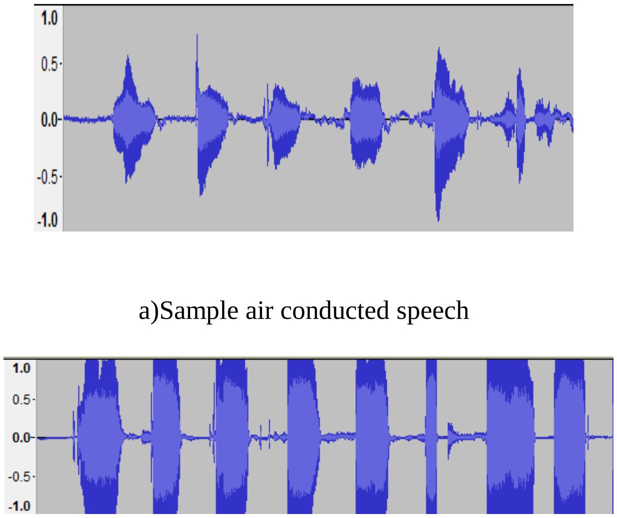
V.BUILDING SPEAKER RECOGNITION SYSTEM

A. Data Collection Process

The speech samples data are recorded for processing in the laboratory environment. The speech sample data from each volunteer speaker is collected one by one using both the TM and NM. Before the actual recording to check speaker adaptability, a 10 sec sample recording is performed to check whether the microphones are placed properly.[2] Volunteer speakers are asked to speak in their natural voice and each recording is checked using audacity to see whether there is any background noise in the signal. If the waveform for the speech signal is good then it is saved. Otherwise, the volunteer speaker is asked to re-record the same content.

B. Speech Data Samples

Below Fig.6. shows the waveform of speech data a)shows the five minutes speech recorded data using normal microphone b)shows the five minutes recorded data using throat microphone.



a)Sample air conducted speech

b)Sample throat conducted speech

Fig.6. Speech Sample data collection with two Microphones

VI. EVALUATION

There are two steps in which speech signal can be evaluated for the creation of models using GMM.

A. Record and play

A command-line sound file recorder which supports several file formats and ALSA soundcard driver with multiple soundcards and multiple devices . It will record a 10-second WAV file with DAT quality on your available soundcard (hw: 0, 0). It is defined as stereo digital audio record for DAT quality with a 48 kHz sampling rate and 16-bit resolution and play the recorded sample.

B. Creation of MFCC features

The speech recognition cannot process directly on speech waveforms it needs tools that has to be represented in a more compact and efficient way. This step is called “pre processing”:

1) The signal is segmented in successive frames , overlapping with each other.

2) Each frame is multiplied by a Hamming windowing function

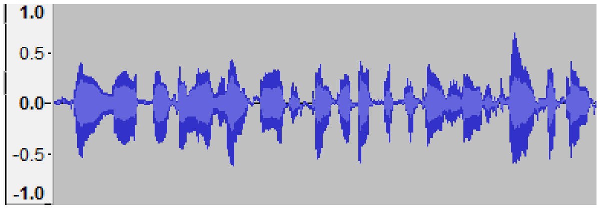
3) A MFCC feature is extracted from each windowed frame.

The conversion of waveform is done based on the HCopy from HTK tool:

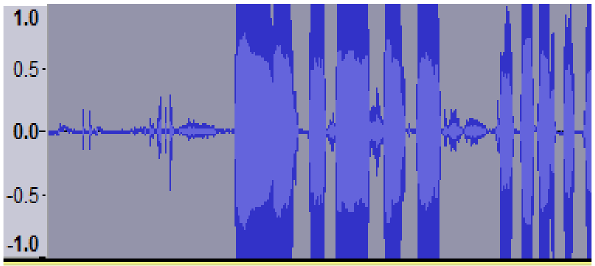
To proceed in this process a configuration file has been set to which the parameters of the coefficient extraction can be done using analysis. and conf.targetlist.txt which specifies the name and location of each waveform to process.

VARIOUS SAMPLES FOR PROCESSING WITH BOTH MICROPHONES

SAMPLE-I

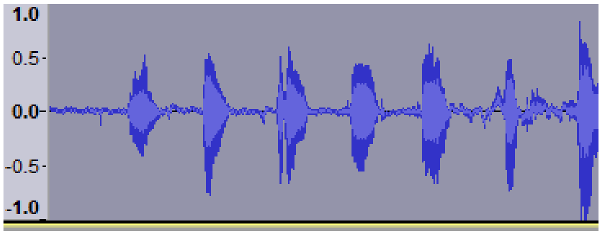


a)Normal Microphone TIMIT Sentence-1

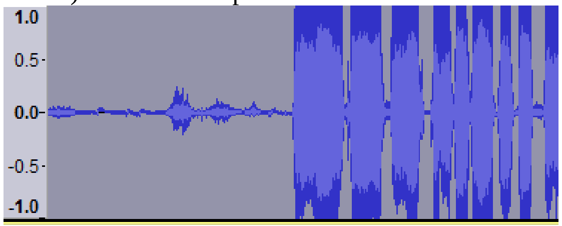


b)Throat Microphone TIMIT Sentence-1

SAMPLE-II

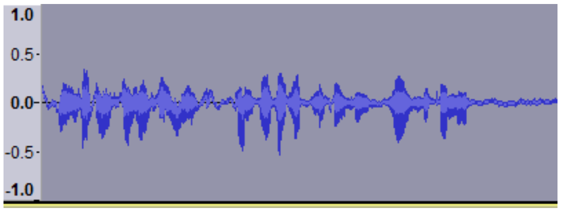


a)Normal Microphone TIMIT Sentence-2

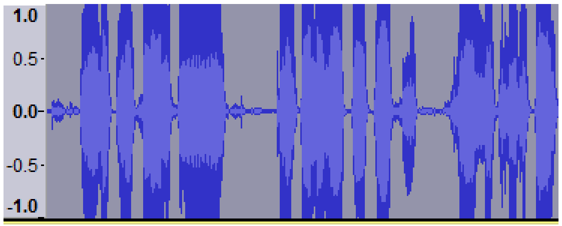


b)Throat Microphone TIMIT Sentence-2

SAMPLE-III



a)Normal Microphone TIMIT Sentence-3



b)Throat Microphone TIMIT Sentence-3

Based on the samples depicted the results are shown in the form of bar charts.

VII. RESULT AND ANALYSIS

The speaker recognition system was built using Alize-Liral toolkit. This system uses 103 speech samples including the throat microphone speech collected from various male and female speaker. The universal back ground(UBM) model was modeled using all the speech samples. To train the system 25 speaker's speech samples are used to modeled as a seperate GMM speaker model by giving speaker IDs like spk0,spk1,spk3..etc.The system is tested by selecting a speech sample against the speaker in the training set.The likelihood scores are calculated from the test sample against the 25 speaker.The maximum likelihood score is used to identify the matched speaker ID.The following Fig.7.shows the likelihood scores against a test sample.

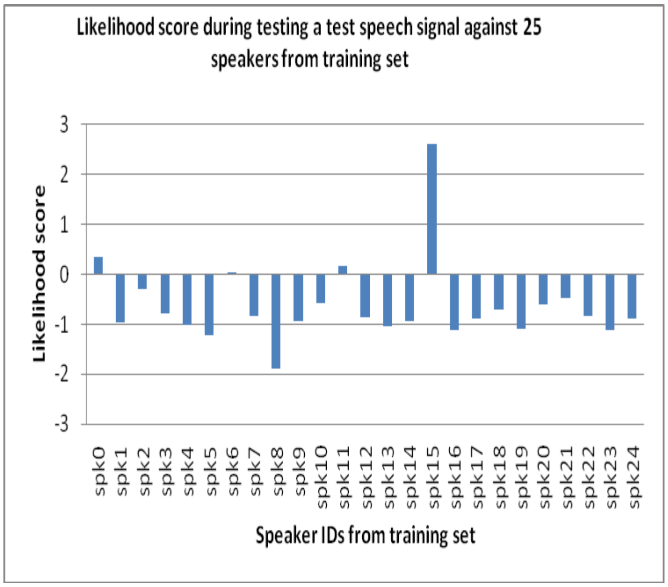


Fig.7.Test Sample comparisons.

In the Fig.7.,the x-axis denotes the speaker IDs of the 25 speakers from the training set. The y-axis denotes the likelihood score calculated against the test sample.

From the above figure it is noted that the speaker Id 'spk15' has the maximum likelihood score among the other speakers in the training set. Thus, It concludes that the test speech sample belong to the speaker of ID spk15 in the training set.

VIII.CONCLUSION

In this paper, the air conducted speech signals are recorded from different speakers using a standard microphone and also with the another mode of speech recording using a throat microphone.The recorded speech signals are preprocessed in which the speechs that are recorded from air microphone has much background noise while the throat microphone is free from background noise and then both speeches are made suitable for the feature extraction process. From these preprocessed speech signals, mfcc features were successfully extracted for generating the GMM speaker model and also with the help of GMM the speakers are trained and tested by depicting the Likelihood scores.

REFERENCES

[1] Jia-Ching Wang,Yu-Hao Chin,Wen-Chi Hsieh,Chang-Hong Lin,Ying-Ren Chen,Siahaan.E, "Speaker Identification With Whispered Speech for the Access Control System", Automation Science and Engineering, IEEE Transactions , vol 12,no 4, pp. 1191-1199, 2015.

[2] Kawthar Yasmine Zergat and Abderrahmane Amrouche, "New Scheme based on GMM-PCA-SVM Modeling for Automatic Speaker Recognition", International Journal of Speech Technology, vol 17, no 4, pp. 373-381, 2014.

[3] Maxim Sidorov, Alexander Schmitt, Sergey Zablotskiy and Wolfgang Minker, "Survey of Automatic Speaker Identification Methods", Proceedings of the Ninth International Conference on Intelligent Environments, pp. 236-239, 2013.

[4] Cherifa S. and Messaoud R,"New Technique to use the GMM in Speaker Recognition System (SRS)", International Conference on Computer Applications Technology, pp. 1-5, 2013.

[5] Seiichi Nakagawa, Longbiao Wang and Shinji Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information" IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 4, 2012.

[6] Rishiraj Mukherjee, Tanmoy Islam and Ravi Sankar, "Text Dependent Speaker Recognition using Shifted MFCC", Proceedings of IEEE Southeast Conference, pp. 1-4, 2012.

[7] Homayoon Beigi, "Fundamentals of Speaker Recognition" Springer Publications, pp. 75-84, 2011.

[8] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, no. 1–3, July 2010.

[9] Tomi Kinnunen and Haizhou Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors", Elsevier Journal of Speech Communication, vol. 52, no 1, pp. 12-40, 2010.

[10] Marcos Faundez-Zanuy and Enric Monte-Moreno, "State-of-the-Art in Speaker Recognition" IEEE Aerospace and Electronic Systems Magazine, vol. 20, no 5, pp. 7-12, May 2005.

[11] Joseph P. Campbell Jr., "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol. 85, no 9, pp. 1437-1462, September 1997.

[12] Robin King, "New Challenges in Automatic Speech Recognition and Speech Understanding", IEEE TENCON, Conference on Speech and Image Technologies for Computing and Telecommunication, pp. 287-294, 1997.

[13] Weng, Z., Li, L., & Guo, D, "Speaker recognition using weighted dynamic MFCC based on GMM", Proceedings - 2010 International Conference on Anti-Counterfeiting, Security and Identification, pp.285–288, 2010.

[14] M. Arun Marx, G.Vinoth, A. Shahina, A. Nayeemulla Khan, "Throat Microphone Speech Corpus for Speaker Recognition",MES Journal of Technology and Management,pp.16-20,2008.