Proceedings of 4th RIT Post Graduates Conference (RIT PG Con-18)
NOVATEUR PUBLICATIONS
JournalNX- A Multidisciplinary Peer Reviewed Journal (ISSN No:2581-4230)
April, 13th, 2018

A SPEAKER RECOGNITION SYSTEM USING GAUSSIAN MIXTURE MODEL

Mr. Ajinkya N. Jadhav
M. Tech Student
Dept. of Computer science and Engineering.
Rajarambapu Institute of Technology, Sakhrale, India
ajinkyajadhav.96@gmail.com

Dr. N. V. Dharwadkar
Head of Department
Dept. of Computer science and Engineering.
Rajarambapu Institute of Technology, Sakhrale, India
nagaraj.dharwadkar@ritindia.edu

Abstract—**Automatic speaker recognition system identifies a person from the information contained in the speech signal. These systems are the most user-friendly means of biometric recognition and are being used in applications like teleconferencing, banking, forensics etc. The accuracy of these depends on the methods used to extract features from the speech signal, modeling methods, classifiers used to identify the speaker and the size of data available for modeling and verifying. Here, the Mel-Scale Frequency Cepstral Coefficient is one of the methods to grab features from wave file of spoken sentences. The Gaussian Mixture Model is a technique applied in the MARF (Modular Audio Recognition Framework) framework to increase outcome estimation. We have presented a Gaussian selection medium for MFCC.**

Keywords-- Speaker Identification, MFCC, GMM.

I. INTRODUCTION

In our daily life, body language, text language, image language and speech are the many forms of communication. However, these forms of speech are always regarded as the strongest forms because of their rich dimensional characteristics. The speaker's gender, attitude, mood, health status and identity also refer to a rich dimension apart from spoken and written language. This information is of the utmost importance to effective communication in today's life.

The development of speech is a study of word signals and methods for developing different speech signals. Voice action can be considered as a particular case of digital signal sort out, because the signal is typically digitized and applied to speech signals. The word processing aspects include acquiring, processing, storing, transmitting, and outputting speech signals. The signal input is called speech recognition, the speech is called synthesis. Word signals are mainly divided into three parts: speech recognition, speech recognition, and speaker recognition [2].

A. Human Speech Production System

Human's uses spoken the language to communicate information which is the most natural. The speech signal transports not only what is being said but also realize personal unique attributes of the speaker. Speaker's specific characters are derived from two components, which are the physiological and behavioral attributes of the speaker.

Understanding the behavior of speech construction will help to identify more successful techniques to isolate the characteristics of the speakers. Figure 1.1 illustrates the representation of members of human's voice creation. From a physiological perspective, speech is produced by an excitation production process. This process works when human is excited to speak, lungs consume the air in it and transfer that air to the vocal tract.

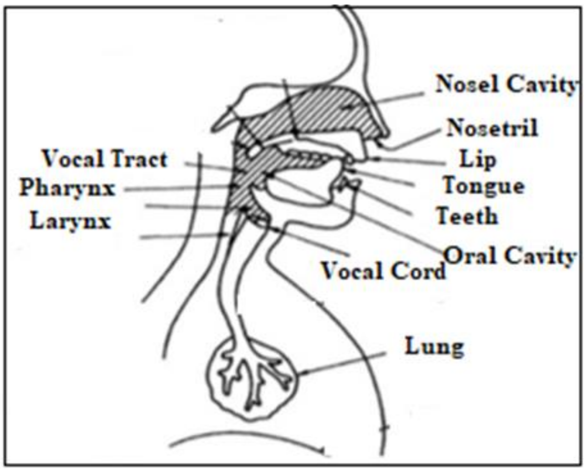


Fig. 1.1.Human Speech Production System

The air current, which is related to the excitation source of the improvised and undirected speech respectively, causes the sound path to resonate, resulting in resonance in its characteristic frequencies (difficult frequencies). The vocal canal begins when the vocal folds open and finish at the tip of lips. It contains three main features, which are the pharynx, cavities like oral or nasal shown in above figure. The frequency of the formula is determined in the form of the acoustic channel, depends on four organs tongue, lips, jaw, and throat. By this characteristics, we control the voice and produce the speech.

B. Speaker Recognition System

Speaker recognition is used for recognizing or identifying from persons individual sound by machine. The system may depend on the text (trained and tested for a particular word or phrase) or independently of the text (without limiting the content). The speaker identification or verification are categorized by depending on final task or decision of machine. Easy-to-access, natural and microphones (inexpensive devices) are used for collecting data is to perform the process of specific applications which are referenced to speaker recognition. The claimed speaker tracking is to locate a given talker's segment in an audio clip or in an automatic teleconference segmenting by potential applications for speaker identification in multi-user systems. In addition, it found it useful in helping the court to discuss and court-applied transcription [7]. In speaker recognition technology, feature extraction is mainly used. Extracting features is a process of holding useful statistics of data from a speech signal while eliminating unwanted signals such as noise. Here, Feature extraction is the conversion of the original acoustic signal into a tightly packed representation of the signal. The series of eigenvectors representing a close-packed speech signal is determined by a feature extraction method. The feature vectors extracted from the original signal in the feature extraction module prominence speaker-specific attributes and vanquish statistical redundancy [9].

This system will perform operations in three phases which are a pre-processing phase, training phase, and decision phase.

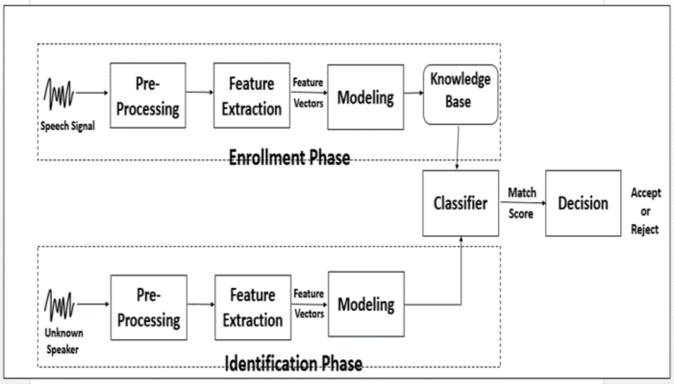


Fig. 1.2. Block-diagram of Speaker Recognition system.

The outputs of the pre-processing phase will be the speech attribute were extracted from the given speech wave. This process is called extraction feature. These withdraw features will be used for training phase to train the system. Throughout the decision-making stage, an unknown speech will be compared or tested with the speech given in the system. A particular speaker will be identified by matching the percentage of the speech with the unknown threshold for speech training. If the match rate is greater than the threshold, only the person's ID will be displayed, otherwise the message "People not available" will be displayed.

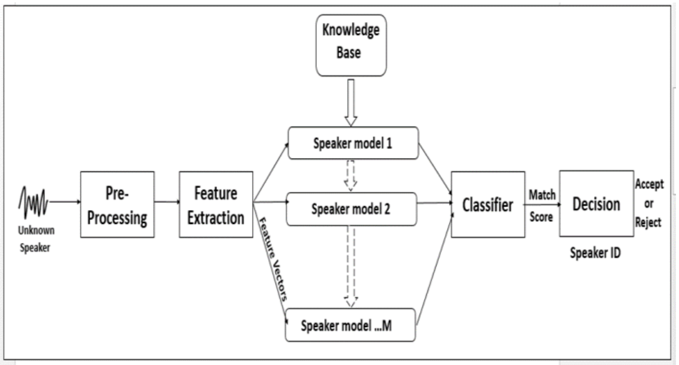


Fig. 1.3. Speaker Identification system.

The checking, verifying and identifying are key applications for identifying speakers. The verification or authentication name is the voice used to verify the certain identity by declared speaker's voice print or vocal tract. On the further work, the task of identifying the speaker is unknown, known as identicalness. In a sense, the speaker validation is a 1: 1 equivalent where the sound of a speaker is equal to one speaker model also called "audio print" or "audio form", while the definition of the speaker is an M: 1 match where Compare the sound against the speaker model M [7].

II. RELATED WORK

A study of existing theories and practices (literature) in the chosen area or domain helps to know about it more deeply. It also helps in identification of gaps or deficiencies in knowledge and in scoping the study by identification of limitations and assumptions. All this helps in framing the problem statement.

The authors Tomi Kinnunen and Haizhou Li [1] implemented an independent speaker remembrance technology, with an important on text-independent placement. He has done work on speaker recognition actively for nearly ten years. The author provides important aspects of a survey of classical as well as an art in various sate methods. The beginning stages are the basics of independent speaker remembrance, with regard to speaker modeling technique and feature extraction method. The advanced computational ability of the technique to handle durability and cycle variability. The recent progression of vectors so as to approach super contains new explosion of feature and reprints the trend of methods. It also provides detail information of current developments and discusses the methodology for assessing speaker recognition systems.

Qing, *et al.*[2] they introduce a system to ameliorate the successfulness of feature parameters, a weighted feature extraction method. The Ear recollection is a type of biometrics methodology, which is most favored and mostlyused. Regulates the average contribution sequence and analyzes each component of the LPCC. Construct on the series, LPCC weighs by every proportion to produce a pressure on feature variables. Matlab implemented the environment of the LPCC properties of the speaker remembrance system which is under. The experimental results of the speaker recognition system show the bestpresentation than the existing models.

In paper [4], JunzoWatada and Hanayuki proposed a Hidden Markov Model approach as an emotion classifier to carry out testing phases using speech data. Audio is a useful and versatile form of communication, where each sound has different frequency characteristics and levels. Audio serves two basic functions for people around the world: signals and contacts. Many problems were found in determining sounds, such as pitch, speed, and accuracy of processing audio data. The research was motivated to identify and analyze human voice in a multi-speaker environment of meeting or indirect conversation.

Asma, Mansour and ZiedLachiri [6] they highlight systematic view to identifying the amplifiers under several emotional conditions based on the multi-vector support seed machine (SVM) workbook. Strengthening the performance of the process of recognizing the emotional speaker has received increasing attention in recent years. The author compared two methods to extract features, to obtain the best accuracy features are used to present a psychological speech in sequence. They used two methods first method is the MFCC and another method is SDCboth are merged withMFCC (SDC-MFCC). This two method were processed by mean and variance attribute. Experiments are performed on the EMOCAP database using two multi-layered SFM approaches one against all (OAA) andone against one (OAO). The outputs obtained shows that SDC-MFCC is superior to ordinary screwdriver performance.

In paper [9], N. Singh, *et al.* author discussed on three main areas of speech technologies which are authentication, surveillance and forensic speaker remembrance. The goal of this research is to introduce the same specific areas where the speech recognition system is used. We also get the information about all application which is related to speech recognition.

The authors S. Paulose and A. Thomas [10] introduces an automatic speaker recognition system identifies a person from the information contained in the speech signal. These systems are the most user-friendly means of biometric recognition and are being used in applications like teleconferencing, banking, forensics etc. The accuracy of these depends on the methods used to extract features from the speech signal, modeling methods, classifiers used to identify the speaker and range of dataset available for training as well as testing. In this paper, recognition systems are implemented using both spectro-temporal features and voice-source features. Classification is done with two different classifiers for an i-vector method and the accuracy rates are compared.

III. DEFINITION AND MODEL

The methods and models used in speaker recognition system are discussed below.

A. Feature-Extraction Technique

The peculiarity of the statement includes the reduction of the number of resources or weights required for a large number of data descriptions. Most seriousproblem is the count of features that make a complex data analysis. Examination with a wide number of variables mostly neededlarge memory and computational power as it can lead to a classification algorithm that is more appropriate for sophisticated and circulating samples for the current sample values. The peculiarity of removing is that the general term of methods for focusing on variables to obtain this problem is still quite accurate.

1) Mel-Frequency Cepstral Coefficients (MFCC):

The activity of sound contains the mel-frequency cepstrum (MFC) assigns the small-range energy spectrum of the voice, related to the sequential cosine conversion of the logarithm capability to the nonlinear frequency scale of the frequency.

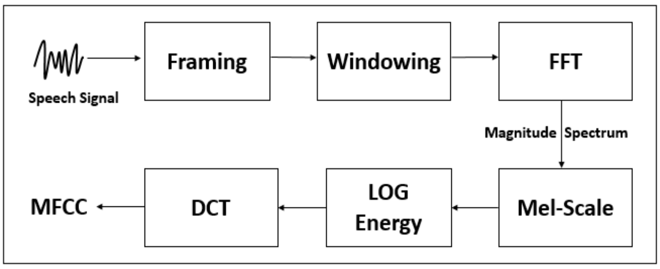


Fig. 3.1. Block diagram of Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency cepstral coefficients (MFCC), which together form MFC. File filter bank energy DCT is the final step to calculating. The filtered banking energies are quite interconnected with one another, as our filter banks are totally the same. The meaning of the covariance matrices of diagonal values is that the DCT Decor refers to the energies that can be used for model models.

MFCC is regularly used as a parameter in soundrecollection systems. MFCC is used to recognize spoken word by telephone automatically. MFCCs are also used in music lyrics as classification and measure the audio similarities.

*a) Framing:*The speech signal is divided into several blocks at the particular duration of 20-30 ms which are frames. The signal is divided into (n) number of samples and frames are separated by (m), where (m) is less than (n). In 20-30 ms the voice of human is constant, so we make the frame up to 30 ms.

b) Hamming Windowing: Each and every frame in preprocessing phase is multiplied with the hamming window sequentially to maintain signal continuously.For eliminating the discontinuity the window function is applied.The spectral distortion is reduced by using the window to make zero at the beginning and end of each frame.

$$Y(m) = X(m) * W(m)$$

W (m) is the window function.

*c) Fast Fourier Transform:*FFT is a process of converting the signal from time domain to the frequency domain. To obtain the magnitude frequency response of each frame we perform FFT. The result we get in a spectrum by applying Fourier transform.

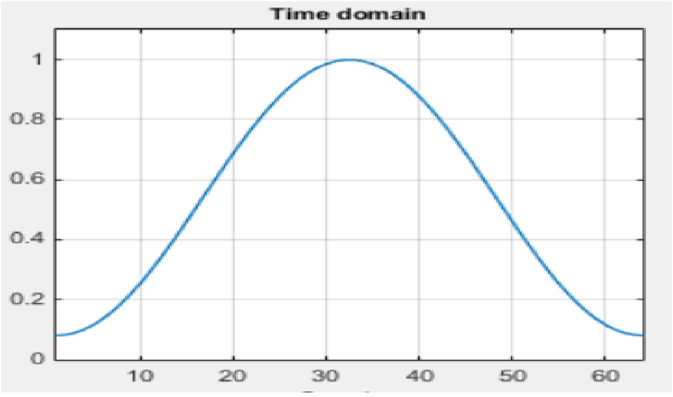


Fig. 3.2.Original Speech signal in the time domain

The above graph represents the original signal. This signal is in time domain format.

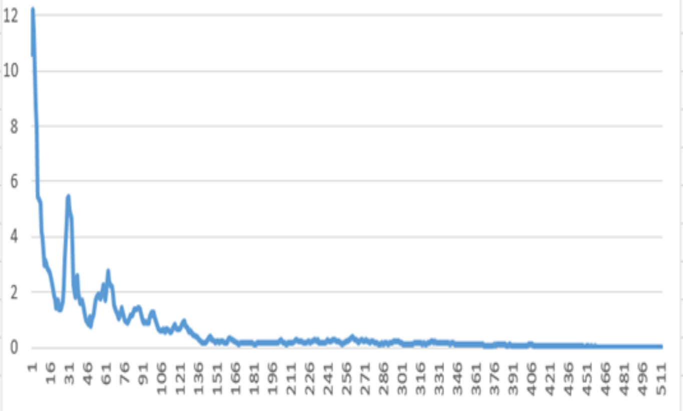


Fig. 3.3.Speech signal in the frequency domain

*d) Filters (Mel-Scale):*The filters are used to eliminate unwanted noise or speech. The triangularly shaped filter is mostly used in preprocessing phase. Fourier transform is used to implement filter bank by transforming window of speech.

e) Discrete Cosine Transform: This transform technique translates a specific sequence of data points, which is alternate of different frequencies in terms of the sum of the cosine functions.The complexity of DCT is also O(nlogn) [4].

$$X_f = 1\sqrt{n}\sum_{i=o}^{n-1} x_i \cos\frac{\pi f(i+0.5)}{n} \quad , F = 0, 1, 2 \ldots, n\text{-}1.$$

The equation used to convert linear scale frequency f to Mel scale frequency is given in equation (1) is as below [10],

$$Mel(f) = 2595 log_{10}(1 + \frac{f}{700})$$

Thus, the following figure represents mel-frequency cepstral coefficient,

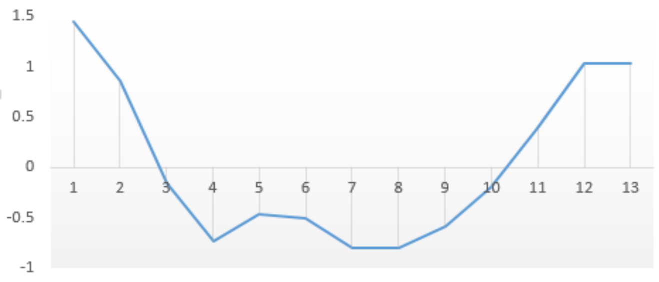


Fig. 3.4. Mel-Scale Cepstral Coefficients

The above graph shows the mel-scale features vectors converted by linear scale frequency.

B. *Classification Technique*

The classification technique is an important activity in speaker recognition system. The formulas and description of it are discussed as below.

1) Gaussian mixture model:

The GMM can be seen as an extension of the VQ model, where groups overlap. That is, the feature vector is not set to the nearest cluster, but has a non-zero probability of the origin of each cluster. GMM is made up of a limited mixture of Gaussian multivariate components. GMM, with k, is characterized by its probability density function.

This model is a heavy sum of Gaussian distributions capable of determining a random separation of supervision.The equation of likelihood method of a GMM for an examination of x is given as below [10],

$$P(x|\boldsymbol{\lambda}) = \sum_{n=1}^{M} w_n p_n(x),$$

The n^{th} Gaussian density$p_n(x)$ contains the w_n as weight.

$$p_n(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_n|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_n)\sum_n^{-1}(x-\mu_n)\}$$

In above equation, $\Sigma_n$and $\mu_n$are the matrices of covarianceand the mean vector of the n^{th} Gaussian, sequentially. And in first equation,$\sum_n^M w_n$= 1 and w_n> 0.

2) Maximum Likelihood Estimation:

MLE's are the elements of the parameters which increases the probability of the observed items. Parameter estimation for GMM using maximum likelihood, $\boldsymbol{\lambda}$ denotes an initial model for ML. The mean and variance are estimated from the known data to maximize likelihood function.

$$\text{F}\ (x_1, x_{2,}x_3, \ldots x_n|\boldsymbol{\lambda}) =$$
$$f(x_1|\boldsymbol{\lambda}) * f(x_2|\boldsymbol{\lambda}) * f(x_3|\boldsymbol{\lambda}) * \ldots f(x_n|\boldsymbol{\lambda})$$

The above equation is joint density function. By this equation we can derive the likelihood function.

$$\therefore \text{L}\ (\boldsymbol{\lambda}|(x)_1, (x)_{2,}(x)_3, \ldots (x)_n) =$$
$$\text{F}\ ((x)_1, (x)_{2,}(x)_3, \ldots (x)_n|\boldsymbol{\lambda}),$$

$$\therefore \text{L}\ (\boldsymbol{\lambda}|(x)_1, (x)_{2,}(x)_3, \ldots (x)_n) = \pi_{i=1}^{n} f(x_i|\boldsymbol{\lambda})$$

Here in above equation we get the likelihood function,

$$Ln\ \text{L}\ (\boldsymbol{\lambda}|x_1, x_{2,}x_3, \ldots x_n) = \sum_{i=1}^{n} Ln\ f(x_i|\boldsymbol{\lambda})$$

It is often more convenient when working with the natural logarithm of likelihood function. GMM are able to build self-clustering boundaries. The mixture model is a probabilistic data which is belongs to the distribution of the mixture model. The density function in the distribution of the mixture, is a convex combination of other probability distribution functions.

$$\therefore P(x) = W_1P_1(x) + W_2P_2(x) + W_3P_3(x) + \cdots W_nP_n(x)$$

The $P_i(x)$ individual density tasks are collected to make density in a mixture density $P(x)$ are called the mixture components and weights $(W)_1, (W)_2, (W)_3, \ldots (W)_n$ linked with each elements are called the mixture coefficients in GMM.

$$P(x) = W_1N(x|\mu_1\Sigma_1) + W_2N(x|\mu_2\Sigma_2) + W_3N(x|\mu_3\Sigma_3) + \cdots W_nN(x|\mu_n\Sigma_n)$$

Each and every elements component of the mixture is a Gaussian distribution with its own parameters and its corresponding variance variables.

3) Expectation Maximization:

This algorithm can be used to estimate the underlying variables, such as those that come from the distribution of the mixture. The EM algorithm is a method to get the maximum probability, evaluates for structure elements when our data is not complete or unexpected. This method is repeated again and again to find the maximum potential task.

The algorithm of expectation maximization as below,

a) First, initialize the $\boldsymbol{\lambda}$parameters for some random values.

b) For each possible value of Z, compute the probability by given$\boldsymbol{\lambda}$.

c) Then, use calculated Z values only to calculate a better estimate of λ parameters.

d) Repeat steps 2 and 3 until convergence.

The process continues until the algorithm is covered on a constant point by creating a better guess using new values.K-mean algorithm is used for clustering or training the data model.

TABLE I: Symbols and their description

Symbols	Description
$P(\boldsymbol{\lambda})$	The likelihood of GMM model λ for an observation$x$.
μ_m	Mean vector.
$\sum_m$	Covariance matrix.
$p_m(x)$	Gaussian density.
Z	The probability of each possible value.
Ln L (x)	Natural logarithm of the likelihood function.
$P(x)$	Mixture components.

IV EXPERIMENTAL SETUP

A. *Training and Testing Data*

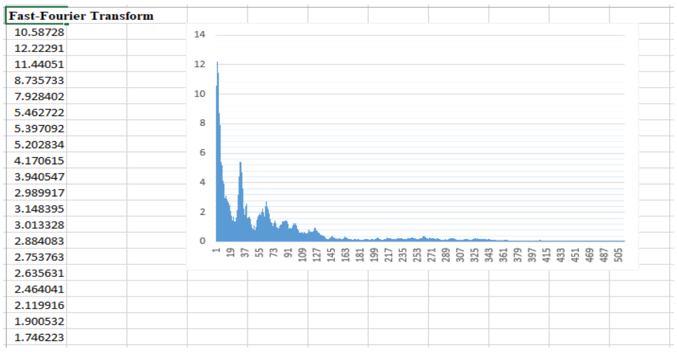
The system performs training on 30 speakers, which contains minimum three to the maximum eight speech samples in wave format of each speaker. The speech is divided into frames on 20ms durations. For smoothing this frames we use Hamming windowing operation. Here, we get 17 feature vectors by applying same filters using MFCC feature extraction technique. Another property like log-energy and magnitude also extracted. For testing, one wave file is passed to each speaker.

B. *Model Training*

The model training is done by using the K-mean algorithm in Gaussian Mixture Model. On the bases of center point, the K-mean algorithm does modeling or clustering. Here, 10 centers are used to cluster or model the dataset.

IIV RESULT AND DISCUSSION

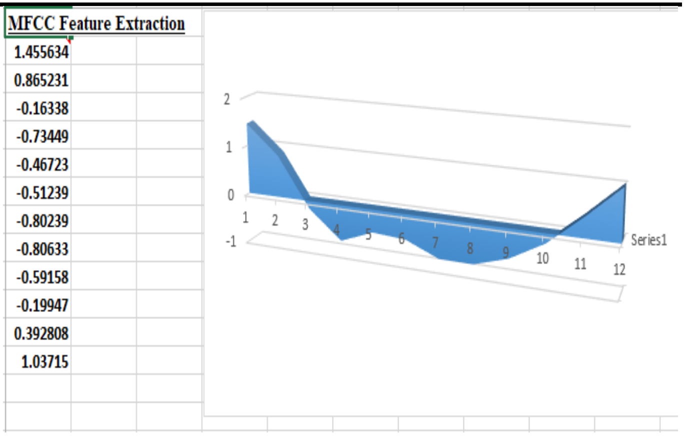
This system mainly works to extractcontinuous signal to get MFCC features and trains that by GMM classifier. The FFT is used to convert the time domain signal to the frequency domain. The frame is divided into 30 ms, on that duration we get stationery values.



Fast-Fourier Transform
10.58728
12.22291
11.44051
8.735733
7.928402
5.462722
5.397092
5.202834
4.170615
3.940547
2.989917
3.148395
3.013328
2.884083
2.753763
2.635631
2.464041
2.119916
1.900532
1.746223

Fig. 5.1. Discrete values of speech

The above figure 5.1 shows the graph of discrete value extracted by a continuous speech by using Fast Fourier Transform. Here we get the 512 discrete values for modeling the classifier.



MFCC Feature Extraction
1.455634
0.865231
-0.16338
-0.73449
-0.46723
-0.51239
-0.80239
-0.80633
-0.59158
-0.19947
0.392808
1.03715

Fig. 5.2. Mel-Frequency Cepstral Coefficient

The above graph represents the mel-cepstrum coefficients, useful for modeling or training dataset. The 12 MFCC coefficients are extracted from the speech sample. On this coefficients the mean and covariance it calculated by GMM for training or modeling the system.

IIIV CONCLUSION

The proposed system is a system to identify claimed speakers from several speech samples. Speaker recognition process is delicate to sound because it can strike the voice signal feature extraction activity. In this paper, we have described the MFCC and GMM system which is used to get correct speaker recognition results. This speaker recognition system is built for 30 speakers in which, the optimal likelihood correlation is tested on different speech samples for detection by using likelihood functions of Gaussian mixture models which simple but productive. Further work is to test the system with more number of speakers and decrease the computational work by applying different preprocessing methods. And maintain the high-accuracy of classifying the speakers.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010.

[2] L. Zhu and Q. Yang, "Speaker Recognition System Based on weighted feature parameter", *Physics Procedia*, vol. 25, pp. 1515-1522, 2012.

[3] F. Bie, D. Wang, J. Wang and T. Zheng, "Detection and reconstruction of clipped speech for speaker recognition", *Speech Communication*, vol. 72, pp. 218-231, 2015.

[4] H. Veisi and H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain", *Speech Communication*, vol. 55, no. 2, pp. 205-220, 2013.

[5] S. Ranjan and J. Hansen, "Curriculum Learning Based Approaches for Noise Robust Speaker Recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 197-210, 2017.

[6] A. Mansour and Z. Lachiri, "SVM based Emotional Speaker Recognition using MFCC-SDC Features", *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 4, 2017.

[7] P. Pal Singh, "An Approach to Extract Feature using MFCC", *IOSR Journal of Engineering*, vol. 4, no. 8, pp. 21-25, 2014.

[8] S. Chougule and M. Chavan, "Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition", *Procedia Computer Science*, vol. 58, pp. 272-279, 2015.

[9] N. Singh, R. Khan and R. Shree, "Applications of Speaker Recognition", *Procedia Engineering*, vol. 38, pp. 3122-3126, 2012.

[10] S. Paulose, D. Mathew and A. Thomas, "Performance Evaluation of Different Modeling Methods and Classifiers with MFCC and IHC

Features for Speaker Recognition", *Procedia Computer Science*, vol. 115, pp. 55-62, 2017.

[11] N. Dehak, P. Dumouchel and P. Kenny, "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2095-2103, 2007.

[12] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448-1460, 2007.

[13] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition", *Speech Communication*, vol. 31, no. 2-3, pp. 193-203, 2000.

[14] M. Alsulaiman, A. Mahmood and G. Muhammad, "Speaker recognition based on Arabic phonemes", *Speech Communication*, vol. 86, pp. 42-51, 2017.

[15] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, vol. 17, no. 1-2, pp. 91-108, 1995.