# Speaker Classification and Recognition using Spectral Features and Gaussian Mixture Models

**Abhinav Sharma[1], Anshu Sharma[2]**
[1] Department of Electronics and Communication Engineering
[2]Department of Physics
[1]Graphic Era University, Dehradun, Uttarakhand
[2]Olympus High School, Dehradun Uttarakhand

***Abstract-****In this work Speaker Recognition has been performed by the extraction of Mel Frequency Cepstral Coefficients (MFCCs). Gaussian Mixture Models technique has been used for the purpose of Classification. An accurate Speaker recognition can play a vital role in the important areas of the society including security identifications, crime investigation etc. The accuracy tests have been performed by varying the number of Mfccs extracted. Extracting more number of features delays the result calculation but may improve the chances of increased accuracy. An optimum combination of number of features and accuracy is achieved. Results show that accuracy is highest (94.28 %) at 21 numbers of MFCCs.*

***Keywords***-Speaker/Speech Recognition, MFCC, GMM, Speech classification

## I. INTRODUCTION

In this technological era, information technology continues making more impact on many aspects of our daily lives. However, the problem of communication between human beings and information processing machines become increasingly important. So far, such communication has been done almost entirely by means of keyboards and screens, but there are substantial disadvantages of this method for many applications. Speech is considered as the most widely used and natural means of communication between humans, and it is an obvious substitute for such means of keyboards and screens for communication process. This simple means of exchanging the information is, in fact, extremely complicated. Although the application of speech in the man-machine interface is growing rapidly, in their present forms machine capabilities for generating and interpreting speech are still incomplete and imperfect.

### I.1-AUTOMATIC SPEECH RECOGNITION

In the field of computer science, speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Some SR systems use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "Speaker Independent" systems. Systems that use training are called "Speaker Dependent" systems.

The field of Automatic Speech Recognition (ASR) is about 60 years old. There have been many interesting advances and developments since the invention of the first speech recognizer at Bell Labs in the early 1950’s. The development of ASR increased gradually until the invention of Hidden Markov Models (HMM) in early 1970’s. Researchers’ contribution were to make use of ASR technology to what can be seen nowadays of various advancements in fields like multi-modal, multi-lingual/cross-lingual ASR using statistical techniques such as HMM, SVM, neural network, etc.

Speech recognition or more commonly known as automatic speech recognition (ASR) was defined as the process of interpreting human speech in a computer. However, ASR was defined more technically as the building of system for mapping acoustic signals to a string of words. In general, all ASR systems aim to automatically extract the string of spoken words from input speech signals.

Despite a lot of advancement in speech recognition technology over many years, the human voice still remains largely unexploited. Voice input has a number of potential benefits, especially for physically disabled people, one of the major limitations of current speech-based interaction methods is their inability to provide fluid and continuous input.

## II. FEATURE EXTRACTION

The Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear’s critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels . The following formula is used to compute the Mels for a particular frequency:

$$mel(f) = 2595 * \log_{10}(1 + f / 700)$$

In speaker independent speech recognition, a premium is placed on extracting features that are somewhat invariant to changes in the speaker. So feature extraction involves analysis of speech signal. Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

A conventional automatic speech recognition (ASR) system can be in two blocks: the feature extraction and the modeling stage. In practice, the modeling stage is subdivided in acoustical and language modeling, both based on HMMs as described in Figure below-

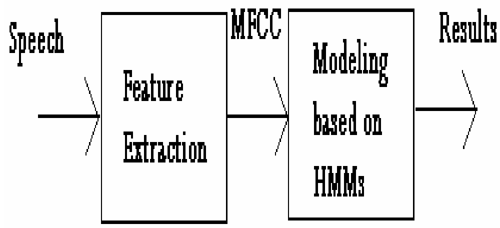

Fig. 1- Simple representation of a conventional ASR

The feature extraction is usually a non-invertible (lossy) transformation, as the MFCC described pictorially in Figure. Making an analogy with filter banks, such transformation does not lead to perfect reconstruction, i.e., given only the features it is not possible to reconstruct the original speech used to generate those features. Computational complexity and robustness are two primary reasons to allow loosing information. Increasing the accuracy of the parametric representation by increasing the number of parameters leads to an increase of complexity and eventually does not lead to a better result due to robustness issues. The better result due to robustness issues. The greater the number of parameters in a model, the greater should be the training sequence.

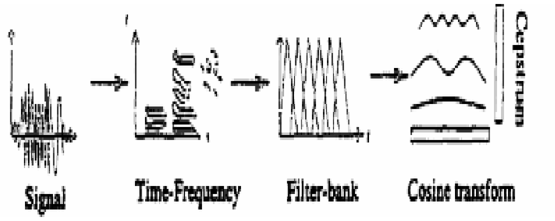

Fig. 2 Pictorial representation of MFCC calculations

Speech is usually segmented in frames of 20 to 30 ms, and the window analysis is shifted by 10 ms. Each frame is converted to 12 MFCCs plus a normalized energy parameter. The first and second derivatives (D’s and DD’s) of MFCCs and energy are estimated, resulting in 39 numbers representing each frame. Assuming a sample rate of 8 kHz, for each 10 ms the feature extraction module delivers 39 numbers to the modeling stage. This operation with overlap among frames is equivalent to taking 80 speech samples without overlap and representing them by 39 numbers. In fact, assuming each speech sample is represented by one byte and each feature is represented by four bytes (float number), one can see that the parametric representation increases the number of bytes to represent 80 bytes of speech (to 136 bytes). If a sample rate of 16 kHz is assumed, the 39 parameters would represent 160 samples. For higher sample rates, it is intuitive that 39 parameters do not allow to reconstruct the speech samples back. Anyway, one should notice that the goal here is not speech compression but using features suitable for speech recognition.

## III. FEATURE CLASSIFICATION AND MATCHING

### III.1-Fundamentals of Gaussian Mixture Model

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data-set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population-identity information.

Some ways of implementing mixture models involve steps that attribute postulated sub-population-identities to individual observations (or weights towards such sub-populations), in which case these can be regarded as types of unsupervised learning or clustering procedures. However not all inference procedures involve such steps.

Mixture models should not be confused with models for compositional data, i.e., data whose components are constrained to sum to a constant value (1, 100%, etc.).

### III.2-General Mixture Model

A typical finite-dimensional mixture model is a hierarchical model consisting of the following components:

1-N random variables corresponding to observations, each assumed to be distributed according to a mixture of K components, with each component belonging to the same parametric family of distributions (e.g., all Normal, all Zipfian, etc.) but with different parameters

2-N corresponding random latent variables specifying the identity of the mixture component of each observation, each distributed according to a K-dimensional categorical distribution

3-A set of K mixture weights, each of which is a probability (a real number between 0 and 1 inclusive), all of which sum to 1

4-A set of K parameters, each specifying the parameter of the corresponding mixture component. In many cases, each "parameter" is actually a set of parameters. For example, observations distributed according to a mixture of one-dimensional Gaussian distributions will have a mean and variance for each component. Observations distributed according to a mixture of V-dimensional categorical distributions (e.g., when each observation is a word from a vocabulary of size V) will have a vector of V probabilities, collectively summing to 1.

In this work speech samples have been collected from 7 speakers and after spectral feature extraction, speaker GMMs have been created which were used for the classification. The testing files have been tested against all the speaker GMMs and recognition have been found out.

## IV. TESTING, EXPERIMENTAL RESULTS AND DISCUSSIONS:

Following operations were performed as methodology of research work:

1-Collection of speaker speech data in different environments.

2-Adding the different prerecorded additive noises to the clean speech data samples.

3-Extraction of distinguished and discriminative features from the collected speech samples using Mel Frequency Cepstral Coefficient (MFCC) technique and producing sets of feature vectors for all 7 speakers.

4-Creating and training the GMM models for all 7 speakers.

5-Testing the performance of GMM models.

### IV.1-Speaker Recognition Performance-

For the testing of random speaker speech samples with the trained models for seven speaker speeches, a folder structure is formed of seven folders for seven speakers, each containing 10 samples of a speaker ( in all 100 test samples) . A percentage confusion matrix is formed as shown in table-1. Sp. 1 in the table represents the test folder 1 having 10 test samples of speaker no. 1. All its test samples go through the process of feature extraction and the feature vectors are sent to 10 GMM models for seven speaker speeches one by one. Now e.g. value 70 in the location (1,1) represents that 70% out of 10 test samples at test folder 1 (Test 1) are recognized as First speaker.

Table-1 Percentage Confusion Matrix (Sp.=Speaker)

| **Sp.** | **Gaussian Mixture Models** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 1 | 70 | 0 | 20 | 0 | 0 | 0 | 10 |
| 2 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 90 | 0 | 0 | 10 |
| 5 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

According to Table-2 e.g. Now e.g. value 7 in the location (1,1) represents that 7 out of 10 test samples at test folder 1 (Test 1) are recognized as the samples of First speaker.

Table-2 Confusion Matrix (Sp.=Speaker)

| **Sp.** | **Gaussian Mixture Models** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 1 | 7 | 0 | 2 | 0 | 0 | 0 | 1 |
| 2 | 0 | 10 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 10 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 9 | 0 | 0 | 1 |
| 5 | 0 | 0 | 0 | 0 | 10 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 10 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |

Table 3 describes the Robustness or the accuracy of the speaker recognition. Here closed test refers that the training speaker speech data and the testing speech data is same. On the other hand open test refers that training speaker speech files and testing speakjer speech files are different.

Table-3 Recognition Robustness Performance (%)

| Speaker | Closed Test | Open Test |
|---|---|---|
| 1 | 90 | 70 |
| 2 | 100 | 100 |
| 3 | 100 | 100 |
| 4 | 90 | 90 |
| 5 | 100 | 100 |
| 6 | 100 | 100 |
| 7 | 100 | 100 |
| **Average** | **97.14** | **94.28** |

Table 4 gives another description of the testing results. Here numbers of MFCCs extracted from the speech files are varied. Results show that extraction of more number of features may improve the accuracy of the recognition but after a certain set the accuracy may start dropping. In this work the accuracy is achieved best with 21 number of features extracted.

Table-4 Speaker Recognition with varying MFCC's (%)

| Speaker | No. of MFCC's | | | | |
|---|---|---|---|---|---|
| | **6** | **8** | **13** | **21** | **29** |
| 1 | 40 | 40 | 60 | 70 | 60 |
| 2 | 50 | 50 | 70 | 100 | 80 |
| 3 | 60 | 60 | 90 | 100 | 60 |
| 4 | 40 | 50 | 60 | 90 | 50 |
| 5 | 30 | 40 | 80 | 100 | 70 |
| 6 | 40 | 50 | 90 | 100 | 90 |
| 7 | 50 | 70 | 80 | 100 | 90 |
| **Average** | **44.2** | **51.4** | **75.7** | **94.2** | **71.4** |

## V. CONCLUSION AND FUTURE WORK

The research covers the importance of speech recognition technology to fasten the process of security identifications. In addition the research achieves and meets the objectives of developing it, and it is hoped that the research will benefit the end users as it is designated for that purpose.

The speech recognition technology has been widely used in this system. As a result, this study manages to show and emphasize the need and importance of such technology in our daily life. This system could be implemented in various business, organizations, financial institutions and many other academic institutions.Biometric identification with speech is a good tool for protection of computer machine.

In the Phase-I of research the speech recognition for seven speakers was developed. As the future work the recognition of speakers of different nationalities and languages will be achieved.

It is very hard to get the exact matches and high accuracy in many cases, because human voice changes from time to time and most importantly in noisy environments.\

The speaker recognition system for seven speakers can be enhanced and modified if spectral and prosodic features are also combined. This will help in achieving more accurate recognition results.

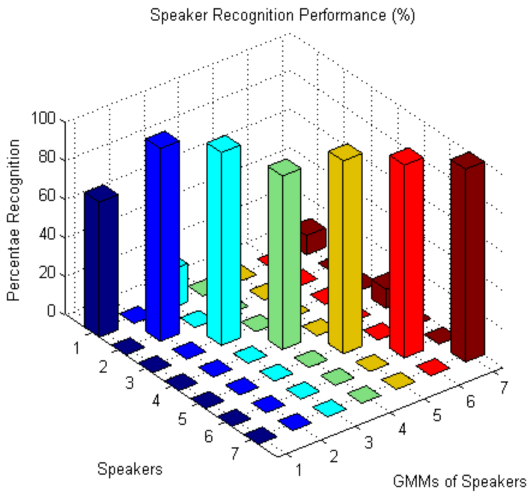

Figure-3 Speaker recognition Performance (%)

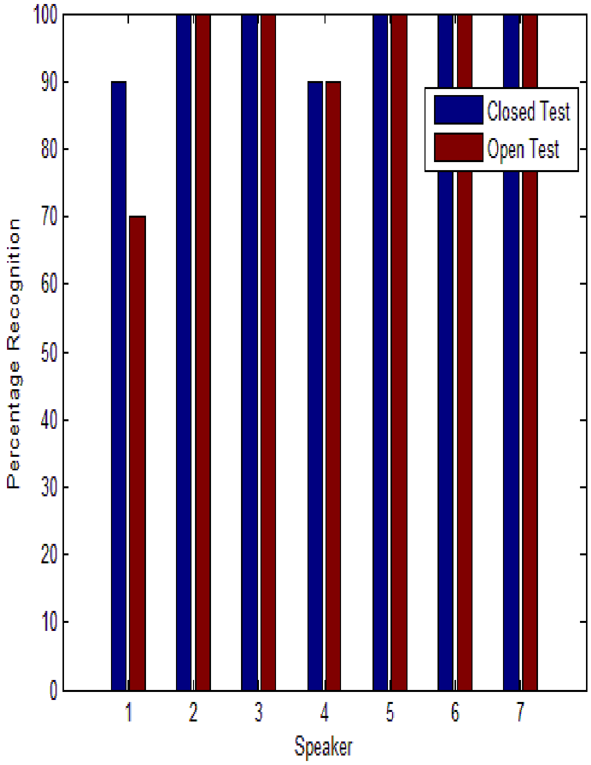

Figure 4-Recognition Robustness Performance (%)

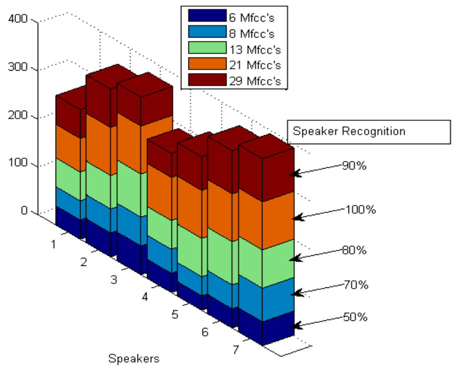

Figure -5 Speaker Recognition with varying no. of MFCC's

## REFERENCES

[1] Shashidhar G. Koolagudi, K.Srinivas Rao, "Recognition of Emotions from Speech using Excitation Source Features", International Journal of Speech Technology, June 2012, Volume 15, Issue 2, pp 265-289

[2] Sandeep Kaur , "Mouse Movement using Speech and Non-Speech Characteristics of Human Voice", International Journal of Engineering and Advanced Technology (IJEAT) ,ISSN: 2249 – 8958, Volume-1, Issue-5, June 2012

[3] Dr.E.Chandra, A.Akila ,"An Overview of Speech Recognition and Speech Synthesis Algorithms", Int.J.Computer Technology & Applications,Vol 3 (4), 1426-1430 ISSN:2229-6093

[4] Ahmad A. M. Abushariah, Reddy Surya Gunawan, " Speech Recognition System using MATLAB ", LAB LAMBERT Aceademic Publishing GmbH & Co. KG, 2011

[5] Ahmad A. M. Abushariah, Teddy S. Gunawan,Othman O. Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia

[6] Susumu Harada, "Harnessing the Capacity of the Human Voice for Fluidly Controlling Computer Interface", University of Washington, 2010

[7] James R Evans, Wayne A Tjoland and Lloyd G Allred, "Achieving a Hands-Free Computer Interface using Voice Recognition and Speech Synthesis ",IEEE AES Systems Magazine, 2000

[8] Susumu Harada and James A Landay, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques", Portland, Oregon, USA: ASSETS, 2006

[9] M Abdeen, H Moshammad and M C E Yagoub, "An Architecture for Multi-Lingual Hands Free Desktop Control System for PC Windows", Niagara Falls, Canada : IEEE , 2008

[10] R Maskeliunas, K Ratkevicius and V Rudzionis, "Voice-based Human-Machine Interaction Modeling for Automated Information Services", ISSN 1392-1215 Electronics and Electrical Engineering, 2011

[11] R Norma Conn and Michael McTear, "Speech Technology: A Solution for People with Disabilities", Savoy Place, London WCPR OBL, UK: IEE, 2000

[12] Minh TU Vo and Alex Waibel "A Multi-Lingual Human-Computer Interface: Combination of Gesture and Speech Recognition", Carnegie Mellon University Pittsburgh, U.S.A, 2009

[13] Susumu Harada, Jacob O Wobbrock, Jonathan Malkin, Jeff A Bilmes James A Landay ,"Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor Control", Boston, MA: Conf. on Human Factors in Computing Systems

[14] M Rahmani, N Yousefian and A Akbari, "Energy-based speech enhancement technique for hands-free communication", ELECTRONICS LETTERS Vol. 45 No. 1, 2009