Voice Recognition Using MFCC And GMM

SHUBHAM GUPTA
SCSE
VELLORE INSTITUTE OF TECHNOLOGY,
CHENNAI, INDIA
shubhamgupta.2018@vitstudent.ac.in

Prof. THOMAS ABRAHAM J V
SCSE
VELLORE INSTITUTE OF TECHNOLOGY.
CHENNAI, INDIA
thomasabraham.jv@vit.ac.in

Abstract— Recognition of speech is the ability to identify spoken words, and recognition of speakers is the ability to identify who is saying those words. Speaker Recognition is one of the main areas of research focused on Speech Signals. Speaking recognition, speech-to-text translation and vice-versa are other important field of study. In this paper, a brief overview of the area of speaker identification, describing its system, various modules for the extraction and modeling of features, applications, underlying techniques and some indications of performance is presented. The Mel Frequency Cepstral Coefficient (MFCC) is regarded as a key factor in the recognition of speakers. Gaussian Mixture Model (GMM) is the most popular model for data training. This considered MFCC as the primary feature with "tuned parameters," and delta- MFCC as a secondary feature and also implemented GMM with some tuned parameters to train our model. This shows that the role of identification of speakers can be performed using MFCC and GMM along with outstanding accuracy in the results of identification / diarization.

Keywords—MFCC,GMM..

I. INTRODUCTION

Speaker Recognition is a person's identification from speech features, voice biometrics which is also called voice recognition. There is a distinction between the recognition of the speaker(the recognition of who speaks) and the recognition of speech recognition of (what is said). These two words are often mixed and, in fact, voice recognition can be used for both. There is a difference between active authentication usually referred to as an authentication and identification of a speaker or a spoken voice, ultimately there is a difference between the speech recognition and who talks and the diarisation of the speaker, recognizing when the same speaker is speaking. Speaker recognition could simplify the work of translating speech into systems trained in voices of particular persons, or it could be used as part of a safety process to authenticate or verify the speaker's identity. Speaker acknowledgement has a history of four decades and uses the acoustically different features of speech. These acoustic patterns reflect anatomy, for instance the size, shape and composition of the throat and mouth. Speaking style for instance voice pitch. The verification of the speaker gained recognition by the speaker and is a behavioral, biometric classification. Where the speaker claims to be certain identity and the voices are used to confirm this claim, it is known as authentication or checking. On the other hand, identification is the task of evaluating and checking the unknown speaker identity as a 1:1 match in a sense. When a voice is matched to a template, a voice printed or a voice model is also matched, whereas the voice identification is 1 and matches when comparing the voice to the templates in the safety perspective.

1. SPEAKER RECOGNITION

A. Speaker Verification

The verification of the speaker shall take the speech of an unknown speaker with his/her claimed identity and determine whether the claimed identity corresponds to the speech. Speaker tests are 1:1 when the voice of one speaker matches one template (also known as a "voice print" or "voice model") or, in other sense of the Pattern matches the claimed speaker model register in the database. If the match meets a certain threshold the claims for identity is confirmed. Using a high threshold, the system achieves high security and prevents the acceptance of impostors, but in the meanwhile it also risks rejecting the genuine person, and vice versa.

B. Speaker Identification

A speaker identification system only takes an unknown speaker's voice, and decides which speaker enrolled best fits the speech. The speaker identification system among the enrolled speakers finds the best matching speaker, and it may be that the unknown speaker is not enrolled. That is why the identification of speakers is accompanied by the speaker verification in many systems. It's one-to-many comparisons (1:N match when the voice is compared to N templates). In the Speaker Identification System, the M speaker models are scored in parallel and most of the speakers are assigned an ID in the database, or none of the above will be reported if and only if the matching score is below a certain threshold and is most likely to be reported in the case of an open speaker, and therefore the decision will be open-set Speaker Identification System.

C. Speaker Text Dependent

A speaker recognition system may be fooled in a text-dependent mode by recording and playback of an enrolled speaker's predefined voice. The system will require the user to utter a randomly prompted text to protect a speaker recognition system against such malicious attack. In most cases, because additional information (text transcription) is given, a text based speaker recognition system outperforms a text Independent speaker recognition system. However, when the true underlying transcription is not given, a text-dependent speaker recognition system can not be used, as in the situation when some are talking openly over the phone. Text-based identification works better for co-operating subjects

D. *Speaker Text Independent*

Text-independent systems are most often used to classify speakers, as they need very little if the speaker cooperates. In this case, the text is different during registration and testing In reality, registration is possible, as in various forensic applications, without the user's knowledge. Since text-independent technologies fail to compare what was said during registration and verification, verification applications often prefer to use speech recognition for evaluating, at the point of authentication, what the user says.

Acoustics and voice analysis techniques are used in text-independent systems

2. Features Extraction

A. *Mel Frequency Cepstral Coefficent*

MFCC are the Mel Frequency Cepstral Coefficients. MFCC takes into account human sensitivity experience at appropriate frequencies by converting the standard scale to Mel Scale, and is thus very suitable for speech recognition tasks (as they are suitable for human perception and the frequency at which humans speak/utter).

Feature extraction and recognition are two essential modules in speech recognition systems. The primary aim of extraction of the function is to identify robust and discriminatory features in the acoustic data. The recognition module decodes the speech input using speech features and acoustic models and generates text results with high precision. The main purpose of the extraction of feature step is to compute a saving sequence of feature vectors that provide a compact representation of the given input signal. First of all, the recording of different speech samples of each word in the vocabulary is done by different speakers. After the speech samples are collected; sampling at a frequency of 20 kHz converts them from analog to digital form. Sampling requires the recording at a regular interval of the speech signals. The data collected are now quantified if it is necessary to eliminate noise in speech samples. Speech samples collected are then passed through the extraction feature, training feature & testing stage. Feature extraction transforms the incoming sound into an internal representation so that the original signal can be reconstructed from it.

MFCC also increasingly finds uses in music information such as gender classification, audio similarity measurement, etc. The basic concept of the MFCC method is shown in the block diagram of Fig. 1

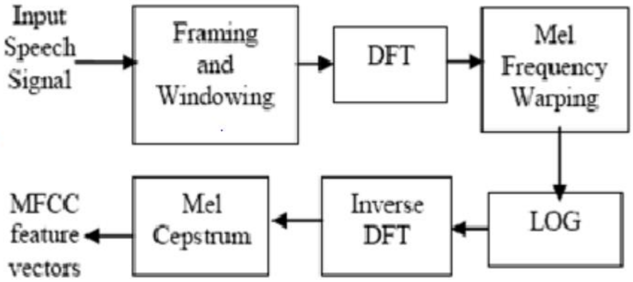


Fig.1: Steps Involved in MFCC Features Extraction

A.1. *Pre-emphasis*

Pre-emphasis is important because voiced parts of the speech signal naturally have a negative spectral slope (attenuation) of around 20db per decade because of the physiological characteristics of speech development. Pre-emphasis raises high frequency energy level. For voiced segments, such as vowels, the lower frequencies have more energy than the higher frequencies. This is called spectral tilt (how vocal folds produce sound) which is related to the glottal source. The boosting of the high frequency energy makes the acoustic model more accessible to information in higher formants. For humans, we're starting to have hearing problems when we can't hear these high-frequency sounds. Also, noise has a high frequency. We're using pre-emphasis to make the system less noise-sensitive later in the process. For some applications, we just need to undo the boost at the end.

Pre-emphasis uses a filter to boost higher frequencies. Below is the before and after signal to boost the high-frequency signal.

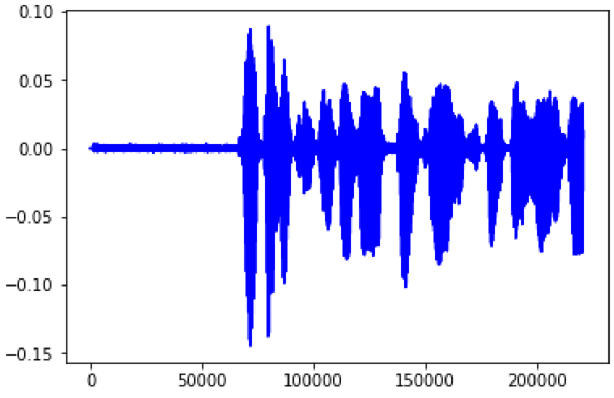


Fig.2: Before Pre-emphasis

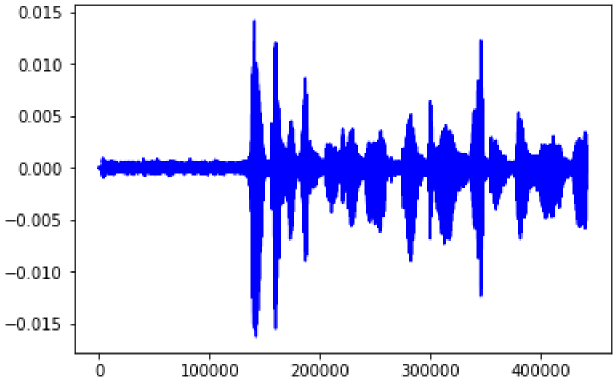


Fig.3: After Pre-emphasis

A.2. *Framing*

We need to split the signal into short-time frames, after pre-emphasis. The explanation for this step is that frequencies in a signal change over time, so in most situations, there's no point in transforming the Fourier throughout the whole signal by losing the signal's frequency contours over time. To prevent this, we can safely assume that the frequencies in a signal are stationary over a very short time period. Thus, by making a Fourier transformation over this short-term frame, we can achieve a good approximation of the frequency contours of the signal by concatenating adjacent frames.

Typical frame sizes range from 20 ms to 40 ms in speech processing, with 50 percent (+ /-10 percent) overlapping between consecutive frames. For the frame size, the common settings are 25 ms, frame size = 0.025 and 10 ms stride (15 ms overlap), frame stride = 0.01.

```
(998, 1102)
```

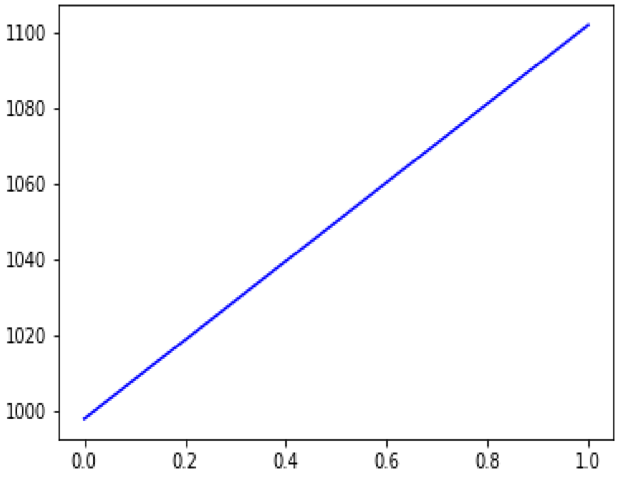
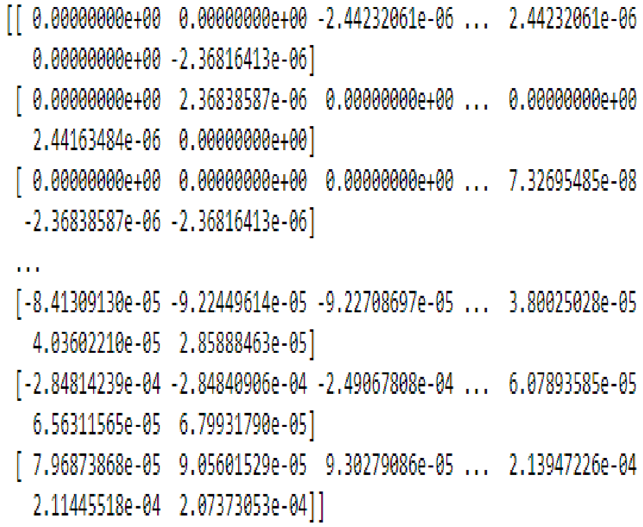


Fig.3. Frame shape and plot

A.3. Windowing

After the signal is sliced into frames, we add to each frame a window function such as the Hamming window. The next step in processing is to window each frame to minimize signal discontinuities at the beginning and end of each frame. The idea is to reduce the spectral distortion by tapering the signal at the beginning and end of each frame via the window. If the window is defined as w(n), 0 ⩽ n ⩽ N-1, where N is the number of samples in each frame, then the result of the signal window is y(n) = x(n) * w(n) where x(n) is the speech signal being processed.



```
[[ 0.00000000e+00  0.00000000e+00 -2.44232061e-06 ...  2.44232061e-06
   0.00000000e+00 -2.36816413e-06]
 [ 0.00000000e+00  2.36838587e-06  0.00000000e+00 ...  0.00000000e+00
   2.44163484e-06  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  7.32695485e-08
  -2.36838587e-06 -2.36816413e-06]
 ...
 [-8.41309130e-05 -9.22449614e-05 -9.22708697e-05 ...  3.80025028e-05
   4.03602210e-05  2.85888463e-05]
 [-2.84814239e-04 -2.84840906e-04 -2.49067808e-04 ...  6.07893585e-05
   6.56311565e-05  6.79931790e-05]
 [ 7.96873868e-05  9.05601529e-05  9.30279086e-05 ...  2.13947226e-04
   2.11445518e-04  2.07373053e-04]]
```

Fig.4: Features of Window

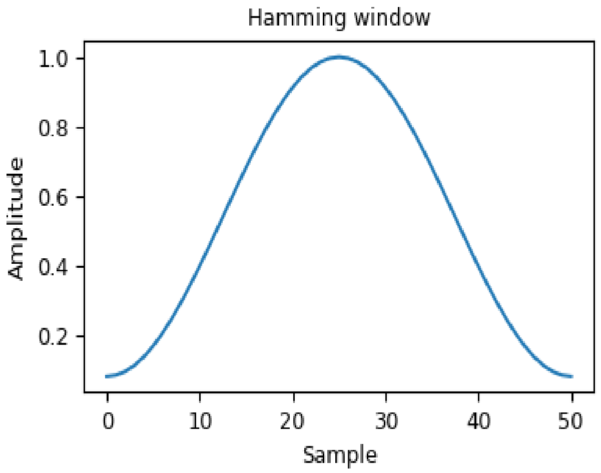
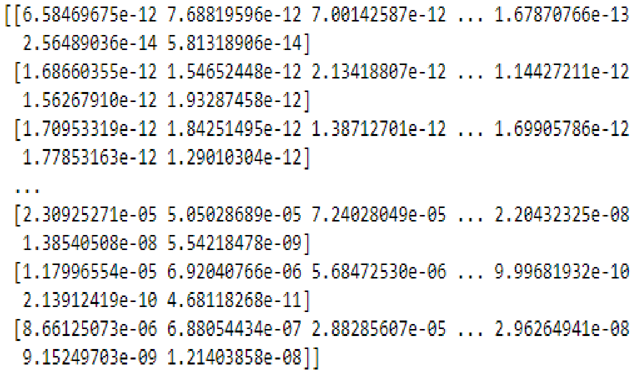


Fig.5 Hamming Window

A.4. Fourier-Transform and Power Spectrum

We now can perform N-point FFT on each frame to compute the frequency spectrum, called Short time Fourier-Transform (STFT), with N usually 256 or 512, NFFT = 512, and then use the following equation to measure the power spectrum (periodogram).

$$P=|FFT(x_i)|_{2}N$$



```
[[6.58469675e-12 7.68819596e-12 7.00142587e-12 ... 1.67870766e-13
  2.56489036e-14 5.81318906e-14]
 [1.68660355e-12 1.54652448e-12 2.13418807e-12 ... 1.14427211e-12
  1.56267910e-12 1.93287458e-12]
 [1.70953319e-12 1.84251495e-12 1.38712701e-12 ... 1.69905786e-12
  1.77853163e-12 1.29010304e-12]
 ...
 [2.30925271e-05 5.05028689e-05 7.24028049e-05 ... 2.20432325e-08
  1.38540508e-08 5.54218478e-09]
 [1.17996554e-05 6.92040766e-06 5.68472530e-06 ... 9.99681932e-10
  2.13912419e-10 4.68118268e-11]
 [8.66125073e-06 6.88054434e-07 2.88285607e-05 ... 2.96264941e-08
  9.15249703e-09 1.21403858e-08]]
```

Fig.6: features of NFFT

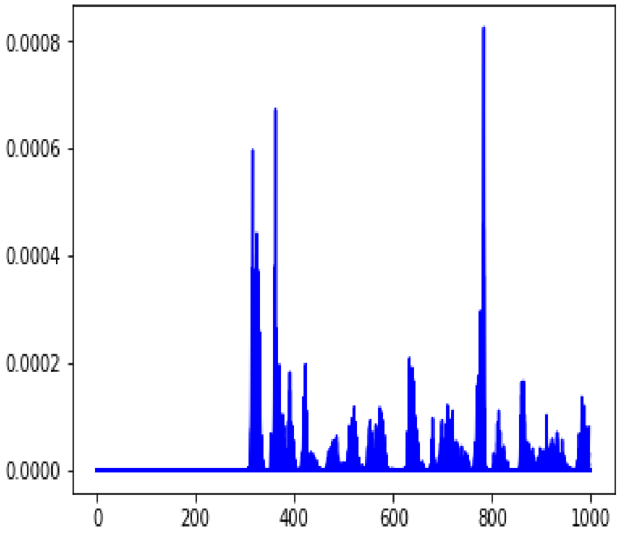
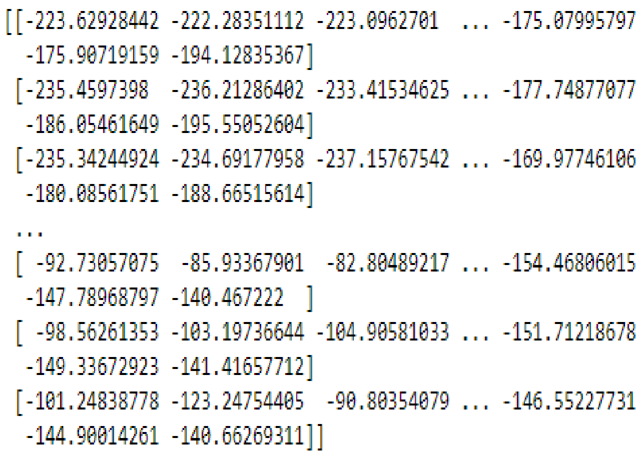


Fig.7 Plot of NFFT

A.4. Filter Banks

The final step towards computing filter banks is to add triangular filters, usually 40 filters, nfilt = 40 to the power spectrum on a Mel-scale to obtain frequency bands. The Mel-scale attempts to imitate the non-linear experience of sound in the human ear by being more selective at lower frequencies and less discriminatory at higher frequencies.
The filter is triangular in the center of the filter bank and has a response of 1 until it hits the middle of two adjacent filters, with a response of 0.



```
[[-223.62928442 -222.28351112 -223.0962701  ... -175.07995797
  -175.90719159 -194.12835367]
 [-235.4597398  -236.21286402 -233.41534625 ... -177.74877077
  -186.05461649 -195.55052604]
 [-235.34244924 -234.69177958 -237.15767542 ... -169.97746106
  -180.08561751 -188.66515614]
 ...
 [ -92.73057075  -85.93367901  -82.80489217 ... -154.46806015
  -147.78968797 -140.467222  ]
 [ -98.56261353 -103.19736644 -104.90581033 ... -151.71218678
  -149.33672923 -141.41657712]
 [-101.24838778 -123.24754405  -90.80354079 ... -146.55227731
  -144.90014261 -140.66269311]]
```

Fig.8 Extracted Filter Banks

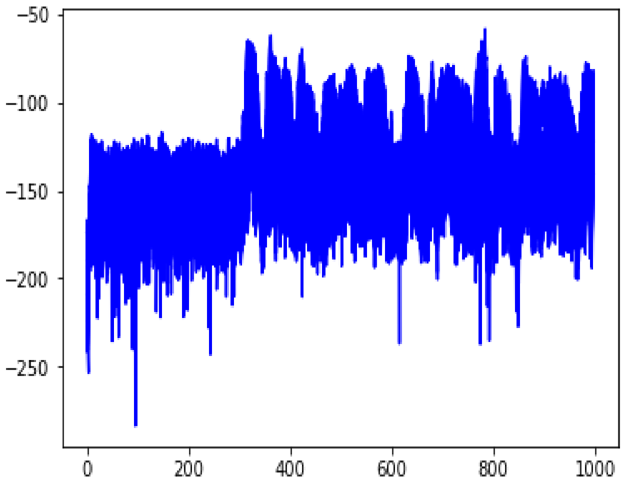


Fig. 9 Filter Bank's Plot

3. Model

A. Gaussian Mixture Model

Gaussian mixture models are a probabilistic model for describing subpopulations normally distributed across an overall population. Generally, mixture models do not need to know which subpopulation a data point belongs to, allowing the model to automatically learn the subpopulations. Since the role of subpopulation is not recognized, this is a form of unsupervised learning.
GMMs were used to extract features from speech data, and were also commonly used in object tracking of multiple objects, where the number of mixture components and their means predict object positions in a audio sequence at each frame. The concept of training a GMM is to approximate the distribution of probabilities in a class by means of a linear combination of the ' k ' Gaussian distribution-clusters, also called the GMM components. The probability of data points for a model (feature vectors) is given as equation:

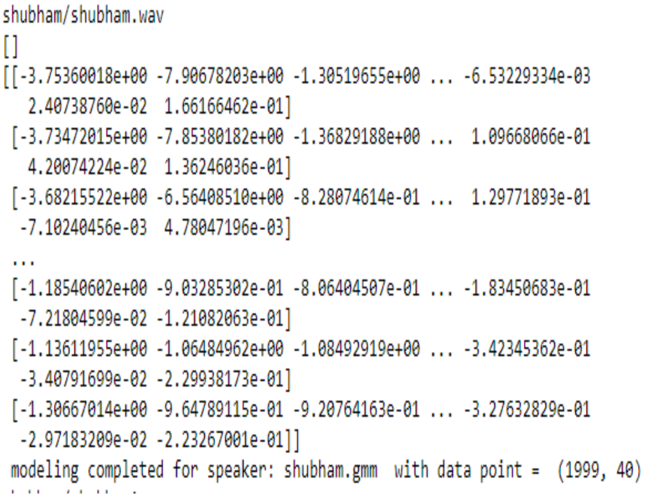
$$P(X|\lambda) = \sum_{k=1}^{K} w_k P_k(X|\mu_k, \Sigma_k) \quad (1)$$

Where the Gaussian distribution is $P_k(X|\mu_k, \Sigma_k)$.

$$P_k(X|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi|\Sigma_k|}} e^{\frac{1}{2}(X-\mu_k)^T \Sigma^{-1}(X-\mu_k)} \quad (2)$$

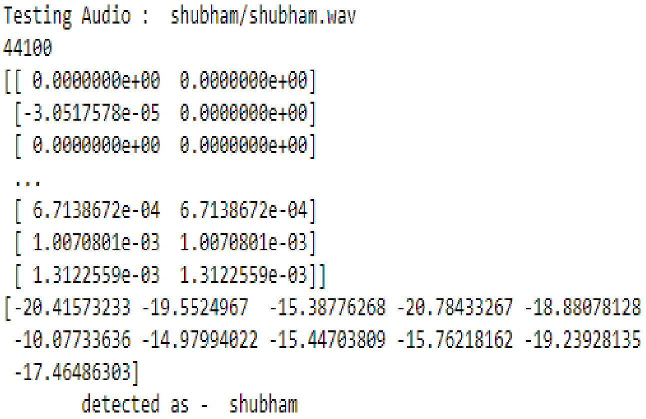
For the parameters mean μ, co-variance matrices Σ and weight w of the k elements, the training data of the class X_i is used.

The GMM object includes the number of n components to be placed on the data, the number of iterations to n_iter in order to determine the co-variance covariance_type parameters to be accepted between the functionality and the number of times n_iter is to occur. The initialization that produced the best results is kept. The function fit) (then estimates the parameters of the model using the EM algo.
The number of iterations needed for the log-likelihood function to converge and the log-likelihood to converge.



```
shubham/shubham.wav
[]
[[-3.75360018e+00 -7.90678203e+00 -1.30519655e+00 ... -6.53229334e-03
   2.40738760e-02  1.66166462e-01]
 [-3.73472015e+00 -7.85380182e+00 -1.36829188e+00 ...  1.09668066e-01
   4.20074224e-02  1.36246036e-01]
 [-3.68215522e+00 -6.56408510e+00 -8.28074614e-01 ...  1.29771893e-01
  -7.10240456e-03  4.78047196e-03]
 ...
 [-1.18540602e+00 -9.03285302e-01 -8.06404507e-01 ... -1.83450683e-01
  -7.21804599e-02 -1.21082063e-01]
 [-1.13611955e+00 -1.06484962e+00 -1.08492919e+00 ... -3.42345362e-01
  -3.40791699e-02 -2.29938173e-01]
 [-1.30667014e+00 -9.64789115e-01 -9.20764163e-01 ... -3.27632829e-01
  -2.97183209e-02 -2.23267001e-01]]
 modeling completed for speaker: shubham.gmm  with data point =  (1999, 40)
```

Fig. 10 Gaussian Mixture Model with Data Point



```
Testing Audio :  shubham/shubham.wav
44100
[[ 0.0000000e+00  0.0000000e+00]
 [-3.0517578e-05  0.0000000e+00]
 [ 0.0000000e+00  0.0000000e+00]
 ...
 [ 6.7138672e-04  6.7138672e-04]
 [ 1.0070801e-03  1.0070801e-03]
 [ 1.3122559e-03  1.3122559e-03]]
[-20.41573233 -19.5524967  -15.38776268 -20.78433267 -18.88078128
 -10.07733636 -14.97994022 -15.44703809 -15.76218162 -19.23928135
 -17.46486303]
        detected as -  shubham
```

Fig.11: Log_likelihood of speakers

CONCLUSION

This paper gives an overview of the speaker recognition that includes various methods of features extraction and model training mainly focused on MFCC and GMM.

Security puts great setbacks for sensitive information in today's world. Recognition of speakers is a multidisciplinary biometrics branch that can be used to classify and validate speakers to protect sensitive information. Therefore to prevent unauthorized access, a voice-based recognition system needs to be developed that provides a solution for financial transaction and protection of personal data that would minimize theft. MFCC features extraction technique and GMM modeling technique use in speaker recognition are discussed in this paper which can be applied to develop a real time application for speaker identification and verification system for confidential data securing in the future.

REFERENCES

[1] George R Doddington, Member, IEEE, "Speaker Recognition- Identifying People by their Voices", proceedings of the IEEE, Vol. 73, no. 11, pp. 1651-1664, November 1985.

[2] Richard D. Peacocke, and Daryl H. Graf, "An Introduction to Speech and Speaker Recognition ", IEEE, pp. 26-33, August 1990.

[3] Tomi kinnunen, Haizhou Li, "An overview of textindependent speaker recognition: From features to supervectors", Speech Communication 52, pp. 12-40, 2010.

[4] Ling Feng , Kgs. Lyngby "Speaker Recognition", Thesis, Technical University of Denmark Informatics and Mathematical Modeling, Denmark, 2004.

[5] J. Wu and J. Yu, "An improved arithmetic of MFCC in speech recognition system," in Electronics, Communications and Control (ICECC), 2011 International Conference on, 2011, pp. 719-722

[6] Yingjie He, Liwei Ding, Yuxian Gong, Yongjin Wang,"Real-time Audio & Video Transmission System Based on Visible Light Communication ",June 2013.