Biometric Voice Recognition system using MFCCand GMM with EM

Rahul Pudurkar
Don Bosco Institute of TechnologyMumbai,
Maharashtra, India

Shruti Patil
Don Bosco Institute of Technology Mumbai,
Maharashtra, India

Gazala Ansari
Don Bosco Institute of TechnologyMumbai,
Maharashtra, India

Shaikh Phiroj
Don Bosco Institute of TechnologyMumbai,
Maharashtra, India

ABSTRACT

The current solutions for passwords that are used for authenticating users can be insecure sometimes and might be hacked easily. Securing data and confidential information is very important in today's world. Biometric is known as a unique biological characteristic of a human being. Using it as a means for securing devices or certain data, has proved to be very useful in recent times. This paper aims to implement a voice-based login authentication system. The voice biometrics (voiceprint) of the user is taken as an input to help authenticate the individual, along with traditional password pin validation, resulting in two-factor authentication. The system tries to identify not only the features of the voice but also collect the signature words of the user (the words that the user speaks differently, e.g., dialects, pronunciation, etc.) and store them in the database. So, when the user is trying to login into the system, a random sentence will be displayed on the screen with their signature words present along with some random words. When the user speaks the sentence, the real-time voiceprint will be compared with the voiceprint present in the database stored during registration. If both these voiceprints match, the individual will be considered an authentic user and will be given permission to access the system. To avoid any kind of malpractice, random words help the system be more secure for taking real-time voiceprints.

General Terms

Machine Learning, Authentication System, Speech Recognition, Natural Language Processing, Two-factor authentication.

Keywords

Voice Recognition, Signature Words, GMM, MFCC, Voice Activity Detection (VAD).

1. INTRODUCTION

The innovation of biometric authentication has helped in the sector of security. To put it simply, a biometric is a unique feature of a human being, such as a fingerprint, iris or eye scan, handprint, or even voiceprint. The popular ways of keeping sensitive information and resources safe and secure are to use passwords, patterns, and PINs. These methods let the users authenticate themselves by entering a "secret" password that they had previously created. Such systems are not proving to be the most secure method of authentication. Biometric characteristics of an individual are unique and can therefore be used to authenticate a user's access to various systems. One of the biometric data types is voice, which studies the distinctive sound waves in a user's voice as he or she speaks. It could be used in various applications, such as in banks, schools, any private organization, or even securing a small device that has vital information in it. These systems could be used to identify authentic users. To secure human-machine interaction and to improve user experience, a voice-based authentication system is implemented in this paper, which lets the user enter into the system using their voice. The system design is divided into three modules. 1) the Registration Module, 2) the Text Generation Module, 3) the Verification Module. During the user's registration into the system, a random sentence is displayed, and the speaker is required to read it aloud.

Before the features of the voice print are extracted, there are some preprocessing phases that are carried out first. The speech signal is processed to remove any background noise that can be present in the voiceprint so that important attributes of the voice can be extracted easily and accuracy can be maintained. When the voiceprints are taken, they are stored in a database. The signature words (unique words: the words that the user speaks differently) are extracted from the voice prints. Each person has a different vocal pitch or frequency of voice [10], controlled by unique brain cells. When, after registration, the user tries to gain access to the system with the help ofthe signature words, a sentence is generated and displayed on the screen. Again, a real-time sample voiceprint is taken analyzed, and compared with the sample present in the database. If both these samples match, the user is considered authentic and is given access to the system; or else, he/she is considered an imposter.

2. LITERATURE REVIEW

In this paper, ASR is proposed based on two important machine learning paradigms: Artificial Neural Network (ANN) and Support Vector Machine (SVM) for the rare and geographically important Indian dialect "Chhattisgarhi" [1]. Conventional feedforward ANNsand SVMs were applied to a dataset of a maximum of 50 isolated words from 15 speakers. The performance of these machine learning paradigms is compared to that of a state-of-the-art Hidden Markov Model (HMM). The tendency of ASR to be speaker dependent and independent has been extensively investigated using speaker variation experiments. In addition, the reliability and stability of the ASR were confirmed by numerical verification. A comprehensive review of ASR techniques from the literature is presented along withASR systems designed for Indian languages.

Sarabjeet Singh [2] has tried to log into the Linux system using voice, and the user's voice here will act as an input to the Linux system at the login time. In this work, three stages are present for the login process: the first step is to provide your text password; the second is speech verification, and the third is identifying the authentic users. In order to prevent record replay problems, random words have been used. MFCC is used as a feature extraction technique. Before extracting features of the voice, Voice Activity Detection (VAD) was done to remove the noise. The K-means algorithm is used for identifying the authentic user. Passphrases are generated randomly every time the user tries to login into the system. An LTSD (long- term speech detection) algorithm is used to identify envelopes of areas where the user is saying random words and ignore the parts wherein noise is present.

Annie Shoup, Tanya Talkar, Jodie Chen, Anubhav Jain ashoup, tjtalkar, jodiec, ajain94 presents an analysis of current voice authentication methods, a security policy as well as two different implementations of ways of supporting biometric two-factor authentication [3]. The implementations provide simple ways to be able to drop in two-factor authentication into existing applications, while also presenting ideas for future work. By presenting a voice biometric-based authentication system, which paves the way for additional work in open-source biometric authentication system.

A literature survey paper wherein the authors have discussed voice and speaker recognition, deep learning, and traditional methods [4]. They have also discussed the datasets that could be used in speaker recognition and voice comparison. The paper gives information aboutdifferent approaches to voice comparison, and an explanation of the generalized process of voice comparison is also given.

M. Reda, N. G. Mohammed and R. A. Abdel Azeem Abul Seoud designed an application called "SVBiComm" (Sign-Voice Bidirectional Communication) which facilitates communication between non-disabled, "Deaf/Dumb" and visionless people. This system can help visionless people to understand words actioned by "deaf/dumb" people, and the message of the blind user will be converted into action representation and received by the deaf user [5]. In this paper the focus has been on which methods have been used for voice extraction and their features. In order to remove noise from therecorded audio, a Cepstral Mean Subtraction (CMS) filter is used. Inthe recognition process, vector quantization is used, which helps in improving this process. MFCC is used for feature extraction techniques.

Speaker Identification Using GMM with MFCC [6] targets the implementation of MFCCs (and Delta MFCCs) extracted features with the GMM model to identify the speaker. The Gaussian MixtureModel was trained by the EM (Expectation-Maximization) algorithm, and it was used to recognize the speaker based on log probability. The use of GMM leads to more flexibility in terms of cluster covariance and lower EER compared to other models.

After analyzing a couple of structures, it was concluded to devise a system that consists of audio characteristic extraction using MFCCs and voice popularity with the aid of authenticating a random passphrase using GMM and EM.

3. PROPOSED ARCHITECTURE

The proposed system is divided into three modules shown in (Figure 1) Registration Module, 2. Text Generation Module 3. VerificationModule

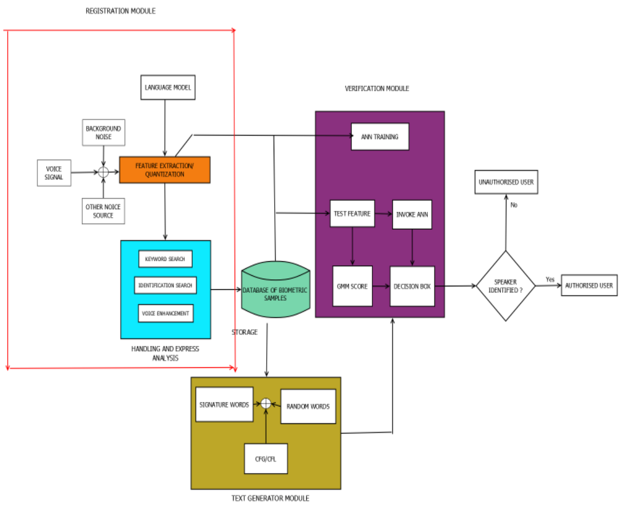


Figure 1. System Architecture

3.1 Registration Module

The registration module is the first step in the system. This module collects all the details of the user, which is used for user authentication. Along with the user's details, a voice print of the user is also taken in this step. To accurately store the voice print of the user, a random sentence is displayed to the user, after which the user is asked to speak the sentence through a microphone. The system then grabs the voiceprint, enhances it and removes any background noises if present. After that, the system looks for signature words (i.e., the words the user speaks differently), which will be stored in the database that will be used for authentication of the user based on pitch and amplitude calculations.

3.2 Text Generation Module

In this module, the signature words, which were stored earlier for each corresponding user, are clubbed along with random words, which are generated using the Random Word library to form a sentence. Thissentence is then displayed to the user, who records his or her voiceprint using a microphone. A long-term spectral divergence (LTSD) algorithm [3] is used to detect the random passphrase. This module isalso used during the registration phase.

3.3 Verification Module

This module helps in the identification and verification of voiceprints. With the help of signature words stored earlier, system validates that the voiceprint is not a system-generated voiceprint given as an input since every time new random words are added. Identification ofthe voiceprint is a very important step in extracting the required features from the voice sample, which can help to accurately authenticate the user. Identifying the individual is done by comparingthe voice imprints with every voiceprint stored in the database. This is known as speaker identification or voice identification.

This is useful for fraud detection, such as checking whether a person claiming to log in to the system is in fact an authorized person trying to access the system or not. The point of identification is most often used to *detect* fraud within a system. On the other hand, verification helps one to understand that the voiceprint provided is the same as that present in the database because it is verified by the system using unique voice features extracted from the user's voiceprint. Verification helps to identify the user and to avoid any false positiveor false negative results from a user.

Verification and identification can be achieved by training the systemwith different samples of voiceprints. This helps the system to identify the fraud and to authenticate users more accurately, eventually giving the best result possible to avoid fraud. Here, the GMM score of a real time voiceprint is compared with the score stored in the database. With the help of this, it is easier to identify whether the user is authenticated or not. The decision box makes thedecision of accepting the user's request to access the system or rejecting it.

4. IMPLEMENTATION

The basic flow diagram (Figure 2) of the system includes registration, VAD, Comparison and validation:

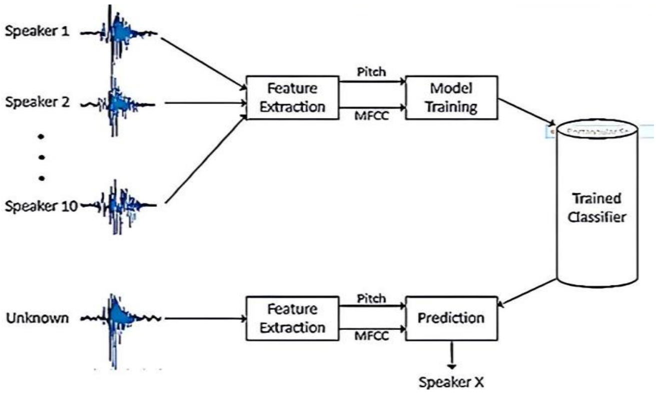


Figure 2. System Architecture

For the registration, the user first needs to enter a username and a desired password. After this, the voice activity detection process is carried out by the system, wherein it removes the background noise if it is present during voice recognition and collects the real-time voice sample. Once the VAD is done, random words will be displayed on the screen, which the user will have to speak out aloud. In the background, IBM Watson converts the user's spoken words into text and generates a fuzzy score; if the fuzzy score exceeds 65, the user is successfully enrolled in the system. MFCC is used for the feature extraction of the real time voice sample provided by the user as an input.

The zero-crossing rate (ZCR) is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive It is widely used to classify percussive sounds. ZCR is defined formally as:

$$zcr = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{R<0} \left(S_t S_{t-1}\right)$$

where S is a signal of length T and $1_{R<0}$ is an indicator function

For authentication the user first needs to provide the username and password which he or she has created during the registration process. Once the username and password are validated a random sentence is displayed on screen along with the signature words clubbed together for the user to speak. This will act as an input to the system. Once theuser gives his/her input to the system through a microphone the systemidentifies the user based on the features of the voice and comparesthe log probabilities of both the samples. based on the GMM score thesystem decides whether the user trying to access the system is authentic or not. If there is a match the user is given permission to access the system and if not the user/unauthorized person will be deprived from the system. The above process is illustrated in (Figure 3).

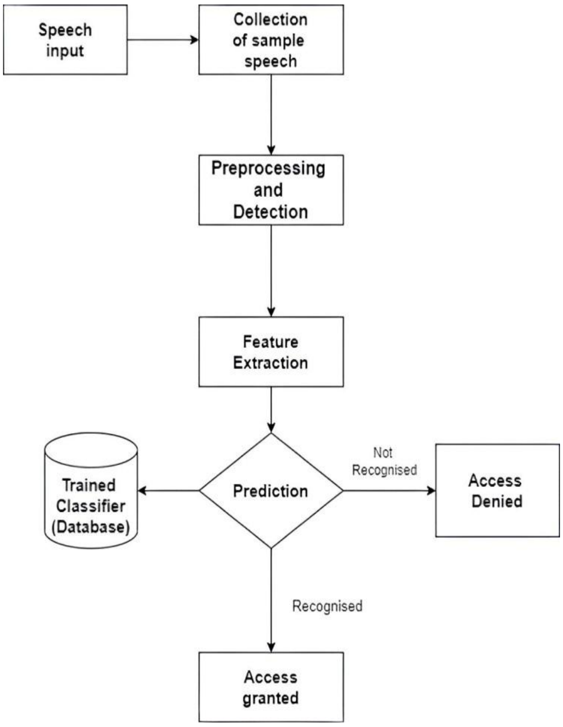


Figure 3. Process of Voice Authentication

4.1 Gaussian mixture models

To understand GMM, basic understanding of clustering is needed.Consider there are sets of data points that need to be put together into several parts or clusters based on the similarities theyshow. In technical terminology of machine learning, this is known as Clustering. There are several methods available for clustering from which one of them is the Gaussian mixture model. GMM is nothing but a probabilistic model that makes an assumption that every data point generated from gaussian distributions with parameters which areunknown.

In voice authentication GMM score of every individual is calculated which helps the system to identify the authenticated user and help to generate better accuracy of the system. Different voice biometric has its different GMM score which plays animportant role to distinguish voiceprint from others efficiently.

The probability distribution [3] is given by the following equation:

$$\mathrm{P(X \mid \lambda)} = \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{w\,i\,Pi\,(X \mid \text{⍰}i}, \sum \mathrm{i)}$$

Where $\mathrm{P_i\,(X \mid \mu_i}, \sum i)$ is the Gaussian distribution

$$\mathrm{Pi\,(X \mid \text{⍰}i}, \sum \mathrm{i)} = \frac{\mathrm{e}^{\frac{1}{2}}\mathrm{(X-\text{⍰}_i)^T \Sigma^{-1} (X-\text{⍰}_i)}}{\sqrt{\mathrm{2M}\,|,\sum \mathrm{i}|}}$$

The training data X of the class λ is used to estimate the parameters mean μ, covariance matrices ∑ and weights *w* of these *i* components.

The extracted voice features can be trained using the GMM (Gaussian Mixture Modeling) by the Expectation-Maximization Algorithm, to train and store them in the database. sklearn. mixture package in Python was used to train the GMM model by the acquired MFCC features. The EM [7] iterates between performing anexpectation (E) step, which creates a function for the expectation of the log-likelihood, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found in the previous step. This method is used to update the parameters of *k* number of distributions.

4.2 MFCC (Mel-frequency cepstral coefficients)

The Mel-Frequency Cepstral Coefficient (MFCCs) is used to extract features from the voiceprint provided as an input. The MFCC feature extraction technique is basically windowing the signal, which means splitting the signal into short time frames. A Discrete Fourier Transform (DFT) is applied to each window. The logarithmic curve that models pitches is applied to the magnitude generated, and the frequencies are wrapped on a Mel scale. And in the last step, inverse DCT is applied. When all these steps are applied to the voiceprint, different features are extracted from it, which helps the system to identify which voiceprint is authentic and which is not. Below formula displays how to calculate delta features from MFCCs, the following equation is used:

$$d_t = \frac{\Sigma_{n=1}^{N} \ n(C_t h_n \ - \ C_{t-n}}{2\Sigma_{n=1}^{N} \quad n}$$

where 'N' is the number of deltas summed over. python_speech_features package was used to acquire the 40-dimensional MFCC features.

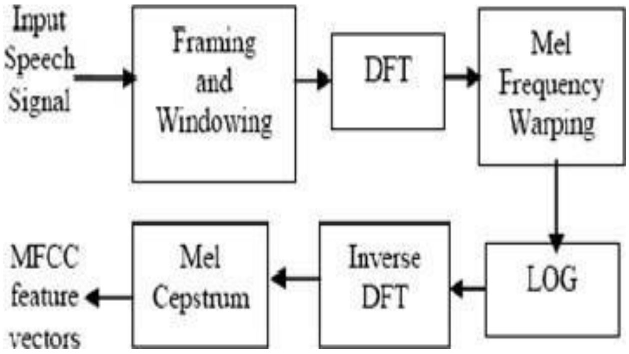


Figure 4. MFCC Flowchart

4.3 VAD (Voice activity Detection)

Voice activity detection is used on the different voice samples to detect the speech of an individual from any other background noises present in the audio file. Therefore, it is also known as voice detection or speech detection. In a speech recognition-based application, VAD is an important technology to help the system understand the voiceprint more accurately and, in turn, provide the correct information the system wants to deliver. It helps to eliminate any background noises or sounds and to only focus on the speech of the individual accessing the system. VAD also helps in maintaining the accuracy of the software and providing the best results. The system uses VAD to ignore the background noise and focus more on the voiceprint present in the recording. Once the voice sample is passed from VAD, it is passed on to the MFCC algorithm for its feature extraction mechanism.

4.4 Voice Matching based on GMM model

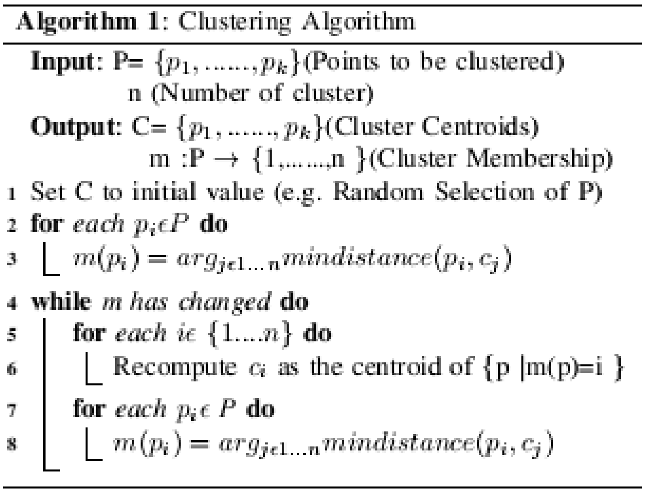
Voice matching is the process of authenticating the registered user by verifying real-time voice with voice samples stored in the database. GMM recognition recognizes the speaker based on log probability. The log-likelihood of the voice sample is compared with the log- likelihood of the stored voice sample of the same speaker. The speaker model with the highest likelihood score is considered as the identifiedspeaker.

4.5 Differential MFCC

To increase robustness of speaker authentication MFCC is improved to differential MFCC. Differential MFCC [6] exploits the property of Mel-Frequency coefficient frames being closely related and works on the differences between the consecutive frames. This ensures that the coefficients generated will be diverse, thus improving the accuracy of authentication.

4.6 K-means Clustering

To reduce the amount of data being processed, K- means clustering algorithm is applied to the coefficients obtained from MFCC. K-means clustering algorithm is an unsupervised learning algorithm that classifies the given dataset through a certain number of clusters. The steps involved in K-means algorithm is shown in the algorithm 1.



```
Algorithm 1: Clustering Algorithm
  Input: P= $\{p_1, ......, p_k\}$(Points to be clustered)
         n (Number of cluster)
  Output: C= $\{p_1, ......, p_k\}$(Cluster Centroids)
         m :P → {1,.......,n }(Cluster Membership)
1 Set C to initial value (e.g. Random Selection of P)
2 for each $p_i \epsilon P$ do
3     $m(p_i) = arg_{j\epsilon 1...n} mindistance(p_i, c_j)$
4 while m has changed do
5     for each $i\epsilon \{1....n\}$ do
6         Recompute $c_i$ as the centroid of {p |m(p)=i }
7     for each $p_i \epsilon P$ do
8         $m(p_i) = arg_{j\epsilon 1...n} mindistance(p_i, c_j)$
```

Explanation of Algorithm 1:

1. Input: Set P is the MFCC Coefficients which are to beclustered by the k-means.
2. n: Number of centroids to be calculated
3. output: C Output array containing the final clusters
4. m: Mapping function between the input point and the cluster
5. Initializes the clusters from the input values
6. The loop in line 2-3 classifies the input data points to the cluster based on the minimum Euclidean distance metric. Euclidean distance calculation needs each data point and cluster point as input.
7. The while loop has two inner loops which will run till the clusters are stabilized.
8. The for loop in line 5-6 will re-compute the centroids for the modified clusters generated in step 6.
9. The for loop in line 7-8 will re-assign the data points to the re-computed clusters to achieve stability

The result of K-means algorithm is a collection of centroids whose size is less in comparison to the total number of MFCC coefficients obtained. This centroids collection is used for uniquely identifyingthe speaker.

5. RESULT AND DISCUSSION

The user is registered in the system with his username, password, andvoiceprint; these characteristics are used to compare it with the real-time individual present in the system. When logging into the system, it asks the user for his or her username and password. After successful authentication based on these traits, the user is taken forward to the next page where his/her voiceprint will be collected. The user's voiceprint forms a pattern of waveform from the words that are displayed to him/her on the screen. These patterns of waveform are different for every individual, which helps in identifying the correct user as shown in (Table 1). The output of feature extraction for the voice imprints are shown in (Figure 5) and (Figure 6). (Figure 5) is the representation of pitch for a voiceprint and (Figure 6) is the representation of VAD [2] using ZCR and Volume.

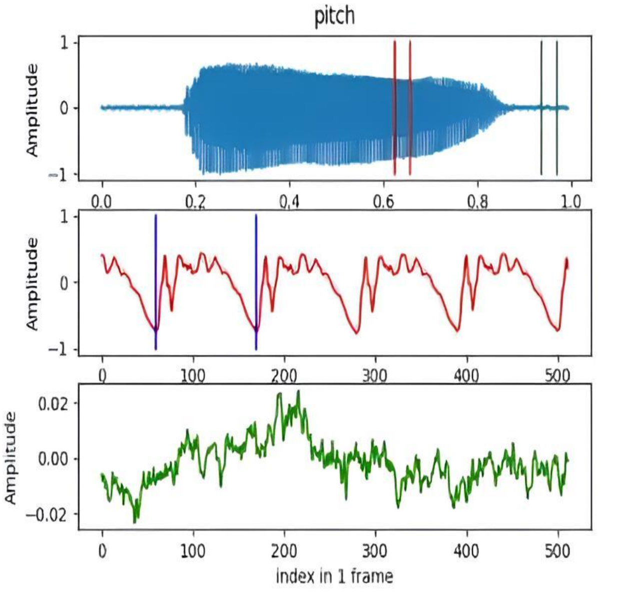


Figure 5. Graphical Representation of Pitch

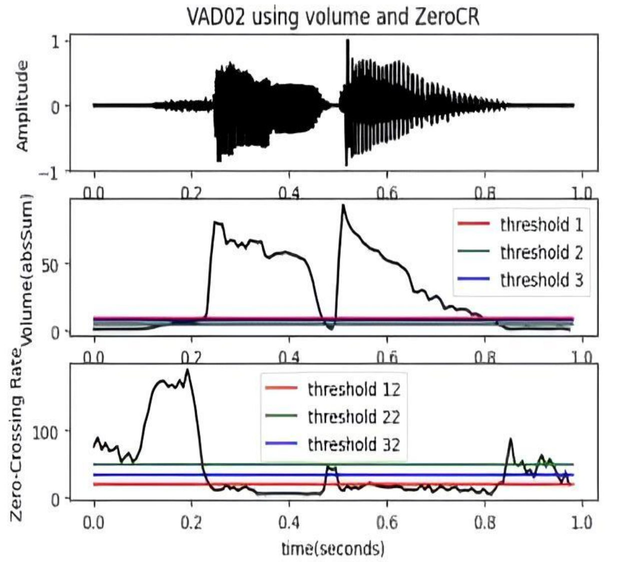


Figure 6. VAD using volume and ZCR

Table 1. GMM Scores Results.

Username	GMM score
User 1	-26.699111111
User 2	-31.690982735
User 3	-27.092616742
User 4	-25.872346291

(Table 1) shows the users GMM scores. For authentication, the user'ssample voice is taken as an input and its features are extracted using MFCC and GMM. From every user's model (GMM file), the log likelihood (the weight, mean, and variance of the mixture components) is calculated, and these GMM scores are stored inside a matrix structure. If the GMM score generated by the user trying to access the system is above the power level, then the user is identified as an authentic user.

The performance of any biometric recognition system is measured experimentally by calculating the Genuine Acceptance Rate (GAR), False Acceptance Rate (FAR) and False Rejection Rate (FRR) along with Total Error Rate (TER). A good system must have a high GAR and low TER. These parameters are determined as follows

FAR (%) = (No of false acceptance)/ (Total Imposter test) ×100

FRR (%) = (No of false rejections)/ (Total Genuine test) ×100

GAR (%) = 100% - FRR%

TER (%) = FAR % + FRR %

Table 2. Results of voice authentication

	GAR	FAR	FRR	TER
Voice	94.33%	0%	2.835%	2.835%

From (Table 2), an observation can be made that this system is more secure than conventional or unimodal authentication with a GAR of 94.33%. Thus, this system can efficiently meet the requirements for practical applications along with traditional password and username login system enabling two-factor authentication.

6. CONCLUSION AND FUTURE SCOPE

In this paper, a real-time voice-based authentication system is developed. This system could be used in various applications where user authentication is required, such as in laptops, smartphones, banking applications, etc. The uniqueness of this system is in the signature words (words that the user speaks differently) that the system learns through training and then uses them to authenticate theusers. The Mel Frequency Cepstral Coefficients (MFCC) are used for feature extraction of the voice, and a Gaussian Mixture Model (GMM) is used for identification of the authentic user along with EM. For removing background noise, voice activity detection is used. From the results, it can be concluded that the system has high accuracy and is more secure than conventional or unimodal authentication.

Coming to the future scope, the concept of "stop words" could be used. Accuracy of the system can be further increased with classified datasets for male and female voice samples trained specifically for complex words and more than one syllable words. Other Biometric measures can be integrated to further enhance the security of the system. Personalized commands can be added for specific users in order to unlock the system. System is trained and tested for English Language and in future the system can be extended to support more than one regional languages.

7. REFERENCES

[1] N. D. Londhe, M. K. Ahirwal and P. Lodha, "Machine learning paradigms for speech recognition of an Indian dialect," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016, pp. 0780-0786, doi:10.1109/ICCSP.2016.7754251.

[2] S. Singh and M. Yamini, "Voice based login authentication for Linux," 2013 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, 2013, pp. 619- 624, doi: 10.1109/ICRTIT.2013. 6844272.xs.

[3] Annie Shoup, Tanya Talkar, Jodie Chen, Anubhav Jain ashoup, tjtalkar, jodiec, ajain94," An Overview and Analysis of Voice Authentication Methods".

[4] N. H. Tandel, H. B. Prajapati and V. K. Dabhi, "Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 459-465, doi: 10.1109/ICACCS48705.2020.9074184.

[5] M. M. Reda, N. G. Mohammed and R. A. Abdel Azeem Abul Seoud, "SVBiComm: Sign-Voice Bidirectional Communication System for Normal, "Deaf/Dumb" and Blind People based on Machine Learning," 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, 2018, pp. 1-8, doi: 10.1109/CAIS.2018.8441985.

[6] Tahira Mahboob, Memoona Khanum, Malik Sikandar HayatKhiyal Ruqia Bibi. "Speaker Identification Using GMM with MFCC" IJCSI International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015.

[7] "A Speaker Recognition System Using Gaussian Mixture Model, EM Algorithm and K-Means Clustering" (2018) Mr. Ajinkya N. Jadhav and Mr. Nagaraj V. Dharwadkar.

[8] R.R. Lawrence, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", In proc. Of IEEE, Vol. 77, No.2, 1989, pp. 257-286.

[9] Shaughnessy, "Interacting with computer by voice automatic speech recognition and synthesis", In proc. Of IEEE, Vol. 9, No.91, 2003, pp.1272-1305.

[10] K. Kolhatkar, M. Kolte, and J. Lele, "Implementation of pitch detection algorithms for pathological voices," In2016 International Conference on Inventive Computation Technologies (ICICT) 2016Aug26 (Vol. 1, pp. 1-5). IEEE.

[11] "Text Independent Speaker Recognition System using GMM"(2012) S G Bagul and Prof R.K. Shastri.