Human Voice Authentication Using Mel-Frequency Cepstral Coefficients and Gaussian Mixture Model

Mrs. Dipti Pawade
Department of Information Technology
K. J. Somaiya College of Engineering, Vidyavihar
Mumbai, India
diptipawade@somaiya.edu

Ms. Avani Sakhapara
Department of Information Technology
K. J. Somaiya College of Engineering, Vidyavihar
Mumbai, India
avanisakhapara@somaiya.edu

Ms. Diya Bakhai
Department of Information Technology
K. J. Somaiya College of Engineering, Vidyavihar
Mumbai, India
d.bakhai@somaiya.edu

Ms. Rujuta Ashtekar
Department of Information Technology
K. J. Somaiya College of Engineering, Vidyavihar
Mumbai, India
rujuta.a@somaiya.edu

Ms. Shruti Tyagi
Department of Information Technology
K. J. Somaiya College of Engineering, Vidyavihar
Mumbai, India
shruti.tyagi@somaiya.edu

***Abstract*— Voice operated devices are becoming popular nowadays. For this it is necessary that voice authentication is secure. In this paper, we address some known attacks like replay, personification and attacks using AI voice bots and limitations like text and language dependency of human voice authentication systems. We have also developed an interactive system to tackle these problems. The system verifies the user by performing voice matching as well as on an intellectual level by asking questions which only humans are able to answer and not any AI bot. In the system, an average user requires around 35 seconds for registration and around 25 seconds for authentication. The system's accuracy comes out to be 97.8% for English speakers and 95% for Hindi speakers.**

Keywords—Speech-to-Text conversion, Classification model, Voice Recognition model, Mel-Frequency Cepstral Coefficients (MFCC), Gaussian Mixture Model (GMM), Human Voice authentication.

I. INTRODUCTION

Biometric authentication uses unique biological characteristics like voice, fingerprints, face etc. to verify the identity of an individual [1]. It is easy to use as the users do not have to remember or recollect passwords. Also it is very difficult to spoof in biometric authentication.The human voice authentication is a type of biometric authentication which has huge potential since it is contactless thus making it more hygienic, convenient and cost effective. Current voice authentication systems have limitations like text dependency which makes them restricted to certain languages and also susceptible to voice replay attack [2, 3]. They are also prone to attacks which use voice impersonation [4] and AI bots. With the rise of technology, the difference between machines and humans is getting negligible. Advancements in artificial intelligence are enabling the bots to behave more human-like. Voice assistants like Google Duplex, Google Assistant, Amazon Alexa, Microsoft Cortana, Apple Siri [5] have highlighted the true potential of what they can do. In Google I/O 2018, Google pitched their new product called Duplex which can carry out a phone call conversation on behalf of human user and the other person won't even know that they are talking to a machine. It's even possible to conduct financial transactions using these assistants. Although this is a big breakthrough in AI, this poses a threat to authentication systems [6]. In the age of voice-based transactions, where entities like banks have adopted voice-based authentication, it is very necessary to improve the security of the voice signature. Important voice transactions need to be verified. Thus we have proposed and developed a human voice authentication system which verifies the entity on an intellectual level and also through voice matching. To prove that the speaker is a human, the user has to answer several logical and personal questions. The logical questions are designed such that it requires human intelligence to answer these questions. These logical questions are tested such that they cannot be answered by current AI voice assistants like google assistants, Alexa, Siri etc. Apart from this, the system is also capable of detecting an intruder using voice signatures. Ultimately, it prevents malicious entities like voice assistants from accessing important information. The proposed system also features bilingual support i.e. it supports English as well as Hindi languages so that it reaches a broad set of users.

This research has following objectives: (1) to provide a reliable human voice authentication system, (2) to implement human voice recognition model, (3) playback attack prevention and aliveness detection, and (4) provide regional language support

II. LITERATURE SURVEY

During the literature survey, we have come across some papers which describe various methodologies and algorithms for voice recognition. We have studied papers that discuss different ways to implement voice-based captcha systems. Also, we have studied papers which exploit the vulnerability in the current voice-based authentication system.

V. Srinivas et al. [7] developed a voice recognition system using Mel Frequency Cepstral Coefficients (MFCC), Gaussian Mixture Model (GMM) and Normalised Least Mean Squares (NLMS). Voice of 50 speakers was locally recorded to build database for training purposes. NLMS adaptive filter diminished the noise in the speech signal

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

which enhanced the overall performance. The rate of voice recognition was 96.96% when using the adaptive filter. Samuel S. Brown et al. [8] have used IBM Watson Developer Cloud 2 Speech to Text services to bypass recaptcha. Using selenium, the authors automated the site session. Former answers were maintained in a customized hash table. Data received from the Watson API was processed. This was then compared with the former answers. After extracting the answers from the data, the extracted digits were entered into the widget using selenium. Thereafter, the answer was verified by the Google server. It gave either a pass or a fail result for the sent answer. In the preliminary tests, the script passed the audio challenge with an accuracy of 35%.

Athira Aroon et al. [9], discussed speaker identification from an array of various voice templates and analyzed the rate of speaker recognition by extracting features like MFCC and GMM. They carried out this research on TIMIT Database. The speech signal was sampled at 16 kHz. Depending on the number of mixture components, the speaker recognition rate ranged from 77% to 98%. Haichang Gao et al. [10] proposed a novel sound based captcha that used the gap between the human voice and synthetic voice. The user read the provided sentence which was selected at random from any book. The generated audio file was scrutinized by the system to determine whether the user is a human being or not. The features of the audio file were extracted using short-term Fourier analysis. Audio was divided into small segments so that different techniques like filters and window functions could be used to boost the accuracy of the system. In addition to this, a few experiments were conducted to determine the critical threshold. Also the coefficients of the three indicators were computed through these experiments. Their study indicated that the success rate of human was approximately 97% while the success rate of attack using Microsoft speech synthesis software was only 4%. More experiments on other speech synthesis softwares were also conducted using the same critical value. The results declared that the success rate of attack software in the case of TextAloud was 48%, and DSpeech was 39%. The experiments also illustrated that on an average the sound based captcha could be dealt within 7.8 seconds. Also most human users could crack it within 14 seconds.

Kevin Bock et al. [11] introduced an automated system called unCaptcha that had the ability to crack and solve ReCaptcha's very hard auditory challenges with a very good rate of success. unCaptcha was evaluated with the help of more than 450 re-Captcha's challenges from live websites and the results proved that it could solve them with an average accuracy of 85.15% in 5.42 seconds. The system operated in four basic steps (1) Extracting the audio sample, (2) Segmenting the audio in sound bites of every digit, (3) Analysis of the sound bites, and (4) Entering the solution of the captcha. For obtaining the audio samples, browser automation software selenium was used. The downloaded audio sample was further analyzed using techniques such as speech recognition service, phonetic mapping, ensembling which generated the solution. This solution was then finally entered into the captcha to bypass the authentication.

A. Das et al. [12] have discussed a multilingual password-based authentication system using cellular phone networks. This system was implemented in two modes:(1) Terminal side authentication,(2) Server-side authentication. For speaker recognition, features were extracted from the spoken password using the MFCC method. MFCC considers speech as a set of images, called "Featurogram" (FGRAM). These FGRAMs efficiently captures the identity of the speaker based on the way a person speaks by extracting speech parameters specific to the speaker. Furthermore, the processing of these FGRAMs takes place in two ways: (1) Sinusoidal model-based Time-normalization and (2) compression by Discrete Cosine Transform (DCT) to form Compressed Feature Dynamics (CFD). Simpler nearest neighbor classifiers are used on these images for classifying a particular voice from other voice samples. The paper also compared the proposed technique with the conventional DWT technique. The authors suggested that the results of the system can be improved by having a greater number of pass phrases and training templates.

Jasmeet Kaur Hundal et al. [13] discussed some prominent feature extraction techniques for voice samples and compared these techniques on the basis of their strengths and weaknesses. The prominent feature extraction techniques discussed were MFCC, Linear Predictive Coding (LPC), Wavelet Transform and Short Time Fourier Transform (STFT). The authors observed that since MFCC feature focuses on energy, it failed to recognize the same words said with different levels of energy. STFT technique was the simplest as it took minimal computation time. It also provided better resistance to noise than MFCC. Using Wavelet Transform comprising of multilevel decomposition deeper view of the signal could be obtained. The complexity of computation for Wavelet Transform was high. Linear Predictive Coding (LPC) faster computation yet it failed to distinguish between words having the same vowel sound. Further for MFCC an accuracy of 83.75% was achieved. STFT gave an accuracy of 86.25%. Accuracy for Wavelet Transform was 87.5% and for LPC it was 85%.

Haipeng Dai et al. [14] proposed a human speech authentication scheme called SpeakPrint based on ultrasound. Traditional human voice authentication is subjected to four distinct types of attacks(1) impersonation, (2) voice synthesis,(3)voice conversion and (4)voice replay. Victoriously, SpeakPrint managed to overcome all these attacks. In contrast to the traditional speech authentication system, SpeakPrint captured the speaking style of the user rather than capturing what the user is speaking. It captured the speaking style of the user by recording mouth and vocal movement through an ultrasound signal simultaneously. SpeakPrint worked on the principle of extracting MFCC features from normal voice frequency and MMSI features from the ultrasound signal. Using a Support Vector Machine (SVM) classifier, it identified different attacks by comparing the MFCC and MMSI feature differences. Using SpeakPrint the experiments were carried out on 40 users. From the results, it was proved that replay attacks could be detected with 100% accuracy. It could also detect replay attack with lip-synching with 99.12% accuracy.

F. G. Barbosa et al. [15] discussed voice-based authentication which is implemented using the SVM algorithm. They treated voice recognition as a classification problem in which one class represents the complete features of the individual and another class represents all the other users who do not cover all these features. The methodology consisted of firstly sampling and encoding of the speech signal in Mel-Cepstral Coefficients and Coefficients of DCT.

Then two dimensional matrices depicting the DCT coefficients were generated. Further, using SVM the matrix elements were classified representing two-dimensional temporal patterns. This methodology lessened the number of parameters through SVM classifier thus reducing the computational load. For classification Radial Basis Function was used. An accuracy of greater than 90% was obtained using SVM algorithm.

S. A. A. Shah et al. [16], put forth an approach which focused on entering the PIN and also recognizing the user through their voice signatures. This enhanced the overall security access to multiple applications. Here, the speaker specific characteristics were extracted using MFCC and feature matching was done using Multi-Layer Perceptron (MLP). The speakers were classified using a feed forward back propagation network model. 50 samples per user was used for experimentation purpose. The results demonstrated that for a speaker, 41 samples were correctly recognized. Hence, false rejection was 18%. The model was also tested to detect imposter with 125 samples per user. The results showed that false acceptance was 14%.

Based on the study of these various papers, we inferred that the existing human voice authentication systems are still vulnerable to attacks. One of the key conclusions derived from the voice recognition papers is that feature extraction and voice matching modules are the crux of voice recognition. Further we inferred that MFCC works best for feature extraction and for voice recognition GMM model gives good results. Table 1 gives the overview of the background research.

TABLE I. OVERVIEW OF THE BACKGROUND RESEARCH

Ref No	Methodology	Results
[7]	MFCC and GMM are used along with an NLMS adaptive filter.	The recognition rate achieved 96.96%.
[8]	Selenium package, IBMs Watson Developer Cloud 2, A customized hash table	In the preliminary tests, the script passed the audio challenge with an accuracy of 35%.
[9]	MFCC, GMM	Recognition rate between 77-98%.
[10]	short-term Fourier analysis, filters and window functions, Indicators such as Short-term energy, Short-term average amplitude, and Short-term zero-crossing rate	Human rate of success is approximately 97% and the success rate of attack software using Microsoft is 4%, success rate of TextAloud is 48%, and of DSpeech is 39%
[11]	Novel phonetic mapping technique, selenium, speech recognition service.	Solved reCaptcha challenges from live websites. Achieved accuracy of 85.15%
[12]	MFCC, FGRAM, Sinusoidal model-based Time Normalization, compression by DCT Neighbor image classifiers	Accuracy DWT:87.5%, VQ:96.9%, MFCC GRAM:99.93%
[13]	MFCC, LPC, Wavelet transform, Short Time Fourier transform (STFT).	Accuracy and computation times are MFCC:83.75%,0.04286, STFT :86.25%,0.00198Wavelet Transform :87.5%,0.21101, LPC: 85%,0.01889
[14]	MMSI, MFCC, SVM classifier	Detects replay attacks with 100% accuracy and detects replay attack with lip-synching with 99.12% accuracy
[15]	MFCC, DCT, SVM	For classification Radial Basis Function is used. With an overall, greater than 90% accuracy was obtained.
[16]	MFCC, MLP, PCM, feed-forward backpropagation network	41 samples were recognized correctly from 50 samples of the speaker

III. METHODOLOGY

In this section the proposed secure human voice authentication system is discussed. This approach not only overcomes the limitations of the traditional authentication method but also provides a way to identify a human voice over AI voice bot. The system is divided into registration phase and authentication phase. Along with it voice recognition module is there which plays a very pivotal role in the authentication process of human voice. Fig. 1 depicts the working of the proposed system.

A. Registration Phase

During the registration phase, the user needs to enter the unique username and valid, unregistered email address. Once it is validated then user is prompted to select two personal security questions. These questions are predefined and stored in the database. User needs to record answers to these questions in his/her own voice. During the authentication phase, the identity of the valid user is confirmed based on two parameters, first whether the answer is matching or not and second, whether the voice is matching or not. After collecting the answers as a voice response, an ajax call is made to the server with audio blob as JSON data to save the recording in a directory with username as its name. The user answer in audio format is converted into the text using speech to text conversion. But authentication based on this type of methodology where every time the answer of the question is same are prone to playback attack where the spoofer can get access to recorded response or they can record the response and playback pretending as a legitimate user. To avoid the playback attack, along with answers of two personal questions, the user also needs to record a pass phrase as an added security measure. A pass phrase consists of 4 to 5 words. In addition to this, three sentences are also recorded for improving the training sample size.

On completing the registration process, these voice samples are used to train the model, which is used in the authentication phase. These responses are used to create a GMM model for the new user. Fig. 2 gives overview of the registration phase.

To provide multilingual support, translation module is implemented. The translation module allows the user to register and login in Hindi language along with English language which makes this system bilingual. For translations, the static content to be translated is wrapped around the ‘trans’ tag. The ‘makemessages’ command in Django is used to create blank messages for the content wrapped in ‘trans’ tag as well as for form validation errors. The ‘translate messages’ command in Django provides the translations for

messages and validation errors. These translations are compiled in .mo file which updates previous translations and adds new translations in templates.

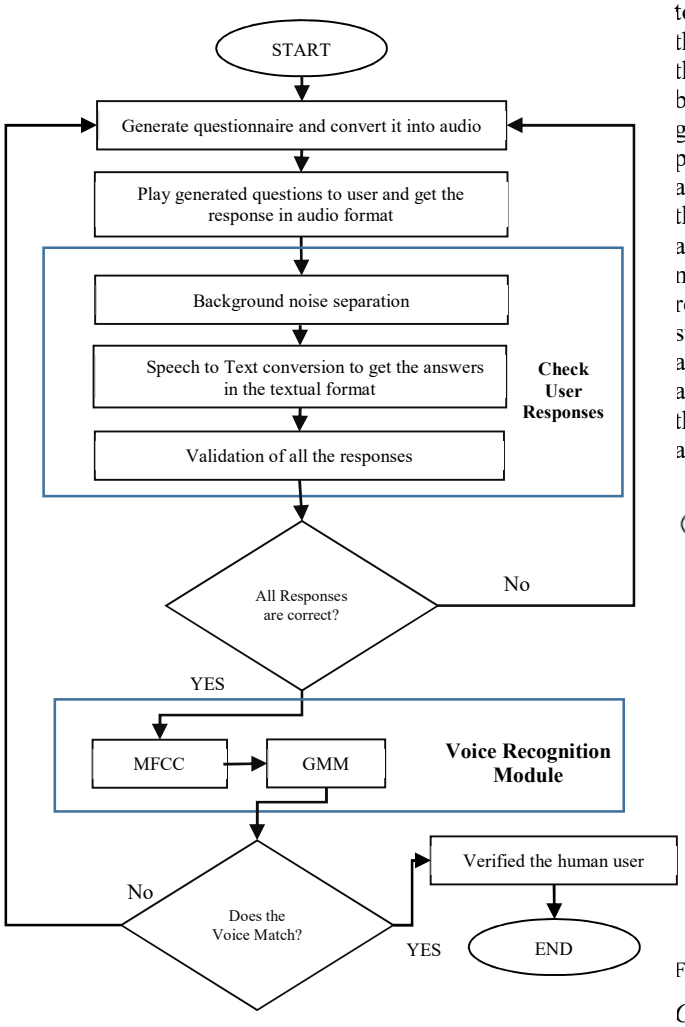

Fig. 1. System Workflow

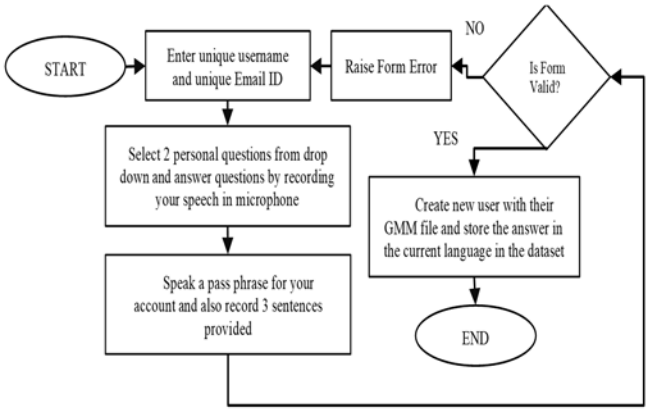

Fig. 2. Registration Process

B. *Authentication Phase*

Fig. 3 represents the authentication process. During the authentication phase, the user is prompted to answer 4 questions. From these 4 questions, 2 questions are the personal security questions that the user had selected in the registration phase and the other 2 questions are generated dynamically. This random and dynamic nature of questions restricts the access of an attacker who has access to the registered user's recorded voice. This makes the system robust against the playback attack. Additionally, the 2 dynamically generated questions require human intelligence to answer these questions. Thus AI voice bots cannot answer these questions. Also through these questions, aliveness of the user is also tested making sure that it is not an AI voice bot. Then the user speaks the pass phrase which he/she had given at the time of registration. Pass phrase is an audio password which is used to provide additional security. Then as a part of the authentication process, to submit the answers, the user need to record the answers in the form of audio for all the questions. On stopping the recording, an ajax call is made to the server with audio blob as JSON data to save the recording in a directory with username as its name. After submission, all the answers including pass phrase are tested against the trained model of each user and if 3 out of 5 answers match the voice of the user trying to authenticate then the login process is successful i.e. the user is authenticated.

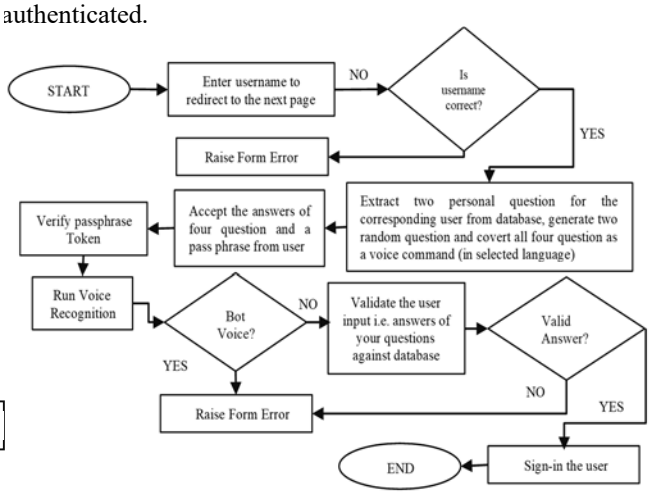

Fig. 3. Authentication Process

C. *Voice Recognition Model*

Voice recognition process is the heart of this approach. The voice recognition process is done using the GMM model. The recorded voice data is a raw audio signal which may have lots of noise. Feeding such a raw audio input to the GMM model can hamper the performance. So, the features are extracted from user's voice samples using MFCC. In order to extract the MFCC features, voice samples are converted into digital signal with a sampling frequency of 16 kHz. Next, pre-emphasis filter removes some glottal effects. Thus to emphasize high frequencies, pre-emphasis filter is used which is given by the transfer function shown in equation (1).

$$H(z) = 1 - cz^{-1} \quad (1)$$

where $0.4 \leq c \leq 0.1$, and c oversees the slope of the filter

After this, the signals are chopped into the frames using Hamming window so that there should not be any noise generation due to sudden fall in amplitude. Then each frame of window is converted into frequency domain by using Discrete Fourier Transform (DFT) given in equation (2).

$$X(k) = \sum_{n=0}^{N-1} \left(x(n) e \frac{-j2\pi nk}{N} \right) \quad (2)$$

where $0 \leq k \leq N-1$, k= total number of points involved in DFT calculation.

The human ears perceive the sound frequency in a different manner as compared to the machine. Thus, using equation (3), Mel filter is applied to map of actual frequency to human perceived frequency.

$$mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (3)$$

Then log function is applied to the output of Mel filter to imitate the human hearing capabilities. Finally Inverse DFT is applied to get the MFCC features. Along with the MFCC features, the delta MFCC features are also calculated as they help in improving the accuracy of the model. A GMM takes as input these MFCCs and derivatives of MFCCs of the training samples of the registered user. It tries to learn their distribution, which is the representation of voice of that user. For each user a GMM model representation is created. In GMM model the density probability is modeled using mixture of densities. The mixture density for the voice of a user represented as a D-dimensional feature vector (x_v) is calculated as weighted sum of N component Gaussian densities given by equation (4).

$$P(x_v) = \sum_{i=1}^{N} P_i\,(x_v) W_i \quad (4)$$

where $P_i(x_v) = \frac{1}{\left[(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}\right]} e^{\left\{-\frac{1}{2}[(x_v-\mu_i)' \sum_i^{-1}(x_v-\mu_i)]\right\}}$

and $\sum_{i=1}^{N} W_i = 1$

For testing part, the voice sample to be tested is given as input to GMM models of all users. Based on this input a log-likelihood score for this sample is calculated for each GMM model in the system using equation (5)

$$\log P(X) = \sum_{j=1}^{R} \log\left[\sum_{i=1}^{N} P_i\,(x_v) W_i\right] \quad (5)$$

where R = Total number of GMM models of all users

Thereafter the GMM model of user with the maximum log-likelihood score is selected as the winner suggesting that the voice sample being tested matches the most with this user. This process is repeated for all the replies of the user in the authentication process. The user needs to reply 5 times in total and based on these replies the authentication process is handled by the authentication module.

IV. Results

The system performance is evaluated on the basis of the operating time. Looking at the usability perspective of the system time required to process the registration and authentication should be as minimum as possible. The system was tested for time taken by a user to register and login. Fig. 4 and Fig. 5 respectively, gives the overview of operational time required to complete the registration and authentication for Hindi and English language. Here RRT stands for Time Required for Registration and ART stands for Time Required for Authentication.

Table 2 gives the comparison of processing times taken to complete registration and authentication process. The average operating times for the registration phase is 35.1 seconds and 36.4 seconds for English and Hindi language respectively, which is more or less the same. Similarly the average time for the authentication phase is 25.5 seconds and 27.6 seconds for English and Hindi language respectively. Again, this depends greatly on the user's command over the respective language.

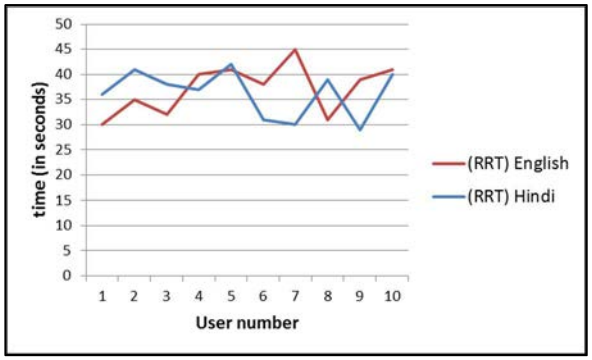

Fig. 4. Comparison of Operating times in Registration Phase

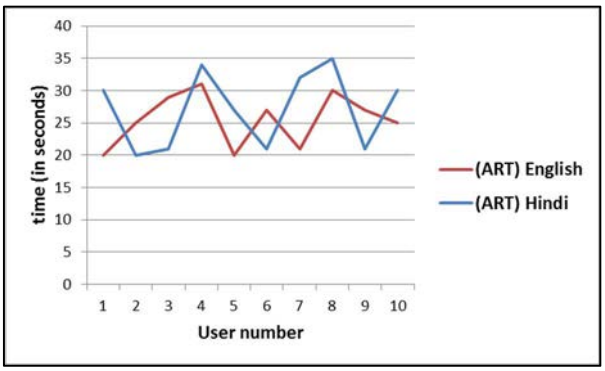

Fig. 5. Comparison of Operating times in Authentication Phase

Next, we measured voice recognition accuracy of our system, i.e. the accuracy we get from the GMM model for recognizing the user. From table 3, it is seen that the accuracy achieved for Hindi language is 95% by correctly classifying 38 voice samples out of 40 samples for 10 users and the accuracy achieved for English language is 97.8% for 183 samples of 40 users. We also tested our system against possible attacks based on voice. In table 4, we have discussed about how our system handles different attacks.

TABLE II. Comparison of Processing Times taken to Complete Registration and Authentication

Language	Phase	Average Time (in seconds)
English	Registration	35.1
	Authentication	25.5
Hindi	Registration	36.4
	Authentication	27.6

TABLE III. Comparison of Accuracies for English and Hindi Languages.

Language	Data Set	Results
English	183 samples of 40 users	179 were classified correctly. 97.8% Accurate.
Hindi	40 samples of 10 users	38 were classified correctly. 95% Accurate.

TABLE IV. POSSIBLE ATTACKS

Type	How System Tackles It
Voice Replay	Random Questions
Voice Impersonation	Pass Phrase
Voice Synthesis	Bot's voice will be recognized using its GMM

V. CONCLUSION

Thus in this paper we have proposed and have discussed the implementation of a secure human voice authentication system using MFCC and GMM. Further the system is also capable of classifying human voice from AI bot voice. The system supports two languages, viz, English and Hindi language. The system is tested for two spoken languages namely English and Hindi with 40 users and 10 users respectively. The performance of the system is measured in terms of accuracy and processing time. From the results, it is observed that the processing time mainly depends on how good the user is having the command over a language. If a user is having good command over both, English and Hindi languages, then the processing time for both languages is almost same. For voice recognition, an accuracy of 97.8% is achieved for English language and 95% accuracy is achieved for Hindi language. The system can be extended further to provide support for more languages. Also the system can be tested further against human voice generated using deep learning.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to Department of Information Technology, K. J. Somaiya College of Engineering, Mumbai, for providing necessary infrastructure and support to carry out this work. Also, we would like to appreciate the contribution of Bro. Jayesh Karvir, Bro. Jay Kanakiya, Bro. Sagar Darji, Bro. Ganesh Patra, LY B. Tech. students, Department of Information Technology, K. J. Somaiya College of Engineering, for the success of this research.

REFERENCES

[1] Debnath Bhattacharyya , Rahul Ranjan , Farkhod Alisherov A. , and Minkyu Choi," Biometric Authentication: A Review", in International Journal of u- and e- Service, Science and Technology Vol. 2, No. 3, September, 2009

[2] Z. Rui and Z. Yan, "A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification", in IEEE Access, vol. 7, pp. 5994-6009, 2019, doi: 10.1109/ACCESS.2018.2889996.

[3] Ren, Y., Fang, Z., Liu, D. et al. "Replay attack detection based on distortion by loudspeaker for voice authentication", in Multimed Tools Appl 78, 8383–8396 , 2019 https://doi.org/10.1007/s11042-018-6834-3

[4] Linghan Zhang, Sheng Tan, and Jie Yang,"Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication", In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17). Association for Computing Machinery, New York, NY, USA, 57–71. https://doi.org/10.1145/3133956.3133962

[5] Koni, Y.J., Al-Absi, M.A., Saparmammedovich, S.A., Lee, H.J,"AI-Based Voice Assistants Technology Comparison in Term of Conversational and Response Time" In: Singh, M., Kang, DK., Lee, JH., Tiwary, U.S., Singh, D., Chung, WY. (eds) Intelligent Human Computer Interaction. IHCI 2020. Lecture Notes in Computer Science(), vol 12616. Springer, Cham. https://doi.org/10.1007/978-3-030-68452-5_39

[6] Badhwar, R.,"Common Sense Security Measures for Voice-Activated Assistant Devices", In: The CISO's Next Frontier. 2021,Springer, Cham. https://doi.org/10.1007/978-3-030-75354-2_31

[7] V. Srinivas, Ch. Santhi Rani, P. Hemakunmar, "Novel Speaker Recognition System using GMM", in International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE), Pp 31-31, Vol 4, Issue 9, September 2017.

[8] Samuel S. Brown, Nicholas DiBari, Sajal Bhatia, "I Am 'Totally' Human: Bypassing the reCaptcha,", in 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2017, pp. 9-12, doi: 10.1109/SITIS.2017.13.

[9] Athira Aroon, S. B. Dhonde. "Speaker Recognition System using Gaussian Mixture Model." In International Journal of Computer Applications 130,2015, pp. 38-40.

[10] Haichang Gao, Honggang Liu, Dan Yao, Xiyang Liu, "An Audio CAPTCHA to Distinguish Humans from Computers." In Third International Symposium on Electronic Commerce and Security 2010, pp. 265-269.

[11] Kevin Bock, Daven Patel, George Hughey, Dave Levin, "unCaptcha: A Low-Resource Defeat of reCaptcha's Audio Challenge", 11th USENIX Workshop on Offensive Technologies (WOOT 17)

[12] A. Das, O. K. Manyam, M. Tapaswi and V. Taranalli, "Multilingual spoken-password based user authentication in emerging economies using cellular phone networks," 2008 IEEE Spoken Language Technology Workshop, 2008, pp. 5-8, doi: 10.1109/SLT.2008.4777826.

[13] Jasmeet Kaur Hundal, Dr. S. T. Hamde, "Some feature extraction techniques for voice based authentication system", in IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017),2017, pp. 419-421. 10.1109/ICPCSI.2017.8392328.

[14] Haipeng Dai, Wei Wang, Alex X. Liu, Kang Ling, Jiajun Sun," Speech Based Human Authentication on Smartphones", 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON),2019 , pp. 1-9.

[15] F. G. Barbosa, W. L. Santos Silva, "Automatic voice recognition system based on multiple Support Vector Machines and mel-frequency cepstral coefficients," 2015 11th International Conference on Natural Computation (ICNC), 2015, pp. 665-670, doi: 10.1109/ICNC.2015.7378069.

[16] S. A. A. Shah, A. u. Asar and S. W. Shah, "Interactive Voice Response with Pattern Recognition Based on Artificial Neural Network Approach," 2007 International Conference on Emerging Technologies, 2007, pp. 249-252, doi: 10.1109/ICET.2007.4516352.