International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 05 | May 2024 www.irjet.net p-ISSN: 2395-0072

Voice Biometrics – A New Outlook to a Traditional Problem

Nakshatra Joshi[1], Mohit Ghadi[2], Sahil Kumar[3], Swati Nandusekar[4], Sunil Patil[5]

[1]Student, Dept. of AI & DS Engineering, KJ Somaiya Institute of Technology, Maharashtra, India
[2]Student, Dept. of EXTC Engineering, KJ Somaiya Institute of Technology, Maharashtra, India
[3]Student, Dept. of EXTC Engineering, KJ Somaiya Institute of Technology, Maharashtra, India
[4]Assistant Professor, Dept. of AI & DS Engineering, KJ Somaiya Institute of Technology, Maharashtra, India
[5]Assistant Professor, Dept. of EXTC Engineering, KJ Somaiya Institute of Technology, Maharashtra, India

---

Abstract - *The traditional method of authentication using a one-time password does not provide enough security for web pages or applications that are a gateway to sensitive information/data. Sensitive data breaches have become a common thing in today's world and have caused great losses to individuals, organizations and nations as well. With advancement in technology, access to OTP has become a piece of cake and sensitive data has become vulnerable. Given the gravity of the situation, a more secure way of authentication has become the need of the hour, therefore voice-based authentication is a better successor to the traditional method and is more secure. In this paper we have collected the voice sample and processed it using the MFCC method to extract the voice print and then used Gaussian Mixture Model to find similarity between the stored voice and the live voice.*

Key Words: **Voice Authentication, MFCC, GMM, Sensitive data breaches, Voice print**

1.INTRODUCTION

In today's world, new cyber-attacks happen every single day, which has resulted in the need for robust and user-friendly authentication systems. Traditional methods of authentication include passwords, PINs and OTPs. These conventional methods are vulnerable to hacking attempts using methods such as social engineering, phishing, and brute force assaults. Voice-based authentication systems offer an excellent solution to address the drawbacks of these methods. It is a biometric method of authentication that makes use of the uniqueness of an individual's voice, as no two people have the same voice, no matter how similar they sound. By utilising distinct vocal characteristics of individuals, including intonation, rhythm, and pronunciation, voice-based authentication systems provide a highly personalized and secure means of identity verification. These systems do not require any fancy hardware, unlike other authentication systems, they only require a regular microphone which is built into almost every device like mobile phones and laptops nowadays. This is why this system holds promise in diverse applications, ranging from mobile devices and smart home assistants to enterprise-level security protocols.

We have adopted Gaussian Mixture Models (GMM) and Mel Frequency Cepstral Coefficients (MFCC) features as the pillars of our method of building the voice-based authentication system. MFCC features enable us to transform the voice input into a compact representation which encapsulates the unique characteristics of an individual's speech. These features are able to capture the subtle features that define an individual's identity. GMM models are a powerful tool for statistical modelling and classification, which are extremely well suited for modelling a person's voice. This helps us create robust templates for genuine user verification while effectively detecting unauthorized attempts. Through the integration of MFCC features and GMM models, our approach aims to build a robust voice-based authentication system.

A. Components:

In a voice-based authentication system, user verification and user registration are two of its important components. When combined, these two elements form a dependable and efficient authentication mechanism.

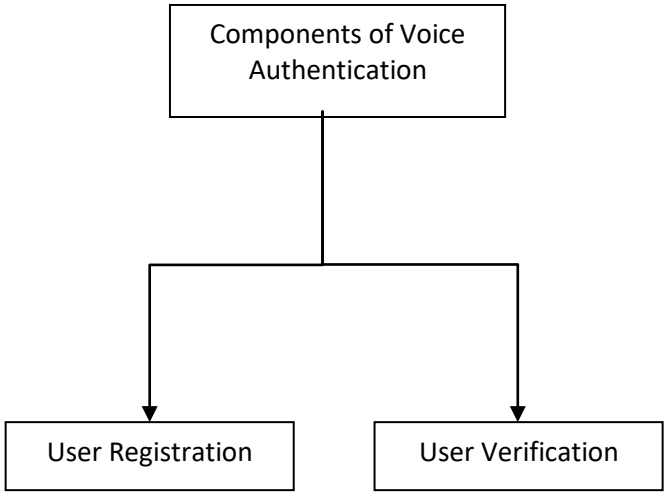

Fig -1: Components of Voice Authentication

i. User Registration

The initial step involves registering users, during which time each person's distinct voice traits are carefully recorded and entered into the system. In order to do this, speech samples must of the user are collected. From these samples, key biometric characteristics, in our case the Mel Frequency Cepstral Coefficients (MFCC) are extracted and saved as voice prints. The voice prints capture the unique vocal characteristics of every user, providing the foundation for further verification processes.

ii. User Verification

At the time of verification, a live sample of the user's voice is captured and the features extracted from this sample are then compared with those captured at the time of registration. On the basis of the result of the comparison, the decision is made as to whether the user has been authenticated or not.

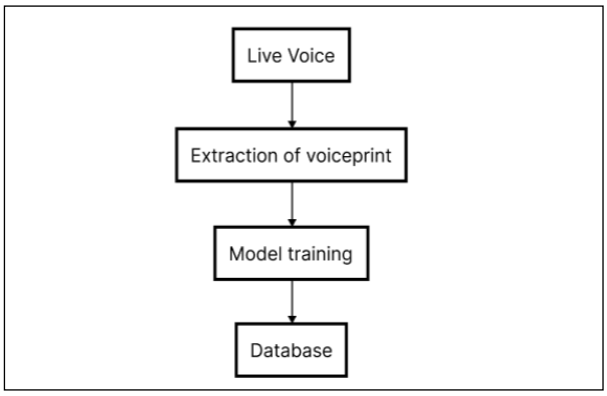

Fig -2: User Registration

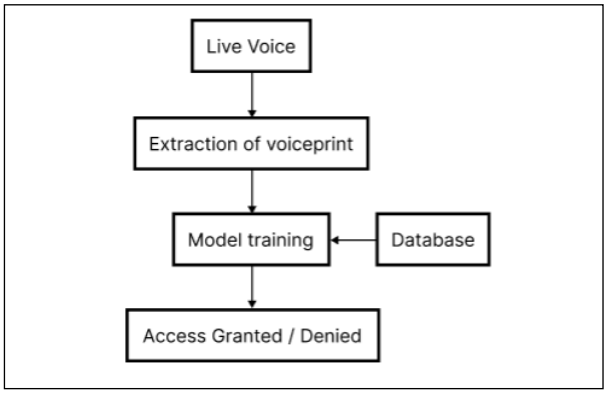

Fig -3: User Verification

B. Methods:

i. Text Dependent

During registration, this method records the user's voice and stores it in a database. Following that, the user has to say the exact same thing that was heard when you registered. It is possible to compare and contrast the two speech samples based on their respective sounds and words. This limits the statements that can be made while the voice is being authenticated. Additionally, the user needs to remember the terms they signed up with.

ii. Text Independent

By using this method, a recording of the user's voice is made at the time of registration and kept in the database. Users are required to provide a voice sample during the verification process. The only thing that is compared is the vocal qualities of the two voice samples—not the content. It is not necessary for the user to retain the registration phrase.

2. LITERATURE SURVEY

Insufficient security measures in voice assistants allow for unwanted access and create serious security risks [1]. In order to increase security, this study presents an MFCC-based authentication system that can differentiate between authorized and unauthorized users. It makes gadgets more accessible by utilizing cross-platform technologies. The framework may be used to address security issues in voice assistant systems; testing on a range of user voices validates this.

The paper explores voice authentication using an anomaly detection algorithm, making it user-friendly and equipment-efficient [2]. The method achieves high accuracy based on the "Common voice" dataset from Mozilla. Language independence is considered.

This study uses MFCC-VQ and MFCC-GMM for text-independent and text-dependent phrases to recognize speakers in Hindi speech [3]. While text-dependent recognition demonstrated better accuracy at 85.49% (MFCC-VQ) and 94.12% (MFCC-GMM), text-independent recognition achieved 77.64% (MFCC-VQ) and 86.27% (MFCC-GMM) accuracy. The study shows the usefulness of these strategies, especially improving accuracy in text-dependent situations for Hindi speech samples, using 15 speakers (10 male, 5 female) across 17 trials each.

The thesis explores the features of the speech database and the authentication system's operation for voice biometric authentication on the Android smartphone platform [4]. MFCC-based techniques are used to extract voice characteristics, and a random shuffling algorithm is employed to produce a voice encryption. GMM successfully trains voiceprint models; using five training samples, it achieves authentication success rates of 89% to 96%. The authentication time is between 210 and 320 ms, indicating excellent accuracy and promptness.

This study discusses a multimodal automated person authentication system that uses voice, face, and visual speech modalities [5]. It identifies and extracts facial characteristics from motion using morphological techniques. Acoustic characteristics (WLPCC) are built using weighted linear prediction cepstral coefficients. Auto associative neural networks (AANN) are used to represent these traits, which are paired with a weighted rule for authentication. When tested using television news data, the system achieves an EER of around 0.45% for 50 persons.

Two strategies are employed for speaker verification: one leverages a support vector machine (SVM) with the cosine kernel, while the other directly exploits cosine similarity [6]. The best results were obtained by combining linear discriminant analysis (LDA) with within-class covariance normalization (WCCN) compensatory approaches.

Gaussian Mixture Models are used in this Voice Authentication project to validate users using audio analysis [7]. Audio content analysis is a means of addressing the issues associated with comprehending complicated audio that contains a variety of information. Device performance and authentication speed are impacted by high-dimensional audio. The project leverages MPEG audio formats, which balance file size and detail, to do this. This method seeks to provide effective authentication across a number of applications, from high-security facilities to online usage, without sacrificing quality.

Table -1: Reviewed Analysis

Contributions	Extracted Features	Classifier	Accuracy
S.F. Ahmed	MFCC	HMM	90%
Berstein S.I.	MFCC	GMM	93%
Ankur Maurya	MFCC	GMM	94.12%
Xinman Zhang	MFCC	GMM	92.75%
S. Palanivel	WLPCC	AANN	90%
P. Kenny	I-Vector	SVM	91%
Vedant Baviskar	MFCC	GMM	91.2%

The paper introduces a voice authentication system based on Hidden Markov Model (HMM) voiceprints, employing text independent speaker recognition for identification without relying on specific spoken text [8].

The voice recognition system utilizes MFCC for text and language-independent voice feature extraction, followed by kNN with k=1 [9]. A new double distance method enhances accuracy to 96.97% compared to 84.85% with normal kNN.

This paper proposes a neural network-based approach for voice recognition and user identification, achieving high accuracies of 94% and 88% in text-dependent and text independent cases, respectively [10]. Language independence is not mentioned.

This paper explores voice authentication using an anomaly detection algorithm, making it user-friendly and equipment-efficient [11]. The method achieves high accuracy based on the "Common voice" dataset from Mozilla. Language independence is considered.

This paper addresses decision making in speaker verification with I-vector technique [12]. A novel threshold estimation method based on multimodality detection and abnormal score pruning improves the recognition performance of the open-set speaker recognition system. Language independence is not mentioned.

This paper focuses on improving Automatic Speech Recognition (ASR) in noisy environments by utilizing MFCC and Shifted MFCC with Vector Quantization and fuzzy modeling techniques [13]. The combined approach achieves 10-20% higher accuracy in challenging conditions like music background and noise. Language independence is not mentioned.

This paper provides a concise overview of ASR, including speaker identification, verification, and diarization [14]. It assesses the performance of current speaker recognition systems, identifies limitations, and proposes potential improvements. Additionally, it highlights some unsolved challenges in the field of speaker recognition.

This paper proposes a Hidden Markov Model (HMM) extension of i-vector approach for text -dependent speaker verification, achieving competitive results on RSR2015 and RedDots databases [15].

3. SYSTEM ANALYSIS

3.1 Analysis of Human Voice

In voice-based authentication, human voice analysis is the process of identifying and analysing distinctive vocal traits in order to confirm the identification of a person. This method is based on the knowledge that every individual has a different voice because of a variety of characteristics, such as speaking patterns, pronunciation habits, and vocal tract shape.

Vocal parameters like pitch and cadence, as well as more subtle qualities like nasal or throat resonances, are analysed by voice biometrics' technology. Voice biometric technologies can transcribe these unique characteristics into a mathematical model by generating a voiceprint. Like fingerprints or facial traits, each individual's voiceprint is distinct and may be used to accurately identify users.

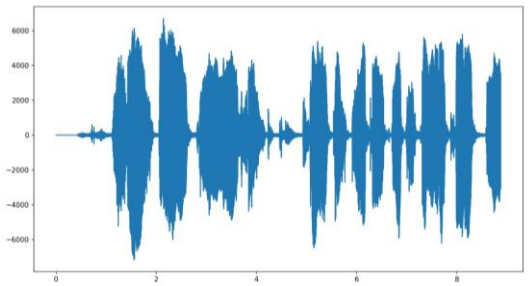

Fig -4: Graphical Analysis of human voice

The above image is a graphical representation of an audio waveform. The various peaks and troughs indicate different sound intensities and frequencies at different points in time, classification, level scale, which ranges from -600 to 600.

3.2 Feature Extraction

The feature extraction process for the voice-based authentication system involves the extraction of voice prints from the user's voice samples. Voiceprints are also known as voice biometrics, are unique representations of an individual's voice. Just as a fingerprint is unique to each person, a voiceprint is distinctive to the speaker, providing a means of identifying and verifying individuals based on their vocal characteristics.

The process of creating a voiceprint involves extracting and analyzing various features from an individual's speech, such as pitch, intonation, cadence, and the unique physiological characteristics of the vocal tract. These features are then used to create a digital representation of the individual's voice, which can be stored and compared for authentication purposes.

Voice prints are created by utilizing the Mel Frequency Cepstral Coefficients (MFCC) technique. Mel Frequency Cepstral Coefficients (MFCC) are a crucial feature widely utilized in speech and audio processing.

```
MFCC coefficients=
[[   7.72771426  118.07905679 -169.07849702 ...  197.24884535
  -36.14652057   91.63829703]
[   7.72771426  118.07905679 -169.07849702 ...  197.24884535
  -36.14652057   91.63829703]
[   7.72771426  118.07905679 -169.07849702 ...  197.24884535
  -36.14652057   91.63829703]
...
[  55.55766982   11.09440862  221.98830988 ... -164.36042252
  221.36251221 -224.86066115]
[  50.76854479   60.57798585   48.18387199 ... -114.12374441
  244.17487178 -116.74762595]
[  28.53498722   69.41668975   -6.4673424  ...  -11.88037935
  148.05837803  -93.83370462]]

Process finished with exit code 0
```

Fig -5: Extracted MFCC values

This image shows 13 MFCC coefficients from an audio file, with each set of coefficients enclosed within square brackets. The values of these numbers are both positive and negative and are displayed with decimal points.

These coefficients capture essential information about the spectral characteristics of the audio signal, emphasizing features that are significant for human speech perception while discarding less relevant information. They capture essential frequency and temporal information, thereby enhancing the system's ability to analyze and process sound data effectively.

These coefficients also have the ability to provide a dense representation of the content and their noise robustness make them a suitable choice for the feature extraction process in the proposed voice-based authentication system.

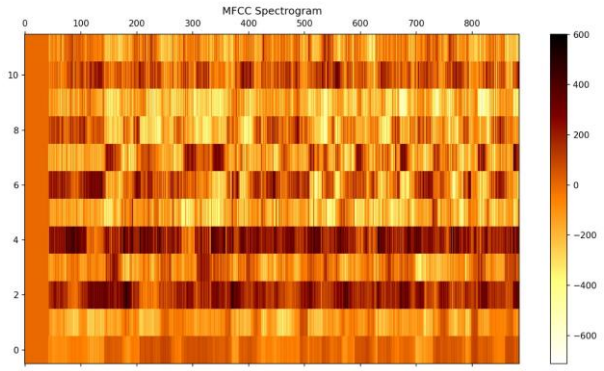

Fig -6: MFCC Spectrogram

Mel-frequency cepstral coefficients (MFCC) spectrogram is a graphical depiction of a sound signal's short-term power spectrum. The cepstral coefficients are shown by the y-axis, which runs from -6 to 10, and the time frames are represented by the x-axis, which runs from 0 to 800. Different frequencies and their intensities are represented by the spectrogram's changing colours, where darker shades denote lower energy levels and brighter colours indicate greater energy levels. A colour bar shown on the right side represents the energy.

A Mel compression spectrogram is a spectrogram that converts frequencies to the Mel scale. The Mel scale is a perceptual scale that closely approximates the human ear's nonlinear frequency response. This conversion is done to better portray sound in a form that is consistent with human auditory perceptions. To make a Mel compression spectrogram, the audio signal is first separated into short segments, and then the Short-Time Fourier Transform (STFT) is applied to each segment to produce a series of frequency spectra.

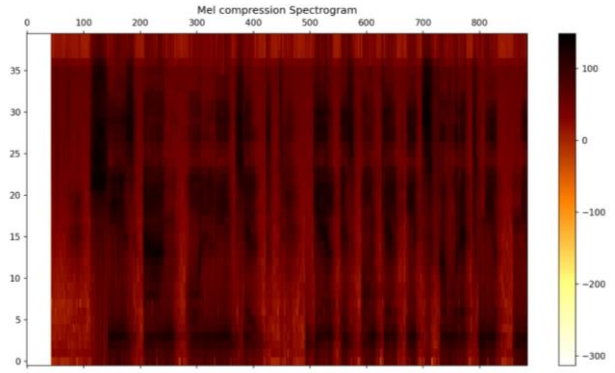

Fig -7: Mel-Compression Spectrogram

The spectra are run through the Mel filter bank, which converts the frequencies to the Mel scale. The Mel compression spectrogram can be built in a few lines of code, making it a simple and effective method for representing audio signals. It is worth noting that the Mel compression spectrogram is widely employed in speech processing and machine learning activities because it gives a more perceptually accurate representation of sound. It captures the significant aspects of the audio signal while lowering the dimensionality of the data, which can be useful for applications like audio compression and classification.

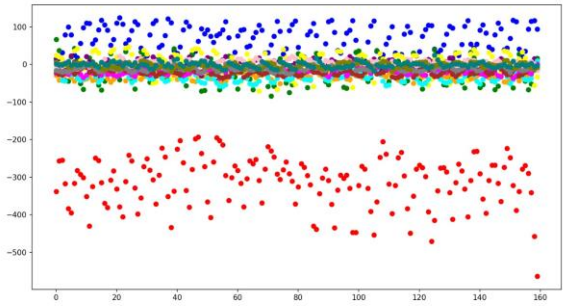

Fig -8: Graph Plot of 13 coefficients of a user

Each individual user's 13 Mel-frequency cepstral coefficients (MFCCs) are shown as a scatter plot in the above image. For a given frame of the audio stream, each dot in the scatter plot stands for a distinct MFCC. Two of the thirteen MFCCs' values specify each dot's location. The plot offers a graphic depiction of these MFCCs' distribution. Because this distribution is specific to each person, it is a very useful part of voice-based authentication system. By examining the distribution of these coefficients, a voice-based authentication system can verify a user's identity based on their voice alone.

3.3 Voice Print Comparison

The process of comparing voice prints is vital for the precise identification of individuals using their vocal characteristics. Gaussian Mixture Models (GMM) are used for this purpose, GMMs use statistical modelling to efficiently match and align voice prints. This method allows the system to recognize and validate the distinct vocal features of each individual, ensuring trustworthy authentication.

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

The GMM approach is effective due to its ability to handle variations in voice patterns. It does this by representing each voice print as a combination of multiple Gaussian distributions, each capturing a different aspect of the voice's characteristics. This makes it possible to account for the natural variability in a person's voice, such as changes due to mood or health, while still maintaining a high level of accuracy in authentication.

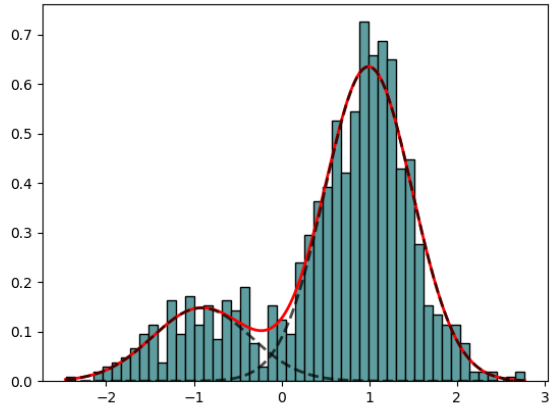

Fig -9: GMM distribution

A GMM is a statistical model that describes the probability distribution of a voice signal. It is made up of a combination of Gaussian distributions, with each Gaussian component representing a separate speaker in the system. The GMM is trained on a large dataset of speech samples from various speakers to determine the features of each speaker's voice. During the authentication process, the GMM weighs the likelihood of a given speech segment belonging to the target speaker against the possibility that it belongs to other speakers in the system. This comparison determines the speaker's authenticity.

GMM-based voice authentication systems are frequently utilized due to their ability to capture the distinct qualities of a speaker's speech. They can provide secure and dependable authentication, especially when paired with other techniques like vocal activity detection and anti-spoofing protection. It is crucial to note that GMM-based speech authentication systems have limitations and are susceptible to attacks like replay and personification. Researchers are always attempting to improve the reliability and security of voice authentication systems. Overall, GMMs are an effective tool

for voice-based identification, allowing computers to reliably validate individuals' identities based on their distinct voice features.

4. SYSTEM ARCHITECTURE

Our system's front end, which handles interface design and user interaction, is at the forefront of the architecture. An intuitive and smooth user experience is achieved through the usage of HTML, CSS, and JavaScript. Individuals can start the authentication process by submitting information via HTML forms. In addition to providing uniformity across various devices and screen sizes, CSS stylesheets improve the interface's visual attractiveness and layout. Instantaneous feedback and validation are made possible by the dynamic handling of user events using JavaScript functions.

The Flask-powered backend architecture is essential for handling authentication flows, processing user requests, and coordinating communication amongst different system components. Built on a lightweight Python web framework, Flask offers a solid basis for managing HTTP requests and creating RESTful APIs. Python scripts are used in the backend system to carry out the fundamental authentication functionality. These scripts interface with the front end to obtain user inputs and start voice verification procedures. The machine learning model for speech analysis may be seamlessly integrated with Flask thanks to its modular architecture, which facilitates effective response processing and data transfer.

The machine learning model at the heart of our system design is in charge of speech verification via the MFCC technique. Mel-Frequency Cepstral Coefficients are fundamental properties of a user's speech that are derived from voice samples and used as feature vectors for authentication.

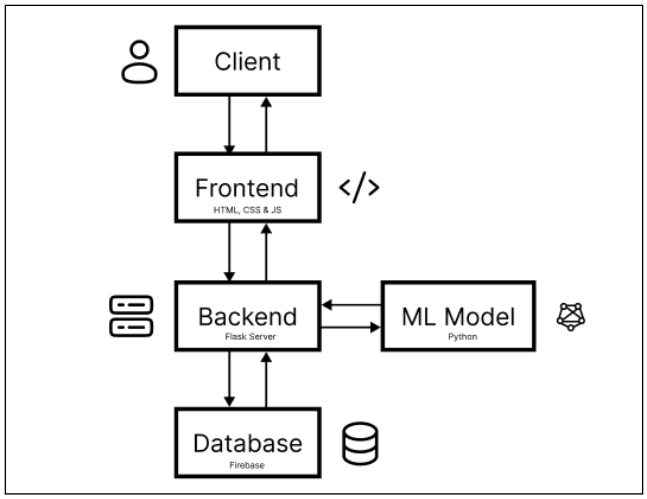

Fig -3: Architecture Design

The main programming language used for developing models is Python, which is combined with libraries like NumPy and TensorFlow to facilitate deep learning applications and numerical calculations. The MFCC technique uses threshold-based authentication to verify user IDs after calculating the logarithmic means from maxima and minima frequencies. Robust authentication performance and flexibility to a variety of speech patterns are guaranteed by the machine learning model's ongoing training and improvement.

User inputs are sent to the backend server via HTTP requests after the frontend interface initiates the authentication process. The machine learning model for voice analysis runs when the backend coordinates the authentication operation. User-provided voice samples go through preprocessing procedures, such as MFCC feature extraction. The validity of the user's voice is then ascertained by comparing the retrieved attributes with pre-established criteria. Access to the targeted service or application is allowed upon successful authentication, indicating the effectiveness of voice-based authentication as a substitute for OTP.

5. RESULTS

VOX CELEB:

Female: Within the VOX CELEB category, the system demonstrated its resilience in successfully confirming the identities of female users by achieving a remarkable 100% accuracy in authenticating female voices.

Male: The system proved to be accurate and consistent in verifying male users, as evidenced by its faultless 100% accuracy record in identifying male voices.

Table -2: Analyzed Results

VOX CELEB		NON-VOX CELEB	
Dataset	**Accuracy**	**Dataset**	**Accuracy**
Male	100%	Training	99.32%
Female	100%	Testing	60%
Samples collected by us: 70; Accuracy: 90%			
OVERALL ACCURACY: 93.85%			

NON-VOX CELEB:

The system demonstrated a high level of proficiency in learning and detecting non-voice celebrity samples during training, as seen by the training data accuracy of 99.32%. However, the testing data accuracy was only 60%. This discrepancy points to a possible area for improvement in the system's ability to authenticate non-voice celebrity samples during testing.

OVERALL ACCURACY:

The overall accuracy of 93.85% reflects the system's strong and reliable performance in authenticating users across different categories, reaffirming its effectiveness as a voice-based authentication solution.

6. CONCLUSIONS

Based on the use of Gaussian Mixture Models (GMMs) and Mel-frequency Cepstral Coefficients (MFCCs) in voice-based authentication, it can be concluded that these techniques are effective in identifying and authenticating users based on their voice. Voice recognition, divided into text-dependent and text-independent categories, relies on various speaker characteristics to differentiate one speaker from another. The successful implementation of MFCC with GMM techniques has enabled voice identification in different languages, applicable in sectors such as voice biometrics for security, voice search, voice-to-text, and voice commands to smart home devices.

The combination of MFCCs and GMMs has also been used for tasks such as emotion recognition in spontaneous speech, speaker height estimation, subglottal resonance estimation, gender recognition, and music indexing and retrieval. These applications demonstrate the versatility and effectiveness of MFCCs and GMMs in analyzing and identifying various aspects of voice data.

In conclusion, the use of GMMs and MFCCs in voice-based authentication has proven to be a robust and versatile approach, with applications ranging from speaker identification to emotion recognition and music retrieval. The success of these techniques in different sectors underscores their potential for further advancements in voice-based authentication and related fields.

REFERENCES

[1] S. F. Ahmed, R. Jaffari, S. S. Ahmed, M. Jawaid, and S. Talpur, "An MFCC-based Secure Framework for Voice Assistant Systems," in Proceedings of International Conference on Cyber Warfare and Security,2022, pp. 58-59. IEEE, 2022.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Bernstein S.I., Kolokoltsev N.K., Ermolaeva V.V. Voice Authantication. Molodoy uchenyy[Young scientist], 2018, no.25. pp. 93-94. Available at: moluch.ru/archive/211/51686/ (accessed 24 April 2019).

[3] A. Mauryaa, D. Kumara, and R.K. Agarwalb, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," in Proceedings of International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, pp. 881-885. IEEE, 2018.

[4] X. Zhang, Q. Xiong, Y. Dai, and X. Xu, "Voice Biometric Identity Authentication System Based on Android Smart Phone," in Proceedings of the 4th International Conference on Computer and Communications, 2019, pp. 1441-1443. IEEE, 2019.

[5] S. Palanivel and B. Yegnanarayana, "Multimodal person authentication using speech, face and visual speech," Speech and Vision Laboratory, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, 2007, pp. 45-55, 2008.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1435–1447, May 2007.

[7] V. Baviskar et al., "Voice-based Login System," in Proceedings of the International Research Journal of Engineering and Technology (IRJET), 2022, pp. 538-540. IEEE, 2022.

[8] R. G. M. Jayamaha et al., "VoizLock – Human Voice Authentication System using Hidden Markov Model," in Proceedings of the IEEE, 2008, pp. 330-335. IEEE, 2008.

[9] Ranny, "Voice Recognition using k Nearest Neighbor and Double Distance Method," in Proceedings of the IEEE, 2016. IEEE, 2016.

[10] M. TunÁkanat, R. Kurban, and Ş. Sağıroğlu, "Multimodal person authentication using speech, face and visual speech," in IJCI Proceedings of International Conference on Signal Processing, 2003, pp. 281-285. IEEE, 2003.

[11] A. Sidorova and K. Kogos, "Voice authentication based on the Russian-language dataset, MFCC method and the anomaly detection algorithm," in Proceedings of the Federated Conference on Computer Science and Information Systems, 2020, pp. 537-540. IEEE, 2020.

[12] C. Zhao, "Threshold Setting Method Based On Multimodality Detection In Speaker Verification System," in 2016 IEEE International Conference on Consumer Electronics-China (ICCE-China), 2016. IEEE, 2016.

[13] P. Bansal, S. A. Imam, and R. Bharti, "Speaker Recognition using MFCC, shifted MFCC with Vector Quantization and Fuzzy," in 2015 International Conference on Soft Computing Techniques and Implementations- (ICSCTI) Department of ECE, FET, MRIU, Faridabad, India, 2015, pp. 41-44. IEEE, 2015.

[14] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," IEEE Access, vol. 9, pp. 79236-79263, 2021. doi: 10.1109/ACCESS.2021.3084299.

[15] H. Zeinali, H. Sameti, and L. Burget, "HMM-Based Phrase-Independent i-Vector Extractor for Text-Dependent Speaker Verification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 7, pp. 1421-1435, July 2017.