Speaker Recognition for User Authentication

Manjunath Inamati
School of Electronics and Communication
KLE Technological University
Hubli, India
manjunathinamati8@gmail.com

Goutami Naragund
School of Electronics and Communication
KLE Technological University
Hubli, India
goutami8296@gmail.com

Shrihari Joshi
School of Electronics and Communication
KLE Technological University
Hubli, India
shrihari2003@gmail.com

Anusha AD
School of Electronics and Communication
KLE Technological University
Hubli, India
anushaad150402@gmail.com

Nirmala S R
School of Electronics and Communication
KLE Technological University
Hubli, India
Nirmala.s@kletech.ac.in

***Abstract*— In the world which requires high security standards we have focused on creating a smart system to make sure only the right person gets access. We use feature extracting method MFCC and DTW to recognize a person's voice, even in noisy situations. This system is like a secret password, but instead of typing, your voice is the key.We've added some clever filters to pick out what makes each voice unique, making our system great at telling one person's voice from another. Imagine it like a super secure room that only opens for one person. If someone else tries to get in, the system says, "Login Failed," keeping everything safe. This is especially useful in places where super tight security is essential. It's our way of making sure only the right people have access, keeping everything secure and sound.**

Keywords—MFCC, DTW, DCT, Mel frequency

I. INTRODUCTION

Human voice carries unique identity for everyone. It has been observed that there is significant variance between every person's speech signals. Consequently, speech signal is also as unique as human fingerprint. To build a user identification system, biometrics speaker recognition technique is inevitable. Biometrics speaker recognition technique automatically recognize the speech of a person based on the features exists in his/her voice signal. There are several available techniques of speech recognition such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Neutral network etc. However, in this paper a biometrics speaker identification system is proposed based on Mel-Scale Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) techniques. Every word of a speech is utilizing the phonetic mixture of a set of vowels, semivowels and consonant discourse sound units. Mel-Scale Frequency Cepstral Coefficients (MFCC) uses spectral based parameter for recognition process. Basically, MFCC is a coefficient. Based on human auditory systems, it represents audio. Using MFCC, Mel scale frequencies of speech Cepstral can be extracted from a voice signal which is equivalent to the MFCC coefficients. With the aid of Dynamic Time Warping (DTW), we can measure how close two voice signals are. Specific user is recognized by analyzing the outputs from the DTW. In this experiment all the experimental results & simulations are performed in MATLAB.

II. LITERATURE SURVEY

The paper [1] project focuses on developing a Voice Recognition system using digital signal processing. The system analyzes input voice features, compares them with prerecorded signals in a database, and displays relevant information. The goal is to enhance security for items like lockers and phones by detecting unauthorized access. The paper [2]explores using nonlinear signal processing for detecting voice disorders. It introduces a modified Grassberger-Proccacia algorithm and surrogate data analysis to quantify chaos in voice signals. The study achieves 95% accuracy in differentiating between healthy and disordered voices. The authors emphasize the potential of these methods in diagnosing and treating voice disorders, highlighting the advantages of nonlinear dynamic approaches over traditional voice analysis methods. The results suggest that surrogate data analysis, particularly using normalized mean sigma deviation, is a promising tool for classifying voice disorders. The paper [3] delves into the role of Voice Activity Detection (VAD) algorithms in forensic speaker verification systems, particularly in handling noisy phone tapping's . The study evaluates the performance of two widely used VAD algorithms under different noise conditions. The authors emphasize the impact of VAD on speaker verification systems and propose a dynamic approach, suggesting that the choice of VAD algorithm should adapt to varying noiselevels. The results indicate that a one-size-fits-all approach toVAD may not be optimal, and a dynamically selected VAD could enhance biometric identification performance. The study uses the Equal Error Rate (EER) as a parameter to assess system performance.

The paper [5] explores speaker recognition for digital forensic audio analysis using the Learning Vector Quantization (LVQ) method. Authored by Danny Bastian Manurung, Burhanuddin Dirgantoro, and Casi Setianingsih from Telkom University, Indonesia, the study addresses the classification of audio samples in forensic evidence to identify suspects based on their voices. The authors develop a prototype application employing the Mel-frequency Cepstral Coefficients (MFCC) method for sound feature extraction and LVQ as the classification method. The training process involves labeling sound features with knownspeakers and updating weights through iterations. The experimental results demonstrate promising accuracy, with the LVQ method achieving a recognition accuracy of 73.33%for the same sentence and 46.67% for different sentences. The study contributes insights into applying machine learningtechniques for speaker recognition in digital forensic scenarios. The paper [6] introduces a novel approach to text- dependent speaker verification, presenting State-GMM models for enhanced accuracy. Specifically, the State-GMM-JFA system exhibits a substantial 44% reduction in Equal Error Rate (EER) under text-prompted conditions. Experimental evaluations conducted on the Wells Fargo dataset showcase the superior performance of the proposed models compared to baseline GMM and HMM systems. Notably, the study underscores the significance of mitigating overtraining effects, especially in specific scenarios, to achieve optimal results.

The authors in [7] addressed the need for robust biometric authentication for Linux and proposed a three-level authentication system: text-based password verification, speech verification, and speaker verification. They highlighted the challenge of the 'record and replay problem' in voice authentication and proposed the use of a random passphrase to address this issue. The system utilized modules for speech and speaker authentication, employing tools like Pocket Sphinx for speech-to-text conversion and Mel Frequency Cepstral Coefficients (MFCC) for speaker verification. Additionally, the authors introduced a differential MFCC approach and applied K-means clustering for data processing to enhance the accuracy of speaker identification. Results from the study showed that the speech authentication achieved an accuracy of 80-85%, with the efficiency affected by noise and homophonic words. Speaker authentication, especially with differential MFCC, showed a significant improvement in authentication rates for male users, reducing authentication failure. The authors in [8] propose a feature fusion approach to enhance speaker identification, especially in noisy conditions. They introduce voice biometrics as a secure means of user identification and address the challenges in speaker identification, particularly in mismatched environments. They suggest combining spectral and time-domain features to improve performance in noisy conditions. The methodology involves a feedforward neural network and a feature vector that combines various features, including MFCC, Pitch, LPCC, ZCR, entropy, and energy. Experiments were conducted using the LibriSpeech dataset, with results showing that the proposed feature vector outperforms other combinations, particularly in the presence of noise. The paper concludes that this feature vector, coupled with deep neural networks, offers promising performance in speaker identification, emphasizing the importance of feature extraction methods in voice biometric systems' effectiveness. This approach has potential applications in various areas, including IoT devices. Overall, the paper aims to enhance the accuracy and robustness of speaker identification in noisy environments. The methodology in [9] involves recording speech signals containing background noise, which is then filtered and normalized. The energy level and zero crossing extraction methods are applied for voiced/unvoiced area detection. Voiced signals are further segmented, and the MFCC method is applied to these segments. The extracted MFCC data serve as inputs for training a neural network.

The preprocessing of the speech signal includes filtering, normalization, energy level detection, and zero crossing extraction. The energy level and zero crossing methods are employed to identify voiced and unvoiced speech areas. The MFCC feature extraction process involves several steps, such as pre-emphasis, framing, windowing, FFT spectrum computation, Mel-spectrum extraction, and Mel-cepstrum extraction. The authors use a backpropagation neural network for training, with 150 input neurons corresponding to the MFCC values. The network achieves a classification accuracy of 98.9%, demonstrating the effectiveness of the proposed approach for speech recognition.

The paper [10] discusses the use of multiple instances or units of the same body trait, such as left and right irises, in biometric systems .It also mentions the use of a single sensor to acquire multiple samples of the same biometric, such as two face views, to account for intra-user variations in the trait. The paper highlights the use of multiple sensors, such as visual and infrared face cameras, to image a single biometric trait .It discusses the processing of the same biometric data using different feature extraction and matching algorithms. The paper mentions the use of evidence provided by different body traits, such as face and finger, for establishing identity. It suggests the possibility of designing a hybrid multibiometric system by integrating a subset of the mentioned scenarios via fusion.

III. METHODOLOGY

The speaker identification process involves several digital steps before using techniques like MFCC and DTW. Initially, users' voice recordings are stored in a database. The identification task is about deciding if a voice sample matches an asserted individual. When a user is to be identified, their voice record is stored. During verification, the recorded speech is processed along with the database voice record through filtering. The processed data matrix is then used with MFCC to extract unique features from the voice signal. Finally, DTW compares this processed voice signal with the stored database voice samples to determine a match.

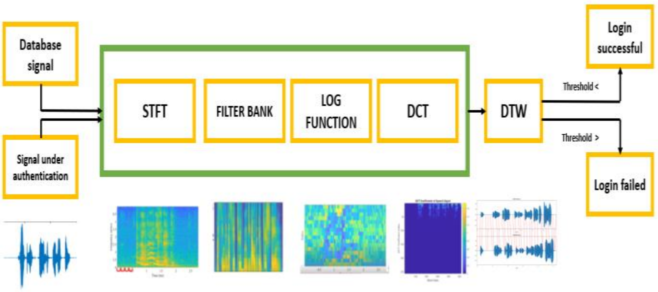


Fig.1.Functional block diagram

The Whole Block Function is a process used in voice signal processing, specifically MFCC (Mel Frequency Cepstral Coefficients). It starts with a voice signal input and goes through several stages including STFT (Short Time Fourier Transform), Filter bank, Log Function, DCT (Discrete Cosine Transform), and DTW (Dynamic Time Warping) to produce a result that can be used for various applications like speech recognition or speaker identification. The image you sent is a simplified diagram of the MFCC process for voice signal processing. It consists of seven blocks connected linearly. The first block labeled “Voice Signal” indicates the input of the process. The second block is “STFT” which stands for Short Time Fourier Transform, a method to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. The third block is “Filter bank”, which filters the frequency components obtained from STFT.

Next is the “LOG FUNCTION” block that applies logarithm to emphasize certain features. Followed by “DCT” or Discrete Cosine Transform which helps in decorrelating the filter bank coefficients and converting them into cepstral domain. Then comes “DTW” or Dynamic Time Warping that aligns sequences in time domain ensuring optimal matching between signals. Finally, there’s “RESULT” indicating output after processing through all these stages.

A. *MFCC*

a. STFT

STFT, or Short-Time Fourier Transform, is a technique used in voice biometrics to look at how the frequency of a voice signal changes over short periods. It's like taking snapshots of the voice at different moments. This helps because when people talk, the pitch and sound vary. Instead of looking at the entire voice recording at once, STFT breaks it into small pieces called frames, kind of like short clips. This method, called windowing, allows us to focus on specific parts of the voice. For each of these short clips, STFT checks the frequency components by using a sliding window, like a small piece moving through the recording, looking at one part at a time. This way, we get a better understanding of how the voice's frequency changes throughout the recording.

b. Triangular Filter bank:

After acquiring the spectrogram through STFT, the subsequent step typically includes the application of a Triangular Filter bank. This filter bank consists of overlapping triangular filters designed to imitate the frequency response of the human ear. Human hearing is particularly attuned to specific frequency ranges, and these triangular filters are adept at capturing such sensitivity. In practice, each triangular filter is employed on a designated frequency range within the spectrogram. The output of the filter bank signifies the energy or magnitude present in those specific frequency bands.

c. Log function :

The introduction of the Mel Frequency scale serves a crucial purpose in aligning with human perception of pitch. The Mel scale is a perceptual representation of pitches that mirrors how humans hear various frequencies. To bridge the gap between the frequencies obtained from the triangular filter bank and the way humans perceive pitch, a mathematical formula is employed for conversion. This transformation utilizes the following formula to convert a frequency (Hertz) to its corresponding Mel frequency. Here, $M(f)$ represents the Mel frequency, and f denotes the frequency in Hertz. In simpler terms, this formula translates the linear frequency scale (Hertz) into the non-linear Mel scale. The logarithmic operation is instrumental in capturing the non-linear way in which humans perceive pitch. Consequently, the resulting Mel frequencies offer a more accurate representation of how our auditory system interprets different pitches.

$$\text{Mel} = 1000 \log2 (1+f)$$

d. DCT (Discrete Cosine Transform)

In the realm of voice biometrics, the DCT (Discrete Cosine Transform) block serves a critical purpose. Its primary goal is to streamline the information contained in the Mel Frequency Cepstral Coefficients (MFCCs) by minimizing redundancy and accentuating key features. This transformation involves converting the logarithmic filter bank energies into a set of coefficients, placing emphasis on the most significant information while simultaneously reducing the dimensionality of the feature vectors. As a result, the output of the DCT block represents the final set of MFCCs, which are widely utilized as distinctive features for tasks such as speaker identification or verification. In essence, the DCT block plays a pivotal role in refining and condensing the essential voice characteristics for more effective and efficient biometric analysis.The MFCC block combines these four steps (a, b, c, d) to efficiently extract and represent key features from the voice signal for biometric applications.

e. DynamicTimeWarping (DTW)

Dynamic Time Warping (DTW) is a technique in voice biometrics that helps compare and line up sequences of sound features. It's especially handy when the timing between signals is different. DTW tackles the challenge of comparing signals with various durations or timing issues. Its main goal is to find the best alignment between the frames of two signals, making the comparison more accurate. Think of it like a smart algorithm that carefully checks how similar the frames of two sequences are, even if their timing is a bit off. DTW can flexibly stretch or compress the time axis, accommodating changes in speech speed. This makes it super useful for tasks like checking if someone's voice matches their identity or recognizing speech. Essentially, DTW makes voice signal comparisons more adaptable to timing differences, improving the accuracy of biometric systems when recognizing and verifying speakers.

B. FILTER :

In the domain of FIR filters, the design process often incorporates various windowing techniques. Window functions, such as the Hamming, Blackman, Barlett, Hanning, and Rectangular windows, are applied to the impulse response of the filter. This helps control the trade-off between the main lobe width and the level of side lobes in the frequency domain. The Hamming window, for instance, minimizes side lobe levels but widens the main lobe, whereas the Blackman window achieves a narrower main lobe with higher side lobe suppression. These windowing techniques play a pivotal role in shaping the frequency response of FIR filters and managing the transition between passband and stopband. They effectively reduce unwanted artifacts such as spectral leakage, enhancing the filter's performance in various applications. Implementation of all the windowing technique for the application has been tested.

Other IIR filters like Chebyshev and Butterworth have also been implemented .Chebyshev Type-1 filter gives ripple at the passband .The Butterworth filter is specifically designed to achieve a maximally flat frequency response in the passband. This is accomplished by minimizing the rate of change of the filter's magnitude response, resulting in a smooth transition between the passband and stopband. Butterworth filters, including an 8th order variant, are often favored in applications where maintaining a flat frequency response in the passband is essential, and the presence of ripples or oscillations in the frequency domain is undesirable.

Here we have used an 8th order Butterworth IIR filter has been selected with a passband frequency range of 300 Hz to 3000 Hz to address distortion in a given input signal. The filter's order indicates a desire for a steeper roll-off, but the potential introduction of phase distortion is a consideration. By setting the frequency range, the filter permits signals within 300 Hz to 3000 Hz to pass through while attenuating frequencies outside this range. The effectiveness of distortion removal is crucial, and careful verification is necessary to ensure that the filter achieves its goal without introducing undesirable noise.This comprehensive evaluation is essential, considering the trade-offs and limitations associated with IIR filters in terms of phase distortion and stability concerns, based on the specific requirements of the application.

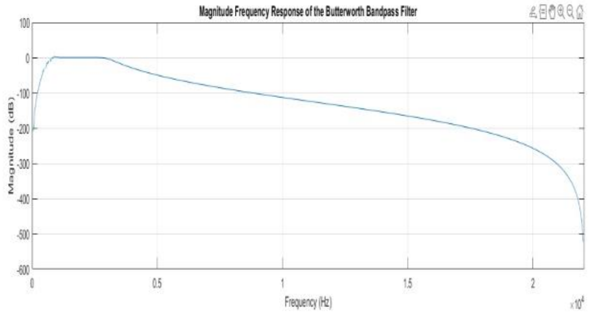
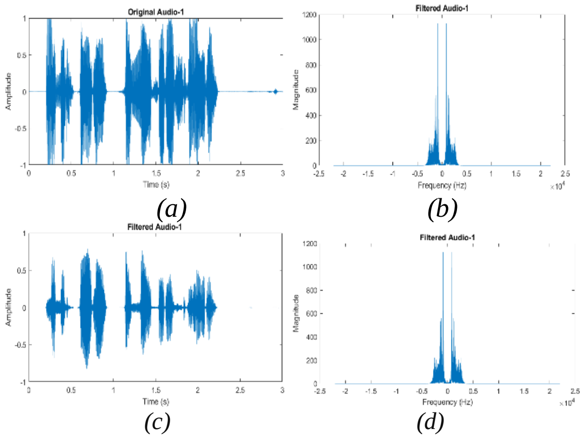


Figure 1.Magnitude frequency response of butterworth filter

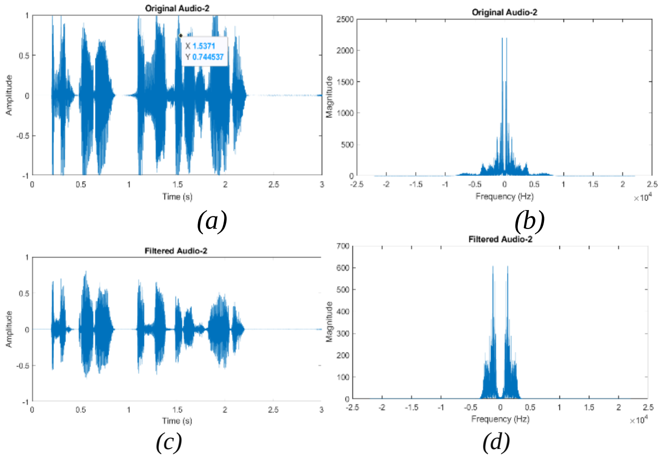
IV. RESULT



(a) (b)

(c) (d)

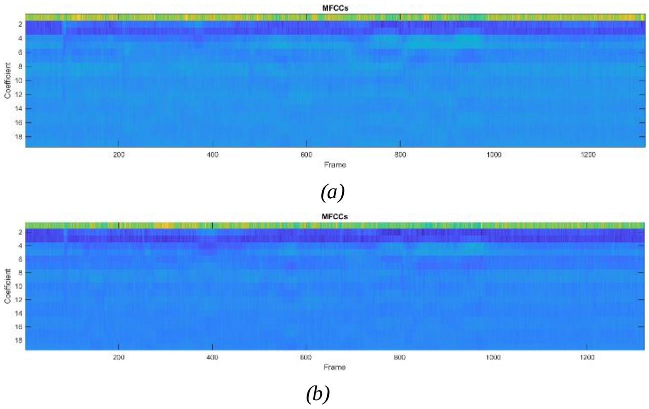
Figure2. (a)original time domain (b)original frequency domain (c)Filtered time domain (d) Filtered frequency domain for user S_1



(a) (b)

(c) (d)

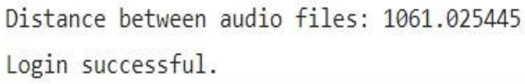
Figure3. (a)original time domain (b)original frequency domain (c)Filtered time domain (d) Filtered frequency domain for user S_2



(a)

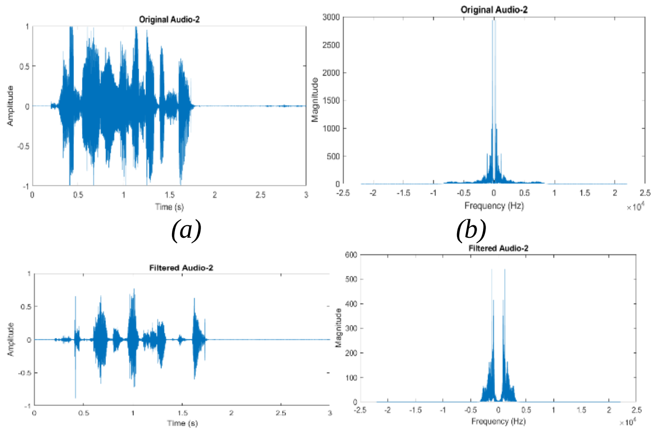
(b)

Figure 4. MFCC plot for (a) S_1 and (b) S_2



```
Distance between audio files: 1061.025445
Login successful.
```

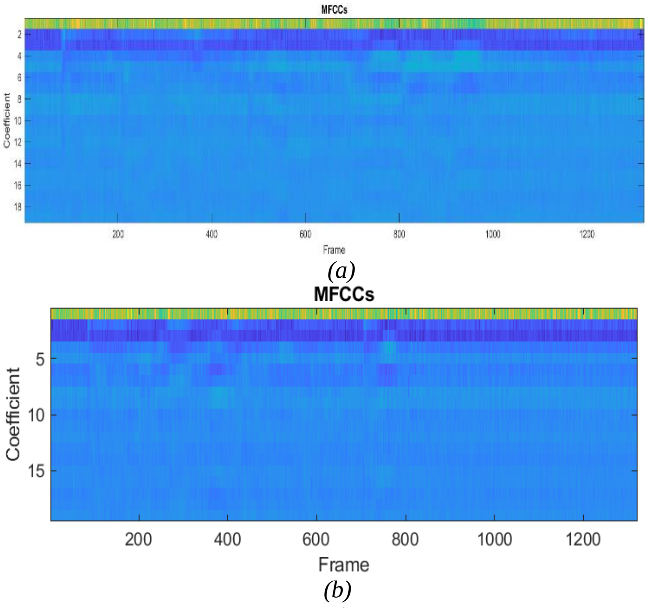
Figure 5. S_1 and S_2 comparison results



(a) (b)

(c) (d)

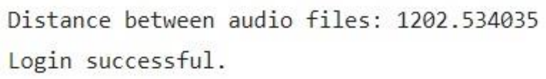
Figure6. (a)original time domain (b)original frequency domain (c)Filtered time domain (d) Filtered frequency domain for user A_2



(a)

(b)

Figure 7. MFCC plot for (a)S_1 and (b) A_2



```
Distance between audio files: 1202.534035
Login successful.
```

Figure 8. S_1 and A_2 comparison result

***Table 1**. The comparative analysis of user S-1 with multiple instances of their own audio*

USER	**DISTANCE**
S_1 with S_1	0
S_1 with S_2	1061.025
S_1 with S_3	1128.6036
S_1 with S_4	1127.8771

Here in the following Table.1 and Table.2 the S_1 stands for the user audio stored in the database. S_2 ,S_3 ,S_4.. stands for same user's audio in different circumstances. The audio of other persons are stored with different names representing them as not a valid user. Based on above table we have decided to take highest value as its threshold.

Table. 1 offers an analysis of user S_1 uttering the same sentence with pitch variations, assessing the model's ability to differentiate between users. The outcomes of the Dynamic Time Warping (DTW) analysis reveal subtle distinctions in the output, indicating that the model correctly identifies the user even amidst pitch variations. This underscores the model's effectiveness in maintaining user recognition accuracy across different pitch conditions.

***Table 2 .**Comparison table with multiple users*

USERS	**DISTANCE**
S_1 with S_1	0
S_1 with S_2	1061.025
S_1 with A_2	1202.534
S_1 with G_2	1244.854
S_1 with M_2	1244.218

Observing the comparison Table.2, a suitable cut-off distance of has been chosen for authenticating users. In this context, a single user's audio signals are compared with multiple users .Notably, the distance between audio signals of user S_1 is significantly larger when compared to their own voice. By employing Dynamic Time Warping (DTW), we can assert that the audio signals of the same person uttering the same sentence multiple times exhibit minimal differences in comparison to the audio signals of other users not present in the database.

It's worth noting that factors such as amplitude, pitch, and voice intensity are crucial considerations in this comparison. These acoustic features contribute to the uniqueness of an individual's voice signature, and DTW serves as a valuable tool in capturing the temporal variations and nuances in the audio signals. The chosen cut-off distance of indicates a threshold beyond which the differences between audio signals are deemed significant for authentication purposes. This approach provides a robust means of verifying users based on their distinct voice patterns while accommodating variations that may naturally occur in speech.

V. CONCLUSION

The primary goal of this research was to design a user identification system using MFCC (Mel-Frequency Cepstral Coefficients) and DTW (Dynamic Time Warping) techniques, focusing specifically on biometric voice identification within a voice database, with an emphasis on a single user's voice. The methodology involved applying both MFCC and DTW algorithms to distinguish a user based on voice signals. These algorithms were employed to differentiate between voice signals of the same speaker and those of different speakers. When the speech signals were identical, the algorithms produced a difference of zero; otherwise, variations in the spectrum occurred due to the dissonance of different voices. Before applying MFCC and DTW, voice samples underwent pre-processing to eliminate noise and extract relevant voice components from the overall recording. MFCC was then used for feature extraction, and DTW was employed to identify differences between two voice signals, leading to successful user identification.

The analysis of experimental data in MATLAB revealed that the system's results were not efficient.To improve efficiency and accuracy, the recommendation is to integrate machine learning models. Utilizing machine learning would optimize the computation process, enabling a more precise comparison of each user with all others in the database. The proposed system could be extended to handle a variable number of users (denoted by 'n') by defining 'n' structures of feature vectors containing multidimensional spaces. This extended version could be treated as a separate project, offering scalability and flexibility to accommodate a larger user base. The integration of machine learning models is advised to enhance the overall efficiency and accuracy of the user identification system.

References

[1] Voice Recognition System on MATLAB for Beginners Using Euclidean Distance Akanksha Singh Thakur1 , Namrata Sahayam2 1ME IV Sem, 2Asst Professor

[2] P. Bratoszewski, G. Szwoch and A. Czyżewski, "Comparison of acoustic and visual voice activity detection for noisy speech recognition," 2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 2016, pp. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] A. Taherkhani, S. A. Seyyedsalehi, A. Mohammadi and M. H. Moradi, "Nonlinear Signal Processing for Voice Disorder Detection by Using Modified GP Algorithm and Surrogate Data Analysis," 2007 IEEE International Symposium on Signal Processing and Information Technology, Giza, Egypt, 2007, pp. 1171-1175, doi: 10.1109/ISSPIT.2007.4458076.

[4] [2]T. Muttaqi, S. H. Mousavinezhad and S. Mahamud, "User Identification System Using Biometrics Speaker Recognition by MFCC and DTW Along with Signal Processing Package," 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 2018, pp. 0079-0083, doi: 10.1109/EIT.2018.8500256.

[5] D. B. Manurung, B. Dirgantoro and C. Setianingsih, "Speaker Recognition For Digital Forensic Audio Analysis Using Learning Vector Quantization Method," 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS), Bali, Indonesia, 2018, pp. 221-226, doi: 10.1109/IOTAIS.2018.8600852.

[6] S. Novoselov, T. Pekhovsky, A. Shulipa and A. Sholokhov, "Text-dependent GMM-JFA system for password based speaker verification," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 729-737, doi: 10.1109/ICASSP.2014.6853692.

[7] S. Singh and M. Yamini, "Voice based login authentication for Linux," 2013 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, India, 2013, pp. 619-624, doi: 10.1109/ICRTIT.2013.6844272.

[8] A. Das, L. P. Roy and S. Kumar Das, "Effectiveness of Feature Collaboration in Speaker Identification for Voice Biometrics," 2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Kolkata, India, 2023, pp. 1-4, doi: 10.1109/ICCECE51049.2023.10085318.

[9] C. K. On, P. M. Pandiyan, S. Yaacob and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," 2006 International Conference on Computing & Informatics, Kuala Lumpur, Malaysia, 2006, pp. 1-5, doi: 10.1109/ICOCI.2006.5276486.

[10] F. G. Barbosa and W. L. Santos Silva, "Multiple Support Vector Machines and MFCCs application on voice based biometric authentication systems," 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015, pp. 712-716, doi: 10.1109/ICDSP.2015.7251968.