Multi-Modal Biometric Access System Using Face, Emotion, and Voice Recognition

Dr. S. V. Rama Rao[1], M. Sankar[2], L. Rishi[3], M. Vinay Kumar[4], K. Guna Veeranna[5], Md. Alfan Babu[6]
[1] *Associate Professor, Dept of Electronics and Communication Engineering, NRI Institute of Technology, Pothavarappadu (V), Agiripalli (M), Eluru (Dt)521212 Andhra Pradesh*
[2,3,4,5,6] *Student, Dept of Electronics and Communication Engineering, NRI Institute of Technology, Pothavarappadu (V), Agiripalli (M), Eluru (Dt)521212, Andhra Pradesh*

Abstract*: The Real-Time Emotion-Based Access Control System has a multi-level authentication system with secure access based on facial recognition, voice, and emotional authentication. It records user's facial photos, voice samples, and verbal passwords during registration through the use of Automatic Speech Recognition (ASR) technology driven by Whisper, FaceNet for facial recognition, Mel-Frequency Cepstral Coefficients (MFCC) in combination with Support Vector Machine (SVM) for voice recognition, and Multi-task Cascaded Convolutional Networks (MTCNN) for facial detection. Further, the system stores the user's average facial emotion for better verification. On authentication, it verifies facial identity, voice identity, verbal password, and emotional match, hence enhancing security controls. Through AI-driven authentication and emotion-based verification, the system enhances access control and reduces unauthorized entries.*

Keywords: Automatic Speech Recognition, Emotion Analysis, Facial Recognition, Mel-Frequency Cepstral Coefficients, Real-Time Authentication, Voice Recognition.

I. INTRODUCTION

In the evolving landscape of information security and identity management, traditional approaches for access control—such as passwords, PINs(Personal Identification Number), physical keys, and access cards—are rapidly becoming obsolete due to increasing threats like phishing attacks, brute-force cracking, social engineering, and credential leakage [4]. These conventional techniques often suffer from vulnerabilities including theft, duplication, sharing, or simple forgetfulness. To counteract these weaknesses and provide enhanced security with convenience, biometric authentication has emerged as a dominant paradigm, offering unique, non-transferable, and user-specific characteristics for identity verification.

Biometrics refers to automated recognition of individuals based on their behavioral or physiological traits such as fingerprint, palmprint, facial structure, iris pattern, voice, or even emotional state [1], [3]. Biometric systems operate in two fundamental modes: enrolment and verification. During enrolment, a biometric template of the user is stored securely in the system database. During verification, the live biometric sample is captured and compared against the stored template to validate identity. While unimodal biometric systems, which rely on a single biometric trait, have gained popularity, they often struggle with challenges like noisy data acquisition, limited distinctiveness, intra-class variations, non-universality, and vulnerability to spoofing attacks [5]. These factors contribute to increased False Acceptance Rate (FAR) and False Rejection Rate (FRR), degrading the reliability and robustness of the overall system. The system's FAR, defined as the probability that an unauthorized person is incorrectly granted access, is calculated as

$$\mathrm{FAR}(\%) = \left(\frac{\mathrm{FA}}{\mathrm{N}}\right) \times 100$$

Conversely, FRR is the probability that an authorized user is incorrectly denied access, calculated as

$$\mathrm{FRR}(\%) = \left(\frac{\mathrm{FR}}{\mathrm{N}}\right) \times 100$$

Due to the limitations of unimodal systems in terms of accuracy, security, and scalability, the use of multimodal biometric systems is increasingly advocated. These systems integrate multiple biometric traits—such as face, voice, fingerprint, palmprint, or emotion—to provide complementary evidence of identity [5], [6]. Multimodal systems are inherently more robust against spoofing and

environmental distortions. For instance, replicating both a person's facial features and voice characteristics simultaneously is significantly more challenging than forging a single trait. This fusion of data sources results in lower error rates, higher accuracy, and improved user confidence.

This paper proposes a novel multimodal biometric secure access system that utilizes facial recognition, voice identification, facial emotion analysis, and speech-based password recognition to offer a robust multi-factor authentication mechanism. The system is developed with a focus on real-time performance, user convenience, and high-security assurance. Unlike conventional systems, it integrates emotional context as part of the authentication layer, ensuring that only a specific emotional state (captured during the lock phase) can unlock the system, thus adding an innovative emotional biometrics factor.

The facial recognition module employs the well-established FaceNet model, which maps facial images into a Euclidean embedding space for identity verification with high precision [7]. For voice recognition, the system extracts Mel-Frequency Cepstral Coefficients (MFCCs) from audio signals and classifies them using a 1D CNN trained on a user-specific dataset [2], [5]. Emotion detection is implemented using a Convolutional Neural Network (CNN) trained on FER2013 to classify emotions such as happy, sad, angry, etc., enhancing the context-awareness of the system [3]. Finally, the system leverages the Whisper model for speech-to-text conversion and validates the spoken password against stored credentials.

The performance of each module was evaluated using metrics like accuracy, FAR, and FRR. Experimental results revealed that the multimodal system significantly outperformed unimodal systems in noisy environments and under partial spoof attempts. The integration of deep learning techniques like Deep Neural Networks (DNN), Backpropagation, and Gabor Filters further enhanced the system's accuracy in minutiae detection and feature extraction [6].

II. PROPOSED MODEL

The proposed system introduces a Multimodal Biometric Authentication Framework that integrates facial recognition, voice-based speaker identification, and speech-to-text password matching, with an added layer of emotion-based validation.

A. Face Image Acquisition and Emotion Detection

During the registration phase, the user is prompted to enter their name, after which the system captures 30–40 facial images using a webcam. These images are preprocessed (normalized, resized) to reduce variations in lighting and scale. Simultaneously, facial emotion detection is performed using a CNN-based classifier trained on the FER2013 dataset. The average detected emotion is stored and later used for emotional consistency checking during access.

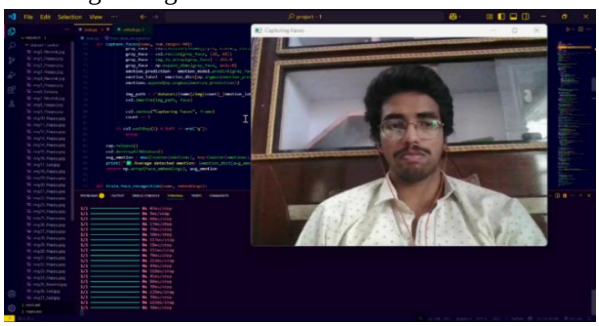

Fig. 1. Capturing Face Samples During Registration

B. Voice Acquisition and Speaker Recognition

The system then records 15 voice samples of the user, each with predefined sentences. Voice features, such as MFCCs, are extracted and used to train a speaker recognition model using a feedforward neural network. This model achieved 83.3% accuracy in identifying users based on their voice profiles.

During authentication, a new voice sample is taken and compared to the registered model to validate the speaker's identity.

Fig. 2. Recording Voice Samples During Registration

C. Speech-Based Password Setup and Verification

After collecting the biometric data, the system asks the user to set a password by speaking it out loud. The spoken password is converted to text using a speech-to-text model and stored securely. During the unlocking process, the user is prompted to speak the password again, which is then re-converted to text and matched with the stored version. A similarity score is computed, and only if it meets the threshold, the password is accepted.

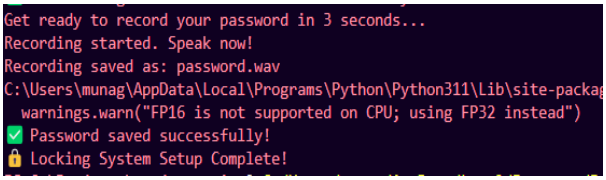

Fig. 3. Speech-Based Password Setup Completion

D. Locking Process Flow

The complete registration (locking) process is visualized in Fig. 4. This includes sequential steps for face capture, voice recording, emotion extraction, and password setup. All the extracted data is stored securely in a user profile for future matching.

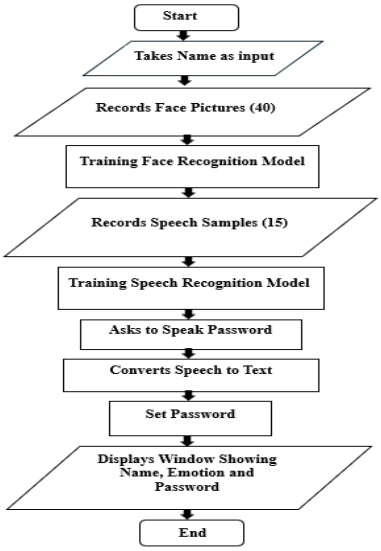

Fig. 4. Flowchart of Locking (Registration) Phase

E. Unlocking Process Flow and Multimodal Matching

During the unlocking phase, the system sequentially performs multiple biometric checks to ensure secure access. It begins with facial recognition. If the detected face matches the stored face model, the system proceeds to validate the emotion detected during unlocking against the average emotion recorded during the initial registration. Only if both face and emotion match, the system then initiates voice-based user identification. Upon successful voice match, the system performs speech-to-text conversion of the spoken password and compares it with the stored password text.

Access is granted only if all the above stages—face recognition, emotion consistency, voice recognition, and password match—are successful. This strict sequential validation ensures a robust multimodal security mechanism resistant to spoofing or single-point failure.

The stepwise design not only enhances security but also simplifies error tracing during failed attempts. If any stage fails, the system immediately halts the process and denies access, thereby reducing unnecessary computational load.

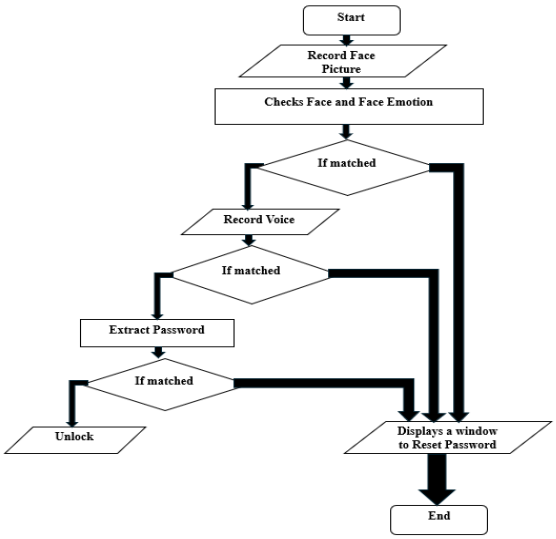

Fig. 5. Flowchart of Unlocking (Authentication) Phase

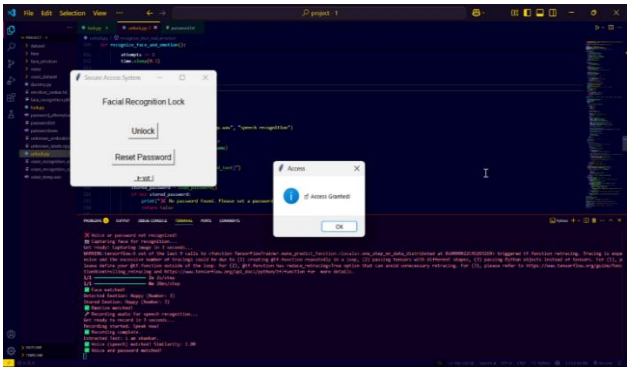

Fig. 6. Real-Time Unlocking Interface Screenshot

III. SYSTEM IMPLEMENTATION

The Secure Access System is designed with a two-phase structure: Enrolment Phase and Identification (Unlocking) Phase. Fig. 4 & 5 illustrates the overall architecture of the proposed multimodal biometric authentication system.

During the Enrolment Phase, the user registers their facial data, voice samples, and a spoken password. Initially, the system captures 30–40 face images through the webcam. Face detection is performed using the MTCNN (Multi-task Cascaded Convolutional Neural Network), and features are extracted through the FaceNet embedding model. These embeddings are stored and trained using a Support Vector Machine (SVM) classifier. Simultaneously, the system detects the emotion in each

captured frame using a CNN-based emotion recognition model trained on the FER2013 dataset. The average emotion value across all images is computed and saved.

In parallel, the user is prompted to speak multiple sample sentences. A total of 15 audio samples are recorded and processed using the Librosa library to extract Mel-frequency cepstral coefficients (MFCCs). These features are used to train a custom Voice Recognition Model that distinguishes the user from others. Furthermore, the user is asked to speak a secret password, which is converted into text using the Whisper Small model. This password text is stored securely for future verification. Additionally, a recovery question and answer are stored as a fallback option.

During the Identification Phase, the system initiates the unlocking process. The user's face is captured in real-time and passed through the face recognition and emotion detection pipeline. The detected emotion is compared against the stored average emotion. If both face and emotion are verified, the system proceeds to record the user's speech for 5 seconds. The spoken password is transcribed using speech-to-text conversion, and matched with the stored password using a string similarity ratio computed by difflib. A similarity threshold of 0.7 is used for acceptance.

If all three modalities—face, emotion, and voice—are successfully verified, access is granted. Otherwise, the user is prompted with the recovery question to regain access.

The system is implemented using Python, with key libraries such as OpenCV, TensorFlow/Keras, Librosa, Tkinter, and sounddevice. Noise in audio signals is reduced using the noisereduce library to improve voice recognition accuracy. All interactions are controlled through a graphical interface built with Tkinter, offering options for unlocking, resetting the password, and exiting.

IV. RESULTS AND DISCUSSIONS

A. Face Recognition Model Performance

The Face Recognition model was implemented using FaceNet and evaluated based on accuracy, confusion matrix, and error rates. The model achieved an accuracy of 94.87%, demonstrating high reliability in person identification. The confusion matrix in Fig. 7confirms minimal misclassifications, with most predictions falling on the diagonal, indicating correct classifications. The False Acceptance Rate (FAR) was 0.00%, ensuring no unauthorized access, while the False Rejection Rate (FRR) was 2.56%, suggesting minor rejections of valid users.

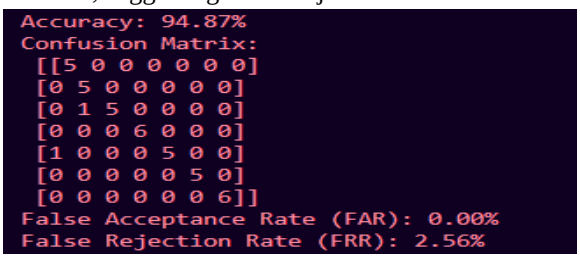

Figure 7: Confusion Matrix of Face Recognition Model

B. Voice Recognition Model Performance

The Voice Recognition model was developed using MFCC feature extraction and a deep learning-based classifier. The training process as shown in Fig. 8, demonstrates a steady improvement in validation accuracy, converging at 83.33% accuracy. The model achieved an Equal Error Rate (EER) of 3.33%, indicating a balance between FAR and FRR.

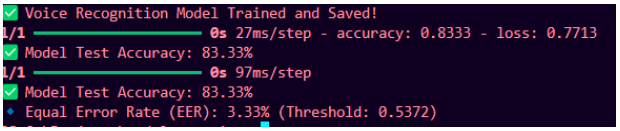

Figure 8: Training Performance of Voice Recognition Model

Additionally, the MFCC visualization in Fig. 9 illustrates the extracted features from speech signals, which play a crucial role in differentiating users based on vocal characteristics.

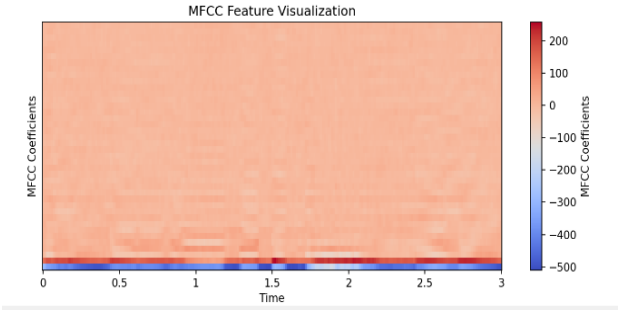

Figure 9: MFCC Feature Visualization

C. Face Emotion Recognition Model Performance

The Face Emotion Recognition model was evaluated using a confusion matrix and standard classification metrics. The confusion matrix Fig. 10 reveals strong classification performance for "Happy" and "Surprised" emotions, with F1-scores of 0.80 and 0.74, respectively. However, emotions such as "Fearful" and "Sad" showed lower recall values, leading to misclassifications.

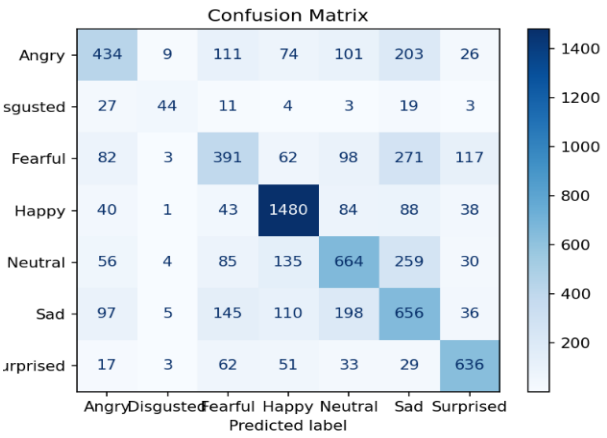

Figure 4: Confusion Matrix of Face Emotion Model

TABLE I. PERFORMANCE METRICS OF FACE EMOTION RECOGNITION MODEL

Emotion	Precision	Recall	F1-Score	Support
Angry	0.58	0.45	0.51	958
Disgusted	0.64	0.40	0.49	111
Fearful	0.46	0.38	0.42	1024
Happy	0.77	0.83	0.80	1774
Neutral	0.56	0.54	0.55	1233
Sad	0.43	0.53	0.47	1247
Surprised	0.72	0.77	0.74	831

Table I summarizes the model's precision, recall, and F1-score, with an overall accuracy of 60%. The weighted-average F1-score of 0.60 indicates moderate classification performance, with improvements required in distinguishing similar expressions.

V. CONCLUSION

Biometrics plays a vital role in ensuring secure and reliable personal authentication. The objective of implementing a robust and intelligent access system has been effectively achieved through the integration of multimodal biometric features, including facial recognition, voice-based person identification, and facial emotion verification. FaceNet-based recognition and emotion-aware access policies significantly enhance the system's security. Voice recognition using MFCC features also contributes to user-specific verification with high precision. The proposed method improves the overall recognition rate, reducing false acceptances and rejections. The experimental results confirm enhanced accuracy and robustness of the developed model in real-time scenarios, thereby validating the effectiveness of the proposed Secure Access System.

REFERENCE

[1] Vaishali, D., Bhargavi, A., Reshma, S.J.A., Krishna, K.S. and Maanesh, M., 2024. Face Recognition based Door Lock System. In 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS). IEEE.

[2] Shofiyah, Z., Mahmudah, H., Santoso, T.B., Puspitorini, O., Wijayanti, A. and Siswandari, N.A., 2022. Voice Recognition System for Home Security Keys with Mel-Frequency Cepstral Coefficient Method and Backpropagation Artificial Neural Network. In 2022 International Electronics Symposium (IES). IEEE.

[3] Tümen, V., Söylemez, Ö.F. and Ergen, B., 2017. Facial emotion recognition on a dataset using convolutional neural network. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE.

[4] Conklin, A., Dietrich, G. and Walz, D., 2004. Password-based authentication: a system perspective. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences. IEEE.

[5] Kar, B., Kartik, B. and Dutta, P.K., 2006. Speech and Face Biometric for Person Authentication. In 2006 IEEE International Conference on Industrial Technology. IEEE.

[6] Al-Majmaie, M.J.R. and Çevik, M., 2022. Deep Learning-Based Biometric System Analysis of Palmprint Images. In 2022 International Conference on Artificial Intelligence of Things (ICAIoT). IEEE.

[7] Schroff, F., Kalenichenko, D. and Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[8] Babu, P.A., Nagaraju, V.S. and Vallabhuni, R.R., 2021. Speech Emotion Recognition System With Librosa. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). IEEE.

[9] Kumar, P. and Li, X., 2023. Interpretable Multimodal Emotion Recognition using Facial Features and Physiological Signals. arXiv preprint arXiv:2306.02845.

[10] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D., 2019. M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues. arXiv preprint arXiv:1911.05659.