A Lightweight Approach Towards Speaker Authentication Systems

Rishi More
Computer Engineering
K.J. Somaiya Institute of Technology

Sahil Deshmukh
Computer Engineering
K.J. Somaiya Institute of Technology

Krishi Chawda
Computer Engineering
K.J. Somaiya Institute of Technology

Dhruv Mistry
Computer Engineering
K.J. Somaiya Institute of Technology

Prof. Abhijit Patil
Computer Engineering
K.J. Somaiya Institute of Technology

***Abstract*—In a world where traditional authentication systems are constrained, the introduction of voice-based authentication provides a viable option. This cutting-edge biometric security method uses unique vocal traits including pitch, tone, and speech patterns to create a unique voiceprint. Its sophisticated and dependable verification procedure, when combined with voice-activated services and secure access systems, not only solves authentication issues but also improves user experience overall. Our suggested method provides a thorough blueprint for developing a lightweight voice authentication system that is based on text. This system leverages an effective encoder model to convert voice spectrograms, making deployment easy even on edge devices with limited resources. Various dimensionality reduction techniques are explored to obtain optimal voice embeddings that capture speaker uniqueness while minimizing model complexity. A key novelty is the application of model compression techniques including lightweight architectures and Siamese Networks to obtain highly condensed voice embeddings for each user, reducing storage and infrastructure costs compared to traditional voice authentication methods. The proposed lightweight spectrogram embeddings leverage language-agnostic acoustic features, enabling language-independent speaker verification. Additionally, dimensionality reduction applied during voice registration allows the capturing of discriminative voice characteristics in a low-dimensional compact feature space. This significantly cuts down the storage requirements and infrastructure costs per user compared to standard voice biometric approaches, while retaining competitive verification performance. The architecture is optimized for responsiveness by leveraging lightweight frameworks. The proposed system delivers competitive voice authentication capabilities while minimizing memory, computational, and energy footprints. This makes the system useful for integration into smart devices and paves the way for ubiquitous voice biometrics.**

***Keywords*—Voice Based Authentication, Language Independent Speaker Verification, Siamese Networks, Voice Embeddings, Convolutional Neural Networks**

I. Introduction

Recent years have witnessed a significant transition in the field of voice authentication, driven by a novel lightweight technique that has the potential to completely change the biometric security environment. This paradigm change is motivated by our lightweight voice authentication model's uniqueness and potential to significantly reduce infrastructure costs and time complexity, in addition to the promise of improved security, customisation, and user identification.

Our method for voice authentication leverages the efficiency benefits provided by [1]Encoder Embeddings and Deep Learning Neural Networks, while focusing on unique voice features analysis. By utilizing these cutting-edge technologies, we have developed an innovative and user-friendly identification confirmation process that raises the bar for effectiveness.

[2] Encoder Embeddings are essential to our systems' ability to quickly comprehend the unique characteristics of each speech because they make the data we use simpler. Our Deep Learning Neural Networks then use these insights to process them intelligently, guaranteeing a dependable and effective authentication procedure that drastically lowers CPU overhead.

Like fingerprints, voice authentication uses distinctive human characteristics as a biometric technique. But our lightweight approach stands out from the crowd since it significantly reduces infrastructure costs and time complexity when compared to conventional authentication techniques. Additionally, because spectrogram embeddings are language independent, our approach expands accessibility and application possibilities.

We utilize the power of Convolutional Neural Networks (CNNs) and technology derived from the VGG-Net model [3] to further optimize our lightweight model. These advancements allow us to identify distinct speech patterns, increasing system accuracy and lowering resource needs at the same time.

The crucial choice of an effective encoder framework is at the center of our project. This approach optimizes computing performance by introducing dimensionality reduction and streamlining the process of converting raw speech samples into spectrograms.

Our project smoothly moves to the design of a user interface (UI) that maximizes the user experience while lowering infrastructure expenses once we have determined the ideal lightweight encoder model. The homepage, registration page,

and login page are examples of important interfaces that have been painstakingly designed to guarantee accessibility and usability.

Our study concurrently investigates the possibility of using NoSQL databases to safely store user-specific embeddings and related identifying data [4]. This database schema is designed to meet the strictest requirements for data security and accessibility, protecting the voice authentication system's integrity and making the most use of the available infrastructure.

In conclusion, the study presents a novel, low-weight voice authentication system supported by cutting-edge technology. We seek to make speech based authentication as commonplace as password or facial authentication by bringing dimensionality reduction, improving security, enabling language independence, and drastically lowering time difficulties and infrastructure costs.

II. Literature Review

Automated speech recognition (ASR) in Indian languages, dysarthric ASR systems, speaker-independent keyword identification, deep network performance degradation, disease diagnosis, automatic voice problem diagnosis, spectrum analysis, Voice Activity Detection (VAD) in noisy environments, and basic principles of speech recognition are just a few of the areas that voice processing research covers. A hybrid CNN-LSTM model is proposed in the context of speaker-independent keyword identification, and it makes use of voice conversion as a data augmentation strategy to improve performance, particularly on small datasets [5]. The model's applications to load forecasting and mood identification highlight the subtleties of speech conversion and its practical uses in industry. In order to solve deep network performance degradation, skip connections are introduced in another study that examines the effects of stacking blocks in a 1-D Convolutional Neural Network (CNN) architecture for overlapping audio identification [6]. An artificial neural network (ANN) with an 89.07% success rate in LOO training is used to investigate disease detection in clinical settings, indicating new avenues for investigation [7].

Neural networks are used for the automatic diagnosis of speech problems, with acoustic analysis serving as the main diagnostic method [8]. Through a comparison of two neural network topologies, this study reveals higher short-term variability in pathological settings and proves the superiority of learning vector quantization. Another study focuses on spectrum analysis and uses the Flower Pollination Algorithm (FPA) and periodograms to test a solution [9]. To improve Voice Activity Detection (VAD) in noisy situations, a novel methodology is presented that outperforms existing approaches in terms of detection costs and achieves higher F1-scores [10].

Using wavelet characteristics as input to a transformer network addresses efficient technology for ASR in Indian languages, leading to lower word mistake rates and effective recognition [11]. For most dysarthric people, transfer learning combined with a customized deep transformer architecture performs better than current approaches in dysarthric ASR systems [12]. A paper that provides an overview of fundamental ideas and notations in voice recognition—defining input (X) as speech features and output (Y) as tokens—concludes the review [13].

Turning now to speaker recognition, current research focuses on advancements in audio and speech processing. The state-of-the-art used to be traditional i-vector systems with probabilistic linear discriminant analysis (PLDA), but deep learning—especially with raw spectrogram inputs—has taken over for speech embedding learning. Research is greatly advanced by the VoxCeleb dataset, which contains over 1 million utterances from 6000 speakers [14].

The VGG-M CNN architecture and Siamese networks with coupled weights for voice verification are two examples of deep CNN architectures being investigated in research employing spectrograms for speaker recognition [14]. Siamese networks, with their lightweight architectures such as MobileNets, hold great potential for efficient implementation on edge devices. Building on these developments, the proposed study creates a lightweight Siamese architecture for speaker verification, which is tested on VoxCeleb and additional datasets. A cheap yet effective protection against adversarial attacks that take advantage of current vulnerabilities in speech recognition systems is provided using spectrogram features and Siamese networks [15]. Spectrograms may be more resilient to adversarial attacks since they can recover phase information that is lost in traditional MFCC features. The study assesses how resistant spectrogram features are to hostile attacks on MFCC-based systems and shows how well Siamese networks trained on large datasets can fend off such attacks [15]. To measure the effectiveness against benchmark attacks, multiple speaker identification pipelines are evaluated, with Siamese networks used in place of standard embeddings and spectrograms used in place of MFCCs [15].

Wearable-based implicit authentication using multi-modal physiological signals was limited in its use due to prior research in the field of biometric authentication focusing on one or two signal modalities. Gaussian classifiers are used in recent studies to investigate motion-based data, PPG, ECG, EDA, and ACC signals for multi-modal biometric identification. CNN-LSTM models and accelerometer and gyroscope features have been studied. Present research focuses on evaluating techniques in controlled real-world settings, applying SVM classifiers to preprocessed features and utilizing Fitbit users' health-related data. Notably, the system presented in this research is the first to integrate end-to-end deep learning algorithms and raw physiological signals for everyday biometric authentication [16].

The risks of stolen biometric templates and the necessity of template security are highlighted, along with privacy problems and limitations in biometric authentication. The article discusses two primary strategies: Cancellable Biometrics (CB) and Biometric Cryptosystems (BCS), emphasizing their possible impact on authentication performance and lack of verifiable security. There is discussion of many access control

models that aim to combine authorization with authentication, such as Mandatory Access Control (MAC), Role-based Access Control (RBAC), and Discretionary Access Control (DAC). The suggested AuthN-AuthZ system uses a hierarchical RBAC model with privacy-preserving Biometric-Capsule-based authentication to solve usability and privacy concerns [17].

A deep learning-based biometric authentication system called CM-PIR, which uses data on chest movements from passive infrared (PIR) sensors, is presented in a different publication. Using a recurrent neural network (RNN) with a 90-second window size, the system achieves 75% accuracy for biometric authentication and % accuracy for stationary human detection. The characteristics consist of acceleration filter coefficients, discrete wavelet transform (DWT), and Fourier transform (FFT). Utilizing the TensorFlow-based Keras deep learning framework, the RNN exhibits strong performance across several domestic environments, indicating its potential for real-world implementation [18].

Siamese-VGG, an inventive technique that surpasses conventional methods, introduces innovations in face tracking. Siamese-VGG improves robustness and generalization by utilizing the first two convolutional layers of VGG-16 for feature extraction. This helps to solve issues like as quick motion, scale changes, rotation, occlusion, and lighting variations. For recognition and fine-tuning, a pre-trained VGG-Face model is employed, and for better performance during training, inner template feature maps are extracted. To improve generalization, L2 regularization is incorporated into the loss function. Siamese-VGG sustains a frame rate of 18.5 frames per second on the Nvidia GTX1070Ti GPU while achieving an 11% overlap improvement in complicated scenarios. Because of its useful speed and accuracy, it may be used for a wide range of tasks. A deep neural network trained for face similarity learning can replace manual feature construction and reliably track faces in complicated backdrops while also adjusting to changes in light, motion blur, and occlusion [18].

A paradigm that makes use of both auditory and visual biometrics is presented in the investigation of crossmodal biometrics for person identification. Traditional biometric systems frequently neglect the possibilities of cross-modal information in favor of single-modal testing and training. This method, which investigates non-parametric density estimation and Hebbian projection matrices for creating speaker-specific models, is especially pertinent to surveillance applications. In order to utilize information from both auditory and visual data during training, the study also presents a model-mapping framework, offering a thorough method for crossmodal person identification [19].

Another study presents a novel use of deep learning techniques to identify cough audio characteristics associated with COVID-19. This is done in conjunction with an exploration into the application of quantum neural networks (QNN) for COVID-19 cough classification, using the VGG-13 architecture. Long run times and reliable data pretreatment are two challenges in quantum machine learning simulations that are addressed. The work applies preprocessing methods including silence detection and log-mel spectrograms to audio from the DiCOVA and COVID datasets. The study controls quantum noise and precision in spite of access issues to quantum computers, providing insights on the possibility of deep learning and quantum neural networks in detecting cough patterns associated with COVID-19 [20].

III. Methodology

The primary goal is to establish an effective system for speaker identification using an efficient pipeline to mel spectrograms and Siamese neural networks.

A. Data Preprocessing

The VoxCeleb dataset, a large collection of voice data with over 7000 speakers, is used as the main data source. VoxCeleb contains short segments of human speech extracted from interview videos uploaded to YouTube, with over 1 million utterances totaling over 2000 hours of audio.

The data preprocessing pipeline involves a systematic approach to transform raw audio data into a format suitable for subsequent analysis. The Librosa Python library provides functions for loading audio data and transforming it into representations that are optimized for audio machine-learning tasks.

The initial step involves loading audio files using Librosa. We first read the audio file and retrieve the audio signals along with their corresponding sampling rate. The preparation of datasets for training and evaluation is performed through numerous steps, including the generation of pairs of mel spectrograms with corresponding labels denoting the same or different speakers. We achieve this by organizing the data and storing it in a CSV file for subsequent model training.

Throughout training and testing, numerous checks are performed to ensure data validity. We inspect the dimensions of the input as well as guarantee that the audio file sizes are within the expected range.

B. Extracting Mel Spectrograms

For speaker recognition, mel spectrograms help capture vocal tract characteristics such as formant frequencies, resonances, articulation, and speaking style. These characteristics manifest themselves as patterns in the mel frequency bands over time. Machine learning models can leverage these patterns to discern different speakers.

Computing mel spectrograms is an essential preprocessing step when training machine learning models for speaker verification. Mel spectrograms provide a compact time-frequency representation where the frequencies are converted to the mel scale using perceptual filterbanks.

The Mel scale better approximates the human auditory system's nonlinear perception of pitches compared to the linear Hertz scale. Research has shown that the logarithmic compression provided by the mel scale makes the features more robust for audio classification compared to normal Short-Time Fourier Transform (STFT) features.

We apply mel filters to map the raw audio to a logarithmic frequency scale that approximates human pitch sensitivity.

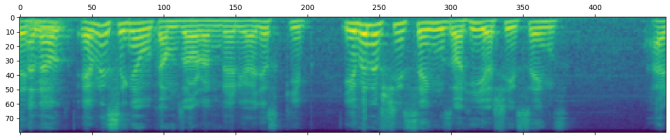

Fig. 1. Mel Spectrogram

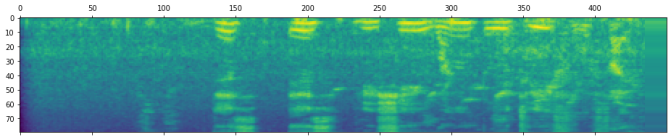

Fig. 2. Padded Mel Spectrogram

Noise in the spectrograms is reduced through smoothing window functions. Storing these preset parameters and transformation matrices in memory optimizes the pipeline by avoiding unnecessary redundant computations every time a new audio clip is processed. By caching these globally, we minimize repetitive calculations and maximize resource efficiency.

To generate the mel spectrograms, the raw waveform is windowed into frames. In order to generate mel spectrograms suited for audio recognition tasks, key parameters need to be set including the Fast Fourier Transform (FFT) size, Mel bands, frame length, frameshift, and frequency cutoffs. The parameters used in our study are 1024 FFT bins, 80 mel bands, 1024 sample frame length, 256 sample frame shifts, 22050 Hz sampling rate, and frequency cutoffs of 0 to 8000 Hz.

The Short-Time Fourier Transform (STFT) is calculated to analyze the frequency content of the audio signal over time using PyTorch. A key parameter for the STFT is the window size, which is set to the sampling rate of the audio signal. This window size determines the time resolution of the STFT. A smaller window provides better time resolution but worse frequency resolution. The hop length is set to 256, which determines how much the window shifts between adjacent STFT frames. A smaller hop size increases the overlap between frames but also increases computational cost.

To minimize edge effects, the dictionary containing hann windows is utilized for windowing. The Hann window smoothly tapers the signal to zero at the edges. Additionally, the center parameter is set to true so that the window was centered at each time step.

The audio signal is padded to match the window length before computing the STFT. The pad_mode is set to 'reflect' so that the padding was a reflection of the signal rather than zeros. This reduces edge discontinuities. The normalized parameter is set to False to provide the raw STFT magnitude rather than a normalized version.

The 1024 FFT size provides sufficiently high frequency resolution. 80 mel bands are chosen to get compact yet discriminative features. The 1024 sample (46 ms at 22050 Hz) frame length captures multiple pitch periods for good frequency resolution. A 256 sample frame shift provides about 50% overlap between frames.

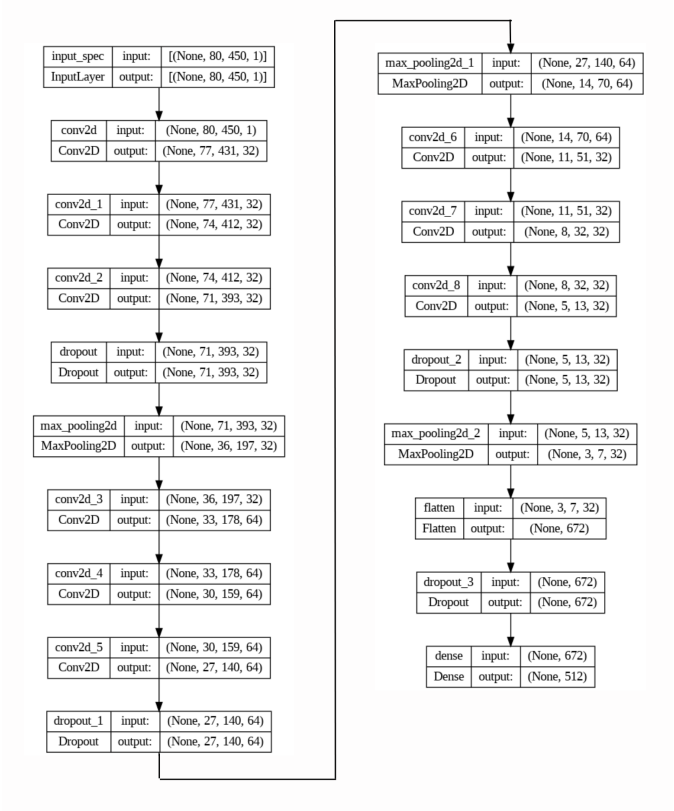

Fig. 3. Tail of Proposed Architecture

We also perform dynamic range compression on the spectrograms to limit the dynamic range using logarithmic transformation. This limits the dynamic range of the data, with an option to adjust the compression factor and clip values.

Spectral normalization is applied to the spectrograms as part of the preprocessing pipeline. Figure 1 showcases a visualization of the resulting normalized spectrogram. To ensure compatibility of the input data during training, the spectrograms are padded and cropped to a standardized shape of 80x450. Figure 2 displays the matrix representation of a padded spectrogram after this preprocessing step.

The result is an efficient input encoding that allows training high-performance deep neural networks for reliable audio analysis. The mel compression also makes the features more compact compared to raw STFT or MFCCs.

C. Model Architecture

We propose a model architecture employing a multi-branch Siamese neural network for feature extraction and comparison. It contains two identical convolutional subnetworks known as tails that share parameters and process separate input samples in parallel. Each tail takes an 80x450 matrix as input and applies a series of convolutional, max pooling, and fully connected layers to encode it into a compact 512-dim embedding vector.

The architecture of the tail network is presented in Figure 3. The tail is designed to extract hierarchical features from

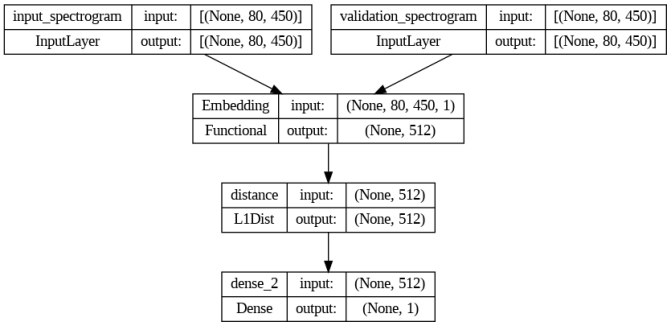

Fig. 4. Head of the Proposed Architecture

the input spectrograms. It contains multiple convolutional layers interspersed with max-pooling layers for downsampling, followed by fully-connected layers. The model architecture consists of the following layers:

- **Input layer:** The input layer takes 450x80 spectrograms as input.The input layer does not perform any computation.
- **Dense layer:** This layer performs high-level reasoning and classification based on the flattened feature vector. It takes the 1D output from the flattening layer as input. The FC layer is densely connected, where each neuron in the layer is connected to every neuron in the previous layer.
- **Dropout layer:** This layer helps to prevent overfitting by randomly dropping out units during training. The final dense layer with a sigmoid activation function produces the embedding vector, which is a low-dimensional representation of the input data that captures its most important features.
- **Convolutional layer:** This layer extracts low-level visual features from the input images. The max pooling layers downsample these representations. The tail has 3 convolutional layers, each with 32 3x3 filters and a ReLU(rectified linear unit) activation function. The convolution strides are set to 1 pixel.

$$\mathrm{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{1}$$

- **Max Pooling layer:**This layer performs downsampling to reduce the spatial dimensions of the feature maps. In the tail, max pooling regions of 2x2 pixels are used with a stride of 2. So each pooling layer downsamples the input by a factor of 2 along width and height. Max pooling condenses the representations and allows the detection of dominant features.
- **Flatten Layer:** This layer transforms the output of the previous convolutional and pooling layers into a 1D vector. It collates all the feature maps from the prior layers and concatenates them into a single long vector.
- **L1 Dist Layer:** This layer computes the manhattan distance between the input embeddings as a measure of their similarity. It takes two input tensors - an input embedding and a validation embedding. These are the latent representations generated by the previous convolutional feature extractor branches.

$$d_{\text{Manhattan}}(E_1, E_2) = \sum_{i=1}^{n} |E_1 i - E_2 i| \tag{2}$$

In the first stage, three consecutive convolutional layers with 32 filters of kernel size (4, 20) and stride 1 perform feature extraction on the (80, 450, 1) input. The outputs are then downsampled by a 2x2 max pooling layer with stride 2. The second stage repeats this process but with 64 filters to extract higher-level features. The resulting representations are further downsampled by another 2x2 max pooling layer. In the third stage, three more convolutional layers with 32 filters are applied, followed by a final 2x2 max pooling layer to aggregate spatial information.

These successive convolutional and pooling layers allow the tail to learn location-invariant features at multiple scales. The pooled feature maps are finally flattened into a vector and passed through fully connected layers for classification. This hierarchical structure allows efficient learning of discriminative spectrogram features for audio scene analysis.

The architecture of the head network is presented in Figure 4. The head module performs speaker verification by computing the distance between two embeddings. During training, the network learns to produce embeddings that are close together for matching speakers while being far apart for different speakers. In practice, the tail module encodes the input spectrograms into fixed-dimensional embeddings. The head module then computes the Manhattan distance between an enrollment embedding and a test embedding.

If the distance is below a threshold, the speakers are considered a match. This two-module architecture allows efficient optimization of the feature extraction and metric learning components for robust speaker verification. The tail can be pre-trained on speaker identification datasets before fine-tuning the head module for verification.

The two-module architecture of the proposed model offers several advantages for speaker verification. First, it allows for efficient optimization of the feature extraction and metric learning components. The tail network can be pre-trained on speaker identification datasets, which are more abundant than speaker verification datasets. This allows the model to learn discriminative spectrogram features without requiring a large amount of labeled speaker verification data.

Second, the two-module architecture makes the model more robust to noise and other variations in the speech signal. The tail network extracts hierarchical features from the input spectrogram at multiple scales, which makes it less sensitive to small perturbations in the input signal. The head module then learns a metric that is specifically designed for speaker verification, which further improves the robustness of the model.

Finally, the model is trained using a generator function that batches and loads MFCC data for training the Siamese network. This generator function efficiently manages data retrieval and aids in the iterative training process.

In summary, our methodology showcases a comprehensive pipeline for voice-based authentication, covering audio data preprocessing, mel spectrogram generation, Siamese network model architecture, and dataset preparation.

IV. Result

Our voice authentication project is a multidimensional accomplishment that embodies a dedication to thorough research and development while producing remarkable accuracy outcomes. The careful gathering of a wide range of speech datasets made sure the model was exposed to a variety of speaking tenors, which enhanced its flexibility in practical situations. We leveraged advanced signal processing techniques, employing functions from the Librosa library to transform audio data into spectrograms. This strategic choice reflects our dedication to adopting cutting-edge tools and frameworks for enhancing model efficiency.

By strategically incorporating speaker-specific information, the triplet-loss function was used to train the encoder architecture. This laid the foundation for future improvements and flexibility in addressing a range of voice authentication difficulties. In order to identify the best architecture for our particular problem statement, we thoroughly investigated Siamese networks and U-Nets as well as other encoding frameworks. This demonstrated our dedication to staying on the cutting edge of deep learning breakthroughs.

Through a meticulous exploration of encoding frameworks like U-Nets and Siamese networks, we identified a model with a training accuracy of 95.52% and a testing accuracy of 96.62%, emphasizing robustness and generalization. Fig. 5 represents the confusion matrix for our model.

TABLE I
Performance Metrics

Metric	Percentage
Precision	98.8%
Recall	94.2%
F1 Score	96.5%

The careful planning that went into the creation of user interfaces for the home, registration, and login pages highlighted how important it was to us to provide a simple, straightforward user experience. Utilizing the model-view-controller architecture of Django facilitated development, guaranteeing both aesthetically pleasing design and effective functioning. This user-centric approach expands the potential applications of voice authentication and is in line with the larger goal of making it accessible and user-friendly.

Our choice to evaluate both SQL and NoSQL databases demonstrated a careful approach to technology selection in

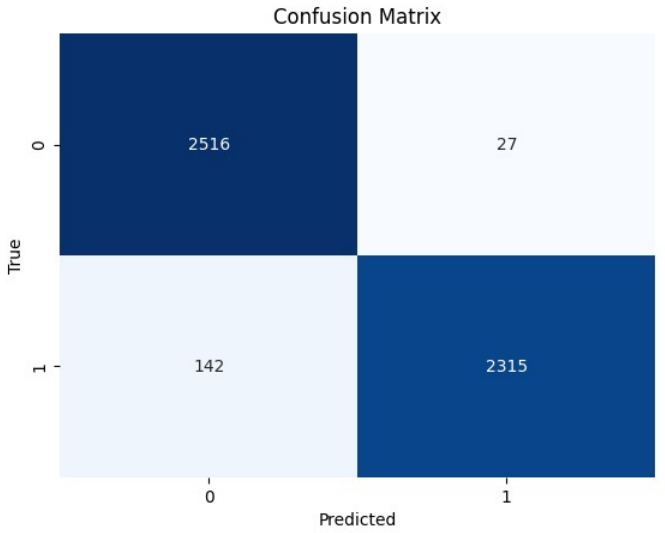

Fig. 5. Confusion Matrix

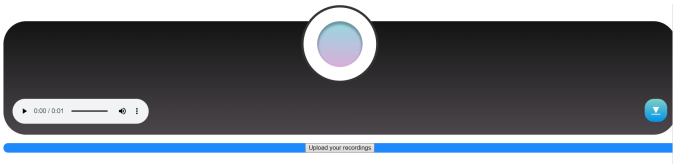

Fig. 6. Recording Page

the field of database administration. The decision to use PostgreSQL was made in light of its effectiveness and suitability for the particular objectives of our project. It ensured safe storage while establishing the framework for future growth and flexibility in response to changing demands.

The efficient back-end and front-end communication made possible by the combination of the Django framework with PostgreSQL also made it possible to implement extra security features like secure user login and encryption methods. Our voice identification system is more reliable overall as a result of this all-encompassing strategy, which also makes it safe, accurate, and easy to use.

Our voice authentication project is a thorough and inno-

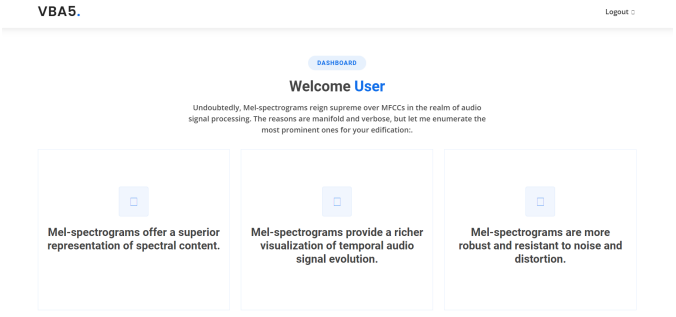

Fig. 7. Authenticated User

vative addition to the industry, with possible uses spanning from identity verification in various real-world contexts to safe access management.

V. Conclusion

The project focused on the implementation and evaluation of voice-based authentication using spectrograms generated through a Siamese neural network model. The utilization of spectrograms provided a unique and effective approach to capturing the intricate details of an individual's voice patterns for authentication purposes.

Through the development and training of a convolutional neural network (CNN)-based model, the system demonstrated promising results in accurately verifying users based on their spectrogram representations. It facilitated the extraction of nuanced features from the spectrograms, enabling the system to distinguish between genuine and unauthorized voices with a high level of accuracy.

The project not only showcased the technical feasibility of voice-based authentication but also highlighted the potential for enhanced security in various applications, such as access control systems, secure transactions, and identity verification processes.

While the system exhibited commendable performance, there is always room for further refinement and optimization. Future work could involve the exploration of larger and more diverse datasets to improve the model's generalization capabilities. Additionally, incorporating real-world environmental factors and variations in speech patterns could enhance the robustness of the authentication system.

The potential applications of this technology in enhancing security and user authentication warrant further research and development in the pursuit of creating more advanced and reliable voice-based authentication systems.

References

[1] H.-W. Lin, V. M. Shivanna, H. C. Chang, and J.-I. Guo, "Real-time multiple pedestrian tracking with joint detection and embedding deep learning model for embedded systems," *IEEE Access*, vol. 10, pp. 51 458–51 471, 2022.

[2] J. Zhang, J. Liss, S. Jayasuriya, and V. Berisha, "Robust vocal quality feature embeddings for dysphonic voice detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1348–1359, 2023.

[3] Y. Jung, Y. Choi, H. Lim, and H. Kim, "A unified deep learning framework for short-duration speaker verification in adverse environments," *IEEE Access*, vol. 8, pp. 175 448–175 466, 2020.

[4] M.-K. He, J. Du, Q.-F. Liu, and C.-H. Lee, "Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1561–1573, 2023.

[5] Y. A. Wubet and K.-Y. Lian, "Voice conversion based augmentation and a hybrid cnn-lstm model for improving speaker-independent keyword recognition on limited datasets," *IEEE Access*, vol. 10, pp. 89 170–89 180, 2022.

[6] M. Yousefi and J. H. L. Hansen, "Block-based high performance cnn architectures for frame-level overlapping speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 28–40, 2021.

[7] C. D. P. Crovato and A. Schuck, "The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 10, pp. 1898–1900, 2007.

[8] J. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[9] D. Połap and M. Woźniak, "Voice recognition by neuro-heuristic method," *Tsinghua Science and Technology*, vol. 24, no. 1, pp. 9–17, 2019.

[10] W. Mu and B. Liu, "Voice activity detection optimized by adaptive attention span transformer," *IEEE Access*, vol. 11, pp. 31 238–31 243, 2023.

[11] T. Choudhary, V. Goyal, and A. Bansal, "Wtasr: Wavelet transformer for automatic speech recognition of indian languages," *Big Data Mining and Analytics*, vol. 6, no. 1, pp. 85–91, 2023.

[12] S. R. Shahamiri, V. Lal, and D. Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3407–3416, 2023.

[13] R. Fan, W. Chu, P. Chang, and A. Alwan, "A ctc alignment-based non-autoregressive transformer for end-to-end automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1436–1448, 2023.

[14] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819302712

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[16] D. Ekiz, Y. S. Can, Y. C. Dardağan, F. Aydar, R. D. Köse, and C. Ersoy, "End-to-end deep multi-modal physiological authentication with smartbands," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14 977–14 986, 2021.

[17] T. Phillips, X. Yu, B. Haakenson, S. Goyal, X. Zou, S. Purkayastha, and H. Wu, "Authn-authz: Integrated, user-friendly and privacy-preserving authentication and authorization," in *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 2020, pp. 189–198.

[18] J. Andrews, A. Vakil, and J. Li, "Biometric authentication and stationary detection of human subjects by deep learning of passive infrared (pir) sensor data," in *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, pp. 1–6.

[19] S. Yuan, X. Yu, and A. Majid, "Robust face tracking using siamese-vgg with pre-training and fine-tuning," in *2019 4th International Conference on Control and Robotics Engineering (ICCRE)*, 2019, pp. 170–174.

[20] M. Esposito, G. Uehara, and A. Spanias, "Quantum machine learning for audio classification with applications to healthcare," in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2022, pp. 1–4.