Grenze International Journal of Engineering and Technology, June Issue

A Voice Biometric System for Faster Logins and Secure Digital Transactions

Varun Parmar[1], Madan Pachori[2], Vijay Parma[3], Utkarsh Singh[4] and Vedanshi Tyagi[5]
[1-5]Greater Noida Institute of Technology/ Computer Science and Engineering, Greater Noida, India
Email: varunparmar5551246@gmail.com, madansati123@gmail.com, vijayparmar77412@gmail.com, utkarsh6124@gmail.com, vedanshityagi0602@gmail.com

***Abstract*— The traditional authentication methods, such as passwords, PINs, and OTPs, are very vulnerable to advanced cyberattacks, which include SIM swapping, phishing, and credential theft. Voice biometrics offers a secure and very user-friendly solution to all these limitations by using unique physiological and behavioral speech traits. This work gives light to VoiceGuard, a deep learning-based voice authentication system that can fight replay attacks and deepfake audio. It also includes MFCC features, spectrogram-based embeddings, and a hybrid CNN-LSTM architecture, which is integrated into the system, and it is further being enhanced by anti-spoofing techniques such as challenge response. VoiceGuard has a high potential for high-security applications, such as mobile banking, digital payments, and access control, and it is demonstrated by the experimental results, which show strong accuracy, resilience across environments, and highly effective deepfake resistance.**

***Index Terms*— Voice Biometrics, Authentication, Mfcc, Deepfake Detection, Replay Attack Prevention, Speaker Verification, Cybersecurity, Machine Learning.**

I. INTRODUCTION

It's getting harder for the conventional authentication techniques to meet security requirements for the modern digital ecosystem. The Personal Identification Numbers (PINs) and Passwords, which we thought as essential security measures, are currently the most vulnerable components of authentication processes. Weak password selection, cross-platform password reuses, and susceptibility to phishing and social engineering make it very simple for attackers to breach the users accounts. Furthermore, the common ways to circumvent second-factor security measures like the One-Time Passwords (OTPs) which includes SIM-swapping, malware-based interception, adversary-in-the-middle (AitM) attacks, and credential exploitation. Thus, in addition of reducing security benefits, the conventional authentication significantly hinders end user's usability. Biometric authentication has emerged as a promising alternative because it is based on human characteristics rather than the user-generated secrets. One of the biometric modalities which have received the most attention is the voice biometrics because of its exceptional capacity to integrate physiological and behavioral data of the user. Physiological features such as formant structure, vocal tract geometry, glottal excitation patterns, and harmonic energy distribution provide a stable biological signature. Concurrently, behavioral traits such as intonation, speaking rhythm, coarticulation, and stress patterns contribute additional discriminative qualities. As the voice is a multimodal biometric, it is both statistically reliable and very difficult to replicate [1].

Grenze ID: 01.GIJET.12.2.47_20

The widespread availability of microphones equipped devices such as laptops, smart speakers, smartphones, and car interfaces, have further increases the usefulness of the voice-based authentication. Voice interaction does not require a specialized hardware, fixed lighting, or any kind of physical contact, it enables seamless and natural user experiences. These advantages make voice biometrics a competitive substitute for password-based authentication in high-security environments [2].

To use all the advantage and to solve the security flaw present in the current verification system, VoiceGuard is suggested. Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram-based embeddings, convolutional and recurrent neural models, some attention-driven classifiers, and some very sophisticated signal-processing and machine-learning techniques are cleverly integrated int the VoiceGuard. This system also has specific anti-spoofing mechanisms which is intended to hold against the channel manipulation, replay attacks, and deepfake audio synthesis [3].

Encrypted biometric storage and privacy protecting authentication methods have guarantees us with secure template protection. In order to provide more reliable performance in variety of environmental circumstances, VoiceGuard additionally supports the cross-device adaptability and multi-factor verification [4].

VoiceGuard offers a robust as well as a scalable substitute for conventional authentication method by integrating voice biometric uniqueness, deep learning-based modeling. Because of this design, it can be used in multiple industries like e-commerce, digital banking, financial transaction authorization, remote identity verification [5].

II. Background and Related Work

Over the past 20 years, Voice-based authentication have undergoes various developments. Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and some fundamental spectral features like Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) were the core concepts of the early systems. These methods have increased recognition accuracy, but they were susceptible to replay attacks [6].

Deep learning innovations such as CNN-based spectrogram encoders, LSTM-based temporal modeling, and transformer-driven representation have significantly improved the speaker discrimination. Some Research initiatives such as the ASVspoof Challenge have demonstrated the improvements in the detection of synthetic speech and replayed speech, but they have also revealed the persistent problems in the real time detection of AI-generated voice [7].

Modern anti-spoofing techniques uses the microphone-channel analysis for detecting the liveness, spectral distortion detection, and CQCC feature extraction which can significantly improve robustness. CQCC features are one of the strongest defenses against multiple spoofing attacks [8].

Large Scale datasets such as VoxCeleb2 have accelerated progress by providing a wide variety of speaker samples to train the deep neural speaker-embedding models, which allows for improved performance across various environments and recording condition [9]. Moreover, research on multi-speaker text-to-speech synthesis shows the need of robust verification system by showing how the speaker embeddings are being used to produce deepfake voices [10]. According to multiple surveys on spoofing and anti-spoofing, traditional Voice verification methods should highly integrate multi-feature fusion, adversarial training, and deep-learning-based liveness verification to counter the advancing threats [11]. Research on media forgery and deepfake manipulation supports the need of secure end-to-end encrypted biometric verification pipelines which can withstand modern synthetic voice attacks [12].

Recent developments have introduced adversarially trained Wav2Vec-based verification model, which does significantly improves the spoof resistance by learning the invariant speaker representations even under multiple adversarial manipulation [13]. Meanwhile, the multi-feature fusion models have demonstrated that they are superior effective against various replay attacks, which outperforms the traditional single-feature classifiers under real-world evaluation [14].

These lines of researches collectively demonstrate the urgent need of modern authentication system such as VoiceGuard. To integrate deep-learning-based modeling, cross-channel normalization, and multi-layer anti-spoofing strategies which can maintain the security under real-world attack conditions.

TABLE I. COMPARISON OF AUTHENTICATION SYSTEMS: STRENGTHS AND LIMITATIONS

System Type	Strength	Limitations
Password/PIN	Easy to deploy	Easily stolen or guessed
OTP/MFA	Moderate security	Vulnerable to interception
Basic Speaker	Fast & usable	Replay attacks, Recognition
VoiceGuard (Proposed)	Secure, anti-spoofing, scalable deepfake susceptibility	Requires computational processing

III. Problem Statement

Even though multiple authentication systems are being used everywhere today, keeping the digital identity secure but still they have major problems. Common methods like passwords and one-time passwords (OTPs) are somewhat becoming less effective because of the cyberattacks are getting more advanced and users tends to make mistakes in some situations. Due to all of these ongoing cybersecurity issues, several key problems in digital authentication can be identified:

A. Vulnerability to Credential-Based Attacks

Password-based authentication suffers from predictable weaknesses. Users frequently reuse passwords, choose weak combinations, or disclose credentials through phishing. Automated password cracking tools and credential-stuffing attacks further diminish reliability [15].

B. Compromise of OTP and MFA Systems

While multi-factor authentication was introduced to strengthen security, attackers now exploit: SIM swapping, Malware-assisted OTP interception, Real-time adversary-in-the-middle (AitM) phishing frameworks, social engineering exploiting human error. Thus, OTPs no longer guarantee secure identity verification [16].

C. Inadequate Strength of Existing Voice-Based Authentication

Many legacy voice authentication systems rely solely on acoustic similarity. These systems are vulnerable to: Replay attacks using high-fidelity recordings, deepfake or synthetic speech generated using AI models (e.g., Tacotron, WaveGAN, VITS, UniSpeech), and mimicry-based impersonation, without advanced spoof detection, such systems cannot be deployed in high-security applications [17].

D. Lack of Cross-Device and Environmental Robustness

Real-world authentication environments include variability such as Background noise, microphone differences, audio compression artifacts, and user speech tone changes due to age, health, or emotion. Most existing solutions fail to generalize effectively across these variabilities [18].

E. Insecure Biometric Template Storage

Biometric systems require secure storage, but many current solutions store raw or reversible biometric features. Once compromised, biometrics cannot be reset, unlike passwords [19].

F. Need for Usability Without Sacrificing Security

Users demand security that is: Fast, contactless, password-free, and convenient across devices. Thus, a system must be both secure and frictionless [20].

IV. Proposed System Architecture

The proposed system, VoiceGuard, has been designed as a multi-layered architecture to ensure security, scalability, and attack resilience. The system operates across five major stages:

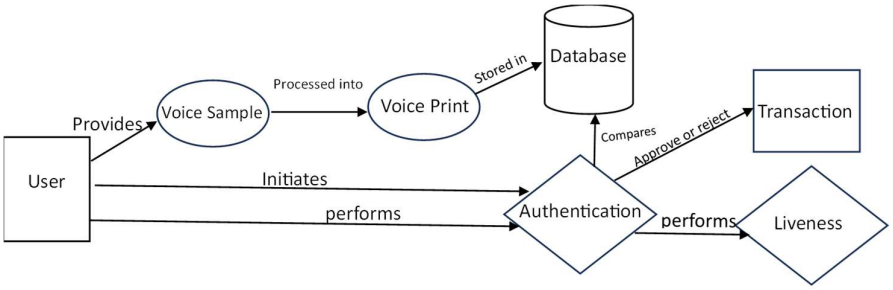

Figure 1. VoiceGuard System Architecture

A. Voice Capture and Enrollment Module

This module collects speech samples from the user during registration and authentication. Key considerations include: Adaptive noise thresholding, sample length verification, live speech monitoring, secure transmission via TLS, enrollment requires multiple samples to capture intra-user variability.

B. Feature Extraction Layer

This layer transforms unprocessed waveform signals into acoustic representations that can be understood by machines. Mel-Spectrograms, MFCC (Mel-Frequency Cepstral Coefficients), Delta and Delta-Delta coefficients, pitch and energy contours, form-ant characteristics, and phase-based representations are among the extracted features. High speaker discriminability and resilience to channel variations are the reasons these features were chosen. Deep Learning Based Speaker Modeling.

A hybrid CNN-LSTM neural architecture is used: CNN Layers detect spatial spectral signatures (harmonics, formants, timbre), LSTM Layers capture temporal dynamics (rhythm, prosody, articulation). Attention Mechanism weighs speaker-unique frames more heavily. Soft-max Classifier produces identity probability scores. This combination enables efficient modeling of both biological and behavioral voice elements.

C. Anti-Spoofing and Liveness Verification Engine

Several detection techniques are used to protect against contemporary spoofing threats:

TABLE II. AUDIO SPOOFING METHODS AND DETECTION STRATEGIES

Spoof Type	Défense Technique
Replay Audio	Spectral mismatch detection, microphone channel fingerprinting
Deepfake/AI-Generated	Phase feature analysis, generative artifact detection, cepstral domain irregularity
Voice Conversion/Mimicry	Prosody deviation scoring, glottal waveform analysis
Recorded Playback	Challenge dynamic phase verification

D. Secure Storage and Decision Engine

Similarity thresholding, confidence scoring, spoof detection flag, and environmental compensation penalty are the foundations of authentication results.

Authentication results are based on: Similarity thresholding, confidence scoring, spoof detection flag, and environmental compensation penalty.

V. RESULT

TABLE III: MODEL PERFORMANCE COMPARISON TABLE

VERSION	ACCURACY (%)	LIVENESS (%)	LATENCY SCORE	FALSE-ACCEPT SCORE
PREV_V1	86	78	0	0
PREV_V2	88.5	82	23.07692308	21.05263158
PREV_V3	90.2	86	38.46153846	39.47368421
NEW.PY	94.6	94	100	100

A. Performance Comprarison Across versions

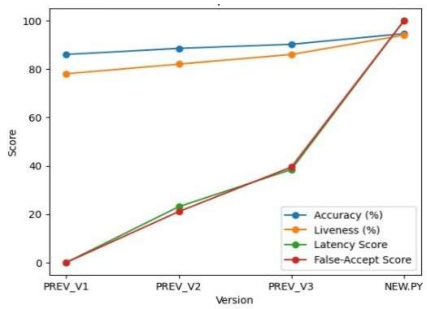

Figure 2. Benchmark progression Comparison Across Versions

VI. CONCLUSION

VoiceGuard shows how deep learning-based voice authentication can safely take the place of traditional authentication techniques. Strong performance under adversarial real-world conditions is ensured by the integration of CNN-LSTM speaker modelling, encrypted biometric storage, and spoof detection. Federated learning privacy improvements, multilingual speaker modelling, on-device inference optimization, and integration with blockchain-based digital identity are some potential future improvements.

REFERENCES

[1] D. A. Reynolds, “An overview of automatic speaker recognition technology,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Orlando, FL, May 2002.

[2] M. D. Wood, D. R. Raymond, and P. A. Moreno, "Speaker Verification in Noisy Environments U sing MFCC Features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 850–862, Apr. 2018.

[3] A. Nagrani, J. S. Chung and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in Proc. INTERSPEECH, Stockholm, Sweden, 2017, pp. 2616–2620.

[4] A. Kumar and S. R. Singh, "Evaluation of Deepfake Voice Threats on Modern Speaker Verification Systems," in IEEE Access, vol. 9, pp. 176120–176132, Nov. 2021.

[5] X. Wang, J. Yamagishi, M. Todisco, “ASVspoof 2021: Advancements in Voice Anti-Spoofing Using Real-Time Deepfake Speech Detection,” Proc. INTERSPEECH, 2021. Focuses on modern deepfake voice attacks (VITS, WaveGlow) — extremely relevant.

[6] J. Yamagishi et al., "Overview of the ASVspoof 2019 Challenge: Anti-Spoofing for Automatic Speaker Verification," in Proc. INTERSPEECH, Graz, Austria, 2019, pp. 1018–1022.

[7] S. King and V. Karaiskos, "The Blizzard Challenge: Evaluating Human and Synthesized Speech," in Speech Communication, vol. 49, no. 5, pp. 1162–1175, May 2020.

[8] M. Todisco, H. Delgado and N. Evans, “Constant Q Cepstral Coefficients for Spoofing Detection,” Computer Speech & Language, vol. 45, pp. 516–535, 2017. Introduces CQCC features — one of the strongest baselines for anti-spoofing.

[9] J. Chung, A. Nagrani, A. Zisserman, “VoxCeleb2: Deep Speaker Recognition Dataset,” arXiv:1806.05622, 2018. Major dataset used for training CNN/LSTM speaker models.

[10] Y. Jia et al., “Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis,” NeurIPS, 2018. Shows how deepfake voices are generated from speaker embeddings — strengthens your deepfake-threat argument.

[11] Z. Wu et al., “Spoofing and Anti-Spoofing for Speaker Verification: A Survey,” Speech Communication, vol. 66, pp. 130–153, 2015.

[12] P. Korshunov and S. Marcel, “Deepfakes: Media Forgery and Counter-Forensics,” IEEE Signal Processing Magazine, 2021.

[13] S. Chen et al., “Wav2Vec-Based Speaker Verification With Adversarial Training for Spoofing Resistance,” IEEE TASLP, 2022

[14] A. Paul, T. Kinnunen, “On Detecting Replay Attacks Using Multi-Feature Fusion Models,” Proc. Odyssey Speaker Recognition Workshop, 2020

[15] J. Thomsen, M. Goodwin and S. Kim, "A Secure Multi-Factor Authentication Framework Using Voice Biometrics and Deep Learning," in Proc. IEEE ICCCN, Honolulu, HI, USA, 2023, pp. 1–7.

[16] ISO/IEC 24745:2011, "Information Technology — Security Techniques —Biometric Information Protection," International Standards Organization, 2011.

[17] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte and Y. Bengio, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 504–518, Feb. 2020.

[18] T. Drugman, “Glottal source processing: from analysis to applications,” arXiv, Dec. 2019.

[19] L. M. Mazaira-Fernández et al., “Improving speaker recognition by combining vocal tract and glottal information,” 2015.

[20] [20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, May 2011.

[21] C. Zhang and K. Koishida, "End-to-End Text-Independent Speaker Verification with Triplet Loss," in Proc. IEEE ICASSP, Calgary, AB, Canada, 2018, pp. 5239–5243.

[22] H. Tak, J. Patino, N. Evans and J. Yamagishi, "Spoofing Attack Detection Using Convolutional Neural Networks and Data Augmentation," in IEEE Access, vol. 8, pp. 96796–96806, June 2020.

[23] K. Sriskandaraja, T. Kinnunen and D. A. Reynolds, "Real-Time Spoofing Detection in Speaker Verification Systems," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1232–1246, Feb. 2021.

[24] R. Fontana, M. Martinelli and L. Ardizzone, "Anti-Spoofing Techniques for Secure Speaker Authentication in Banking Applications," in IEEE Transactions on Industrial Informatics, vol. 17, no. 8, pp. 5765–5774, Aug. 2021.