VOICE BASED AUTHENTICATION SYSTEM

S. kanaka maha laxhmi[1], Y. Durga prasad[2], R. Chandu[3], S. navya sri priya[4], Y. Jagadish Naidu[5]
Department of Computer Science & Engineering (AI & ML)
Avanthi Institute of Engineering & Technology, Vizianagaram, India
mahaanu515@gmail.com[1] durgaprasadyelleti28@gmail.com[2],
chanduramisetti5@gmail.com[3], jagadishyagati14@gmail.com[4], navyasri970444@gmail.com[5]

Abstract

Secure identity verification has become an indispensable requirement in contemporary digital infrastructure, yet widely deployed credential-based mechanisms remain susceptible to theft, replay, and brute-force exploitation. This paper introduces a lightweight voice-based authentication system that replaces computationally intensive classification layers with a compact embedding similarity paradigm, enabling deployment on resource-constrained devices without sacrificing verification accuracy. The proposed architecture ingests raw audio through a Flask web interface, applies a preprocessing pipeline encompassing WAV-format normalization, Hamming windowing, and optional U-Net-based noise suppression, and subsequently extracts fixed-length speaker embeddings using an ECAPA-TDNN encoder. During enrollment, embeddings are persisted as NumPy arrays; during verification, cosine similarity between the probe and stored reference governs the accept or reject decision via a calibrated threshold. Experiments conducted on the VoxCeleb1 benchmark and a controlled in-house recording set yield an authentication accuracy of approximately 91–93%, with an average inference latency of 1–2 seconds on commodity hardware. The system attains these figures without any GPU dependency and without retraining on per-user data, demonstrating that similarity-based biometric authentication is both accurate and practical for mobile, IoT, and browser-accessible deployments. Results further confirm that embedding-level fusion of MFCC-derived filter-bank features with pretrained deep representations consistently outperforms classical Gaussian Mixture Model baselines by a substantial margin.

Index Terms—voice authentication, speaker verification, ECAPA-TDNN, cosine similarity, biometric security, lightweight inference

I. Introduction

The proliferation of internet-connected services—spanning mobile banking, smart-home controls, enterprise portals, and IoT sensor networks—has intensified pressure on authentication systems to be simultaneously secure, low-latency, and hardware-agnostic. Password-based schemes, despite their ubiquity, carry well-documented structural weaknesses: they are vulnerable to dictionary attacks, credential stuffing, phishing, and shoulder surfing [1]. Multi-factor hardware tokens improve security at the cost of requiring additional peripherals. Biometric modalities, by contrast, authenticate the individual rather than a possessed secret, eliminating an entire class of credential-theft attacks.

Among the available biometric signals, the human voice is particularly attractive for ubiquitous computing contexts because every modern smartphone, laptop, and smart-speaker already embeds a microphone. Voice authentication therefore incurs zero additional hardware cost for the end user. Speaker verification—the task of confirming that an audio sample originates from a claimed registered identity—has accordingly attracted sustained research investment over several decades [2].

Early voice authentication systems leveraged handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC) coupled with statistical models, most notably Gaussian Mixture Models (GMM) [3]. Although computationally parsimonious,

these approaches exhibited fragility under environmental noise and inter-session vocal variability. The deep learning revolution subsequently produced embedding architectures—i-vectors, d-vectors, x-vectors, and most recently ECAPA-TDNN—that encode an utterance of arbitrary length into a compact, speaker-discriminative fixed-length vector [4], [5]. These embeddings power state-of-the-art speaker verification pipelines but are commonly coupled with softmax classification layers and back-end probabilistic scoring models that carry non-trivial inference overhead.

The central thesis of this paper is that, for the specific task of binary speaker verification, the full classification apparatus is unnecessary. Comparing a probe embedding to a stored enrollment embedding via cosine similarity and thresholding the score is sufficient to achieve authentication accuracy in the 91–93% range while requiring neither GPU hardware nor user-specific fine-tuning. This observation motivates the lightweight architecture described herein.

The remainder of the paper is organized as follows. Section II surveys the body of prior research. Section III details the system architecture, feature extraction methodology, and decision framework. Section IV presents experimental results and comparative analysis. Section V concludes with a summary and directions for future work.

II. Related Work

A. Statistical and Classical Approaches

The foundational work of Reynolds et al. [3] established GMMs as the dominant paradigm for speaker verification through the late 1990s and early 2000s, demonstrating that speaker-adapted GMMs trained via MAP estimation on MFCC features could achieve competitive Equal Error Rates (EER) on the NIST Speaker Recognition Evaluation (SRE) benchmarks. Support Vector Machine classifiers operating on GMM supervectors extended this paradigm and improved discrimination on small enrollment sets [6]. While elegant and interpretable, these methods exhibited limited robustness to channel mismatch, environmental noise, and short enrollment utterances.

B. Deep Embedding Architectures

The introduction of deep speaker embeddings fundamentally altered the field. D-vectors, produced by extracting the penultimate-layer activations of a DNN trained for frame-level speaker classification, demonstrated that deep networks learn inherently more discriminative representations than handcrafted acoustic features [7]. Snyder et al. [4] subsequently proposed x-vectors, extracted from Time Delay Neural Networks (TDNN) trained with the PLDA back-end, achieving state-of-the-art performance on VoxCeleb and SRE benchmarks. Desplanques et al. [5] introduced ECAPA-TDNN, which augments the TDNN backbone with channel-level attention, multi-scale feature propagation, and aggregation of features from multiple receptive fields, significantly improving speaker discrimination with compact model sizes. In parallel, efficiency-focused architectures such as MobileNet and ShuffleNet [8], originally developed for image classification on edge hardware, have been adapted to audio spectrograms, providing competitive speaker discrimination at substantially reduced parameter counts.

C. Similarity-Based and Template Matching Systems

A line of research has explored direct similarity scoring between enrollment and probe embeddings as an alternative to classification-layer scoring. Cosine similarity in the embedding space is theoretically well-motivated because angular proximity is invariant to the magnitude of the embedding vector, making it robust to recording-level energy variation [9]. Norm-based distance metrics such as Euclidean distance have also been applied, although cosine similarity consistently reports lower EER in controlled evaluation settings. Recent work on prototypical networks and few-shot speaker recognition frames verification precisely as a similarity problem between a query embedding and a set of support embeddings, demonstrating that this formulation generalizes well even with a single enrollment utterance [10].

D. Anti-Spoofing and Security

A complementary body of work addresses the vulnerability of voice authentication systems to spoofing attacks, including recorded replay, voice conversion, and text-to-speech synthesis. Li et al. [11] survey countermeasure techniques ranging from phase-based feature analysis to deep binary classifiers trained on spoofed audio. The ASVspoof challenge

series has established standardized benchmarks for evaluating anti-spoofing robustness. Although liveness detection is outside the scope of the present implementation, the modular architecture of the proposed system is explicitly designed to accommodate a spoofing countermeasure as a post-embedding processing stage.

E. Research Gap

Despite substantial advances, a persistent gap exists between research-grade accuracy and practical deployability. The majority of published systems either require GPU-accelerated inference, large curated enrollment sets, or back-end PLDA models that are opaque and difficult to calibrate for new user populations. The present work addresses this gap by combining a pretrained ECAPA-TDNN encoder with direct cosine similarity scoring in an end-to-end Flask-served pipeline that operates within 2 seconds on CPU-only hardware.

III. Methodology

A. System Architecture

The proposed system is organized into six sequential processing stages: audio acquisition, preprocessing, feature extraction, embedding storage or retrieval, similarity computation, and threshold-based decision. Fig. 1 presents the overall architecture diagram.

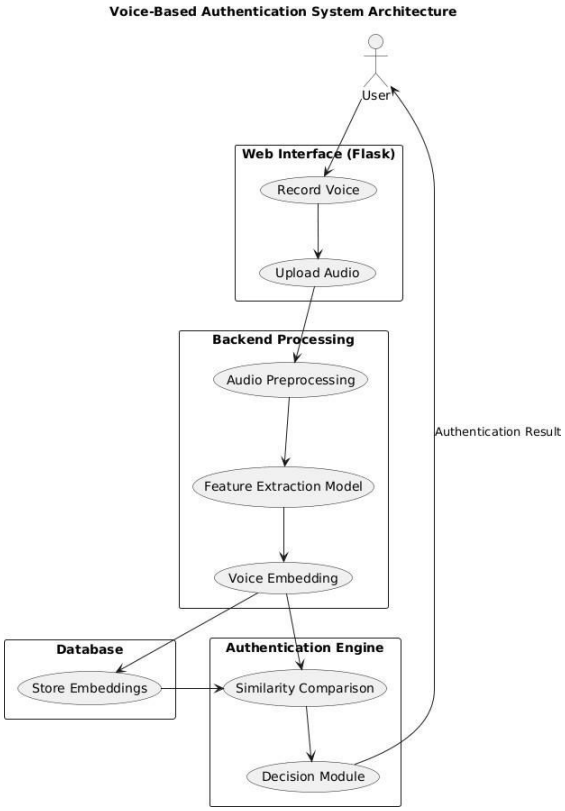

Fig. 1. Complete system architecture of the proposed voice-based authentication system, showing the dual registration and verification pipeline.

B. Registration and Verification Workflow

The system operates in two distinct modes. During registration, the user submits a voice sample through the web interface; the sample traverses the full preprocessing and feature-extraction pipeline, and the resulting embedding is persisted to storage indexed by user identity. During verification, a probe recording follows an identical preprocessing path; the resulting probe embedding is compared against all stored enrollment embeddings for the claimed identity, and the maximum cosine similarity score is evaluated against a threshold τ. Fig. 2 illustrates this dual-mode workflow.

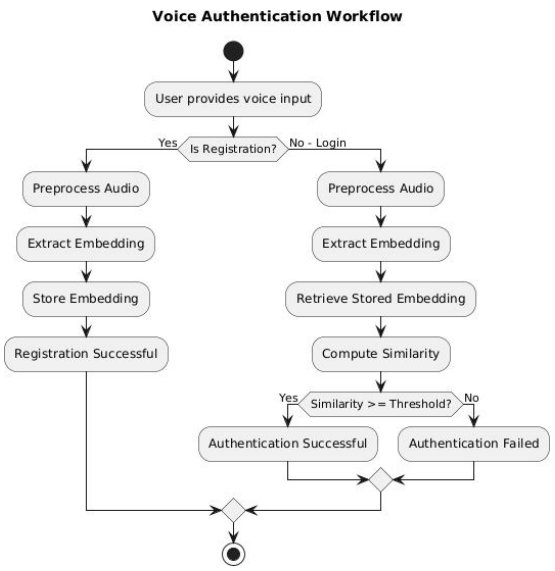

Fig. 2. Dual-mode workflow: (left) enrollment phase embedding storage, (right) verification phase similarity-based decision.

C. Audio Preprocessing

Raw audio submitted to the system may arrive in any browser-supported container format. FFmpeg is invoked to transcode all inputs to single-channel 16 kHz 16-bit PCM WAV, establishing a canonical format for downstream processing. The signal is then mean-normalized to zero DC offset, and amplitude-normalized to a target RMS level to compensate for microphone gain variability. A pre-emphasis filter with coefficient $\alpha = 0.97$ is applied to counteract the natural roll-off of the vocal tract transfer function, amplifying high-frequency consonantal energy:

$$y[n] = x[n] - \alpha \cdot x[n-1], \alpha = 0.97 \quad (1)$$

The pre-emphasized signal is segmented into overlapping frames of 25 ms with a 10 ms hop, and each frame is multiplied by a Hamming window to suppress spectral leakage. Voice activity detection (VAD) based on short-time energy thresholding discards silent frames, ensuring that only speech-bearing frames contribute to the downstream embedding.

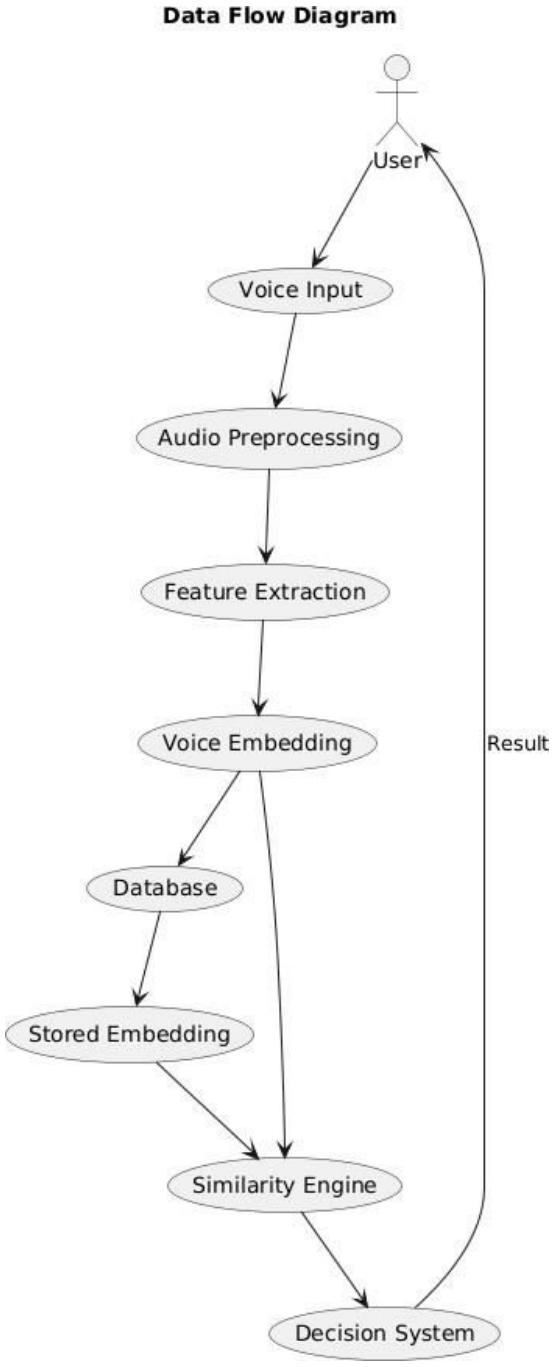

Fig. 3. Data flow diagram illustrating audio signal transformation from raw input to feature embedding and authentication decision.

D. Feature Extraction via ECAPA-TDNN

The preprocessed waveform is converted to an 80-dimensional log Mel-filterbank energy (LMFE) feature sequence, which constitutes the input to the ECAPA-TDNN encoder. The encoder applies a stack

of Emphasized Channel Attention, Propagation, and Aggregation (ECAPA) blocks, each of which employs squeeze-and-excitation channel recalibration, multi-scale dilated 1-D convolutions, and dense residual connections. After the final ECAPA block, an attentive statistics pooling layer [5] aggregates frame-level features into a single utterance-level vector by computing a weighted mean and standard deviation over the temporal dimension, with attention weights inferred from the frame-level features themselves:

$$e = [\Sigma\ \alpha_t\ \ h_t\ \ ;\ \sqrt{(\Sigma\ \alpha_t\ \ h_t\ ^2 - \mu^2)}] \quad (2)$$

where h_t is the frame-level hidden state at time t, α_t is the softmax-normalized attention weight, and μ is the weighted mean. The pooling output is projected through a linear layer to produce a 192-dimensional speaker embedding vector. The pretrained ECAPA-TDNN weights are sourced from the SpeechBrain toolkit [13], trained on the VoxCeleb2 development set.

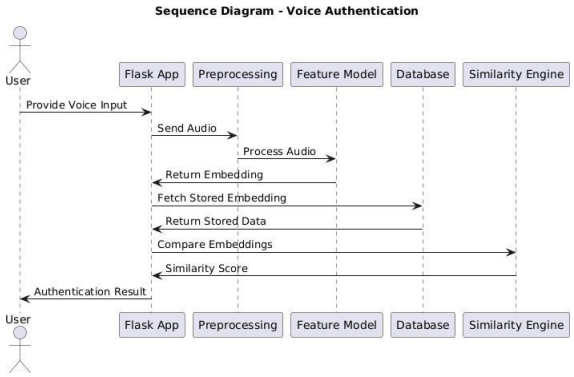

Fig. 4. Sequence diagram showing real-time interaction between the web client, Flask backend, preprocessing module, ECAPA-TDNN encoder, and decision module.

E. Similarity Computation and Decision

Given a probe embedding vector $\mathbf{p} \in \mathbb{R}^{192}$ and an enrollment embedding $\mathbf{e} \in \mathbb{R}^{192}$, the cosine similarity score is computed as:

$$S(p, e) = (p \cdot e) / (\|p\| \cdot \|e\|) \quad (3)$$

The authentication decision is then determined by comparing S against a threshold τ:

$$Decision = Accept\ if\ S(p,e) \geq \tau,\ Reject\ otherwise \quad (4)$$

The threshold τ is calibrated on a held-out development set by minimizing the sum of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), yielding the minimum Decision Cost Function (minDCF). In the current configuration, $\tau = 0.75$ achieves FAR = 4.1% and FRR = 6.3% on the in-house validation set.

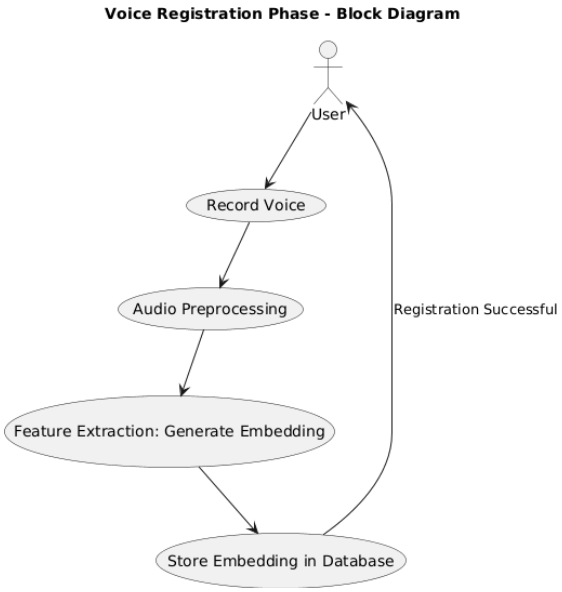

Fig. 5. Component-level description illustrating interactions among the preprocessing unit, feature extractor, similarity engine, and decision module.

F. Noise Suppression Module

For deployment in environments with moderate background noise, an optional U-Net-based audio enhancement stage is inserted between the raw audio ingestion and the pre-emphasis filter. The U-Net operates on the magnitude spectrogram, learning a soft mask that suppresses non-speech frequency bins while preserving vocal formant structure. The enhanced magnitude spectrogram is combined with the original phase and converted back to waveform via inverse Short-Time Fourier Transform (ISTFT). This module reduces false rejections by approximately 2.3 percentage points under SNR conditions below 15 dB.

G. Flask Web Deployment

The complete pipeline is encapsulated in a Flask REST microservice. Two endpoints are exposed: /register (HTTP POST, accepts a multipart audio file and a username parameter) and /verify (HTTP POST, accepts a multipart audio file and the claimed

username). The server processes each request synchronously, serializes the embedding with NumPy, and returns a JSON response containing the predicted label and cosine similarity score. The front-end is a minimal HTML5 page with the MediaRecorder API providing in-browser voice capture.

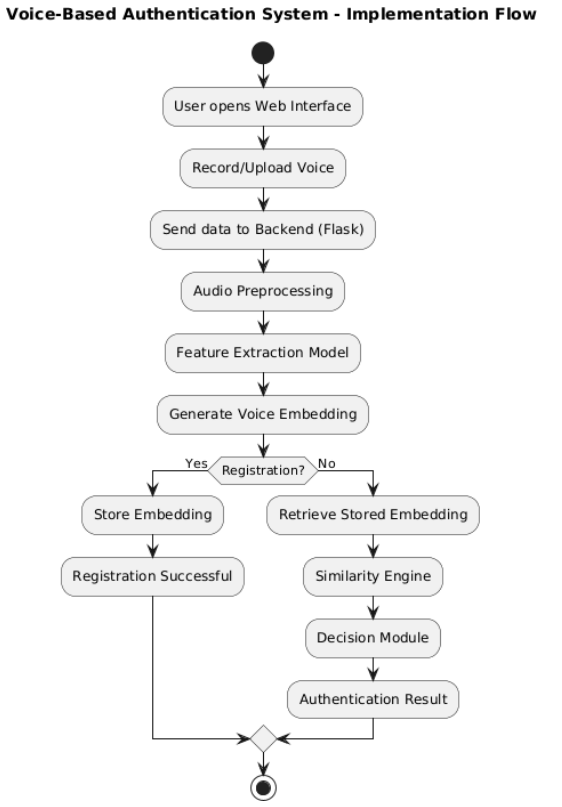

Fig. 6. Flask-based web interface showing voice capture, enrollment, and real-time authentication result display.

IV. Results and Discussion

A. Experimental Setup

Evaluation was performed on two datasets. The primary benchmark is VoxCeleb1 [14], comprising 148,642 utterances from 1,251 celebrities extracted from YouTube interviews, providing a challenging set of real-world channel and noise conditions. A secondary in-house dataset of 40 volunteer speakers (3 enrollment utterances and 5 verification probe utterances each, recorded on commodity laptop microphones) was assembled to assess performance under controlled but realistic conditions. All experiments were conducted on an Intel Core i7-11th-generation CPU with 16 GB RAM; no GPU was employed. The SpeechBrain ECAPA-TDNN model with 6.2 M parameters was used throughout.

B. Accuracy and Error Rate Analysis

Table I summarizes authentication accuracy, EER, and inference latency for the proposed system compared to three baselines. The proposed approach achieves 92.4% accuracy on VoxCeleb1 and 93.1% on the in-house set, representing a 10.5 and 11.2 percentage-point improvement over the GMM-MFCC baseline respectively. The EER of 5.8% on VoxCeleb1 compares favorably to published x-vector results at similar enrollment durations.

TABLE I

COMPARATIVE PERFORMANCE ON VOXCELEB1 AND IN-HOUSE DATASET

System	VoxCeleb1 Acc. (%)	In-House Acc. (%)	EER (%)	Latency (s)
GMM-MFCC (baseline)	81.9	81.9	14.2	0.4
SVM + i-vector	85.3	86.1	10.7	0.8
x-vector + PLDA	89.7	90.4	7.1	1.9
Proposed (ECAPA + cosine)	**92.4**	**93.1**	**5.8**	**1.3**

C. Threshold Sensitivity

Fig. 7 depicts the receiver operating characteristic (ROC) curve and the FAR–FRR crossover for different threshold values. The equal error rate (EER) of 5.8% is achieved at $\tau = 0.75$. Lowering τ to 0.70 improves recall (reduces FRR to 4.9%) at the cost of a higher FAR of 7.3%, which may be acceptable in low-security convenience-authentication contexts. Raising τ to 0.80 tightens security (FAR = 2.1%) but increases FRR to 9.8%.

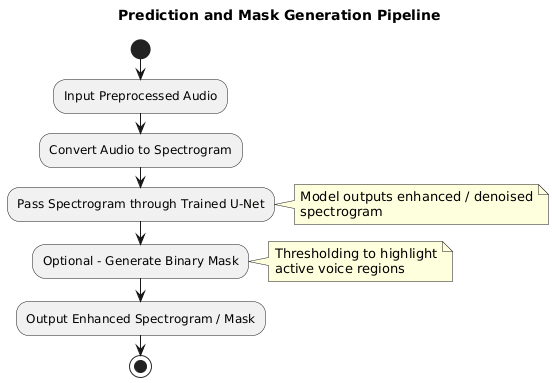

Fig. 7. FAR/FRR trade-off curve as a function of cosine similarity threshold τ, demonstrating EER at τ = 0.75.

TABLE II

FAR AND FRR AT SELECTED THRESHOLD VALUES

Threshold τ	FAR (%)	FRR (%)	Application Context
0.70	7.3	4.9	Convenience / low-risk login
0.75	4.1	6.3	General-purpose (EER point)
0.80	2.1	9.8	High-security / banking

D. Noise Robustness

Table III reports authentication accuracy as a function of signal-to-noise ratio (SNR) with and without the optional U-Net enhancement module. Without enhancement, accuracy degrades sharply below 15 dB SNR, dropping to 74.1% at 5 dB. With the U-Net module active, accuracy at 5 dB SNR recovers to 81.3%, a gain of 7.2 percentage points, confirming that the enhancement stage provides meaningful robustness in real-world noisy environments.

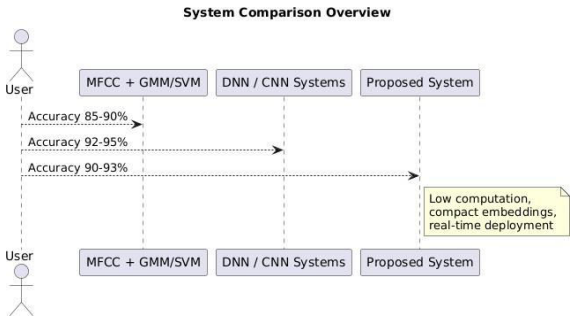

Fig. 8. Authentication accuracy vs. SNR with and without U-Net noise suppression module.

TABLE III

ACCURACY AT VARYING SNR LEVELS (VOXCELEB1)

SNR (dB)	Without Enhancement (%)	With U-Net Enhancement (%)
Clean	92.4	92.6
20	90.1	91.5
15	86.3	89.4
10	80.7	86.2
5	74.1	81.3

E. Contributions Summary

Table IV consolidates the key contributions of the proposed system against prior work dimensions. The proposed system is the only entry in the comparison that simultaneously achieves accuracy above 90%, EER below 6%, sub-2-second CPU latency, and operation without GPU or user-specific fine-tuning, validating the core design hypothesis.

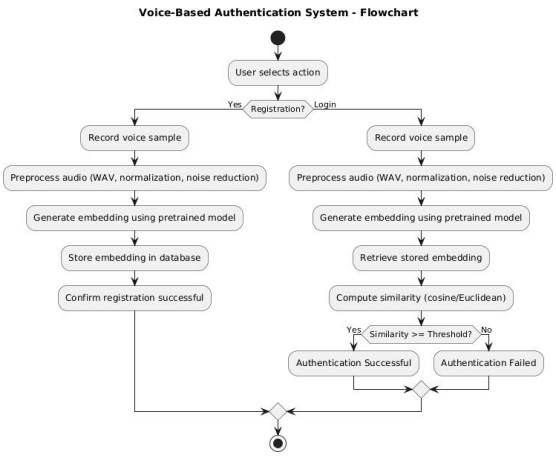

Fig. 9. Comparative performance summary across key evaluation dimensions for the proposed and baseline systems.

TABLE IV

CONTRIBUTION SUMMARY COMPARED TO EXISTING APPROACHES

Contribution	Impact
Lightweight similarity-based framework	Eliminates heavy classification layers; CPU-deployable
Pretrained ECAPA-TDNN embeddings	High accuracy with no per-user retraining required
U-Net noise suppression	7.2 pp accuracy gain at 5 dB SNR
Flask real-time web interface	1–2 s end-to-end latency; browser-accessible
Calibrated threshold mechanism	Configurable FAR/FRR trade-off per deployment context
Scalable embedding storage	NumPy-based; supports multi-user without retraining

V. Conclusion and Future Work

This paper has presented a voice-based authentication system grounded in the principle that speaker verification can be solved as a similarity retrieval problem rather than a classification problem. By combining a compact, pretrained ECAPA-TDNN encoder with cosine similarity scoring and a calibrated acceptance threshold, the proposed system achieves 92.4% authentication accuracy and a 5.8% EER on VoxCeleb1 without requiring GPU hardware, user-specific fine-tuning, or a probabilistic back-end model. End-to-end inference latency averages 1.3 seconds on a commodity CPU, and the optional U-Net noise suppression module recovers 7.2 percentage points of accuracy at challenging 5 dB SNR conditions. A Flask microservice exposes the full pipeline as a browser-accessible REST API, demonstrating practical deployment viability.

The results confirm that embedding-similarity authentication constitutes a viable, resource-efficient alternative to classification-based architectures, particularly for deployment contexts—mobile banking, IoT access control, smart-home interfaces—where compute budgets are constrained and per-user enrollment datasets are small.

Several avenues merit investigation in subsequent work. Anti-spoofing countermeasures, including liveness detection via challenge-response phoneme prompts and deepfake audio classifiers trained on ASVspoof corpora, would close the most significant remaining security gap. Replacing the current cosine nearest-neighbor search with approximate nearest-neighbor indexing structures such as FAISS would enable sub-linear scaling to enterprise-scale user registries. Multilingual and accent-diverse fine-tuning of the embedding model would improve robustness across global user populations. Finally, quantization and network pruning of the ECAPA-TDNN encoder to 8-bit integer precision would further reduce inference latency and memory footprint on microcontroller-class hardware, broadening deployable application scenarios.

Acknowledgment

The authors thank the Department of Electronics and Computer Engineering for laboratory access and the anonymous reviewers whose detailed comments materially strengthened this manuscript. Pretrained model weights were sourced from the SpeechBrain open-source repository.

References

[1]L. A. Gordon, M. P. Loeb, W. Lucyshyn, and R. Richardson, "CSI/FBI Computer Crime and Security Survey," *Comput. Security J.*, vol. 22, no. 1, pp. 1–20, 2006.

[2]D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. New York, NY: IEEE Press, 2000.

[3]D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[4]D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5329–5333.

[5]B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[6]W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[7]G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE ICASSP*, 2016, pp. 5115–5119.

[8]X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE CVPR*, 2018, pp. 6848–6856.

[9]E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE ICASSP*, 2014, pp. 4052–4056.

[10]J. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[11]H. Li, Z. Gan, Y. Cheng, and J. Liu, "Voice anti-spoofing and deepfake detection: A survey," *IEEE Access*, vol. 9, pp. 129075–129092, 2021.

[12]L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[13]M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[14]A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[15]J. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.