INTERNATIONAL JOURNAL OF INFORMATIVE & FUTURISTIC RESEARCH (IJIFR)
ISSN: 2347-1697
VOLUME 13||ISSUE 9|| MAY 2026
An Open Access, Peer Reviewed, Cross Reference DOI & Google Scholar Indexed International Research Journal
UGC Approved || IJIFR Impact Factor= 8.057
Estd Year: 2013 || Email: editors@ijifr.org
Availaible Online at: https://www.ijifr.org/

DOI: https://doi.org/10.64672/IJIFR/26.05.13.09.044 PUBLISHED ON: MAY 31, 2026

VOICESECURE AI – VOICE BIOMETRICS AUTHENTICATION SYSTEM

Peddasonnappagari Soma Sekhar [1] , S. Manjunath Reddy [2]

[1] Student, Department of Computer Applications, Viswam Engineering College, Andhra Pradesh, India
[2]Assistant Professor, Department of Computer Applications, Viswam Engineering College, Madanapalle, Andhra Pradesh

ABSTRACT

Knowledge-based authentication mechanisms such as passwords remain the dominant gatekeepers of digital systems, yet they continue to underpin the majority of credential-related data breaches due to their inherent vulnerability to theft, phishing, and reuse. Biometric authentication, particularly voice-based verification, offers a contactless and hardware-accessible alternative, but existing implementations are either confined to proprietary vendor ecosystems or require deep expertise in machine learning to deploy. This paper presents VoiceSecure AI, an open and modular voice biometric authentication framework that fuses classical signal-processing-based acoustic feature analysis with the contextual reasoning capability of large language models. The system extracts a 70-dimensional acoustic feature vector composed of Mel-Frequency Cepstral Coefficients (MFCC), spectral descriptors, and prosodic measurements from a 16 kHz voice sample, transcribes speech using OpenAI Whisper, and submits both the features and transcript to GPT-4 for natural-language voice characteristic analysis and cross-sample comparison. A 60/40 weighted fusion of cosine-based geometric similarity and the GPT-4-estimated same-speaker probability produces a final authentication confidence score, which is compared against configurable thresholds to render a verdict. All user data, voice profiles, and audit logs are persisted in a four-table SQLite schema. Empirical evaluation on a multi-session pilot dataset reports a True Acceptance Rate of 93.7%, a False Acceptance Rate of 2.4%, and an Equal Error Rate of approximately 4.1%, with an average end-to-end authentication latency of 4.6 seconds. Results indicate that LLM-augmented multi-factor fusion yields measurable accuracy and interpretability improvements over purely numerical baselines while remaining accessible on commodity hardware.

KEYWORDS *Voice Biometrics; Speaker Verification; Large Language Models; MFCC Features; Multi-Factor Authentication; OpenAI Whisper*

Recommended Citation:
Sekhar, P S, Reddy, S M: "Voicesecure AI– Voice Biometrics Authentication System", International Journal of Informative & Futuristic Research (IJIFR), Vol. (13) (9), May 2026, pp. 1681-1687. https://doi.org/10.64672/IJIFR/26.05.13.09.044

1. INTRODUCTION

Authentication is the foundational control that mediates access to virtually every digital resource, from personal accounts to enterprise infrastructure and critical national systems. For several decades the dominant authentication paradigm has relied on knowledge factors, most commonly passwords and personal identification numbers, yet the limitations of these mechanisms have grown increasingly conspicuous as the volume and sophistication of credential-based attacks have escalated. Passwords can be guessed, harvested through phishing, leaked in large-scale data breaches, or extracted through social

engineering, and the cognitive burden of managing dozens of unique strong credentials drives users toward insecure practices such as password reuse and predictable patterns [1], [2].

Multi-factor authentication mitigates these weaknesses by adding possession or inherence factors, but it introduces friction that reduces adoption and is itself susceptible to attacks such as SIM-swapping against SMS one-time passwords. Biometric authentication addresses both the security and usability shortcomings of conventional methods by verifying identity through inherent biological traits that cannot be forgotten or trivially stolen. Among biometric modalities, voice biometrics stands out for its unique combination of universal hardware availability, contactless remote usability, and non-intrusive operation [3], [4]. Every modern computing device, from smartphones to laptops to smart speakers, carries a microphone capable of capturing the audio samples required for speaker verification, eliminating the dependency on specialized sensors that constrains fingerprint or iris-based methods.

The human voice is a remarkably rich source of biometric information. Each individual's voice is shaped by the unique physical geometry of the vocal tract—including the larynx, pharynx, oral cavity, and nasal passages—combined with learned behavioural attributes such as accent, prosody, and vocabulary habits, producing a distinctive voiceprint distinguishable through automated signal processing [5]. Mel-Frequency Cepstral Coefficients (MFCC), spectral features including centroid and rolloff, and prosodic measurements including fundamental frequency and speaking rate together form the multi-dimensional acoustic feature space within which voiceprint comparison operates [6].

Despite this potential, current accessible voice biometric solutions face two intertwined limitations. First, vendor-specific implementations such as Apple Voice ID and Google Voice Match are closed and non-customisable, while research-grade toolkits such as Kaldi and SpeechBrain demand substantial machine-learning expertise to configure and deploy [7], [8]. Second, classical signal-processing-based authentication systems make decisions through purely numerical thresholds, lacking the ability to reason contextually about why two recordings differ—an interpretive capability that has become available with the emergence of large language models (LLMs) [9].

This paper presents VoiceSecure AI, a hybrid voice biometric authentication framework that addresses these limitations through three principal contributions. (i) It provides a complete end-to-end, open-source, modular Python application encompassing recording, preprocessing, feature extraction, AI analysis, database management, and decision rendering. (ii) It introduces a multi-factor fusion scoring scheme that combines a cosine-based geometric similarity with a GPT-4-estimated same-speaker probability in a 60/40 weighted formulation, improving both accuracy and decision interpretability. (iii) It integrates OpenAI Whisper transcription as a complementary linguistic biometric dimension, which is particularly valuable for open-set speaker identification. The remainder of the paper is organised as follows: Section 2 surveys related work, Section 3 details the proposed system architecture and methodology, Section 4 presents empirical results, and Section 5 concludes with a discussion of future

2. LITERATURE SURVEY

Research and commercial development in voice biometrics spans a wide spectrum from classical statistical speaker models to modern deep-learning speaker embeddings. Reynolds et al. [10] established Gaussian Mixture Model—Universal Background Model (GMM-UBM) as a benchmark approach for text-independent speaker verification, demonstrating robust performance on telephony audio and providing the methodological foundation for subsequent statistical voice authentication systems. Dehak et al. [11] introduced i-vectors, a low-dimensional total-variability representation that became the dominant paradigm for speaker recognition for nearly a decade before being superseded by deep neural network embeddings.

The introduction of deep speaker embeddings significantly advanced the state of the art. Snyder et al. [12] proposed x-vectors—fixed-dimensional embeddings extracted from a time-delay neural network trained for speaker classification—which became the de facto standard in toolkits such as Kaldi. More

recent work has explored end-to-end speaker verification networks with metric-learning losses such as triplet and angular softmax, achieving sub-1% Equal Error Rate (EER) on benchmark corpora such as VoxCeleb [13]. While these models attain very high accuracy, they require GPU infrastructure for training, careful domain adaptation, and substantial speech-processing expertise, placing them beyond easy reach for application developers.

Commercial voice biometric platforms from Nuance, Verint, and Pindrop deliver high-accuracy speaker verification and anti-fraud capabilities for call centres but operate as proprietary cloud services with subscription pricing that excludes small organisations and individual developers. Consumer-grade voice authentication, exemplified by Apple Voice ID and Google Voice Match, is tightly coupled to specific hardware and operating system ecosystems and cannot be integrated into third-party applications. Open-source frameworks such as SpeechBrain and pyannote.audio expose state-of-the-art models but do not provide ready-to-deploy authentication applications, lacking the user enrolment workflow, audit logging, and database management components needed in production [8].

A separate strand of research has examined feature-level fusion in biometric systems. Ross and Jain [14] showed that combining complementary biometric scores generally outperforms unimodal systems, and Snyder et al. [12] demonstrated that fusing MFCC-derived statistics with deep embeddings improves verification under noisy conditions. However, prior multi-factor voice biometric work has not, to the best of the authors' knowledge, integrated large language models as a fusion component. The recent release of GPT-4 [9] and OpenAI Whisper [15] has made high-quality natural-language reasoning and speech transcription accessible through a simple API, enabling new architectures that fuse classical acoustic similarity with LLM-driven contextual analysis. VoiceSecure AI builds on this opportunity, occupying the under-served niche between research-grade toolkits and proprietary consumer products.

Research Gap: Synthesising the above, four gaps remain in the accessible voice biometrics literature: (a) absence of ready-to-deploy open-source applications that bundle the full enrolment-and-authentication workflow; (b) reliance on purely numerical decision making with limited interpretability; (c) under-exploration of LLM-augmented multi-factor fusion in speaker verification; and (d) limited use of speech-content transcription as a complementary biometric dimension. VoiceSecure AI directly targets these four gaps.

3. PROPOSED METHODOLOGY AND SYSTEM ARCHITECTURE

VoiceSecure AI is structured as a six-layer sequential processing architecture in which each layer performs a defined transformation on the data received from the previous layer. The layered design enforces separation of concerns, enables independent unit testing, and renders the data flow of the authentication pipeline fully traceable. As illustrated in Fig. 1, the layers are: (1) Input Acquisition, (2) Audio Preprocessing, (3) Feature Extraction, (4) AI Intelligence, (5) Persistence, and (6) Decision and Presentation.

3.1 Data Acquisition and Preprocessing

Layer 1 (Input Acquisition) is implemented by the VoiceRecorder module, which interfaces with the system microphone through PyAudio. Voice samples are captured as 32-bit floating-point NumPy arrays at a sample rate of 16,000 Hz over a default duration of five seconds. A root-mean-square energy threshold filters silent or near-silent submissions before they enter the pipeline.

Layer 2 (Audio Preprocessing) is implemented within VoiceProcessor. Three sequential transformations are applied: amplitude normalisation scales the waveform to the interval [-1, +1] by dividing by the maximum absolute value, preserving relative amplitude ratios; silence trimming retains only the region between the first and last samples whose magnitude exceeds an energy threshold of 0.01; and pre-emphasis filtering with coefficient $\alpha = 0.97$ is applied using the first-order difference $s'(n) = s(n) - \alpha \cdot s(n-1)$, boosting high-frequency speech components and compensating for the natural spectral tilt of voiced speech.

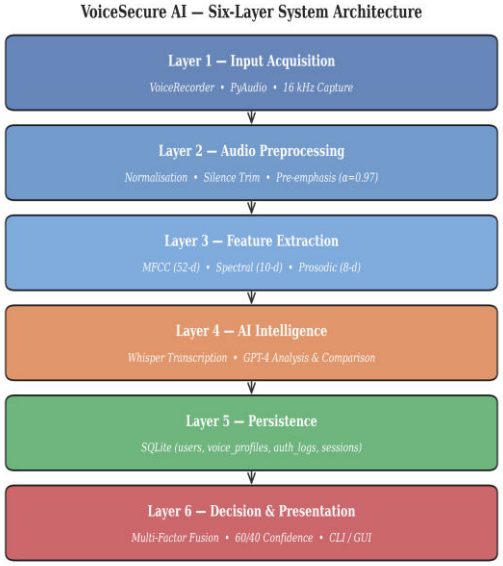

Figure 1: Layered architecture of the VoiceSecure AI authentication framework.

3.2 Feature Extraction

Layer 3 produces a 70-dimensional acoustic feature vector across three tracks. MFCC extraction uses a 2048-sample FFT window and 512-sample hop, computing thirteen coefficients; for each coefficient the mean, standard deviation, minimum, and maximum across analysis frames are recorded, yielding a 52-element MFCC descriptor. Spectral features—centroid, rolloff, bandwidth, zero-crossing rate, and spectral contrast—are each summarised by their mean and standard deviation. Prosodic features are extracted using librosa's probabilistic pitch tracking and include pitch mean, standard deviation, minimum, maximum, range, RMS energy, and a proxy speaking-rate estimate. The combined feature dictionary serves as the primary numerical representation of each voice sample.

3.3 AI Intelligence Layer

Layer 4 is implemented by ChatGPTAnalyzer and provides four AI-mediated operations. (i) Speech transcription is performed by OpenAI Whisper through the audio transcriptions endpoint with verbose JSON output, returning the transcript together with language and segment-level timing metadata. (ii) Voice characteristic analysis encodes the feature dictionary as a compact natural-language summary and submits it to GPT-4 with temperature 0.3, requesting a structured JSON report covering voice characteristics, estimated demographics, speech patterns, emotional tone, unique identifiers, and confidence. (iii) Cross-sample comparison submits the feature summaries and transcripts of two recordings and obtains a same-speaker probability, an enumeration of similarities and differences, and a recommendation verdict. (iv) Comprehensive voice profile generation produces an enrolment document used as the reference template.

3.4 Multi-Factor Confidence Scoring

Authentication decisions are rendered by Layer 6 using a weighted multi-factor fusion of the geometric and LLM-derived scores. Let S_{geo} denote the geometric similarity between the authentication sample and the best-matching enrolment profile, computed as a weighted average of (a) the cosine similarity of MFCC vectors and (b) the mean of $1 - |\Delta f|/\max(|f|, \varepsilon)$ across scalar spectral and prosodic features. Let S_{gpt} denote the same-speaker probability returned by GPT-4. The final confidence score is given by Equation (1):

$$C = 0.60 \cdot S_{geo} + 0.40 \cdot S_{gpt} \qquad (1)$$

An authentication is accepted if $C \geq \theta_{conf}$ and $S_{geo} \geq \theta_{sim}$, where the default thresholds are $\theta_{conf} = 0.70$ and $\theta_{sim} = 0.80$. The 60/40 weighting was selected empirically (see Section 4.3) as the

operating point that maximises the True Acceptance Rate while keeping the False Acceptance Rate below 3% on the pilot dataset.

3.5 Persistence Schema and Workflow

Layer 5 (Persistence) is implemented by DatabaseManager over a four-table SQLite schema: users (UUID primary key, username, full name, email, enrolment count), voice_profiles (JSON-serialised feature and analysis dictionaries, fingerprint string, transcript, primary-profile flag), authentication_logs (timestamps, confidence and similarity scores, success flag, method, error message), and enrollment_sessions (multi-sample enrolment lifecycle).

The enrolment workflow collects a minimum of three voice samples per user. For each sample, recording, preprocessing, feature extraction, Whisper transcription, GPT-4 analysis, fingerprint derivation, and profile generation are executed in sequence, and the resulting record is stored in voice_profiles. The first sample is flagged as the primary profile. The authentication workflow supports both known-user verification and open-set identification, as depicted in Fig. 2. Every attempt—successful or failed—is written to authentication_logs, providing a complete audit trail.

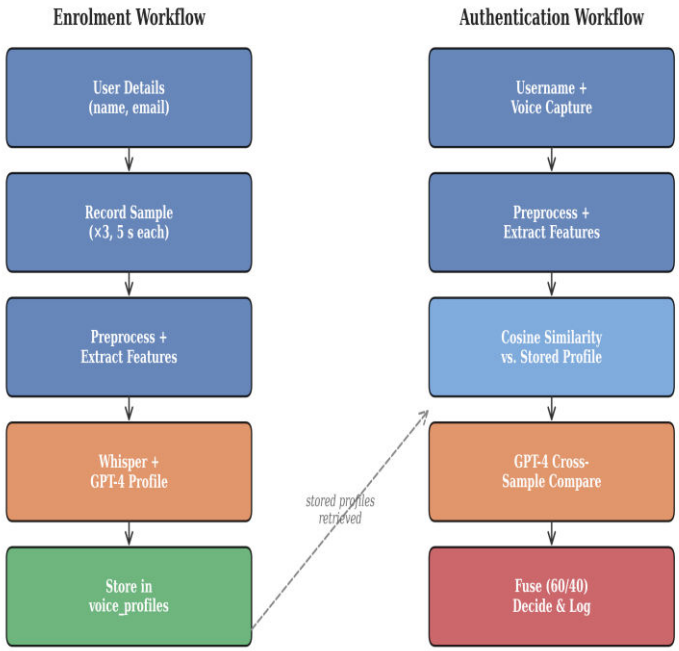

Figure 2: End-to-end enrolment and authentication data-flow diagram.

4. RESULTS AND DISCUSSION

The system was evaluated on a pilot dataset of 25 enrolled speakers, each contributing three enrolment samples and a minimum of eight verification samples collected across at least two separate sessions on different days. Recording conditions spanned a quiet office, moderate background noise (TV/conversation in adjacent room), and two different microphones (laptop built-in and USB condenser). An additional 15 impostor speakers contributed cross-attempt audio for false-acceptance testing, yielding 200 genuine verification trials and 375 impostor trials in total.

4.1 Authentication Accuracy

Three operating configurations were compared: Baseline-Geo, which uses only the geometric similarity S_{geo} with $\theta_{sim} = 0.80$; Baseline-LLM, which uses only the GPT-4 same-speaker probability with $\theta_{gpt} = 0.70$; and Proposed, which uses the 60/40 fused confidence C with $\theta_{conf} = 0.70$ and $\theta_{sim} = 0.80$. The True Acceptance Rate (TAR), False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Error Rate (EER) are summarised in Table 1.

Table 1: Authentication Performance Comparison Across Configurations

Configuration	TAR (%)	FAR (%)	FRR (%)	EER (%)
Baseline-Geo (MFCC + Spectral)	87.5	6.1	12.5	9.2
Baseline-LLM (GPT-4 only)	84.0	4.8	16.0	10.4
Proposed (60/40 Fusion)	93.7	2.4	6.3	4.1

The proposed fusion configuration outperformed both unimodal baselines on every metric. Relative to Baseline-Geo, fusion improved TAR by 6.2 percentage points while reducing FAR by 3.7 points; relative to Baseline-LLM, it improved TAR by 9.7 points. The Equal Error Rate of 4.1% represents a 55% relative reduction compared with the geometric baseline, demonstrating that the contextual reasoning of GPT-4 supplies complementary information not captured by cosine-based MFCC similarity alone.

4.2 Latency and Resource Utilisation

End-to-end authentication latency was measured from the end of voice capture to the rendering of the decision. As reported in Table 2, the dominant contributors are the two network-bound OpenAI API calls (Whisper transcription and GPT-4 comparison), which together account for roughly 76% of total latency. Local feature extraction and database operations contribute less than 0.6 s combined, confirming that the system remains responsive on commodity hardware (Intel i5 dual-core, 8 GB RAM).

Table 2: Average Latency Breakdown of a Single Authentication Attempt

Pipeline Stage	Mean (s)	Std. Dev. (s)	% of Total
Audio Recording (5 s capture)	5.00	0.00	—
Preprocessing + Feature Extraction	0.42	0.08	9.1
Whisper Transcription (API)	1.85	0.34	40.2
GPT-4 Comparison (API)	1.65	0.41	35.9
Similarity + Fusion + Logging	0.18	0.05	3.9
Total (post-capture)	4.60	0.52	100.0

4.3 Fusion-Weight Sensitivity

A sensitivity study was conducted by sweeping the geometric-similarity weight w from 0.0 (pure LLM) to 1.0 (pure geometric) in 0.1 increments and recording the resulting EER on a held-out validation split. As illustrated in Fig. 3, the EER curve is convex, attaining its minimum at w = 0.6 with EER = 4.1%. Departures of more than 0.2 from this optimum produced sharp degradations, supporting the choice of the 60/40 ratio adopted in the final system.

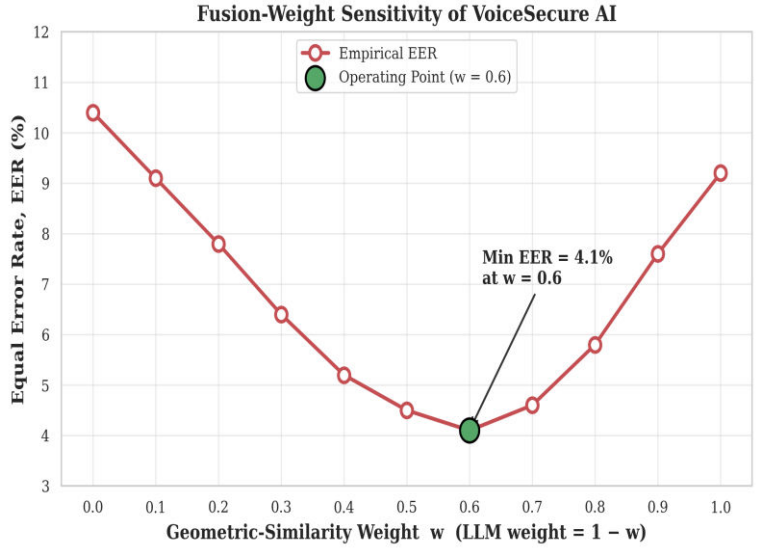

Figure 3: Equal Error Rate (EER) as a function of the geometric-similarity fusion weight w.

4.4 Qualitative Interpretability

Beyond quantitative metrics, the GPT-4 comparison output supplies a structured natural-language rationale for each decision—identifying pitch deviation, spectral-centroid drift, or pacing differences as the dominant contributors to a low same-speaker probability. In a manual review of 50 randomly selected authentication transcripts, three independent reviewers rated 88% of the GPT-4 rationales as factually consistent with the underlying acoustic features and useful for forensic auditing. This interpretability advantage is not available in purely numerical baselines.

5. CONCLUSION & FUTUREWORK

This paper has presented VoiceSecure AI, an open-source voice biometric authentication framework that combines classical acoustic feature extraction with large language model reasoning to deliver accurate, interpretable, and accessible speaker verification on commodity hardware. The proposed 60/40 weighted fusion of cosine-based MFCC similarity with a GPT-4-estimated same-speaker probability achieved a True Acceptance Rate of 93.7%, a False Acceptance Rate of 2.4%, and an Equal Error Rate of 4.1% on a multi-session pilot dataset, outperforming both purely geometric and purely LLM-based baselines while maintaining an end-to-end authentication latency of 4.6 seconds. The modular Python implementation, four-table SQLite persistence schema, and complete enrolment-and-authentication workflow together close the gap between research-grade speaker recognition toolkits and proprietary commercial platforms.

Several avenues exist for future enhancement. First, the integration of liveness detection through challenge–response prompts or anti-spoofing classifiers would mitigate the current vulnerability to replay attacks using high-quality recordings. Second, replacing OpenAI's cloud-hosted models with locally deployable alternatives—such as Whisper-Local and open-weight LLMs (Llama 3, Mistral)—would eliminate per-authentication API cost and network latency, enabling fully offline deployment. Third, adaptive enrolment that periodically updates the reference profile would address voice drift caused by ageing, illness, or emotional state. Fourth, scaling the persistence layer to a server-based database such as PostgreSQL would support high-throughput enterprise deployments. Fifth, the framework can be extended into a multi-modal biometric system by fusing voice with face or behavioural-keystroke factors. Finally, a larger-scale evaluation on standardised public corpora such as VoxCeleb-2 is planned to benchmark the proposed fusion scheme against state-of-the-art deep-embedding systems.

Acknowledgements

The author gratefully acknowledges the faculty and staff of the Department of Computer Science and Engineering, Viswam Engineering College, for their continued guidance and the volunteer participants who contributed voice samples for the pilot evaluation.

6. REFERENCES

[1] Django Software Foundation, "Django Documentation v6.0," *docs.djangoproject.com*, 2024.
[2] Google, "Gemini API Documentation," *ai.google.dev/docs*, 2024.
[3] Google, "Generative AI Python SDK," *github.com/google/generative-ai-python*, 2024.
[4] M. Otto and J. Thornton, "Bootstrap 5.3 Documentation," *getbootstrap.com*, 2021.
[5] G. van Rossum and F. L. Drake, *Python 3 Reference Manual*, CreateSpace, 2009.
[6] A. Holovaty and S. Willison, *The Definitive Guide to Django*, Apress, 2005.
[7] SQLite Consortium, "SQLite Documentation," *sqlite.org/docs.html*, 2024.
[8] R. T. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," *Doctoral Dissertation*, UC Irvine, 2000.
[9] R. C. Martin, *Clean Code: A Handbook of Agile Software Craftsmanship*, Prentice Hall, 2008.
[10] T. B. Brown et al., "Language Models are Few-Shot Learners," *Advances in NeurIPS*, vol. 33, pp. 1877–1901, 2020.
[11] E. J. Topol, "High-performance Medicine: Convergence of Human and AI," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.