IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

"VoiceGuard: An Explainable AI Framework for Synthetic Voice Authentication"

Ms. Divya Arun Patil[1], Prof. Shivam Limbhare[2], Prof. Manoj Vasant Nikum[3]

Research Scholar, Master of Computer Applications, Shri JRIT, Dondaicha, KBCNMU Jalgaon, Maharashtra, India[1]

Assistant Professor, Master of Computer Applications, Shri JRIT, Dondaicha, KBCNMU Jalgaon, Maharashtra, India[2]

HOD, Master of Computer Applications, Shri JRIT, Dondaicha, KBCNMU Jalgaon, Maharashtra, India[3]

Abstract: Advances in neural text-to-speech and voice conversion have made synthetic speech increasingly convincing, raising concerns about impersonation and the reliability of voice-based verification. VoiceGuard is proposed as an explainable artificial intelligence framework for assessing the authenticity of recorded speech. The design combines standardized audio preprocessing, Mel-frequency cepstral coefficients (MFCCs), log-Mel spectrogram representations, and a convolutional neural network (CNN) classifier. A separate explanation layer is intended to highlight influential time-frequency regions and provide interpretable evidence alongside a real-versus-synthetic prediction. This paper defines the proposed system architecture, reproducible evaluation protocol, database organization, and analysis criteria. Evaluation on established anti-spoofing benchmarks is planned; no trained-model performance or experimental accuracy is claimed at this stage. The resulting specification provides a practical foundation for subsequent implementation and validation of an auditable synthetic-voice screening application.

Keywords: VoiceGuard, Synthetic Speech Detection, Voice Cloning, Text-to-Speech, MFCC, Mel-Spectrogram, CNN, Explainable AI, ASVspoof.

I. INTRODUCTION

Synthetic speech systems can reproduce speaker characteristics with increasingly natural rhythm, intonation, and timbre. Although these technologies support accessibility and creative applications, they also complicate decisions about whether a recording originated from a person or a speech-generation pipeline. Audio-only verification is particularly challenging because compression, background noise, and recording equipment can alter the same acoustic properties that detection models use.

VoiceGuard addresses this problem as an audio-authenticity screening tool rather than a replacement for identity verification. The framework aims to distinguish bona fide speech from spoofed speech and to explain the evidence contributing to a classification. Its intended workflow includes recording upload, feature extraction, inference, confidence reporting, and an interpretable time-frequency visualization. The central research question is whether a compact acoustic classifier with transparent explanations can offer useful synthetic-speech screening across varied recording conditions.

II. LITERATURE SURVEY

1. **Automatic Speaker Verification Spoofing:** The ASVspoof challenge series established common datasets and evaluation practices for detecting replay, speech synthesis, and voice conversion attacks. Its logical-access track is especially relevant to synthetic-speech countermeasures.
2. **Acoustic Representations:** MFCCs summarize spectral-envelope information, while log-Mel spectrograms retain a time-frequency map suitable for convolutional feature learning. These representations offer complementary perspectives but may respond differently to noise and codecs.
3. **Deep Learning for Anti-Spoofing:** CNN-based classifiers can learn local spectrotemporal patterns associated with generated speech. Benchmark studies also show that performance may change considerably between known and previously unseen synthesis methods.
4. **Explainable Classification:** Model explanations such as SHAP, LIME, and gradient-based activation maps can help users inspect which features influenced a prediction. Such explanations require validation: a visually plausible heatmap is not itself proof of authenticity.

 This work is licensed under a Creative Commons Attribution 4.0 International License

III. RESEARCH METHODOLOGY

The proposed study follows a staged, reproducible development plan: dataset preparation; acoustic preprocessing; baseline feature extraction; model training; explanation generation; and held-out evaluation. Training and test partitions must be separated according to official benchmark protocols to reduce leakage from speakers, recordings, and synthesis systems.

3.1 Architectural Design

A browser interface will accept a short audio file and send it to a Python-based API. The backend will validate the file, convert it to a standard mono sample rate, normalize signal amplitude as appropriate, and derive MFCC and log-Mel features. A PyTorch CNN will estimate the probability of bona fide or synthetic speech. An explanation service will produce a time-frequency attribution view, and an SQLite store will retain analysis metadata without requiring permanent storage of raw audio.

3.2 Core Analytical Modules

1. **Audio Preprocessing:** Supported recordings are decoded, resampled, trimmed or padded to a consistent duration, and checked for silence or corruption. Preprocessing parameters will be fixed before evaluation.
2. **Spectral Feature Extraction:** Librosa will compute MFCC vectors and log-Mel spectrograms using documented window, hop, and frequency settings. Spectrograms will be standardized using statistics derived from the training split only.
3. **Deep Classification:** A compact CNN with convolution, pooling, and fully connected layers will classify the input. Class imbalance and calibration will be considered during training and validation.
4. **Explainability:** Gradient-based class activation mapping or SHAP will identify influential spectrogram regions. Explanations will be compared against model perturbation tests to assess their faithfulness.

3.3 Operational Workflow & Database Schema

SQLite Table	Core Schema Elements	Functional Role in VoiceGuard
analysis_jobs	id, filename, created_at, status	Tracks submitted audio analyses.
audio_features	id, job_id, feature_version, duration	Records extraction metadata and settings.
predictions	id, job_id, predicted_label, score, model_version	Stores classifier outputs and model provenance.
explanations	id, prediction_id, method, artifact_path	Links interpretable plots to predictions.

IV. EVALUATION PLAN

VoiceGuard has not yet been implemented or experimentally evaluated. The following scenarios and metrics are prespecified so that subsequent development can be compared against the design without retroactively choosing favorable results.

4.1 Planned Experimental Scenarios

Test Scenario	Component	Evaluation Criterion
Bona fide vs. synthetic speech	MFCC/log-Mel CNN	Accuracy, precision, recall, F1, EER on held-out data
Unseen synthesis methods	Cross-condition test partition	Generalization gap versus in-domain evaluation
Codec and noise robustness	Audio preprocessing and model	Change in EER under controlled distortions
Explanation fidelity	Attribution/occlusion analysis	Prediction change when salient regions are masked
Application integration	Upload, API, database	Valid response, error handling, reproducible records

 This work is licensed under a Creative Commons Attribution 4.0 International License

4.2 Metrics and Reporting

Classification performance will be summarized with a confusion matrix, precision, recall, macro-F1, and equal error rate (EER). EER is the operating point at which false-accept and false-reject rates coincide. The evaluation will document dataset splits, class counts, preprocessing parameters, random seeds, model checkpoints, and confidence intervals where feasible. No numerical results are included because experiments have not yet been run.

V. DISCUSSION AND ANALYSIS

1. **Practicality:** MFCCs and log-Mel features can be computed with widely available audio libraries, making the framework suitable for a student-built prototype. The actual inference speed and hardware requirements remain to be measured.
2. **Explainability:** Spectrogram attributions may make model decisions easier to inspect, but their reliability should be tested with feature perturbations and human review rather than inferred from appearance alone.
3. **Limitations:** Synthetic speech detectors can fail on novel generators, noisy clips, short utterances, and heavily compressed recordings. A VoiceGuard prediction should therefore be presented as decision support, not conclusive forensic proof.

Data Flow & System Architecture Diagram

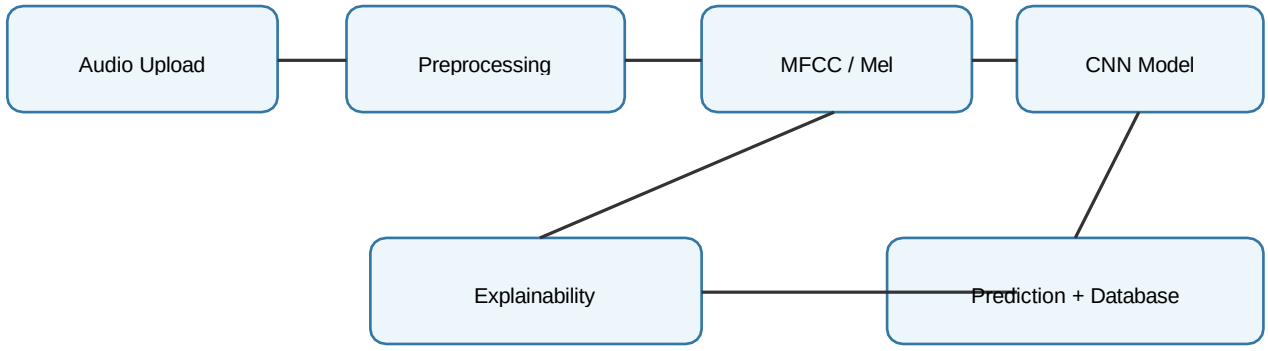


Fig. 1 VoiceGuard – Proposed Synthetic Voice Authentication Architecture & Data Flow

VI. CONCLUSION

This paper specifies VoiceGuard as an explainable framework for detecting synthetic and cloned speech using acoustic features and a CNN-based classifier. The planned architecture combines reproducible signal processing, model-versioned predictions, and interpretable spectrogram attribution. Its research contribution will depend on subsequent implementation and independently verifiable benchmark evaluation. Future work includes cross-dataset robustness testing, confidence calibration, and comparison with established anti-spoofing baselines.

REFERENCES

[1] M. Todisco et al., “ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection,” Proc. Interspeech, 2019.
[2] J. Yamagishi et al., “ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection,” ASVspoof 2021 Workshop, 2021.
[3] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” Proc. Python in Science Conference, 2015.
[4] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” NeurIPS, 2019.
[5] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” NeurIPS, 2017.
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” KDD, 2016.
[7] R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” ICCV, 2017.