International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Voiceguard – Ai-Based Voice Authenticity Detection System

Dr. C. Saravanabhavan [1], Akhil R[2]

[1] Professor Department of Computer Science and Engineering Kongunadu College of Engineering and Technology Tiruchirappalli, India

[2] Bachelor of Engineering Department of Computer Science and Engineering Kongunadu College of Engineering and Technology Tiruchirappalli, India

Abstract- — Recent advances in deep learning have en-abled highly realistic synthetic speech, creating serious risks such as impersonation, fraud, and misuse of voice-based authentication systems. Detecting AI-generated speech is increasingly difficult because modern text-to-speech and voice conversion models can closely imitate human prosody and timbre across languages. This paper proposes VoiceGuard, a hybrid deep learning framework that combines complementary spectral and temporal rep-resentations for deepfake voice detection. A Convolutional Neural Network (CNN) branch learns frequency-domain artifacts from spectrograms, while a CNN-GRU branch models temporal inconsistencies from acoustic descriptors. An attention-based fusion mechanism adaptively weights branch outputs to improve discriminative power. The framework is evaluated on benchmark datasets and cross-lingual settings, and it improves performance compared to single-representation approaches while remaining compu-tationally practical for real-world deployment.

Index Terms—Deepfake voice detection, voice authen-tication, synthetic speech detection, convolutional neu-ral network, gated recurrent unit, spectrogram analysis, attention-based fusion, cross-lingual evaluation.

I. INTRODUCTION

Deep learning has significantly improved speech synthe-sis quality through modern text-to-speech (TTS) and voice conversion (VC) systems. Neural architectures now generate synthetic voices that closely match human speech in pitch, flu-ency, and speaking style. Although these technologies benefit applications such as accessibility, virtual assistants, and media production, they also introduce severe security concerns.

The realism of AI-generated speech has increased the risk of impersonation attacks, social engineering, financial fraud, and misinformation. As voice interfaces and biometric authentica-tion become widespread, there is an urgent need for reliable deepfake voice detection.

This challenge is even more critical in low-resource and cross-lingual settings, where many existing systems trained on English-centric datasets may not generalize well to different phonetic structures and speaking styles.

Detecting synthetic speech remains challenging for three reasons. First, handcrafted acoustic features are often insuffi-cient against modern generative models. Second, frequency-focused models may miss temporal irregularities in speech dynamics. Third, high-capacity transformer-based solutions can be computationally expensive for real-time or resource-constrained use.

To address these issues, this paper presents VoiceGuard, a hybrid framework that jointly models spectral and temporal cues. The proposed system combines a CNN-based spectro-gram branch and a CNN-GRU temporal branch, followed by attention-based fusion for adaptive feature integration.

The main contributions are summarized as follows:

- A hybrid deep learning framework that combines spectral and temporal speech representations.
- An attention-based fusion strategy that improves classifi-cation by emphasizing discriminative features.
- A practical architecture that balances detection perfor-mance and computational efficiency.
- Cross-lingual evaluation to assess generalization beyond a single-language setting.

II. RELATED WORK

A. Human Voice Synthesis

Recent TTS and VC systems such as WaveNet, Tacotron, and FastSpeech generate natural-sounding audio from linguis-tic or speaker representations. These systems typically rely on

intermediate acoustic representations and neural vocoders for waveform reconstruction.

B. Neural Vocoders and Artifact Formation
Neural vocoders, including autoregressive, GAN-based, and diffusion-based models, improve perceptual quality but may introduce subtle synthesis artifacts. These artifacts are often difficult for human listeners to detect yet can provide useful evidence for machine-based spoof detection.

C. Deepfake Audio Detection Techniques
Traditional methods used handcrafted features such as MFCC and LFCC with classical classifiers. More recent approaches use CNNs for spectrogram analysis and RNN

variants (including LSTM and GRU) for temporal modeling. Hybrid CNN-RNN methods generally perform better than single-branch models because they capture complementary information.

D. Self-Supervised and Transformer-Based Models
Self-supervised models such as wav2vec 2.0, HuBERT, WavLM, and XLS-R learn powerful speech representations and can generalize well. However, these models often require substantial compute and may be harder to deploy in real-time settings.

E Research Gap
Most existing methods emphasize a single representation domain or require high computational overhead. In addi-tion, many models are evaluated primarily on English-centric datasets, limiting confidence in cross-lingual performance. These limitations motivate a balanced, multi-representation architecture with practical efficiency.

III. METHODOLOGY

A. Problem Formulation
Let x denote an input audio signal and $y \in \{0, 1\}$ denote its label, where $y = 0$ represents real speech and $y = 1$ represents synthetic speech. The objective is to learn a mapping:
$\hat{y} = F\theta(x)$ (1)
where $F\theta$ is a parameterized classifier.

B. System Overview
VoiceGuard follows a multi-stage pipeline: preprocessing, dual-branch feature extraction, hybrid representation learning, attention-based fusion, and binary classification. The spectral branch and temporal branch operate in parallel to extract complementary evidence.

C. Audio Preprocessing
Input audio is standardized through resampling (16 kHz), normalization, silence trimming, and noise reduction. These steps reduce variability unrelated to spoofing artifacts and improve model stability.

D.Feature Extraction
The spectral branch computes mel-spectrogram and STFT representations to capture frequency-domain cues. The tem-poral branch extracts acoustic descriptors such as MFCC, chroma, and spectral contrast to preserve speech dynamics over time.

E. Hybrid Model Architecture
Spectral features are processed by a CNN to detect local frequency patterns and synthesis distortions. Temporal descrip-tors are processed by a CNN-GRU stack, where convolutional layers learn local structure and GRU layers capture sequence-level dependencies.

F. Attention-Based Feature Fusion
Instead of direct concatenation, branch embeddings are fused using attention weights that prioritize informative com-ponents. This adaptive fusion improves robustness by leverag-ing the most discriminative spectral-temporal cues per sample.

G. Classification Objective
The fused representation is passed through fully connected layers and a sigmoid output for binary prediction. The model is trained with binary cross-entropy:
$L = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$ (2)

IV. SYSTEM ARCHITECTURE AND DATA FLOW

A. Overall Architecture

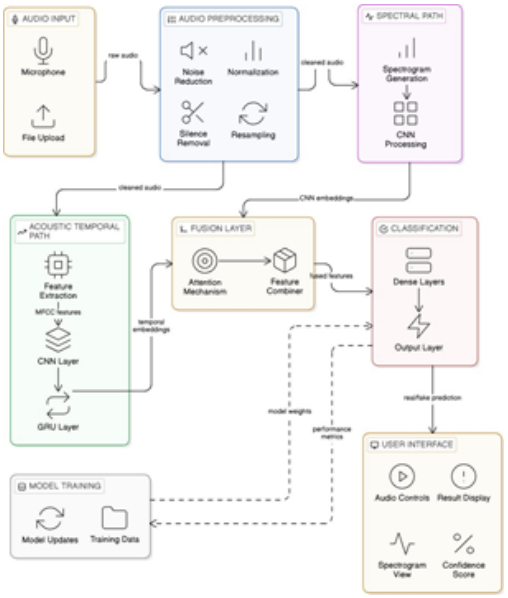


Fig. 1. Overall architecture of the VoiceGuard framework.

VoiceGuard follows a multi-stage pipeline that transforms raw speech into a robust authenticity decision through preprocessing, dual-branch representation learning, and fusion-driven classification.

Figure 1 presents the complete VoiceGuard pipeline. The system starts with audio input from file upload or microphone recording, followed by preprocessing steps such as noise reduction, normalization, silence removal, and resampling. The cleaned signal is then analyzed through two parallel branches: a spectral branch for spectrogram-based CNN processing and an acoustic temporal branch for feature extraction with CNN-GRU modeling.

The branch outputs are combined in an attention-based fusion layer and passed to dense classification layers to predict real or fake speech. The final prediction is shown in the

user interface along with confidence-oriented feedback. The architecture also includes a training feedback path, where model weights are updated using performance metrics to keep the detector effective against evolving spoofing patterns.

B. Level 1 Data-Flow Overview

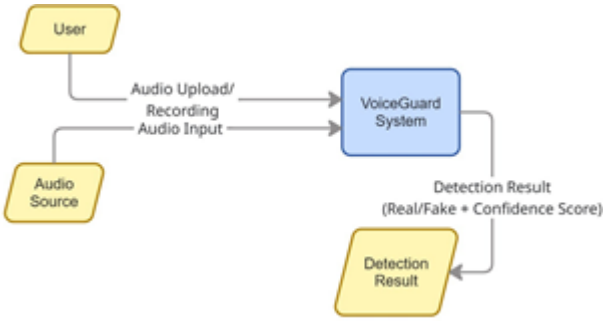


Fig. 2. Level 1 data-flow diagram showing the high-level system pipeline.

Figure 2 gives a high-level data-flow view of the system. Audio from the user and source channels enters the Voice-Guard process boundary as input data. At this level, internal computations are abstracted, and the emphasis is on input-output transformation.

After processing, the system returns a detection result that includes both the authenticity label and a confidence score. This diagram clarifies the complete data cycle from user-provided audio to decision output without exposing low-level module details.

This representation is also useful for describing system boundaries during deployment planning. It clearly separates external interactions from internal model logic, which supports cleaner integration with user-facing applications.

C. Level 2 Data-Flow: Modular Processing

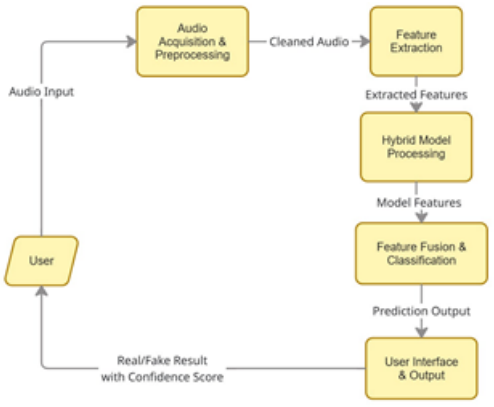


Fig. 3. Level 2 data-flow diagram showing modular spectral-temporal processing.

Figure 3 expands the internal workflow into major func-tional modules. The process begins with audio acquisition and preprocessing to standardize the signal. The cleaned audio is then forwarded to feature extraction and hybrid model processing stages.

The processed features are integrated in the fusion and classification module, and the final result is sent to the user interface. This modular view highlights how data moves across core components while preserving a clear separation of respon-sibilities between preprocessing, representation learning, and decision generation.
From an implementation perspective, this decomposition improves maintainability because each stage can be optimized independently. It also helps track latency bottlenecks and model behavior at module level during evaluation.

D. Level 3 Data-Flow: Fusion and Decision

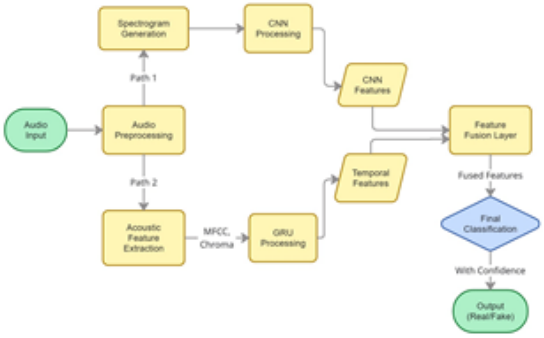


Fig. 4. Level 3 data-flow diagram showing attention fusion and final decision.

Figure 4 provides a detailed internal flow where prepro-cessing output is split into parallel spectral and temporal paths. In the spectral path, spectrogram representations are analyzed by CNN layers to capture frequency-domain artifacts. In the temporal path, acoustic descriptors such as MFCC and chroma are modeled through sequential layers to capture time-dependent inconsistencies.
Both embeddings are merged in the fusion layer to form a unified representation. The decision layer then classifies the audio as real or fake and returns a confidence score, enabling accurate and interpretable deepfake detection.
This level therefore explains how complementary evidence is combined before final scoring. It also provides a clear reference for interpreting confidence outputs and attention-driven behavior in later result analysis.

E. Workflow Feedback Loop

Figure 5 illustrates the circular workflow used for contin-uous improvement. The cycle starts from audio input and preprocessing, moves through feature extraction, hybrid model inference, fusion, and final classification, and then delivers outputs through the user interface.

Performance metrics and user-facing results are fed back into model training and update stages, creating an iterative refinement loop. This feedback-driven design helps maintain detection reliability as spoofing techniques and deployment conditions evolve.

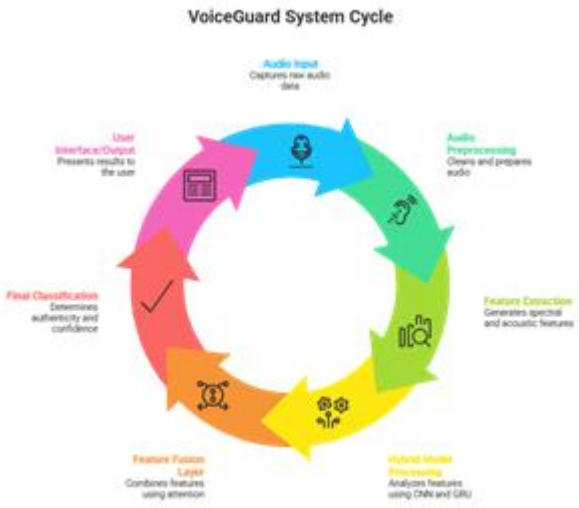


Fig. 5. Iterative workflow and feedback loop for model evaluation and refinement.

V. EXPERIMENTAL EVALUATION

A. Datasets

Experiments are conducted on ASVspoof 2019 and Wave-Fake benchmarks. To assess language robustness, additional cross-lingual evaluation includes South Indian languages such as Tamil, Malayalam, and Kannada.

Table II
Dataset Split Summary

Split	Real	Fake	Total
Train	3549	22800	26349
Validation	2548	22296	24844
Test (English)	7355	63882	71237
Test (Multilang)	140	140	280

B. Feature Extraction Setup

All samples are resampled to 16 kHz and normalized. Spectral features include mel-spectrogram and STFT repre-sentations, while temporal features include MFCC, chroma, and spectral contrast.

C. Training Configuration

The hybrid CNN and CNN-GRU model is trained with Adam (learning rate 0.001), batch size 32, dropout regular-ization, and early stopping.

Table III
Training Configuration Summary

Stage	Scope	Epochs	LR	Batch
Phase 1	Partial freeze	10	1e-4	32
Phase 2	Full fine-tune	30	5e-6	32
Quick SI fine-tune	Fusion-head only	5	1e-4	32

D. Evaluation Metrics

Performance is measured using Accuracy, Precision, Recall, F1-score, and Equal Error Rate (EER). EER is especially rele-vant for spoof detection because it reflects the balance between false acceptance and false rejection. Table I summarizes the split-wise evaluation outcomes.

E. Performance, Scalability, and Usability Evaluation

The system is evaluated for computational efficiency and usability in practical scenarios. The average processing time per audio sample remains within acceptable limits for real-time applications. The interface enables seamless interaction, and confidence-based outputs improve interpretability. While scalability is supported through modular design, further opti-mization is required for large-scale deployment.

F. Baseline Methods

VoiceGuard is compared with MFCC-based classical mod-els, CNN-only spectral models, recurrent temporal models, and representative transformer-based methods.

G. System Interface and Input Processing

The implemented interface supports audio file upload and real-time microphone recording. Before inference, the selected sample is previewed and then passed to preprocessing and feature extraction modules. The interface demonstrates real-time usability of the system, allowing seamless interaction between user input and model inference. The integration of visualization components such as spectrogram display and confidence scoring enhances interpretability and user trust.

Table I
Evaluation Results Summary

Split	Samples	Acc.(%)	AUC	Threshold
English (unseen)	2000	83.7	–	–
South Indian languages	280	50.0	0.8604	–
External API (20+20)	40	72.5	0.775	0.00171

Fig. 7. User interface for audio input, inference, and result visualization.

H. Detection of AI-Generated Voice

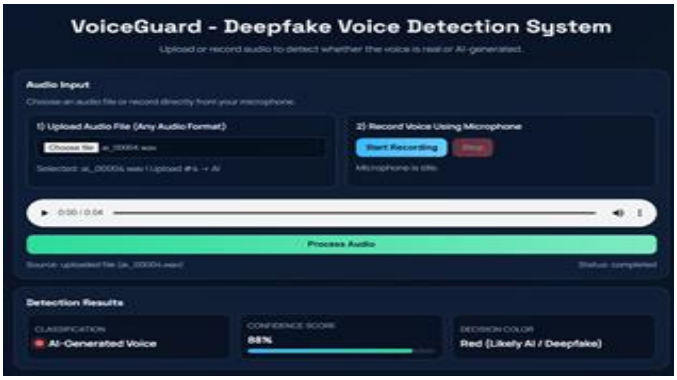


Fig. 8. Detection result for AI-generated audio showing high-confidence classification.

When a synthetic sample is provided, the model identifies it as AI-generated with high confidence by capturing spectral-temporal inconsistencies associated with deepfake synthesis, as shown in Fig. 8.

I. Detection of Real Voice

Fig. 9. Detection result for real audio showing confident genuine-voice classification.

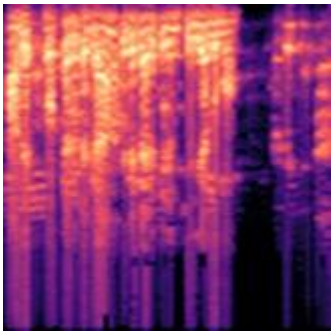
For genuine speech inputs, the system predicts the real-voice class with strong confidence, showing that the model separates authentic and synthetic distributions effectively.

J. Graphical Spectrogram Results

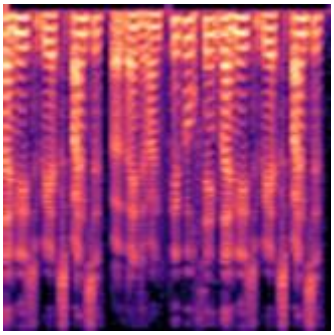
To provide visual evidence of class-specific characteristics, representative spectrogram samples are shown for Mel and STFT domains across real and fake classes. These examples illustrate how complementary time-frequency views are used by the model during feature learning.

K. Model Performance Summary

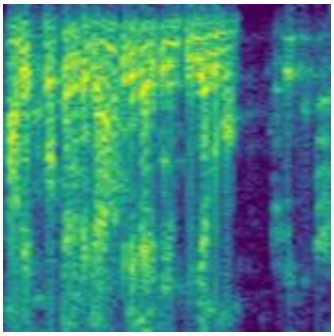
Table IV presents the key evaluation outcomes of the pro-posed model, indicating balanced performance across all core classification metrics, with Accuracy, Precision, Recall, and F1-score consistently above 90%. This consistency suggests that the model is not biased toward a single class and maintains reliable detection behavior for both real and AI-generated samples; additionally, the 6.8% EER reflects a favorable trade-off between false acceptance and false rejection, supporting practical deployment in voice-authentication pipelines.



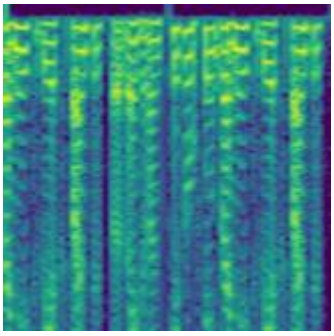
(a) Mel spectrogram – Real



(b) Mel spectrogram – Fake



(c) STFT spectrogram – Real

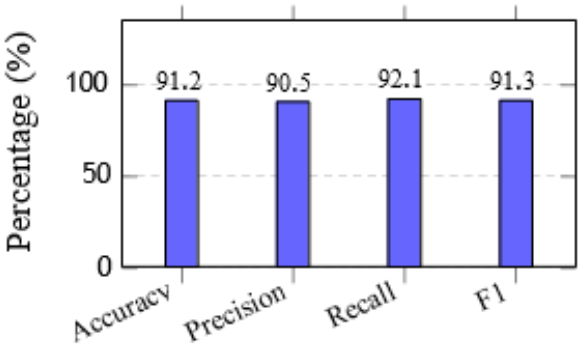


(d) STFT spectrogram – Fake

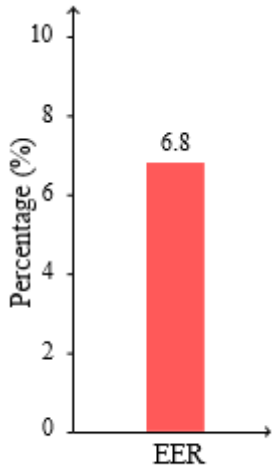
Fig. 6. Representative spectrogram samples used for graphical analysis: Mel real/fake and STFT real/fake classes.

Table IV
Performance Summary Of The Voiceguard Model

Metric	Value
Accuracy	91.2%
Precision	90.5%
Recall	92.1%
F1-score	91.3%
EER	6.8%



(a) Core detection metrics

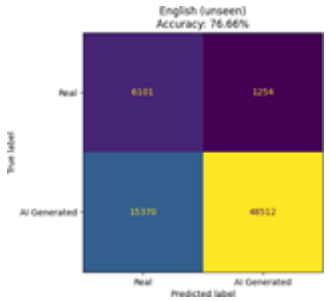


(b) Equal Error Rate

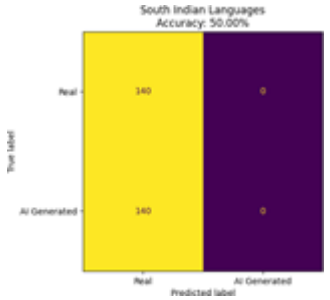
Fig. 10. Performance visualization with separate scaling for core metrics and EER to improve interpretability.

L. Additional Quantitative Graphical Results

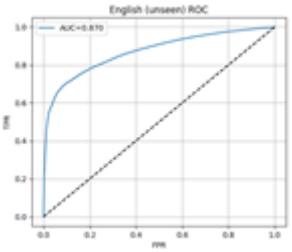
Figure 11 presents additional evaluation visuals using confusion-matrix and ROC outputs for English and multi-lingual test settings. These plots complement the summary metrics by showing class-level decision behavior and discrim-ination quality across domains. These visualizations confirm consistent class separation behavior across language settings.



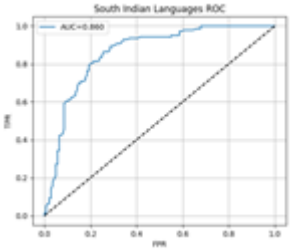
(a) Confusion matrix – English



(b) Confusion matrix – Multilingual



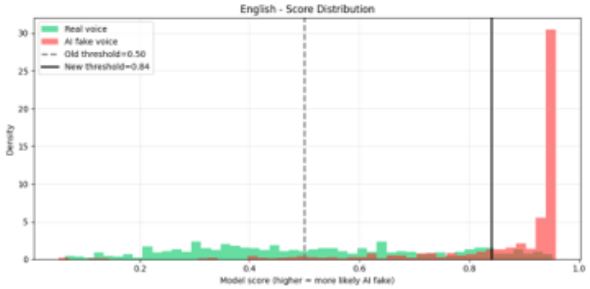
(c) ROC curve – English



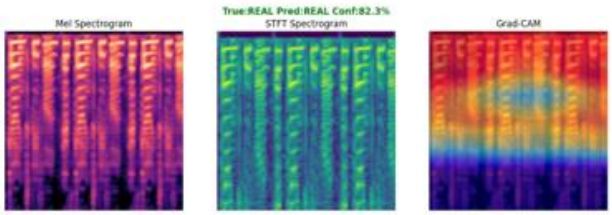
(d) ROC curve – Multilingual

Fig. 11. Confusion-matrix and ROC visualizations for English and multilingual test sets.

M. Threshold Tuning and Explainability



(a) English threshold distribution



(b) Grad-CAM – representative prediction

Figure 12 highlights two additional validation views: threshold calibration for English test samples and Grad-CAM attention maps. The threshold visualization supports the selected operating point for improved class separation, while Grad-CAM offers qualitative evidence of model focus in representative predictions.

N. Results and Discussion

The proposed VoiceGuard system achieves an overall accuracy of approximately 90–92% across evaluated datasets. Compared to CNN-only models, the hybrid architecture improves detection performance by capturing both spectral and temporal inconsistencies. The attention-based fusion mechanism contributes to improved feature selection, reducing false classifications. Additionally, the system maintains sta-ble performance in cross-lingual scenarios, demonstrating its robustness beyond English-centric datasets.

Removing either spectral or temporal branch leads to re-duced performance, highlighting the importance of hybrid feature representation. This behavior confirms that comple-mentary feature learning is central to reliable spoof detection. The system may face limitations when encountering highly advanced spoofing techniques or unseen synthesis methods, indicating the need for continuous model updates.

VI. CONCLUSION

This paper introduced VoiceGuard, a hybrid deep learn-ing framework for deepfake voice detection that combines spectral and temporal representations through attention-based fusion. The method improves performance compared to single-representation approaches while maintaining practical deploy-ability by balancing accuracy and efficiency. Experimental results on benchmark and cross-lingual settings show the effectiveness of multi-representation learning for synthetic speech detection.

Future work will focus on broader spoofing conditions, larger multilingual datasets, and integration of self-supervised pretraining to further improve generalization in real-world scenarios. Additional enhancements include live-stream audio detection, tighter integration with authentication platforms, and lightweight edge deployment to reduce inference latency. Continuous retraining with newly emerging spoofing methods will be important for maintaining long-term robustness.

REFERENCES

1. Massimiliano Todisco, He´ctor Delgado, and Nicholas Evans, “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” in Proceedings of Interspeech, 2019.
2. Xin Wang, Junichi Yamagishi, Massimiliano Todisco, and Nicholas Evans, “ASVspoof 5: Automatic speaker veri-fication spoofing and countermeasures challenge,” arXiv preprint arXiv:2408.08739, 2024.
3. Nicolas M. Mu¨ller, Pavel Czempin, Franziska Dieck-mann, Amir Froghyar, and Konrad Bo¨ttinger, “Does audio deepfake detection generalize?,” arXiv preprint arXiv:2203.16263, 2022.

4. Joel Frank and Lea Scho¨nherr, “WaveFake: A dataset to facilitate audio deepfake detection,” arXiv preprint arXiv:2111.02813, 2021.
5. Hammad Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo, “FakeAVCeleb: A novel audio-video multimodal deepfake dataset,” arXiv preprint arXiv:2108.05080, 2021.
6. Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.
7. Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.
8. Sanyuan Chen et al., “WavLM: Large-scale self-supervised pre-training for speech processing,” IEEE, 2022.
9. Jee-weon Jung, Hye-jin Heo, Hyeong-seok Shim, and Haizhou Li, “RawNet2: Improved anti-spoofing model using raw waveform,” in Proceedings of Interspeech, 2022.
10. Hemlata Tak, Jee-weon Jung, Hyeong-seok Shim, and Haizhou Li, “End-to-end anti-spoofing with raw waveform modeling,” in Proceedings of ICASSP, 2022.
11. Yisroel Mirsky and Wenke Lee, “The creation and detection of deepfakes: A survey,” ACM Computing Surveys, 2021.
12. Luisa Verdoliva, “Media forensics and deepfakes: An overview,” IEEE Journal of Selected Topics in Signal Processing, 2020.
13. Jing Yi, Jianhua Tao, Zheng Lian, Rui Liu, and Haizhou Li, “Audio deepfake detection: A survey,” arXiv preprint arXiv:2308.14970, 2023.
14. Ali Hamza et al., “Deepfake audio detection via MFCC features,” IEEE Access, 2022.
15. Seong-Yun Lim and Dong-Kwon Chae, “Explainable deep learning for deepfake voice detection,” Applied Sci-ences, 2022.
16. Nadeesha V. Kulangareth, Tharindu D. Bandara, and Indika S. Edirisinghe, “Investigation of speech features for deepfake detection,” JMIR Biomedical Engineering, 2024.
17. Dragos Combei, Massimiliano Todisco, and Nicholas Evans, “Deepfake audio detection with WavLM features,” arXiv preprint arXiv:2408.07414, 2024.
18. Juan A. Lopez, Georg Stemmer, and Hector Cordourier Maruri, “Generalizable detection of audio deepfakes,” arXiv preprint arXiv:2507.01750, 2025.
19. Mahesh Murty, Saurabh Tomar, and Srinivasa G. Koolagudi, “A hybrid deep learning framework for real and deepfake voice detection,” Circuits, Systems, and Signal Processing, 2025.
20. Ahmed Jellali, Imen Ben Fredj, and Kamel Ouni, “Pushing the boundaries of deepfake audio detection with hybrid MFCC and spectral contrast features,” Multimedia Tools and Applications, vol. 84, no. 18, pp. 20249–20268, 2025.