Generator Guided Synthetic Speech for Robust Deepfake Detection

Abdullah Al Muhit
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
muhit009@vt.edu

Ishika Khokhani
Department of Computer Science
Virginia Tech
Alexandria, VA, USA
ishikakhokhani@vt.edu

Abstract

Current audio deepfake detectors degrade sharply under distribution shift. We propose a controllable synthetic-speech framework that enhances robustness and interpretability by pairing a conventional detector with a lightweight LLM-based rationale generator. The curated combination of real and synthetic speech data allows evaluation on ASVspoof and related benchmarks for generalization and explainability. Expected outcomes include improved robustness, cross-generator consistency, and interpretable reasoning for spoof identification.

Keywords

Synthetic Speech, Deepfake Detection, LLM Fine-tuning, Explainability, Audio Forensics

1 Introduction and Related Work

Audio deepfakes, namely synthetic or cloned voices generated through Text-to-Speech (TTS) or Voice Conversion (VC) –pose serious challenges to security and trust. Recent incidents involving cloned celebrities or political voices demonstrate the risks of misinformation and fraud. Existing deepfake detection models, such as **AASIST** and **RawNet2** , achieve strong accuracy on known spoofing generators but degrade significantly under unseen conditions.

Although data augmentation through pitch-shifting or time-stretching provides shallow diversity, it does not sufficiently expose models to novel spoofing artifacts. Furthermore, most detectors operate as black boxes without offering meaningful explanations. Our project aims to design a framework that integrates deepfake detection with lightweight LLM-based interpretability, enabling robust and transparent decision-making.

2 Proposed Contribution

We will explore a hybrid architecture combining audio-based spoof detection with textual rationalization. A lightweight LLM (e.g., LLaMA-3 1B or Mistral 0.7B) will be coupled with pre-trained detection backbones (AASIST/RawNet2) to generate short natural-language explanations for predictions. This design enables users to interpret why a given audio sample is classified as real or fake while improving the robustness of the model through semantic reasoning.

2.1 Datasets

For this work, we utilize publicly available benchmark datasets widely used in the study of synthetic speech and deepfake detection. The **ASVspoof 2019** and **ASVspoof 2021** corpora form the core of our experiments, containing bona fide and spoofed audio samples generated through TTS and VC algorithms. Each utterance is accompanied by speaker metadata and ground-truth labels, enabling fine-grained evaluation of detector performance.

The training and validation subsets will be prepared by resampling all audio to **16 kHz**, converting signals into **log-mel spectrograms**, and normalizing durations for consistent input. These representations are compatible with AASIST and RawNet2 architectures, which rely on spectral features. In addition, auxiliary datasets such as **WaveFake** and **Fake-or-Real Speech** may be used to enhance diversity and test cross-dataset generalization. Each sample's transcript and embeddings will be preserved to support potential multimodal extensions with the LLM-based rationale generator.

3 Evaluation

Our evaluation emphasizes both detection performance and interpretability. Primary quantitative metrics include **Accuracy**, **Precision**, **Recall**, **F1-score**, and **Equal Error Rate (EER)**—a key benchmark for spoofing detection. We further report **Receiver Operating Characteristic (ROC)** curves and **Area Under Curve (AUC)** values to illustrate trade-offs between false acceptance and rejection rates.

The evaluation procedure compares baseline detectors (AASIST, RawNet2) with models enhanced through feature-level modifications and LLM-based reasoning. Cross-corpus generalization, such as training on ASVspoof 2019 and testing on ASVspoof 2021, will assess robustness to unseen spoofing algorithms. Explainability will be evaluated via LLM-generated rationales, qualitatively analyzed by human reviewers for coherence and acoustic alignment.

3.1 Evaluation Plan

The evaluation plan consists of three major components. First, **baseline replication** will validate reproducibility of published results for AASIST and RawNet2 on ASVspoof datasets. Second, **robustness analysis** will measure performance under cross-dataset and cross-generator settings, focusing on shifts in spoofing types and recording conditions. Finally, **explainability evaluation** will involve human inspection of generated rationales and alignment with model activations to determine interpretive accuracy.

This evaluation plan aims to demonstrate measurable improvements in both robustness and transparency. Metrics such as EER, AUC, and confusion matrices will provide quantitative validation, while qualitative reasoning outputs reflect interpretability from a user perspective.

4 Primary Experiments

The experimental workflow is organized into three progressive stages:

(1) **Baseline Reproduction:** Replicate state-of-the-art detectors (AASIST, RawNet2) using ASVspoof 2019/2021 to establish a solid benchmark for comparison.
(2) **LLM-Assisted Explanation Module:** Integrate a compact, open-weight LLM to generate concise textual rationales conditioned on extracted spectrogram or latent features. These explanations will improve interpretability while maintaining model efficiency.
(3) **Cross-Generator and Ablation Studies:** Evaluate models on unseen spoofing types to test domain robustness. Conduct ablation experiments by removing or altering key spectral bands to assess their effect on detection reliability.

Performance across all experiments will be statistically compared using paired significance testing to ensure consistent improvement.

5 Work Distribution

Table 1: Work allocation among project members.

Member	Responsibilities
Ishika	Dataset preparation, preprocessing, and baseline setup using AASIST and RawNet2.
Muhit	LLM integration for explainability, evaluation of results, and report documentation.
Both	Model training, result analysis, writing, and presentation preparation.

Acknowledgments

We would like to thank the course instructor **Dr. Chandan Reddy** and teaching assistants for their continued guidance during this project. We also acknowledge the contributors of the ASVspoof datasets and the open-source research communities behind AASIST, RawNet2, and lightweight LLM architectures that serve as the foundation of this work.

References

[1] Todisco, M. et al. (2019). ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge. *Interspeech 2019.*
[2] Jung, J. et al. (2022). AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. *IEEE TASLP.*
[3] Tak, H. et al. (2021). End-to-End Anti-Spoofing with RawNet2. *IEEE Transactions on Information Forensics and Security.*
[4] Touvron, H. et al. (2024). LLaMA 3: Open Efficient Foundation Language Models. *arXiv preprint arXiv:2405.13015.*
[5] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Communications of the ACM* 50, 1 (January 2007), 36–44. DOI:https://doi.org/10.1145/1188913.1188915
[6] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79).* ACM Press, New York, NY, 226–236. DOI:https://doi.org/10.1145/567752.567774
[7] Ian Editor (Ed.). 2007. *The Title of Book One* (1st ed.). The Name of the Series One, Vol. 9. University of Chicago Press, Chicago, IL. DOI:https://doi.org/10.1007/3-540-09237-4
[8] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd ed.). Wiley, New York, NY.