International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 09 | Sep 2026 www.irjet.net p-ISSN: 2395-0072

Deepfake Audio Detection Using AASIST: A Deep Learning Approach for Audio Anti-Spoofing

Dr. Pranjali Patil[1], Kashish Gupta[2], Saurabh Ramesh Jha[3], Ritesh Chakravarti[4]

[1]*Assistant Professor, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*
[2]*Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*
[3]*Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*
[4]*Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India*

Abstract - *This paper presents a reproducible framework for deepfake audio detection using AASIST, a graph-attention architecture that integrates spectral and temporal evidence for binary bona fide-versus-spoof classification. The study clarifies dataset selection, preprocessing, reference training configuration, leakage control, evaluation metrics, baseline comparison, and error analysis for an ASVspoof-based experiment. It also distinguishes the official AASIST reference setting on ASVspoof 2019 Logical Access from an adapted evaluation on ASVspoof 2021. Published benchmark values are included only as reference evidence; no new experimental measurements are claimed. The proposed protocol emphasizes EER and minimum t-DCF, reproducible reporting, cross-dataset generalization, and analysis of attack and acoustic conditions.*

Key Words: AASIST, deepfake audio detection, audio anti-spoofing, ASVspoof, graph attention, EER, min t-DCF

1. INTRODUCTION

In this study, deepfake audio denotes synthetic or manipulated speech that can be presented as genuine speech. Two important generation paradigms are text-to-speech (TTS), which generates speech from text, and voice conversion (VC), which changes the perceived speaker characteristics while retaining the linguistic content. The resulting attacks may contain subtle spectral, temporal, phase, excitation, or vocoder-related artefacts that are difficult for human listeners to identify [1].

The central problem is the reliable discrimination of bona fide speech from AI-generated or manipulated speech under conditions that may differ from the training environment. A detector can perform strongly on a benchmark while still learning dataset-specific artefacts and failing on unseen synthesis systems, codecs, channels, or acoustic conditions. Consequently, a useful anti-spoofing study must evaluate not only classification performance but also generalization and error characteristics.

Deepfake audio can undermine voice authentication, facilitate social engineering and impersonation, and support the spread of fabricated statements. In forensic and media-verification settings, an automated countermeasure can provide an additional technical signal about whether an audio recording contains characteristics associated with synthetic generation. The motivation for using AASIST is its explicit modeling of interactions between spectral and temporal artefacts through graph attention [2].

The study has five objectives: (1) review existing audio deepfake and anti-spoofing approaches; (2) explain and implement the AASIST architecture; (3) prepare a reproducible benchmark-based preprocessing and evaluation pipeline; (4) compare AASIST with established baselines; and (5) analyze generalization, error patterns, and practical limitations.

RQ1: How effectively can AASIST distinguish bona fide speech from synthetic/manipulated speech? RQ2: Does modeling spectral-temporal interactions provide an advantage over simpler convolutional or conventional baselines? RQ3: Which attack or acoustic conditions contribute most to detection errors? RQ4: How well does the trained countermeasure generalize to conditions that differ from those represented during training?

The study is limited to speech/audio deepfake detection and binary bona fide-versus-spoof classification. It does not address image or video deepfake detection, nor does it attempt to reconstruct or improve the generative systems used to create spoofed speech.

2. LITERATURE REVIEW AND RESEARCH GAP

Deepfake technology uses learned representations to synthesize or manipulate media. In speech, TTS and VC systems can generate or transform vocal characteristics while preserving intelligible linguistic content. Earlier systems often exposed obvious artefacts, whereas modern neural synthesis can suppress many perceptual cues,

shifting detection toward subtle signal-level inconsistencies.

Earlier anti-spoofing systems commonly used hand-crafted acoustic representations with statistical classifiers [1]. More recent systems employ CNNs, recurrent networks, end-to-end waveform models, self-supervised speech encoders, and attention mechanisms [2]–[5]. These approaches differ in whether they explicitly represent spectral structure, temporal context, or long-range interactions.

MFCCs, LFCCs, CQCCs, spectrograms, and learned waveform representations have all been used in speech anti-spoofing [1], [3], [5]. A key design issue is whether a representation exposes the small differences introduced by synthesis systems. End-to-end models reduce dependence on manually engineered features, but they can also learn shortcuts tied to a particular dataset or generation pipeline.

ASVspoof is a community benchmark for developing and evaluating spoofing countermeasures [6], [7]. ASVspoof 2021 contains three tracks: Logical Access (LA), Physical Access (PA), and Speech Deepfake (DF). LA includes TTS and VC attacks transmitted through telephony/VoIP-related conditions, while DF focuses on detecting fake speech generated by TTS and VC without the speaker-verification component. The official evaluation resources provide keys, metadata, and scripts for EER and minimum tandem detection cost evaluation [8].

AASIST introduces a heterogeneous stacking graph-attention mechanism [2] to model artefacts that may occur in different temporal and spectral intervals. Its architecture uses graph operations and a readout mechanism to integrate information from these heterogeneous regions. The original publication reports that AASIST improved over competing systems and also introduced AASIST-L, a lightweight variant. Its best reported ASVspoof 2019 LA result is an EER of 0.83% and min t-DCF of 0.0275 [2].

The original AASIST reference training, validation, and evaluation setup uses the ASVspoof 2019 Logical Access dataset, not ASVspoof 2021 [2], [6]. Therefore, a study that evaluates AASIST on ASVspoof 2021 should describe this as an adapted or extended evaluation rather than implying that it exactly reproduces the original experiment. This distinction is important for scientific reproducibility and fair comparison.

A benchmark result alone does not establish robustness to future synthesis systems. The literature continues to show concerns about cross-dataset performance, interpretability, and robustness to domain mismatch. The present study therefore emphasizes attack-condition analysis and clearly separates published AASIST benchmark results from results generated in the authors' own experiment.

3. METHODOLOGY

3.1 Research Design and Dataset

The study follows a quantitative experimental design. The independent variable is the audio condition/class (bona fide or spoof), while the principal dependent measures are EER and min t-DCF, supplemented by accuracy, precision, recall, and F1-score when a fixed classification threshold is reported. The workflow comprises data preparation, preprocessing, model training, validation, evaluation, and error analysis.

For an ASVspoof 2021 extension, the researcher should select the intended track explicitly. The Speech Deepfake (DF) track is the most direct match for a pure deepfake-detection study because it is designed around bona fide and spoofed speech generated using TTS and VC. The Logical Access (LA) track is appropriate when robustness to communication/channel effects is also part of the research question. The chosen track, release/version, metadata files, and inclusion criteria must be recorded in the final experimental report.

3.2 Data Preprocessing and Partitioning

Audio is converted to a consistent representation before model inference. The supplied draft specifies mono conversion, 16-kHz resampling, fixed-length processing, and amplitude normalization. For strict replication of the official AASIST implementation, the reference configuration uses 64,600 samples rather than exactly 64,000. This corresponds to approximately 4 seconds at 16 kHz. Short signals are padded and long signals are cropped/segmented according to the implementation. All preprocessing decisions must be identical between training and evaluation except for explicitly documented training-time augmentation.

Training, development, and evaluation data must remain disjoint. Speaker identities and attack conditions should be checked for unintended overlap when the research question requires speaker-independent or attack-independent evaluation. The final test/evaluation set must not be used for hyperparameter selection. Any additional cross-condition experiment should be described separately from the primary benchmark result.

3.3 AASIST Architecture

Note: Fig. 1 is an original schematic prepared for this manuscript; it summarizes the architecture described by Jung et al. [2] and is not a reproduction of the source figure.

The AASIST pipeline can be summarized as: raw waveform → learned front-end feature extraction → spectro-temporal feature map → heterogeneous graph construction → graph attention and graph operations → graph-level readout → binary classification. The central design principle is that spoofing artefacts can be localized in different time-frequency regions and may interact across those regions. Graph attention provides a mechanism for learning which relationships are informative rather than relying only on local convolutional neighborhoods.

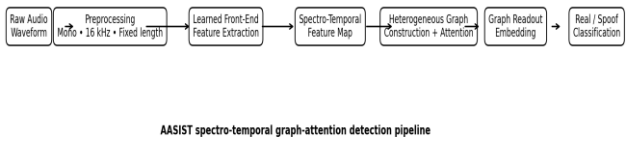

Fig. 1: Proposed AASIST-based spectro-temporal deepfake audio detection pipeline [2].

3.4 Training Configuration and Baselines

The official AASIST configuration uses Adam with a base learning rate of 1×10^-4, cosine learning-rate scheduling, a batch size of 24, and a maximum of 100 epochs [2]. The supplied reference configuration also specifies a minimum learning rate of 5×10^-6, weight decay of 1×10^-4, and deterministic cuDNN settings. The official implementation uses weighted cross-entropy during training. These values should be treated as the reference configuration; any change made for an ASVspoof 2021 experiment must be reported explicitly.

The comparison should include at least one conventional/deep CNN baseline and one strong end-to-end or self-supervised baseline where computational resources permit. RawNet2 provides a relevant end-to-end baseline [5], while the official ASVspoof resources provide challenge baselines and evaluation materials [8]. Using an established baseline is preferable to an ad hoc implementation because it improves reproducibility.

3.5 Evaluation and Experimental Procedure

Equal Error Rate (EER) is the operating point where false-accept and false-reject rates are approximately equal; lower is better. Minimum t-DCF [9] measures the cost of a countermeasure when considered in tandem with an automatic speaker-verification system and is an official ASVspoof evaluation measure [8]. Accuracy, precision, recall, and F1-score may be reported as supplementary threshold-dependent measures, but they should not replace EER/min t-DCF for benchmark comparison.

1) download and verify the dataset and official keys; 2) prepare metadata; 3) apply preprocessing; 4) initialize the AASIST configuration; 5) train using only training data; 6) select the model using development performance; 7) run the final evaluation once the configuration is frozen; 8) compute EER/min t-DCF using the official evaluation tools where applicable; 9) generate a confusion matrix at a clearly stated threshold; and 10) analyze errors by attack type and acoustic condition.

4. RESULTS AND DISCUSSION

4.1 Published Benchmark Evidence

Published benchmark evidence is included to contextualize the proposed evaluation. On the ASVspoof 2019 LA evaluation set, the AASIST paper reports best single-run EER values of 0.83% for AASIST, 0.99% for AASIST-L, and 1.06% for RawGAT-ST; the corresponding min t-DCF values are 0.0275, 0.0309, and 0.0335 [2]. These are published reference results, not measurements generated by the present authors.

Table 1. Published ASVspoof 2019 LA results [2]

System	Parameters	min t-DCF	EER (%)
AASIST	297K	0.0275	0.83
AASIST-L	85K	0.0309	0.99
RawGAT-ST	437K	0.0335	1.06

No new experimental run is claimed in this manuscript. Accordingly, accuracy, EER, min t-DCF, confusion-matrix counts, and baseline scores are not presented as original measurements. The published AASIST benchmark values in Table 1 provide a reproducible reference point for the proposed evaluation.

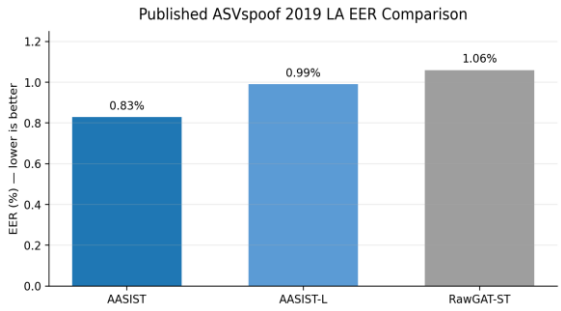

Fig. 2: Published ASVspoof 2019 LA EER comparison; lower is better [2].

4.2 Planned Experimental Validation and Baseline Comparison

When the authors perform the proposed experiment, the final results should be reported with the exact dataset track, model checkpoint, evaluation split, preprocessing configuration, and thresholding rule so that the measurements are reproducible.

AASIST should be compared with selected baselines using the same evaluation split and metric definitions. Published values from different datasets or protocols should be kept separate from newly generated measurements.

4.3 Error Analysis and Discussion

For a completed experimental run, a 2×2 confusion matrix should be reported at a stated decision threshold. True positives and true negatives must be defined consistently with the chosen positive class, and the manuscript should state whether “spoof” or “bona fide” is treated as positive.

Errors should be grouped into false positives, in which bona fide speech is incorrectly classified as spoof, and false negatives, in which spoofed speech is accepted as bona fide. The analysis should additionally identify whether errors concentrate in particular synthesis families, signal durations, channels, codecs, or acoustic conditions. This is more informative than reporting a single aggregate accuracy value.

If AASIST obtains lower EER than the selected baseline under the same protocol, the discussion should relate the improvement to its ability to integrate spectral and temporal information rather than simply attributing the gain to “deep learning.” If performance deteriorates on unseen conditions, that result should be treated as evidence of domain sensitivity rather than as a failure of the entire architecture. Any claim of generalization must be supported by a genuinely disjoint evaluation condition.

4.4 Limitations

The principal limitations are benchmark dependence, possible domain mismatch, computational requirements, and the rapidly changing nature of speech synthesis. A detector trained on one generation ecosystem may encounter new artefacts when exposed to later synthesis models. In addition, benchmark performance does not by itself establish forensic admissibility or certainty about the provenance of an individual recording.

5. CONCLUSION AND FUTURE WORK

This study presents a reproducible framework for deepfake audio detection using AASIST, a graph-attention architecture designed to integrate spectro-temporal evidence. The methodology clarifies dataset selection, preprocessing, reference training configuration, evaluation metrics, leakage control, baseline comparison, and error analysis. A critical replication distinction is also established: the official AASIST reference experiment uses ASVspoof 2019 Logical Access, whereas evaluation on ASVspoof 2021 constitutes an adapted benchmark setting. The published benchmark evidence reported in this paper should therefore be interpreted as reference performance until the proposed experiment is independently executed.

The revised study contributes (1) a structured explanation of AASIST for deepfake audio detection; (2) a reproducible experimental protocol; (3) a clear separation between original benchmark evidence and new experimental evidence; (4) an evaluation framework centered on EER and min t-DCF; and (5) an error-analysis plan focused on attack type and environmental robustness.

Future work should examine newer voice-cloning and speech-synthesis systems, cross-dataset evaluation, multilingual speech, compressed/telephony audio, lightweight real-time variants, and interpretable detection. It could also test whether analyzing voiced and unvoiced regions separately improves spectro-temporal countermeasures.

6. REFERENCES

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and Countermeasures for Speaker Verification: A Survey,” Speech Communication, vol. 66, pp. 130–153, 2015, doi: 10.1016/j.specom.2014.10.005.

[2] J.-W. Jung et al., “AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks,” in Proc. IEEE ICASSP, Singapore, 2022, pp. 6367–6371, doi: 10.1109/ICASSP43922.2022.9747766.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. 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.

[4] S. Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," IEEE J. Sel. Top. Signal Process., vol. 16, no. 6, pp. 1505–1518, 2022, doi: 10.1109/JSTSP.2022.3188113.

[5] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End Anti-Spoofing with RawNet2," in Proc. IEEE ICASSP, Toronto, ON, Canada, 2021, pp. 6369–6373, doi: 10.1109/ICASSP39728.2021.9414234.

[6] X. Wang et al., "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," Computer Speech & Language, vol. 64, Art. no. 101114, 2020, doi: 10.1016/j.csl.2020.101114.

[7] X. Liu et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 30, pp. 2823–2837, 2022, doi: 10.1109/TASLP.2022.3188392.

[8] ASVspoof Challenge, "ASVspoof 2021 challenge resources," 2021. [Online]. Available: https://www.asvspoof.org/index2021.html. Accessed: Sep. 10, 2026.

[9] T. Kinnunen et al., "t-DCF: A Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification," in Proc. Odyssey 2018, Les Sables d'Olonne, France, pp. 312–319, 2018, doi: 10.21437/Odyssey.2018-44.