Multi-Stream Artifact Detection for Audio Deepfake Detection via Temporal, Spectral, and Phase Cues

Damarhafni Rahmannabel Nadim Pramono, Bagus Tris Atmaja, and Sakriani Sakti
Nara Institute of Science and Technology, Japan
E-mail: {damarhafni.pramono.do0, bagus.tris}@naist.ac.jp, ssakti@is.naist.jp

***Abstract*—Audio deepfakes produced by modern text-to-speech and voice-conversion systems pose growing threats to security and trust. This work proposes the Artifact Detection Module (ADM), a multi-stream neural classifier that decomposes deepfake detection into three complementary artifact hypotheses: temporal, spectral, and phase. Each hypothesis is addressed by a dedicated residual multilayer perceptron sub-module operating on mean-pooled utterance-level features from self-supervised learning embeddings, mel-filterbank energies, and short-time Fourier transform-derived phase features, respectively. Branch scores are fused using a concatenation-based projection by default, while alternative fusion strategies are analyzed through ablation experiments. ADM is evaluated on ASVspoof 2019 Logical Access, ASVspoof 2021 Logical Access, ASVspoof 2021 Deepfake, and the In-the-Wild dataset using equal error rate. Although the temporal branch is the strongest individual stream, three-branch fusion with maximum aggregation achieves the best overall result, with equal error rates of 0.12%, 6.01%, 3.15%, and 4.72% on the four evaluation sets, respectively.**

***Index Terms*—Audio deepfake detection, anti-spoofing, multi-stream learning, self-supervised speech representation, spectral features, phase features.**

I. Introduction

Audio deepfake detection aims to identify manipulated speech that is intended to appear authentic. Recent advances in text-to-speech (TTS), voice conversion (VC), and related speech generation technologies have made realistic fake audio increasingly accessible [1]. While these technologies enable useful applications such as personalized assistants, creative voice cloning, and language-learning tools, their misuse can threaten security, privacy, and trust [1]. In fraud or impersonation scenarios, undetected deepfakes can have serious consequences; for example, a synthetic voice clone could be used to impersonate a bank customer. Moreover, both humans and machines are struggling to identify convincing deepfakes [2], and recent listening evidence suggests that modern audio deepfakes may erode trust in genuine speech as well as deceive listeners [3]. Effective audio deepfake detection is therefore essential for protecting audio-based communication.

This problem is amplified by the growing use of voice as a biometric signal in speaker identification, speaker verification, and voice-based profiling. Recent work argues that human voice can be treated as a distinctive biometric entity when measurable vocal-production characteristics are considered, with voice uniqueness depending on how such characteristics are quantized [4]. If voice carries biometric identity information, synthetic voice cloning can directly undermine authentication, attribution, and trust, consistent with broader concerns about deepfakes as attack vectors against recognition systems [2]. Thus, audio deepfake detection supports both media forensics and the protection of biometric voice systems.

Despite recent progress, many detectors emphasize a single modeling perspective or feature family. Prior studies show that spoofing artifacts are not confined to one acoustic view: learned speech representations can encode temporal and low-level acoustic patterns, while hand-crafted descriptors can reveal spectral and prosodic irregularities [5], [6]. Work on cross-dataset generalization and bona fide distribution modeling further shows that performance can shift with attack types, corpus, and recording conditions [7], [8]. These observations motivate a detector that jointly analyzes multiple cues rather than relying on one representation to capture all spoofing evidence.

Motivated by these limitations, this work formulates audio deepfake detection as a multi-stream artifact detection problem. The proposed ADM explicitly separates temporal, spectral, and phase artifact hypotheses. The temporal stream uses self-supervised learning (SSL) embeddings to capture utterance-level speech dynamics, the spectral stream models mel-filterbank envelope artifacts, and the phase stream targets waveform-level phase inconsistencies introduced by synthesis and vocoding.

By fusing these streams, ADM is designed to remain robust when one artifact type is weak, absent, or dataset-dependent.

The main contributions of this work are three parts:

- ADM is introduced as a lightweight multi-stream residual MLP framework for detecting temporal, spectral, and phase artifacts in audio deepfakes.
- SSL layer selection, stream-branch combinations, and fusion strategies are systematically analyzed to clarify which cues contribute most to generalization.
- A locked ADM configuration is evaluated on ASVspoof 2019 LA, ASVspoof 2021 LA, ASVspoof 2021 DF, and ITW, demonstrating competitive performance against recent methods on both benchmark and real-world evaluation sets.

II. Related Work

Previous works in audio deepfake detection have explored several directions such as improving generalization [7], modeling real-speech distributions to classify out-of-distribution samples [8], or identifying acoustic features correlated with real and fake speech, either using hand-crafted features [5] or SSL features [6]. PoinHierNet improves generalization by learning hierarchical representations that reduce domain-sensitive variation and better capture relationships between attacks and datasets within a single stream [7], while FPM frames deepfake detection as an anomaly-detection problem by modeling bona fide speech and treating spoofed samples as deviations from that distribution [8]. Feature-oriented studies also show that spoofing artifacts may appear in different acoustic forms: hand-crafted features can expose spectral and prosodic irregularities [5], while SSL models provide hierarchical representations, with recent analyses showing that their lower-level acoustic and temporal features are particularly discriminative for deepfake detection [6]. However, these methods often focus on a single feature family or representation, which can miss artifacts outside that family. Synthetic speech artifacts may arise across multiple domains, and no single perspective is sufficient to capture all of them. This limitation becomes important in real-world scenarios where generation methods vary widely, motivating a framework that fuses multiple representations for robust detection.

III. Methods

A. Model Architecture and Parameters

The proposed ADM is a **multi-stream neural classifier** for binary audio deepfake detection (real vs. synthetic speech). It decomposes the detection problem into three distinct artifact hypotheses (temporal, spectral, and phase) and combines their evidence scores into a single utterance-level decision. In the base configuration, ADM uses three active branches, including time, spectral, and phase, and produces one utterance-level logit for binary classification. The overall ADM architecture is shown in Fig. 1. The model is designed to remain lightweight: each branch is a residual MLP backend rather than a large sequence model, while the heavier speech representation learning is handled by the SSL feature extractor.

1) Feature streams: The model uses mean-pooled utterance-level features from three streams: (i) 1024-dimensional wav2vec2-XLSR-300m embeddings for temporal and micro-prosodic cues, with layers 0–24 considered in the ablation [9]; (ii) mel-filterbank energies for spectral and vocoder artifacts; and (iii) approximately 257-dimensional STFT phase features for synthesis-induced phase dynamics.

2) Sub-module architecture: Each stream is processed by a dedicated residual MLP with LayerNorm and GELU activations. Given a stream-level input vector $\mathbf{x}$, each branch computes

$$\begin{aligned} \mathbf{h}_1 &= \mathrm{GELU}(\mathrm{LN}(W_1\,\mathbf{x})),\\ \mathbf{h}_2 &= \mathbf{h}_1 + \mathrm{GELU}(\mathrm{LN}(W_2\,\mathbf{h}_1)),\\ \mathbf{h}_3 &= \mathrm{GELU}(\mathrm{LN}(W_3\,\mathbf{h}_2)),\\ s &= W_{\mathrm{out}}\,\mathbf{h}_3 \in \mathbb{R}. \end{aligned} \tag{1}$$

The scalar s is the branch-level evidence score for the bona fide (real) class. Dropout (p=0.2) is applied after hidden transformations for regularization. The spectral and phase sub-modules use `LazyLinear` for the first projection so that input dimensionality can be inferred at runtime. LayerNorm is preferred over BatchNorm because training batches can be small, imbalanced, or drawn from heterogeneous corpora.

3) Multi-stream fusion: Let $s_i \in \mathbb{R}$ denote the scalar branch score from branch $i \in \{\text{time}, \text{spectral}, \text{phase}\}$. The model supports five fusion strategies: simple average (*avg*), element-wise maximum (*max*), learnable softmax weighting (*weighted*, $\hat{y} = \sum_i \mathrm{softmax}(\mathbf{w})_i\, s_i$), linear projection of the score vector (*concat*), and gated fusion (*gated*):

$$\hat{y} = \sum_i \big[\sigma(W_g\,\mathbf{s})\big]_i \cdot s_i, \tag{2}$$

where $\mathbf{s} = [s_1, \ldots, s_N]^\top$ and $W_g \in \mathbb{R}^{N\times N}$. Concatenation-based score projection is used as the default ADM configuration because it learns a direct affine combination of branch evidence while preserving

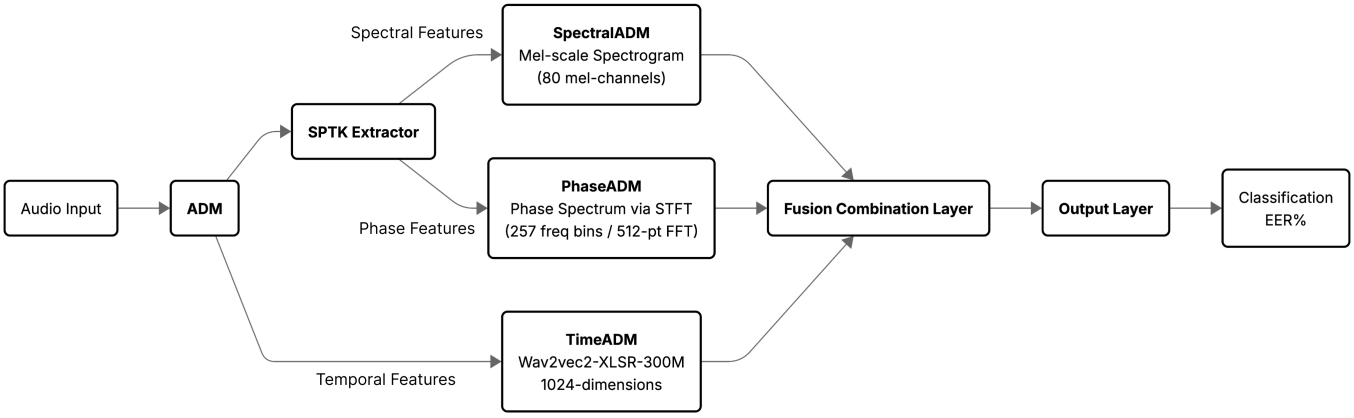


Fig. 1. Overview of the proposed ADM architecture for audio deepfake detection. The model extracts temporal SSL embeddings, spectral mel-filterbank features, and STFT-derived phase cues from an input waveform, processes each stream with a dedicated residual MLP branch, and fuses branch-level evidence for utterance-level real/fake classification.

branch-specific score information. The remaining fusion rules are retained for ablation; in particular, gated fusion enables input-dependent branch weighting, while max fusion tests whether the strongest branch evidence is sufficient for robust utterance-level decisions.

4) Training Loss and Inference Rule: The output $\hat{y}$ is an unbounded logit trained with focal loss [10]

$$\mathcal{L} = -\alpha(1 - p_t)^{\gamma} \log p_t, \tag{3}$$

where α=0.25 and γ=3.0. Focal loss down-weights easy samples so that training concentrates on more difficult bona fide/spoof decisions. At inference, the utterance is assigned to *real* if $\sigma(\hat{y}) > \tau$ and to *fake* otherwise, where τ=0.65 in the base configuration.

IV. Experiments

A. Dataset and Evaluation Metrics

ADM is evaluated on four fixed test sets used in prior benchmarks: ASVspoof 2019 Logical Access (LA) evaluation [11], ASVspoof 2021 LA, ASVspoof 2021 Deepfake (DF) [12], and In-the-Wild (ITW) [13]. These sets are kept consistent across benchmark comparisons and ablations, with training performed on a separate multi-corpus pool.

Training uses ASVspoof 2019 LA training and development data, HABLA, HAD, PartialSpoof [14], and FoR-2sec. ASVspoof 2019, derived from VCTK, provides disjoint training, development, and evaluation partitions; its LA scenario includes text-to-speech and voice-conversion attacks, with known attacks for training/development and unseen attacks for evaluation [11].

Every corpus is used without oversampling, subsampling, or corpus-level reweighting, giving a 75.8%/24.2% split or a 3.1:1 fake-to-real ratio.

ASVspoof 2021 is used only for evaluation. Its LA set adds channel and transmission variability, while its DF set targets manipulated and compressed speech distributed in online-style conditions [12]. ITW provides an external real-world test set with bona fide and spoofed audio from 58 celebrities and politicians, totaling 20.7 and 17.2 hours, respectively [13].

The remaining training corpora broaden coverage: HABLA contributes Latin American Spanish anti-spoofing data [15], FoR-2sec provides short real/synthetic utterances from modern synthesis systems [16], and HAD [17] plus PartialSpoof [14] add localized manipulations through half-truth or partially spoofed utterances. All corpora are mapped to binary labels, *fake* and *real*; for ITW, *spoof* and *bona-fide* are mapped to *fake* and *real*, respectively.

Performance is measured by Equal Error Rate (EER), where false acceptance and false rejection rates are equal. EER is widely used in speech deepfake detection, including ASVspoof and ADD, and is more informative than accuracy for imbalanced evaluation sets [18].

B. Implementation Details

All experiments use the ADM configuration described in Section III and Fig. 1, trained for 50 epochs. The pipeline reads waveform files and extracts SSL embeddings and SPTK-based acoustic descriptors. The SSL stream uses `facebook/wav2vec2-xls-r-300m`, yielding a 1024-dimensional representation per selected layer; layer 8 is used in the base configuration, while

layers 0–24 are varied in the layer ablation. The acoustic stream uses mel-filterbank energies and STFT-derived phase features with FFT length 512, frame length 400 samples, frame period 160 samples, and 80 mel channels. The phase representation has 257 frequency-bin features. All features are robustly scaled before training.

All corpora are loaded from CSV metadata files and mapped to the binary label set {*fake*, *real*}; for ITW, *spoof* and *bona-fide* are mapped to *fake* and *real*, respectively. The benchmark evaluation sets remain disjoint from the multi-corpus training pool.

The ADM backend uses temporal, spectral, and phase branches fed by SSL, mel-filterbank, and STFT phase representations, respectively. The base configuration uses concatenation-based score projection, with average, maximum, weighted, and gated fusion retained for ablation. The model is optimized with AdamW using a learning rate of 5×10^{-6}, weight decay of 0.1, mini-batches of 32 utterances, and a three-epoch warmup. Gaussian feature noise with standard deviation 0.03 and gradient clipping at norm 1.0 are used for robustness. Focal loss uses $\alpha = 0.25$ and $\gamma = 3.0$, and inference assigns the *real* class when $\sigma(\hat{y}) > 0.65$. Because EER is computed from the score ranking, the reported results are independent of τ. All reported EERs are single-run values obtained under the identical training protocol described above.

V. Results and Discussion

A. Benchmark with Previous Works

Table I compares ADM with recent audio deepfake detection methods on the same benchmark test sets. A single configuration is locked in advance (SSL layer 8, temporal, spectral, and phase branches, max fusion, and τ=0.65) and is applied unchanged to all four held-out evaluation sets, so no per-dataset model selection enters the comparison. Under this locked setting, ADM reaches 0.12% EER on ASVspoof 2019 LA and 4.72% on ITW, the second-best in-the-wild result among the compared methods. These results suggest that the proposed artifact-focused streams are effective for controlled logical-access evaluation and remain promising for in-the-wild deepfake detection, while the ASVspoof 2021 LA and DF results highlight the importance of further robustness studies under channel and compression variability.

TABLE I
Results on benchmark datasets in terms of EER (%) on the same training sets; lower is better.

Method	19LA	21LA	21DF	ITW
OCKD [19]	0.39	0.90	2.27	7.68
Wav2Vec+Linear [20]	0.22	3.63	3.65	16.17
LSR+LSA [21]	0.15	1.19	2.43	5.92
Poin-HierNet [7]	0.11	0.94	1.40	4.91
EBM [22]	0.26	1.21	1.89	-
SLIM [23]	0.20	-	4.40	10.70
FPM [8]	10.60	-	-	2.70
ADM (Ours)	0.12	6.01	3.15	4.72

A single ADM setting (SSL layer 8, T+S+P branches, max fusion, and τ=0.65) is used for all four evaluation sets. Section V-B reports design ablations rather than dataset-specific model selection.

B. Ablation Studies

To provide a controlled analysis, the ablation studies are performed in three consecutive stages: *(i)* SSL layer selection, *(ii)* branch combination analysis using the selected SSL layer, and *(iii)* fusion strategy comparison using the best SSL layer and branch setting. The evaluation flow is summarized as

$$\text{SSL Layer } \{0, 2, \ldots, 24\} \rightarrow \text{Best layer} \rightarrow \text{Branch study} \rightarrow \text{Temporal+Spectral+Phase branches} \rightarrow \text{Fusion study.} \quad (4)$$

1) Effect of SSL Layer Selection: Table II reports the effect of varying the SSL layer used by the temporal stream. All three streams are enabled in this study, while the spectral and phase feature settings are kept fixed. The 8th SSL layer gives the best overall cross-dataset trend, especially on 21DF and ITW, and is therefore fixed for the subsequent branch and fusion ablations. Performance degrades markedly at layers 0–2 on the three cross-domain sets, while deeper layers do not consistently reduce EER. This pattern indicates that intermediate SSL representations provide a better balance between low-level acoustic detail and higher-level abstraction for cross-dataset detection.

TABLE II
SSL layer ablation for the temporal stream in ADM. Results are shown in EER (%); lower is better.

SSL Layer	19LA	21LA	21DF	ITW
24	0.51	10.97	7.29	14.62
20	0.12	8.63	5.75	12.12
18	1.56	7.77	5.75	12.85
16	1.38	9.66	10.48	10.34
14	0.86	9.39	6.93	8.78
12	0.71	9.19	10.66	7.22
10	0.46	7.26	5.71	7.23
8	0.39	8.04	4.82	4.79
6	0.13	9.28	13.20	4.73
4	0.26	12.35	11.15	8.35
2	0.21	17.96	19.41	20.56
0	0.37	22.55	28.06	34.76

2) Effect of Different Branches: Using SSL layer 8, Table III reports the evaluated stream combinations. Single-stream results quantify individual cue discrimination, whereas combined streams assess cross-stream complementarity. The temporal stream is strongest alone, whereas spectral and phase cues depend on the preservation of synthesis artifacts. Adding phase reduces ITW EER from 5.47% to 5.22%, indicating that codec-free audio may preserve useful vocoder phase cues. Thus, all three streams are retained so that fusion can exploit reliable evidence from each utterance.

TABLE III
STREAM-BRANCH ABLATION IN ADM USING THE SELECTED SSL LAYER FROM TABLE II. RESULTS ARE SHOWN IN EER (%); LOWER IS BETTER. T: TEMPORAL, S: SPECTRAL, P: PHASE.

Branch	19LA	21LA	21DF	ITW
T	0.02	6.14	3.38	5.47
S	14.14	36.74	38.99	43.75
P	28.67	43.25	47.21	40.37
T+S	1.43	8.44	4.25	8.05
T+P	0.14	6.43	4.17	5.22
S+P	12.78	39.91	38.89	37.12
T+S+P	0.39	8.04	4.82	4.79

3) Effect of Different Fusion Strategies: Using the full T+S+P branch configuration with SSL layer 8, the fusion rules introduced in Section III are compared. Although concatenation is ADM's default fusion rule, max fusion performs best overall, suggesting that the strongest branch score is more robust when the informative stream varies with recording conditions. Max fusion is therefore adopted in Table I, together with SSL layer 8, all three branches, and τ=0.65.

TABLE IV
FUSION-STRATEGY ABLATION IN ADM USING THE SELECTED SSL LAYER AND FULL STREAM CONFIGURATION. RESULTS ARE SHOWN IN EER (%); LOWER IS BETTER.

Fusion Strategy	19LA	21LA	21DF	ITW
weighted	0.59	7.67	5.53	4.76
avg	1.07	7.77	4.64	6.01
concat	0.39	8.04	4.82	4.79
gated	0.12	7.55	4.63	5.11
max	0.12	6.01	3.15	4.72

VI. CONCLUSIONS

This paper presented ADM, a lightweight multi-stream detector that combines SSL, mel-filterbank, and STFT phase features to model temporal, spectral, and phase artifacts in audio deepfakes. By fusing branch-level evidence, ADM captures complementary cues related to synthesis, vocoding, compression, and recording conditions.

Experiments on ASVspoof 2019 LA, ASVspoof 2021 LA, ASVspoof 2021 DF, and ITW show competitive performance, with a configuration locked in advance achieving EERs of 0.12%, 6.01%, 3.15%, and 4.72%, respectively. Ablations indicate that SSL layer selection strongly affects generalization, the temporal stream is the strongest individual cue, phase evidence contributes where the vocoder phase structure survives to the evaluation signal, and max fusion is the most robust way to combine them.

Future work will explore adaptive stream selection, stronger phase-aware representations, training for compressed or unconstrained audio, voice-uniqueness-inspired biometric features [4], and explainable analyses such as branch attribution and time–frequency saliency under broader cross-corpus and open-set evaluations.

ACKNOWLEDGMENT

Part of this work was supported by JSPS KAKENHI Grant Number JP24K0296.

REFERENCES

[1] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, no. 7, 2025, ISSN: 14248220. DOI: 10.3390/s25071989.

[2] A. Firc, K. Malinka, and P. Hanáček, "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors," *Heliyon*, vol. 9, no. 4, e15090, 2023, ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e15090.

[3] N. M. Müller and W. H. Choong, *Eroding trust in real speech: A large-scale study of human audio deepfake perception*, 2026.

[4] R. Singh and B. Raj, "Human Voice is Unique," 2025.

[5] A. Chaiwongyen, N. Songsriboonsit, and S. Duangpummet, "Contribution of Timbre and Shimmer Features to Deepfake Speech Detection," November, 2022, pp. 97–103, ISBN: 9786165904773.

[6] Y. El Kheir, Y. Samih, S. Maharjan, T. Polzehl, and S. Möller, "Comprehensive Layer-wise Analysis of SSL Models for Audio Deepfake Detection," in *Find. Assoc. Comput. Linguist. NAACL 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 4070–4082. DOI: 10.18653/v1/2025.findings-naacl.227.

[7] M. Yang et al., "Generalizable Audio Deepfake Detection via Hierarchical Structure Learning and Feature Whitening in Poincaré sphere," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 2255–2259, 2025, ISSN: 29581796. DOI: 10.21437/Interspeech.2025-1594.

[8] E. Coletta, D. Salvi, V. Negroni, D. U. Leonzio, and P. Bestagini, "Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching," in *EUSIPCO*, Mar. 2025, ISBN: 2200308000.

[9] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, *Unsupervised cross-lingual representation learning for speech recognition*, 2020.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020, ISSN: 19393539. DOI: 10.1109/TPAMI.2018.2858826.

[11] J. Yamagishi et al., "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," Tech. Rep., 2019. DOI: 10.21437/Interspeech.2019-1623.

[12] X. Liu et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2507–2522, 2023, ISSN: 23299304. DOI: 10.1109/TASLP.2023.3285283.

[13] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, no. September, pp. 2783–2787, 2022, ISSN: 29581796. DOI: 10.21437/Interspeech.2022-108.

[14] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, *PartialSpoof Database - Partially Spoofed Audio Dataset for Anti-spoofing*, version 1.2, May 2021. DOI: 10.5281/zenodo.5766198.

[15] P. A. T. Flórez, R. Manrique, and B. P. Nunes, "HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-Augus, no. August, pp. 1963–1967, 2023, ISSN: 29581796. DOI: 10.21437/Interspeech.2023-2272.

[16] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," *2019 10th Int. Conf. Speech Technol. Human-Computer Dialogue, SpeD 2019*, 2019. DOI: 10.1109/SPED.2019.8906599.

[17] J. Yi et al., *Half-truth: A partially fake audio detection dataset*, 2023.

[18] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A Survey on Speech Deepfake Detection," *ACM Comput. Surv.*, vol. 57, no. 7, Feb. 2025, ISSN: 0360-0300. DOI: 10.1145/3714458.

[19] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, "One-Class Knowledge Distillation for Spoofing Speech Detection," in *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process.*, IEEE, Apr. 2024, pp. 11 251–11 255, ISBN: 979-8-3503-4485-1. DOI: 10.1109/ICASSP48485.2024.10446270.

[20] X. Wang and J. Yamagishi, "Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2023-June, 2023, ISBN: 9781728163277. DOI: 10.1109/ICASSP49357.2023.10094779.

[21] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "Generalizable Audio Deepfake Detection via Latent Space Refinement and Augmentation," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2025, ISSN: 15206149. DOI: 10.1109/ICASSP49660.2025.10888328.

[22] S. B. Kim et al., "Enhancing Audio Deepfake Detection by Improving Representation Similarity of Bonafide Speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 2250–2254, 2025, ISSN: 29581796. DOI: 10.21437/Interspeech.2025-422.

[23] Y. Zhu, S. Koppisetti, T. Tran, and G. Bharaj, "SLIM: Style-Linguistics Mismatch Model for Generalized Audio Deepfake Detection," *Adv. Neural Inf. Process. Syst.*, vol. 37, no. NeurIPS, 2024, ISSN: 10495258. DOI: 10.52202/079017-2168.