Decay-Region Group Delay as a Forensic Cue for AI-Generated Impulsive Sounds

JaeHyeong Chang*, Chengzhe Sun*, and Siwei Lyu†
Institute for Artificial Intelligence and Data Science
University at Buffalo, Buffalo, NY, USA
{jchang46, csun22, siweilyu}@buffalo.edu
*Equal contribution. †Corresponding author.

***Abstract*—We investigate whether AI-generated impulsive sounds can be distinguished from real ones through group delay analysis. Our central finding is that AI-generated impulsive sounds show near-identical onset-region group-delay distributions but exhibit measurably different group-delay behavior in the late decay region: decay-region KL divergence reaches 0.322 compared to near-zero onset divergence (0.022). Cross-band GD variability achieves single-feature AUC = 0.720, and a Random Forest (RF) over nine decay-region features reaches AUC = 0.884 under sample-disjoint evaluation. A group delay map used as a standalone 2D input to CNN classifiers achieves 90–94% accuracy, demonstrating that group delay carries substantial discriminative information. Under generator hold-out, CNN and transformer classifiers show highly variable AUC (0.457–0.918). The group delay RF achieves the highest average hold-out accuracy among the evaluated methods (66.7%) and avoids extreme below-random collapse, although its average AUC (0.731) is lower than CNN avg (0.762) and AST (0.772). Parameter sensitivity analysis across 27 STFT configurations confirms that the RF AUC remains stable (0.700–0.847, std = 0.035). These results suggest that decay-region group delay can serve as a physically interpretable forensic cue that complements magnitude-based classifiers, while broader validation remains necessary.**

***Index Terms*—AI-generated impulsive sounds, audio forensics, group delay, decay-region analysis, generator generalization.**

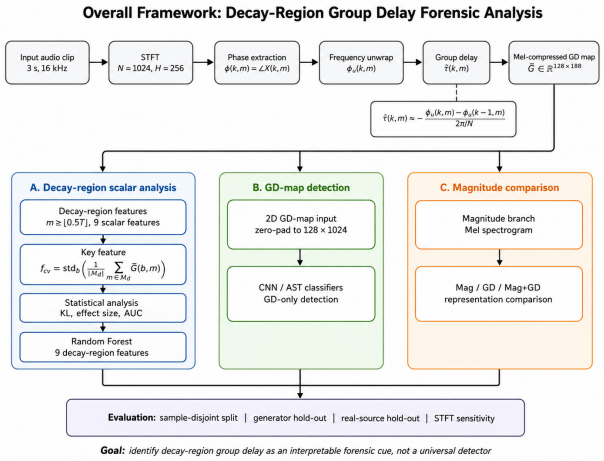

Fig. 1. Overall framework for decay-region group delay forensic analysis. The pipeline processes a 3-second audio clip through STFT and phase unwrapping to extract a mel-compressed group delay map $\bar{G}$, which feeds three analysis branches: (A) decay-region scalar features with Random Forest classification, (B) GD-map-based CNN/AST detection, and (C) magnitude comparison. All branches are evaluated under sample-disjoint split, generator hold-out, real-source hold-out, and STFT sensitivity protocols.

I. Introduction

Generative audio models such as ElevenLabs, Stable Audio [1], and AudioLDM2 [2] can now synthesize impulsive sounds with increasing realism, raising concerns for forensics and misinformation detection [3], [4]. Unlike speech deepfake detection, AI-generated environmental sound forensics remains comparatively underexplored. We study this problem through the lens of *group delay* — the negative frequency-derivative of the phase spectrum — which characterizes how different frequency components are temporally delayed by an acoustic system. Real impulsive sounds are physically constrained: shock wave propagation, environmental reflection, and energy dissipation impose structured group delay patterns especially in the late decay region. We hypothesize that generative models, optimized for perceptual plausibility rather than physical consistency, may not fully reproduce this behavior.

The primary contribution of this paper is not a new state-of-the-art detector, but rather a demonstration that **decay-region group delay is an important forensic cue** for AI-generated impulsive sound detection. We make four contributions. First, we identify decay-region group delay inconsistency as a forensic cue observed across the three tested generators, supported by KL divergence, effect size, and classifier experiments. Second, group delay maps achieve 90–94% standalone detection accuracy in the sample-disjoint setting. Third, we show that CNN and transformer classifiers exhibit highly variable AUC under generator hold-out (0.457–0.918), while the group delay RF avoids extreme below-random collapse and achieves higher average accuracy, suggesting that physically grounded features may provide complementary forensic cues. Fourth, parameter sensitivity analysis across 27 STFT configurations confirms robustness of the RF classifier (AUC std = 0.035).

II. Related Work

A. Audio Deepfake Detection

The detection of AI-generated speech has been extensively studied through the ASVspoof challenge series [3], which has driven development of countermeasures based on spectral, cepstral, and phase features. Recent surveys [4] highlight that magnitude-based classifiers achieve high in-distribution

accuracy but often fail to generalize across unseen spoofing systems. Our work extends this line of research to the less-studied domain of impulsive sounds, where physical constraints on signal decay impose additional forensic structure.

B. Phase and Group Delay in Audio Analysis

The importance of phase information in signal reconstruction and analysis has long been recognized [5]. Group delay — the negative frequency-derivative of the phase spectrum — has been applied to speaker gender identification [6] and speech processing [7], where it captures temporal structure not visible in magnitude spectrograms. Phase-based features including group delay have been explored for speech deepfake detection [8], demonstrating that phase information complements magnitude features. Our work extends this to impulsive sound forensics, focusing specifically on decay-region group delay inconsistencies rather than full-spectrum phase features. To the best of our knowledge, decay-region group delay has not been systematically studied as a forensic cue for AI-generated impulsive sounds.

C. Generative Audio Models

Recent text-to-audio and sound synthesis models such as AudioLDM2 [2] and Stable Audio [1] can generate perceptually convincing impulsive sounds. These models are optimized for perceptual quality rather than physical accuracy, motivating our hypothesis that they may fail to reproduce physically constrained signal properties such as decay-region group delay structure.

III. Dataset

We construct 15,000 clips (7,500 real / 7,500 fake) at 16 kHz, 3 seconds, focusing on explosion and impact-type impulsive sounds. Table I details sources and augmentation ratios. Real samples are drawn from FSD50K [9] and SESA [10]; fake samples from ElevenLabs, Stable Audio, and AudioLDM2. Augmentation applies time-shift, SpecAugment masking [11], and noise mixing identically to both classes.

A **sample-disjoint** split ensures augmented variants of the same source file remain in the same partition, preventing augmentation leakage across Train/Val/Test = 12000/1500/1500 (balanced). We acknowledge that the high augmentation ratio for SESA (6.4×) and the source mismatch between real (FSD50K/SESA field recordings) and fake (synthesized audio) may introduce domain-level biases beyond real/fake distinction; we address this in the Limitations.

Because large-scale, controlled real impulsive sound datasets are difficult to obtain publicly, our dataset necessarily combines available real impulsive sound sources with generated samples. We therefore frame this work not as a definitive benchmark for universal impulsive sound deepfake detection, but as an initial forensic study showing that decay-region group delay contains discriminative and physically interpretable cues. Accordingly, results should be interpreted as evidence for a forensic cue rather than corpus-independent benchmark performance.

TABLE I
Dataset construction.

Class	Source	Orig.	Final	×
Real	FSD50K	1,122	3,750	3.3
Real	SESA	585	3,750	6.4
Fake	ElevenLabs	1,119	2,500	2.2
Fake	Stable Audio	1,000	2,500	2.5
Fake	AudioLDM2	1,000	2,500	2.5

TABLE II
Real-source hold-out: GD RF performance under real-source shift.

Condition	AUC	Acc (%)
FSD50K → SESA (hold-out)	0.849	76.47
SESA → FSD50K (hold-out)	0.857	77.00
Mixed → Mixed (reference)	**0.878**	**79.33**

A. Real-Source Bias Control

To verify that our results are not driven by source-specific recording artifacts in the real set, we conducted a real-source hold-out experiment. In the FSD50K→SESA condition, the GD RF is trained using FSD50K real samples and a sample-disjoint subset of fake samples, and tested using SESA real samples and held-out fake samples. The reverse condition (SESA→FSD50K) swaps only the real source while preserving sample-disjoint fake partitions. Table II reports GD RF performance under each condition.

The GD RF achieves AUC = 0.849–0.857 under real-source shift vs. AUC = 0.878 for the mixed reference. A sanity check classifying FSD50K vs. SESA real samples yields AUC = 0.737, confirming partial source separability; results should therefore be interpreted as initial evidence for a forensic cue rather than a source-independent detector.

IV. Group Delay Representation

A. Group Delay Definition

Group delay is formally defined as the negative derivative of the phase spectrum with respect to angular frequency [7], [6]:

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega}. \tag{1}$$

This quantity characterizes the time delay experienced by each frequency component as it passes through a system. In real impulsive sounds, $\tau(\omega)$ in the decay region reflects structured acoustic propagation constraints imposed by physical wave propagation and environmental response. The phase spectrum carries complementary information to magnitude [5], and we approximate Eq. (1) via finite differences of the unwrapped STFT phase [12], as described below.

B. STFT and Phase

Given a 3-second waveform $x[n]$ at 16 kHz, the Short-Time Fourier Transform is:

$$X(k,m) = \sum_{n=0}^{N-1} x[n+mH]\, w[n]\, e^{-j2\pi kn/N}, \tag{2}$$

where $N = 1024$, hop $H = 256$, and $w[n]$ is a Hann window. The instantaneous phase is:

$$\phi(k, m) = \angle X(k, m). \tag{3}$$

C. Group Delay Approximation

We discretize Eq. (1) via finite differences of the frequency-unwrapped phase ϕ_u:

$$\hat{\tau}(k, m) \approx -\frac{\phi_u(k, m) - \phi_u(k-1, m)}{2\pi/N}. \tag{4}$$

Values are clipped at $\tau_{\max} = 500$ samples (≈31 ms at 16 kHz) for numerical stability:

$$\tilde{\tau}(k, m) = \text{clip}\big(\hat{\tau}(k, m),\, -\tau_{\max},\, \tau_{\max}\big). \tag{5}$$

D. Mel-compressed Group Delay Map

As a practical representation, a mel filterbank $\{H_b(k)\}_{b=1}^{128}$ compresses the frequency axis:

$$G(b, m) = \sum_k H_b(k)\, \tilde{\tau}(k, m), \tag{6}$$

normalized to $[-1, 1]$:

$$\bar{G}(b, m) = \frac{G(b, m)}{\max_{b,m} |G(b, m)| + \epsilon}. \tag{7}$$

For a 3-second clip at 16 kHz with hop $H = 256$, this yields $\bar{G} \in \mathbb{R}^{128\times188}$, zero-padded to 128×1024 only to match the input resolution expected by image-based classifiers; scalar GD features are computed from the original 128×188 representation. The same temporal zero-padding procedure was applied consistently across all compared image-based representations (Mag, GD, Mag+GD) to avoid representation-specific input-size differences.

E. Decay-Region Features

We define the late temporal region as frames $m \geq \lfloor 0.5T \rfloor$, where $T = 188$ is the total number of frames, and refer to it as the decay region, as impulsive energy predominantly occurs in the earlier portion of the three-second clips. The most discriminative scalar feature is **cross-band GD variability**:

$$f_{\text{cv}} = \text{std}_b\left(\frac{1}{|\mathcal{M}_d|}\sum_{m\in\mathcal{M}_d} \bar{G}(b, m)\right), \tag{8}$$

measuring the standard deviation of per-band mean group delay in the decay region. AI-generated samples show *higher* cross-band GD variability than real impulsive sounds, suggesting less stable or less physically consistent phase-derivative behavior across frequency bands. Real impulsive sounds show *lower* and more structured variability, consistent with physically constrained decay. The **Decay-to-Onset KL divergence** quantifies temporal distributional shift:

$$f_{\text{KL}} = D_{\text{KL}}(p_{\text{decay}} \,\|\, p_{\text{onset}})\,. \tag{9}$$

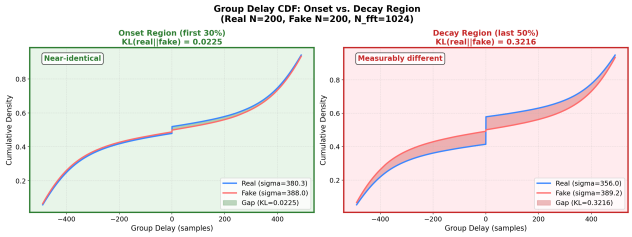

Fig. 2. Group delay CDF: onset region (left, KL = 0.022) shows near-identical real/fake distributions, while the decay region (right, KL = 0.322) reveals measurably different distributions. Shaded area indicates the gap between real and fake CDFs.

TABLE III
KL DIVERGENCE: REAL VS. FAKE.

Region / Band	**KL**
Overall	0.2426
Low freq (0–500 Hz)	0.2339
Mid freq (500–2000 Hz)	0.2417
High freq (2000–8000 Hz)	0.2437
Onset (first 30%)	0.022
Decay (last 50%)	**0.322**

V. Group Delay Forensic Analysis

A. Global vs. Decay-Region Separability

Figure 2 shows onset and decay region CDFs; Table III additionally reports global and frequency-band KL values, all of which are approximately 0.24, indicating limited global separability. As shown in Fig. 2 (left panel), the onset region shows near-identical distributions (KL = 0.022), indicating that onset GD alone does not distinguish the two classes. The decay region (right panel) reveals measurably larger divergence (KL = 0.322), localizing forensic artifacts in late decay.

B. Scalar Feature Analysis

Table IV reports effect sizes and AUC ($N = 15{,}000$, 5-fold CV, all $p < 0.001$). Cross-band GD variability (f_{cv}) achieves the highest discriminative power (Cohen's $d = 0.646$, AUC = 0.720): AI-generated samples show higher cross-band GD variability than real impulsive sounds, suggesting less stable or less physically consistent phase-derivative behavior. Real impulsive sounds show lower and more structured variability. The Decay-to-Onset KL divergence shows striking distributional separation (real: $\mu = 0.923$, $\sigma = 1.795$; all generators: $\mu \approx 0.010$, $\sigma = 0.010$). Its standalone AUC of 0.570, however, reflects the high variance in the real set ($\sigma = 1.795$): while the mean difference is large, the real distribution is highly spread, reducing class separability at the individual sample level. The Decay-to-Onset KL should therefore be interpreted as a **distribution-level forensic cue** rather than a strong standalone sample-level classifier. A Random Forest [13] over all nine features achieves AUC = 0.884 under sample-disjoint evaluation (all 15,000 samples; note this differs from the mixed real-source reference AUC = 0.878 in Table II, which uses an 80/20 train/test split on a subset). Note

TABLE IV
DECAY-REGION SCALAR FEATURES ($N = 15{,}000$, ALL $p < 0.001$).

Feature	Cohen's d	Cliff's δ	AUC
Decay variance	0.588	0.261	0.651
Decay entropy	0.502	0.010	0.521
Decay abs. mean	0.578	0.204	0.628
Decay/onset var. ratio	0.036	0.096	0.573
Decay→onset KL	0.482	0.114	0.570
Decay→onset shift	0.084	0.102	0.530
Cross-band GD var. (f_{cv})	**0.646**	**0.424**	**0.720**
Temporal fluct.	0.531	0.054	0.566
Global variance	0.604	0.306	0.675

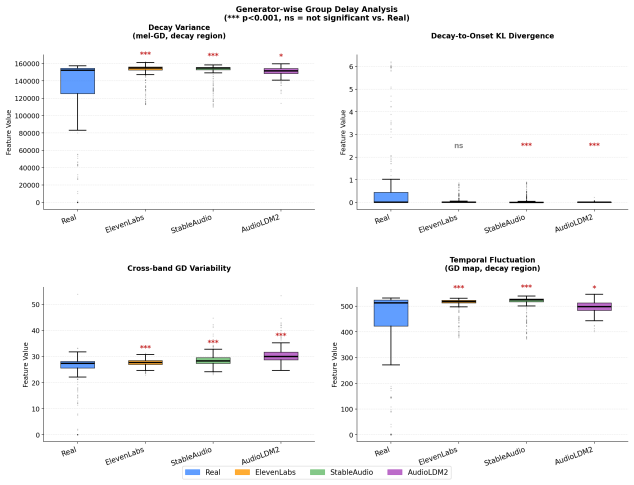

Fig. 3. Generator-wise group delay analysis ($^{***}p < 0.001$, $^{*}p < 0.05$, ns = not significant vs. real). The most pronounced separation appears in Decay-to-Onset KL (top-right), where Stable Audio and AudioLDM2 differ significantly ($p < 0.001$) while ElevenLabs shows a weaker, non-significant trend. The compact fake distributions may partly reflect generator-specific homogeneity; these results are interpreted as evidence of a candidate forensic cue rather than a source-independent detector.

that the KL values in Fig. 2 measure distributional divergence between real and fake samples within each temporal region, whereas the Decay-to-Onset KL feature (f_{KL}) in Table IV measures within-sample temporal shift between onset and decay distributions.

C. Consistency Across Generators

Figure 3 shows group delay feature distributions across the three tested generators. The most visually pronounced separation appears in the Decay-to-Onset KL feature (top-right), where Stable Audio and AudioLDM2 cluster near zero while real samples show large variability ($\mu = 0.923$, $p < 0.001$); ElevenLabs shows a weaker but directionally consistent trend (not significant on this feature alone). Cross-band GD variability and decay variance show statistically significant but partially overlapping differences across the tested generators, with varying significance levels. These results suggest a systematic tendency in decay-region group delay behavior of the tested generators, though the pattern is not uniformly strong across all features and generators.

TABLE V
STFT PARAMETER SENSITIVITY (27 CONFIGS, AVG. OVER CLIP VALUES).

N_{fft}	hop	f_{cv} AUC	RF AUC
512	128	0.710	0.801
512	256	0.557	0.755
512	512	0.569	0.768
1024	128	0.730	0.841
1024	256*	0.687	0.828
1024	512	0.584	0.752
2048	128	0.728	0.822
2048	256	0.725	0.804
2048	512	0.696	0.777

*Default configuration used in the main experiments.

TABLE VI
FEATURE REPRESENTATION (SAMPLE-DISJOINT, 15 EPOCHS, ACCURACY %).

Model	Mag	GD only	Mag+GD
ResNet50	97.00	90.40	**97.40**
EfficientNet-B2	**98.93**	92.67	98.33
CNN14	**98.93**	94.20	98.73
AST	**98.27**	93.67	97.60

D. Parameter Sensitivity Analysis

We evaluated group delay features across 27 STFT configurations: $N \in \{512, 1024, 2048\}$, hop $\in \{128, 256, 512\}$, clip $\in \{250, 500, 1000\}$ ($N_{\text{sample}} = 500$ per class, 5-fold CV). Table V reports results averaged over clip values for each (N, hop) pair.

The RF AUC ranges 0.700–0.847 (std = 0.035) across all 27 configurations. Cross-band GD variability shows higher sensitivity (std = 0.069), with lower performance for $N = 512$; for $N \geq 1024$ the AUC stabilizes at 0.686–0.734. The default ($N = 1024$, hop = 256) is representative of the stable region.

VI. DETECTION EXPERIMENTS

A. Feature Representation Comparison

We trained ResNet50 [14], EfficientNet-B2 [15], CNN14 [16], and AST [17], [18] with three inputs for 15 epochs under the sample-disjoint protocol: **Mag** (mel-spectrogram, 1-ch), **GD** (group delay map $\bar{G}$, 1-ch), **Mag+GD** (2-ch). Results are shown in Table VI.

Models were trained with AdamW [19]. **GD-only achieves 90–94%** standalone accuracy, demonstrating substantial discriminative information in group delay maps. Mag+GD improves ResNet50 by 0.4 pp; higher-capacity models show a ceiling effect. The gap between in-distribution and hold-out performance suggests that only part of GD's discriminative information transfers across unseen generators.

B. Generator Hold-out Evaluation

Fake samples from one generator were excluded entirely from training and used only for testing. Each test set was balanced between real and the held-out fake generator. Table VII reports accuracy and AUC for all evaluated models.

TABLE VII
GENERATOR HOLD-OUT: ACC (%) AND AUC. **BOLD**: BEST ACC OR BEST AUC PER HOLD-OUT CONDITION. CNN AVG DENOTES THE AVERAGE OVER RESNET50, EFFICIENTNET-B2, AND CNN14; AST IS REPORTED SEPARATELY AS A TRANSFORMER-BASED MODEL.

Hold-out	Method	Acc	AUC
AudioLDM2	ResNet50	**61.2**	**0.714**
	EfficientNet-B2	54.7	0.586
	CNN14	55.2	0.457
	AST	57.2	0.659
	GD RF	58.6	0.580
Stable Audio	ResNet50	64.1	0.829
	EfficientNet-B2	63.1	0.907
	CNN14	69.0	**0.918**
	AST	69.1	0.906
	GD RF	**73.6**	0.832
ElevenLabs	ResNet50	59.1	**0.825**
	EfficientNet-B2	54.3	0.820
	CNN14	54.1	0.806
	AST	53.6	0.753
	GD RF	**67.8**	0.782
Average	CNN avg (ResNet/EffNet/CNN14)	59.4	0.762
	AST	59.9	**0.772**
	GD RF	**66.7**	0.731

CNN and transformer (AST) classifiers show **highly variable AUC** (0.457–0.918): CNN14 reaches AUC = 0.918 for Stable Audio hold-out but collapses to AUC = 0.457 for AudioLDM2 — below random. AST similarly ranges from 0.659 to 0.906, suggesting that both CNN and transformer architectures may rely substantially on generator-specific artifacts rather than fully generator-generalizable real/fake cues. The GD RF does not achieve the highest average AUC (0.731 vs. 0.762 for CNN avg and 0.772 for AST), but it avoids the extreme below-random collapse observed for CNN14 (AUC = 0.457) and yields the highest average accuracy (66.7% vs. 59.4% for CNN avg and 59.9% for AST).

The AudioLDM2 hold-out is the most challenging condition for all methods; the GD RF also struggles here (AUC = 0.580), suggesting that group delay cues do not transfer uniformly across generator architectures. These results indicate that group delay features provide complementary forensic cues, though further validation is needed.

C. Supplementary: Phase-Flow CRNN Attention

A lightweight Phase-Flow CRNN [20] (2.24M parameters) processes temporal phase-flow via BiGRU [21] with temporal attention, achieving 92.33% in-distribution accuracy. Its attention concentrates on late decay regions (Figure 4), providing qualitative support for the decay-region group delay finding.

VII. CONCLUSION

Real impulsive sounds undergo physically constrained group delay evolution from onset to decay (Eq. 1). The tested generators show substantially lower within-sample Decay-to-Onset KL values than real impulsive sounds, which exhibit greater variability. GD-only maps achieve 90–94% standalone accuracy in the sample-disjoint setting. Under generator hold-out, CNN and transformer classifiers show highly variable

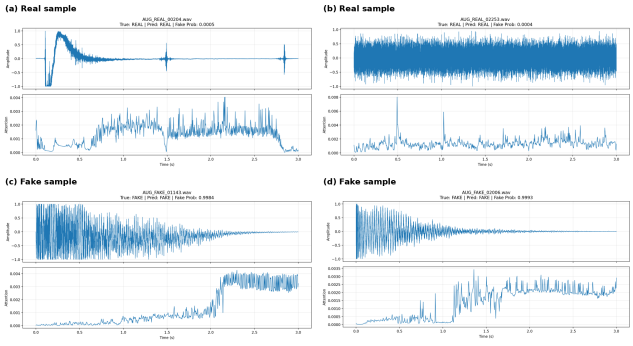

Fig. 4. Phase-Flow CRNN temporal attention (note: y-axis scales differ across panels). Real samples (a,b) show relatively diffuse or transient attention patterns, whereas fake samples (c,d) exhibit concentrated attention in the late decay region. This pattern is consistent with the decay-region group delay findings.

AUC (0.457–0.918), while the group delay RF achieves the highest average accuracy and avoids extreme below-random collapse, despite not achieving the highest average AUC. Parameter sensitivity analysis confirms RF robustness across 27 STFT configurations (AUC std = 0.035). Phase-Flow CRNN attention provides qualitative support for the decay-region locus. These results show that *high in-distribution accuracy does not imply generator-generalizable forensic understanding*. Phase-based representations — specifically decay-region group delay — offer a physically interpretable and complementary lens for AI-generated impulsive sound forensics.

VIII. LIMITATIONS AND FUTURE WORK

Several limitations should be acknowledged. Real data relies on augmentation (SESA: 6.4×), and source/domain mismatch may introduce biases (FSD50K vs. SESA GD AUC = 0.737), though the real-source hold-out experiment in Section III-A shows this bias is modest. Only three generators are evaluated, and generalization is not yet uniform: the GD RF collapses to near-chance performance on AudioLDM2 hold-out (AUC = 0.580), while remaining more stable for Stable Audio and ElevenLabs. This pattern, together with the Decay-to-Onset KL feature's gap between strong distribution-level separation (real $\mu = 0.923$ vs. generators $\mu \approx 0.010$) and weaker sample-level discriminability (AUC = 0.570), suggests decay-region group delay is best understood as a distribution-level forensic signature rather than a universal sample-level classifier – a distinction we believe is itself a useful finding for future detector design. The STFT parameter sensitivity analysis, based on a smaller subset (500 samples per class), and the absence of adversarial robustness testing (e.g., against phase randomization) are natural next steps, alongside extending evaluation to sound categories beyond explosions and impacts and to clip durations other than 3 seconds. $N_{\text{fft}} \geq 1024$ is recommended based on current results. We view these as concrete, addressable directions rather than fundamental obstacles to the core finding that decay-region group delay carries forensically meaningful signal.

REFERENCES

[1] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *Proc. 41st International Conference on Machine Learning (ICML)*, 2024.

[2] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.

[4] J. Yi, J. Tao, R. Fu, Z. Tian, C. Yuan, J. Wang, Y. Zhao, D. Dai, C. Li, and G. An, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[5] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[6] K.-H. Lee, S.-I. Kang, J.-H. Song, and J.-H. Chang, "Group delay function for improved gender identification," in *Proc. Interspeech*, Brisbane, Australia, 2008, p. 1513.

[7] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 3, pp. 610–622, 1984.

[8] J. Xue, C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao, "Audio deepfake detection based on a combination of F0 information and real plus imaginary spectrogram features," in *Proc. 30th ACM International Conference on Multimedia (MM)*, Lisbon, Portugal, 2022, pp. 4779–4788.

[9] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[10] T. Spadini, "Sound events for surveillance applications (SESA)," 2019, zenodo dataset, Version 1.0.0.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[12] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[16] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[17] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[20] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2392–2396.

[21] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.