The Perceived Fragility of Explanations in Audio Models: Manipulation of Attribution with Unchanged Predictions

Piotr Kitłowski [1] **Dominik Wiącek** [1] **Mateusz Modrzejewski** [1]

Abstract

This paper investigates the fragility of post-hoc explanation methods in audio deepfake detection. While previous work on explanation manipulation focused on images using standard L_p metrics, we introduce a psychoacoustic framework that optimizes inaudible perturbations to decouple model attributions from final classifications. We evaluate this vulnerability across state-of-the-art architectures under strict prediction-preserving constraints. By evaluating the manipulation cost through domain-specific perceptual audio quality metrics alongside explanation alignment criteria, our framework demonstrates that an adversary can systematically distort automated explanation heatmaps while preserving the predicted deepfake label. Full code available at: https://github.com/cncPomper/Audio-XAI

1. Introduction

The proliferation of synthetic audio, driven by advances in generative models, has made deepfake detection critical. To foster trust in these systems, Explainable AI (XAI) methods are deployed to highlight the acoustic artifacts driving a model's decision. However, this reliance introduces a security vulnerability: the fragility of the explanations themselves. If an adversary can manipulate the system to provide a deceptive justification while maintaining a prediction, the credibility of the interpretability framework is compromised.

While the fragility of attribution maps has been extensively demonstrated in the computer vision domain (Ghorbani et al., 2019; Dombrowski et al., 2019; Heo et al., 2019), its implications for audio models remain largely unaddressed. Furthermore, vision-based attacks measure the manipulation cost using L_p norms, which do not correlate with human auditory perception (Abdullah et al., 2021). In the audio domain, an attack on interpretability is viable if the adversarial perturbation remains imperceptible to the ear. Early investigations into the plausibility of audio explanations under adversarial conditions (Prinz et al., 2023) highlighted the need for more domain-specific, perceptually bounded constraints. More recently, efforts to establish robust baselines for audio deepfake explanations have exposed structural limitations in post-hoc methods like LRP (Grinberg et al., 2025). We expand on this line of thought by demonstrating that these limitations extend beyond inherent inaccuracy; rather, the explanations themselves can be systematically and imperceptibly manipulated by an adversary.

To bridge this gap, this study investigates the stability of post-hoc explanation methods against masked perturbations. We aim to determine whether XAI techniques provide robust interpretations of audio data or can be decoupled from the classifier's decision boundary.

Our main contributions are as follows:

- We introduce a novel optimization framework that successfully adapts explanation-targeted adversarial attacks to the audio domain.
- We incorporate a dynamic psychoacoustic masking threshold into the loss function, encouraging large attribution-map changes while penalizing perturbations that exceed a psychoacoustic masking threshold or change the model's final prediction.
- We provide a focused empirical evaluation across diverse architectures using the SONICS deepfake dataset (Rahman et al., 2025), leveraging domain-specific perceptual metrics to assess perceptual transparency.

2. Methodology

2.1. Analyzed XAI Methods

We investigate two paradigms of post-hoc attribution: **Grad-CAM** (Selvaraju et al., 2017) and **Layer-wise Relevance Propagation (LRP)** (Bach et al., 2015). While Grad-CAM highlights regions using final-layer gradients, LRP follows a conservation rule to backpropagate relevance scores to

[1]Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland. Correspondence to: Piotr Kitłowski <01149528@pw.edu.pl>, Dominik Wiącek <01169141@pw.edu.pl>, Mateusz Modrzejewski <mateusz.modrzejewski@pw.edu.pl>.

the input spectrogram. Relying on both first-order (Grad-CAM) and structurally constrained (LRP) methods allows us to examine whether explanation fragility appears across audio architectures rather than being an artifact of a single algorithm.

2.2. Dataset and Target Models

All experiments were conducted using the **SONICS** dataset, a large-scale corpus designed for audio deepfake detection. To ensure the observed explanation fragility is not an artifact of a single architecture, we evaluated three models employing distinct feature extraction paradigms on time-frequency representations. As a baseline for classical convolutional architectures that rely on local spectrogram patterns, we used **VGGish** (Hershey et al., 2017). To represent modern, self-attention-based architectures capable of modeling global context, we selected the **Audio Spectrogram Transformer (AST)** (Gong et al., 2021). Finally, we used **SpecTTTra**, denoted as `spectttra-gamma-5s` (Rahman et al., 2025), a recent attention-based architecture specifically proposed alongside the SONICS dataset for detecting synthetic music by capturing long-range temporal dependencies. Analyzing this diverse set enables a comparative assessment of the stability of attribution maps across convolutional and token-based processing regimes.

2.3. Audio Sample Perturbation

We randomly sampled 100 recordings from the dataset, independent of the label, and applied three attack methodologies targeting the input space x with a bounded perturbation δ. Since attacking attribution mechanisms requires second-order derivatives, the Adam optimizer was used across all methods to ensure convergence.

1. Standard Projected Gradient Descent (PGD): As a baseline, we implemented an L_∞-bounded PGD attack (Madry et al., 2018) aimed solely at minimizing the structural similarity between the original and perturbed attribution maps, without considering perceptual audio quality.

2. X-Shift Attack (Adapted): Originally designed for vision-language models (Babadi & Karimipour, 2026), we adapted this spatial displacement strategy to the audio domain. The objective forces the explanation to assign maximum relevance to a designated, irrelevant target region M_{target}, steering it away from the originally salient time-frequency patches M_{orig}.

3. Psychoacoustic Noise Modeling (Ours): To constrain perturbations according to a psychoacoustic masking model, we propose a custom optimization framework. The total loss function $\mathcal{L}(\delta)$ forces map displacement while maintaining rigorous acoustic and predictive constraints:

$$\begin{aligned}\mathcal{L}(\delta) = &\mathcal{L}_{explain}(\delta) \\ &+ \lambda_{aud}\mathcal{L}_{audibility}(\delta) \\ &+ \lambda_{pred}\mathcal{L}_{pred\_preserve}(\delta)\end{aligned} \tag{1}$$

The $\mathcal{L}_{explain}(\delta)$ term minimizes the cosine similarity between the original and perturbed attribution maps. Crucially, to operationalize a psychoacoustic audibility constraint, the threshold penalty is formalized as $\mathcal{L}_{audibility}(\delta) = \mathbb{E}[\max(0, 20\log_{10}|\mathcal{F}(\delta)| - T(x))^2]$, where $T(x)$ is the static masking threshold pre-computed from the clean input. This exclusively penalizes the spectral energy of δ that exceeds the bounds of human perception. Because manipulating explanations requires differentiating through gradients (second-order derivatives), we optimize the total loss using Adam rather than standard sign-based methods. This is coupled with a margin-based hinge loss ($\mathcal{L}_{pred\_preserve}$) to penalize changes in the original prediction and a hard waveform amplitude constraint $\delta \in [-\varepsilon, \varepsilon]$.

2.4. Evaluation Metrics

To assess perceptual transparency, audio fidelity was evaluated across the SONICS dataset using **PEAQ** (Thiede et al., 2000), **ViSQOL** (Chinen et al., 2020), **Zimtohrli** (Alakuijala et al., 2025), **CDPAM** (Manocha et al., 2021), **PESQ** (Rix et al., 2001), and **STOI** (Taal et al., 2011). High values across these indicators corroborate the minimal energy of the injected noise. Finally, explanation fragility was quantified by measuring the discrepancy between original and perturbed attribution maps using **Cosine Similarity** and **Top-10 Overlap**.

2.4.1. Audio Fragility Score

To summarize the vulnerability of explanation maps, we introduce the Audio Fragility Score (AFS_{stable}). Unlike binary attack success rates, AFS_{stable} offers a continuous measure of attribution displacement, conditioned on preserving the predicted class and maintaining high perceptual quality. For a given sample i, the metric is defined as:

$$\begin{aligned}AFS_i^{stable} &= \left(1 - \frac{C_i + T_i}{2}\right)\mathbf{1}[\hat{y}_i^{orig} = \hat{y}_i^{adv}]Q_i, \\ C_i &= \cos(A_i^{orig}, A_i^{adv}), \\ T_i &= \text{Top10}(A_i^{orig}, A_i^{adv}).\end{aligned} \tag{2}$$

The first term measures the magnitude of the explanation shift by averaging the cosine similarity and Top-10 overlap between the original (A_i^{orig}) and perturbed (A_i^{adv}) attribution maps. The indicator function $\mathbf{1}[\cdot]$ acts as a strict penalty, zeroing the score if the predicted class $\hat{y}$ changes. Finally, $Q_i \in [0, 1]$ represents the normalized perceptual quality score of the adversarial audio. Consequently, an AFS^{stable} approaching 1.0 signifies a highly successful, imperceptible

manipulation of the attribution map. In contrast, a score of 0.0 indicates a failure to shift the explanation, an altered model prediction, or unacceptable acoustic degradation.

3. Results

3.1. Global Results

A practical attack on explainability must remain imperceptible. As shown in Table 1, unconstrained optimization methods such as **PGD** severely degrade audio quality (e.g., PESQ ≈ 2.8) and introduce easily audible artifacts. While **X-Shift** maintains acceptable fidelity, our **Psychoacoustic** framework preserves high objective perceptual quality (ViSQOL > 4.1, CDPAM ≥ 0.98) by bounding noise within human masking thresholds. These results suggest that explanations for deepfake detection can be substantially manipulated while preserving predictions and maintaining high objective perceptual quality. To systematically compare the vulnerability of the models and the efficacy of the adversarial strategies, we aggregated the sample-level AFS^{stable} scores into a unified ranking system. For each test sample, the model-attack combinations were ranked based on their explanation displacement, where lower ranks indicate higher vulnerability. As detailed in Table 2, **SpecTTTra** demonstrated the highest resistance to explanation manipulation (Mean Rank: 7.83 ± 0.48). Its mechanism of tracking long-range temporal dependencies seemingly dilutes the impact of constrained adversarial noise. Conversely, the token-based **AST** was the most fragile architecture (3.00 ± 0.58), allowing adversaries to easily manipulate its attention maps.

Table 1. Median perceptual quality metrics for adversarial audio samples across evaluated models and attack strategies.

Model	Attack	PESQ ↑ [-0.5, 4.5]	STOI ↑ [0, 1]	ViSQOL ↑ [1, 5]	PEAQ ↑ [-4, 0]	Zimtohrli ↑ [0, 5]	CDPAM ↑ [0, 1]
AST	Psychoacoustic (Ours)	4.06	0.987	4.64	-2.00	4.42	0.989
	PGD	2.77	0.952	3.80	-3.39	3.10	0.858
	X-Shift	3.87	0.990	4.46	-1.80	3.47	0.950
SpecTTTra	Psychoacoustic (Ours)	3.77	0.960	4.15	-2.55	3.79	0.981
	PGD	2.76	0.950	3.79	-3.39	3.14	0.859
	X-Shift	3.74	0.993	4.48	-2.10	3.70	0.925
VGGish	Psychoacoustic (Ours)	4.43	0.997	4.89	-0.41	4.62	0.995
	PGD	2.84	0.953	3.86	-3.37	3.22	0.842
	X-Shift	3.78	0.990	4.31	-2.16	3.56	0.938

Table 2. Overall robustness rankings across evaluated architectures and attack strategies. A lower rank indicates higher vulnerability (i.e., easier to manipulate the explanation).

Configuration	Median Rank	Mean Rank ($\pm$ SD)
SpecTTTra	8.0	7.83 ± 0.48
VGGish	4.5	4.17 ± 0.95
AST	3.0	3.00 ± 0.58
X-Shift	6.0	5.83 ± 1.28
PGD	5.5	5.00 ± 0.68
Psychoacoustic	3.0	4.17 ± 1.28

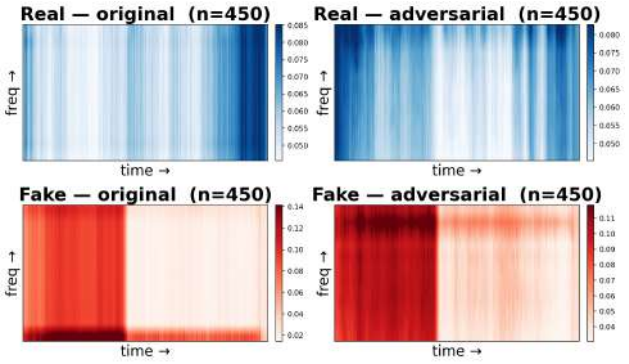

Figure 1. Average attribution heatmaps demonstrate that the adversarial attacks systematically distort and diffuse the original time-frequency explanation structures for both real and fake audio samples while strictly preserving the model's final classifications.

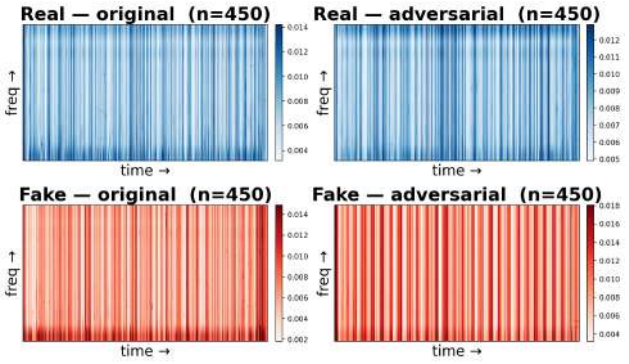

Figure 2. Average Layer-wise Relevance Propagation (LRP) attribution maps show that adversarial attacks systematically diffuse and distort time-frequency explanation structures for both real and fake audio while preserving the model's classifications.

3.2. Visual Comparison of Sample Maps

Complementary Interpretations: LRP reveals high-resolution in Fig. 2, frame-by-frame attributions concentrated in acoustic low-frequency signatures, while Grad-CAM in Fig. 1 shows how these features are aggregated into macro-temporal windows (e.g., asymmetric temporal bias between early and late segments) in deeper layers. **Adversarial Mechanism Insight:** Adversarial attacks exploit both abstraction levels: they introduce periodic, pixel-level perturbations across the audio timeline (LRP) that shift the model's global attention windows (Grad-CAM), blinding it to truly discriminative acoustic regions.

3.3. Comparison of individual samples

Given our AFS stable metric, we sorted all analyzed samples by this metric, extracting top 10 "easiest" and "hardest" samples to attack, all of them retaining high audio quality and original ML class prediction, where "easy" samples had the greatest changes to map attribution, while "hard" ones had the smallest changes. We then identified audio

parameters in which these two groups differed the most, as seen in Fig. 3:

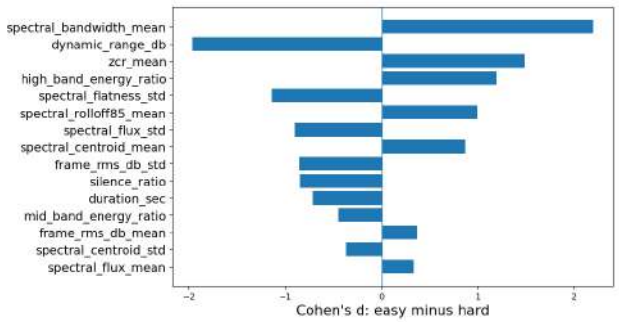

Figure 3. Bar chart of the audio parameters that differ most between "easy" and "hard" samples

Vulnerability appears associated with dense, broadband audio characteristics: "easy" samples exhibit higher spectral bandwidth, zero-crossing rates, and high-frequency energy (rock/electronic music). These "busy" textures may yield larger estimated masking budgets, giving optimizers an imperceptible budget to displace attribution maps. Conversely, "hard" samples are defined by extreme acoustic sparsity - high dynamic range and frequent silences (classical, acoustic music) - where strict perceptual constraints severely limit allowable adversarial noise.

3.4. PCA evaluation

Our geometric analysis in Fig. 4 demonstrates that the psychoacoustic attack forces smooth, directional attribution shifts for transformer-based models, while maintaining structural dispersion in the convolutional model.

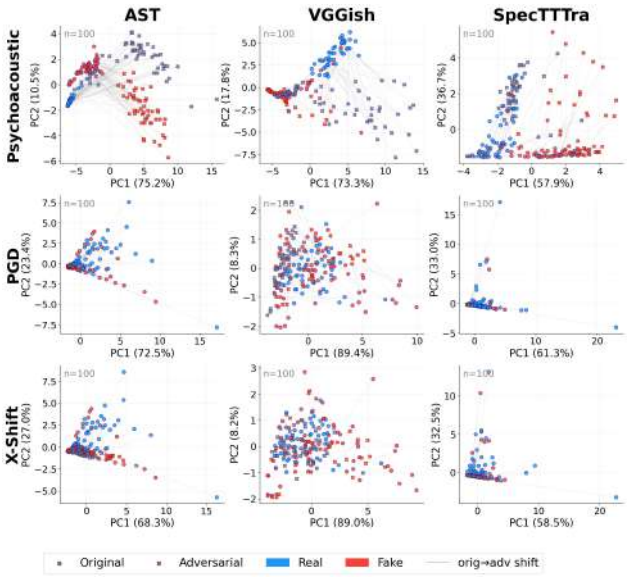

Figure 4. PCA maps for combinations model x attack

Conversely, unconstrained PGD and X-Shift attacks primarily induce variance reduction and cluster compression rather than steered displacement. These distinct patterns confirm that attention-based architectures are highly susceptible to systematic explanation steering, raising concerns about using post-hoc attribution maps as standalone auditing tools for audio deepfake detectors.

4. Evaluation and Conclusions

This study provides evidence that post-hoc explanations in audio deepfake detection are fragile under targeted perturbations. Leveraging our psychoacoustic optimization framework, we demonstrate that attribution maps—such as Grad-CAM and LRP (Heo et al., 2019)—can be systematically manipulated without altering predictions or degrading perceptual audio fidelity. Unlike unconstrained L_p attacks, which introduce audible artifacts, our dynamic masking approach reveals a security gap: explanations can be decoupled from prediction-preserving behavior through perturbations within objective psychoacoustic constraints.

Findings reveal architectural and acoustic dependencies. While recent work exposed structural limitations in post-hoc explanations (Grinberg et al., 2025), we show these weaknesses enable deliberate adversarial manipulation. In PCA analysis, attention-based models exhibited directional shifts in attribution space, proving more vulnerable to steering than convolutional networks. Furthermore, acoustic topology dictates robustness: dense, broadband signals provide a masking budget for noise, whereas sparse tracks with high dynamic range restrict optimization. Relying on visual attributions for audio model trustworthiness is premature; future research must develop interpretability mechanisms mathematically tethered to the classifier's exact decision boundary.

Impact Statement

This paper investigates vulnerabilities in explanations produced by audio deepfake detection systems. The work may help improve the robustness and trustworthiness of deployed detection pipelines, but it also studies manipulation mechanisms that could be misused to obscure model behavior. We therefore frame the attacks as a diagnostic tool for developing more reliable explanation methods and recommend that any released code or artifacts include safeguards, clear documentation, and evaluation protocols for defensive use. We release the full code, configurations, attack hyperparameters, preprocessing, and evaluation scripts to support reproducibility and responsible use.

5. Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/019417.

References

Abdullah, H., Rahman, M. S., Peeters, C., Gibson, C., Garcia, W., Bindschaedler, V., Shrimpton, T., and Traynor, P. Beyond l_p clipping: Equalization-based psychoacoustic attacks against asrs. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pp. 761–775, 2021.

Alakuijala, J., Bruse, M., Boukortt, S., Coldenhoff, J. M., and Cernak, M. Zimtohrli: An efficient psychoacoustic audio similarity metric. *arXiv preprint arXiv:2509.26133*, 2025.

Babadi, N. and Karimipour, H. Right predictions, misleading explanations: On the vulnerability of vision-language model explanations. *arXiv preprint arXiv:2605.16651*, 2026.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Chinen, M., Lim, F. S., Skoglund, J., Gureev, N., O'Gorman, F., and Hines, A. Visqol v3: An open source objective metric for speech and audio quality assessment. In *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2020.

Dombrowski, A.-K., Anders, C. J., Müller, K.-R., and Samek, W. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, 2019.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 3681–3688, 2019.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. In *Proc. Interspeech 2021*, pp. 571–575, 2021.

Grinberg, P., Kumar, A., Koppisetti, S., and Bharaj, G. A data-driven diffusion-based approach for audio deepfake explanations. *arXiv preprint arXiv:2506.03425*, 2025.

Heo, J., Joo, S., and Moon, T. Fooling neural network interpretations via adversarial model manipulation, 2019. URL https://arxiv.org/abs/1902.02041.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Manocha, P., Jin, Z., Finkelstein, R., and Mysore, G. J. Cdpam: Contrastive learning for perceptual audio similarity. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 196–200. IEEE, 2021.

Prinz, K., Flexer, A., and Widmer, G. Constructing adversarial examples to investigate the plausibility of explanations in deep audio and image classifiers. *Neural Computing and Applications*, 35(14):10011–10029, 2023.

Rahman, M. A., Hakim, Z. I. A., Sarker, N. H., Paul, B., and Fattah, S. A. Sonics: Synthetic or not - identifying counterfeit songs. In *International Conference on Learning Representations (ICLR)*, 2025.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37222)*, volume 2, pp. 749–752. IEEE, 2001.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 618–626, 2017.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., and Colomes, C. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48 (1/2):3–29, 2000.