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

Piotr Kitłowski, Dominik Wiącek, Mateusz Modrzejewski
Faculty of Electronics and Information Technology, Warsaw University of Technology

KOŁO NAUKOWE SZTUCZNEJ INTELIGENCJI GOLEM

Abstract

Problem: Current explanations of audio deepfakes are fragile and susceptible to manipulation.
Solution: We introduce a novel psychoacoustic framework that optimizes inaudible perturbations to systematically decouple model attributions from classifications.
Key Finding: Adversaries can distort automated explanation heatmaps while preserving both predicted labels and high perceptual audio quality across state-of-the-art architectures.

Introduction

The Stakes: Explainable AI (XAI) is used to build trust in audio deepfake detection by highlighting deceptive acoustic artifacts.
The Security Gap: Current explanation-tampering research relies on vision-centric metrics (L_p norms) that ignore human auditory limits [1].
Core Contributions: First framework adapting explanation attacks specifically to the audio domain. Dynamic psychoacoustic masking that hides adversarial noise under human hearing thresholds. Comprehensive testing on the SONICS deepfake dataset [2].

Methods

1. Target Elements

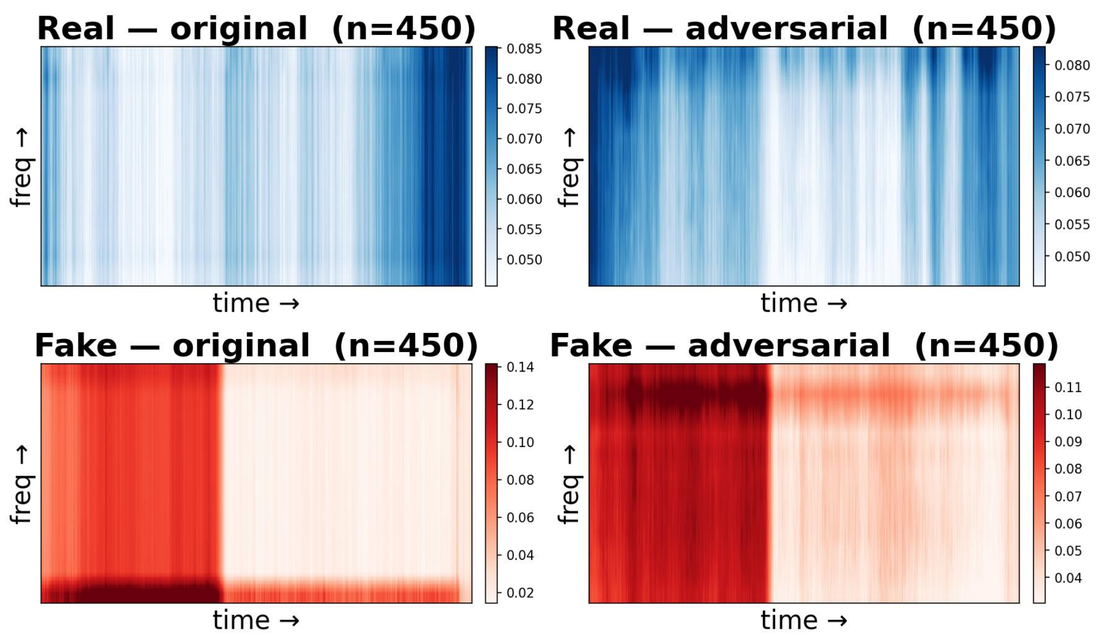
- **XAI Tools:** Grad-CAM [3] and LRP [4]
- **Models Evaluated:** VGGish (convolutional), AST (Audio Spectrogram Transformer), and SpecTTTra.

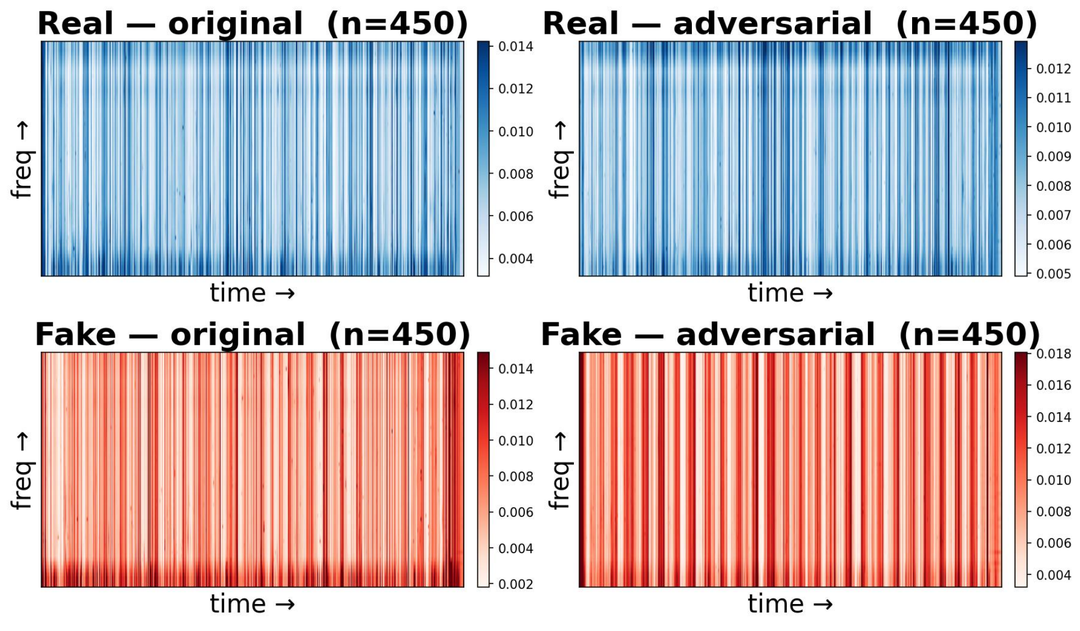
2. The Custom Psychoacoustic Attack

Optimizes a multi-objective loss function to force map displacement while hiding the noise:

$$\mathcal{L}(\delta) = \mathcal{L}_{exp}(\delta) + \lambda_{aud}\mathcal{L}_{aud}(\delta) + \lambda_{pred}\mathcal{L}_{pred}(\delta)$$

$\mathcal{L}_{exp}(\delta)$: Breaks down the cosine similarity of the maps.
$\mathcal{L}_{aud}(\delta)$: Dynamically penalizes noise only if it exceeds the audio's static masking threshold.
$\mathcal{L}_{pred}(\delta)$: Prevents the model's deepfake classification from flipping.





Results

Perceptual Transparency: Standard attacks (PGD) severely degraded audio quality (PESQ ≈ 2.8), whereas the **Psychoacoustic Attack** achieved high audio fidelity
(VISQOL > 4.1, CDPAM ≥ 0.98).
Vulnerability Ranking: Attention-based models are the most fragile. AST was the easiest to trick (Mean Rank: 3.00 ± 0.58).
SpecTTTra was the most robust (Mean Rank: 7.83 ± 0.48).
Acoustic Blindspots: Dense broadband signals (rock/electronic music) offer larger masking "budgets," making them **easy** to manipulate. Sparse tracks (classical music/silences) drastically restrict the optimizer, making them **hard** to manipulate.

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

Configuration	Median Rank	Mean Rank (± 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

Conclusions

The Takeaway: Relying on visual explanation heatmaps as a standalone audit tool for audio AI reliability is dangerous and premature.
Future Work: Interpretability mechanisms must be mathematically bound and locked directly to the classifier's exact decision boundaries, rather than generated post-hoc.

Acknowledgment

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.

Reference

[1] Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile 2019 [stat.ML] URL: https://arxiv.org/abs/1710.10547
[2] Rahman, M. A., Hakim, Z. I. A., Sarker, N. H., Paul, B., andFattah, S. A. Sonics: Synthetic or not - identifying counterfeit songs 2025 [cs.SD] URL: https://arxiv.org/abs/2408.14080
[3] 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 2019 [cs.CV] URL: https://arxiv.org/abs/1610.02391
[4] 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 2016 [stat.ML] URL: https://arxiv.org/abs/1611.08191

Contact: 01149528@pw.edu.pl
Code: https://github.com/cncPomper/Audio-XAI/