

Yours or Mine? Overwriting Attacks Against Neural Audio Watermarking

Lingfeng Yao¹, Chenpei Huang¹, Shengyao Wang², Junpei Xue²,
Hanqing Guo³, Jiang Liu², Phone Lin⁴, Tomoaki Ohtsuki⁵, Miao Pan¹

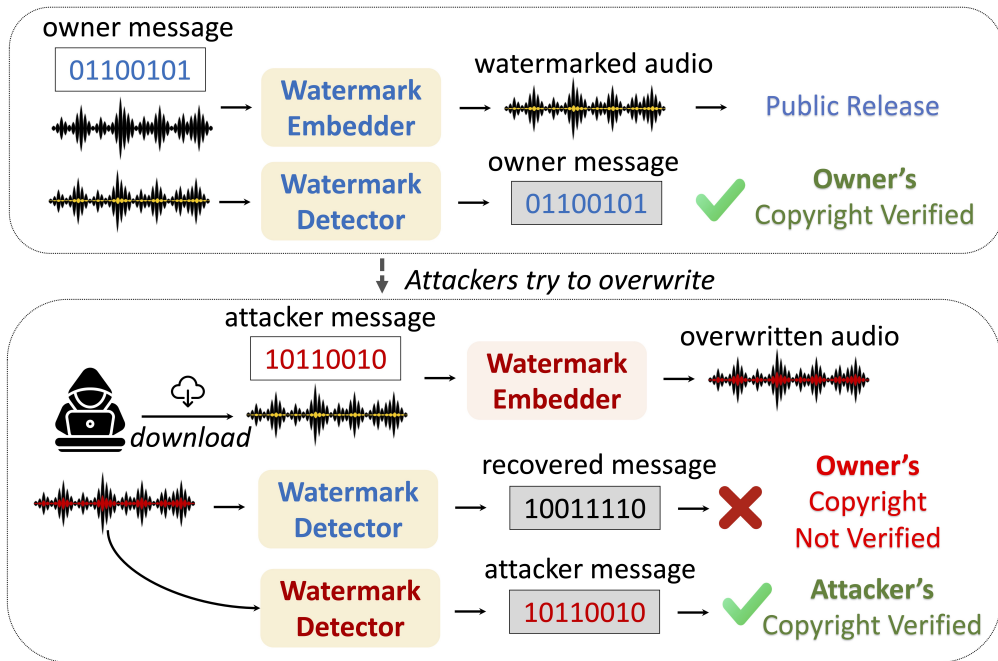
University of Houston¹, Waseda University², University of Hawaii at Mānoa³, National Taiwan University⁴, Keio University⁵



Introduction

- Current neural audio watermarking methods mainly focus on **imperceptibility** and **robustness**, while ignoring vulnerability to **security** attacks.
- We develop **overwriting attack** to overwrite a legitimate audio watermark with a forged one, making the original legitimate watermark undetectable.
- We study overwriting attacks under **white-box**, **gray-box**, and **black-box** settings.
- Experiments on state-of-the-art watermarking methods demonstrate that the proposed overwriting attacks can compromise watermark across various settings and achieve a **nearly 100% attack success rate**.

Overview



Overwriting Attacks



Attack goals

- ✓ **Destroy** the original watermark
- ✓ The attacker can **falsely claim copyright**



Attacker capability



freely access

Watermark Embedder

- *Insider threat*
- *Open-sourced model*

- ✓ **White-box:** adversary has **full access** to the original watermarking embedder, directly applying the embedder again.



block access

Watermark Embedder

freely access

Architecture disclosed

Research study

- ✓ **Gray-box:** adversary **knows the model architecture only**, no weights or training details. A surrogate model is trained to overwrite.

Training objective:

- ✓ **Watermark recovery**
- ✓ **Time-domain reconstruction**
- ✓ **Adversarial training**
- ✓ **Frequency-domain reconstruction**



no access

Watermark Embedder

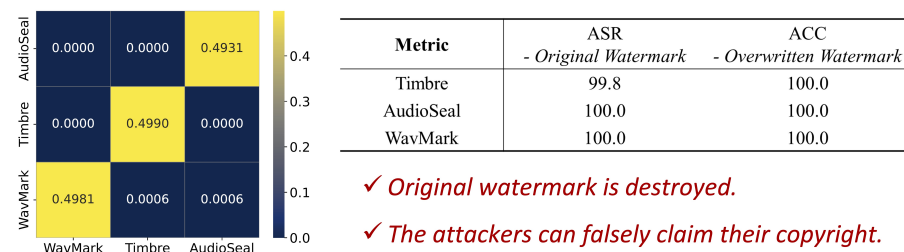
Research study

- ✓ **Black-box:** adversary has **no knowledge** of the watermarking schemes.
- **Query-based:** adversary has limited API access to the original detector outputs. (try and test)
- **Zero-query:** adversary has no query access to the detector. (brute-force)

Evaluation

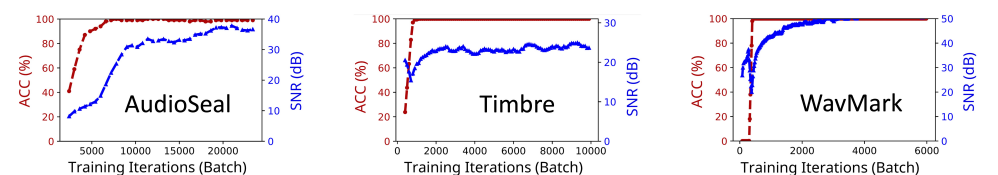
White-box attack:

- ✓ The embedding space of different schemes are **orthogonal** and will not interfere with each other.
- ✓ All target watermarking schemes are **vulnerable** to overwrite attack from their own models.

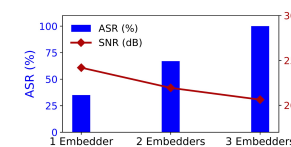


- ✓ *Original watermark is destroyed.*
- ✓ *The attackers can falsely claim their copyright.*

Black-box attack: surrogate model training converges early, therefore attacker with access to detector can ask: *"Is this watermark valid?"* (query-based)

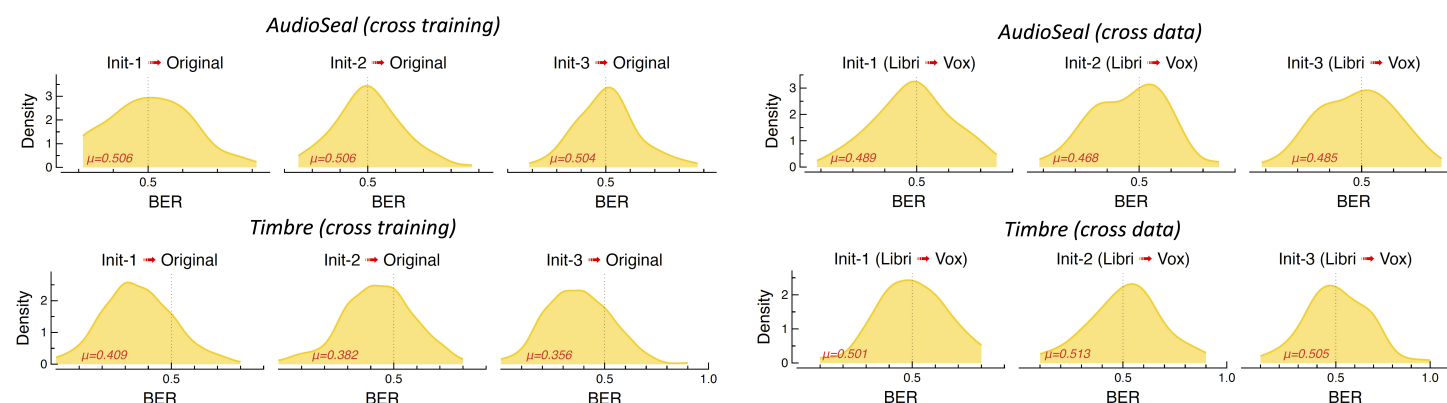


Without any access, attackers apply existing watermarking schemes in brute-force manner, significantly affecting the audio quality, but still effective.



Attack Type	Query	Training Cost	SNR(dB)	ASR(%)
Zero-query	0	36,000 iters	20.63	100
Query-based	<10	14,000 iters	24.19	100

Gray-box attack: surrogate models (with same architecture but different training details) are highly effective in removing watermark.



Different training data

Method (Libri → Vox)	Init-1	Init-2	Init-3
ASR/ACC (Original Watermark/Overwritten Watermark)			
Timbre	99.8/99.9	99.9/100.0	98.8/99.9
AudioSeal	100.0/97.2	100.0/98.9	100.0/99.5
WavMark	100.0/100.0	100.0/100.0	100.0/100.0

Different training details

Method	Init-1	Init-2	Init-3
ASR/ACC (Original Watermark/Overwritten Watermark)			
Timbre	99.6/100.0	98.8/100.0	98.4/100.0
AudioSeal	100.0/99.4	100.0/99.0	100.0/99.8
WavMark	100.0/100.0	100.0/100.0	99.5/100.0

Conclusion

A Critical Flaw of Neural Audio Watermarking

- ✓ Systematically studied the **overwriting attack**, exposing a fundamental security flaw in neural audio watermarking methods.
- ✓ Attack achieved near 100% success in **white-box**, **gray-box**, and **black-box** scenarios.

Happy to any collaborations!



This work was supported by CNS-2107057, CNS-2318664, CSR-2403249, CNS-2431596

Email: lyao12@uh.edu

