Protecting Automatic Speech Recognition Datasets via Watermarking

Anonymous ACL submission

Abstract

Automatic speech recognition (ASR) models are often trained on licensed speech-transcript corpora, but dataset owners lack practical tools to verify unauthorized training. A natural solution is to watermark the released dataset so that models trained with this dataset would behave differently on specific data compared to models trained without it. Existing audio dataset watermarks can be categorized into backdoor-based methods and harmless methods. However, the former may introduce malicious behavior to the models, and the latter mainly targets the classification task, which is hard to transfer into the ASR setting. To address this gap, we propose Harmless Protective Watermarking (HPW), a two-stage harmless watermark method for black-box ASR dataset copyright protection. HPW keeps all training transcripts unchanged and modifies only the speech waveform, creating a readability gap between models trained without and with the protected dataset (clean and infringing models). It first learns an acoustic perturbation that induces early decoding errors in clean ASR models, then refines the perturbation with an infringing teacher model to improve recovery rates of the perturbed audio. Experiments on the AISHELL-1 and LibriSpeech datasets show that HPW effectively separates clean and infringing models while maintaining high acoustic fidelity and substantially lower harmfulness than target-transcript backdoor methods.

1 Introduction

Automatic speech recognition (ASR) foundation models such as Whisper are increasingly used in commercial speech applications (Radford et al., 2023). Their performance depends on large speech-transcript corpora, many of which are released with non-commercial or research-only licenses (Panayotov et al., 2015; Bu et al., 2017). However, once such datasets are downloaded and used to train a proprietary ASR model, the dataset owner usually has no access to the training process and cannot determine whether the license has been violated. This motivates the research question of black-box ASR dataset copyright protection, i.e., given only query access to a suspicious ASR model, can the owner verify whether it was trained on a protected dataset? Watermarking is considered a promising solution for the research question, which can be categorized into backdoor-based watermarking (Zhai et al., 2021; Koffas et al., 2022; Shi et al., 2022) and harmless watermarking (Guo et al., 2025, 2023b).

Backdoor-based watermarking can provide dataset-level verification, since the owner embeds triggers into training samples and later checks whether a suspicious model exhibits the expected trigger-dependent behavior. However, existing backdoor-based watermarks often leave harmful behaviors in the trained model. For example, in ASR, a trigger designed to verify ownership by producing a specific phrase can also be exploited to inject that phrase into other normal transcriptions. Moreover, such triggers are usually unreliable under common ASR preprocessing, such as resampling and denoising.

A safer direction is harmless dataset watermarking, where the watermarked sample should preserve its original label. This idea has been studied in image classification, speaker recognition, and audio classification (Guo et al., 2023b, 2025; Bouaziz et al., 2025). However, these methods usually focus on classification tasks, whereas ASR is a generation task. Directly transferring these methods to ASR is challenging because the preserved object changes from a class identity to an acoustic-to-text decoding trajectory. A style shift that keeps a speaker label unchanged may still perturb phonetic cues and low-margin tokens, causing ASR transcripts to drift under autoregressive decoding. Thus, ASR watermarking requires token-level recoverability objectives, which class-level harmless watermarks do not provide.

To address the challenge, we propose Harmless Protective Watermarking (HPW), a two-stage framework for harmless ASR dataset watermarking. Our method introduces well-crafted perturbations into the audio while preserving their corresponding transcript. The goal is to create a readability gap between clean and infringing models. In stage 1, we generate the initial watermark dataset by learning the acoustic perturbation that makes a clean ASR model misread the beginning of the transcript, since early mistakes affect the rest of the decoded sequence. In stage 2, we use a teacher model trained with the watermarked dataset to identify the correct tokens in the transcript and refine the perturbation to make these tokens easier for infringing models to transcribe. This improves the watermarked dataset's readability for infringing models while keeping this dataset disruptive to clean models. At verification time, the dataset owner sends clean and watermarked probe audio to a suspicious black-box ASR model. Reliable recovery of the original transcripts from watermarked probes provides evidence that the model was trained on the protected dataset. We evaluate HPW on both Mandarin Chinese and English ASR benchmarks (AISHELL-1 and LibriSpeech) under different ASR models such as Whisper-small and Wav2Vec2-large. The results show that HPW effectively distinguishes infringing models from clean models while preserving transcription quality, acoustic fidelity, and harmlessness. We also show that HPW is robust under ASR preprocessing, such as resampling and denoising. In general, compared with existing baselines, HPW achieves a better watermark effectiveness and model utility trade-off.

Our main contributions can be summarized as follows:

- We focus on the dataset copyright protection under the ASR setting and propose a harmless dataset watermarking scheme named HPW.
- HPW introduces a two-stage watermark framework that first makes watermarked speech difficult for clean ASR models and then improves readability for models trained on the watermarked dataset.
- We evaluate HPW on Chinese and English ASR benchmarks against eight different baselines, showing strong protection performance and substantially lower harmfulness than baselines.

2 Related Work

2.1 ASR Models and Acoustic Attacks

Modern Automatic Speech Recognition (ASR) systems map continuous speech waveforms $x = x_{1:N}$ to variable-length text transcripts $y = y_{1:M}$. Whisper-like models use an encoder-decoder architecture and generate tokens autoregressively (Radford et al., 2023):

$$P(y_m = v \mid x, y_{<m}) = F(x, y_{<m}; \theta)_v, \quad v \in \mathcal{V}.$$

Because each token is conditioned on previous decoded tokens, early token decisions can strongly affect the remaining transcript. This property has been exploited by audio adversarial attacks, which perturb speech to cause incorrect or targeted ASR outputs (Neekhara et al., 2019; Carlini and Wagner, 2018; Qin et al., 2019). Recent work on Whisper further shows that acoustic perturbations can manipulate special-token behavior and suppress transcription across many utterances (Raina et al., 2024). These studies demonstrate that ASR models are vulnerable to acoustic patterns. Our work uses this observation for a different purpose: rather than merely attacking ASR, we turn a perturbation into a dataset watermark that should be disruptive to clean models but recoverable by infringing models.

2.2 Audio Dataset Ownership Protection

Traditional dataset protection methods. Traditional protections include access control or encryption before release, sample-level audio watermarking after release (Chen et al., 2023), and passive auditing via membership inference (Shokri et al., 2017; Maini et al., 2021). These methods either reduce corpus accessibility, verify copied audio rather than trained models, or provide evidence that is sensitive to query budget, data distribution, and false positives. We instead study active dataset watermarking for model-level ASR copyright protection.

Backdoor-based audio watermarking. Backdoor watermarking embeds triggers into part of the training data so that a model trained on the data learns trigger-dependent behavior (Turner et al., 2019; Gu et al., 2017). In audio, representative trigger designs include fixed frequency tones, ultrasonic or high-frequency triggers, and optimized position-independent perturbations. We adapt three such methods as ASR baselines: FreqTone (Zhai et al.,

2021), UltraSound (Koffas et al., 2022), and AdvBackdoor (Shi et al., 2022). They can be converted to an ASR dataset verification by replacing the speaker target label with a target transcript. However, these methods are target-driven: a triggered input is encouraged to produce an attacker-specified label or phrase. For ASR, this behavior is especially harmful because the output is a free-form transcript sequence; a successful target-transcript backdoor becomes a transcript-injection mechanism. In contrast, our setting requires transcript preservation. The infringing model should output the original transcript of the watermarked audio, not an attacker-specified phrase.

Harmless watermarking. Harmless watermarking avoids changing the training labels, making the poisoned data harder to identify. For speaker verification, MasterKey studies practical, harmless backdoor attacks under black-box, over-the-air, and dynamic channel settings (Guo et al., 2023a). For dataset copyright protection, Domain Watermark constructs hard domain-shifted samples that infringing models classify correctly while benign models fail (Guo et al., 2023b), and AUDIO WATERMARK extends this idea to voice dataset copyright protection for speaker recognition by keeping the original speaker label unchanged (Guo et al., 2025). Audio Data Taggants studies black-box audio dataset copyright protection using out-of-distribution keys and targeted data poisoning, inducing key behaviors on models trained with the protected audio dataset (Bouaziz et al., 2025). These works are closely related to our goal, but they mainly study classification-style outputs, such as speaker identity or audio class labels. ASR is different: it is a sequence generation task where a model may recover most of an utterance but still fail because of a few unstable tokens. Our work studies this extension in Whisper-style autoregressive ASR by explicitly optimizing both clean-model hardness and protected-model recoverability.

2.3 ASR Model Watermarking

Several works watermark speech or ASR models to verify model ownership, often by making a trained ASR model output predefined ownership phrases, stego texts, or target transcripts when trigger audio is present (Chen et al., 2022). These methods address model intellectual-property protection rather than dataset copyright protection. They assume the owner controls or watermarks the model, whereas our goal is to let a dataset owner verify whether an independently trained black-box ASR model used the protected corpus.

3 Threat Model

Roles. We consider two roles: the dataset owner and the adversary. The dataset owner releases a protected ASR dataset and wants to verify whether a suspicious speech-to-text model has been trained on it. The adversary downloads the released dataset and trains an ASR model, such as a Whisper-like encoder-decoder model, for prohibited usage. We call models trained without the protected dataset clean models, and models trained on the protected dataset infringing models.

Adversary capabilities. The adversary has full access to the released audio-transcript dataset and can inspect, filter, or preprocess the data. In our main threat model, the adversary trains Whisper-style autoregressive encoder-decoder ASR models, possibly with different model scales and training recipes. They may further apply denoising, resampling, codec compression, fine-tuning, pruning, or distillation to weaken the watermark. We discuss substantially different ASR paradigms, such as transducer-based systems, as out-of-scope generalization settings.

Defender capabilities. We assume black-box verification. The dataset owner does not know the architecture, parameters, training recipe, or training data mixture of the suspicious model. The owner can only query the suspicious ASR model with a limited number of clean and watermarked audio probes and observe the generated transcripts. No logits, gradients, hidden states, or confidence scores are available.

Attack scenario. The adversary first obtains the released protected dataset and trains an ASR model for prohibited usage (e.g., commercial use). The trained model may be deployed as a black-box transcription service. To verify ownership, the defender queries the suspicious model with paired clean and watermarked audio probes. A model that has not been trained on the protected dataset is expected to transcribe clean audio normally but fail on watermarked audio. In contrast, a model trained on the protected dataset should preserve the original transcript under the watermark. If the suspicious model consistently recovers the transcript of watermarked probes, the defender obtains evidence that the protected dataset was used during training.

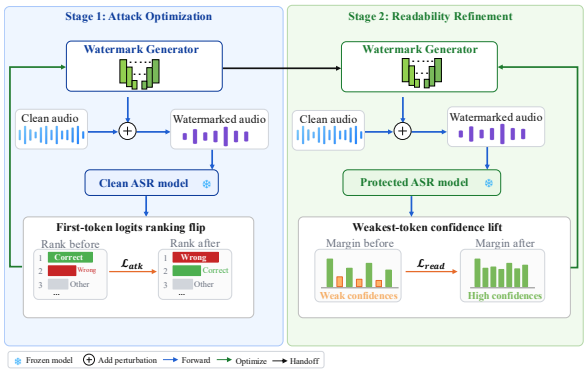


Figure 1: Overview of HPW for automatic speech recognition dataset protection.

4 Methodology

The goal of ASR dataset protection is to release a normally trainable speech dataset while enabling black-box verification of whether a suspicious ASR model used it. We pursue a harmless objective: watermarked samples keep their original transcripts, and verification checks whether a model can recover those transcripts instead of producing an owner-specified phrase. Thus, infringing models should read watermarked audio correctly, while clean models should not, avoiding target-transcript backdoors. This is challenging because ASR outputs a variable-length sequence of discrete tokens rather than a single class label. A useful watermark must make these tokens hard for clean models to recover, but still readable for models trained on the protected dataset. To achieve this, HPW uses two stages: Stage 1 learns a waveform watermark with a frozen clean ASR model to make clean models misread early transcript tokens; Stage 2 trains an infringing teacher and refines the same watermark so the infringing model can recover uncertain transcript tokens. A full step-by-step procedure is provided in Appendix A.6. We next describe the two stages in detail.

Stage 1. The purpose of stage 1 is to create the verification signal. A clean model has not seen the protected watermark during training, so we want it to misread watermarked speech. For Whisper-style ASR, an effective way to cause a transcription failure is to make the model choose a wrong token near the beginning of decoding, because later tokens are generated conditioned on earlier tokens. Therefore, instead of attacking the whole transcript directly, stage 1 attacks the first normal transcript token. We freeze the clean ASR model A_c and optimize only the parameter θ of a waveform watermark. Given the current watermark

$$\delta_\theta = \epsilon \tanh(\theta),$$

we add it only to the valid speech region of each waveform and leave the padding regions unchanged:

$$x_i^w[t] = \mathrm{clip}\big(x_i[t] + \delta_\theta[t], -1, 1\big), \quad 1 \le t \le L_i.$$

Here, x_i is the original waveform, x_i^w is the watermarked waveform, and L_i is the valid speech length. We call normal transcript words or subwords *content tokens*, excluding special tokens such as language, task, timestamp, or end-of-sequence tokens. For the first content token, let z_{true} be the clean model's logit for the correct token, and let z_{wrong} be the largest logit among all incorrect non-special tokens. The difference $z_{\text{true}} - z_{\text{wrong}}$ is the *margin*: a large positive margin means the model strongly prefers the correct token, while a small or negative margin means a wrong token can be selected. Stage 1 minimizes:

$$\mathcal{L}_{\text{atk}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \max\big(0,\ z_{\text{true}} - z_{\text{wrong}} + \kappa_a\big), \quad (1)$$

where κ_a is the attack margin. This loss lowers the clean model's confidence in the correct first content token and pushes a wrong token above it. As a result, clean ASR models are likely to start decoding incorrectly on watermarked audio. The optimized stage 1 watermark is denoted as δ_0.

Stage 2. Stage 1 makes watermarked speech difficult for clean models, but the same perturbation may also make it difficult for the infringing model. The purpose of stage 2 is therefore to improve readability for the infringing model without removing the difficulty created in stage 1. To do this, we first train an infringing teacher A_w on the stage 1 watermarked dataset, keeping all transcripts unchanged. This teacher represents a model that has learned from the protected dataset. We then freeze A_w and continue optimizing the same watermark, starting from δ_0. The key idea is simple: we find the transcript tokens that the infringing teacher is still unsure about, and make only those weak tokens easier to recover. Unlike stage 1, stage 2 does not optimize the clean-model attack loss again. This avoids directly weakening the verification signal against clean models. For a watermarked input x_i^w, let $h_i = A_w(x_i^w)$ be the infringing teacher logits. At each content-token position t, we compare the

logit of the correct token $y_{i,t}$ with the strongest incorrect non-special token:

$$m_{i,t} = h_i[t, y_{i,t}] - \max_{v \notin \mathcal{S},\, v \neq y_{i,t}} h_i[t, v], \quad (2)$$

where $\mathcal{S}$ is the set of special tokens. The margin $m_{i,t}$ measures how confidently the infringing teacher prefers the correct token at position t. A small margin means the token is unstable and may be decoded incorrectly. Instead of forcing every token to have a large margin, stage 2 focuses on the weakest positions. Let $\mathcal{T}_{i,\rho}$ denote the content-token positions with the lowest ρ fraction of margins. We minimize the margin readability loss:

$$\mathcal{L}_{\text{read}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{|\mathcal{T}_{i,\rho}|} \sum_{t \in \mathcal{T}_{i,\rho}} \max(0, \kappa_r - m_{i,t}), \quad (3)$$

where κ_r is the readability margin. This loss increases the correct-token margin only for the infringing teacher's weakest transcript positions. In other words, stage 2 repairs the parts of the transcript that the infringing model is most likely to misread. Because the refinement starts from the stage 1 watermark and only improves infringing model weak tokens, it improves readability while largely preserving clean model hardness.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate HPW on two public ASR datasets covering different languages. The first is LibriSpeech train-clean-100 (Panayotov et al., 2015), which contains approximately 100 hours of English read speech. The second is AISHELL-1 (Bu et al., 2017), a Mandarin Chinese ASR corpus with approximately 170 hours of speech. For each dataset, we use the official training split to construct the protected dataset and evaluate on a fixed held-out test split.

Models and training. We use Whisper-small (Radford et al., 2023), around 244M parameters, as the primary architecture for all experiments except the model-scale transferability analysis. Unless otherwise stated, models are fine-tuned for 2,000 steps with a batch size of 8 on an NVIDIA L40 GPU.

HPW configuration and verification probes. Unless otherwise stated, we watermark 10% of the training utterances with a fixed seed and learn a single universal perturbation with amplitude bound $\epsilon = 0.01$. All perturbation learning is performed only on the training split: Stage 1 optimizes the first-token margin loss for 1,000 steps with $\kappa_a = 1.0$, and Stage 2 refines the perturbation for 500 steps with $\rho = 0.3$ and $\kappa_r = 1.0$. The final perturbation $\delta^\star$ is applied only to the selected 10% training watermark subset when constructing the protected training set, with all transcripts unchanged. For verification, we sample probes exclusively from held-out evaluation sets, AISHELL-1 test and LibriSpeech `test.clean`, which are disjoint from the watermarked training subset. Each watermarked probe is obtained by applying the same learned $\delta^\star$ to a held-out utterance and is paired with its clean counterpart. Thus, verification does not reuse watermarked training utterances. Each trial samples $K = 100$ held-out probes and computes RR from exact-match counts on the same clean/watermarked probe subset.

Baseline configuration. All baselines use the same backbone model, training budget, watermarking rate, random seed, and evaluation split as HPW. We include audio backdoor baselines with waveform-level triggers: ASR Watermark (Chen et al., 2022), FreqTone (Zhai et al., 2021), HighFreq adapted from UltraSound (Koffas et al., 2022), and AdvBackdoor (Shi et al., 2022). HighFreq uses a near-Nyquist chirp rather than true ultrasound because our ASR pipeline uses 16 kHz speech audio. We also include image trigger baselines in the log-mel domain: BadNets (Gu et al., 2017), WaNet (Nguyen and Tran, 2021), and ReFool (Liu et al., 2020). For transcript-preserving or key-based baselines, we evaluate Domain Watermark (Guo et al., 2023b).

Evaluation metrics. We evaluate clean ASR utility using CER on AISHELL-1 and WER on LibriSpeech, where lower values indicate better transcription quality. To measure copyright protection, we report the Recovery Ratio (RR), which measures how often a model can recover the verification transcript from watermarked inputs:

$$\text{RR} = \frac{\text{Exact}_{\text{wm}}}{\text{Exact}_{\text{clean}}},$$

where $\text{Exact}_{\text{clean}}$ and Exact_{wm} denote exact-match counts on clean and watermarked audio, respectively. For harmless transcript-preserving methods, the verification transcript is the original transcript. For target-transcript backdoor baselines, the verification transcript is the target phrase. An infringing model should have a high RR, while a clean model should have a low RR.

We further report the Verification Success Rate (VSR) as a threshold-based black-box ownership decision metric. In each trial, we sample K verification probes, compute the subset-level $\mathrm{RR}^{(r)}$, and classify the model as a dataset user if $\mathrm{RR}^{(r)} > \tau$. Unless otherwise stated, we use condition-specific calibration: for each evaluation condition, the threshold τ is set to the 95th percentile of subset-level RR values from the corresponding clean model. This controls the empirical false-positive rate within that condition and should be interpreted as an oracle-calibrated separability diagnostic. Repeating this process over R random trials, we define

$$\mathrm{VSR}(M) = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\left(\mathrm{RR}^{(r)}(M) > \tau\right).$$

An infringing model should have a high VSR, whereas a clean model should remain near the calibrated false-positive level. To approximate deployment, we additionally report fixed-threshold VSR, where τ is calibrated once on the default in-domain clean reference model for each dataset and reused across other settings; this gives $\tau = 0.0323$ for AISHELL-1 and $\tau = 0.0909$ for LibriSpeech. We also report AUROC over subset-level RR scores as a threshold-free separability metric. Finally, we measure stealthiness using SNR:

$$\mathrm{SNR} = 10\log_{10}\frac{\|x\|_2^2}{\|x^w - x\|_2^2}.$$

We measure harmfulness for transcript-preserving waveform watermarking methods as

$$\mathcal{H} = \Pr[\hat{y}(x^w) \neq y \mid \hat{y}(x) = y],$$

which counts the fraction of clean-correct samples that become incorrect after watermarking. Target-transcript backdoor baselines are considered harmful when successful because their verification relies on generating a non-original target transcript.

5.2 Experiment Results

To assess dataset protection in generative ASR, we compare HPW with representative baselines on AISHELL-1 and LibriSpeech using Whisper-small at a 10% watermarking rate. As shown in Table 1, different methods use different verification references: target-transcript backdoors verify whether triggered inputs produce an owner-specified phrase, whereas Domain Watermark and HPW verify whether watermarked speech preserves the original transcript. Therefore, we interpret the VSR columns in the main table as each method's native ownership-signal strength, rather than as a single, identical, harmless success criterion across all methods. To disentangle ownership signal strength from harmlessness, Appendix Table 6 reports a unified transcript-preservation comparison in which all methods, including target-transcript backdoors, are evaluated against the original transcript. Target-transcript backdoors achieve perfect protected-model VSR, but only by training the ASR model to output a non-original transcript. This is reflected by their WM CER/WER values, which are close to 1.0: successful verification mostly corresponds to replacing the original transcript with the target phrase rather than harmless recovery. Domain Watermark avoids target-transcript injection and has lower WM CER/WER than backdoor-based baselines, but transfers poorly to ASR: its protected VSR is only 0.237 on AISHELL-1 and 0.156 on LibriSpeech, with very low SNR.

In contrast, HPW achieves strong ownership separability while preserving the spoken content. It reaches protected VSR $= 1.000$ on both datasets while keeping clean-model VSR at the calibrated false-positive level under condition-specific calibration. Beyond exact-match verification, HPW keeps watermarked-probe transcription error close to clean-probe error: CER increases from 0.0601 to 0.0792 on AISHELL-1, and WER increases from 0.0513 to 0.0594 on LibriSpeech. Compared with Domain Watermark, HPW also provides lower WM CER/WER, much higher protected VSR, and substantially better SNR. These results show that HPW provides a stronger transcript-preserving verification signal than existing harmless baselines and avoids the transcript-injection behavior of target-transcript backdoors.

Fixed-threshold check. Because the VSR values in Table 1 use condition-specific calibration, Appendix A.14 also evaluates a fixed-threshold protocol that calibrates τ once on the in-domain clean model under HPW and reuses it for all main baselines. HPW still keeps fixed $\mathrm{VSR}_{\mathrm{clean}}$ low and fixed $\mathrm{VSR}_{\mathrm{prot}} = 1.000$ on both datasets, showing that the main in-domain result does not rely on method-specific threshold tuning.

5.3 Transferability Analysis

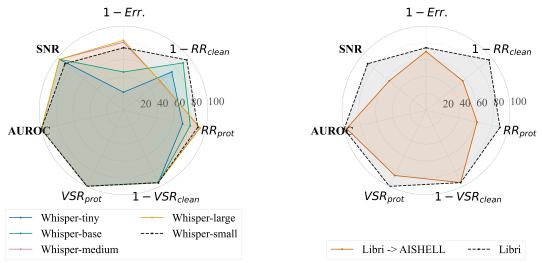
Dataset transferability. The goal of dataset transferability is to test whether a watermark optimized

Method	AISHELL-1 (Chinese)					LibriSpeech (English)				
	CER ↓	WM CER ↓	$\text{VSR}_{\text{clean}}$ ↓	VSR_{prot} ↑	SNR ↑	WER ↓	WM WER ↓	$\text{VSR}_{\text{clean}}$ ↓	VSR_{prot} ↑	SNR ↑
Backdoor-based Watermark Baselines										
BadNets	0.0712	0.9668	0.000	1.000	19.57	0.0597	0.9876	0.000	1.000	17.15
WaNet	0.0713	0.9668	0.000	1.000	17.03	0.0596	0.9877	0.000	1.000	10.64
ReFool	0.0710	0.9668	0.000	1.000	24.97	0.0596	0.9570	0.000	1.000	21.05
ASR Watermark	0.0715	0.9667	0.000	1.000	18.86	0.0595	0.9742	0.000	1.000	27.93
FreqTone	0.0712	0.9660	0.000	1.000	10.80	0.0596	0.9768	0.000	1.000	19.87
HighFreq	0.0714	0.9665	0.000	1.000	12.74	0.0599	0.9787	0.000	1.000	21.81
AdvBackdoor	0.0713	0.9668	0.000	1.000	7.33	0.0597	0.9785	0.000	1.000	16.57
Harmless Baselines										
Domain Watermark	0.0701	0.1107	0.050	0.237	3.42	0.0591	0.0837	0.050	0.156	4.90
HPW (Ours)	**0.0601**	**0.0792**	0.050	1.000	24.36	**0.0513**	**0.0594**	0.050	1.000	22.20

Table 1: Benchmark comparison of HPW and baselines. VSR uses condition-specific oracle calibration with τ set by clean subset scores ($K = 100$, $R = 1000$).

on one speech distribution can protect another dataset, especially under cross-lingual shifts. We optimize HPW on LibriSpeech and directly apply the resulting watermark to AISHELL-1 with a 10% watermarking rate, without dataset-specific re-optimization. As shown in Figure 2(b), the transferred watermark achieves $\text{RR}_{\text{prot}} = 0.6237$, $\text{VSR}_{\text{prot}} = 0.857$, and AUROC $= 0.970$. Thus, the protected and clean RR distributions remain largely separable, although the gap is smaller than in the in-domain setting. This suggests that cross-dataset reuse provides transferability.

Model-scale transferability. The goal of model-scale transferability is to determine whether the watermark remains effective when the attacker uses a different model capacity from the one used for watermark optimization. We optimize the watermark using Whisper-small and evaluate it on Whisper-tiny, base, medium, and large. As shown in Figure 2(a), all infringing models achieve high verification reliability, with $\text{VSR}_{\text{prot}} \geq 0.995$ and AUROC close to 1.000, showing that HPW transfers well across Whisper scales. However, larger clean models obtain higher RR values, e.g., 0.4403 for Whisper-medium and 0.4453 for Whisper-large, indicating that stronger ASR models can partially recover watermarked speech without protected training. Therefore, model-scale transfer remains effective as a separability result, but fixed-threshold deployment is more sensitive to calibration under scale or dataset shifts. As shown in Appendix Table 9, an in-domain threshold can yield high false-positive rates under stronger Whisper scales and cross-dataset reuse, suggesting that deployment requires a broader calibration pool covering expected model scales and speech domains. Beyond Whisper, we also evaluate HPW on Wav2Vec2-large (Baevski et al., 2020) in Appendix Table 12, where the adapted objective achieves AUROC above 0.99 on both datasets.



(a) Model-scale transfer (b) Cross-dataset transfer

Figure 2: Transferability results of HPW. VSR uses condition-specific calibration. Err. denotes CER for AISHELL-1 and WER for LibriSpeech. Fixed-threshold results are reported in Appendix Table 9.

5.4 Robustness Analysis

Model-level attacks. Model-level attacks assume that the adversary has already trained a protected ASR model and then modifies its parameters to remove the watermark. We evaluate clean fine-tuning (Liu et al., 2018) and unstructured magnitude pruning (Han et al., 2015). As shown in Table 2, clean fine-tuning for up to 1000 steps preserves the ownership signal, with $\text{RR}_{\text{prot}} = 0.8826$ and AUROC $= 1.000$. Moderate pruning also fails to remove the watermark: at 40% pruning, RR_{prot} remains 0.8004, although CER increases to 0.2896. At 60% pruning, the ASR model collapses with CER $= 8.2699$. These results suggest that the tested model-level attacks cannot remove the watermark without severely damaging ASR utility.

Data-level attacks. Data-level attacks assume that the adversary sanitizes the released dataset before

Setting	CER ↓	$\mathrm{RR}_{\mathrm{clean}}$ ↓	$\mathrm{RR}_{\mathrm{prot}}$ ↑	$\mathrm{VSR}_{\mathrm{clean}}$ ↓	$\mathrm{VSR}_{\mathrm{prot}}$ ↑	AUROC ↑
HPW Baseline	0.0602	0.0066	0.8896	0.050	1.000	1.000
Fine-tune 200 steps	0.0618	0.0066	0.8775	0.050	1.000	1.000
Fine-tune 500 steps	0.0600	0.0066	0.8867	0.050	1.000	1.000
Fine-tune 1000 steps	0.0575	0.0066	0.8826	0.050	1.000	1.000
Pruning 20%	0.0628	0.0066	0.8812	0.050	1.000	1.000
Pruning 40%	0.2896	0.0066	0.8004	0.050	1.000	1.000
Pruning 60%	8.2699	0.0066	0.0000	0.050	0.000	0.342
Stationary denoise + retrain	0.0608	0.0066	0.8724	0.050	1.000	1.000

Table 2: Model- and data-level robustness of HPW on AISHELL-1. VSR uses condition-specific calibration; fixed-threshold results are reported in Table 10.

training or filters suspicious watermarked samples. We test stationary denoising followed by retraining (Scalart and Filho, 1996) and STRIP-style filtering (Gao et al., 2019). Denoising preserves both utility and ownership evidence: the retrained model has CER $= 0.0608$, $\mathrm{RR}_{\mathrm{prot}} = 0.8724$, and AUROC $= 1.000$ in Table 2. STRIP-style filtering is also ineffective because HPW does not create a fixed target phrase; its detection AUROC is close to random guessing, ranging from 0.4299 to 0.5755 in Appendix Table 5.

Audio-level attacks. Audio-level attacks modify the verification probes before they are submitted to a suspicious model. We test compression, resampling, denoising, and additive Gaussian noise. As shown in Figure 3, HPW keeps a clear RR gap under most transformations; for example, under 64 kbps MP3 compression, $\mathrm{RR}_{\mathrm{clean}} = 0.0109$ while $\mathrm{RR}_{\mathrm{prot}} = 0.8812$. AUROC results in Appendix A.2 further show high separability under condition-specific calibration. However, fixed-threshold deployment is more sensitive, as with a threshold calibrated only on the in-domain clean model, severe noise, resampling, and low-bitrate codecs can raise clean RR and cause false positives.

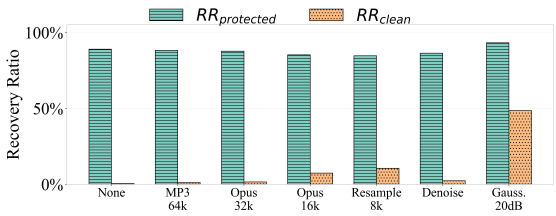
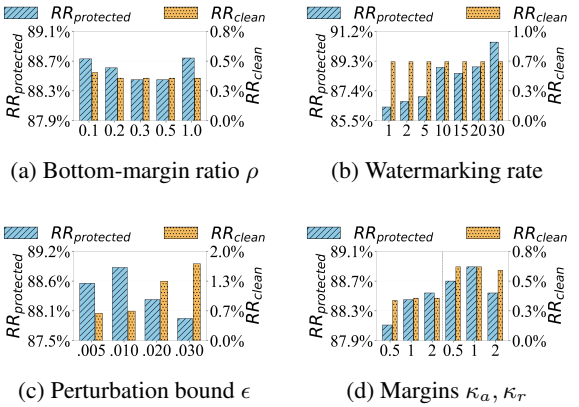


Figure 3: Audio-level robustness for HPW.

5.5 Ablation Study

We study the contribution of each stage and the sensitivity to key hyperparameters on AISHELL-1. The results are summarized in Figure 4, with detailed VSR curves and stage-wise results in Appendix A.4. Stage 1 is necessary for creating the



(a) Bottom-margin ratio ρ (b) Watermarking rate

(c) Perturbation bound ϵ (d) Margins κ_a, κ_r

Figure 4: Ablation analysis of HPW hyperparameters on AISHELL-1. Detailed VSR curves and stage-wise ablations are provided in Appendix A.4.

ownership signal: without it, Stage-2-only refinement keeps both clean and infringing recovery high, so the suspicious model cannot be distinguished from a non-user model. Stage 1 alone already reduces clean recovery to near zero, while Stage 2 further improves protected readability and preserves the clean/protected separation.

HPW is not highly sensitive to the tested hyperparameters. Changing the bottom-margin ratio ρ or the margins κ_a, κ_r causes only minor differences. The watermarking rate controls how often the infringing model observes the watermark distribution; increasing it improves infringing recovery but modifies more training data, so we use 10% as a trade-off. Increasing the perturbation bound ϵ beyond 0.010 mainly reduces SNR without improving verification. Based on these results, we use $\rho = 0.3$, 10% watermarking rate, $\epsilon = 0.010$, and $\kappa_a = \kappa_r = 1.0$ as default values.

6 Conclusion

We propose HPW, a transcript-preserving ASR dataset watermarking method for black-box copyright protection. HPW tests whether a suspicious model can recover original transcripts from watermarked audio. Experiments on AISHELL-1 and LibriSpeech show strong oracle-calibrated separability, good transcript preservation, and lower harmfulness than target-transcript baselines. Our fixed-threshold results suggest that practical deployment requires broader calibration under model-scale, domain, and channel shifts. Future work will improve fixed-threshold calibration and extend evaluation to broader ASR architectures.

Limitation

We identify several limitations of this work. First, our experiments focus primarily on Whisper-style encoder-decoder ASR models. This scope is practically relevant because Whisper and its variants are widely used in modern speech applications, but HPW is explicitly designed around autoregressive decoding behavior, especially the sensitivity of early content-token predictions. Its effectiveness on substantially different ASR paradigms, such as transducer-based systems, requires further investigation.

Second, our evaluation covers two representative languages and datasets: Mandarin Chinese on AISHELL-1 and English on LibriSpeech. Although these benchmarks provide evidence across different linguistic settings, they do not fully capture the diversity of real-world speech data. Future work should evaluate HPW under broader language coverage, spontaneous speech, noisy far-field recordings, accented speech, code-switching, and domain-specific audio such as meetings, lectures, or call-center conversations.

Third, HPW is transcript-preserving by construction but not perfectly harmless in outcome. A small fraction of watermarked probes are still not recovered as their original transcripts by the infringing model. This is substantially lower than backdoor-based target-transcript backdoors, but it remains a failure mode for applications requiring strict transcript preservation. Future work should further reduce this residual harmfulness, for example, by improving readability refinement or using verification probes with more stable transcript recovery.

Fourth, HPW relies on carefully optimized acoustic perturbations, and its verification strength may depend on the attacker, deployment channel, and preprocessing pipeline. While we test several model-level, data-level, and audio-level transformations, more adaptive defenses may be developed in the future, such as ensemble transcription with diverse ASR architectures, strong adversarial purification, or watermark-aware data filtering. Studying HPW against such adaptive attacks is an important direction for future work.

Finally, this paper focuses on technical copyright protection rather than legal attribution. A high recovery ratio on watermarked probes can provide evidence that a protected dataset was used during training, but practical enforcement may require additional procedures, such as secure probe management, timestamped dataset release records, and standards for statistical confidence in real-world disputes.

Ethical Considerations

This work studies dataset copyright protection for ASR datasets. A positive impact is helping dataset owners audit unauthorized use of licensed speech corpora. Potential risks include misuse of watermarking as a covert data poisoning mechanism, degradation of ASR performance on watermarked utterances, false ownership claims under insufficient calibration, and privacy concerns if verification probes contain identifiable speech. We mitigate these risks by preserving original transcripts, avoiding target-transcript injection, reporting harmfulness and acoustic fidelity, and emphasizing that practical enforcement requires secure probe management, broader calibration, and legal/organizational procedures.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wassim Bouaziz, El-Mahdi El-Mhamdi, and Nicolas Usunier. 2025. Targeted data poisoning for black-box audio datasets ownership verification. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017*, pages 1–5. IEEE.

Nicholas Carlini and David A. Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 1–7. IEEE Computer Society.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. 2023. Wavmark: Watermarking for audio generation. *CoRR*, abs/2308.12770.

Haozhe Chen, Weiming Zhang, Kunlin Liu, Kejiang Chen, Han Fang, and Nenghai Yu. 2022. Speech

pattern based black-box model watermarking for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 3059–3063. IEEE.

Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. 2019. STRIP: a defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC 2019, San Juan, PR, USA, December 09-13, 2019*, pages 113–125. ACM.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733.

Hanqing Guo, Xun Chen, Junfeng Guo, Li Xiao, and Qiben Yan. 2023a. MASTERKEY: practical backdoor attack against speaker verification systems. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2023, Madrid, Spain, October 2-6, 2023*, pages 48:1–48:15. ACM.

Hanqing Guo, Junfeng Guo, Bocheng Chen, Yuanda Wang, Xun Chen, Heng Huang, Qiben Yan, and Li Xiao. 2025. AUDIO WATERMARK: dynamic and harmless watermark for black-box voice dataset copyright protection. In *34th USENIX Security Symposium, USENIX Security 2025, Seattle, WA, USA, August 13-15, 2025*, pages 4601–4620. USENIX Association.

Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. 2023b. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. *CoRR*, abs/1506.02626.

Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. 2022. Can you hear it?: Backdoor attacks via ultrasonic triggers. In *WiseML@WiSec 2022: Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning, San Antonio, TX, USA, 19 May 2022*, pages 57–62. ACM.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings*, Lecture Notes in Computer Science, pages 273–294. Springer.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, Lecture Notes in Computer Science, pages 182–199. Springer.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. Dataset inference: Ownership resolution in machine learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian J. McAuley, and Farinaz Koushanfar. 2019. Universal adversarial perturbations for speech recognition systems. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 481–485. ISCA.

Tuan Anh Nguyen and Anh Tuan Tran. 2021. Wanet - imperceptible warping-based backdoor attack. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Yao Qin, Nicholas Carlini, Garrison W. Cottrell, Ian J. Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 5231–5240. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 28492–28518. PMLR.

Vyas Raina, Rao Ma, Charles McGhee, Kate M. Knill, and Mark J. F. Gales. 2024. Muting whisper: A universal acoustic adversarial attack on speech foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7549–7565. Association for Computational Linguistics.

Pascal Scalart and Jozue Vieira Filho. 1996. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on*

Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996, pages 629–632. IEEE Computer Society.

Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. 2022. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *ACM MobiCom '22: The 28th Annual International Conference on Mobile Computing and Networking, Sydney, NSW, Australia, October 17 - 21, 2022*, pages 583–595. ACM.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2019. Label-consistent backdoor attacks. *CoRR*, abs/1912.02771.

Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. 2021. Backdoor attack against speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 2560–2564. IEEE.

A Appendix

A.1 Query-Budget Verification Results

Dataset	# Probes	τ	$\mathrm{VSR}_{\mathrm{clean}} \downarrow$	$\mathrm{VSR}_{\mathrm{prot}} \uparrow$	Gap $\uparrow$
AISHELL-1	20	0.0769	0.034	1.000	0.966
AISHELL-1	50	0.0357	0.043	1.000	0.957
AISHELL-1	100	0.0278	0.050	1.000	0.950
AISHELL-1	200	0.0233	0.050	1.000	0.950
LibriSpeech	20	0.1667	0.046	1.000	0.954
LibriSpeech	50	0.1155	0.050	1.000	0.950
LibriSpeech	100	0.0910	0.050	1.000	0.950
LibriSpeech	200	0.0729	0.049	1.000	0.951

Table 3: Decision-level black-box copyright protection of HPW under different query budgets. The threshold τ is selected as the 95th percentile of the clean recovery-ratio distribution. VSR is computed over $R = 1000$ random query subsets.

A.2 Threshold-Free AUROC Results for Robustness Studies

Setting	AUROC $\uparrow$	Mean $\mathrm{RR}_{\mathrm{clean}}$	Mean $\mathrm{RR}_{\mathrm{prot}}$
Baseline protected	1.000	0.0067	0.8856
Fine-tuning 200 steps	1.000	0.0065	0.8775
Fine-tuning 500 steps	1.000	0.0071	0.8868
Fine-tuning 1000 steps	1.000	0.0071	0.8834
Magnitude pruning 20%	1.000	0.0067	0.8818
Magnitude pruning 40%	1.000	0.0064	0.8021
Magnitude pruning 60%	0.342	0.0060	0.0000
Stationary denoise + retrain	1.000	0.0069	0.8746
Gaussian noise 20 dB	1.000	0.4862	0.9278
Resampling 16k–8k–16k	1.000	0.1051	0.8474
MP3 64 kbps	1.000	0.0114	0.8847
Opus 16 kbps	1.000	0.0722	0.8562
Opus 32 kbps	1.000	0.0146	0.8774
Stationary denoising	1.000	0.0235	0.8643

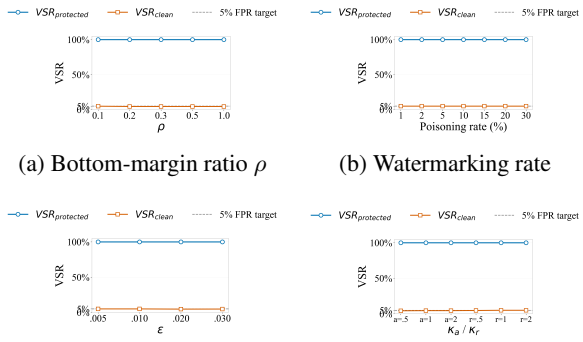
Table 4: Threshold-free separability under robustness settings. AUROC is computed from subset-level RR distributions with infringing models as positives and clean models as negatives.

A.3 STRIP Detection Results

Detection Metric	AUROC
Target Consistency	0.4299
Pairwise CER Variance	0.5633
Average Reference CER	0.5755

Table 5: STRIP-style anomaly detection results on AISHELL-1. AUROC close to 0.5 indicates near-random detection.

A.4 VSR Results for Ablation Studies



(a) Bottom-margin ratio ρ (b) Watermarking rate

(c) Perturbation bound ϵ (d) Margin parameters

Figure 5: VSR results for hyperparameter ablations on AISHELL-1. VSR is computed with $K = 100$, $R = 1000$, and a data-driven threshold selected as the 95th percentile of the corresponding clean RR distribution.

A.5 VSR Results for Robustness Studies

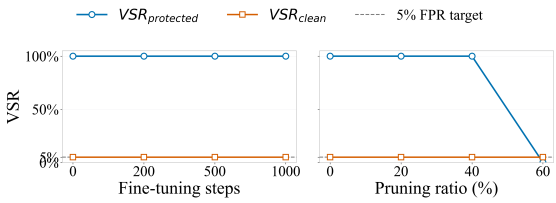


Figure 6: VSR results under model-level attacks on AISHELL-1. VSR is computed with $K = 100$, $R = 1000$, and a data-driven threshold selected from the corresponding clean RR distribution.

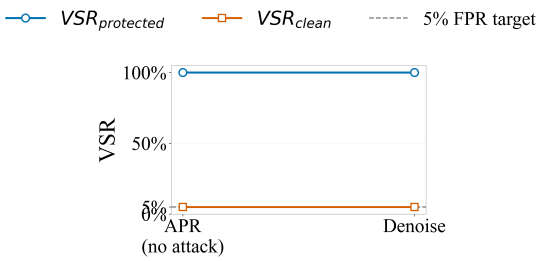


Figure 7: VSR results under data-level denoising sanitization on AISHELL-1. VSR is computed with $K = 100$, $R = 1000$, and a data-driven threshold selected from the corresponding clean RR distribution.

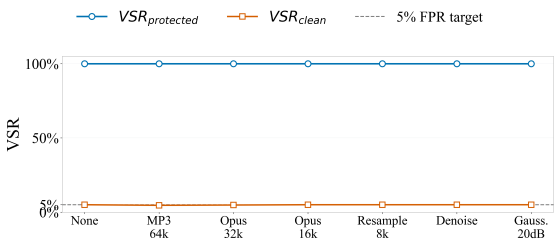


Figure 8: VSR results under audio-level transformations on AISHELL-1. For each transformation, the threshold is calibrated using the corresponding clean RR distribution under the same verification condition.

A.6 HPW Optimization Procedure

Algorithm 1 HPW watermark optimization

Require: Training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; watermark rate p; clean model A_c; perturbation bound ϵ; attack margin κ_a; readability margin κ_r; bottom-margin ratio ρ; stage 1 steps T_a; stage 2 steps T_r

Ensure: Protected dataset $\mathcal{D}^w$

Select watermark subset $\mathcal{I}_w \subset \{1, \ldots, N\}$ with $|\mathcal{I}_w| = \lfloor pN \rfloor$
Initialize a waveform perturbation δ
Freeze the clean model A_c
for $s = 1, \ldots, T_a$ **do**
    Sample a mini-batch $\mathcal{B} \subseteq \mathcal{I}_w$
    Construct $x_i^w = x_i + \delta$ for each $i \in \mathcal{B}$ with length masking
    Update δ by minimizing $\mathcal{L}_{\text{atk}}$ in Eq. (1)
    Project δ to $[-\epsilon, \epsilon]$ and clip x_i^w to the valid waveform range
end for
Set $\delta_0 \leftarrow \delta$
Construct intermediate dataset

$$\mathcal{D}^0 = \{(x_i, y_i) : i \notin \mathcal{I}_w\} \cup \{(x_i + \delta_0, y_i) : i \in \mathcal{I}_w\}$$

Fine-tune infringing teacher A_w on $\mathcal{D}^0$ with the standard automatic speech recognition loss
Freeze A_w
Initialize $\delta \leftarrow \delta_0$
for $s = 1, \ldots, T_r$ **do**
    Sample a mini-batch $\mathcal{B} \subseteq \mathcal{I}_w$
    Construct $x_i^w = x_i + \delta$ for each $i \in \mathcal{B}$ with length masking
    Compute token margins $m_{i,j}$ for content-token positions j using frozen A_w
    Select bottom-margin token set $\mathcal{T}_{i,\rho}$
    Update δ by minimizing $\mathcal{L}_{\text{read}}$ in Eq. (3)
    Project δ to $[-\epsilon, \epsilon]$ and clip x_i^w to the valid waveform range
end for
Set $\delta^\star \leftarrow \delta$
Construct final protected dataset

$$\mathcal{D}^w = \{(x_i, y_i) : i \notin \mathcal{I}_w\} \cup \{(x_i + \delta^\star, y_i) : i \in \mathcal{I}_w\}$$

return $\mathcal{D}^w$

A.7 Watermarked-Probe Transcription Quality

Dataset	Method	Clean Err. ↓	WM Err. ↓	WM EM Fail ↓	Δ Err. ↓
AISHELL-1	BadNets	0.0712	0.9668	1.0000	0.8956
AISHELL-1	WaNet	0.0713	0.9668	1.0000	0.8955
AISHELL-1	ReFool	0.0710	0.9668	1.0000	0.8958
AISHELL-1	ASR Watermark	0.0715	0.9667	0.9999	0.8952
AISHELL-1	FreqTone	0.0712	0.9660	0.9997	0.8948
AISHELL-1	HighFreq	0.0714	0.9665	0.9999	0.8951
AISHELL-1	AdvBackdoor	0.0713	0.9668	1.0000	0.8955
AISHELL-1	Domain Watermark	0.0701	0.1107	0.6074	0.0406
AISHELL-1	HPW	0.0601	0.0792	0.4902	0.0191
LibriSpeech train-clean-100	BadNets	0.0597	0.9876	1.0000	0.9280
LibriSpeech train-clean-100	WaNet	0.0596	0.9877	1.0000	0.9281
LibriSpeech train-clean-100	ReFool	0.0596	0.9570	0.9977	0.8974
LibriSpeech train-clean-100	ASR Watermark	0.0595	0.9742	0.9973	0.9147
LibriSpeech train-clean-100	FreqTone	0.0596	0.9768	0.9996	0.9172
LibriSpeech train-clean-100	HighFreq	0.0599	0.9787	0.9992	0.9189
LibriSpeech train-clean-100	AdvBackdoor	0.0597	0.9785	0.9996	0.9188
LibriSpeech train-clean-100	Domain Watermark	0.0591	0.0837	0.6725	0.0246
LibriSpeech train-clean-100	HPW	0.0513	0.0594	0.5649	0.0081

Table 6: Protected-model transcription quality on clean and watermarked verification probes. Err. denotes CER for AISHELL-1 and WER for LibriSpeech. All errors are computed against the original transcript, including backdoor-based baselines, to measure transcript preservation rather than target-phrase success.

WM EM Fail is the direct exact-match failure rate on all watermarked probes and is not identical to $\mathcal{H}$ in Table 1, which measures the recovery drop relative to clean-probe exact matches. For backdoor-based baselines, the high WM Err. and WM EM Fail values indicate that triggered probes are transcribed as the watermark target rather than preserving the original transcript. In contrast, HPW keeps the watermarked-probe CER/WER close to the clean-probe error while still enabling verification.

A.8 Failure Taxonomy for HPW

Dataset	Failure type	% of failures	% of all probes
AISHELL-1	Clean already wrong, WM similar	59.7	29.3
AISHELL-1	Clean wrong, WM worse	18.1	8.9
AISHELL-1	Clean correct, minor WM error	8.6	4.2
AISHELL-1	Clean correct, moderate WM error	7.9	3.9
AISHELL-1	Clean wrong, WM better but not exact	5.0	2.5
AISHELL-1	Clean correct, severe WM error	0.7	0.3
LibriSpeech	Clean already wrong, WM similar	74.1	41.8
LibriSpeech	Clean wrong, WM worse	9.3	5.3
LibriSpeech	Clean correct, minor WM error	8.9	5.0
LibriSpeech	Clean correct, moderate WM error	3.9	2.2
LibriSpeech	Clean wrong, WM better but not exact	3.4	1.9
LibriSpeech	Clean correct, severe WM error	0.4	0.2

Table 7: Taxonomy of HPW exact-match failures on watermarked probes. We separate failures already present on clean probes from errors introduced or worsened by the watermark. Severe watermark-induced failures from clean exact-match probes are rare.

The taxonomy shows that most exact-match failures do not correspond to severe watermark-induced semantic corruption. On AISHELL-1, 59.7% of failures are already present on clean probes with similar watermarked outputs, and only 0.3% of all probes are clean-correct cases that become severe watermarked errors. On LibriSpeech, the corresponding severe watermark-induced rate is 0.2%. This supports the interpretation that HPW's remaining exact-match failures are mostly local or pre-existing ASR errors, although severe failures are not eliminated completely.

A.9 Fixed-Threshold Verification Results

Protocol. The VSR results in the main tables use condition-specific calibration, where the threshold τ is selected from the corresponding clean RR distribution under each evaluation condition. This protocol evaluates whether protected and clean models are separable when a matched calibration distribution is available. To approximate a more operational deployment setting, we further evaluate a fixed-threshold protocol. For each dataset, we calibrate τ_{fix} once using the default in-domain clean model with $K = 100$ and $R = 1000$, and then reuse the same threshold across transferability, robustness, and ablation settings. This gives $\tau_{\text{fix}} = 0.0323$ for AISHELL-1 and $\tau_{\text{fix}} = 0.0909$ for LibriSpeech. Successful fixed-threshold verification requires high fixed VSR_{prot} and low fixed $\text{VSR}_{\text{clean}}$. High values for both indicate false positives rather than successful copyright protection.

Transferability. Table 9 reports fixed-threshold VSR under cross-dataset, model-scale, and non-Whisper architecture shifts. The in-domain fixed threshold works well in the matched in-domain setting, where fixed VSR_{prot} is 1.000 and fixed $\text{VSR}_{\text{clean}}$ remains low. However, transfer settings often increase clean RR substantially. As a result, fixed $\text{VSR}_{\text{clean}}$ becomes high for most model-scale, cross-dataset, and non-Whisper transfer settings. This indicates that threshold transfer can cause high false-positive rates under large distribution or architecture shifts. Therefore, fixed-threshold deployment under such shifts requires a broader and more conservative clean calibration pool, and non-Whisper architectures remain outside the main threat model of HPW.

Dataset	Type	Reference	Watermarked output	Comment
AISHELL-1	Local error	应用于汽车医疗等行业	遥用于汽车医疗等行业	Local character substitution at the beginning; the main content remains unchanged.
AISHELL-1	Moderate error	带动了区域板块的扩张	带动了七亿百块的扩张	Several characters differ, but partial content such as 带动了 and 扩张 remains.
AISHELL-1	Severe error	提升公共交通设施装备水平	仍然供应交通涉事光碑水中	Rare severe content mismatch with substantial semantic drift.
LibriSpeech	Tokenization	SAID SHE POINTING TO THE PLAYTHINGS SEE	SAID SHE POINTING TO THE PLAY THINGS SEE	Word split changes exact match while preserving the spoken content.
LibriSpeech	Tokenization	REJOICE IN THY GROWTH SAID THE SUNBEAMS	REJOICE IN THY GROWTH SAID THE SUN BEAMS	Word split changes exact match while preserving the spoken content.
LibriSpeech	Severe error	VERSE TWO	FIRST TUBE	Rare severe content mismatch on a very short utterance.

Table 8: Representative HPW failure cases on watermarked probes. These examples illustrate that many failures are local token or word differences, while a small number are severe transcription errors.

Setting	Dataset	$\mathbf{RR}_{\text{clean}}$	$\mathbf{RR}_{\text{prot}}$	τ_{fix}	Fixed $\mathbf{VSR}_{\text{clean}}$	Fixed $\mathbf{VSR}_{\text{prot}}$
HPW in-domain	AISHELL-1	0.0070	0.8890	0.0323	0.030	1.000
Whisper-tiny	AISHELL-1	0.2674	0.7252	0.0323	1.000	1.000
Whisper-base	AISHELL-1	0.0951	0.8101	0.0323	0.887	1.000
Whisper-medium	AISHELL-1	0.4379	0.9264	0.0323	1.000	1.000
Whisper-large	AISHELL-1	0.4466	0.9347	0.0323	1.000	1.000
LibriSpeech → AISHELL-1	AISHELL-1	0.4396	0.6269	0.0323	1.000	1.000
HPW in-domain	LibriSpeech	0.0396	0.9044	0.0909	0.036	1.000

Table 9: Fixed-threshold VSR for transferability settings. τ_{fix} is calibrated once on the default in-domain clean model for each dataset and reused across transfer settings. RR values are subset-level means over $R = 1000$ trials with $K = 100$. High fixed VSR_{prot} is meaningful only when fixed $\text{VSR}_{\text{clean}}$ remains low.

Robustness. Table 10 reports fixed-threshold VSR for model-level, data-level, and audio-level robustness settings on AISHELL-1. The fixed threshold remains reliable for model-level attacks and data-level denoise sanitization: fixed VSR_{prot} remains 1.000, while fixed $\text{VSR}_{\text{clean}}$ stays close to the in-domain false-positive level. In contrast, several audio-level transformations substantially raise clean RR and therefore cause high fixed-threshold false-positive rates. For example, resampling and Opus 16 kbps yield fixed $\text{VSR}_{\text{clean}}$ of 0.986 and 0.905, respectively. This does not contradict the condition-specific AUROC and VSR results in the main text; instead, it shows that threshold transfer is sensitive to channel shifts. Practical deployment should therefore use channel-aware calibration or reject verification probes whose channel conditions differ substantially from the calibration setting.

Setting	$\mathbf{RR}_{\text{clean}}$	$\mathbf{RR}_{\text{prot}}$	Fixed $\mathbf{VSR}_{\text{clean}}$	Fixed $\mathbf{VSR}_{\text{prot}}$	Gap
Model- and data-level attacks					
HPW in-domain	0.0070	0.8890	0.030	1.000	0.970
Fine-tuning 200 steps	0.0065	0.8775	0.031	1.000	0.969
Fine-tuning 500 steps	0.0071	0.8868	0.037	1.000	0.963
Fine-tuning 1000 steps	0.0071	0.8834	0.033	1.000	0.967
Magnitude pruning 20%	0.0067	0.8818	0.032	1.000	0.968
Magnitude pruning 40%	0.0064	0.8021	0.031	1.000	0.969
Magnitude pruning 60%	0.0064	0.0000	0.029	0.000	-0.029
Stationary denoise + retrain	0.0069	0.8746	0.033	1.000	0.967
Audio-level transformations					
Gaussian noise 20 dB	0.4862	0.9278	1.000	1.000	0.000
Resampling 16k–8k–16k	0.1051	0.8474	0.986	1.000	0.014
MP3 64 kbps	0.0114	0.8847	0.085	1.000	0.915
Opus 16 kbps	0.0722	0.8562	0.905	1.000	0.095
Opus 32 kbps	0.0146	0.8774	0.143	1.000	0.857
Stationary denoising	0.0235	0.8643	0.284	1.000	0.716

Table 10: Fixed-threshold VSR for robustness settings on AISHELL-1. The fixed threshold $\tau_{\text{fix}} = 0.0323$ is calibrated once on the in-domain clean model and reused across robustness settings. RR values are subset-level means over $R = 1000$ trials with $K = 100$. High fixed $\text{VSR}_{\text{clean}}$ indicates false positives under threshold transfer.

Threshold sweep. We also sweep $\tau \in \{0.05, 0.10, 0.20, 0.30, 0.50, 0.75, 0.90\}$ to understand the threshold-sensitivity of the verification statistic. In-domain verification is stable across a broad range of thresholds. On AISHELL-1, VSR_{prot} remains 1.000 for $\tau \leq 0.50$ and is still 0.992 at $\tau = 0.75$, while $\text{VSR}_{\text{clean}}$ becomes 0 for $\tau \geq 0.10$. On LibriSpeech, VSR_{prot} also remains 1.000 for $\tau \leq 0.50$, with $\text{VSR}_{\text{clean}} = 0$ for

$\tau \geq 0.20$. However, shifted settings require much larger thresholds to control false positives, and very large thresholds can reduce protected detection. For example, under AISHELL-1 Whisper-medium transfer, $\tau = 0.50$ reduces $\mathrm{VSR}_{\mathrm{clean}}$ to 0.143 while keeping $\mathrm{VSR}_{\mathrm{prot}} = 1.000$, but $\tau = 0.90$ lowers $\mathrm{VSR}_{\mathrm{prot}}$ to 0.736. These results confirm that fixed-threshold deployment should be calibrated to the expected model family and channel conditions rather than assumed to transfer universally.

A.10 Stage-wise Ablation Results

Setting	CER ↓	$\mathrm{RR}_{\mathrm{clean}}$ ↓	$\mathrm{RR}_{\mathrm{prot}}$ ↑	$\mathrm{VSR}_{\mathrm{clean}}$ ↓	$\mathrm{VSR}_{\mathrm{prot}}$ ↑	SNR ↑
Stage 2 only	**0.0601**	0.9995	0.9985	0.046	0.053	**44.54**
Stage 1 only	0.0609	0.0069	0.8755	0.049	**1.000**	24.32
Stage 1+2	**0.0601**	**0.0066**	0.8889	0.050	**1.000**	24.36

Table 11: Ablation on the necessity of the two training stages on AISHELL-1.

Evaluation on Wav2Vec2-large CTC. To further evaluate whether the transcript-preserving verification principle extends beyond Whisper-style encoder-decoder ASR, we instantiate HPW on Wav2Vec2-large (Baevski et al., 2020) with a Connectionist Temporal Classification (CTC) ASR head, which predicts frame-level token distributions without autoregressive decoding. In this setting, the watermark is optimized, trained, and verified entirely within the Wav2Vec2-large CTC pipeline. Since CTC does not use autoregressive decoding, Stage 1 replaces the first-token attack with an untargeted CTC objective that lowers the clean model's likelihood of the original transcript on watermarked inputs. After training an intermediate protected Wav2Vec2-large model on the Stage 1 watermarked dataset, Stage 2 refines the perturbation by improving low-confidence aligned transcript positions under the infringing teacher. As in the main setting, all watermarked utterances keep their original transcripts, and verification remains black-box: the verifier only observes the decoded transcripts of paired clean and watermarked probes.

Dataset	Err. ↓	$\mathrm{RR}_{\mathrm{clean}}$ ↓	$\mathrm{RR}_{\mathrm{prot}}$ ↑	$\mathrm{VSR}_{\mathrm{clean}}$ ↓	$\mathrm{VSR}_{\mathrm{prot}}$ ↑	AUROC ↑	SNR ↑
AISHELL-1	0.0564	0.0847	0.7463	0.049	0.972	0.993	23.91
LibriSpeech	0.0478	0.1124	0.7826	0.050	0.986	0.997	22.84

Table 12: Black-box copyright protection on Wav2Vec2-large CTC. The watermark is optimized and verified directly on Wav2Vec2-large.

Table 12 shows that HPW also produces a clear ownership signal in a CTC-based ASR system. Although the infringing recovery ratio is lower than in the Whisper encoder-decoder setting, the protected and clean models remain well separated. On AISHELL-1, Wav2Vec2-large reaches $\mathrm{RR}_{\mathrm{prot}} = 0.7463$ versus $\mathrm{RR}_{\mathrm{clean}} = 0.0847$; on LibriSpeech, it reaches $\mathrm{RR}_{\mathrm{prot}} = 0.7826$ versus $\mathrm{RR}_{\mathrm{clean}} = 0.1124$. The AUROC values are above 0.99 on both datasets, indicating strong threshold-free separability. These results suggest that the transcript-preserving verification principle is not tied to a single ASR implementation, while architecture changes can still reduce the average recovery margin and should be considered during calibration.

A.11 Artifact Licenses, Terms of Use, and Intended Use

We used two public ASR datasets from Hugging Face. AISHELL-1 (`AISHELL/AISHELL-1`) is listed on Hugging Face with an Apache-2.0 license. Its dataset card describes it as an open-source Mandarin speech corpus released by Beijing Shell Shell Technology Co., Ltd., with recordings from 400 speakers from different accent regions in China, downsampled to 16 kHz, and free for academic use. The LibriSpeech clean-100 dataset used in our experiments (`nguyenvulebinh/libris_clean_100`) is listed on Hugging Face with a CC-BY-4.0 license. We use the `train.clean.100` split for training and the `test.clean` split for evaluation. The Hugging Face dataset page reports 28.5k rows for `train.clean.100` and 2.62k rows for `test.clean`.

We use these datasets only for research evaluation of ASR dataset copyright protection. We do not redistribute the original datasets. Any derived watermarked data created in our experiments remains subject to the licenses and terms of the underlying datasets and will not be redistributed unless permitted by the corresponding license and terms of use. We use existing public ASR benchmark datasets and do not collect new speech data or recruit new speakers. We rely on the dataset creators' release and documentation for the original data collection, consent, and anonymization procedures. Our work does not infer speaker identities, protected attributes, or other sensitive personal attributes, and our watermarking procedure does not add new semantic content to the transcripts. We did not perform an additional large-scale manual

audit for personally identifying information or offensive content. Any derived watermarked audio remains subject to the privacy, distribution, and usage conditions of the underlying datasets.

A.12 Software Implementations and Package Settings

We implemented our experiments in PyTorch 2.2.2 with CUDA 12.1. Whisper fine-tuning and inference were based on Hugging Face Transformers 4.51.3 using the `openai/whisper-small` checkpoint; the standalone OpenAI Whisper package was not used. Datasets were loaded using Hugging Face Datasets 2.21.0. Audio loading, resampling, and waveform processing used librosa 0.11.0, SoundFile 0.13.1, NumPy 1.26.4, and SciPy 1.17.1. WER and CER were computed using jiwer 4.0.0 together with our own wrapper functions. For English transcripts, text normalization lowercases text, removes ASCII punctuation, collapses repeated whitespace, and strips leading/trailing spaces. For Chinese transcripts, Unicode punctuation is removed, and whitespace is deleted before character-level comparison. Audio compression experiments used FFmpeg 4.4.2 with the bitrate settings reported in Section 5.4. Unless otherwise specified, package parameters followed their default settings.

A.13 Use of AI Assistants

We used AI assistants to support code development and paper polishing. In particular, AI assistants were used for coding and debugging assistance, improving the clarity and grammar of the manuscript. All research ideas, experimental designs, implementation decisions, experimental results, analyses, and final claims were developed, checked, and approved by the authors. ChatGPT was not listed as an author and was not used to independently generate, validate, or interpret the experimental results.

A.14 Fixed-Threshold Verification Results for Main Baselines

Protocol. The VSR results in the main table use condition-specific calibration, where τ is selected from the corresponding clean RR distribution for each method and dataset. As a conservative single-threshold diagnostic for the main benchmark, we also evaluate a shared fixed threshold: for each dataset, we calibrate τ_{fix} once on the in-domain clean model under `HPW`, giving $\tau_{\text{fix}} = 0.0323$ for AISHELL-1 and $\tau_{\text{fix}} = 0.0909$ for LibriSpeech. We then reuse this threshold for all main-table methods using subset-level RR scores with $K = 100$ and $R = 1000$. This shared-threshold protocol is not intended to be an optimal threshold for every baseline, since different watermark families use different verification targets and RR scales. Instead, it tests whether `HPW`'s main in-domain separation remains reliable without method-specific threshold tuning. Successful fixed-threshold verification requires high fixed VSR_{prot} and low fixed $\text{VSR}_{\text{clean}}$; high values for both indicate false positives under the shared threshold.

Under this shared fixed-threshold protocol, target-transcript backdoor baselines still show strong native target-phrase ownership signals, because their clean RR is essentially zero while their protected RR is large. However, this separability comes from target-phrase injection rather than original-transcript recovery, as reflected by their near-1.0 WM CER/WER values. Domain Watermark exceeds the shared threshold for both clean and infringing models, leading to high false-positive rates under this diagnostic. In contrast, `HPW` preserves low fixed $\text{VSR}_{\text{clean}}$ while maintaining fixed $\text{VSR}_{\text{prot}} = 1.000$ on both datasets, indicating that its in-domain main-table result does not rely on method-specific threshold tuning.

Method	AISHELL-1 (Chinese)					LibriSpeech (English)				
	CER ↓	**WM CER ↓**	**Fixed VSR$_{\text{clean}}$ ↓**	**Fixed VSR$_{\text{prot}}$ ↑**	**SNR ↑**	**WER ↓**	**WM WER ↓**	**Fixed VSR$_{\text{clean}}$ ↓**	**Fixed VSR$_{\text{prot}}$ ↑**	**SNR ↑**
Backdoor-based Watermark Baselines										
BadNets	0.0712	0.9668	0.000	1.000	19.57	0.0597	0.9876	0.000	1.000	17.15
WaNet	0.0713	0.9668	0.000	1.000	17.03	0.0596	0.9877	0.000	1.000	10.64
ReFool	0.0710	0.9668	0.000	1.000	24.97	0.0596	0.9570	0.000	1.000	21.05
ASR Watermark	0.0715	0.9667	0.000	1.000	18.86	0.0595	0.9742	0.000	1.000	27.93
FreqTone	0.0712	0.9660	0.000	1.000	10.80	0.0596	0.9768	0.000	1.000	19.87
HighFreq	0.0714	0.9665	0.000	1.000	12.74	0.0599	0.9787	0.000	1.000	21.81
AdvBackdoor	0.0713	0.9668	0.000	1.000	7.33	0.0597	0.9785	0.000	1.000	16.57
Harmless Baselines										
Domain Watermark	0.0701	0.1107	1.000	1.000	3.42	0.0591	0.0837	1.000	1.000	4.90
HPW (Ours)	**0.0601**	**0.0792**	0.030	1.000	24.36	**0.0513**	**0.0594**	0.036	1.000	22.20

Table 13: Shared fixed-threshold VSR for HPW and baselines. τ_{fix} is calibrated once on the in-domain clean model under HPW for each dataset and reused for all methods ($\tau_{\text{fix}} = 0.0323$ for AISHELL-1 and 0.0909 for LibriSpeech; $K = 100$, $R = 1000$). This is a single-threshold diagnostic rather than an optimal per-method deployment threshold, since different watermark families use different verification targets. The remaining columns follow the same definitions as Table 1.