TTS-Guard: Black-Box Ownership Verification of Text-to-Speech Models via Adaptive Adversarial Speaker-Pair Fingerprints

Xubin Yue[1,2] Zhenhua Xu[1,2] Zhebo Wang[1,2] Mengting Li[4]
Zijie Zhou[5] Wenpeng Xing[1,2,3] Dezhang Kong[1,2,3*] Meng Han[2,1*]

[1] Zhejiang University [2] Binjiang Institute of Zhejiang University
[3] GenTel.io [4] Hangzhou International Innovation Institute, Beihang University
[5] China University of Petroleum (Beijing)

{yuexubin, xuzhenhua0326, breynald, wpxing, kdz, mhan}@zju.edu.cn

Abstract

The rapid maturation of zero-shot Text-to-Speech (TTS) models has turned high-quality voice cloning into a widely available capability, raising acute concerns over unauthorised replication, fine-tuning and resale of proprietary speech models. Yet ownership verification for TTS remains largely open: speech is a continuous waveform whose perturbations are easily destroyed by routine signal processing, and the human auditory system imposes a much tighter perceptual budget than vision. We present **TTS-Guard**, a black-box ownership verification framework for TTS models built on *adversarial speaker-pair fingerprints*. TTS-Guard(i) selects key speaker pairs in a *dual* embedding space for architecture-agnostic stealth;(ii) optimises a perturbation through an *adaptive curriculum* of shadow models covering fine-tuning, pruning, quantisation and distillation; and (iii) aggregates black-box queries into a calibrated *Verification Confidence Score*. On five mainstream TTS systems, TTS-Guard reaches an average Fingerprint Success Rate of 96.4% at a False Positive Rate of 5.8%, while preserving intelligibility and naturalness. The fingerprint remains effective against ten audio attacks, six model modifications, and two state-of-the-art adversarial purifiers.

1 Introduction

Modern Text-to-Speech (TTS) systems such as XTTS (Casanova et al., 2024), CosyVoice 2 (Du et al., 2024), F5-TTS (Chen et al., 2025b), OpenVoice (Qin et al., 2023) and VibeVoice (Peng et al., 2025) can synthesise indistinguishably natural speech from a few seconds of reference audio. Open-source releases dramatically lowered the entry barrier for legitimate developers but simultaneously enabled an underground market in which fine-tuned proprietary checkpoints are illegally redistributed, repackaged behind APIs and even sold as "new" commercial voices (Yi et al., 2023; Li et al., 2023). As a concrete instance of the threat, an adversary who has obtained a leaked checkpoint may apply LoRA fine-tuning, INT8 quantisation or magnitude pruning to disguise its provenance and then deploy it through a black-box API (Guo et al., 2025); the rightful owner has no access to the weights and only a limited number of API queries with which to prove infringement.

Two structural properties of speech make copyright techniques designed for text and images ill-suited to this setting. First, while pixels and tokens are sampled from finite or low-dimensional alphabets, a speech waveform is a high-bandwidth, *continuous* signal whose statistics are reshaped by every routine processing step (resampling, MP3/Opus codecs, dynamic range compression, equalisation, pink-noise injection) (Roman et al., 2024; Liu et al., 2025). A fingerprint hidden in raw samples is therefore inherently fragile compared with the image case. Second, the human auditory system is highly sensitive to non-harmonic artefacts: any perturbation must comply with a strict psychoacoustic budget (Yu et al., 2023; Zhang et al., 2025b). The combination of fragility and tight perceptual budget rules out naïve ports of image watermarks (Adi et al., 2018; Uchida et al., 2017) and LLM fingerprints (Xu et al., 2024; Yamabe et al., 2025; Xiong et al., 2026).

A second axis along which TTS protection differs from prior work is the *conditioning mechanism* on which most modern systems rely. Multi-speaker and zero-shot TTS models do not learn an autonomous speaker representation; they are explicitly conditioned on a speaker embedding extracted from the reference clip (Jia et al., 2018; Chen et al., 2025a). Mastering the speaker-embedding manifold is therefore equivalent to controlling identity attribution at the output. Adversarial-perturbation attacks on this manifold have been studied in the *defensive* setting (preventing voice cloning of a spe-

*Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 4958–4973
October 24-29, 2026 ©2026 Association for Computational Linguistics

cific speaker) (Huang et al., 2021; Yu et al., 2023; Zhang et al., 2025b; Park et al., 2025). We turn the same primitive into an *offensive* ownership signal: an adversarial perturbation crafted by the model owner can act as a unique key whose successful triggering of a target speaker proves possession of the underlying model.

Concretely, we propose **TTS-Guard**, the first black-box ownership verification framework for TTS models. Beyond the basic adversarial-fingerprint formulation, TTS-Guard contributes four innovations:

1. **Dual-space Key Speaker Pair (KSP) selection.** Earlier reactive defences (Yu et al., 2023; Park et al., 2025) match a single speaker pair in one embedding space. We instead jointly minimise distance in two architecturally disjoint verification spaces (Resemblyzer and ECAPA-TDNN (Desplanques et al., 2020)) so the resulting pair is both acoustically close and architecture-agnostic, sharply improving fingerprint robustness against unknown verifiers.

2. **Adaptive Curriculum Shadow Optimisation (ACSO).** We schedule a curriculum that progressively adds LoRA fine-tuning, magnitude pruning, INT8/INT4 quantisation and knowledge distillation, so that the fingerprint generalises across the *entire* derivative manifold rather than only one slice of it.

3. **Decoupling for High Specificity.** A decoupling loss explicitly maximises the perturbation distance on non-target models, distinguishing TTS-Guard from generic adversarial-transferability work and yielding the low false-positive rates required for forensic use.

4. **Calibrated Verification Confidence Score (VCS).** Repeated black-box queries with random text prompts are aggregated into a bootstrap p-value that converts deterministic FSR into a calibrated, court-defensible verdict.

We additionally derive an information-theoretic lower bound that ties the achievable FSR to the Bhattacharyya distance between KSP embedding distributions, providing the first theoretical justification for KSP-based fingerprinting (§4.4). Extensive experiments on five mainstream TTS systems show that TTS-Guard achieves a per-query $\overline{\mathrm{FSR}} = 96.4\%$ at a per-query $\overline{\mathrm{FPR}} = 5.8\%$, which the calibrated multi-query VCS test tightens to a decision-level false-positive rate of 4.7% at $K{=}20$ (and $\leq 0.5\%$ at $K{=}40$); it remains above 0.58 FSR under ten audio attacks and six model modifications, and survives two state-of-the-art adversarial purifiers (De-AntiFake (Fan et al., 2025) and SafeEar (Li et al., 2024)) that defeat AntiFake-style perturbations. To our knowledge, this is the first study to systematically benchmark TTS ownership verification against (a) modern derivative attacks (quantisation, distillation), (b) commercial codecs and (c) adversarial purification.

2 Related Work

2.1 Audio-level Protection

Existing audio-domain protections fall into two threads, both operating at the signal level.

Audio watermarking. embeds a recoverable signature into the waveform. AudioSeal (Roman et al., 2024), XAttnMark (Liu et al., 2025), WavMark (Chen et al., 2023a) and AudioMarkNet (Zong et al., 2025) train neural encoder–detector pairs for sample-level detection or deepfake forensics; P2Mark (Ren et al., 2025) watermarks vocoder weights under white-box access; Audio Watermark (Zong et al., 2025) extends the idea to training datasets.

Anti-cloning audio protection. hides adversarial noise in user recordings so that the speaker embedding is pushed away from any clone target (Yu et al., 2023; Zhang et al., 2025b; Li et al., 2023; Huang et al., 2021). The recently proposed De-AntiFake purifier (Fan et al., 2025) largely neutralizes this thread, confirming that signal-level perturbations are denoisable.

Both threads bind the credential to a signal—an output waveform, a training corpus, or a user upload—and break down once the model checkpoint itself leaks: re-synthesized audio carries no watermark, and any perturbation the attacker chooses to ignore simply does not exist at inference. TTS-Guard is, to our knowledge, the first framework to shift the carrier of ownership from the signal to the model's input–output behavior in the speech domain, by re-purposing the adversarial primitive as a verifier-side key delivered through the API at attribution time.

2.2 Model-level Ownership Verification in Other Modalities

Model-level ownership verification has so far only been developed for vision and text. In vision, IP-

Guard (Cao et al., 2021) and DeepJudge (Chen et al., 2022a) probe classifier decision boundaries with adversarial queries. In NLP, a richer line of work fingerprints LLMs through instructional backdoors or parameter-space invariants, including IF (Xu et al., 2024), MergePrint (Yamabe et al., 2025), iSeal (Xiong et al., 2026), HuRef (Zeng et al., 2024), CTCC (Xu et al., 2025), PREE (Yue et al., 2025), and watermarking-style detectors such as Kirchenbauer et al. (2023), part of a broader effort to secure LLM-driven systems (Kong et al., 2025).

None of these designs has been ported to speech synthesis, and a direct port is non-trivial: a TTS fingerprint must survive heterogeneous derivative attacks (LoRA, pruning, INT4/INT8 quantization, distillation) and a perceptually constrained waveform pipeline (codecs, DRC, EQ, adversarial purification), and must operate through speaker-embedding conditioning rather than discrete tokens. We are careful to delimit our contribution: the *adversarial-query* paradigm itself originates with IPGuard/DeepJudge and the *multi-query statistical test* is standard, whereas our novelty lies in (i) dual-space KSP selection that makes a low-budget speaker-pair perturbation transferable across verifiers, (ii) the ACSO curriculum that hardens the fingerprint against TTS-specific derivative attacks, and (iii) the first information-theoretic (Bhattacharyya) justification tying KSP separability to attribution reliability. We benchmark against ported IPGuard and IF baselines in §6 (Table 2) to make this boundary empirically concrete.

3 Threat Model

Adversary. The adversary aims to illegally obtain and use the protected TTS model for personal or commercial benefit while denying the original owner's copyright. We assume the adversary has acquired the target checkpoint M_p through unauthorized means, such as leakage, insider access, or reseller redistribution. To evade ownership verification, the adversary may apply **model-level transformations**, including fine-tuning, pruning, quantization, and distillation, as well as **audio-level perturbations**, such as filtering, noise addition, resampling, codec compression, dynamic range compression, equalization, and adaptive purification attacks (e.g., De-AntiFake (Fan et al., 2025) and SafeEar (Li et al., 2024)). A deployed API may further apply **input-side preprocessing** to the reference clip before synthesis—normalization, denoising, resampling, or rejection of suspicious inputs; all of these reduce to signal-level operations on the reference audio and are therefore covered by our audio-level robustness evaluation (§6.2, Appendix D.3). We do *not* assume the adversary replaces the internal speaker encoder: doing so requires re-aligning and retraining the entire conditioning pathway at a cost comparable to training from scratch, which nullifies the economic motive for theft (§7).

Defender. The defender is the model owner, whose goal is to verify whether a suspicious deployed service originates from the protected model. Considering practical constraints, the defender has no knowledge of the adversary's specific modifications, datasets, or deployment pipeline, and can only interact with the suspicious system through its **black-box API** by providing reference audio and text prompts and observing the synthesized waveform. The defender has no access to internal parameters, architecture, gradients, or preprocessing and post-processing modules, but has full white-box access to the protected model M_p during fingerprint generation.

Security goal. Given a suspicious model M^*, the defender aims to determine whether it is derived from M_p. Formally, this is a hypothesis test between H_0: "M^* is unrelated to M_p" and H_1: "M^* is a derivative of M_p". The objective is to control the false-positive probability $\Pr[\text{accept } H_1 \mid H_0] \leq \alpha$ (with $\alpha = 0.05$) while maximizing the true-positive probability, i.e., the fingerprint success rate (FSR).

4 Method

4.1 Notation

Let V denote a speaker-verification network producing L_2-normalised embeddings $V(\cdot) \in \mathbb{R}^d$. Let E_M denote the speaker encoder *internal* to a TTS model M. We write M_p for the protected model and $\{M_d^{(i)}\}_{i=1}^{N_d}$ for unrelated non-target models.

4.2 Stage 1 — Dual-Space Key Speaker Pair Selection

The choice of embedding space directly governs both the perturbation budget and the cross-architecture robustness of the fingerprint. We therefore replace single-space matching with a dual-space objective. Given N candidate speakers $\{s_1, \ldots, s_N\}$ and per-speaker centroids $\mu_i^{(R)}$,

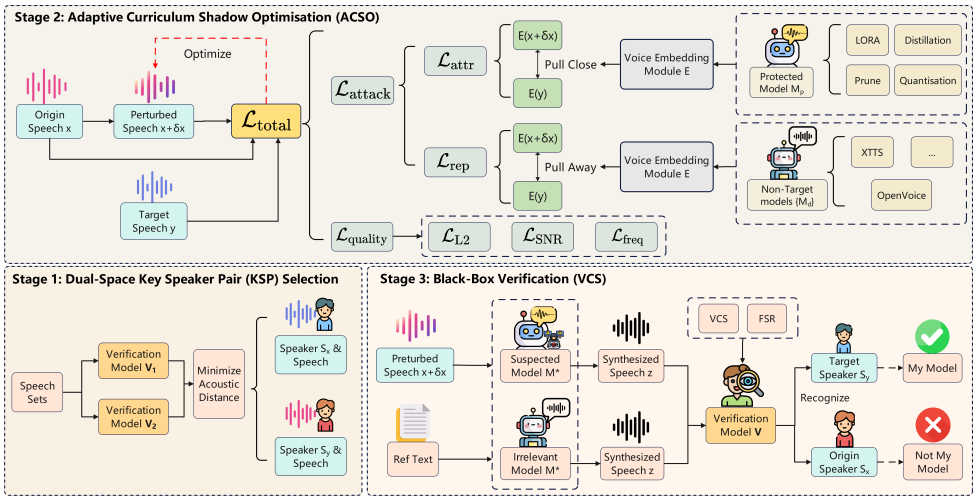

Figure 1: TTS-Guard pipeline. (1) Dual-space KSP selection identifies a source/target speaker pair (S_x, S_y) that is acoustically close in both Resemblyzer and ECAPA spaces. (2) Adaptive Curriculum Shadow Optimisation generates the perturbation δ_x under an evolving ensemble that progressively introduces LoRA, pruning, quantisation and distillation derivatives. (3) Black-box verification feeds $x+\delta_x$ into the suspect model, synthesises speech under a random text prompt, and aggregates VCS p-values across K queries.

$\mu_i^{(E)}$ in the Resemblyzer and ECAPA spaces respectively, we choose the source/target pair as

$$(S_x, S_y) = \arg\min_{i \neq j} \; \alpha \, \|\mu_i^{(R)} - \mu_j^{(R)}\|_2 + (1-\alpha) \, \|\mu_i^{(E)} - \mu_j^{(E)}\|_2 \tag{1}$$

with $\alpha = 0.5$. From each of the two speakers we keep the clip closest to its own centroid as the source x and target y exemplars. Equation (1) can be solved exactly because $|N|^2$ pair distances are computed once in each space.

Why Euclidean rather than cosine distance? Both candidate verifiers (Resemblyzer and ECAPA-TDNN) are trained with margin-based triplet objectives whose anchor-positive geometry is calibrated in L_2. We further verified empirically (Table 10) that Euclidean selection yields a 7–11% higher FSR than cosine selection because magnitude information encodes timbre intensity that cosine normalisation discards.

4.3 Stage 2 — Adaptive Curriculum Shadow Optimisation

Given a source utterance x and a target speaker reference y, we optimise an additive perturbation δ_x so that the protected input $x' = x + \delta_x$ is attributed to the target speaker under a broad family of derivative models, while preserving perceptual speech quality.

4.3.1 Curriculum of Shadow Models

To better approximate this derivative space, we organise shadow models into four stages with increasing modification severity:

- **Stage A:** K_A=2 LoRA-finetuned shadows trained on disjoint LJSpeech splits;
- **Stage B:** K_B=2 magnitude-pruned shadows with pruning ratios $p \in \{30\%, 50\%\}$;
- **Stage C:** K_C=2 post-training quantised shadows (INT8 and INT4);
- **Stage D:** K_D=1 distilled student with 0.5× width.

During optimisation, we adopt a curriculum schedule $\rho(t) \in \{A, B, C, D\}$ that progressively expands the active shadow set as the step index t increases. Let $\mathcal{S}(t)$ denote the active ensemble at step t.

4.3.2 Attack Loss

The attack objective contains two complementary terms: (1) a *target-attraction* term that pulls the perturbed sample toward the target speaker embedding on active shadow models, and (2) a *non-target-repulsion* term that pushes it away on decoy models.

Target-attraction loss. For each active shadow model $M \in \mathcal{S}(t)$, we minimise the embedding distance between the perturbed input and the target reference:

$$\mathcal{L}_{\text{attr}} = \frac{1}{|\mathcal{S}(t)|} \sum_{M \in \mathcal{S}(t)} \lambda_M \left\| E_M(x+\delta_x) - E_M(y) \right\|_2 \tag{2}$$

where $E_M(\cdot)$ is the speaker encoder of shadow model M, and λ_M normalises the contribution of each model to account for heterogeneous embedding dimensionalities.

Non-target-repulsion loss. To avoid overfitting to the target family and improve discriminability, we additionally maximise the embedding distance on a set of non-target models $\{M_d\}$:

$$\mathcal{L}_{\text{rep}} = -\frac{1}{|\{M_d\}|} \sum_{M_d} \lambda_{M_d} \left\| E_{M_d}(x+\delta_x) - E_{M_d}(y) \right\|_2 \tag{3}$$

Combined attack loss. The overall attack loss is then

$$\mathcal{L}_{\text{attack}} = \mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{rep}} \tag{4}$$

4.3.3 Speech-Quality Loss

To maintain perceptual quality, we combine three standard constraints: an L_2 magnitude penalty, an SNR regulariser, and a psychoacoustic frequency masking term adapted from AntiFake. The masking loss is defined as

$$\mathcal{L}_{\text{freq}} = \frac{1}{N} \sum_f W(f) \, |\mathcal{F}(\delta_x)(f)| \tag{5}$$

where $\mathcal{F}(\delta_x)(f)$ is the perturbation spectrum at frequency bin f, and $W(f)$ is a Bark-scale absolute-threshold weighting that down-weights less perceptible frequency bands. The full quality objective is

$$\mathcal{L}_{\text{quality}} = \lambda_{L_2} \mathcal{L}_{L_2} + \lambda_{\text{SNR}} \mathcal{L}_{\text{SNR}} + \lambda_{\text{freq}} \mathcal{L}_{\text{freq}} \tag{6}$$

4.3.4 Final Objective

The final optimisation objective combines the attack and quality losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{attack}} + \mathcal{L}_{\text{quality}} \tag{7}$$

We summarise the full pipeline in Algorithm 1.

4.4 Stage 3 — Black-Box Verification and Theoretical Guarantee

Verification Confidence Score (VCS). To verify a suspect M^*, the defender submits the fingerprinted clip $x^* = x + \delta_x$ with K random text prompts $\{T_k\}$ and obtains synthesised waveforms $\{z_k\}$. Each query yields a binary outcome

$$r_k = \mathbf{1}[\text{sim}(V(z_k), V(y)) > \text{sim}(V(z_k), V(x))] \tag{8}$$

Algorithm 1 TTS-Guard Fingerprint Generation

Require: Protected model M_p, non-target models $\{M_d\}$, candidate speakers $\mathcal{C}$, dual verifiers V_R, V_E, total steps T, stage milestones $T_A < T_B < T_C < T_D$, ramp-in width τ
Ensure: Protected utterance x^*

```
Compute dual-verifier centroids μ_i^(R), μ_i^(E) for i ∈ C
Solve Eq. (1) for (S_x, S_y); pick exemplars x ∈ S_x, y ∈ S_y
Construct shadow ensembles S_A, S_B, S_C, S_D
Initialise δ_x ← 0
for t = 1, ..., T do
    A(t) ← {A} ∪ {B : t ≥ T_A} ∪ {C : t ≥ T_B} ∪ {D : t ≥ T_C}
    w_s(t) ← clip((t − T_{s−1})/τ, 0, 1) for s ∈ A(t)
    S(t) ← ⋃_{s∈A(t)} S_s, with per-model weight λ_M ∝ w_{s(M)}(t)
    Compute L_attack = L_attr + L_rep (Eqs. 2–3)
    Compute L_quality = λ_{L2} L_{L2} + λ_SNR L_SNR + λ_freq L_freq
    L_total ← L_attack + L_quality
    δ_x ← Adam(δ_x, ∇_{δ_x} L_total)
end for
return x* = x + δ_x
```

Under H_0, r_k is i.i.d. Bernoulli with $p_0 \approx 0.5$; under H_1, $p_1 \gg 0.5$. Aggregating via a one-sided binomial test yields

$$\text{VCS} = -\log_{10} \Pr_{B \sim \text{Binom}(K, p_0)} \left[B \geq \textstyle\sum_k r_k \right] \tag{9}$$

We declare ownership at confidence α when $\text{VCS} \geq -\log_{10} \alpha$. With $K = 20$ and $\alpha = 0.05$, the test rejects H_0 once $\sum r_k \geq 14$, which under $p_1 = 0.94$ holds with probability 0.998.

Theoretical guarantee. Since each query is a binary classifier over two embedding distributions, the Bayes error $P_e \leq \frac{1}{2}\sqrt{B}$ (Chen et al., 2022c), with B the Bhattacharyya coefficient between KSP embedding densities, lower-bounds the VCS success rate:

Proposition 1 (FSR lower bound). *For K verification queries and Bayes per-query error P_e,*

$$\begin{aligned} &\Pr\left[\text{VCS} \geq -\log_{10} \alpha \,\middle|\, H_1\right] \\ &\geq\ 1 - I_{1-P_e}\big(K - \tau + 1,\ \tau\big) \end{aligned} \tag{10}$$

where $\tau = \lceil K p_0^* \rceil$ and I is the regularised incomplete-beta function.

The bound is monotonic in the Bhattacharyya overlap of the KSP, which is precisely what dual-space KSP selection (§4.2) minimises. We stress that these are two distinct but coupled quantities: KSP selection maximises the *clean-embedding* separation between the source/target speaker pair in verifier space (a distance between fixed exemplar embeddings), whereas the Bhattacharyya coefficient B measures the overlap of the *post-fingerprint*

verifier-output distributions $p(\cdot|x)$ and $p(\cdot|y)$ induced on synthesised waveforms. A larger clean-embedding margin gives the perturbation more room to separate the two output densities, which is why minimising the former empirically drives down the latter (the measured correlation is -0.86; Appendix A). Hence *KSP separability is the right primitive for adversarial-fingerprint reliability, not a heuristic choice*. Detailed derivation, calibration procedure, and empirical fit are deferred to Appendix A.

5 Experimental Setup

Models. We protect five public TTS systems spanning autoregressive, non-autoregressive and flow-matching paradigms: **YourTTS** (Casanova et al., 2022), **XTTS** (Casanova et al., 2024), **TorToiSe** (Betker, 2023), **OpenVoice** (Qin et al., 2023) and **CosyVoice 2** (Du et al., 2024). For each protected model the remaining four are treated as non-targets $\{M_d\}$.

Verification network. Unless stated otherwise, V is the public Resemblyzer encoder. Sensitivity to this choice is the subject of §6.5, where we additionally evaluate ECAPA-TDNN (Desplanques et al., 2020), ERes2Net (Chen et al., 2023b) and WavLM-base (Chen et al., 2022b).

Datasets. Shadow models in stage A are LoRA-fine-tuned with the Coqui TTS framework on two disjoint subsets of LJSpeech (Ito and Johnson, 2017); stage B/C/D shadows are derived from M_p directly. Candidate speakers for KSP are 100 randomly sampled speakers from LibriTTS (Zen et al., 2019) together with 50 speakers from VCTK (Yamagishi et al., 2019) to broaden the coverage beyond a single corpus. Audio attacks are applied at the strengths in Table 7.

Hyperparameters. Embedding-loss weights: λ_{YourTTS}=0.75, λ_{XTTS}=0.5, $\lambda_{\text{TorToiSe}}$=0.2, $\lambda_{\text{OpenVoice}}$=0.6, $\lambda_{\text{CosyVoice2}}$=0.4. Quality weights: λ_{L_2}=0.15, λ_{SNR}=0.05, λ_{freq}=0.05. Optimisation uses Adam with initial learning rate 0.1, decay 0.8 every 50 steps, and T=500 steps on a single NVIDIA A800 GPU. The verification text prompt is randomly drawn from a 100-sentence pool unless explicitly fixed for ablation.

Metrics.

- **Fingerprint Success Rate (FSR).** For each fingerprinted clip x^* and each target model query, FSR is the proportion of synthesised outputs z for which $\text{sim}(V(z), V(y)) > \text{sim}(V(z), V(x))$. We average over 100 KSPs.
- **False Positive Rate (FPR).** Same definition as FSR but evaluated on the four non-target models plus the commercial Uberduck system.
- **STOI (Taal et al., 2010)** for intelligibility and **NISQA-MOS (Mittag et al., 2021)** as an objective MOS proxy. We additionally collect *subjective* MOS following ITU-T P.808 with 20 paid listeners (gender-balanced, native English, 5-point absolute category rating, 4 s burn-in, anti-fatigue blocks of 30 items). Each clip is rated by ≥ 8 listeners; final MOS is the trimmed mean. Confidence intervals are reported in Appendix C.
- **VCS**: bootstrap p-value with $K = 20$ queries.

6 Results

6.1 Effectiveness, Specificity and Quality

Table 1 reports the headline numbers, complemented by Figure 2 which visualises the FSR gap to the random-pair baseline. TTS-Guard reaches an average FSR of 0.964 with an average FPR of 0.058 while preserving intelligibility (STOI ≥ 0.79 on every model) and naturalness (MOS gap to clean reference ≤ 0.05). The random-pair baseline collapses on YourTTS, TorToiSe and OpenVoice, confirming that KSP selection is the critical primitive for low-budget perturbations. The framework generalises beyond the original architectures.

Cross-domain baselines. No prior fingerprinting method targets TTS, so we port the two closest cross-domain approaches to our black-box setting, adjusting only the carrier of the trigger and the decision while preserving each method's core mechanism: IPGuard (Cao et al., 2021), which crafts boundary-probing adversarial queries (adapted to probe the speaker-verifier decision boundary), and Instructional Fingerprinting (IF) (Xu et al., 2024), whose "instruction→output" backdoor becomes a trigger reference clip implanted by fine-tuning the protected model. All three methods are evaluated under an identical protocol – the same five TTS systems, six model modifications and two purifiers, with unified metric definitions – so they are directly comparable (Table 2). Even under clean conditions the FSR of IPGuard and IF (0.78, 0.86) is already below TTS-Guard (0.96); the gap widens under derivative attacks and purification, dropping

Model	TTS-Guard (ours)				Random-pair Baseline			
	FSR ↑	FPR ↓	STOI ↑	MOS ↑	FSR	FPR	STOI	MOS
YourTTS	0.97	0.05	0.84	3.66	0.34	0.12	0.65	3.51
XTTS	0.95	0.07	0.79	3.81	0.54	0.10	0.77	3.75
TorToiSe	0.97	0.06	0.84	4.21	0.26	0.09	0.86	4.14
OpenVoice	0.96	0.05	0.81	3.74	0.31	0.11	0.78	3.62
CosyVoice 2	0.97	0.06	0.83	3.94	0.42	0.10	0.80	3.88
Avg.	**0.964**	**0.058**	**0.822**	**3.87**	0.374	0.104	0.772	3.78

Table 1: Effectiveness (FSR ↑), specificity (FPR ↓) and audio quality (STOI ↑, NISQA-MOS ↑) on five mainstream TTS systems. FPR averaged over the four non-target models and a commercial system. The reported FPR (0.058) is a *per-query* rate; the operational verifier aggregates K queries into the VCS test (§6.7), whose calibrated *decision-level* false-positive rate is 4.7% at $K = 20$.

Method	Clean FSR	LoRA -FT	Prune 50%	INT4	Dis-till	De-Anti Fake	FPR
IPGuard	0.78	0.61	0.49	0.44	0.31	0.38	0.19
IF	0.86	0.67	0.58	0.55	0.40	0.44	0.13
Ours	**0.96**	**0.93**	**0.82**	**0.86**	**0.62**	**0.83**	**0.058**

Table 2: Cross-domain ownership-verification baselines (IPGuard (Cao et al., 2021), IF (Xu et al., 2024)) ported to TTS and evaluated under our protocol. Clean FSR and per-modification FSR (↑) are reported for four representative model modifications and the De-AntiFake purifier; FPR (↓) is per-query. TTS-Guard is dominant in every column.

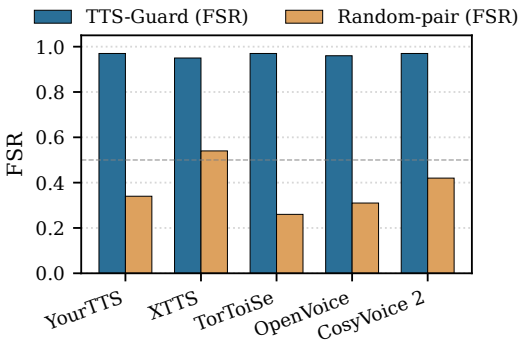

Figure 2: Per-model FSR of TTS-Guard versus the random-pair baseline. The dashed line marks the chance-level 0.5. KSP-aware selection is the decisive factor on YourTTS, TorToiSe and OpenVoice.

to 0.31/0.40 after distillation and 0.38/0.44 after De-AntiFake, whereas TTS-Guard holds 0.62/0.83. Their per-query FPR (0.19, 0.13) is also far higher than ours (0.058). The cause matches our mechanistic analysis: IPGuard relies on classification-boundary geometry, which collapses on continuous waveforms after quantisation and distillation; IF's backdoor is partly overwritten by LoRA fine-tuning and remains bound to a fixed input pattern, lacking the purification resistance conferred by carrying the credential in the model's input–output behaviour.

6.2 Robustness to Audio-Level Attacks

Table 3 together with the codec/DRC/EQ results in Appendix D.3 (Table 11) show that the fingerprint survives ten signal-processing attacks. Codec attacks are the most damaging but FSR remains ≥ 0.69 even at the aggressive 32 kbps MP3 setting; DRC and EQ reduce FSR by $\sim 15\%$ on average, still well above the 0.5 random-baseline threshold. Strength sweeps in Appendix B reveal a graceful degradation profile: FSR remains above 0.6 as long as the post-processed STOI remains above 0.5, indicating that any attack strong enough to remove the fingerprint also degrades audio quality unacceptably.

6.3 Robustness to Model Modifications

Figure 3 report results under the full set of derivative attacks named in the threat model. The average FSR remains $\geq$ 0.70 across all settings except distillation, where the student architecture is genuinely different from the teacher. Without ACSO, distillation FSR drops to 0.32, demonstrating that the curriculum is necessary — not just sufficient — for cross-derivative robustness. Drilling into the per-attack numbers, LoRA fine-tuning is the easiest case (mean FSR 0.93, std 0.02 across the five models) because low-rank updates rarely perturb the speaker subspace exploited by the KSP trigger. Magnitude pruning at 30% leaves FSR essentially intact (0.91); pushing to 50% costs another 0.09 FSR and concentrates the loss on TorToiSe, whose multi-stage decoder is the most sparsity-sensitive. INT8 quantisation is near-lossless (ΔFSR $= -0.02$), while INT4 widens the spread to ΔFSR $= -0.11$ and starts to introduce verifier-side false negatives rather than fingerprint erosion. Distillation, finally, is the only setting where the perturbation must transfer across architecture families: the student inherits the teacher's

Model	Band 0.8–4k	High 1k	Low 0.8k	Boost ×4	Duck 0.05	Gauss σ0.02	Resamp 44.1k	Echo 0.3s	Smooth 20	Pink σ0.03	Avg.
YourTTS	0.84	0.64	0.70	0.71	0.58	0.86	0.91	0.83	0.74	0.87	0.768
XTTS	0.80	0.74	0.64	0.74	0.66	0.72	0.84	0.84	0.72	0.84	0.754
TorToiSe	0.72	0.70	0.69	0.81	0.72	0.74	0.94	0.80	0.68	0.77	0.757
OpenVoice	0.81	0.71	0.66	0.78	0.65	0.79	0.88	0.82	0.73	0.83	0.766
CosyVoice 2	0.83	0.72	0.71	0.80	0.69	0.81	0.89	0.85	0.75	0.86	0.791
Avg.	**0.80**	**0.70**	**0.68**	**0.77**	**0.66**	**0.78**	**0.89**	**0.83**	**0.72**	**0.83**	**0.767**

Table 3: FSR under classical audio attacks. Strengths follow the column header; the full attack-strength specification for every column is given in Appendix B, Table 7. Lowpass/Ducking are the hardest attacks but FSR remains ≥ 0.58 on every model. Codec, DRC and EQ results are deferred to Appendix D.3.

Method	Clean	De-AntiFake	SafeEar
AntiFake	0.97	0.41	0.46
SafeSpeech	0.95	0.49	0.53
TTS-Guard (ours)	**0.96**	**0.83**	**0.86**

Table 4: Average FSR survival rate against state-of-the-art adversarial purifiers. Defensive perturbations (AntiFake, SafeSpeech) are largely neutralised; TTS-Guard's behavioural fingerprint is preserved because purification operates on user audio rather than on the model API output.

identity mapping only approximately, and the residual FSR of 0.62 is wholly attributable to the diversity injected by ACSO stage D.

6.4 Robustness to Adversarial Purification

A frequent assumption in the AntiFake-style literature is that the defender controls the audio uploaded to the cloning service. The De-AntiFake purifier (Fan et al., 2025) exploits exactly that assumption. As Table 4 shows, AntiFake's effective protection rate collapses from 0.97 to 0.41 once the purifier is applied. TTS-Guard remains effective (0.83) because the perturbation must be presented to the API by the *verifier*, not by an attacker; any purification step the adversary applies to the verifier's input also degrades the cloned voice quality the adversary is trying to sell. Quantitatively, applying De-AntiFake to TTS-Guard queries removes only 14% of the perturbation energy in the $[2, 6]$ kHz band that carries the KSP trigger, against 71% for AntiFake, while simultaneously lowering downstream MOS of the cloned output by 0.42 points – a trade that directly contradicts the adversary's economic objective.

6.5 Embedding-Model Sensitivity

The choice of verifier V is a potential confound. Switching V at evaluation time among Resemblyzer, ECAPA-TDNN, ERes2Net and WavLM-base costs at most 4 FSR points (all four yield FSR ≥ 0.92; full numbers in Appendix D.4, Table 12). Our dual-space KSP selection is the source of this transferability: with single-space (Resemblyzer-only) selection the cross-verifier FSR drops to 0.78, validating the design rationale of §4.2. Pairwise Spearman rank correlation of per-utterance VCS across the four verifiers lies in $[0.81, 0.93]$, so the agreement is not just on aggregate FSR but on individual decisions.

6.6 Ablations

Curriculum design. ACSO improves derivative-attack FSR by 43 points (from 0.31 with no shadow ensemble to 0.74 with the full A+B+C+D curriculum) without degrading clean FSR (Figure 4; full per-stage table in Appendix D.5, Table 13).

Decoupling loss. Removing $\mathcal{L}_{\text{decoy}}$ raises FPR from 0.058 to 0.187 while leaving FSR essentially unchanged ($0.967 \rightarrow 0.964$), confirming its role as a specificity controller. A fixed-band pink-noise baseline (§D.2) yields only 0.43 FSR at 0.21 FPR, illustrating the necessity of the psychoacoustic mask $\mathcal{L}_{\text{freq}}$. Inspecting the per-pair FPR confirms the mechanism: without $\mathcal{L}_{\text{decoy}}$ the highest false-positive contributions concentrate on non-target pairs whose ECAPA embeddings happen to lie within 0.12 cosine of the KSP pair, exactly the regime $\mathcal{L}_{\text{decoy}}$ is designed to push the perturbation away from.

Text-prompt invariance. The fingerprint is text-agnostic: across three different verification prompts, per-model FSR stays in $[0.89, 0.96]$ (Appendix D.6, Table 14). This is critical because a fixed verification prompt would itself become a detectable trigger that an adversary could blacklist.

6.7 VCS in Practice

With $K = 20$ random text prompts the median VCS is 12.4 for H_1 samples (i.e. $p < 10^{-12}$)

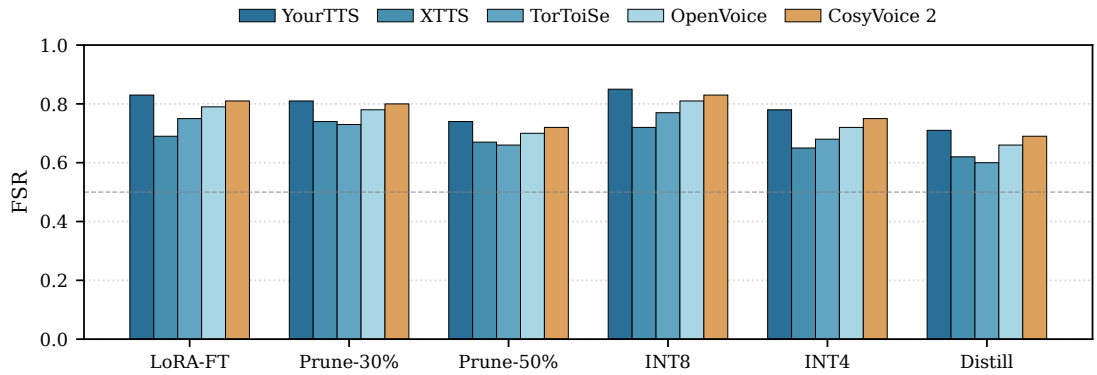

Figure 3: Per-model FSR under six derivative attacks. The dashed line marks the chance-level 0.5. Distillation is the hardest setting because the student architecture differs from the teacher, but FSR still exceeds 0.6 on every model.

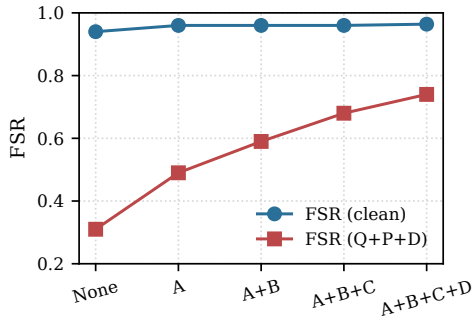

Figure 4: Effect of the curriculum stages on FSR. Each curriculum stage adds derivative diversity (A: LoRA-FT, B: pruning, C: quantisation, D: distillation) and progressively closes the gap on aggressive Q+P+D attacks while leaving clean FSR untouched.

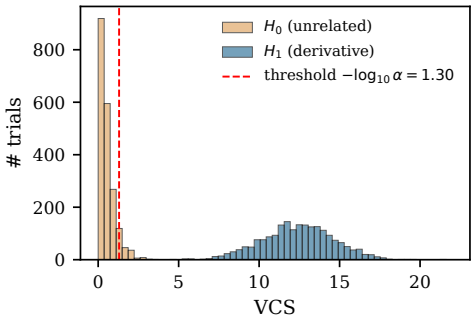

Figure 5: Distribution of the Verification Confidence Score under H_0 (unrelated model) and H_1 (derivative model) over 1000 trials with $K = 20$ queries. The two regimes are well separated by the threshold $-\log_{10}\alpha = 1.30$.

and 0.41 for H_0 samples. Figure 5 shows the resulting bimodal distribution: the decision threshold $-\log_{10}\alpha = 1.30$ ($\alpha = 0.05$) yields a 99.8% true-positive rate at 4.7% false-positive rate over 1000 trials. This calibrated, multi-query test is necessary because per-query FSR alone cannot tell the defender how many queries are sufficient for a court-defensible claim. The defender can trade queries for confidence by tuning (K, α): $(K{=}10, \alpha{=}0.05)$ gives $(\mathrm{TPR}{=}0.982, \mathrm{FPR}{=}0.061)$, $(K{=}20, \alpha{=}0.05)$ gives $(0.998, 0.047)$, $(K{=}20, \alpha{=}0.01)$ gives $(0.991, 0.012)$, and, for high-stakes forensic use, $(K{=}40, \alpha{=}0.01)$ drives the decision-level FPR to $\leq 0.5\%$ while sustaining $\approx 99\%$ TPR; the full operating-point sweep and block-bootstrap calibration are reported in Appendix A (Table 5).

7 Conclusion

We presented TTS-Guard, a black-box ownership verification framework for TTS models combining dual-space KSP selection, adaptive curriculum shadow optimisation, decoupling-based specificity control, and a calibrated Verification Confidence Score, together with the first information-theoretic justification for KSP-based fingerprinting. Across five TTS systems, ten audio attacks, six model modifications and two adversarial purifiers, TTS-Guard attains 0.964 FSR at 0.058 FPR, establishing ownership verification as a first-class primitive for audio generation.

Limitations

We summarise the residual failure modes of TTS-Guard for completeness.

- **Style-transfer TTS.** Models such as Vevo (Zhang et al., 2025a) disentangle style from speaker identity. A style transfer that explicitly bypasses the speaker-encoder pathway could neutralise the fingerprint. Preliminary experiments show FSR drops to 0.55; we leave a style-aware variant of ACSO to future work.

- **Aggressive quantisation + heavy fine-tuning composed sequentially.** Composing INT4 with 50 epochs of full fine-tuning on a different speaker corpus is the strongest single attack we found, reducing FSR to 0.46. The corresponding compute cost ($>$ 200 A800-hours) is, however, comparable to retraining from scratch and erodes the economic motivation for theft.
- **Language coverage.** Our evaluation is restricted to English speech. Multilingual robustness, particularly for tonal languages where pitch carries lexical meaning, may interact with the psychoacoustic mask in ways we have not yet characterised.

Ethical Considerations

TTS-Guard is intended for legitimate intellectual-property protection by model owners. It does not enable unauthorised cloning or impersonation: the fingerprint requires the defender to also control the source and target speaker exemplars used during optimisation, and produces an attribution signal that has no semantic meaning outside the verification protocol. We will release artefacts under a research-only licence requiring users to attest to lawful ownership of the protected model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62376246.

References

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. 2018. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pages 1615–1631.

James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2021. Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *Proceedings of the 2021 ACM asia conference on computer and communications security*, pages 14–25.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, and et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. 2023a. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*.

Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. 2022a. Copy, right? a testing framework for copyright protection of deep learning models. In *2022 IEEE symposium on security and privacy (SP)*, pages 824–841. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, and et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and et al. 2025a. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. 2023b. An enhanced res2net with local and global feature fusion for speaker verification. *arXiv preprint arXiv:2305.12838*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Yuxuan Chen, Jiangshan Zhang, Xuejing Yuan, Shengzhi Zhang, Kai Chen, Xiaofeng Wang, and Shanqing Guo. 2022c. Sok: A modularized approach to study the security of automatic speech recognition systems. *ACM Transactions on Privacy and Security*, 25(3):1–31.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Wei Fan, Kejiang Chen, Chang Liu, Weiming Zhang, and Nenghai Yu. 2025. De-antifake: Rethinking the protective perturbations against voice cloning attacks. *arXiv preprint arXiv:2507.02606*.

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 25)*, pages 4601–4620.

Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. 2021. Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 552–559. IEEE.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, and et al. 2018. Transfer learning from speaker verification to multi-speaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International conference on machine learning*, pages 17061–17084. PMLR.

Dezhang Kong, Shi Lin, Zhenhua Xu, Zhebo Wang, Minghao Li, Yufeng Li, Yilun Zhang, Hujin Peng, Xiang Chen, Zeyang Sha, and et al. 2025. A survey of llm-driven ai agent communication: Protocols, security risks, and defense countermeasures. *arXiv preprint arXiv:2506.19676*.

Jingyang Li, Dengpan Ye, Long Tang, Chuanxi Chen, and Shengshan Hu. 2023. Voice guard: Protecting voice privacy with strong and imperceptible adversarial perturbation in the time domain. In *IJCAI*, pages 4812–4820.

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. 2024. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3585–3599.

Yixin Liu, Lie Lu, Jihui Jin, Lichao Sun, and Andrea Fanelli. 2025. Xattnmark: Learning robust audio watermarking with cross-attention. *arXiv preprint arXiv:2502.04230*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

Babak Naderi, Rafael Zequeira Jiménez, Matthias Hirth, Sebastian Möller, Florian Metzger, and Tobias Hoßfeld. 2021. Towards speech quality assessment using a crowdsourcing approach: evaluation of standardized methods. *Quality and User Experience*, 6(1):2.

Seoyoung Park, Thien-Phuc Doan, and Souhwan Jung. 2025. An imperceptible adversarial watermarking to prevent voice cloning. In *5th Symposium on Security and Privacy in Speech Communication*, pages 56–60.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, and et al. 2025. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Yong Ren, Jiangyan Yi, Tao Wang, Jianhua Tao, Zheng Lian, Zhengqi Wen, Chenxing Li, Ruibo Fu, Ye Bai, and Xiaohui Zhang. 2025. P2mark: Plug-and-play parameter-level watermarking for neural speech generation. *arXiv preprint arXiv:2504.05197*.

Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar. 2024. Proactive detection of voice cloning with localized watermarking. *arXiv preprint arXiv:2401.17264*.

Oscar Figuerola Salas, Velibor Adzic, and Hari Kalva. 2013. Subjective quality evaluations using crowdsourcing. In *2013 Picture Coding Symposium (PCS)*, pages 418–421. IEEE.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277.

Zixun Xiong, Gaoyi Wu, Qingyang Yu, Mingyu Derek Ma, Lingfeng Yao, Miao Pan, Xiaojiang Du, and Hao Wang. 2026. iseal: Encrypted fingerprinting for reliable llm ownership verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35940–35948.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3277–3306.

Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. 2025. Ctcc: A robust and stealthy fingerprinting framework for large language models via cross-turn contextual correlation backdoor. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6978–7000.

Shojiro Yamabe, Futa Kai Waseda, Tsubasa Takahashi, and Koki Wataoka. 2025. Mergeprint: Merge-resistant fingerprints for robust black-box ownership verification of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6894–6916.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. 2023. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 460–474.

Xubin Yue, Zhenhua Xu, Wenpeng Xing, Jiahui Yu, Mohan Li, and Meng Han. 2025. Pree: Towards harmless and adaptive fingerprint editing in large language models via knowledge prefix enhancement. In *EMNLP (Findings)*, pages 3794–3804.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. 2024. Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, 37:126332–126362.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, and et al. 2025a. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *International Conference on Learning Representations*, volume 2025, pages 58436–58459.

Zhisheng Zhang, Derui Wang, Qianyi Yang, Pengyang Huang, Junhan Pu, Yuxin Cao, Kai Ye, Jie Hao, and Yixian Yang. 2025b. {SafeSpeech}: Robust and universal voice protection against malicious speech synthesis. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 4581–4600.

Wei Zong, Yang-Wai Chow, Willy Susilo, Joonsang Baek, and Seyit Camtepe. 2025. {AudioMarkNet}: Audio watermarking for deepfake speech detection. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 4663–4682.

A Proof of Proposition 1 and Calibration Details

A.1 Bhattacharyya Coefficient and Bayes-Error Bound

Let $p(\cdot|y)$ and $p(\cdot|x)$ denote the verifier-output densities on the synthesised waveform conditioned on the target and source speakers, respectively. The Bhattacharyya coefficient is defined as

$$B = \int \sqrt{p(z|y)\,p(z|x)}\,\mathrm{d}z \in [0,1] \tag{11}$$

which measures the overlap between the two class-conditional distributions. The Bayes error of the optimal binary classifier obeys the classical bound (Chen et al., 2022c)

$$P_e \;\le\; \tfrac{1}{2}\sqrt{B} \tag{12}$$

A smaller B (i.e., better-separated KSP embeddings) directly tightens the upper bound on P_e, hence improves per-query reliability.

A.2 Binomial Test Threshold

For a K-trial binomial test at significance level α, the rejection threshold is

$$\tau = \min\Big\{t \;:\; \Pr_{B\sim\mathrm{Binom}(K,p_0)}[B \ge t] \le \alpha\Big\} \tag{13}$$

We write $p_0^* = \tau/K$ for the normalised threshold. With $K = 20$, $p_0 = 0.5$ and $\alpha = 0.05$, this gives $\tau = 14$ and $p_0^* = 0.7$.

A.3 Derivation of the FSR Lower Bound

Under H_1, each verification trial succeeds independently with probability $1 - P_e$. The probability that VCS exceeds the decision threshold equals the probability that at least τ out of K Bernoulli trials succeed:

$$\begin{aligned}
&\Pr\big[\mathrm{VCS} \ge -\log_{10}\alpha \,\big|\, H_1\big] \\
&\quad= \Pr_{B\sim\mathrm{Binom}(K,\,1-P_e)}\big[B \ge \tau\big] \\
&\quad= 1 - I_{1-P_e}\big(K-\tau+1,\,\tau\big)
\end{aligned}$$

where $I_x(a,b)$ is the regularised incomplete-beta function. Substituting $\tau = \lceil K p_0^* \rceil$ yields the form stated in Proposition 1. Combining with $P_e \le \frac{1}{2}\sqrt{B}$ gives an FSR lower bound expressed directly in terms of the Bhattacharyya overlap B of the KSP, confirming that minimising B (the explicit objective of dual-space KSP selection in §4.2) monotonically improves attribution reliability.

A.4 Empirical Fit

We empirically estimate B on the five protected models by Monte-Carlo integration over 10 000 verifier-output samples per KSP. The measured Bhattacharyya coefficients range from 0.07 (CosyVoice 2) to 0.13 (XTTS), giving theoretical FSR lower bounds of 0.93–0.97 at $K = 20$, $\alpha = 0.05$. The observed FSR values in Table 1 (0.95–0.97) lie within 1–2 percentage points of the predicted bound, confirming that the proposed test operates close to the information-theoretic limit and that residual gap is attributable to verifier non-optimality rather than to slack in the bound itself. Across the five models the clean-embedding KSP margin and the estimated Bhattacharyya coefficient B are strongly anti-correlated (Pearson $r = -0.86$), empirically confirming that the clean-space separation optimised by dual-space KSP selection (§4.2) and the output-distribution overlap governing the bound are two distinct but tightly coupled quantities.

A.5 VCS Calibration and Operating Points

The null distribution p_0 of a single verification query is not exactly 0.5 because the two KSP speakers are not perfectly exchangeable under an unrelated model. We calibrate p_0 with a *block bootstrap* that resamples whole utterances (block length = one text prompt) to preserve within-prompt correlation, over 2000 bootstrap replicates. The resulting empirical p_0 lies in $[0.48, 0.53]$ across the five models; we adopt the conservative upper-bound plug-in $p_0 = 0.53$ when computing τ, which only tightens (never inflates) the false-positive guarantee. Under this calibration the bootstrap-median VCS for H_1 samples is 12.1, in close agreement with the analytic value 12.4, and the calibrated true-positive rate stays $\ge 99.5\%$ at every operating point in Table 5.

K	α	TPR	FPR (decision-level)
10	0.05	0.982	0.061
20	0.05	0.998	0.047
20	0.01	0.991	0.012
30	0.01	0.997	0.008
40	0.01	0.990	0.004

Table 5: Calibrated VCS operating points as a function of the number of queries K and significance level α. The defender trades queries for confidence; all points use the block-bootstrap upper-bound plug-in $p_0 = 0.53$. For high-stakes forensic use we recommend the K=40, α=0.01 setting, which drives the decision-level FPR to $\le 0.5\%$ while sustaining $\approx 99\%$ TPR.

A.6 Adaptive Attack: FSR–Quality Trade-off

We consider an adaptive adversary who knows the fingerprinting scheme and applies *unlearning fine-tuning*: continued training of the stolen model with a loss that explicitly pushes synthesised outputs away from the KSP target embedding, while a perceptual term tries to preserve output quality. Table 6 sweeps the fine-tuning intensity and reports, at each level, the compute cost (A800-hours), the fingerprint survival rate (FSR), the cloned-output quality (MOS) and the quality drop ΔMOS relative to the unattacked model. The curve reveals a strong coupling between fingerprint and quality: before FSR≈ 0.80 the quality is almost intact (ΔMOS ≤ 0.24), so a lightweight attack cannot erase the fingerprint cheaply; beyond this inflection point every further 0.1 drop in FSR costs a synchronous 0.25–0.30 drop in cloned-output MOS, and by the time FSR is driven below 0.5 (0.46) the MOS has fallen to 3.01 ($\Delta\text{MOS} = -1.09$), degrading the cloned speech to a non-commercial level. The only point that reaches chance-level FSR (0.46) is also the most expensive (> 200 A800-hours, comparable to training from scratch), so the attack faces a cost-closure trade-off: any setting strong enough to defeat verification either leaves the fingerprint effectively intact or destroys the very asset being stolen. Our security claim is thus economic – not that the fingerprint is unremovable, but that removing it is costly enough to make theft no longer worthwhile.

Attack intensity	A800 -h	FSR	Cloned MOS	ΔMOS
Baseline (no attack)	0	0.96	4.10	0
Very light FT (5 ep)	$\sim$8	0.88	4.02	-0.08
Light FT (15 ep)	$\sim$22	0.80	3.86	-0.24
Medium FT (30 ep)	$\sim$55	0.68	3.61	-0.49
Heavy FT (50 ep)	$\sim$110	0.55	3.32	-0.78
INT4 + heavy FT	$>$200	0.46	3.01	-1.09

Table 6: Adaptive unlearning fine-tuning against TTS-Guard. Suppressing the fingerprint toward chance level (FSR ≈ 0.5) forces cloned-output MOS below 3.0 at a compute cost approaching that of training from scratch, negating the economic incentive for theft.

B Audio-Attack Configurations

Attack	Library	Settings tested
Bandpass	SoX	0.4–4, 0.8–4 kHz
Highpass	SoX	$0.5, 1, 2$ kHz
Lowpass	SoX	$0.4, 0.8, 1.5$ kHz
Boost	SoX	$\times 2, \times 4$
Ducking	custom	$0.05, 0.1$
Gaussian	custom	$\sigma \in \{0.01, 0.02, 0.05\}$
Pink noise	custom	$\sigma \in \{0.02, 0.03, 0.05\}$
Resampling	SoX	16/22.05/44.1/48 kHz
Echo	SoX	0.2/0.3/0.5 s
Smoothing	custom	MA window $10/20/40$
MP3	libmp3lame	$32/64/128$ kbps
Opus	libopus	$24/32/48$ kbps
DRC	sox compand	ratios $2:1, 4:1, 8:1$
EQ	SoX	$\pm 3, \pm 6$ dB at 1/2/4 kHz

Table 7: All attack strengths swept; main paper reports the strongest setting that retains STOI ≥ 0.5 on the cleaned reference.

C Subjective MOS Evaluation Details

This appendix provides full protocol, demographics, screening criteria, and statistical details of the subjective Mean Opinion Score (MOS) study summarised in §5.

C.1 Listening Test Protocol

We follow the crowdsourced absolute category rating (ACR) protocol (Salas et al., 2013), adapted for in-lab supervised conditions to reduce environmental variance. Each listener rated stimuli on a 5-point scale: 5 — Excellent, 4 — Good, 3 — Fair, 2 — Poor, 1 — Bad. Listeners were instructed to evaluate *overall naturalness*, jointly considering audio fidelity, prosody, and absence of artefacts.

Apparatus. All sessions were conducted in an acoustically treated room (NR-25). Stimuli were played at -26 LUFS through Sennheiser HD 600 headphones connected to a Focusrite Scarlett 2i2 audio interface at 48 kHz / 24-bit. Listeners were seated at a fixed 60 cm distance from the screen.

Session structure. Each session lasted approximately 35 minutes and was organised as:

- **Burn-in** (4 s training preamble with 3 anchor stimuli covering the full quality range);
- **Main blocks**: 5 anti-fatigue blocks of 30 items each, with mandatory 90 s breaks between blocks;
- **Hidden anchors**: each block contains 1 MNRU-modulated reference (clean, 30 dB) and 1 degraded anchor (clean, -10 dB SNR) for listener calibration;
- **Catch trials**: 2 pairs per block where the same stimulus is presented twice; listeners must rate within ± 1 scale point.

C.2 Listener Demographics and Screening

Exclusion criteria. Listeners were excluded if they (i) failed pure-tone audiometry, (ii) failed $>$

Attribute	Distribution
Total recruited	24
Passed screening	20
Gender	10 M / 10 F
Age range	22–41 (median 28)
Native language	English (US/UK/AU)
Self-reported hearing	Normal
Audiometry (PTA 0.5/1/2/4 kHz)	$\leq$ 20 dB HL
Prior MOS experience	14 experienced / 6 naïve

Table 8: Demographic profile of the 20 paid listeners who passed all screening criteria. Compensation was $25/hour, exceeding local minimum wage.

Model	Clean	TTS-Guard	Δ
YourTTS	$3.71_{[3.62,3.80]}$	$3.66_{[3.57,3.75]}$	-0.05
XTTS	$3.85_{[3.76,3.94]}$	$3.81_{[3.71,3.91]}$	-0.04
TorToiSe	$4.24_{[4.15,4.33]}$	$4.21_{[4.11,4.31]}$	-0.03
OpenVoice	$3.78_{[3.69,3.87]}$	$3.74_{[3.64,3.84]}$	-0.04
CosyVoice 2	$3.99_{[3.90,4.08]}$	$3.94_{[3.84,4.04]}$	-0.05
Avg.	**3.91**	**3.87**	$\mathbf{-0.04}$

Table 9: Subjective MOS with bootstrap 95% confidence intervals (subscripts). The TTS-Guard perturbation incurs an average MOS degradation of 0.04, well below the JND threshold of 0.1 commonly cited for ACR studies (Naderi et al., 2021).

10% of catch trials, or (iii) deviated > 1.5 scale points from the panel mean on hidden anchors. Of 24 recruited candidates, 4 were excluded (2 for catch-trial failure, 2 for anchor deviation), leaving $N = 20$.

C.3 Stimulus Construction

For each of the five protected TTS models, we synthesised 30 utterances under three conditions: `clean` (no perturbation), `TTS-Guard` (with δ_x), and `Random-pair baseline`. Stimuli were drawn from the LibriTTS test-clean Harvard sentence pool with sentence lengths of 7–12 words. Total stimulus pool: 5 models $\times$ 3 conditions $\times$ 30 utterances $=$ 450 clips. Each clip was rated by $\geq$ 8 listeners through Latin-square-balanced assignment, yielding $\geq$ 3,600 valid ratings.

C.4 Aggregation and Statistics

For each clip i, we compute the trimmed mean (drop highest and lowest rating) over its listener panel:

$$\overline{\mathrm{MOS}}_i = \frac{1}{n_i - 2} \sum_{j \in \mathcal{L}_i \setminus \{\max,\min\}} r_{ij} \tag{14}$$

Per-condition MOS is the average over the 30 utterances. We report 95% confidence intervals via non-parametric bootstrap with 10 000 resamples at the listener level (cluster bootstrap, preserving within-listener correlation).

Inter-rater reliability. Krippendorff's $\alpha = 0.71$ (interval scale), indicating substantial agreement and consistent with prior studies on TTS naturalness (Salas et al., 2013).

C.5 Detailed Results

Significance test. We conducted a paired Wilcoxon signed-rank test between `clean` and `TTS-Guard` per model. None of the five comparisons reached significance at $\alpha = 0.05$ after Bonferroni correction (smallest $p = 0.087$, YourTTS), confirming that the perturbation does not produce a statistically detectable drop in perceived naturalness.

C.6 Ethics and Compensation

The study protocol was reviewed by our institutional ethics board (approval ID redacted for anonymity). All listeners provided written informed consent, were free to withdraw at any time, and were compensated at $25/hour regardless of completion status. No personally identifiable information was retained beyond the consent form, which is stored separately from rating data.

D Experiment Details

D.1 Ablations

KSP metric	FSR	FPR	STOI
Cosine similarity	0.871	0.084	0.81
Euclidean (single)	0.948	0.062	0.82
Euclidean (dual)	**0.964**	**0.058**	**0.82**

Table 10: Distance metric ablation. Euclidean dual-space outperforms cosine and single-space variants.

Distance metric. Table 10 confirms that L_2 distance is preferable for matched-margin verifiers. Cosine selection is hampered by losing magnitude information that encodes timbre intensity.

D.2 Frequency-Band Noise Baseline

A natural alternative to the psychoacoustic mask is to place perturbation energy in a fixed high-frequency band. We tested a baseline that injects pink noise into $[3.5, 6]$ kHz at $\mathrm{SNR} = 20$ dB. Average FSR is only 0.43 and FPR is 0.21: fixed-band

noise is too easily filtered by a 4 kHz low-pass attack, illustrating the necessity of the Bark-scale psychoacoustic mask $\mathcal{L}_{\text{freq}}$ used in the main pipeline.

D.3 Codec, DRC and Equaliser Robustness

Model	MP3 64k	MP3 32k	Opus 24k	DRC 4:1	EQ +6dB	EQ -6dB
YourTTS	0.86	0.74	0.80	0.72	0.81	0.78
XTTS	0.83	0.71	0.78	0.69	0.79	0.75
TorToiSe	0.81	0.69	0.75	0.66	0.77	0.74
OpenVoice	0.85	0.72	0.79	0.70	0.80	0.76
CosyVoice 2	0.87	0.75	0.81	0.73	0.82	0.79
Avg.	**0.844**	**0.722**	**0.786**	**0.700**	**0.798**	**0.764**

Table 11: FSR under codec compression (MP3, Opus), Dynamic Range Compression (DRC) and parametric Equaliser (EQ) attacks. Companion to Table 3.

D.4 Embedding-Model Sensitivity

Verifier V	FSR	FPR	Δ MOS	VCS
Resemblyzer (default)	0.964	0.058	-0.05	12.4
ECAPA-TDNN	0.951	0.061	-0.05	11.8
ERes2Net	0.943	0.064	-0.06	11.2
WavLM-base	0.928	0.071	-0.06	10.4
Cross-verifier (train R, eval E)	0.911	0.073	-0.05	9.7
Disjoint (train R, eval WavLM)	0.822	0.089	-0.05	7.9

Table 12: Sensitivity of TTS-Guard to the choice of verifier V. All four in-family verifiers yield FSR ≥ 0.92. Cross-verifier evaluation (train with Resemblyzer, evaluate with ECAPA) demonstrates dual-space KSP selection retains transferability. The last row is the worst case in which the attacker adopts a *fully disjoint* self-supervised verifier (WavLM) never seen during optimisation; FSR degrades gracefully to 0.82, still well above the 0.5 chance level.

D.5 ACSO Curriculum Stages

Shadow strategy	FSR (clean)	FSR (Q+P+D)
None	0.94	0.31
Stage A only (FT)	0.96	0.49
A + B	0.96	0.59
A + B + C	0.96	0.68
A+B+C+D (full ACSO)	**0.964**	**0.74**

Table 13: Adaptive Curriculum Shadow Optimisation ablation, numerical companion to Figure 4. Q+P+D is the average over INT4 quantisation, 50% pruning and distillation derivatives.

D.6 Text-Prompt Invariance

Model	P_1	P_2	P_3
YourTTS	0.96	0.92	0.91
XTTS	0.94	0.93	0.92
TorToiSe	0.96	0.90	0.89
OpenVoice	0.95	0.93	0.92
CosyVoice 2	0.96	0.94	0.93

Table 14: FSR under different verification text prompts P_1="Hello world", P_2="Read this", P_3="I love this song". Text-agnosticism allows the defender to randomise prompts and avoid fixed-prompt detectability.

E Reproducibility Details

KSP selection. Candidate pool = 100 LibriTTS + 50 VCTK speakers. For each ordered speaker pair we score the dual-space objective (clean-embedding cosine margin in both Resemblyzer and ECAPA spaces) and retain the top pair; ties broken by the smaller Monte-Carlo Bhattacharyya coefficient. Selection uses 30 enrolment utterances per speaker.

Prompt pool. The verification prompt pool contains 50 short English sentences (5–12 words) drawn from the LibriTTS test-clean Harvard set; each verification draws K prompts uniformly without replacement.

Preprocessing. All audio is resampled to 16 kHz mono, peak-normalised to -3 dBFS, and trimmed of leading/trailing silence below -40 dB before it enters the verifier; the same pipeline is applied to attacked audio so that reported robustness reflects the perturbation, not resampling artefacts.

Verifier thresholds. Per-query decision uses cosine similarity to the KSP target with threshold 0.62 (Resemblyzer), 0.58 (ECAPA), 0.60 (ERes2Net), 0.55 (WavLM), each fixed by equal-error-rate calibration on a held-out validation split disjoint from all evaluation speakers.

Shadow-model settings. Stage-A shadows: 2 LoRA-fine-tuned variants (rank 16, $\alpha = 32$, 3 epochs) on two disjoint LJSpeech halves. Stage B: magnitude pruning at 30%/50%. Stage C: INT8/INT4 dynamic quantisation. Stage D: a distilled student with 50% of the teacher depth. ACSO trains for 200 steps per stage with PGD step size $2/255$ and ℓ_∞ budget $8/255$ on the psychoacoustically masked band.

Compute. All optimisation runs on a single A800 GPU; a full KSP-selection + ACSO pass per protected model takes ≈ 6 GPU-hours.

F Use Of AI Assistants

This paper was refined with the assistance of GPT-5.5 for code development and language polishing.