FakeMark: Deepfake Speech Attribution With Watermarked Artifacts

Wanying Ge; Xin Wang; Junichi Yamagishi
National Institute of Informatics, Tokyo, Japan

CVPR JUNE 3-7, 2026 DENVER COLORADO

TL;DR

We propose FakeMark, a watermarking method that injects system-specific artifacts instead of predefined bitstrings to attribute deepfake speech.

FakeMark remains robust under severe distortions like neural codecs and cross-dataset domain shifts.

Motivation

Accountability: Attributing deepfake speech to its source system is essential for identifying the generative models used in synthetic audio generation.

Generalization Gap: Existing classifier-based solutions often fail to generalize to samples from unseen domains or datasets.

Watermark Vulnerability: Conventional bitstring-based watermarking can be easily compromised by common distortions like codec compression or malicious removal attacks.

Proposed Solution: FakeMark aligns the watermark message with the learned artifact patterns of specific systems. This "dual-cue" design ensures attribution remains effective even if one cue (injected or intrinsic) is elusive or removed.

Contact:
Wanying Ge
gewanying@nii.ac.jp

Proposed framework

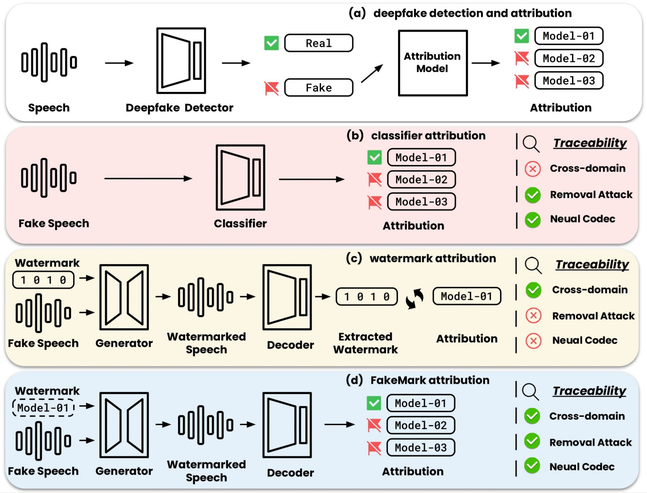

Figure 1. Overview of (a) deepfake detection and attribution, (b) classifier-based attribution, (c) watermarking-based attribution, and (d) proposed FakeMark.

Training pipeline of FakeMark

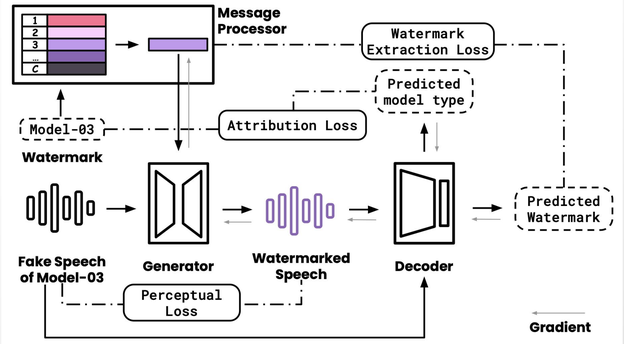

Figure 2. FakeMark is optimized using a multi-objective training strategy to ensure both attribution accuracy and audio quality: **Attribution Loss:** Guides the decoder to accurately identify the deepfake source system by jointly leveraging intrinsic deepfake artifacts and the injected watermark. **Watermark Extraction Loss:** Ensures the system-specific watermark message can be reliably recovered from the watermarked signal, even after it has undergone various channel distortions. **Perceptual Loss:** Utilizes multi-resolution STFT and GAN-based adversarial losses to maintain high acoustic similarity to the original signal and ensure the watermark remains imperceptible.

Attribution accuracy results cross distortions and attacks

In-domain
MLAAD v5 test set (Table 1)

	PROPOSED					
	FakeMark^A	FakeMark^T	AudioSeal	Timbre	MMS-300M	ResNet34
input →	waveform	spectrogram	waveform	spectrogram	waveform	spectrogram
None (clean)	1.00	1.00	1.00	1.00	1.00	0.97
CODEC						
SpeechTokenizer	0.85	0.99	0.10	0.94	0.92	0.88
FACodec	0.91	0.85	0.17	0.82	0.92	0.79
WavTokenizer	0.33	0.47	0.09	0.19	0.39	0.71
VOCODER						
HiFi-GAN	0.91	1.00	0.09	1.00	0.94	0.92
Vocos	0.98	1.00	0.12	1.00	0.98	0.97
BigVGAN	0.99	1.00	0.28	1.00	1.00	0.97

Cross-dataset
ASVspoof5 + TIMIT-TTS (Table 2)

	PROPOSED					
	FakeMark^A	FakeMark^T	AudioSeal	Timbre	MMS-300M	ResNet34
input →	waveform	spectrogram	waveform	spectrogram	waveform	spectrogram
None (clean)	1.00	1.00	1.00	1.00	0.07	0.12
CODEC						
SpeechTokenizer	0.58	0.94	0.07	0.90	0.07	0.10
FACodec	0.87	0.32	0.08	0.85	0.08	0.05
WavTokenizer	0.06	0.02	0.03	0.21	0.07	0.07
VOCODER						
HiFi-GAN	0.88	1.00	0.08	1.00	0.07	0.11
Vocos	0.98	1.00	0.09	1.00	0.03	0.11
BigVGAN	1.00	1.00	0.19	1.00	0.06	0.11

Findings

Robustness to Reconstruction: Unlike methods like AudioSeal, which fail under neural codec and vocoder reconstructions, FakeMark maintains high attribution accuracy across these distortions.

Cross-Domain Evaluation: FakeMark achieves near-perfect attribution on unseen cross-dataset samples where SSL-based classifier baselines (e.g., MMS-300M) fail completely.

Resistance to Removal: FakeMark demonstrates 100% robustness to averaging-based removal attacks.

Spectrogram vs. Waveform: Models that process spectrogram features are generally more robust to distortions than those processing raw waveforms.