DEEP AUDIO WATERMARKS ARE SHALLOW: LIMITATIONS OF POST-HOC WATERMARKING TECHNIQUES FOR SPEECH

Patrick O'Reilly & Bryan Pardo
Northwestern University
patrick.oreilly2024@u.northwestern.edu, pardo@northwestern.edu

Zeyu Jin & Jiaqi Su
Adobe Research
zejin@adobe.com, jsu@adobe.com

ABSTRACT

In the audio modality, state-of-the-art watermarking methods leverage deep neural networks to allow the embedding of human-imperceptible signatures in generated audio. The ideal is to embed signatures that can be detected with high accuracy when the watermarked audio is altered via compression, filtering, or other transformations. Existing audio watermarking techniques operate in a *post-hoc* manner, manipulating "low-level" features of audio recordings after generation (e.g. through the addition of a low-magnitude watermark signal). We show that this post-hoc formulation makes existing audio watermarks vulnerable to transformation-based removal attacks. Focusing on speech audio, we (1) unify and extend existing evaluations of the effect of audio transformations on watermark detectability, and (2) demonstrate that state-of-the-art post-hoc audio watermarks can be removed with no knowledge of the watermarking scheme and minimal degradation in audio quality.

1 INTRODUCTION

Recent generative models of audio have made the creation of realistic synthetic voices increasingly accessible (Eskimez et al., 2024; Chen et al., 2024). Unsurprisingly, advances in the capabilities of such models have been accompanied by a number of documented harms. For example, audio generative models have been used to "clone" or impersonate the voices of individuals without consent, enabling fraud (Moon, 2023; Edwards, 2022) and introducing challenges to existing intellectual property frameworks (Coscarelli, 2023).

Key to these harms is the inability of human listeners to distinguish between real and synthetic voices with any meaningful accuracy (Barrington & Farid, 2024). In response, researchers have proposed a number of methods for automatically identifying synthetic audio, among them *watermarking*. Watermarking methods embed a detectable signature in media to convey provenance information, and have traditionally been employed for intellectual property protection (Kirovski & Malvar, 2003). Recently, watermarking methods have gained attention for their potential to aid in the identification of synthetic media produced by generative models, including text (Kirchenbauer et al., 2023), images (Wen et al., 2023), video (Fernandez et al., 2024a), and audio (San Roman et al., 2024). Watermarking tends to outperform classification methods, which struggle with limited robustness and generalization beyond models observed during training (Liu et al., 2024b; Müller et al., 2022).

State-of-the-art (SOTA) audio watermarks operate in a *post-hoc manner* by embedding signatures in instances after generation (San Roman et al., 2024; Chen et al., 2023; Liu et al., 2024a; O'Reilly et al., 2024). In general, these post-hoc watermarks are restricted to embedding signatures via low-magnitude perturbations of generated audio, often at signal-to-noise ratios (SNR) on the order of 20-30dB, to avoid degrading perceived quality. This negatively impacts detection performance when watermarked audio is substantially modified or *transformed* (Liu et al., 2024e). Transformations

may be adversarial and intended to remove watermarks; standard processing stages applied for data transmission and storage, such as codec compression; or common steps taken in editing media with consumer-facing software, such as speeding or slowing speech, applying equalization, or adding room-tone with noise.

In this work, we examine the detectability of today's SOTA audio watermarks when watermarked audio is transformed, and thus the effectiveness of these watermarks in identifying synthetic media under real-world conditions and in the presence of motivated adversaries. We focus on speech audio in particular, due to the demonstrated harms of existing speech generative models and concomitant need for robust speech watermarking techniques.

Our contributions are as follows:

1. We unify and extend existing evaluations of the effect of audio transformations on watermark detectability
2. We demonstrate that state-of-the-art post-hoc audio watermarks can be removed with no knowledge of the watermarking scheme and minimal degradation in audio quality [1]

Notably, we find that two classes of transformation – neural network-based low-bitrate audio codecs and denoisers – suppress detection rates to near zero across all evaluated watermarking methods. Through this work, we hope to shed light on the limitations of existing deep neural network-based audio watermarks that embed signatures through "shallow" post-hoc perturbations, and to correspondingly motivate the development of more robust watermarking methods for audio.

1.1 Digital Audio Watermarking

Digital audio watermarking methods typically consist of two algorithms – an *embedding* algorithm that hides a signature in audio, and a *detection* algorithm that determines whether given audio contains a signature. Commonly, the embedding algorithm encodes a message within the embedded signature (e.g. a binary string of fixed length), and the detection algorithm attempts to recover this message; a decision as to whether audio contains a signature can then be rendered based on the similarity of the recovered message to a known watermark message (Chen et al., 2023; Liu et al., 2024a; O'Reilly et al., 2024), under the assumption that messages "recovered" from un-watermarked audio will be uniformly distributed over the set of all possible messages of the chosen length. Alternatively, a detection algorithm may directly produce a score indicating confidence that given audio bears a signature (Juvela & Wang, 2024b), or apply both approaches simultaneously (San Roman et al., 2024; Liu et al., 2025).

Traditional audio watermarking methods utilized interpretable signal-processing techniques for embedding and detecting signatures, e.g. via pseudorandom modulation of the spectrogram and correlation measures (Kirovski & Malvar, 2003; Tai & Mansour, 2019; Hua et al., 2016). More recently, researchers have proposed methods to learn embedding and detection algorithms directly from data using deep neural networks (Pavlović et al., 2022; Liu et al., 2023; San Roman et al., 2024; Chen et al., 2023; Liu et al., 2024a; O'Reilly et al., 2024; Li et al., 2024). In this formulation, an *embedder network* processes a given audio recording to embed a signature, while a *detector network* predicts the presence of a signature in a given audio recording and/or the message contained within the signature. Both networks are typically trained together to cooperatively embed and detect signatures. In general, deep neural network-based watermarking methods outperform signal-processing methods in their ability to embed human-imperceptible signatures that remain detectable when watermarked audio is transformed, in part due the incorporation of simulated transformations into the training process (Liu et al., 2023; O'Reilly et al., 2024; Chen et al., 2023).

Similar to their signal-processing predecessors, these deep neural network-based audio watermarks operate in a post-hoc manner. The methods of Liu et al. (2023), Pavlović et al. (2022), Chen et al. (2023), San Roman et al. (2024), Li et al. (2024), Singh et al. (2024b), O'Reilly et al. (2024), Liu et al. (2024a), and Liu et al. (2025) embed signatures via low-magnitude perturbations of existing audio at waveform or spectrogram representations. While attempts have been made to integrate watermarks into the process of speech audio generation, the resulting methods still operate in a fundamentally post-hoc manner. Cho et al. (2022), Juvela & Wang (2024b), Juvela & Wang (2024a), and

[1] We provide audio examples at `https://deep-watermark.github.io/`

Liu et al. (2024d) incorporate watermarks into neural network vocoders that map low-dimensional audio representations to high-fidelity waveform reconstructions; however, because the vocoders are trained with signal-level objectives to penalize the divergence between watermarked audio and the original audio from which vocoding representations are extracted, the resulting signatures still manifest as low-magnitude perturbations or are otherwise highly correlated with the source audio – they can not alter high-level speech attributes such as pitch, pronunciation, or timing [2]. Other works embed existing post-hoc watermarks in either the training data of audio generative models or in the latent decoders used to map generated instances to the data space (Zhou et al., 2024b;a; Wang et al., 2024; Roman et al., 2024; Singh et al., 2024a), and thus inherit the detection limitations of these post-hoc watermarks. Finally, Ji et al. (2025) use a masked language model operating on spectrogram tokens to replace selected audio frames with tokens drawn from a watermarked subset. While this approach hypothetically allows for the embedding of signatures within high-magnitude but realistic perturbations of an audio signal, in practice the signature is concealed in only a small subset of audio frames to avoid degrading audio quality (and is thus low-magnitude). In Figure 1, we visualize the signatures embedded in audio by selected neural network-based watermarks; we describe these watermarks in more depth in Section 3.1.

1.2 Evaluations of Audio Watermark Robustness

While the aforementioned works conduct evaluations of detection performance under various audio transformations, the choice of transformations and evaluation metrics differs significantly from work to work. For example, some works consider only signal-processing transformations such as additive noise and gain scaling (Pavlović et al., 2022; Chen et al., 2023), while others consider neural network-based transformations (O'Reilly et al., 2024; Liu et al., 2024a; 2025). Additionally, some works measure detection performance in terms of message recovery accuracy (Pavlović et al., 2022; Chen et al., 2023), and others in terms of the accuracy of discrimination between watermarked and un-watermarked audio (Juvela & Wang, 2024b;a; O'Reilly et al., 2024). Benchmarks such as AudioMarkBench (Liu et al., 2024c) and OmniSealBench (Research, 2024) partially bridge these differences but still focus mainly on signal-processing transformations and optimization-based attacks. Specifically, AudioMarkBench evaluates 15 transformations: 8 signal-processing transformations, 2 signal-processing codecs, 3 adversarial optimization attacks, and 2 first-generation neural codecs; OmniSealBench adds to these 6 signal-processing transformations. Our work differs from these previous works in the following key aspects: **(1)** we focus on watermark detection rather than message recovery; **(2)** we consider a much larger number of neural network-based transformations (14 vs. 2), including vocoder and denoiser transformation categories not included in AudioMarkBench or OmniSealBench; **(3)** we extend the codec bitrate evaluation from AudioMarkBench to more recent low-bitrate neural codecs developed specifically for speech, which better preserve audio quality; **(4)** we do not include optimization attacks, as we find that adversaries can more easily remove watermarks while retaining audio quality through the aforementioned single-pass neural network-based transformations; **(5)** we set detection score thresholds (Section 3.4) individually per transformation rather than on clean audio, preventing saturated error rates and allowing for clearer distinctions between transformation strengths (O'Reilly et al., 2024; Wen et al., 2023). **Through these choices, we hope to focus attention away from high-effort optimization-based attacks, and towards neural network-based transformations that can passively remove existing post-hoc watermarks while maintaining excellent audio quality**.

2 Audio Transformations

This work evaluates the efficacy of a variety of audio transformations in removing watermarks. These transformations vary in the degree to which they modify a given audio recording, the manner in which this modification is performed, and the amount of effort and expense involved in their application. We try to focus our evaluations on realistic transformations that may be passively applied to watermarked audio as it is transmitted and stored, as well as to attacks that can be undertaken by adversaries with reasonable motivation and resource constraints. We now describe and motivate the transformations considered in this work.

[2] While we find the method of Juvela & Wang (2024a) results in larger perturbations as measured by waveform difference (see Figure 1), the watermark signature is highly correlated with the source, especially at the spectrogram.

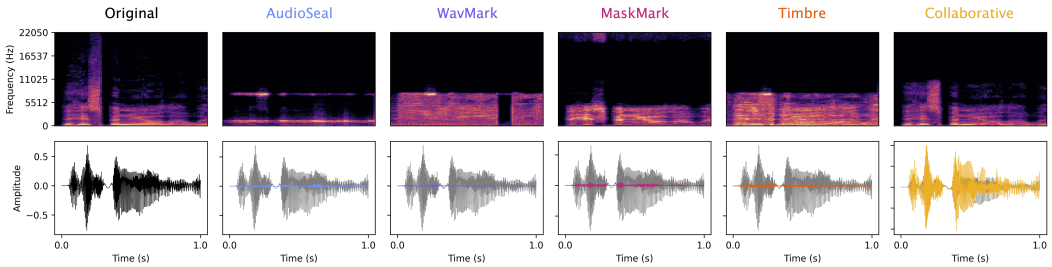

Figure 1: **Watermark signatures**. Given a short audio excerpt ("Original"), we apply five watermarking methods described in Section 3.1 ("AudioSeal," "WavMark," "MaskMark," "Timbre," and "Collaborative") and visualize the difference between original and watermarked audio at the spectrogram (top row) and waveform (bottom row). For each watermarking method, the original waveform is shown in grey and the difference in color.

2.1 Signal-Processing Transformations

The majority of audio watermarking works consider simple signal-processing transformations such as playback speed change, filtering, waveform dropout, additive noise, reverberation, pitch and timescale modification, phase shift, and nonlinear distortion (San Roman et al., 2024; Chen et al., 2023; O'Reilly et al., 2024; Liu et al., 2024a). These transformations are easy to implement and apply efficiently while preserving the quality and intelligibility of transformed audio – for instance, small phase shifts and changes in playback speed are often imperceptible. Signal-processing transformations may also occur naturally, e.g. re-recording of watermarked audio will likely introduce reverberation, background noise, and filtering. These factors mean that in practice, signal-processing transformations may be encountered as cascades of many individual transformations, potentially increasing the difficulty of watermark detection.

2.2 Audio Codec Compression

Audio data is often compressed with codecs for efficient transmission and storage. To maximize perceived audio quality at high compression ratios, codecs are designed to discard perceptually irrelevant information within an audio signal while preserving perceptually relevant information. This process may remove post-hoc watermarks that conceal low-magnitude signatures in perceptually irrelevant regions of audio, and as a result recent works have explored the detection performance of audio watermarks under both traditional signal-processing codecs (mp3, 1993; Valin et al., 2012) and neural network-based codecs (Kumar et al., 2023b; Défossez et al., 2022). In general, existing watermarks appear to fare better against signal-processing codecs than neural codecs (San Roman et al., 2024; Liu et al., 2024a; Juvela & Wang, 2024b; O'Reilly et al., 2024), although direct comparisons of SOTA watermarks on multiple neural codecs are scarce in the literature. In addition to the general-purpose neural codecs studied in previous works, we consider for the first time recent *low-bitrate* neural codecs developed for high-ratio compression of speech.

2.3 Neural Vocoders

Similar to codecs, neural network-based vocoders reconstruct speech audio from compressed representations such as mel-spectrograms (Kong et al., 2020). While vocoders are an important component of many speech generative models (Eskimez et al., 2024; Chen et al., 2024), it is unlikely that watermarked generated audio will be transformed via additional vocoding unless an adversary specifically uses a vocoder to attempt to remove watermarks. Compared to codecs, a smaller number of works have explored watermark robustness to vocoders (O'Reilly et al., 2024; Liu et al., 2024a).

2.4 Neural Denoisers

Neural network "denoisers" are often employed to restore noisy or otherwise corrupted audio (Défossez et al., 2020; Fu et al., 2021; Hu et al., 2020; Yang et al., 2024). An attacker might seek to remove audio watermarks by adding noise to watermarked audio and then processing with

a denoiser, in hopes the denoiser will remove the low-magnitude watermark signature along with the noise while preserving the perceptual quality of the original recording. While there exist a number of conventional signal-processing algorithms for removing noise, recent neural network-based denoisers show improved tolerance for high noise levels and are capable of removing complex non-stationary perturbations from speech signals (Défossez et al., 2020; Yang et al., 2024) – both potentially useful properties for watermark removal. A small number of audio watermarking works have explored robustness to neural denoisers (López-López et al., 2024; O'Reilly et al., 2024).

2.5 Watermark-Aware Transformations

While recent works have explored the vulnerability of neural network-based audio watermarks to overwriting attacks (Liu et al., 2024a) that flood watermarked audio with multiple signatures and optimization-based attacks (San Roman et al., 2024; Liu et al., 2025) that craft adversarial examples to fool detectors, both approaches assume some degree of knowledge of the attacked watermarking algorithms. We find that this knowledge is not necessary for removing watermarks in practice, and instead focus on the watermark-agnostic transformations listed in the previous subsections.

3 Experiments

We collect examples of the transformation types listed in Subsections 2.1 - 2.4 from the audio watermarking literature and conduct experiments measuring their efficacy in removing audio watermarks while preserving the overall quality of watermarked audio. In addition to these existing transformations, we also evaluate more recent and as-yet unstudied codecs, vocoders, and denoisers. Finally, we demonstrate how the audio degradation incurred through transformations can be reduced through band-splitting to target the frequency regions where existing watermarks operate.

3.1 Watermarking Methods

We evaluate five SOTA neural network-based watermarking methods for audio, described here.

AudioSeal (San Roman et al., 2024) embeds watermarks via a residual (i.e. additive) waveform perturbation predicted using an Encodec-like (Défossez et al., 2022) neural network, and is trained for simultaneous detection and message recovery with sample-level temporal resolution. Detection is performed by averaging sample-level detector scores over a segment of audio.

WavMark (Chen et al., 2023) embeds watermarks via a residual waveform perturbation predicted using an invertible convolutional neural network whose weights are shared with a detector network. Detection is performed by measuring the bit accuracy of a recovered message against a known watermark message. Robustness to temporal distortions (e.g. time shift, speed change) can be improved via a localization scheme using additional encoded message bits.

Timbre-Watermark (Liu et al., 2024a) embeds watermarks via a convolutional neural network operating on the magnitude spectrogram; watermarked audio is obtained by inverting the watermarked magnitude spectrogram using the original (un-watermarked) phase. An embeded message can be recovered by averaging a frame-level detector network's predictions over time, and detection by measuring the bit accuracy of a recovered message against a known watermark message. For AudioSeal, WavMark, and Timbre-Watermark, we use the implementations provided in OmniSeal-Bench (Research, 2024).

MaskMark (O'Reilly et al., 2024) operates similarly to Timbre-Watermark but embeds watermarks via a multiplicative mask at the magnitude spectrogram rather than via convolution. Detection is also performed analogously, but using a network with a larger receptive field.

Finally, **Collaborative Watermark** (Juvela & Wang, 2024b) embeds watermarks by synthesizing audio waveforms from a mel-spectrogram representation using a watermarked HiFiGan (Kong et al., 2020) vocoder. We use a variant (Juvela & Wang, 2024a) trained to be robust to processing with Descript Audio Codec (Kumar et al., 2023b).

	Transformation	Quality Preservation			Watermark Detection (TPR@1%FPR) ↓				
		ASR-CER ↓	SIM ↑	SQUIM-MOS ↑	Timbre	WavMark	AudioSeal	MaskMark	Collaborative
Sig.-Proc.	Speed	0.03	0.93	4.81	0.27	0.00	0.15	1.00	1.00
	Reverb	0.01	0.93	3.73	1.00	0.98	0.96	0.93	0.83
	Pitch shift	0.01	0.86	4.67	0.17	0.00	0.04	1.00	0.71
	Low-pass	0.00	0.79	3.93	1.00	1.00	1.00	1.00	0.00
	High-pass	0.01	0.80	2.53	1.00	1.00	1.00	1.00	1.00
	Time stretch	0.03	0.90	4.67	1.00	0.44	0.35	1.00	0.74
	Equalizer	0.01	0.97	3.91	1.00	1.00	0.99	0.85	0.86
	Noise (0dB SNR)	0.08	0.71	2.66	0.00	0.00	0.85	0.10	0.00
	Noise (10dB SNR)	0.02	0.83	2.83	0.85	0.00	1.00	0.91	0.02
	Noise (20dB SNR)	0.01	0.90	4.58	1.00	0.04	1.00	1.00	0.06
Codec	MP3	0.00	0.87	4.54	1.00	1.00	1.00	1.00	0.64
	OPUS	0.00	0.97	4.80	1.00	1.00	1.00	1.00	0.79
	DAC	0.00	0.93	4.77	0.92	0.00	1.00	0.22	0.69
	ESC	0.00	0.96	4.81	0.99	0.00	1.00	0.25	0.69
	SpectralCodec	0.00	0.95	4.76	0.56	0.00	0.01	0.64	0.59
	Encodec	0.00	0.92	4.79	0.68	0.00	1.00	0.15	0.26
	Encodec-VoiceCraft	0.01	0.88	4.72	0.06	0.00	0.01	0.07	0.02
	SpeechTokenizer	0.01	0.85	4.59	0.03	0.00	0.00	0.04	0.37
	TiCodec	0.01	0.77	4.66	0.02	0.00	0.01	0.03	0.41
	FACodec	0.01	0.88	4.79	0.01	0.00	0.00	0.03	0.44
Vocoder	BigVGAN	0.00	0.99	4.78	1.00	0.00	0.29	1.00	0.69
	DiffWave	0.01	0.82	3.97	0.98	0.00	0.00	0.93	0.04
	HiFi-GAN	0.01	0.95	4.71	0.80	0.00	0.03	0.38	0.54
	BigVGAN2	0.00	0.99	4.82	1.00	0.00	0.47	1.00	0.81
	Vocos	0.00	0.98	4.77	1.00	0.00	0.14	1.00	0.48
Denoiser	Denoiser (0dB SNR)	0.10	0.63	4.60	0.00	0.00	0.02	0.01	0.00
	Denoiser (10dB SNR)	0.02	0.73	4.71	0.01	0.00	0.03	0.01	0.00
	Denoiser (20dB SNR)	0.01	0.81	4.75	0.04	0.00	0.04	0.03	0.01

Table 1: **Watermark detection under transformations**. Under *Quality Preservation*, we evaluate the audio degradation incurred by each transformation in terms of intelligibility ("ASR-CER"), speaker identity similarity ("SIM"), and overall quality ("SQUIM-MOS"). Cells are shaded darker where degradation is minimized. Arrows in column headers indicate the optimal direction to minimize audio degradation. Under *Watermark Detection*, we evaluate the effect of each transformation on the true-positive detection rate of watermarks, given a fixed 1% false-negative rate ("TPR@1%FPR"). Lower values indicate stronger watermark removal. Cells are shaded darker as the true positive detection rate decreases. The transformations that most effectively and inconspicuously remove watermarks are therefore shaded in every column.

3.2 DATASET

We embed watermarks in 200 five-second voiced segments from the `clean` subset of the DAPS dataset (Mysore, 2015), consisting of clean speech from 10 male and 10 female speakers. The DAPS dataset provides a reasonable approximation of high-quality synthetic speech for our purposes, as all utterances in the `clean` subset were recorded professionally at 44.1kHz with very little audible noise. DAPS is not present in the training data of any of the evaluated watermarks.

We choose a five-second segment length to allow for fair comparisons, and to ensure that transformation strength, not segment length, is the dominant factor in detection performance. While some works (San Roman et al., 2024) consider larger evaluation sets, we are not interested in fine-grained distinctions between true positive rates but rather in quantifying the poor overall performance of existing watermarking methods on speech recordings under transformations. In this case, comparing detection scores on 400 examples (200 un-watermarked, 200 watermarked) for each transformation and watermarking approach suffices.

The evaluated watermarks were not all designed to operate at the same sample rate. To account for differences in sample rate, we adopt the method proposed by O'Reilly et al. (2024):

1. Audio is split into two bands via low-pass and high-pass filters at the Nyquist frequency of the watermarking method's native sample rate
2. The low band is resampled to the watermarking method's native sample rate and the watermark applied
3. The watermarked low band is resampled to the original audio rate, normalized to account for gain change, and summed with the high band

In this manner a watermark can be applied to audio beyond the watermark's native sample rate. Detection can then be performed by resampling audio to the watermark's native sample rate, discarding

un-watermarked high-frequency content. We find that this method has essentially no effect on the detection performance of the evaluated watermarks, while allowing for embedding and detection in high-quality speech.

3.3 TRANSFORMATIONS

We consider the following transformations from the categories described in Sections 2.1 - 2.4.

Signal-processing transformations: we apply Gaussian noise at 20dB, 10dB, and 0dB SNR; equalization across 6 bands with random gains sampled in $[-1, 1]$dB; low-pass filtering at 4kHz; high-pass filtering at 500Hz; random pitch shift sampled in $[-1, 1]$ semitones; reverb using impulse responses sampled from the MIT dataset (Traer & McDermott, 2016); and playback speed change and phase-vocoding time-stretch at factors sampled in $[-0.95, 1.05]$. We find that a large number of signal-processing transformations explored in previous works such as phase shift, waveform dropout, gain scaling, and sample-level quantization have virtually no effect on detection performance and therefore omit results to save space (Liu et al., 2024c; Chen et al., 2023; O'Reilly et al., 2024).

Audio codec compression: we apply the signal-processing codecs MP3 (mp3, 1993) and OPUS (Valin et al., 2012) at 24 kbps. We also apply the following neural network-based codecs: Encodec (Défossez et al., 2022) at 24 kbps; ESC (Gu & Diao, 2024) at 9.0 kbps; Descript Audio Codec (DAC) (Kumar et al., 2023b) at 8 kbps; Spectral-Codec (Langman et al., 2024) at 6.9 kbps; FACodec (Ju et al., 2024) at 4.8 kbps; SpeechTokenizer (Zhang et al., 2024) at 4 kbps; TiCodec (Ren et al., 2024) at 3 kpbs; and the "VoiceCraft" Encodec variant of Peng et al. (2024) at 2.2 kbps. Notably, ESC, Spectral-Codec, FACodec, SpeechTokenizer, TiCodec, and Encodec-VoiceCraft were developed specifically for speech audio. For codecs that operate below 44.1kHz (namely Encodec, ESC, FACodec, SpeechTokenizer, TiCodec, and Encodec-VoiceCraft), we use the band-splitting method described in Section 3.2.

Neural vocoders: we apply the mel-spectrogram vocoders BigVGAN (Lee et al., 2023), BigVGAN2 (gil Lee & Valle, 2024), DiffWave (Kong et al., 2021), HiFi-GAN (Kong et al., 2020), and Vocos (Siuzdak, 2023). With the exception of BigVGAN2, all evaluated vocoders are processed with the aforementioned band-splitting method to account for sample rates below 44.1kHz.

Neural denoisers: we build on the denoising attack of López-López et al. (2024), which was shown to be effective against the neural network-based audio watermarks of San Roman et al. (2024) and Chen et al. (2023). While López-López et al. utilize a discriminative DCCRN (Hu et al., 2020) denoiser model and a single noise level, we utilize a more recent generative denoiser model (Yang et al., 2024) and consider multiple noise levels of 20dB, 10dB, and 0dB SNR to examine the trade-off between attack strength and audio quality degradation.

3.4 METRICS

We are primarily interested in discrimination between real and synthetic audio. Although information recovery and attribution are desirable capabilities for watermarking methods, we find that the prerequisite task of simply detecting the presence of a watermark is already a high a hurdle to clear when watermarked audio is transformed. Therefore, we adopt the standard approach of measuring the achievable true-positive rate at a fixed false-positive rate based on detection scores obtained over a balanced dataset of watermarked and un-watermarked audio (O'Reilly et al., 2024; Liu et al., 2025). While watermarking methods may be required to operate at false-positive rates below 0.1% to function practically in large-scale production environments (San Roman et al., 2024; Fernandez et al., 2024b), we find that existing methods fail to detect watermarked audio under many transformations even when allowing much larger false-positive rates. We therefore use the true-positive rate at a fixed 1% false-positive rate ("TPR@1%FPR").

Whereas previous works (López-López et al., 2024) quantify the audio degradation incurred by transformations using invasive metrics such as PESQ (Rix et al., 2001), these metrics do not correlate well with human perception for transformations that compromise precise alignment (e.g. speed change). Instead, we measure the audio degradation incurred by transformations along three interpretable axes: **speech intelligibility**, as measured by the character error rate between transcripts predicted by a HuBERT-based (Hsu et al., 2021) speech recognition system for watermarked audio before and after transformation ("ASR-CER"); **speaker identity preservation**, as measured by the

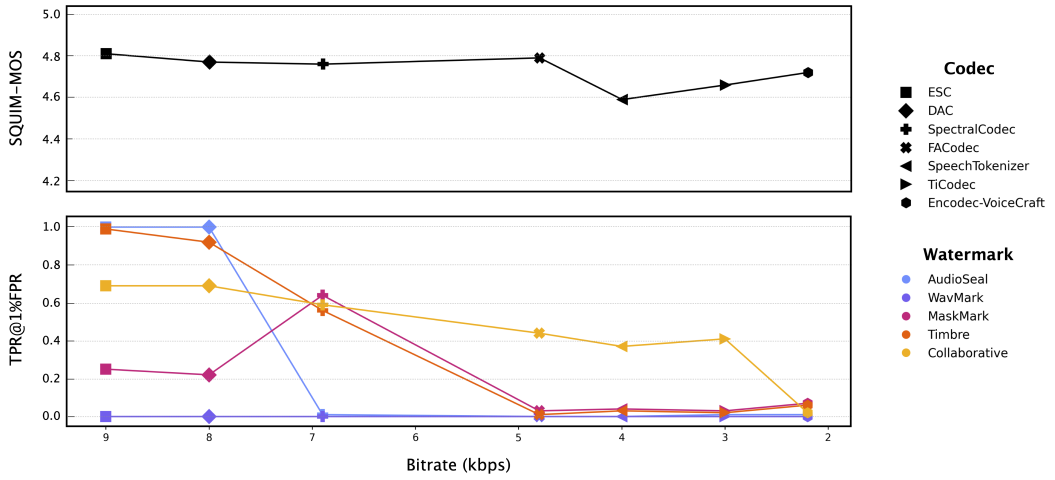

Figure 2: **Robustness to low-bitrate codecs**. We plot the achievable true-positive rate at a fixed 1% false-positive rate ("TPR@1%FPR") for each watermarking method under transformation via low-bitrate neural codecs. Watermark detection declines at low bitrates while audio quality is generally preserved, as indicated by SQUIM-MOS scores.

embedding-space cosine similarity between embeddings produced by a speaker recognition system (Heo et al., 2020) for watermarked audio before and after transformation ("SIM"); and **audio quality similarity** as measured through mean opinion scores (MOS) predicted by the automated SQUIM (Kumar et al., 2023a) metric, with watermarked audio serving as a non-matching quality reference for transformed watermarked audio ("SQUIM-MOS"). Lower ASR-CER implies better intelligibility, higher SIM implies better speaker identity preservation, and higher SQUIM-MOS implies that the transformation introduces less-perceptible artifacts or changes to channel characteristics. ASR-CER has a lower bound of 0, SIM ranges from -1 to 1, and SQUIM-MOS ranges from 1 to 5.

The results of our evaluation are presented in Table 1. We plot a subset of these results for transformations that leverage low-bitrate neural audio codecs in Figure 2. In Figure 3, we illustrate the application of the denoiser attack to audio watermarked with the method of Liu et al. (2024a).

4 Discussion

We summarize key trends in our experimental results below.

Post-hoc watermarks lack general robustness. No evaluated watermark demonstrates general robustness across all classes of transformation. While some watermarks fare well against some classes (e.g. Timbre-Watermark and MaskMark against vocoders and signal-processing transformations), all watermarks have clear weaknesses that can be exploited by adversaries to thwart detection. Moreover, complementary weaknesses can be leveraged to construct combined attacks. For example, an adversary could apply speed change to foil Timbre-Watermark, WavMark, and AudioSeal, and then apply DAC to additionally foil MaskMark.

Neural network-based transformations are strong watermark removers. Neural network-based codecs, vocoders, and denoisers are capable of significantly reducing watermark detectability while preserving audio quality as quantified by our metrics. For example, our denoiser attack brings detection rates to near zero for all watermarks even when operating in its weakest configuration (20dB SNR) and allowing a generous 1% false-positive rate. Such transformations can be applied by leveraging off-the-shelf neural network models without any re-training or modification, making them a convenient option for even low-resource adversaries.

Low-bitrate neural audio codecs pose a particular challenge for speech watermarking. Neural audio codecs increasingly leverage the structured nature of human speech to achieve large compression ratios and correspondingly low bitrates. To maintain high perceptual quality as bitrate decreases, these codecs must discard perceptually irrelevant details (e.g. fine-grained spectral struc-

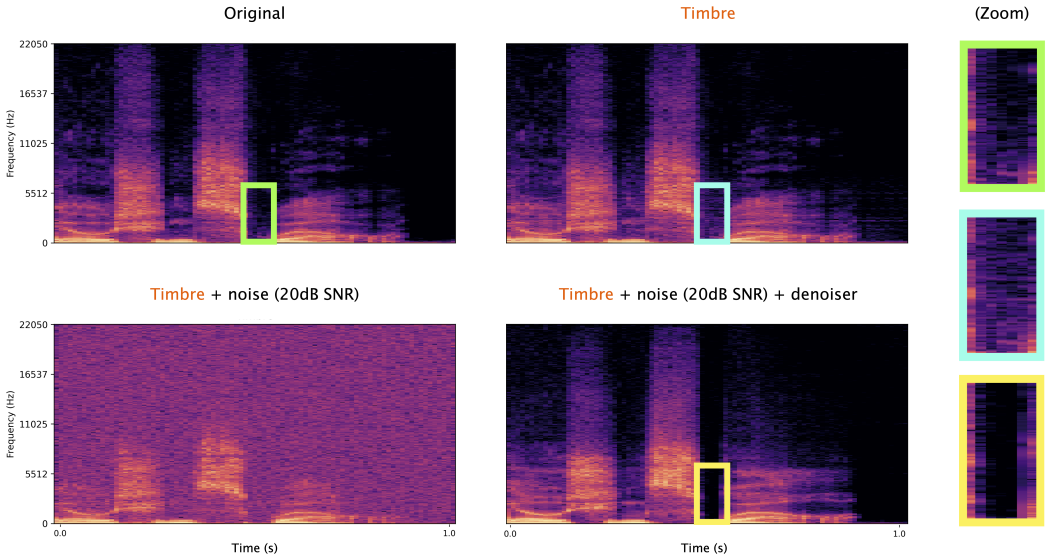

Figure 3: **Denoiser attack**. We illustrate the application of the denoiser attack to Timbre-Watermark with spectrogram plots. Artifacts introduced by the watermark (blue rectangle) can be seen more clearly when compared to the original spectrogram (green rectangle); these are removed by the application of noise followed by the denoiser (yellow rectangle).

ture) during the encoding process, and then "hallucinate" plausible reconstructions of these details during the decoding process. Existing post-hoc watermarks generally embed signatures through low-magnitude perturbations of perceptually irrelevant details, and are thus erased through an encoding-decoding pass. In Figure 2, we show that watermark detection performance under neural codec reconstruction generally decreases with the bitrate, even as codecs maintain high audio quality. This class of transformation is worth exploring further, as researchers continue to develop neural codecs capable of operating at even lower bitrates than those evaluated in this work, often with comparable or better audio quality (Ji et al., 2024; Parker et al., 2024). Moreover, if such codecs are integrated into digital media platforms due to their strong performance, they may inadvertently remove post-hoc watermarks applied to track synthetic audio.

5 Conclusion

In this work, we demonstrate that SOTA post-hoc audio watermarks can be effectively and inconspicuously removed from speech through a number of neural network-based transformations. In particular, low-bitrate speech codecs and neural network-based denoisers reduce the detection rates of all evaluated watermarks to near zero even when allowing a large 1% false-positive rate.

We stress that the effectiveness of neural network-based transformations in removing existing post-hoc audio watermarks does not mean these watermarking methods are without utility. Fernandez et al. (2024b) convincingly argue that watermarks need not be maximally robust to serve as detection mechanisms and deterrents against the misuse of generative models. Moreover, in large-scale production environments where high false-positive rates can incur significant costs, watermark-based systems for synthetic media detection may have to balance many considerations beyond robustness to removal attacks (San Roman et al., 2024). Through this work, we hope to emphasize that there are simple and broadly effective watermark removal transformations available to adversaries *right now*, and that future works claiming to develop robust watermarks should account for these transformations in their evaluations or else show why they may be considered out-of-scope.

6 Acknowledgements

This work was supported in part by NSF Award Number 2222369.

REFERENCES

Mp3 (mpeg layer iii audio encoding), 1993. URL https://www.iso.org/standard/22412.html.

Sarah Barrington and Hany Farid. People are poorly equipped to detect ai-powered voice clones. *arXiv preprint arXiv:2410.03791*, 2024.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. URL https://arxiv.org/abs/2410.06885.

Yongbaek Cho, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributable-watermarking of speech generative models. In *Proc. ICASSP*, 2022.

J. Coscarelli. An a.i. hit of fake 'drake' and 'the weeknd' rattles the music world. *The New York Times*, 2023.

A. Défossez, G. Synnaeve, and Y. Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.

A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Benj Edwards. Bruce willis denies selling deepfake rights to deepcake. *"Ars Technica"*, 2022. URL https://arstechnica.com/information-technology/2022/10/bruce-willis-denies-selling-deepfake-rights-to-deepcake/.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Pierre Fernandez, Hady Elsahar, I. Zeki Yalniz, and Alexandre Mourachko. Video seal: Open and efficient video watermarking. *arXiv preprint arXiv:2412.09492*, 2024a.

Pierre Fernandez, Anthony Level, and Teddy Furon. What lies ahead for generative ai watermarking. In *Proceedings of the ICML 2024 Workshop on Generative AI and Law*, 2024b.

S. Fu, C. Yu, T. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao. Metricgan+: An improved version of metricgan for speech enhancement. In *Interspeech*, 2021.

Sang gil Lee and Rafael Valle. Achieving state-of-the-art zero-shot waveform audio generation across audio types, sep 2024. URL https://developer.nvidia.com/blog/achieving-state-of-the-art-zero-shot-waveform-audio-generation-across-audio-types/. NVIDIA Developer Blog.

Yuzhe Gu and Enmao Diao. Esc: Efficient speech coding with cross-scale residual vector quantized transformers, 2024.

Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung. Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153*, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Proc. InterSpeech 202*, 2020.

Guang Hua, Jiwu Huang, Yun Q Shi, Jonathan Goh, and Vrizlynn LL Thing. Twenty years of digital audio watermarking—a comprehensive review. *Signal processing*, 128:222–242, 2016.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, Wen Wang, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Shengpeng Ji, Ziyue Jiang, Jialong Zuo, Minghui Fang, Yifu Chen, Tao Jin, and Zhou Zhao. Speech watermarking with discrete intermediate representations. In *AAAI*, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Lauri Juvela and Xin Wang. Audio codec augmentation for robust collaborative watermarking of speech synthesis. *arXiv preprint arXiv:2409.13382*, 2024a.

Lauri Juvela and Xin Wang. Collaborative watermarking for adversarial speech synthesis. In *Proc. ICASSP*, 2024b. arXiv preprint arXiv:2309.15224.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning (ICML)*, 2023.

D. Kirovski and H.S. Malvar. Spread-spectrum watermarking of audio signals. *IEEE Transactions on Signal Processing*, 51(4):1020–1033, 2003. doi: 10.1109/TSP.2003.809384.

J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP*, 2023a.

Rithesh Kumar, Prem Seetharaman, Ishaan Kumar Alejandro Luebs, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. In *NeurIPS*, 2023b.

Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Boris Ginsburg. Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis, 2024.

S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR*, 2023.

Pengcheng Li, Xulong Zhang, Jing Xiao, and Jianzong Wang. IDEAW: Robust neural audio watermarking with invertible dual-embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4500–4511, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.emnlp-main.258`.

C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang, and N. Yu. Dear: A deep-learning-based audio re-recording resilient watermarking. In *AAAI*, 2023.

Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. Detecting voice cloning attacks via timbre watermarking. In *Network and Distributed System Security Symposium*, 2024a. doi: 10.14722/ndss.2024.24200.

Hongbin Liu, Youzheng Chen, Arun Narayanan, Athula Balachandran, Pedro J. Moreno, and Lun Wang. Can deepfake speech be reliably detected? *arXiv preprint arXiv:2410.06572*, 2024b.

Hongbin Liu, Moyang Guo, Zhengyuan Jiang, Lun Wang, and Neil Zhenqiang Gong. Audiomarkbench: Benchmarking robustness of audio watermarking. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024c.

Weizhi Liu, Yue Li, Dongdong Lin, Hui Tian, and Haizhou Li. Groot: Generating robust watermark for diffusion-model-based audio synthesis. In *ACM MM*, 2024d.

Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:22410.05470*, 2024e.

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

Alvaro López-López, Eros Rosello, and Angel M. Gomez. Speech watermarking removal by DNN-based speech enhancement attacks. In *Proceedings of IberSPEECH*, 2024.

M. Moon. A new ai voice tool is already being abused to make deepfake celebrity audio clips. *Engadget*, 2023.

G. J. Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? - a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8), 2015.

Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Proc. InterSpeech 2022*, 2022.

Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. Maskmark: Robust neural watermarking for real and synthetic speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Julian D Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu. Scaling transformers for low-bitrate high-quality speech coding. *arXiv preprint arXiv:2411.19842*, 2024.

K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski. Robust speech watermarking by a jointly trained embedder and detector using a dnn. *Digital Signal Processing*, 122:103381, 2022. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2021.103381. URL https://www.sciencedirect.com/science/article/pii/S1051200421004206.

Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973v1*, 2024.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chuyuan Zhang, and Junzuo Zhou. Fewer-token neural speech codec with time-invariant codes. In *ICASSP*, 2024.

Facebook Research. Omnisealbench. https://github.com/facebookresearch/omnisealbench, 2024.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001.

Robin San Roman, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel. Latent watermarking of audio generative models. *arXiv preprint arXiv:2409.02915*, 2024.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre D´efossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. *International Conference on Machine Learning (ICML)*, 2024.

Mayank Kumar Singh, Naoya Takahashi, Wei-Hsiang Liao, and Yuki Mitsufuji. LOCKEY: A novel approach to model authentication and deepfake tracking. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024a.

Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, and Yuki Mitsufuji. SilentCipher: Deep Audio Watermarking. In *Proc. InterSpeech 2024*, 2024b.

H. Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2023.

Y. Tai and M. F. Mansour. Audio watermarking over the air with modulated self-correlation. In *ICASSP*, 2019.

J. Traer and J. H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

J.-M. Valin, K. Vos, and T. B. Terriberry. Definition of the opus audio codec. *IETF RFC 6716*, 2012. URL `https://tools.ietf.org/html/rfc6716`.

Helin Wang, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. Ssr-speech: Towards stable, safe and robust zero-shot text-based speech editing and synthesis. *arXiv preprint arXiv:2409.07556*, 2024.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Haici Yang, Jiaqi Su, Minje Kim, and Zeyu Jin. Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens. In *Interspeech*, 2024.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. In *ICLR*, 2024.

Junzuo Zhou, Jiangyan Yi, Yong Ren, Jianhua Tao, Tao Wang, and Chu Yuan Zhang. Wmcodec: End-to-end neural speech codec with deep watermarking for authenticity verification. *arXiv preprint arXiv:2409.12121*, 2024a.

Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, Ye Bai, Chu Yuan Zhang, Yong Ren, and Zhengqi Wen. Traceablespeech: Towards proactively traceable text-to-speech with watermarking. In *Proc. InterSpeech*, 2024b.