CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Limitations of watermarking AI-generated speech using AudioSeal
Authors	Faziludeen, Shameer;Sankar, Arun;De Leon, Phillip L.;Roedig, Utz
Publication date	2025
Original Citation	Faziludeen, S., Sankar, A., De Leon, P. L. and Roedig, U. (2025) 'Limitations of watermarking AI-generated speech using AudioSeal', 7th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS 2025), Pittsburgh, PA, USA, 11-14 November 2025.
Type of publication	Conference item
Rights	© 2025, the Authors. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-10 14:33:40
Item downloaded from	https://hdl.handle.net/10468/18360

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Limitations of Watermarking AI-Generated Speech using AudioSeal

1st Shameer Faziludeen
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
sfaziludeen@ucc.ie

2nd Arun Sankar M. S.
Department of Electronic Engineering and Communications
South East Technological University
Carlow, Ireland
arun.sankar@setu.ie

3rd Phillip L. De Leon
Department of Electrical Engineering
University of Colorado Denver
Denver, Colorado, USA
Phillip.DeLeon@ucdenver.edu

4th Utz Roedig
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
u.roedig@ucc.ie

***Abstract*—AI-generated speech is currently of such high quality that it is indistinguishable from a genuine human speaker. Expert listeners or purpose-built detectors are no longer able to reliably distinguish between the two. Thus, it has been proposed that AI systems which generate speech embed a secondary signal or watermark that allows identification. AudioSeal is currently the most advanced watermarking algorithm proposed for this purpose and its resilience against common channel and coding effects has been demonstrated. In this paper, we present approaches which compromise AudioSeal, making it unusable in practical settings. First, we describe two methods that result in a shifting of the detector score distribution for watermarked speech toward the distribution for unwatermarked speech. Second, we describe a method that uses AudioSeal watermarks generated for a particular speaker's signal on a different speaker's signal, i.e. unmatched watermarks. These unmatched watermarks, which could be imposed on genuine human speech, are also inaudible, resilient, and result in a shift of the detector score distribution away from unwatermarked speech. Considering both approaches, we observe that AudioSeal watermarks cannot be used to reliably identify AI-generated speech from genuine human speech due to overlapping score distributions. While our results are specific to AudioSeal, it casts doubt on the approach of watermarking in general to identify AI-generated speech.**

***Index Terms*—Watermarking, AI Generated Speech, Deep Fake, AudioSeal, Audio Watermarking, Watermark Attacks**

I. Introduction

It is now relatively easy to build a high quality AI speech generation model for a specific individual using only a few seconds of audio recordings [1]. Thereafter, speech signals can be generated for this specific individual such that expert listeners or purpose-built detectors are no longer able to reliably distinguish between AI generated speech and genuine human speech. As a result, people may be misled when listening to generated or "deep fake" speech. For example, using sophisticated attacks such as speech synthesis, systems using voice interaction or voice based authentication can be circumvented [2]. However, generated speech also has several positive applications [3]. For example, generated speech can help individuals with speech impairments or may be used in education settings to generate teaching materials attributed to a specific educator.

To safeguard AI-generated speech against misuse it has been proposed that a watermark be superimposed onto the speech so that it may be identified as AI-generated speech [4], [5]. The idea here is that stakeholders providing generative systems add watermarks. Many application domains could be protected using the aforementioned watermarking approach. Consider a large social media platform where users share audio (and/or video) content. Users could be protected from deep fakes as a watermark embedded in any uploaded content could be easily spotted. Rogue players will not adhere to such an arrangement but it is assumed that a majority of stakeholders, and especially the ones developing generative systems, have an interest to comply. Future regulation may also require watermarking or other technology to combat deception.

AudioSeal [4] is currently the most advanced watermarking technique and it has been shown to outperform all other watermarking techniques such as WavMark and Timbre. Recent work [6] has proposed a set of benchmarks for evaluating the robustness of audio watermarking against watermark removal and watermark forgery and also confirmed the robustness of AudioSeal compared to other methods. AudioSeal can easily distinguish watermarked speech signals from unwatermarked signals. The AudioSeal detector outputs a score for a speech signal which is close to one for a signal containing a watermark and close to zero for an unwatermarked signal. Thus, a decision threshold is easily found to distinguish both signals with perfect accuracy [4]. Even if a watermarked signal is

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 19/FFP/6775 and 13/RC/2077_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

subjected to edits (transformations such as noise or filters) near perfect classification accuracy is possible. Edits to the signal may reduce the detection score, however, the watermarked signals can usually still be distinguished from unwatermarked signals with near perfect accuracy. Some edits can reduce the detection score dramatically such that watermarked and unwatermarked signals cannot be distinguished. However, in these cases, the speech signal is distorted significantly and a manipulation is obvious. For our work, we utilize the objective metric, PESQ (Perceptual Evaluation of Speech Quality) [7] for assessing such distortion. Speech quality degradation which can be perceived by humans will be reflected in the PESQ score making it a reliable measure widely employed in quality testing, including by the authors in AudioSeal [4].

In this paper, we demonstrate that it is possible to: i) provide an edit that *decreases* detection scores of watermarked signals significantly while retaining signal quality (referred to as compromised watermarked signals) and ii) propose techniques to *increase* detection scores of speech signals (referred to as using unmatched watermarks). As a result, score distributions of watermarked signals and unmatched watermarked signals overlap into each other. AudioSeal becomes a compromised approach to watermarking AI-generated speech signals. The specific contributions of the paper are:

- *Reducing detection scores of watermarked signals*: We describe two methods that allow us to reduce the detection score of watermarked speech while maintaining quality. The first method is based on re-applying a watermark and assumes access to the watermark generator. The second method uses a common speech enhancement method which puts no constraints on the attacker. Recent work has employed speech enhancement [8], [9] as an attack, however this requires a deep learning based model to be effective unlike our enhancement method.
- *Increasing detection scores with unmatched watermarked signals*: We show that it is possible to add an unmatched watermark to any speech signal to produce unmatched watermarked signals. These unmatched watermarked signals exhibit a high detection score while maintaining speech quality. We also show that the average detection score of a pool of unmatched watermarked signals can be significantly improved if we further assume access to the detector output.
- *Compromising AudioSeal*: We illustrate the impact of the above techniques on a hypothetical social media platform which uses AudioSeal for watermarking AI-generated speech content. To the best of our knowledge, ours is the first attempt to consider realistic attack scenarios and to provide ROC curves which illustrate resulting error rates that significantly degrade system performance.

In the next section we provide a background on AudioSeal and provide an evaluation of its performance. In Section III we describe methods for reducing detection scores of watermarked signals while in Section IV we outline methods to increase detection scores of unwatermarked signals. Section V discusses the findings and specifically we describe the impact of our work on practical application scenarios. Section VI discusses related work. Finally, we provide several conclusions in Section VII.

II. Background

A. AudioSeal

Among the different audio watermarking techniques, AudioSeal [4] is considered a superior model in various respects, followed closely by Wavmark [5]. Audio watermarking approaches are generally categorized into zero-bit, aimed at detecting the presence of a watermark, and multi-bit, which offers additional details during detection, such as the watermarking model used and other metadata. AudioSeal supports both zero-bit and multi-bit watermarking and is capable of high-speed detection, making it well-suited for real-time applications.

AudioSeal comprises a generator and a detector that are jointly trained, enabling the watermarking method to adapt to the detector. This approach also involves training the generator on unwatermarked samples, enhancing its performance in practical scenarios. The training process focuses on two objectives: minimizing the perceptual differences between original and watermarked samples, and maximizing detection scores. To improve robustness against signal modifications, the training samples undergo time-domain data augmentation. Additionally, a localized loss function is employed to facilitate sample-level watermark detection, allowing the identification of watermarked segments within an audio signal [4].

The perceptual quality of watermarked signals produced by both AudioSeal and Wavmark is comparable, with minimal noticeable differences to listeners. AudioSeal incorporates auditory masking during its training process, which enhances the model's robustness against various types of signal modifications. As a result, AudioSeal outperforms Wavmark in handling edits such as noise addition and low-pass filtering. However, its performance falls short with high-pass filtering. This limitation stems from Wavmark's approach, which embeds the watermark in the high-frequency components of the audio signal, making it more resilient to high-pass filtering compared to AudioSeal.

B. Notation

We begin by introducing slightly different notation from [4], where we denote the watermark as

$$\delta_j = G(s, m_j) \tag{1}$$

where s is the speech signal, m_j is a 16 bit message, and $G(\cdot)$ is the watermark generator. The watermarked signal is specific to the 16 bit input message and is given by

$$y_j = s + \delta_j. \tag{2}$$

For sample n in y_j, the detector D outputs a soft decision

$$p[n] = D\left(y_j[n]\right). \tag{3}$$

A hard decision on whether y_j is watermarked is obtained by time-averaging $p[n]$ to obtain a score and thresholding, i.e. $\overline{p} > \theta$ where θ is a decision threshold. In [4], the authors set $\theta = 0.5$ in order to maximize accuracy, defined as the ratio of the True Positive Rate (TPR) to the False Positive Rate (FPR), against various edits. For simplicity, in this paper, we do not consider recovery of the message.

C. Baseline Results using the TIMIT Corpus

For the experiments in this paper, we use two portions of the TIMIT corpus [10] as follows*. The first portion is from the publicly-available TIMIT sample [11] and is composed of 16 speakers each with 10 utterances (typically 3-5 s in duration); the public TIMIT sample is a subset of the TIMIT train set. For each utterance, we generate 10 watermarks using random messages and apply to the corresponding utterance for a total of 1,600 (matched) watermarked signals. This set generated from the first portion constitutes the watermarked signal set we use for our experiments. The second portion uses 160 randomly-chosen speakers from the TIMIT test set each with 10 utterances for a total of 1,600 speech signals; the test set does not have speaker overlap with the train set. This second portion constitutes the unwatermarked signal set we use for our experiments. When watermarks generated from the first portion are applied to utterances from the second portion, we have unmatched watermarked signals.

Figure 1 shows the distributions of detection scores for the 1600 watermarked and unwatermarked speech signals. The detection scores are computed by passing the watermarked/unwatermarked signals through the AudioSeal detector and the average detection score for the respective histogram bins are computed. As described in [4], detection score is near 1 for watermarked signals and near 0 for unwatermarked signals. With a threshold set to $\theta = 0.5$, detection accuracy is maximized to 100% for the TIMIT sample. Using a subset of VoxPopuli, i.e. 10,000 speech signals each of duration 10s and generating 10k watermarked/10k unwatermarked signals from that set for each edit considered including filtering and noising, the authors report a detection accuracy of 96% [4].

In addition, we investigate previously-reported results for two edits which degrade accuracy - white noise and highpass filtering. For our experiments with the TIMIT sample, we adjust the detection threshold to maximize accuracy. Results with white noise addition broadly follow the trends reported in [4] with a drop in accuracy visible only with extreme PESQ degradation. The results are included in Table I. For the High Pass Filter (HPF) edit, the accuracy using TIMIT is broadly similar to the accuracy reported in [4], i.e. 0.50 vs. 0.61. We also show the results for one of our suggested approaches for reducing detection score, speech enhancement in Table I (See Section III-B). Although two recent works [8], [9] have utilized deep learning based speech enhancement to defeat watermarking, our approach demonstrates that a simple and widely available enhancement approach works to this end.

*Although watermarks are intended to be applied to AI-generated speech, we follow the approach in [4] and [6] that use genuine speech as a surrogate for AI-generated speech since the quality is similar.

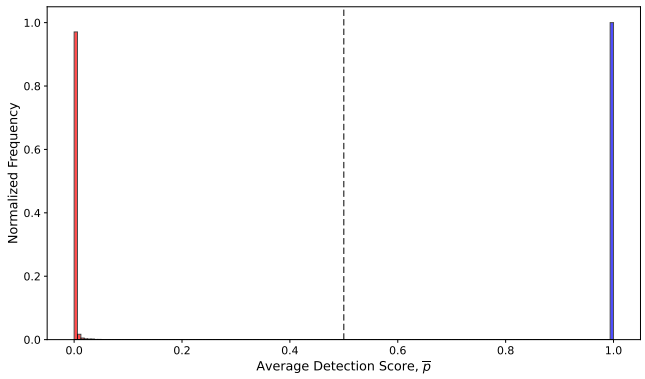


Fig. 1. Distributions of detection scores for 1,600 watermarked speech signals (shown in blue) and 1,600 unwatermarked speech signals (shown in red) from the TIMIT sample corpus. With a threshold, $\boldsymbol{\theta = 0.5}$, detection accuracy is 100%.

In this case the mean detection score drops below 0.5 while preserving the PESQ. In the next sections, we discuss its usage in detail focusing on its impact in compromising AudioSeal when used in combination with other approaches.

D. Threat Model

The adversary's goal is to reduce the detection score for watermarked AI-generated speech or increase the detection score for genuine human speech to defeat the watermarking system. Attacks assume different levels of access to the watermarking system. We describe and label the assumptions here and use the labels as reference throughout the paper.

(AG) Generator access assumption: The attacker has access to the generator as a black box, i.e. they can create a watermark as denoted by (1) but do not have access to the internal workings of the generator. They can pass a speech signal to it and obtain a watermark signal matched to the input signal. This type of access is a realistic assumption in many situations. For example, an online platform providing watermarking as a service would allow users to watermark AI content. It may also be the case that through a leak the generator model is available.

(AD) Detector access assumption: The attacker can obtain the average detection score $\bar{p}$ from the watermarked signal.

TABLE I
DETECTION RESULTS FOR EDITS [4] - WHITE NOISE, HIGHPASS FILTER, AND AUDIBLE NOISE SUPPRESSION, I.E. SPEECH ENHANCEMENT. FOR EACH EDIT, THE THRESHOLD θ IS ADJUSTED TO MAXIMIZE DETECTION ACCURACY. ALSO PROVIDED ARE PESQ OF SPEECH WITH WATERMARK AND DETECTION SCORE $\overline{p}$ WITH THE EDIT.

Edit	θ	**Accuracy (TPR, FPR)**	**PESQ**	$\overline{\mathbf{p}}$
Noise ($\sigma = 0.001$)	0.623	1.00 (1,0)	2.11	0.96
Noise ($\sigma = 0.01$)	0.003	0.933 (0.902,0.036)	1.069	0.143
Noise ($\sigma = 0.05$)	7.03×10^{-5}	0.573 (0.22,0.069)	1.029	0.004
HPF ($f_c = 1500$ Hz)	0	0.5 (1,1)	2.26	0.0008
Speech enhancement	0.147	1.00 (1,0)	3.98	0.47

When the detector is embedded in an online platform (back-end) this condition will be difficult to match. Users may submit signals for watermark testing but do not have access to the detector itself. If the detector is available to everyone to apply to signals this condition is easily met. The authors of AudioSeal [4] suggest that the detector can be made publicly available.

(AM) Message access assumption: The attacker has knowledge of the encoded message m_j within a signal. This 16 bit number may be used to identify a specific AI speech generation service or provider and the same message will be used for many signals. It is likely that m_j is not a secret and is known. The detector provides $p[n]$ and also the message m_j. If the detector is available (AD) any message can be extracted. Thus, assumption (AD) implies (AM).

(AW) Watermark access assumption: The attacker has access to some watermarks. A watermark could be obtained from various sources. For example, the attacker could supply a signal to a watermarking service resulting in a watermarked signal. Then the attacker can subtract the original signal from the watermarked signal to obtain a watermark. They may also have access to a leaked watermark signal. If the attacker has access to the generator (AG) they can produce watermarks. Thus, assumption (AG) implies (AW).

The main attack types we consider is similar to the no-box scenario considered in audiomarkbench [6] with the adversary having close to no access to the watermarking system including detector output. These attacks are considered in the context of realistic application scenarios in Section V.

III. Reducing the Detection Score of a Watermarked Speech Signal

While AudioSeal has shown outstanding resilience to common channel and codec effects, our experiments consider two approaches not reported in [4]. These approaches are designed around modifying the watermarked signal, y_j such that $\overline{p} < \theta$ thus rendering the watermark undetectable. In [6], the authors refer to this approach as "watermark removal" although the watermark is not removed per se but rather the detection score is lowered. In the first approach, we assume access to the watermark generator (Assumption AG) and knowledge of the encoded message (Assumption AM), i.e. white-box setting [6]. In the second approach we do not assume access to the watermark generator or knowledge of the message (while also not requiring any access to the detector output), i.e. no-box setting [6].

A. Reducing Detection Score with Generator Access

In [6], the authors pose an optimization problem for the white-box setting and propose an algorithm to reduce the detection score through repeated iterations. Given the white-box setting, we propose a far simpler method that also achieves the goal of reducing the detection score with similar results but does not require iteration.

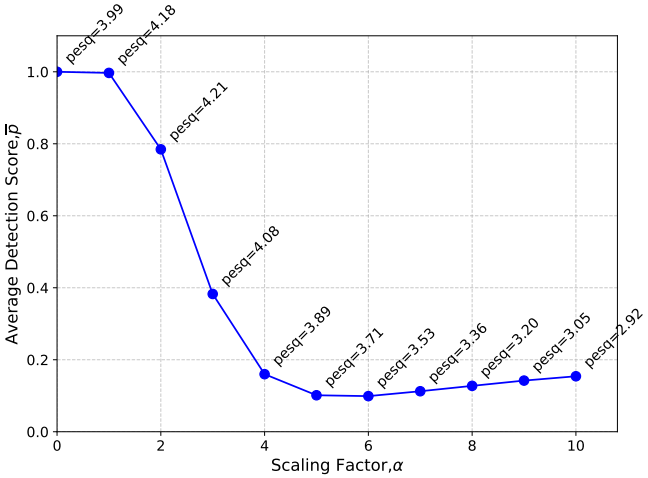


Fig. 2. For the watermark subtraction method, this plot shows the relation between average detection scores for different values of α [see in (5)]. Labeled on each datapoint is the PESQ score of the watermarked speech signal. As an example, for $\alpha = 5$ the detection score falls below 0.1 while having little impact on PESQ.

Assuming generator access (Assumption AG) and knowledge of the encoded message (Assumption AM), we generate a new watermark

$$\delta'_j = G(y_j, m_j) \tag{4}$$

which is an estimate of δ_j and apply it to the watermarked signal as in

$$y'_j = y_j - \alpha\delta'_j \tag{5}$$

where α is a scale factor. The idea is to produce a new watermark δ'_j that is similar to the original watermark δ_j and via subtraction, scale out enough of the original watermark so as to reduce the detection score below the threshold. We assume m_j is known and hence we can generate δ'_j as in (4). The choice of α in (5) should strike a balance between lowering the detection score of y'_j while preserving PESQ, using s as the reference.

Figure 2 shows detection scores for a range of α (0 to 10 in steps of 0.1) with each data point showing the PESQ value. As explained in Section II-C, 1600 watermarked signals are generated and used for this experiment. For each watermarked signal, at each α, the subtraction attack is carried out and the attacked watermarked signal is passed through the detector. The detection scores and PESQ are computed and the average value is calculated over the 1600 signals. For $\alpha = 3$, we see the average detection score falls below 0.5 while not reducing PESQ (compared to $\alpha = 0$). Furthermore, for $\alpha = 5$, we see the average detection score falls below 0.1 while slightly reducing PESQ (< 0.3 points when compared to $\alpha = 0$). These results suggest if the watermark generator is accessible, this approach is a very simple and viable attack. The histogram of detection scores for $\alpha = 5$ is shown in Figure 3. Although our approach could be implemented iteratively as in [6] for possibly better performance, we have shown a non-iterative approach that can use a fixed value for α (see Figure 3).

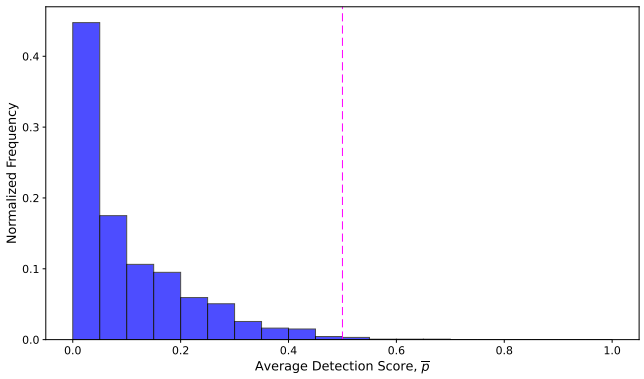


Fig. 3. Histogram of the detection scores for the watermark subtraction method with $\alpha = 5$. Compared with Figure 1, this method allows an adversary to shift the score distribution resulting in lower average detection score.

B. Reducing Detection Score without Generator Access

In [6], the authors pose a different optimization problem for the black-box setting and investigate the Hop Skip Jump and Square attacks to solve the problem. For both these attacks, the optimization is run over 10,000 iterations. We propose a far simpler method in the no-box setting that in many cases also achieves the goal of reducing the detection score but does not require iteration. As we will show in Section V, this result combined with our results in increasing detection score, i.e. referred to as "watermark forgery" in [6] is sufficient to compromise AudioSeal.

We do not assume access to the watermark generator and view (2) as an additive noise model. With this view, we apply a common speech enhancement method, audible noise suppression [12] directly to y_j. Figure 4 shows the histogram of detection scores after speech enhancement on the 1,600 watermarked signals. The watermarked signals used are generated as explained in Section II-C. Speech enhancement is done for each of the 1600 watermarked signals used, and the enhanced watermarked signal is passed through the detector to obtain the detection scores. With a detection threshold $\theta = 0.5$, approximately 62% of enhanced signals have a watermark which is no longer detectable.

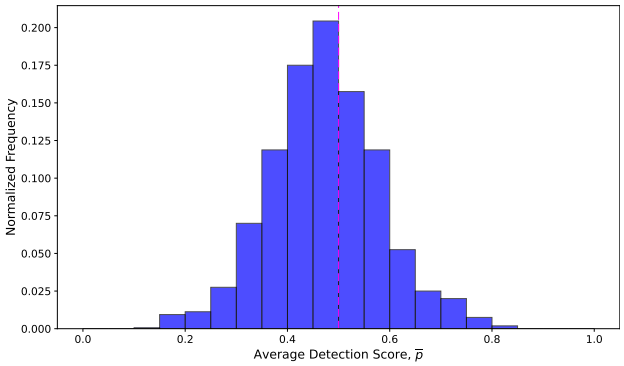


Fig. 4. Histogram of detection scores after speech enhancement (audible noise suppression). With a detection threshold, $\boldsymbol{\theta = 0.5}$, 61.69% of enhanced signals have a watermark which is no longer detectable.

IV. Increasing the Detection Score of Genuine Speech with an Unmatched Watermark

We now consider applying δ_j generated as in (1) but to a speech signal u from a different speaker to generate the watermarked signal

$$x = u + \delta_j. \tag{6}$$

In the case of (2), we say the watermark is matched to the speech signal s whereas in (6), the watermark is not matched to the speech signal u, i.e. unmatched watermark. In this approach, a watermark could be obtained from various sources (Assumption AW), i.e. in the wild and imposed on genuine human speech even though the watermark is generated for a different speaker. Use of an unmatched watermark obtained without direct access to the generator has not been previously considered [6] and is a realistic scenario given that watermarks generated as in (1) could be made available.

If the unmatched watermark is shorter than the speech signal, it is repeated to the length of the speech signal, and for the other way round where the watermark is longer than the speech signal, the watermark is trimmed to the length of the speech signal. We consider two experiments where the watermark is selected randomly from the first portion of the speech data (as explained in Section II-C) and applied to a speech signal in the second portion of the speech data. In the first experiment (Section IV-A), we assume no access to the detector while creating the unmatched watermarked signals, while in the second experiment (Section IV-B), we assume access to the detector output i.e. we can obtain the average detection score $\bar{p}$ from the watermarked signal (Assumption AD).

A. Increasing Detection Score without Detector Access

In the first experiment, we assume no access to the detector, making this scenario similar to the no-box setting in audiomarkbench [6]. Here, we use the PESQ measure as a surrogate for the detection score. This assumption is viable because in the ideal case having the presence of a watermark (embedded) in the speech signal, the PESQ is highest since the watermark will be matched to the speech signal. This case will also have the highest detection score which as we have shown in Section II-C is close to 1. Extending this to the case of unmatched watermarking, the more matched the speech signal and the watermark are, the higher will be the PESQ and the detection score. Hence, we devise a two stage strategy for unmatched watermark addition based on this assumption. In the first stage, we select 10 random watermarks initially. Then we add each of these 10 unmatched watermarks to the speech signal under consideration as in (6) and compute the PESQ in each case. We choose the unmatched watermark which gives the best PESQ score among these. In the second stage, for that particular selected unmatched watermark, we then proceed as follows. For a given PESQ, we scale this unmatched watermark appropriately and add it to the speech signal so that the actual PESQ is within ± 0.1 of our target PESQ. 1600 unmatched watermarked signals are created in

this manner. We then determine the detection scores for these unmatched watermarked signals as a function of PESQ. The results are shown in Figure 5 where the points indicate the measured average PESQ values and average detection scores.

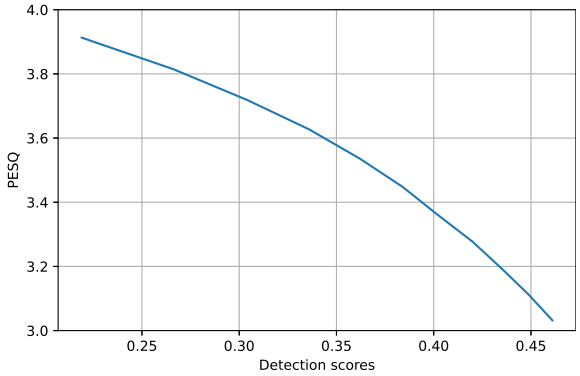


Fig. 5. Average PESQ vs average detection scores for unmatched watermarked signals.

It is observed that as the scaling factor for watermark increases, the PESQ score decreases and the detection score increases. Hence the PESQ and detection scores are inversely related. This relates to the impact of scaling on watermark strength for a particular watermark and does not contradict our assumption of higher PESQ being indicative of better watermark match to a speech signal and hence higher detection probability among a selection of unmatched watermarks.

With a given PESQ of 3.4 (tolerance = 0.1), which maintains perceptual quality, it is possible to increase the detection score to 0.4 on average without perceptual distortion. The histogram of the detection scores is shown in Figure 6. We see that the detection score can be raised beyond the default detection threshold ($\theta = 0.5$) for 22.38% of the target speakers' utterances implying that even without having access to the detector, it is possible to use unmatched watermarked signals that can increase the false positives of the detector.

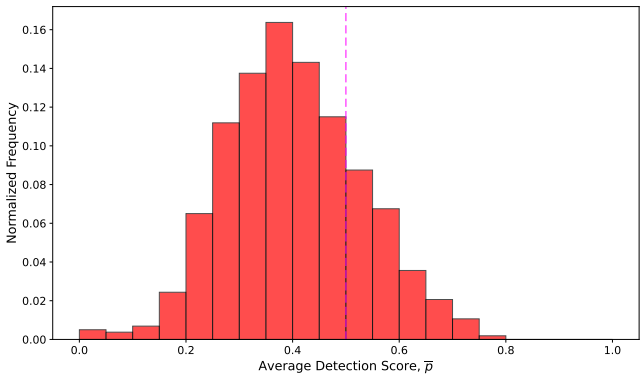


Fig. 6. Histogram of detection scores on selecting an unmatched watermark from ten randomly-chosen unmatched watermarks based on best PESQ based match to a speech signal. A target PESQ of 3.4 is used (with tolerance of 0.1). The histogram shows 22.38% of detection scores increase beyond a 0.5 threshold.

B. Increasing Detection Score with Detector Access

In [6], the authors consider "watermark forgery" in order to increase the detection score and use the same optimization approach for the watermark removal, black-box case. Given the black-box setting, we propose a simpler method that in many cases also achieves the goal of reducing the detection score but does not require iterative optimization.

In this experiment, 10 random watermarks are selected as in the aforementioned experiment (Section IV-A) and each applied to the speech signal. Here, we assume access to the detector (Assumption AD). For each watermark, scaling is performed similar as in the second stage of the previous approach (Section IV-A) so as to match the given PESQ within a tolerance of ± 0.1. The unmatched watermarked signal with the highest detection score from among the ten signals so generated is chosen. This process is repeated to generate 1600 unmatched watermarked speech signals. Figure 7 shows the histogram of detection scores for this set. The results indicate that detection scores surpass the threshold ($\theta = 0.5$) for 60.06% of the target speakers' utterances. This demonstrates a significant increase in false positives by the detector compared to the previous experiment. Further increasing the detection scores may be possible by applying a larger number, i.e. more than 10 of unmatched watermarks to the speech signal and choosing the signal with the highest detection score.

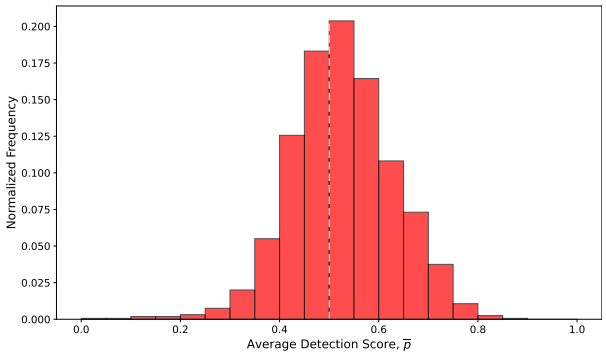


Fig. 7. Histogram of detection scores when choosing the watermark with the highest detection score from among 10 randomly-chosen unmatched watermarks and combining these with the speech signal to achieve a target PESQ of 3.4 (with tolerance of 0.1). The histogram shows 60.06% of detection scores increase beyond the default threshold of 0.5.

V. Discussion

A. General Findings

In Section III, we demonstrated how to decrease the detection score of the watermarked signal with and without generator access. As developed in Section III-A, with generator access (Assumption AG) it is possible to lower the detection score considerably using a simple subtraction attack, with the average detection score falling below 0.1. As developed in Section III-B, without generator access (which is a general approach) it is possible to lower the detection score considerably using speech enhancement with audible noise suppression. In

this case, we were able to lower the average detection score below the 0.5 threshold.

In Section IV, we demonstrated how to increase the detection score of a speech signal that uses an unmatched watermark (Assumption AW). As developed in Section IV-A, without detector access we can obtain detection scores greater than the 0.5 threshold for about 22.38% of signals. As developed in Section IV-B, with detector access (Assumption AD) the proportion increases to 60.06%.

Given the methods we have presented to increase and decrease detection scores, we consider Figure 8 which shows overlapping histograms of the approach described in Section III-B (which does not assume generator access) and Section IV-A (which does not assume detector access). This points to the existence of a practical scenario where the accurate detection of watermarked/unwatermarked signals is difficult due to the overlap in the distributions of detection scores.

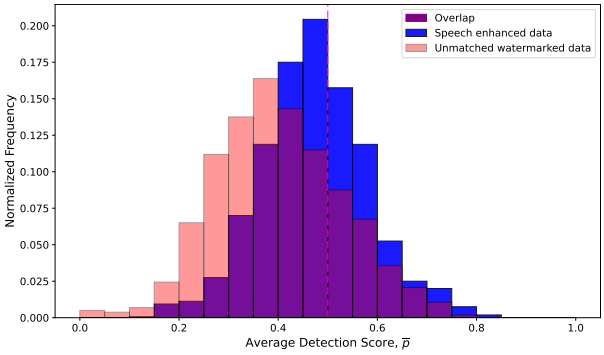


Fig. 8. Histograms of detection scores for the unmatched watermark approach which increases detection scores and speech enhancement approach which decreases scores for watermarked signals. The overlapping detection scores (shown in purple color) illustrate the practical challenges with AudioSeal as detection can be manipulated.

B. Application Scenarios

Consider now a large social media platform where users share audio (and/or video) content. The aim would be to protect users from deep fakes using watermarks while ensuring genuine, human speech is not watermarked. Compliance would guarantee that generated content contains an embedded watermark which is easily identified and thus a user would be provided with a warning that particular content is AI-generated.

We now need to consider that adversaries are also present and they may compromise watermarked signals as described in Section III. We assume a more general method of reducing detection scores without generator access using the speech enhancement approach from Section III-B. The assumed capability of the adversary is similar to that of the no-box scenario presented in audiomarkbench [6]. Thus, the social media platform now also contains compromised watermarks. Furthermore, adversaries may also use unmatched watermarks on genuine speech using the method described in Section IV-A, i.e. we do not assume the adversary has access to the social media platform's detector. Thus, the social media platform may contain AI-generated speech with (undetectable) compromised watermarks and genuine speech with (detectable) unmatched watermarks thus creating both false negatives and positives. Finally, we also assume that a social media platform may also contain non-speech signals e.g. music, animal sounds, background noise, machine sounds, etc. We only consider in this scenario our Assumption (AW); the attacker has access to some unmatched watermarks.

In summary, we consider a social media platform hosting many types of signals:

A (Unwatermarked) speech signals
B Watermarked speech signals
C Compromised watermarked speech signals
D Unmatched watermarked speech signals
E (Unwatermarked) non-speech signals

We assume that content on the platform is composed of sets A,B,C,D, and E in various proportions. By defining these proportions we can now analyze different scenarios and the tuple S defines the content mix. For example, $S = (0.90, 0.06, 0.2, 0.2, 0.0)$ describes a scenario where 90% of content is speech signals, 6% of content is watermarked speech signals, and the remainder are composed of 2% each of compromised watermarked signals and unmatched watermarked signals, and no non-speech signals.

The usefulness of AudioSeal for a scenario can be evaluated by analyzing the resulting ROC curve. The balance between False Positive Rate (FPR) and True Positive Rate (TPR) can be shown and we can determine the EER as metric of general usability of the system. The FPR describes how often a genuine signal is marked as deep fake (i.e. AI generated); the TPR describes how often a deep fake is identified as such. While one case may be considered more problematic than the other, both situations are balanced at the EER.

For each scenario in this section, we use a total of 1600 signals with proportions described by the tuple. Sets A and B are generated using the TIMIT dataset as described in Section II-C. Set C is drawn from the set in Section III-B. Set D is drawn from the set in Section IV-A with the unmatched watermarked signals generated satisfying a PESQ of 3.4 (tolerance = 0.1). Set E is selected from among non-speech signal categories available from the freesounds platform [13].

Scenario 1 - Equal Distribution: For the first scenario, we consider $S_1 = (0.25, 0.25, 0.25, 0.25, 0.0)$ where an equal amount of Sets A, B, C, and D are present. This scenario would put considerable effort on the adversary as they would have to generate considerable compromised watermarked signals (Set C) and unmatched watermarked speech signals (Set D). While Sets C and D may be easily generated, an adversary may be limited by the amount of signals that can be uploaded from one source or the platform may require payment for content upload and this may pose a constraint. For this scenario, the ROC curve is shown in Figure 9 and has an $\text{EER}_{S1} = 0.188$. In practice, this would mean that when assuming S_1, the EER point would mean that 18.8% of deep fakes are not recognized or missed and that a user receives an erroneous deep fake

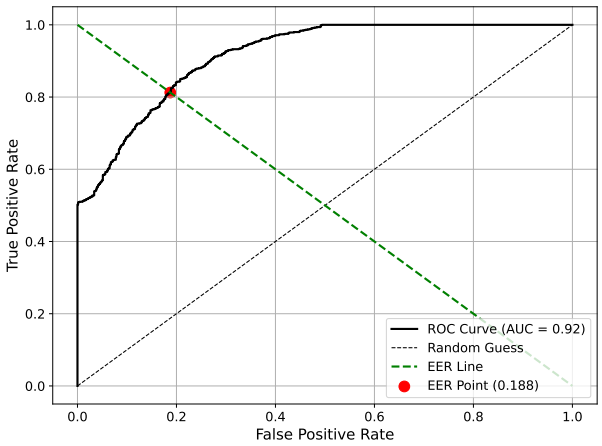


Fig. 9. Scenario 1 with $\mathbf{S}_1$ = **(0.25,0.25,0.25,0.25,0.0)**, i.e. equal proportions of unwatermarked signals, watermarked signals, compromised watermarked, and unmatched watermarked signals. With S_1, the EER = 18.8% is unacceptable.

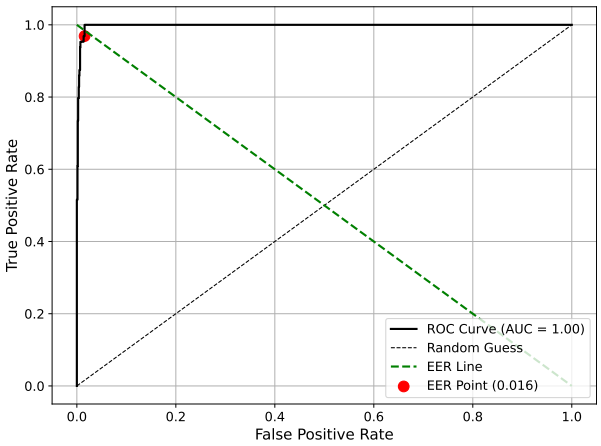


Fig. 10. Scenario 2 with $\mathbf{S}_2$ = **(0.94,0.02,0.02,0.02,0.0)**, i.e. higher proportion of unwatermarked (genuine) speech signals. This is a realistic scenario having EER=1.6%.

warning for 18.8% of genuine content. Clearly, this result is likely unacceptable since many deep fakes are missed and there are many false alarms.

Scenario 2 - Biased Unwatermarked (Genuine) Speech: For the second scenario, we consider $S_2 = (0.94, 0.02, 0.02, 0.02, 0.0)$ which may be considered more realistic. The majority of content (94%) are unwatermarked speech signals (Set A) while 2% is watermarked speech signals (Set B). We also consider 2% each, compromised watermarked speech signals (Set C) and unmatched watermarked signals (Set D). For this scenario the ROC curve is shown in Figure 10. and has an $\text{EER}_{S2} = 0.016$.

While the EER is lower, compared to Scenario 1, it may still be unacceptable, i.e. one percent of deep fakes are not recognized and one percent of genuine content is flagged as a deep fake. The average length of a TikTok video has a duration of 42.7 seconds [14]. The average TikTok user spends 58 minutes per day on this platform, consuming 81 standard videos per day. Assuming S_2, approximately once per day a user would be falsely warned about a deep fake and approximately once per day view a deep fake which they believe is genuine content. We believe these error rates are not acceptable, in general.

Scenario 3 - Non-Speech Signals: Finally for the third scenario, we consider $S_3 = (0.47, 0.02, 0.02, 0.02, 0.47)$ which is similar to scenario S_2 but instead of assuming 94% unwatermarked signals we split this evenly into unwatermarked speech signals (Set A) and non-speech signals (Set E). This represents a social media platform with both speech content and non-speech (music or environment sounds) content. For this scenario, the Receiver Operating Characteristic (ROC) curve is shown in Figure 11 and has an $\text{EER}_{S3} = 0.019$. We observed that inclusion of (unwatermarked) non-speech signals minorly change the ROC curve from Scenario 2 thus concerns for Scenario 2 apply to Scenario 3.

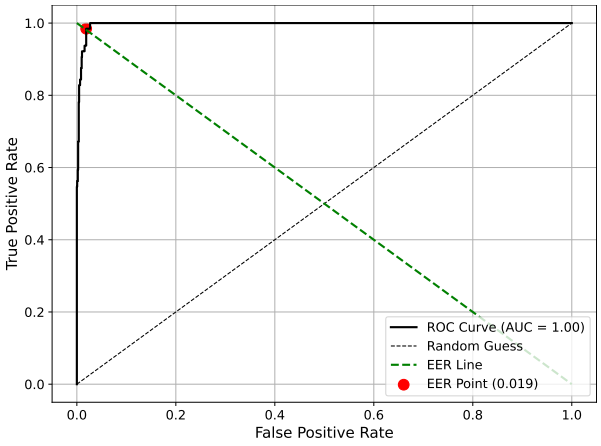


Fig. 11. Scenario 3 with $\mathbf{S}_3$ = **(0.47,0.02,0.02,0.02,0.47)**, i.e. inclusion of non-speech signals in the content.

VI. Related Work

Previously, watermarking has mainly focused on copyright protection and Digital Rights Management (DRM) [15]. Audio watermarking is an established field and focus has been on imperceptibility and resilience to transformations in the audio processing chain or by deliberate modifications [16].

Early works were focused on developing watermarks for image data. Drawing inspiration from spread spectrum techniques used in communication systems, Cox et al. [17], [18] introduced a spread spectrum based watermarking technique. The authors demonstrated the viability of this approach on digital images. Boney et al. [19] developed a temporal and frequency masking based approach for audio signals, with a more robust version developed by Swanson et al. [20]. The spread spectrum based approach to audio signals was developed by Kirowski et al. [21], [22]. Transform domain audio watermarking approaches based on Discrete Cosine Transform (DCT) [23] and Discrete Wavelet Transform (DWT) were developed [24].

With the increase in computational capabilities, deep learning based methods came into use for watermarking algorithms [25]. Kandi et al. [26] developed a Convolutional Neural Network (CNN) based watermarking scheme for image data. With the increasing ability of generative AI for realistic fake speech generation even with limited training data such as zero shot voice synthesis [27], which could beat anti-spoofing approaches, watermarking of AI generated speech became a necessity. Audio watermarking strategies using adversarial deep learning networks with an embedder/encoder - decoder type architecture were developed [28]. Such networks work in an adversarial framework with the encoder attempting to embed imperceptible watermarks and the decoder/detector (adversary) attempting to detect the watermark, with the embedder working similarly to autoencoders. Liu et al. [29] developed a framework robust to audio re-recording attacks. Chen et al. [5] developed wavmark which introduced an invertible neural network based framework for audio watermarking, bringing in curriculum based learning and weighted attack handling. AudioSeal [4], improved upon the existing state of the art and was the first approach solely focussed on AI speech watermarking, providing watermark detection at a sample level.

Recent works have looked at robustness of watermarking approaches to attack scenarios. Audiomarkbench [6] looks at attacks in no-box, black-box, and white-box settings. The authors used human speech data from Librispeech [30] and the common voice dataset [31] for analysis. O'Reilly et al. [9] considered a set of transformations including standard signal processing techniques such as pitch shift and reverberation, codecs including neural codecs such as DAC (descript audio codec) and neural vocoders. Lopez et al. [8] and O'Reilly et al. [9] considered the use of speech enhancement using deep neural networks for watermark removal. Human speech data available from DAPS [32] and TIMIT [10] datasets has been used for the experiments. The robustness of watermarking approaches to neural codecs has been studied in depth by Ozer et al. [33].

VII. Conclusions

In this paper, we have investigated a state-of-the-art watermarking system, AudioSeal. In our investigation, we have shown two methods which may be used to reduce detection scores of watermarked signals. The first method requires access to the watermark generator and simply subtracts an estimate of the watermark from the watermarked signal. This results in the average detection score falling well below the detection threshold while maintaining perceptual quality. The second method does not require generator access and uses a common speech enhancement technique. This results in approximately 62% of watermarked signals no longer being detectable.

Furthermore, we have shown two methods which may be used to increase detection scores of unwatermarked signals by adding an unmatched watermark. The first method does not require access to the watermark detector and we demonstrate 22.38% of detection scores exceed the detection threshold. In the second method, which requires access to the detector, we demonstrate 60% of detection scores exceed the detection threshold.

The methods presented allow an adversary to decrease detection scores of watermarked signals or increase detection scores of unwatermarked signals, causing overlap in the score distributions making accurate detection difficult. To illustrate the impracticality of watermarking AI-generated speech using AudioSeal, we consider a hypothetical application scenario on a social media platform and found that the performance results are unacceptable.

References

[1] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 33, pp. 705–718, 2025.

[2] P. Cheng and U. Roedig, "Personal voice assistant security and privacy—a survey," *Proc. IEEE*, vol. 110, no. 4, pp. 476–507, 2022.

[3] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: generation, detection, and applications," *Int. J. Multimed. Info. Retr.*, vol. 11, no. 3, pp. 219–289, 2022.

[4] R. San Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon, and T. Tran, "Proactive detection of voice cloning with localized watermarking," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[5] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "Wavmark: Watermarking for audio generation," *arXiv preprint arXiv:2308.12770*, 2023.

[6] H. Liu, M. Guo, Z. Jiang, L. Wang, and N. Gong, "Audiomarkbench: Benchmarking robustness of audio watermarking," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 52 241–52 265, 2024.

[7] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2001, pp. 749–752. [Online]. Available: https://doi.org/10.1109/ICASSP.2001.941023

[8] Á. López López, Á. M. Gómez García, E. Roselló Casado *et al.*, "Speech watermarking removal by DNN-based speech enhancement attacks," *in Proc. IberSPEECH*, 2024.

[9] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Deep audio watermarks are shallow: Limitations of Post-Hoc watermarking techniques for speech," *CoRR*, vol. abs/2504.10782, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2504.10782

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, p. 16, 1988.

[11] Kaggle, "TIMIT Corpus Sample (LDC93S1)," 2005. [Online]. Available: https://www.kaggle.com/datasets/nltkdata/timitcorpus

[12] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[13] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACM Int. Conf. Multimed. (ACMMM)*, 2013, pp. 411–412.

[14] Statista, "Tiktok video duration 2024," 2024, accessed: April 1, 2025. [Online]. Available: https://www.statista.com/statistics/1485205/tiktok-video-duration/

[15] F. Hartung and F. Ramme, "Digital rights management and watermarking of multimedia content for m-commerce applications," *IEEE Commun. Mag.*, vol. 38, no. 11, pp. 78–84, 2000.

[16] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, "Twenty years of digital audio watermarking—a comprehensive review," *Signal Process.*, vol. 128, pp. 222–242, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168416300263

[17] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," in *Proc. IEEE Int. Conf. Image Proc. (ICIP))*, vol. 3. IEEE, 1996, pp. 243–246.

[18] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, 1997.

[19] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems (ICMCS)*. IEEE, 1996, pp. 473–480. [Online]. Available: https://doi.org/10.1109/MMCS.1996.535015

[20] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Process.*, vol. 66, no. 3, pp. 337–355, 1998.

[21] D. Kirovski and H. Malvar, "Robust spread-spectrum audio watermarking," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, vol. 3. IEEE, 2001, pp. 1345–1348.

[22] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1020–1033, 2003.

[23] Y. Yan, H. Rong, and X. Mintao, "A novel audio watermarking algorithm for copyright protection based on DCT domain," in *Int. Symp. Elect. Commerce and Security (ISECS)*, vol. 1. IEEE, 2009, pp. 184–188.

[24] X.-Y. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4835–4840, 2006.

[25] J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. Molloy, "Protecting intellectual property of deep neural networks with watermarking," in *Proc. ASIACCS 2018.* ACM, 2018, pp. 159–172.

[26] H. Kandi, D. Mishra, and S. R. S. Gorthi, "Exploring the learning capabilities of convolutional neural networks for robust image watermarking," *Comput. Secur.*, vol. 65, pp. 247–268, 2017.

[27] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2022, pp. 2709–2720.

[28] K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski, "Robust speech watermarking by a jointly trained embedder and detector using a dnn," *Dig. Sig. Process.*, vol. 122, p. 103381, 2022.

[29] C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang, and N. Yu, "Dear: A deep-learning-based audio re-recording resilient watermarking," in *Proc. AAAI Conf. Art. Intell.*, vol. 37, no. 11, 2023, pp. 13 201–13 209.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[31] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. Lang. Resources Eval. Conf. (LREC)*. European Language Resources Association, 2020. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[32] 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 Process. Lett.*, vol. 22, no. 8, pp. 1006–1010, 2015. [Online]. Available: https://doi.org/10.1109/LSP.2014.2379648

[33] Y. Özer, W. Choi, J. Serrà, M. K. Singh, W.-H. Liao, and Y. Mitsufuji, "A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs?" *arXiv preprint arXiv:2505.19663*, 2025.