PRODUCTIONIZING AUDIO WATERMARKING FOR SHORT-FORM VIDEO

Elias Kokkinis*, **Dimitris Koutsaidis***, **Elias Lumpert***, **Stergios Terpinas***, **George Tzoupis***
*Equal contribution, alphabetical order.
Meta Inc.
1 Hacker Way, Menlo Park, CA 94025, USA
{eliaskokkinis,dkoutsaidis,eliaslumpert,sterpinas,gtzoupis}@meta.com

ABSTRACT

In this work, the application of audio watermarking in large scale short-form video platforms is explored by addressing challenges particularly focusing on minimizing watermark audibility while maximizing detectability. Experimental results are presented, discussing approaches to improve imperceptibility such as using a mixing gain for the watermark signal and applying it only on speech segments. The experiments also examine the impact of multiple audio encodings and music mixing on watermark detectability, proposing solutions to enhance robustness.

1 INTRODUCTION

Within video-centric online platforms, there is an increasing need for effective verification of AI generated content [Murphy et al. (2024), Zhao et al. (2024)]. Audio watermarking has emerged as a promising solution to address this challenge. By embedding an inaudible signal into the audio content, audio watermarking enables the identification of generated content [Roman et al. (2024)]. The productionization of audio watermarking at-scale poses further challenges.

The problem of audio watermarking incorporates two equally important dimensions: minimizing watermark audibility and maximizing detectability. The two tasks are not independent to each other and optimizing only for one could degrade the other. Thus, we need to optimize for both concurrently to reach an optimal behaviour. Recent developments in the field showed important improvements in both directions. For example, AudioSeal achieves better imperceptibility and higher detectability than previous models (Roman et al. (2024)).

While there is increasing interest in this topic [Chen et al. (2024), Liu et al. (2023)], its practical application in industry settings can be further explored. In the complex environment of at-scale video production and delivery, the watermarked signal might be processed with different processing steps over time as it passes through various system components. When a watermark has been applied to an audio signal, any further processing is considered an attack to the watermark and can reduce the detection rate or render the watermark undetectable. In production systems it's not uncommon for an audio signal to go through multiple encoding steps, mixed with other signals, etc. This increases the requirements for watermark robustness in practice. Furthermore, audio watermarking lacks standardized guidelines and regulations [Uddin et al. (2024)]. In the absence of clear laws and industry standards the development and integration of watermarking remains at the responsibility of the platforms.

In this paper, we explore the challenges of audio watermarking integration in at scale video platforms and propose solutions. In the audibility experiment section (2.1), we discuss two approaches to reduce the perceived quality degradation after watermarking and present their performance in terms of detectability and audibility. In the following two experiment sections, we explore concurrent attacks such as multiple encodings (2.2) and mixed-in music (2.3) and their effect on watermark detectability.

2 Experiments

Watermark generation was performed using an internal AudioSeal checkpoint [Roman et al. (2024)], fine-tuned for short-form video at 48kHz, with a workaround operating at 16kHz as described in 2.2. A preliminary internal checkpoint trained at 16kHz was used in 2.1.2. The watermark mixing gain was set to 0.5, except in 2.1.1 where various gains were investigated. Detectability was aggregated using the maximum detector output probability over the entire signal.

Encoding in 2.2 and 2.1.1 refers to a two-stage process: AAC-LC (1st pass) followed by HE-AAC transcoding (2nd pass) [ISO (2019)]. The same two-stage encoding process was used in the section 2.1 as well.

We used ViSQOL Chinen et al. (2020) to evaluate the audio quality of the watermarked signal compared to the non-watermarked reference. The metric outputs a value between 1 and 5, where a higher value indicates better perceptual score. Detectability results are reported in terms of accuracy (Acc.), obtained for the threshold that gives best accuracy, and area under the ROC curve (AUC), obtained on a balanced set of samples.

We used a dataset consisting of 100 synthesized speech signals generated from Harvard Sentences Rothauser (1969) and 100 music signals from MedleyDB (Bittner et al., 2014).

2.1 Audibility

Even though there is much work done to make the watermark imperceptible while being detectable, we have observed that mixing the watermark signal as is with the original audio can result in audible artifacts (Fig. 1). This issue is more prominent in silent parts of the original audio. In an effort to improve the perceived audio quality we tried to make the watermark as transparent as possible through two different post-processing approaches:

(a) mixing the watermark $w[n]$ with the original audio $x[n]$ using a gain $g \in (0, 1]$ to reduce its overall audibility (Eq. 1)

$$y[n] = x[n] + g \cdot w[n] \tag{1}$$

(b) leveraging a voice activity detection (VAD) mask $v[n] \in \{0, 1\}$ to apply the watermark only on active speech parts while keeping the silent parts intact (Eq. 2)

$$y[n] = x[n] + g \cdot v[n] \cdot w[n] \tag{2}$$

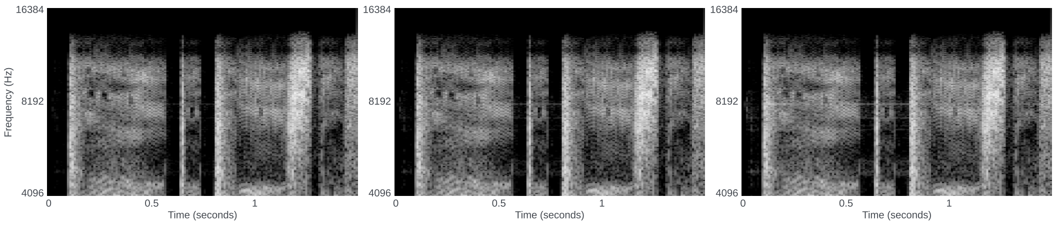

Figure 1: Spectrograms of watermarked files. Focusing near the 8kHz frequency region we can see how the watermark becomes less audible as the mixing gain g decreases.

2.1.1 Mixing Gain for Watermark

We have tested 3 different options for the mixing gain $g \in \{0.25, 0.5, 1.0\}$. We have verified the expected behaviour where the watermark is less audible when the gain is lower. Detectability remains at similar levels (Tab. 1). When the output audio $y[n]$ is encoded, detectability accuracy is reduced for lower g values which is obvious from the results with $g = 0.25$. A good trade-off can be achieved when using $g = 0.5$, where the outputs are robust to detectability degradation from encodings while also making the watermark less perceptible improving the users' experience. ViSQOL increments are within the 95% confidence intervals, leaving it uncertain whether the metric was able to capture the audibility improvement.

Table 1: Accuracy, AUC and ViSQOL (Median 95% CI) values for the various mixing gain values of g.

Gain	Acc.	AUC	ViSQOL
0.25	73.50%	80.25%	4.60 [4.58, 4.61]
0.50	90.50%	96.69%	4.60 [4.58, 4.61]
1.00	99.50%	99.99%	4.59 [4.58, 4.60]

Table 2: Accuracy, AUC and ViSQOL (Median 95% CI) values for applying watermark everywhere versus application on active speech.

Config	Acc.	AUC	ViSQOL
All	79.50%	86.84%	4.59 [4.57, 4.61]
Speech	78.50%	85.86%	4.60 [4.58, 4.61]

2.1.2 Watermark on Active Speech

Using the VAD mask $v[n]$ to selectively apply the watermark only on active speech parts further improves the perceptual quality of the output audio. The watermark signal is easier to perceive on silent parts of the original audio where the content has lower volume. By focusing only on the speech parts for applying watermark (Fig. 2), we ensure that it is more transparent to the user while also being able to be detected in the most important parts of the audio file, i.e. parts with generated speech. Our experiments show that using this approach can further improve audibility while keeping the detection at similar levels (Tab. 2).

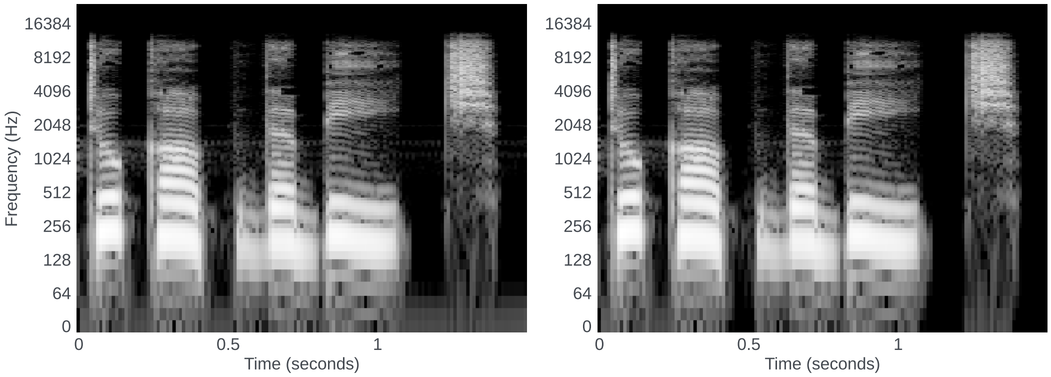

Figure 2: Spectrograms of watermarked files. Left: The watermark is applied everywhere and is particularly evident in the silent parts after first second. Right: The watermark is only applied on active speech areas.

2.2 Encodings as an attack

In production systems audio is often encoded to reduce file size and improve transmission efficiency. This encoding process can compromise the audio signal, particularly in high frequency ranges. Certain encoding strategies, such as those employing psycho-acoustic modeling, can lead to substantial loss of information in these ranges, as they prioritize human perception over accurate representation of high-frequency components [ISO (2019)].

Our experiments revealed that watermarking generation would produce a watermark signal in a frequency range affected by encoding schemes typically used in the industry [Fig. 3]. The result was a significant drop in detection probability for encoded watermarked files, especially for multiple encoding passes and certain encoding schemes. The same encoding schemes produce higher detection probabilities in non-watermarked files (in other words false positive detections), ultimately resulting in low detection accuracy [Tab. 3]. To address this issue, we employed a workaround using resampling and calculating the watermark at a lower operating sample rate. This approach improved the robustness of the watermarking approach in regards to audio encoding [Tab. 3].

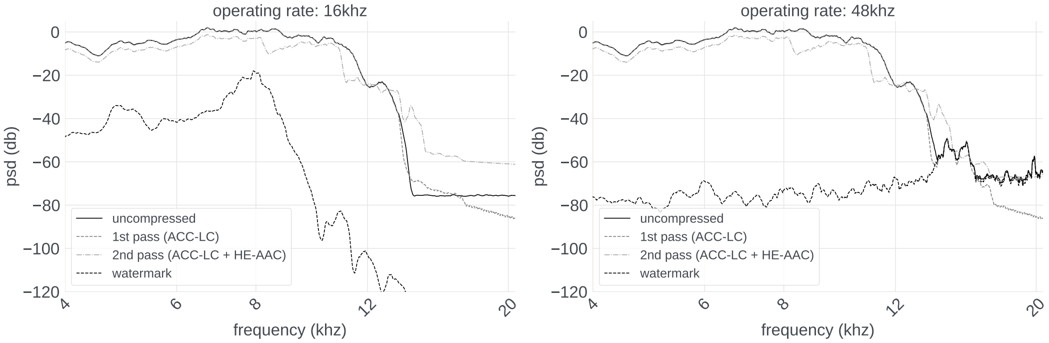

Figure 3: Power Spectral Density (PSD) calculated using a short-time Fourier transform with 2048 samples, 512 sample hop length, averaged across the whole file and calculated at a 48 kHz sample rate. Each plot contains the watermarked synthetic speech encoded with different encoding parameters, with corresponding watermark. Left and right plots show watermark generation at different operating sample rates (16 kHz and 48 kHz).

2.3 Music as an attack

Short-form videos on social media platforms exhibit a wide range of audio-visual characteristics, making it challenging to develop a robust watermarking system. Although AudioSeal generalizes and is robust to such diverse audio content (Roman et al., 2024), the effect of mixing music with speech has not being thoroughly evaluated as an attack in the context of a production system where multiple encoding passes also occur. Prior work indicates that watermarking both the speech and non-vocal music signals before mixing, leads to higher detection accuracy (Roman et al., 2024).

For the case of our production system we evaluate two configurations of the music mixing attack, based on whether or not the music signal is mixed before watermarking and double encoding. The two configurations are showcased in Fig. 4. Both configurations were tested on 3 different Signal-to-Noise Ratios (SNR) (5dB, 10dB, and 15dB). Results presented in Tab. 4, indicate that mixing the music signal before applying the watermark and the two encodings should be preferred.

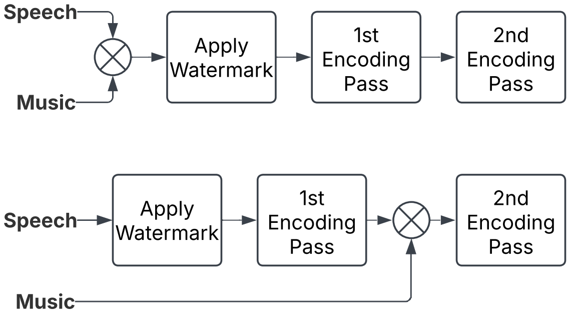

Figure 4: (Top) Music mixing before applying Watermark and encodings. (Below) Music mixing after watermark and 1st pass encoding is applied to the speech signal.

3 Conclusions

In this work, the application of audio watermarking in large scale short-form video platforms is explored by addressing challenges particularly focusing on minimizing watermark audibility and maximizing detectability. Experiments showed how post-processing on the watermark helps im-

Table 3: Results of the Encodings as an attack experiment.

Encoding	Op. rate	Acc.	AUC
n/a	16 kHz	100.00%	100.00%
n/a	48 kHz	100.00%	100.00%
1st pass	16 kHz	**100.00%**	**100.00%**
1st pass	48 kHz	99.50%	99.98%
2nd pass	16 kHz	**90.50%**	**96.69%**
2nd pass	48 kHz	53.50%	50.36%

Table 4: Results of the Music as an attack experiment.

Mix Config	SNR	Acc.	AUC
Before WM	5dB	**79.95%**	**85.34%**
After WM	5dB	70.50%	78.54%
Before WM	10dB	**79.00%**	**83.98%**
After WM	10dB	73.50%	82.15%
Before WM	15dB	**79.00%**	**87.30%**
After WM	15dB	75.50%	82.17%
No Music	-	89.00%	96.69%

prove perceptual quality while ensuring that the generated speech content can still be detected with high probability.

More specifically, using a mixing gain for the watermark signal and applying the watermark only on active speech parts makes it more transparent to the user without degrading the detection performance. Furthermore, since the industry often employs various encoding formats and multiple encoding passes, high frequency content of any watermark is vulnerable to be lost during encoding. By focusing the frequency content of the watermark in the lower range, we ensure that no information is lost and retain high detectability in the face of encodings. The evaluation of applying watermark on speech and music mixtures showed the significance of considering the signal path in the production system and in our case the mixing of music before watermark generation was preferable. Since music is ubiquitous in user uploaded content, it is important to ensure high detectability in the face of such an attack to the watermark.

Assessing the audibility of a watermark proved difficult and improving upon existing objective quality metrics could be beneficial. We can see that ViSQOL results show small arithmetic differences which implies that the metric is not very sensitive to the distortion introduced by the watermark. Finally, we believe that introducing specific guidelines for setting guardrails on the performance of a watermark detector will aid in setting informed legal requirements for a watermarking model that is used in a production system.

References

ISO/IEC 14496-3:2019 - Information technology – Coding of audio-visual objects – Part 3: Audio, 2019.

Rachel Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb audio: A dataset of multitrack audio for music research, October 2014. URL https://doi.org/10.5281/zenodo.1649325.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. Wavmark: Watermarking for audio generation, 2024. URL https://arxiv.org/abs/2308.12770.

Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. Visqol v3: An open source production ready objective speech and audio metric. In *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2020.

Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. Detecting voice cloning attacks via timbre watermarking, 2023. URL https://arxiv.org/abs/2312.03410.

M. Murphy, R. Metz, and M. Bergen. Biden audio deepfake spurs AI startup ElevenLabs—valued at $1.1 billion—to ban account: 'We're going to see a lot more of this'. Fortune, January 2024. URL https://fortune.com/2024/01/27/aifirm-elevenlabs-bans-account-forbiden-audio-deepfake/.

Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar. Proactive detection of voice cloning with localized watermarking, 2024. URL https://arxiv.org/abs/2401.17264.

Ernst H Rothauser. Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969.

Mohammad Shorif Uddin, Ohidujjaman, Mahmudul Hasan, and Tetsuya Shimamura. Audio Watermarking: A Comprehensive Review. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 15(5), 2024.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. Sok: Watermarking for ai-generated content, 2024. URL https://arxiv.org/abs/2411.18479.