Productionizing Audio Watermarking for Short-Form Video

Elias Kokkinis, Dimitris Koutsaidis, Elias Lumpert, Stergios Terpinas, George Tzoupis

Meta Platforms Inc.

Abstract

- Exploring audio watermarking in large scale short-form video platforms.
- Addressing challenges particularly focusing on minimizing watermark audibility while maximizing detectability.
- Presenting experimental results and discussing approaches to improve imperceptibility, detectability and enhance robustness.

Audibility

Two approaches to minimize audibility of the watermark, while trying to assure high detectability:

- Applying a gain to lower the watermark level
  - experimented with gain values (0.25, 0.50 and 1.0)
  - 0.50 proved to show best trade off
- Using Voice Activity Detection (VAD) to selectively apply the watermark on active speech minimizing audibility by avoiding to watermark silent regions

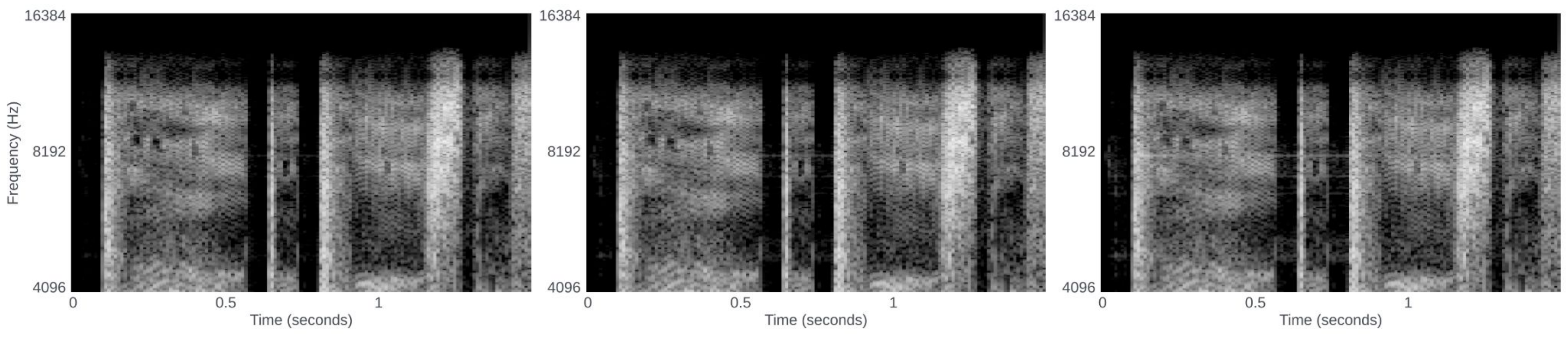

Figure 1: Spectrograms of watermarked files. We can see how the watermark becomes less audible as the mixing gain g decreases.

Config	Accuracy	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 1: Accuracy, AUC and ViSQOL (Median 95% CI) values for various mixing gains

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

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

Encoding 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

- Audio encoding can alter the signal, potentially affecting detectability of watermarks with high-frequency components
- Multiple encoding passes and certain schemes are more prone to expose this issue
- We employed a workaround using resampling, producing and applying the watermark at a lower operating sample rate

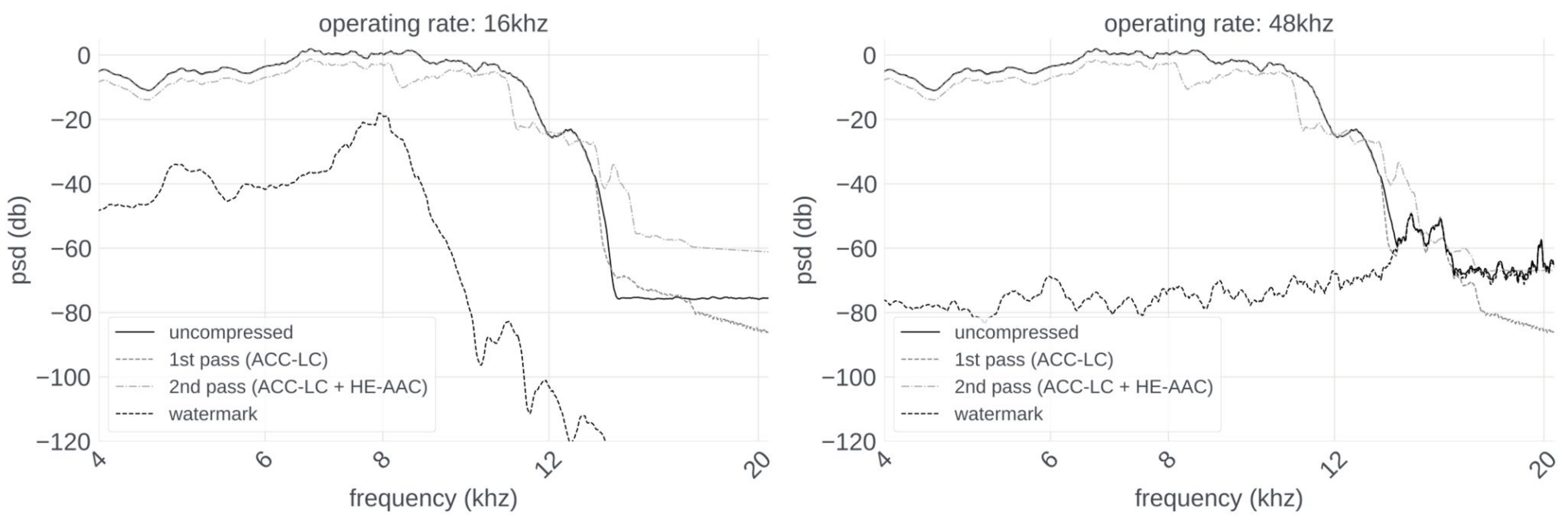

Figure 2: Power Spectral Density (PSD) of the watermarked synthetic speech encoded with different encoding parameters, with corresponding watermark. Left and right show watermark generation at different operating sample rates (16 kHz and 48 kHz).

Encoding	Op. rate	Accuracy	AUC
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 3: Results of the Encodings as an attack experiment

Mic Config	SNR	Accuracy	AUC
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%

Table 4: Results of the Music as an attack experiment

Music as an Attack

- Evaluating the effect of mixing music with speech as an attack in the context of a production system where multiple encoding passes also occur
- Two configurations of the music mixing attack, based on whether or not the music signal is mixed before watermarking and double encoding. Both configurations were tested on 3 different Signal-to-Noise Ratios (SNR) (5dB, 10dB, and 15dB).

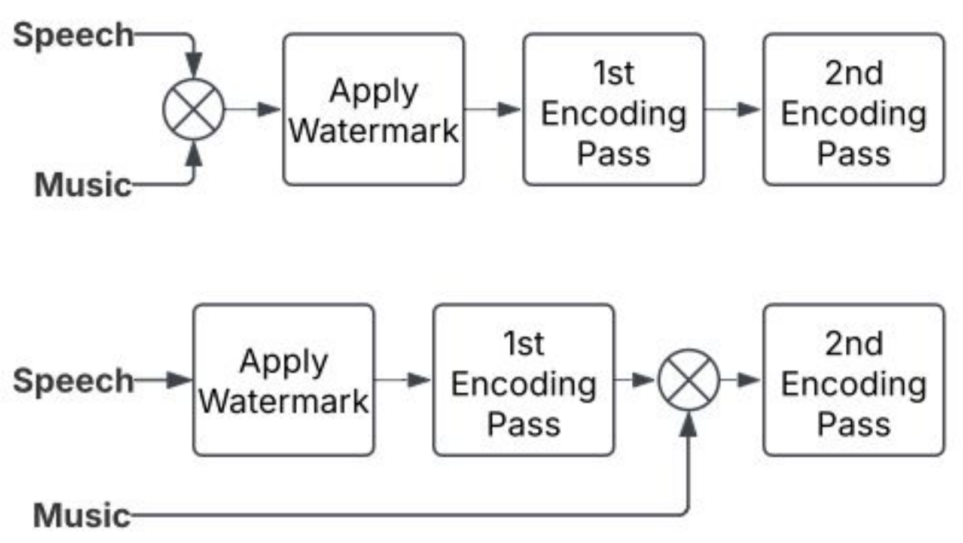

Figure 3: (Top) Music mixing before applying Watermark and encodings. (Below) Music mixing after watermark and 1st pass encoding is applied to the speech signal.

Conclusions

- 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.
- 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.
- Considering the preferred signal path and apply encodings, music signals, and watermarks strategically is important to ensure maximum detectability.
- Assessing the audibility of a watermark proved difficult and improving upon existing objective quality metrics could be beneficial.

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 Eleven-Labs—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/.5

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