The AI Waterfall : Case Study in Integrating Machine Learning and Security

Patrick BAS Jan BUTORA

Centre de Recherche en Informatique, Signal et Automatique de Lille
CNRS, Université Lille, Centrale Lille, Avenue Henri Poincaré, 59655 Villeneuve d'Ascq, France

Résumé – Si le tatouage numérique existe depuis plus de 30 ans, les solutions les plus populaires actuellement reposent sur des systèmes d'apprentissages de bout en bout. En prenant comme exemple un système de tatouage à l'état de l'art permettant de localiser précisemment la génération de paroles, nous montrons que l'opacité du système appris permet de facilement enlever le signal de tatouage sans aucunement dégrader le contenu. Cette attaque illustre à quel point il est actuellement difficile de formaliser la contrainte de la sécurité multimedia comme une fonction de coût à optimiser. Il ne faut donc pas oublier que dans le domaine de la sécurité, les sommets de l'IA cotoient souvent des chutes d'eau.

Abstract – If digital watermarking has been around for over 30 years, the most popular solutions currently rely on end-to-end machine learning. Taking as an example a state-of-the-art watermarking recognition system that can accurately locate word generation, we show that the opacity of the learned system allows for easy removal of the watermarking signal without affecting the content. This attack illustrates how it is currently difficult to formalize the multimedia security constraint as a cost function to be optimized. Regarding security, it is important to remember that AI summits are often close to waterfalls.

1 Introduction

The first academic paper on *Digital Watermarking* has been published in 1993 [17] by Tirkel *et al.* and since then the Google Scholar website indexes more than 100k publications on this topic. The scientific routes of this vast domain have been numerous, the most popular being (1) the cancellation of the interference between the host signal and the watermark [6], (2) the robustness to different operations such as compression, blurring, noise addition, content processing [8] and geometrical transforms [2], and (3) the tradeoff between robustness and perceptual distortion [1].

In 2005, Cayre *et. al.* paved the road for the security analysis of different watermarking schemes [5]. Contrary to robustness benchmarks, security attacks consider the role of a potential adversary, who represents a dynamic actor willing to do his best to attack the watermarking system by designing potential exploits such as removing the watermark while minimizing the distortion, copying the watermark on another content, or estimating the secret key used during the embedding phase from a set of watermarked contents.

With the rise of deep learning techniques in 2015, a new category of watermarking schemes has emerged : in 2018 the HiDDeN scheme [21] proposed end-to-end message embedding and detection for digital images using neural deep learning. Here the training minimizes an objective function composed of two terms, one measuring the distortion between the host and watermarked image, the other measuring the bit error rate between the embedded and decoded payload. Two networks are then trained, one to embed the message and another to decode it. Thanks to data-augmentation, this generation of schemes can be robust to different processes because, with an appropriate architecture, the system can learn to adapt its embedding strategy to be able to decode the embedded watermark.

However, this generation of end-to-end neural watermarking systems suffers from two important drawbacks :

1. Because the system relies on a large number of parameters, it misses explainability, *i.e.* the embedding strategy, which in classical systems consists of designing the watermark signal, choosing the watermarking domain, defining the detection function, is not straightforward to understand.

2. Very often no secret key is used, which means that the access to the watermark message is often public, or relies only on the training procedure, which means that the entropy of the key can be very small [3] and consequently the system can be very easily attacked.

Note that the difficulty of taking into account security and explainability constraints is not limited to neural watermarking, other ML systems like Large Langage Models [11], classifiers [10] or distributed learning [20] can also suffer from security attacks.

In this paper, we analyze the security of a popular and recent audio watermarking system proposed by Roman *et al.* [15] belonging to this new category of ML-driven schemes, and we show that if the proposed scheme is both extremely performant in term of robustness and detection accuracy, it is also terribly easy to remove the watermark with a success rate larger than 99% while decreasing the distortion w.r.t. the original content at the same time.

The paper is organized as follows, Section 2 presents the architecture, training strategy, and performances of the target scheme, Section 3 analyses the features of the watermark using time-frequency analysis, Section 4 proposes two possible attacks on the scheme, one to remove the watermark signal, and another to inject it. Finally, Section 5 highlights the dangers of formalizing watermarking as a machine-learning problem and proposes different perspectives to cope with this issue.

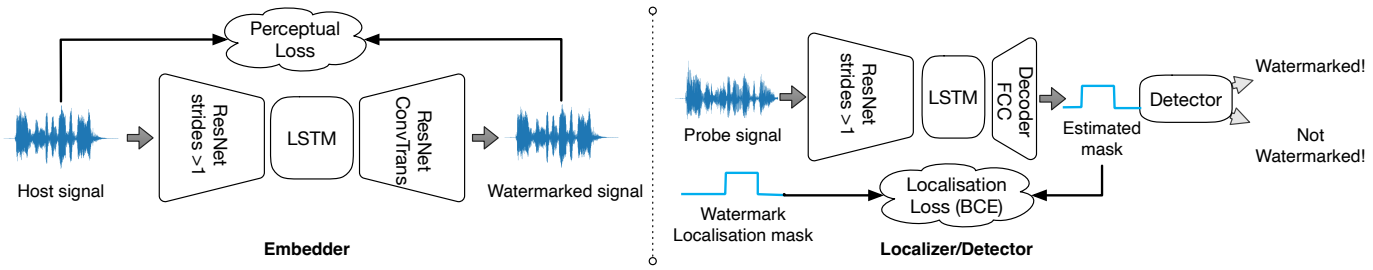


FIGURE 1 : Presentation of the AudioSeal system : the two neural structures, the watermark *embedder* and *decoder*, are trained using a double loss composed of two terms. The perceptual loss tends to minimize the distortion between the host and the original signal, and the localization loss minimizes the binary cross entropy (BCE) between a binary mask representing the localization of the watermark and the extracted one. Finally, the watermark is detected as soon as the watermark is localized on the whole test signal. These two neural structures are both similar to auto-encoders, composed of convolutional residual networks with stride/downsampling or transposed convolutions/upsampling operations, LSTM and FCC.

2 The AudioSeal system

The AudioSeal system [15] finds its legacy in a high fidelity neural audio coding method called *EnCodec* [9]. It is designed as an encoder-decoder architecture that combines transformers, LSTM, and vector quantization. The whole system, which includes a network embedding the watermark and a network localizing the watermark, is illustrated in Figure 1.

For watermarking, the quantization step of *EnCodec* is not considered, but the auto-encoder is fine-tuned to modify the latent representation of the audio signal and to produce a watermarked signal. Both the encoder and the decoder are modified by the learning procedure described below.

One interesting feature of the system is the possibility to localize the watermark, one of its potential applications being the detection and localization of voice-cloning. In this case, the synthetic voice is watermarked and the localization enables to detect where the signal has been tempered. A second network, localizing the watermarked signal, is consequently trained jointly with the first one. It is also an encoder-decoder that predicts a binary mask representing the estimated localization of the watermark signal at the sample level (usually 1/16k sec).

Note that the watermark is either zero-bit - *i.e.* a mark is embedded, not a message - or multi-bit. Without loss of generality on the security analysis detailed in the next section, we consider here only zero-bit embedding.

This double structure is trained considering a dual loss :

- A perceptual loss based on the *time-frequency loudness*, which computes the difference between the loudness function[1] applied on the original audio signal and the watermarked signal. The rationale here consists of exploiting the masking effect of the host signal to increase the watermark power on loud portions of the host signal.
- A localization loss representing the localization accuracy of the estimated mask. Here the binary cross entropy (BCE) between a binary mask indicating where the watermark is embedded and the estimated probability of the sample to be watermarked is used. The estimated localization mask is a thresholded version of the localization probability with a threshold of 0.5.

The watermark detection function computes the average predicted mask on a given duration and the sample is considered as watermarked if the average estimated mask, denoted D, is larger than 0.5, which is practically associated with a false positive rate of 10^{-3}.

The noteworthy robustness of the trained system is largely due to an extensive data-augmentation policy which includes 14 processes including low/high/band-pass filtering, resampling, echo process, addition of pink or white noise, and AAC or MP3 encoding applied on 4500 hours of audio samples. Different masking operations, such as adding silences, copy-paste, or using the original samples, are also considered data augmentation.

The benchmark of the system shows remarkable performance, surpassing its close competitor WavMark [7]. It indeed offers 100% detection accuracy when the audio signal is unprocessed, or after processes such as band-pass filtering, echo, pink noise addition, resampling, ACC and MP3 encoding. Except for highpass filtering with an accuracy of 61%, all the other 14 tested processes offer an accuracy above 91%.

3 Analysis of the embedding scheme

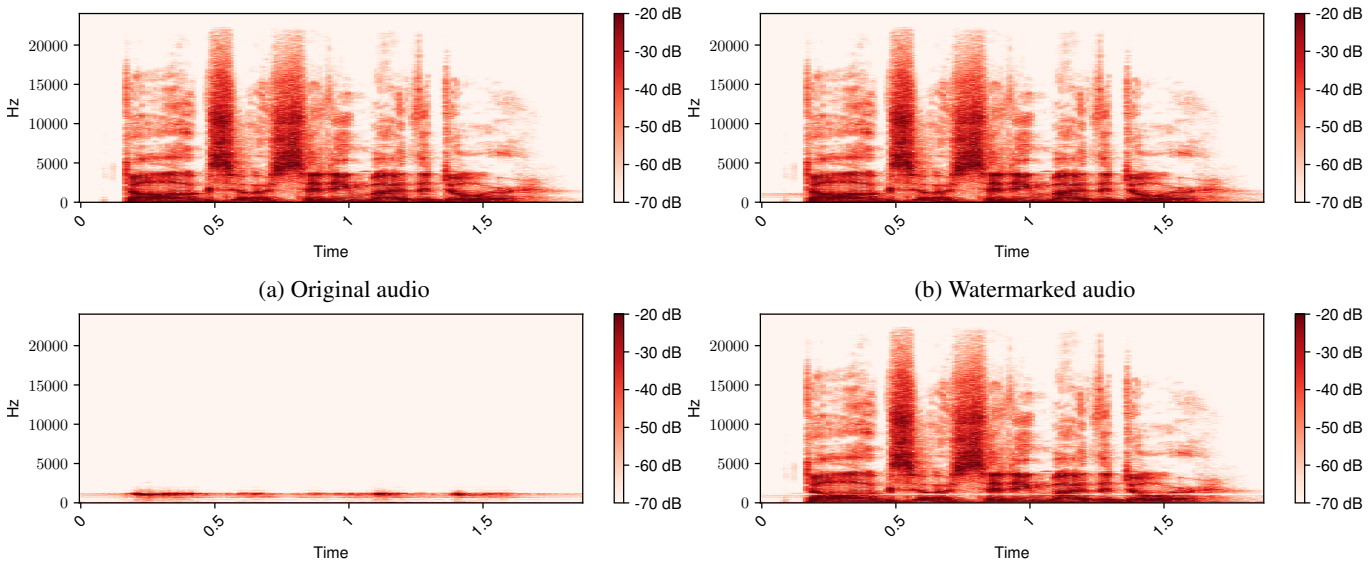
3.1 Discussion

When we first reviewed AudioSeal, we were impressed by its robustness, especially since besides data augmentation, nothing specific was designed at the embedding or detection side to be robust to synchronization issues linked to echo addition or cutting, to audio compression, or to noise addition. Before end-to-end neural watermarking systems such as this one, specific solutions for these different problems were deployed.

For example, before the deep-learning age, to cope with audio removal or cutting, specific synchronization patterns or hidden echos needed to be added as part of the watermark (see e.g. [18] and [19]), the possibility to localize the watermark at the sample level usually required the use of advanced coding systems such as Quantization Index Modulation [6] associated with error correcting codes [13]. Additionally, these embedding mechanisms needed, to be robust to amplitude scaling, to be associated with appropriate methods such as the use of a quantization step following the dynamic of the signal [12]. These contributions often required elaborated mechanisms that are difficult to simulate with neural networks.

As watermarking connoisseurs, we also noticed that contrary to the AudioSeal and WavMark systems, "old-school" water-

[1]This function is defined by the ITU-R BS.1770-4 recommendations [16].



(a) Original audio

(b) Watermarked audio

(c) Watermark signal

(d) Attacked audio (watermark removal)

FIGURE 2 : Short Time Fourier Transforms (STFT) of the original (a), watermarked (b), and attacked (d) signals, plus the watermark (c). Sampling rate = 48 kHz, window = 1024, and for visualization purposes, the maximum power spectrum is clipped between -20 dB and -70 dB with a reference power of 10. (d) shows the effect of the band-stop filter to remove the watermark (zoom-in).

marking systems use an extra parameter, namely a secret key, which enables to increase the security level of the system preventing an adversary from removing the watermarked content with very low distortion or to copy the watermark elsewhere. These attacks, if possible, would cause a genuine audio signal to be detected as generated from voice cloning, or on the contrary, cause a generated signal to be detected as genuine.

The absence of the security constraint is also reflected by the training procedure which takes into account only two constraints objectified by the two losses : a loss on the perceptual distortion, a loss on the robustness (the BCE) associated with appropriate data-augmentations, but no formal loss on security which is an inherent constraint in watermarking [4].

It consequently came to our mind that, in the end, AudioSeal might be less elaborate than expected, and we decided to pursue a small time-frequency analysis of the watermarked signals.

3.2 Time Frequency Analysis

Using the default reference implementation, we visualize the Short Time Fourier Transform (STFT a.k.a. Spectrograms) of the original audio, watermarked audio and watermark signal, which are presented respectively on Figure 2 in plots a, b and c.

Using the appropriate rendering (power spectrum in the range -20 dB and -70 dB), we can immediately notice that the watermark signal is a very narrow band-pass signal with a power positively correlated with the host power, and with a frequency around 1.1 kHz.

These two observations are coherent with the facts that : (i) The watermark is not audible : to take advantage of the masking effect, one has to scale the watermark power with respect to the host power ; (ii) the robustness of the system is rather low when facing high-pass filtering (see Section 2) : the watermark has a strong low-pass component.

Based on this simple observation, we are able to design two rather powerful attacks described in the next section.

	Ratio of Detected Contents	Ratio of PESQ≥ 4
Watermarked	100%	45.0%
Removal attack	0.6%	95.0%
Copy attack	81.1%	87.7%

TABLE 1 : Performances of the two presented attacks.

4 Two possible attacks

4.1 Removal attack

The attack is simple and consists of applying a Butterworth band-stop filter of order 5 between 1 kHz and 1.2 kHz.

We conducted this attack on 1135 voice samples coming from Kaggle[2] and the success rate (detection score $D > 0.5$) equals 99.4%. To evaluate the perceptual distortion of this removal attack, we also computed the *Perceptual Evaluation of Speech Quality Wide-Band* (PESQ-WB) score [14] which, when higher than 4.0, is associated with inaudible distortion. To compute the score, all files are downsampled to 16 kHz. Results are presented in Figure 3 and Table 1. If on the test samples, 45.0% of the watermarked samples are above this threshold, 95.0% of attacked samples are qualified. The increase of this rate plainly confirms that we successively removed the watermark signal, a PESQ-WB of 4.5 representing no distortion between the original and tested signal.

4.2 Copy attack

To inject a watermark-like signal into the original audio, we simply modified the magnitude of its Fourier transform by enhancing the frequency $f = 1.1$ kHz. We increased the magnitude at frequencies $+f$ and $-f$ by a constant value $c = 10$

[2]https://www.kaggle.com/datasets/farneetsingh2/audio-files

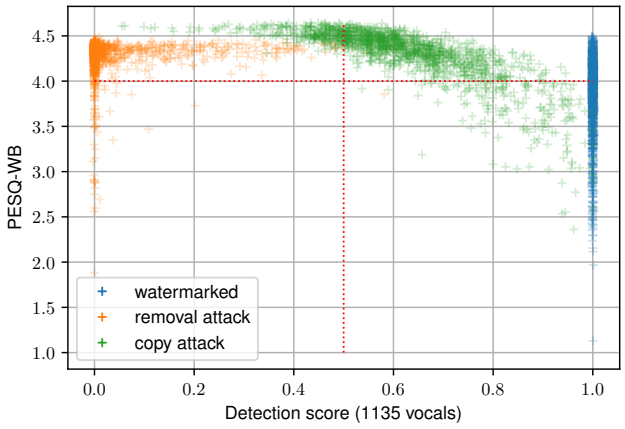


FIGURE 3 : Detection score VS distortion score (PESQ-WB) for watermarked and attacked contents after a removal attack and a copy attack.

which amounts to adding a cosine of frequency f and magnitude c to the signal. While a more clever approach could also modify the phase, our simple attack reaches 81.1% success rate of being detected as watermarked, demonstrating its effectiveness.

Moreover, 87.7% of attacked samples achieve a PESQ-WB score above the threshold 4, which is a significant improvement compared to the results obtained on the watermarked content.

5 Learned lessons

Mixing Machine Learning and Security is not an easy task. Using as a toy example a popular audio watermarking scheme we show that a very simple time-frequency analysis and basic processes can be designed to remove or copy the watermark with a very low distortion budget.

When considering security, *i.e.* the fact that an adversary can actively try to break the system, this end-to-end optimization process is difficult to design since this constraint is difficult to formalize as an objective function when the attack surface is too wide. This is for example the case in watermarking, with synchronization, estimation, copy, removal attacks, ... Note that in steganography, the security constraint is easier to formalize since the detectability can be objectified by a discriminator, or for example by a distance between the distribution of Cover and Stego contents.

Another possibility can also stem from the learning strategy : supposing that the used augmentations are diverse and exhaustive enough, the learning should be done concerning the most harmful augmentation, *i.e.* the wort case attack, and not w.r.t. the average set of augmentations. As highlighted in Section 2, the studied system was very robust on average, but not very robust to high-pass filtering. This trade-off induced a security hole.

Last and least : when related to watermarking, we strongly believe that security cannot be obtained without the use of a secret key : if no key is used, the detector is similar to a public system which can be openly attacked, for example with oracle attacks.

6 Acknowledgement

This work was also supported by a French government grant managed by the *Agence Nationale de la Recherche* under the France 2030 program, reference ANR-22-PECY0011.

Références

[1] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. Optimum decoding and detection of multiplicative watermarks. *IEEE Transactions on Signal Processing*, 51(4) :1118–1123, 2003.

[2] P. Bas, J.-M. Chassery, and B. Macq. Geometrically invariant watermarking using feature points. *IEEE transactions on image Processing*, 11(9) :1014–1028, 2002.

[3] P. Bas and T. Furon. A new measure of watermarking security : The effective key length. *IEEE Trans. Inf. Forensics and Security*, 8(8) :1306–1317, 2013.

[4] P. Bas, T. Furon, F. Cayre, G. Doërr, and B. Mathon. *Watermarking security : fundamentals, secure designs and attacks*. Springer, 2016.

[5] F. Cayre, F. Caroline, and F. Teddy. Watermarking security : Theory and practice. *IEEE Trans. Sig. Process.*, 53(10) :3976–3987, 2005.

[6] B. Chen and G. W. Wornell. Quantization index modulation : A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inf. Theory*, 47(4) :1423–1443, 2001.

[7] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei. Wavmark : Watermarking for audio generation. *CoRR*, 2023.

[8] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia content. *IEEE Trans. Image Process.*, 6(12) :1673–1687, 1997.

[9] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *stat*, 1050 :20, 2015.

[11] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz. Not what you've signed up for : Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023.

[12] Q. Li and I. J. Cox. Using perceptual models to improve fidelity and provide resistance to valumetric scaling for quantization index modulation watermarking. *IEEE Transactions on Information Forensics and Security*, 2(2) :127–139, 2007.

[13] R. Martínez-Noriega, M. Nakano, B. M. Kurkoski, and K. Yamaguchi. High payload audio watermarking : toward channel characterization of mp3 compression. *J. Inf. Hiding Multim. Signal Process.*, 2(2) :91–107, 2011.

[14] I. Rec. P. 862 "perceptual evaluation of speech quality". *International Telecommunication Union, Geneva*, pages 1–30, 2001.

[15] R. S. Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon, and T. Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43180–43196, 2024.

[16] B. Series. Algorithms to measure audio programme loudness and true-peak audio level. *International Telecommunication Union Radiocommunication Assembly*, 2011.

[17] A. Z. Tirkel, G. Rankin, R. Van Schyndel, W. Ho, N. Mee, and C. F. Osborne. Electronic watermark. *Digital Image Computing, Technology and Applications (DICTA'93)*, pages 666–673, 1993.

[18] S. Wu, J. Huang, D. Huang, and Y. Q. Shi. Efficiently self-synchronized audio watermarking for assured audio data transmission. *IEEE Transactions on Broadcasting*, 51(1) :69–76, 2005.

[19] Y. Xiang, I. Natgunanathan, D. Peng, W. Zhou, and S. Yu. A dual-channel time-spread echo method for audio watermarking. *IEEE Transactions on Information Forensics and Security*, 7(2) :383–392, 2011.

[20] C. Xie, K. Huang, P.-Y. Chen, and B. Li. Dba : Distributed backdoor attacks against federated learning. In *International conference on learning representations*, 2019.

[21] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei. Hidden : Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018.