SECURITY ISSUES IN DIGITAL AUDIO WATERMARKING

Michiel van der Veen[(1)], Aweke Negash Lemma[(2)], Fons Bruekers[(1)], Javier Aprea[(2)], and Ton Kalker[(1)]

[(1)] Philips Research Laboratories / [(2)] Philips Digital System Laboratories
Prof. Holstlaan 4, 5656 AA, Eindhoven
The Netherlands

ABSTRACT

Security of watermarking algorithms fundamentally differs from security of cryptographic algorithms. In this paper security issues are put in perspective and a method is presented to improve the security of a class of watermarking algorithms.

1. INTRODUCTION

Digital watermarking is a technology for imperceptibly conveying information through an existing audio-visual channel. The technology is often associated with copyright and copy-protection applications. Here, not only parameters like perceptual quality, robustness to common signal processing attacks, complexity, or capacity are important, but also its security. This resulted in a number of studies on this topic [1,2,3,4].

Up to this point, the usage of the term watermark security is somewhat vague. In addition, it is often confused with the term *cryptographic security*. In [5] a solid attempt was made to setup a framework for discussing watermark security. It is intended to describe some issues of watermark security at an abstract and mathematical level. In this paper we adopt some of these concepts and translate them into practical issues. Although these ideas hold for any multimedia format, we shall limit our examples to audio watermarking.

2. WHAT IS WATERMARK SECURITY?

To start our discussions on watermark security we will first give our definition of the term "watermark security". To this end, we will follow the framework described in [5] and start with a solid definition of robust watermarking [6]:

> *"Robust watermarking is a mechanism to create a communication channel that is multiplexed into original content (the host data). It is required that, (i) the perceptual degradation of the marked content with respect to the original is minimal, and (ii) the capacity of the watermark channel degrades as a smooth function of the degradation of the marked content".*

This definition makes explicit that watermarking is a method to create a communication channel on top of existing multi-media signals. In traditional cryptographic language, it means that it is a method of creating a communication channel to convey information from Alice to Bob. This reference brings us to the distinction between digital watermarking and cryptography. It is well known that cryptographic tools may be deployed to secure messages conveyed in a communication channel [7]. However, in many cases, the use of these tools for the "watermarking channel" is somewhat limited. This is true in particular for applications, such as copy protection, where only a few messages are used (e.g. 'copy-never', 'copy-once' and 'copy-allowed'). For such limited message sets, cryptography does not provide an optimum solution because (i) it is either always possible to guess the message by exhaustive search, or (ii) there is not even a need to hide the message. For example, for a multimedia signal that is marked 'copy-never' it is not the intention to hide the message 'copy-never', but to guarantee that the watermark detector reads this message faithfully (even from a degraded signal).

Considering that watermarks are used to convey information whereas cryptography is intended to add a security layer on top of the watermark channel, we could define watermark security as follows [5]:

> *"Watermark security refers to the inability of unauthorised users to either (i) remove, (ii) detect and estimate, (iii) write, or (iv) modify the original watermark message".*

We explicitly mention that watermark security is not related to the semantic of the message, but only with its physical presence. Cryptographic tools may be deployed to protect the semantics of the message. Therefore, the general aim of a malicious attacker is to either (i) remove (ii) change or (iii) embed a watermark such that its effectiveness is undermined, its detectability is disabled, or its general application is discouraged.

Adopting the Kerkhoffs' principle [7] of cryptography, we assume that an attacker has knowledge of all the details of the embedding and detection algorithms. The security of the watermark should therefore rely only in the

secrecy of the keys. Knowledge of this key and knowledge of the algorithm are both necessary for using or breaking the algorithm.

3. WATERMARK ATTACKS

Formally, any operation that affects the detectability of a watermark can be considered as an attack on the watermark signal. Focusing on audio watermarking, examples of 'attacks' that may occur in standard situations are: DA/AD conversion and related time-scale modifications [9], audio coding, equalization, and many more. The primary intention of these processing operations is not to remove the watermark, or make it undetectable. Therefore we make a distinction between the standard 'attacks' and malicious 'attacks'. The latter is defined as attacks with the focus on removing the watermark, or render it undetectable while maintaining an acceptable audio quality level.

A wide collection of watermark attacks is described in literature. Following [5] and the definition of security in the previous paragraph, we can make the following classification of malicious attacks:

- *Watermark removal*,
- *Watermark detection or estimation*,
- *Watermark writing*

The relevance of the individual attack scenarios differs for different application scenarios. This becomes clear if we classify watermarks based on their applications. For example, in [8] a distinction is made between *Copyright, Fingerprinting, Copy control*, *Annotation* and *Integrity* watermarks.

In the following we will briefly address each of the attacks mentioned above in light of some application scenarios.

Watermark Removal

In removal attacks, the malicious attacker is not necessarily concerned with the semantics of the message representing the watermark signal. Very often he does not care or he already knows the message. Therefore removing the message is his primary objective. Removing the message does not necessarily mean removing the watermark, but can mean modifying the audio signal such that the perceptual quality is retained, but that the watermark detector engine fails to detect. Furthermore, it can also mean to confuse the detector such that subsequent actions differ from what was originally intended by the watermark embedder.

An example that falls in this category is the *collusion attack*. This typically is a threat for fingerprint watermarks. Multiple copies of the same content are distributed, and each one contains a different fingerprint watermark. Collusion is the process of averaging different copies and thereby reducing the watermark energy compared to the host signal. Depending on the watermark design, a few copies could already be sufficient for a successful collusion attack. Note that with this attack, knowledge of the message and knowledge of the watermark is not necessary.

Watermark Detection or Estimation

The second category is the unauthorized watermark estimation or detection attack. Their main purpose is to determine the principal modifications to the host signal, which represent the watermark signal. In most cases, an estimation attack boils down to approximate the watermark by clever filtering of the host signal and taking the difference between the estimated original and the watermarked item [4]. Note that we realize that stand-alone watermark estimation is not an attack. It is merely a "tool" for subsequent attacking the watermark via e.g. watermark removal, unauthorized watermark detection or unauthorized watermark writing. However, for the purpose of this paper we decided to combine watermark detection and watermark estimation as one class of attacks.

An example of an attack, which falls in this category, is *averaging*. This attack makes use of the fact that the watermark is embedded with some redundancy - i.e. some systems embed portions of the watermark codes repeatedly. Focusing on audio watermarking, the averaging attack could involve temporal or spectral averaging, depending on the method of embedding. For example if we know that particular spectral components are modified according to a fixed pattern, we can make use of this knowledge to estimate it. Although these kinds of attacks are potential threats for all watermark applications, we explicitly mention the copy-control watermark. Usually it constitutes only a few bits of information, which is constant for a relatively long time segment. This could be a starting point for successful watermark sequence estimation.

Watermark Writing

In the last class of attacks we consider the problem of unauthorized writing of a watermark in content. An example where this could happen is in a copy-control application where a malicious attacker tries to embed a watermark 'copy-always' in a signal. A well-known example is the copy-attack [3]. This attack typically succeeds for watermark signals that are independent of the content to be marked.

An example is the spread-spectrum technique, where the spread spectrum noise depends only on the message to

be embedded (possibly with some perceptual adaptation). A malicious attacker could relatively easily estimate the watermark and embed it in another signal.

4. AUDIO WATERMARK ALGORITHM

In the following sections we will explain the implications of some of the concepts discussed thus far on a practical watermark algorithm. For that purpose, we will use the audio watermark algorithm described in [9]. The principal embedding strategy of this algorithm is displayed in Figure 1. A watermark *w* is calculated based on the host signal *x*, the message and a secret key in the watermark generation block A. The final watermark signal is then weighted according to a parameter α and added to the host. Although not relevant to this paper, the embedding strength α is controlled via a perceptual model assuring inaudible distortions.

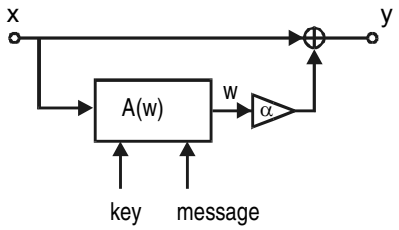

Figure 1: Basic embedding scheme.

The watermark is designed to modify the magnitudes of the Fourier coefficients. Without additional measures, a 1-bit message could be embedded by continuously modifying the Fourier coefficients according to a constant pattern. Therefore detection of the watermarks also operates in the Fourier domain. It relies on accumulating a sufficient amount of signal *y*, and correlating its Fourier coefficients *Y* with the initial modifications.

An obvious malicious attack to this general scheme is to accumulate a sufficient amount of watermarked host signal so as to estimate the watermark signal. Then, use this estimate to render the watermark undetectable, either by adding an additional watermark, or by degrading it. This relatively simple example shows the potential of malicious attacks. In the following section, we will give an overview of countermeasures to these attacks.

5. COUNTERMEASURES

We first give a brief overview of techniques that can be considered as countermeasures to one or more of the malicious attacks cited above. Secondly, we propose a technique that enhances the security of any watermarking system against the mentioned malicious attacks.

Overview of known security measures

As a first category, we consider a set of techniques that embed watermarks in the portions of the host signal, where this portion is selected according to a certain pre-defined random sequence usually referred to as the *stego-key*. One outstanding example is the frequency hopping based echo-hiding technique presented as technology A in the SDMI public challenge [10]. The main shortcoming of such approaches is that the security is based on the secrecy of the stego-key and shares the same danger of being estimated using the same techniques that are designed to estimate the secret watermark sequence. Since the stego-keys are normally very short, the extra effort needed to resolve this security is not significant. This has been successfully demonstrated in [10] where technology A was successfully defeated in the SDMI public challenge.

As a second category we consider techniques that perform transformations on watermarked signals so as to overcome watermark attacks. One technique from this group is the one given in [11] that is intended to resist collusion attacks. Here, non-equal phase modifications are applied to signals originating from the same host signal but carrying different watermark sequence such that when they are added up the audio quality is severely degraded. A potential drawback of this approach is that one can easily estimate the phase difference between the two versions of the audio. To this end, we demonstrate a simple mathematical model how this can be put into effect.

Let x_1 be a signal x carrying the watermark w_1 and let x_2 be the same signal carrying the watermark w_2. Let us now assume that y_1 and y_2 are phase-modified versions of x_1 and x_2, respectively, i.e.,

$$y_1(t) = x_1(t - \tau_1(t))$$
$$y_2(t) = x_2(t - \tau_2(t))$$

Since τ_1 and τ_2 are slowly varying phases, one can easily estimate the difference τ_1-τ_2. To this end, define a frequency time dependent ration $R(t,f)$ such that

$$R(t,f) = \frac{Y_1(t,f)}{Y_2(t,f)} \approx \frac{X_1(t,f)e^{j2\pi\tau_1(t)}}{X_2(t,f)e^{j2\pi\tau_2(t)}}.$$

Now, let the filter h be such that its transfer function $H(t,f)$ is given by

$$H(t,f) = \frac{R(t,f)}{|R(t,f)|}.$$

If we pass $y_2(t)$ through this filter we get a matching phase behavior with $y_1(t)$ which will not add up destructively.

Alternative measure against averaging attack

The problem of the foregoing approaches is that they don't address the key issue for the success of the malicious attacks – the watermark estimation issue. We believe that the inability to estimate the watermark secret key (sequence) is a fundamental factor in improving the watermark security. Thus, in the following, we propose a true random time- frequency- space- and/or secret key-multiplexed embedding approach that significantly improves the watermark security and at the same time resolves the shortcomings of the known approaches.

The idea is to embed a watermark by randomly varying the embedding parameters. Here, unlike the stego-key, the random sequence need not be unique or known to the detector. This means that, since the randomness of the embedding parameters is not bound to a certain pattern, it is very difficult, if not impossible, to estimate them. One can appreciate that the improvement in security in this instance is a tradeoff between detectability and the inability for averaging. That is, when trying to defeat the averaging problem, we make it difficult for the detector to use averaging to improve its performance. Thus, the rate of the parameter change should be such that it still allows reliable detection performance. Depending on the application, one can put more weight to one or the other.

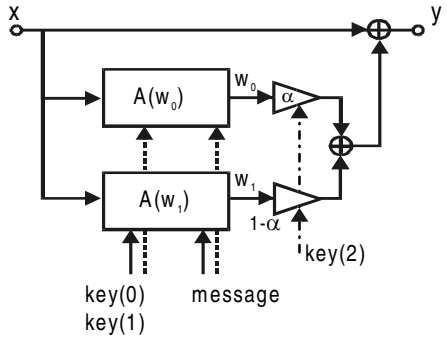

Figure 2: Watermark embedding scheme based on principles of Figure 1, but with additional randomization.

An implementation related to the algorithm described in section 4, is illustrated in Figure 2. The apparent security is improved by introducing additional randomization. An example is given for the insertion of either watermark w_0, or w_1. Both watermarks could be different watermarks generated with a different key, or may even originate from completely different algorithms. The choice of both watermarks is controlled via the parameter α and an additional secret key. In this setup, the embedded pattern changes over time, making it more difficult to estimate the secret sequence. We explicitly mention that we do not make the system proofed secure, but we only introduce additional difficulties for malicious attackers.

6. CONCLUSIONS

A clear definition of security in watermarking systems is discussed and a way to improve upon is presented. Because of true randomness of the embedding parameters, the security against averaging attack is improved.

7. REFERENCES

[1] F. Deguillaume, G. Csurka, and T. Pun, *Countermeasures for Unintentional and Intentional Video Watermarking Attacks*, in IS&T/SPIE Electronic Imaging 2000, San Jose, CA, USA, 2000.

[2] F. Hartung, J.K. Su and B. Girod, *Spread-spectrum watermarking: Malicious attacks and counterattacks*, In Proc. SPIE Security and Watermarking of Multimedia Content, San Jose, CA, USA, 1999.

[3] M. Kutter, S. Voloshynovskiy and A. Herrigel, *Watermark Copy Attack*, In Proc. SPIE Security and Watermarking of Multimedia Content, San Jose, CA, USA, 2000.

[4] G.C. Langelaar, R.L. Lagendijk, and J. Biemond, *Removing spatial spread spectrum watermarks by non-linear filtering*, in Proc. EUSIPCO 98, Rhodos, Greece, September 1998.

[5] T. Kalker, *Considerations on Watermark Security*, IEEE Workshop on Multimedia Signal Processing, October, 2002.

[6] P. Moulin and J. O'Sullivan, *Information-theoretic analysis of information hiding*, IEEE Transactions on Information Theory, preprint, 1999.

[7] B. Schneier, *Applied Cryptography* 2nd Edition, John Wiley& Sons Inc., New York, 1996.

[8] J. Dittman, P. Wohlmacher, and K. Nahrstedt, *Using Cryptography and Watermarking Algorithms*, IEEE Multimedia, Oct-Dec 2001.

[9] M. van der Veen, F. Bruekers, J. Haitsma, T. Kalker, A.W. Lemma and W. Oomen, *Robust, multi-functional and high-quality audio watermarking technology*, Audio Engineering Society, Presented at the 110th AES convention, 2001. paper no. 5345.

[10] S. A. Craver, M. Wu and B. Liu, *Reading Between the lines: Lesson from the SDMI Challenge,* Proceedings of the 10-th USENIX Security Symposium, Washington DC, August 13-17, 2001.

[11] US patent number US 6145081, *Method and apparatus for preventing removal of embedded information in cover signals,* Verance Corporation.