RANDOMIZING THE REPLACEMENT ATTACK

Darko Kirovski and Zeph Landau
Microsoft Research, One Microsoft Way, Redmond, WA 98052

ABSTRACT

Billions of dollars allegedly lost to piracy of multimedia have recently triggered the industry to rethink the way music and movies are distributed. As encryption is vulnerable to re-recording, currently all copyright protection mechanisms tend to rely on watermarking. In order to analyze the security of such systems, recently, a new breed of replacement attacks has been proposed that strongly affects most modern watermarking systems. A typical replacement attack relies upon the observation that multimedia content is often highly repetitive. Thus, the attack procedure replaces each signal block with another, perceptually similar block computed as a combination of other similar blocks found either within the same media clip or within a library of media clips. In this paper, we demonstrate that by randomizing the attack algorithm, its performance can be improved in almost all aspects: attack efficacy, distortion, speed, and size of the look-up media library. We describe the logistics of the new attack and an exemplary implementation against a spread-spectrum data hiding technology for audio signals.

1. INTRODUCTION

Significantly increased levels of multimedia piracy over the last decade have put the movie and music industry under pressure to deploy a standardized anti-piracy technology. The goal is to enforce copyright protection via **content screening** on client media players. A media player would refuse to play copyright protected content for which the user does not hold a license. In a **forensic marking** scenario, each distributed copy is marked with a unique fingerprint. Forensic analysis of pirated material is then performed on a trusted secure server in the presence of the original content. Here, players are not modified. Both systems have inherent problems. Content screening demands public key watermarking [1], whereas fingerprinting suffers from exceptionally low collusion resistance [2]. In addition, the embedded watermark in both applications must survive an arbitrary signal processing attack which preserves perceptual fidelity of the targeted content.

1.1. Rationale Behind the Attack

Recently, Kirovski and Petitcolas proposed the first version of the **replacement attack** which aims at reducing the correlation of a watermarked signal with its watermark by replacing each original watermarked block of the multimedia signal with another perceptually similar block [3]. The replacement is computed as a combination of other signal blocks that are perceptually similar but not tainted with the same watermark bits as the original marked block. Similar attacks preceded the replacement attack in [4] and [5].

Assuming a highly repetitive multimedia content, which is the case with most music (example illustrated in Figure 1) and video, the replacement attack adds a nearly marginal noise to the marked content. However, although the noise appears to be zero-mean i.i.d. gaussian, it wipes out an enormous percentage[1] of the correlation [3]. An additive and truly gaussian noise of equivalent variance would not alter the correlation beyond its original statistics. The surprising effect of the replacement attack stems from the redundancy which exists in multimedia content. By using this information, one can recreate the protected multimedia while being marginally dependent upon the originally embedded watermark.

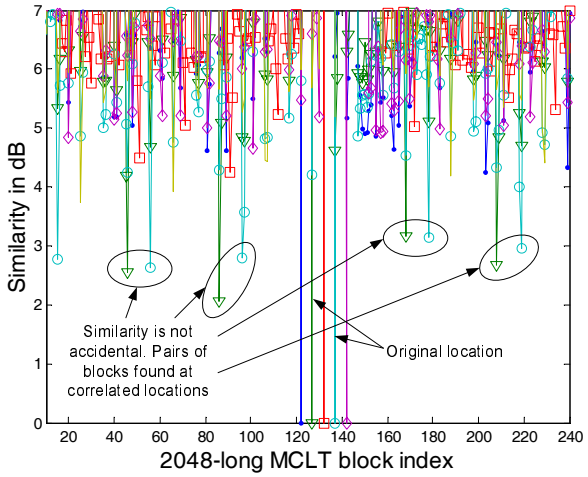

Fig. 1. Music self-similarity: a similarity diagram for five different 2048-long MCLT blocks within a techno clip with 240 MCLT blocks. Zero-similarity denotes equality. Abscissa x denotes the index of a particular MCLT block. The ordinate denotes the similarity $||x, B_i||$ of the corresponding block x with respect to the selected five blocks with indices $B_i|i = \{122, 127, 132, 137, 142\}$.

The rationale behind the attack is simple. A given watermarked block $\mathbf{x} + \mathbf{w} \in \{\mathbb{R}\}^N, \mathbf{w} \in \{\pm 1\}^N$ of N samples represents a point in the N-dimensional space. With no loss of generality[2], the original content can be modelled as a zero-mean gaussian random variable of i.i.d. samples $\mathbf{x} = \mathcal{N}(0, \sigma_x)^N$. Based on the statistics of spread spectrum watermarking [6], by adding additive white gaussian noise $\mathbf{n} = \mathcal{N}(\mu, \sigma)^N$ to the marked content, the expected normalized correlation value of $E[(\mathbf{x} + \mathbf{w} + \mathbf{n}) \cdot \mathbf{w}] = E[(\mathbf{x} + \mathbf{w}) \cdot \mathbf{w}] = 1$ remains intact for any mean μ and variance σ^2. Operator $(\cdot)$ is defined as a normalized inner product of two vectors: $\mathbf{x} \cdot \mathbf{y} = N^{-1} \sum_{i=1}^{N} x_i y_i$. We denote all points $\mathbf{y} \in \{\mathbb{R}\}^N$ in a ball $Y(\varepsilon) : \{||\mathbf{y} - (\mathbf{x} + \mathbf{w})|| \leq \varepsilon\}$ at minimal Euclidean distance $\varepsilon \geq 1$ from $\mathbf{x} + \mathbf{w}$ as "perceptually valid" points. The goal of the attacker is to drive $\mathbf{x} + \mathbf{w}$ with some noise pattern $\mathbf{n}_a$ into a point of attack $\mathbf{z} = \mathbf{x} + \mathbf{w} + \mathbf{n}_a \in Y(\varepsilon)$ which is not correlated with $\mathbf{w}$, i.e., $E[\mathbf{z} \cdot \mathbf{w}] = 0$. Although there are many points in Y which are not correlated with $\mathbf{w}$, for example $\mathbf{x}$ is by definition

[1] Authors report up to 90% correlation reduction as a result of the attack.
[2] Due to the Central Limit Theorem.

one of them as $E[\mathbf{x} \cdot \mathbf{w}] = 0$, it is difficult to find in a single trial an $\mathbf{n}_a$ which cancels out the effect of $\mathbf{w}$. For example, random $\mathbf{n}_a = \mathbf{n}$ provably does not affect $E[(\mathbf{x} + \mathbf{w} + \mathbf{n}_a) \cdot \mathbf{w}]$.

There are three standard approaches to find $\mathbf{n}_a$. The first one is to use the estimation attack $\mathbf{n}_a = \text{sign}(\mathbf{x} + \mathbf{w})$ which has limited success if $\mathbf{w}$ is not redundantly embedded [1]. If the watermark detector is available to the adversary as in the case of content screening, she can launch a multi-trial search for $\mathbf{n}_a$ which quickly finds an optimal attack vector to achieve its goal [7]. Finally, the third approach is to randomly bend the space dimensions to inflict difficulty positioning $\mathbf{w}$ in the same direction as it had when embedded to $\mathbf{x}$ [8, 9]. This type of an attack can be prevented using redundancy while embedding $\mathbf{w}$ [10], however, such a solution is prone to the estimation attack [11].

The replacement attack assumes that multimedia is repetitive enough so that for a given point $\mathbf{x}+\mathbf{w}$ of sufficient dimensionality N, there exists at least one other point $\mathbf{a}$ within the same media clip that is within the ball of interest $Y(\varepsilon)$. By definition, if $\mathbf{a}$ is not tainted with $\mathbf{w}$, replacing $\mathbf{x}+\mathbf{w}$ with $\mathbf{a}$, intuitively, removes the targeted correlation. However, if $\mathbf{a} = \mathbf{x} + \mathbf{w}$, then $E[\mathbf{a} \cdot \mathbf{w}] = 1$. With the increase of N, this case can be made arbitrarily unlikely. Although perceptual repetitiveness in audio [3] and in particular video [5] is significant, the above assumption is still strong as it is unlikely to expect that for any high-dimensional point in the media clip there exists a redundant one.

To address this issue, the replacement attack as described in [3], creates $\mathbf{a}$ in two steps. First, it finds a set $\mathbf{B}$ of K points $\mathbf{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$ in the media clip closest to $\mathbf{x} + \mathbf{w}$ and then, it computes a least-squares approximation of $\mathbf{x} + \mathbf{w}$ using a linear combination of vectors in $\mathbf{B}$. The resulting approximation is the replacement vector $\mathbf{a}$. For large N, it is safe to assume that none of the vectors in $\mathbf{B}$ are correlated with $\mathbf{w}$. However, the correlation of their linear combination with $\mathbf{w}$ increases with the increase of K. Theoretic analysis of this process is exceedingly difficult as it is hard to model content similarity due to its diverse nature. To address this issue, the replacement attack assumes that there is a "safe distance" $\alpha \leq ||\mathbf{a} - (\mathbf{x} + \mathbf{w})||$, that $\mathbf{a}$ needs to have with respect to $\mathbf{x} + \mathbf{w}$ in order to be non-correlated. This is a model that has proven to yield solid results for small K [3].

We expand upon this similarity model by assuming that $\mathbf{a} = \mathbf{x} + \mathbf{d} + \mathbf{v}$, where $\mathbf{d}$ is the similarity noise in the original content modelled as a zero-mean normal random variable of i.i.d. samples $\mathbf{d} = \mathcal{N}(0, \sigma_d)$ and $\mathbf{v} \in \{\pm 1\}^N$ is the watermark added to the original block $\mathbf{x} + \mathbf{d}$.

Theorem 1. ***Computing*** $E[\mathbf{a} \cdot \mathbf{w}]$***.*** *Assuming mutual independence of* $\mathbf{x}$*,* $\mathbf{d}$*,* $\mathbf{v}$*, and* $\mathbf{w}$*, if* $||\mathbf{a} - (\mathbf{x} + \mathbf{w})|| = \alpha$*, then:*

$$2E[\mathbf{a} \cdot \mathbf{w}] = 2 + \sigma_d^2 - \alpha. \quad (1)$$

Proof. From $||\mathbf{a} - (\mathbf{x} + \mathbf{w})|| = \alpha$ and $E[\mathbf{x} \cdot \mathbf{w}] = 0$, we conclude that $2E[\mathbf{a} \cdot \mathbf{w}] = 1 + ||\mathbf{a} - \mathbf{x}|| - \alpha$. From $||\mathbf{a} - \mathbf{x}|| = ||\mathbf{d}|| + 1$, we derive Eqn.1. □

From Th.1, we conclude that if we want to drive $E[\mathbf{a} \cdot \mathbf{w}]$ to a small value, the minimal distance α must be driven close to $2+\sigma_d^2$. Needless to say, the efficacy of the attack is highly determined by the similarity metric σ_d^2, which represents the variance of the difference among most similar blocks in a given media clip. By using least-squares approximation of $\mathbf{x} + \mathbf{w}$ using various similar blocks from the media clip, we effectively reduce σ_d at the cost of increasing the correlation of $\mathbf{a}$ and $\mathbf{x} + \mathbf{w}$. Careful execution of this process is the key to the success of the replacement attack.

In the remainder of this paper, using randomization of the basic primitives of the replacement attack, we address several trade-offs and problems that inherently exist in realistic attack scenarios. In the next section, we present the new techniques and the logistics behind the proposed steps. Finally, in the last section, we demonstrate how it can be applied on a spread spectrum audio watermarking technology.

2. RANDOMIZED REPLACEMENT ATTACK

The replacement attack is not limited to a type of content or to a particular watermarking algorithm. For example, systems that modulate secrets using spread-spectrum [6] and/or quantization index modulation (QIM) [12] are all prone to the replacement attack. For brevity, the analysis of the attack in this paper is restricted to direct sequence spread spectrum watermarks. In order to launch the attack successfully, the adversary does not need to know the details of the watermark codec. This assumption is convenient for the adversary compared to the knowledge that other attacks mentioned in the previous section mandate.

Given a signal $\hat{\mathbf{x}} \in \{\mathbb{R}\}^M$ and a corresponding watermark $\hat{\mathbf{w}} \in \{\pm 1\}^M$, the attack performs the following steps:

I. partition $\hat{\mathbf{x}} + \hat{\mathbf{w}}$ into overlapping blocks $\mathbf{x} + \mathbf{w}$ of length N,

II. for each block $\mathbf{x} + \mathbf{w}$, find a set $\mathbf{B}$ of K perceptually most similar blocks in $\hat{\mathbf{x}} + \hat{\mathbf{w}}$ that do not overlap $\mathbf{x} + \mathbf{w}$,

III. compute the replacement block $\mathbf{a}$ as a least-squares linear approximation of $\mathbf{x} + \mathbf{w}$ using blocks from $\mathbf{B}$, and

IV. replace $\mathbf{x} + \mathbf{w}$ with $\mathbf{a}$.

2.1. Attack Trade-offs

Considering the issues related to the replacement attack and presented in the previous section, we identify several important trade-off decisions that the adversary needs to make before applying the attack. The trade-offs reflect on the following important performance metrics: reduction in correlation, distortion, and speed.

T.1 POINT DIMENSIONALITY N has a profound effect on $E[\mathbf{a} \cdot \mathbf{w}]$. By increasing N, the adversary reduces the likelihood that vectors in $\mathbf{B}$, as well as their linear combinations, are correlated with $\mathbf{w}$. On the other hand, significantly increased N reduces the expectation on the cardinality of $\mathbf{B}$ as it increases σ_d, thus, reducing attack effectiveness. One heuristic is that N should be maximized for a given clip so to still produce "perceptually valid" matches.

T.2 SELECTION OF α. Increased α improves all aspects of attack performance except distortion. This parameter should be maximized for a given perceptual quality.

T.3 SIZE M OF THE LOOK-UP MEDIA LIBRARY $\hat{\mathbf{x}} + \hat{\mathbf{w}}$ determines the complexity of the attack and can significantly improve σ_d.

2.2. Randomizing the Attack

2.2.1. STEP I: Signal Partitioning.

For improved perceptual quality of the resulting multimedia clip, the protected signal $\mathbf{z} = \hat{\mathbf{x}} + \hat{\mathbf{w}}$ is partitioned into a set of blocks $\Pi = \{\mathbf{p}_1, \ldots, \mathbf{p}_P\}$, where each block $\mathbf{p}_i = \{h_j z_{1+(i-1)N/2+j},$

$j = 1 \ldots N\}$ overlaps its neighbors and is windowed with an analysis windowing function $\mathbf{h} \in \{\mathbb{R}\}^N$ that yields perfect reconstruction with its synthesis counterpart. With no loss of generality, we assume that $\hat{\mathbf{x}} + \hat{\mathbf{w}}$ is an one-dimensional signal such as audio.

2.2.2. STEP II: Search for the Substitution Base.

Finding perceptually similar blocks of certain music or video content is a challenging and computationally expensive task. In this paper we restrict our focus to audio, although video is in many cases a much better source of repetitive content within a single recording. For example, within a common scene, its objects experience geometric transformations significantly more frequently than changes in appearance. In general, repetition is often a principal part of composing music and is a natural consequence of the fact that distinct instruments, voices and tones are used to create a soundtrack. Thus, it is likely to find similarities within a single musical piece, an album of songs from a single author, or in instrument solos.

For each point $\mathbf{p}_i$, we want to find a set $\mathbf{B}_i$ of K best matched blocks in $\mathbf{z}$ denoted as $\mathbf{B}_i = \{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$ with individual points $\mathbf{b}_j = \mathbf{h}\{z_{s_j}, \ldots, z_{s_j+N-1}\}$ where s_j indexes the location of $\mathbf{b}_j$ in $\mathbf{z}$. Before we define the search process, we adopt normalized and squared Euclidean distance between two N-dimensional points $\mathbf{a}$ and $\mathbf{b}$ as a similarity metric:

$$\phi(\mathbf{a}, \mathbf{b}) = ||\mathbf{a} - \mathbf{b}|| = \frac{1}{N} \sum_{k=1}^{N} [a_k - b_k]^2 . \quad (2)$$

Although in realistic attack scenarios the similarity function is masked with the perceptual model for improved matching, in this section we disregard this effect. Next, note that maximized normalized correlation corresponds to minimal Euclidean distance in L^2. Thus, the search for top K matches in $\mathbf{z}$ against each $\mathbf{p}_i$ can be conducted in the following way. We first compute the normalized block convolution of $\mathbf{p}_i$ with respect to $\mathbf{z}$. This can be done rather fast using the Fast Fourier Transform and the overlap-add fast convolution method [13]. The complexity of this step is $\mathcal{O}(M \log_2 N)$. The top K correlated blocks in $\mathbf{z}$ that do not overlap $\mathbf{p}_i$ constitute the substitution base $\mathbf{B}_i$ for $\mathbf{p}_i$.

2.2.3. Step III: Computing the Replacement.

This step of the algorithm is crucial as it resolves the trade-offs related to the selection of α and the inherent σ_d - the two most important metrics of the attack. First, we review the restrictions of the attack, which are different compared to the ones in [3]. We restrict that each sample a_i of the replacement block $\mathbf{a}$ is at a "safe" and "perceptually valid" distance from the sample p_i it is replacing in $\mathbf{p}$. More formally:

$$(\forall a_i \in \mathbf{a})\ \alpha \leq |a_i - p_i| \leq \varepsilon. \quad (3)$$

We discuss several randomized algorithms for computing $\mathbf{a}$ such that the above constraints are satisfied.

Algorithm A1 computes the replacement block $\mathbf{a}$ in several steps. In the first step, it generates c random and distinct subsets of r blocks from $\mathbf{B}$. We denote these subsets $\mathbf{S}_1$ through $\mathbf{S}_c$. For each of these subsets, **A1** computes the least squares approximation of $\mathbf{p}$. More formally, for a given $\mathbf{S}_i = \{\mathbf{s}_1, \ldots, \mathbf{s}_c\}$ we are seeking for λ_i such that $||\mathbf{S}_i\lambda_i - \mathbf{p}||$ is minimized. Optimal λ_i is computed simply as:

$$\lambda_i = (\mathbf{S}_i^T \mathbf{S}_i)^{-1} \mathbf{S}_i^T \mathbf{p}, \quad (4)$$

which yields the following replacement candidate vector $\mathbf{a}_i = \mathbf{S}_i\lambda_i$. Samples from $\mathbf{a}_i$ can be categorized into two categories: ones that satisfy Eqn.3 and the ones that do not.

To address this issue, we introduce a binary coverage matrix $\mathbf{q}_i = \{0, 1\}^N$ associated with each $\mathbf{S}_i$. We set to $q_{i,j} = 1$ if sample $a_{i,j}$ satisfies Eqn.3 and vice versa. We define as effective dimensionality $\bar{N}(\mathbf{a}_i)$ the number of samples a given replacement $\mathbf{a}_i$ covers $\bar{N}(\mathbf{a}_i) = \sum_{j=1}^{N} q_{i,j}$. Heuristically, we are already driven by the assumption that the larger the effective dimensionality, the stronger the effect of the attack on the resulting correlation. Hence, we can model the goal of our replacement algorithm as a combinatorial optimization problem. **A1** aims to cover as many as possible samples from $\mathbf{p}$ using as few as possible vectors from the set $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_c\}$. This problem is better known as minimum cover and is NP-hard [14]. **A1** solves it using a greedy heuristic which iteratively selects vectors from $\mathbf{A}$ that cover maximum number of remaining uncovered samples. There may be samples that cannot be covered by any $\mathbf{a}_i$ - their values are set to the corresponding values of the marked content $\mathbf{z}$ in order to minimize distortion.

Parameters K, c, and r strongly influence the performance of **A1**. By increasing r, we reduce σ_d at the cost of stronger correlation of each $\mathbf{a}_i$ and $\mathbf{z}$. Empirically, we have received best results for small r, usually in the order of $r \in [1, 20]$. Once r is set, we determine the average effective dimensionality $\bar{N}$ in $\mathbf{A}$. The higher the $\bar{N}$, the more candidate trials c **A1** can afford to test. Again empirically, we have achieved solid results with $c \approx \bar{N}$. Finally, the size of the substitution database is kept large at $K \sim \bar{N}$.

Algorithm A2 is a significantly slower, but still randomized, version of **A1** and the algorithm presented in [3]. It is based on the observation that the blocks in $\mathbf{A}$ are highly redundant because they are searched using a common criterion (Eqn.2). Using only these blocks in linear combinations restricts strongly the search space. A better, but still not optimal strategy in representing $\mathbf{p}$ as accurately as possible using a constant number of blocks from $\mathbf{z}$, is to use a variant of Gram-Schmidt orthonormalization (GSO) [15].

Hence, **A2** iteratively performs the following process. It finds the first $\mathbf{A} = \mathbf{a}_1 = \mathbf{B}_1$ with $K = 1$. The most similar point $\mathbf{a}_1$ is subtracted from $\mathbf{p}$ as $\mathbf{p} - \lambda_1\mathbf{a}_1$, where λ_1 is a scalar equal to the normalized correlation of $\lambda_1 = \mathbf{p} \cdot \mathbf{a}_1/(||\mathbf{p}||\ ||\mathbf{a}_1||)$. In the subsequent iteration, **A2** computes the similarity of $\mathbf{p} - \lambda_1\mathbf{a}_1$ with $\mathbf{z}$ as described in Subsection 2.2.2, finds the best match $\mathbf{a}_2$ and subtracts it from the remainder as $\mathbf{p} - \lambda_1\mathbf{a}_1 - \lambda_2\mathbf{a}_2$, where $\lambda_2 = (\mathbf{p} - \lambda_1\mathbf{a}_1) \cdot \mathbf{a}_1/(||\mathbf{p} - \lambda_1\mathbf{a}_1||\ ||\mathbf{a}_1||)$. This procedure is iterated while $||\mathbf{p} - \sum \lambda_i\mathbf{a}_i|| > \alpha$.

The above version of **A2** has the problem that not all samples of the final replacement $\mathbf{a} = \sum \lambda_i\mathbf{a}_i$ obey the constraint in Eqn.3. In order to address this problem, we adjust **A2** to discard samples that satisfy Eqn.3 in the subsequent iterations. In addition, we consider the top K similar blocks in each iteration as it is not case that the closest point provably has the highest effective dimensionality. These two adjustments marry **A1** with GSO to best describe **A2**. Similarly, one can advertise **A1** as a low cost version of GSO because it does not need to perform the similarity search in each iteration. Finally, in this manuscript, we fail to provide detailed description of the algorithms due to brevity. Instead, we present an application of the attack on an off-the-shelf spread spectrum scheme for audio watermarking.

Table 1. Response of a spread-spectrum watermark detector to the replacement attack. Attack parameters are $K = 200$, $r = 5$, $c = 100$, $\alpha = 2.5dB$ and $\varepsilon = 4dB$. The results are obtained for five full songs in different genres. Watermark amplitude equals one. The search space was drastically smaller than in [3] at only 10 seconds of audio from the same song. The table presents information collected from 100 different tests for each test clip: σ_x^2 is signal variance after a moving average filter, $E[\mathbf{a} \cdot \mathbf{w}]^*$ is the correlation response obtained from the watermark detector and normalized with respect to the sample coverage, the total number of samples altered by the attack, the total number of audible samples, their ratio, and finally, the consequent distortion resulting from the attack.

Parameter	Techno	Jazz	Rock	Vocals	Classical	Average
σ_x	3.77	3.71	4.34	4.60	5.49	–
$E[\mathbf{a} \cdot \mathbf{w}]^*$	0.07	0.47	0.50	0.42	0.43	0.4
Covered samples	87364	59126	74847	67716	46693	67149
Audible samples	101359	74673	91974	77985	56410	80480
Coverage [%]	86.19	79.18	81.37	86.83	82.77	83.27
Noise [dB]	2.67	2.56	2.55	2.74	2.60	2.60

3. RANDOMIZED REPLACEMENT FOR AUDIO

Since most psycho-acoustic models operate in the frequency spectrum [16], we launch the replacement attack in the logarithmic (dB) frequency domain. The set of signal blocks Π is created from the coefficients of a modulated complex lapped transform (MCLT) [16]. The MCLT is a $2\times$ oversampled DFT filter bank, used in conjunction with analysis and synthesis windows that provide perfect reconstruction. We consider MCLT analysis blocks with 2048 transform coefficients. Each block of coefficients is normalized and psycho-acoustically masked using an off-the-shelf masking model [16]. Similarity is explored exclusively in the audible part of the 2-7.2KHz frequency spectrum. This is the spectrum commonly used to hide watermarks [10]. Because of psycho-acoustic masking, the actual similarity function in Eqn.2 is not commutative. A replacement block is always masked with the psycho-acoustic mask of the replaced block. Watermarks are spread over 240 consecutive MCLT blocks.

Table 1 shows the response of a spread-spectrum watermark detector to a version of the replacement attack guided by **A1**. Attack parameters are $K = 200$, $r = 5$, $c = 100$, $\alpha = 2.5$dB and $\varepsilon = 4$dB. The results are obtained for five full songs in different genres. Watermark amplitude equals one. The search space was drastically smaller than in [3] at only 10 seconds of audio taken from the same song. The table presents information averaged over 100 different tests for each test clip. Parameter σ_x^2 is signal variance (of $\mathbf{x}$) after applying a moving average filter [10]. Row marked as $E[\mathbf{a}\cdot\mathbf{w}]^*$ is the correlation response obtained from the watermark detector and normalized with respect to the sample coverage. We also present the coverage of the total number of samples altered by the attack, the total number of audible samples, and their ratio. Finally, we demonstrate that the attack on the average caused distortion marginally higher than α. For a search window almost 4 times smaller than in [3] and a signal distortion at only 2.5dB, we have succeeded to almost completely remove the watermark in one song (93% watermark removal) and more than half the correlation in the remaining four clips.

In summary, we identify two potential prevention strategies against a replacement attack. The first one is to enforce a hiding primitive to identify rare parts of the content at watermark embedding time and mark only these blocks. With the randomization of the replacement attack presented in this paper, this attack poses a great computational challenge and reduces significantly the practical significance of such a scheme. Second, in the case of spread-spectrum watermarks, longer and stronger watermarks and increased detector sensitivity may enable watermark detection at lower thresholds (detection thresholds at 0.05–0.1). Unfortunately, such a solution results in a significantly lowered robustness with respect to de-synchronization and estimation attacks.

4. REFERENCES

[1] D. Kirovski, H. Malvar, and Y. Yacobi, "A dual watermarking and fingerprinting system," *ACM Multimedia*, 2002.

[2] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Transactions on Information Theory*, vol.44, no.5, pp.1897–1905, 1998.

[3] D. Kirovski and F.A.P. Petitcolas, "Replacement attack on arbitrary watermarking systems," *ACM Workshop on Digital Rights Management*, 2002.

[4] D. Kirovski and F.A.P. Petitcolas, "Blind pattern matching attack on audio watermarking systems," *ICASSP*, 2002.

[5] C. Rey, G. Doeer, J.-L. Dugelay, and G. Csurka, "Toward generic image dewatermarking," *ICIP*, 2002.

[6] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," *Information Hiding Workshop*, pp.183–206, 1996.

[7] T. Kalker, J. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," *ICIP*, pp.425–429, 1998.

[8] R.J. Anderson and F.A.P. Petitcolas, "On the limits of steganography," *IEEE Journal on Selected Areas in Communications*, vol.16, pp.474–481, 1998.

[9] A. Briassouli and P. Moulin, "Detection-Theoretic Anaysis of Warping Attacks in Spread-Spectrum Watermarking," *ICASSP*, 2003.

[10] D. Kirovski and H. Malvar, "Robust covert communication over a public audio channel using spread spectrum," *Information Hiding Workshop*, pp.354–368, 2001.

[11] M.K. Mihcak, M. Kesal and R. Venkatesan, "Crypto-Analysis on Direct Sequence Spread Spectrum Methods for Signal Watermarking and Estimation Attacks," *Information Hiding Workshop*, 2002.

[12] B. Chen and G.W. Wornell, "Quantisation index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol.47, no.4, pp.1423–1443, 2001.

[13] A.V. Oppenheim and R.W. Schafer, "Discrete-Time Signal Processing," *Prentice-Hall*, 1989.

[14] M.R. Garey and D.S. Johnson, "Computers and Intractability," *W.H. Freeman*, 1979.

[15] H. Cohen, "A Course in Computational Algebraic Number Theory," *Springer-Verlag*, 1993.

[16] H. Malvar, "A modulated complex lapped transform and its application to audio processing," *ICASSP*, 1999.