Replacement Attack on Arbitrary Watermarking Systems

Darko Kirovski[1] and Fabien A.P. Petitcolas[2]

[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
[2] Microsoft Research, 7 J. J. Thomson Avenue, Cambridge, CB3 0FB, UK
{darkok,fabienpe}@microsoft.com

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. A watermark is an imperceptive secret hidden into a host signal. In this paper, we analyze the security of multimedia copyright protection systems that use watermarks by proposing a new breed of attacks on generic 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. Assuming the blocks used to compute the replacement are marked with distinct secrets, we show that if the computed replacement block is at some minimal distance from the original marked block, large portion of the embedded watermark is irreversibly removed. We describe the logistics of the 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. A content screening platform, for example, aims at disabling free downloads from centralized and peer-to-peer file-sharing networks – e.g., Napster alone had orchestrated almost 3 billion downloads of sound clips in February 2001. Several industry-wide initiatives have had little success in establishing a content screening standard [1–3].

1.1 Content Screening

The problem of ensuring copyright at the client side lies in the fact that traditional data protection technologies such as encryption or scrambling cannot be applied as they are prone to digital or analog re-recording (copying). Thus, almost all modern copyright protection mechanisms tend to rely to a certain extent on **watermarks**, imperceptive marks hidden in host signals. In a typical content screening system, the client's media player searches the content for hidden information. If the secret mark is found, the player must verify, prior to playback, whether it has a license to play the content. By default, unmarked content is considered as unprotected and is played without any barriers. A key technology required for content screening is public-key watermarking, that is, a marking scheme where breaking a single player or a relatively large subset of players does *not* compromise the security of the entire system. Such a system, potentially efficient for content screening, has been detailed in [4]. If breaking a single player does not pose a significant security threat, the main target of the adversary is finding a signal processing primitive that removes the watermark or prevents a detector to find it. Several attack mechanisms surveyed in [5] have been largely successful in setting up robustness benchmarks for watermarking technologies. However, none of the attack technologies that do not rely on having access to the watermark detector, remove watermarks without any hope that an irreversible or preventing action is possible.

1.2 The Replacement Attack

In this manuscript, we propose an 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 which 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. We call this type of an attack: a **replacement attack**. The rationale behind this attack is the fact that the replacing block, if at certain minimal distance from the original marked block, conveys little correlation with respect to the watermark embedded in the replaced block as it is created from data that is independent with respect to this watermark. Thus, the newly created content preserves the perceptual similarity with respect to the original clip, while irreversibly cleared of the correlation with the originally embedded watermark. The strategy of this new attack paradigm is simple:

$\boxed{1}$ partition the content into overlapping low-granularity signal blocks,
$\boxed{2}$ for each block B find a subset S of K most perceptually similar blocks,
$\boxed{3}$ compute a block R as a combination of blocks from S, such that the Euclidean distance between R and B is minimal, and
$\boxed{4}$ replace B with R.

In step 2, perceptually similar blocks are originally searched within the original media clip. The search is constrained to a part of the media clip which is assumed to be marked with a different secret compared to block B. If the computed replacement block R is at an Euclidean distance which is higher than some predefined fidelity constraint (e.g., $|B - R| < 4\text{dB}$), the adversary can alternatively seek replacement blocks in an external multimedia library. The distance between R and B must not be small, because the replaced block in that case preserves certain correlation proportional to the similarity. For example, if $B = R$ then the attack does not affect the existence of the watermark. Thus, if R is at a distance which is closer than a certain lower bound, it is recomputed as to increase the similarity beyond that bound.

Finding perceptually similar blocks of certain music or video content is a challenging task. With no loss of generality, 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 both background and 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. In this paper, we explore the challenges of the replacement attack and show how it can be launched on audio content.

2 Logistics of the 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) [7] are all prone to the replacement attack. In order to launch the attack successfully, the adversary does not need to know the details of the watermark codec. The adversary needs to reduce the granularity of integral blocks of data such that no block contains enough information from which a watermark can be identified individually. Note that watermark detection involves processing large amount of data (for example, reliable and robust detection of audio watermarks requires at least several seconds of audio [8]). Thus, blocks considered for replacement must be at least one order of magnitude smaller than watermark length. For both audio and video, this requirement is not difficult to satisfy as typically blocks of $256 - 2048$ transform coefficients for audio or bitmaps of up to 64×64 pixels for video are considered for pattern matching.

In the remainder of this section, we assume that coefficients of the marked signal are replaced only with other coefficients of the same signal. It is straightforward to redefine the attack such that coefficients from external signal vectors are considered as a substitution base.

The **host signal** to be marked $\mathbf{x} \in \mathcal{R}^N$ can be modeled as a vector, where each element $x_i \in \mathbf{x}$ is a zero-mean independent identically distributed normal random variable[1] with standard deviation σ_x: $x_i \sim \mathcal{N}(0, \sigma_x)$. The replacement

[1] This model is adopted for the purpose of analyzing the watermark detector. Reality shows that the model is not memoryless as parts of the signal tend to repeat, slightly distorted, both in music and video.

attack is not restricted to a particular signal model; we use the Gaussian assumption to analyze certain properties of the attack. A **watermark** is defined as an arbitrary pseudo-randomly generated vector $\mathbf{w} \in \mathcal{R}^N$, where each element $w_i \in \mathbf{w}$ is a random variable with standard deviation $\delta \ll \sigma_x$. For example, if direct sequence spread-spectrum is used for watermark modulation then $\mathbf{w} \in \{\pm\delta\}^N$. We assume that the watermark $\mathbf{w}$ is mutually independent with respect to $\mathbf{x}$. The **marked signal** $\tilde{\mathbf{x}}$ is created as $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{w}$. The replacement attack receives as input the marked signal $\tilde{\mathbf{x}}$ and outputs its modification $\tilde{\mathbf{x}}'$.

2.1 Attack steps

Signal partitioning. In the initial step of the attack, the watermarked content $\tilde{\mathbf{x}}$ is partitioned into a set $\mathcal{B}$ of overlapping blocks, where each block B_p represents a sequence of m samples of $\tilde{\mathbf{x}}$ starting at $\tilde{x}_{(B_p)}$. (B_i) denotes the index of the first sample in the i-th block B_i. For an overlap ratio of η, the total number of blocks equals $n = \lceil \frac{N-m}{1-\eta} \rceil$. The higher the overlap, the larger the search-space for the replacement attack. We want to select the overlap such that:

1. consecutive blocks do not have similar perceptual characteristics – this upper bound on block overlap aims at reducing the search space – and
2. for two consecutive blocks B_p and B_{p+1} starting at $\tilde{x}_{(B_p)}$ and $\tilde{x}_{(B_{p+1})}$ respectively, the block starting at $\tilde{x}_a$, $a = [(B_p) + (B_{p+1})]/2$ is not perceptually similar to B_p or B_{p+1}.

Similarity function. This is the core function of the replacement attack. It takes as an input a pair of blocks B_p and B_q and returns a real number $\phi(B_p, B_q) \geq 0$ that quantifies their similarity. Block equality is represented as $\phi(B_p, B_q) = 0$. The adversary can experiment with a number of different functions. In this section, we restrict similarity to the root-mean-square distortion between blocks:

$$\phi(B_p, B_q) = \frac{1}{\sqrt{m}} ||B_p - B_q|| = \sqrt{\frac{1}{m} \sum_{i=0}^{m-1} \left[\tilde{x}_{i+(B_p)} - \tilde{x}_{i+(B_q)}\right]^2}. \quad (1)$$

Search for the substitution base. This step is repeated for each block (target B) of the original media clip. The goal is to find a subset S of K perceptually similar blocks to B. An additional constraint is that the subset S is searched in part of the signal which is marked with watermark bits different with respect to the watermark present in the target. Usually, watermark bits are replicated within certain vicinity of the target [8], which excludes several neighboring blocks to the target from the search process. The following principles guide the search process.

1. **Lower bound on similarity** – the computed block R replacing B must be at a certain minimal distance from B, i.e. $\phi(B, R) \geq \alpha$. This requirement

stems from the fact that replacing a signal block with another, exceptionally similar block has only nominal impact on watermark existence. Parameter α depends on the watermark amplitude as well as the type of watermark modulation (e.g., direct-sequence spread spectrum or QIM). It is discussed further in Subsection 2.2.

2. **Upper bound on similarity** – the upper limit ensures that the resulting clip has a preserved high fidelity with respect to the marked copy. If the search procedure cannot find a subset of blocks that can linearly combine to create a replacement block R that is sufficiently similar to B, $\phi(B, R) \leqslant \beta$, then $R = B$, i.e. replacement does not occur. Note that the search procedure is not limited to the host media clip – the subset of similar blocks can be extracted from a large library of media clips.

Based on the above principles, the algorithm for computing the replacement block R takes the following steps. In the first step, the algorithm finds two pools of blocks, S' and S. The first pool S' contains all blocks from the substitution database which are at distance $(\forall B_i \in S')\phi(B_i, B) < \alpha$. Frequently, with the exception of electronically generated music, this pool is empty as it is hard to find exceptionally similar blocks in music performed by humans. The second pool S contains K most similar blocks to B that are at distance $(\forall B_i \in S)\phi(B_i, B) \geq \alpha$. Parameter K should be significantly smaller than the length of a block. For 256- to 2048-long audio blocks, values within $10 \leqslant K \leqslant 50$ result in good balance for fidelity and performance. Although parameter K can, in general, be variable across blocks, in our experiments we consider only constant K. The complexity of finding S is linearly proportional to the size of the replacement database.

```
for each block $B$
  find $S' \subset \mathcal{B} | (\forall B_i \in S')\phi(B_i, B) < \alpha$
  find $S \subset \mathcal{B} | (\forall B_i \in S)\phi(B_i, B) \geq \alpha$
  create matrix $\mathbf{s}$ such that each row in $\mathbf{s}$ is a distinct block from $S$
  compute $R' = \mathbf{s}(\mathbf{s}^T\mathbf{s})^{-1}\mathbf{s}^T B$
  depending on $\phi(R', B)$ and $S'$, set $R$ according to rules (i-iv)
  replace $B$ with $R$
end for
```

In the next step, the replacement block R is computed from the selected blocks in S such that its similarity with respect to R is maximized. More formally, we construct a matrix $\mathbf{s} \in \mathbb{R}^{K \times m}$ where each row of this matrix represents one block from S. We aim to compute a vector A such that $||\mathbf{s}A - B||$ is minimized. The least-squares solution to this set of overdetermined linear equations, commonly called pseudo-inverse of $\mathbf{s}$, equals $A = (\mathbf{s}^T\mathbf{s})^{-1}\mathbf{s}^T B$. A temporary replacement block R' is now computed as $R' = \mathbf{s}A$. Four cases can occur:

(i) the temporary replacement block R' satisfies the requirements, e.g., $\alpha \leqslant \phi(B, R') \leqslant \beta$, in which case the replacement equals $R = R'$, *(ii)* R' is too distorted and subset S' is empty, e.g., $\phi(B, R') > \beta$ and $S' = \varnothing$, in which case no replacement occurs $R = B$ for preserved signal fidelity,

(iii) R' is too distorted and S' is not empty, e.g., $\phi(B, R') > \beta$ and $S' \neq \varnothing$, in which case R' and a randomly chosen block T from S' are mixed as $R = (1-q)T + qR'$ such that $\phi(R, B) = \alpha$, and

(iv) R' is too similar to B, e.g., $\phi(B, R') < \alpha$, in which case R' and a randomly chosen block T from S are mixed as $R = (1-q)T + qR'$ such that $\phi(R, B) = \alpha$.

The mixing parameter q enforces the desired similarity $\phi(B, R) = \alpha$ in the last two cases if:

$$q = \frac{\varepsilon^2 - \sqrt{\alpha^2(\vartheta^2 + \varepsilon^2) - \vartheta^2\varepsilon^2}}{\vartheta^2 + \varepsilon^2}, \tag{2}$$

where $||R' - B||^2 = m\vartheta^2$ and $||T - B||^2 = m\varepsilon^2$ under the assumption that $T - B$ and $R' - B$ are mutually independent[2].

Block substitution. In the final step, each block B of the original watermarked signal is replaced with the corresponding computed replacement R to create the output media clip $\tilde{\mathbf{x}}'$.

2.2 Determining α for Spread-Spectrum Watermarks

Lets assume that vector $\mathbf{x} + \mathbf{w}$ is deemed similar to and replaced by vector $\mathbf{y} + \mathbf{v}$, where $\mathbf{x}$ and $\mathbf{y}$ are original signals marked with two distinct watermarks $\mathbf{w}$ and $\mathbf{v}$, where $\mathbf{w}, \mathbf{v} \in \{\pm\delta\}^m$. All vectors are assumed to have the same length as a single block: m. In addition, we assume that the watermarks are spread-spectrum sequences, which means that watermark $\mathbf{w}$ is detected in a signal $\mathbf{z}$ by matched filtering: $C(\mathbf{z}, \mathbf{w}) = \mathbf{z}^{\mathrm{T}}\mathbf{w}$. If $\mathbf{z}$ has been marked with $\mathbf{w}$, $\mathrm{E}[C(\mathbf{z}, \mathbf{w})] = m\delta^2$, otherwise $\mathrm{E}[C(\mathbf{z}, \mathbf{w})] = 0$, with variance $\mathrm{Var}[C(\mathbf{z}, \mathbf{w})] = m\sigma_z^2$. Watermark is detected if $C(\mathbf{z}, \mathbf{w})$ is greater than a certain detection threshold τ. In order to have symmetric probability of a false alarm and misdetection, τ is commonly set to $m\delta^2/2$. From the requirement for two blocks to be eligible for substitution:

$$E[||(\mathbf{y} + \mathbf{v}) - (\mathbf{x} + \mathbf{w})||^2] = E[||\mathbf{y} - \mathbf{x}||^2] + 2m\delta^2 - 2E[C(\mathbf{v}, \mathbf{w})] \geqslant m\alpha^2, \tag{3}$$

we can compute the expected resulting correlation $\mathrm{E}[C(\mathbf{y} + \mathbf{v}, \mathbf{w})]$ under the assumption that vectors $\mathbf{v}$ and $\mathbf{w}$ are independent with respect to $\mathbf{x}$ and $\mathbf{y}$:[3]

$$E[C(\mathbf{y} + \mathbf{v}, \mathbf{w})] \leqslant \frac{1}{2}E[||\mathbf{y} - \mathbf{x}||^2] + m(\delta^2 - \frac{\alpha^2}{2}) \tag{4}$$

[2] In case *(iv)* the two vectors are not mutually independent as R' is dependent upon T. To address this issue, we select T as the block from S which has the smallest absolute value of the corresponding coefficient in the vector A that builds R'.

[3] $\mathrm{E}[C(\mathbf{y}, \mathbf{v})] = \mathrm{E}[C(\mathbf{x}, \mathbf{v})] = \mathrm{E}[C(\mathbf{y}, \mathbf{w})] = \mathrm{E}[C(\mathbf{x}, \mathbf{w})] = 0$.

Assuming that there exists true repetition of the original content $\mathbf{y} = \mathbf{x}$, then setting $\alpha \geqslant \delta\sqrt{2}$ would set the expected minimum correlation to zero after substitution. If $\mathbf{y} \neq \mathbf{x}$, then α needs to be additionally increased to compensate for the effect of $E[||\mathbf{y}-\mathbf{x}||^2]$ on resulting correlation. Quantifying this compensation analytically is difficult as it depends upon the self-similarity of the targeted content.

3 A Replacement Attack for Audio

In this section, we demonstrate how the generic principles behind the replacement attack can be applied against an audio watermarking technology. We first describe how an audio signal is partitioned and pre-processed for improved perceptual pattern matching. Next, we analyze the similarity function we used for our experiments. The effect of the replacement attack on direct-sequence spread-spectrum watermark detection is presented in the following sections.

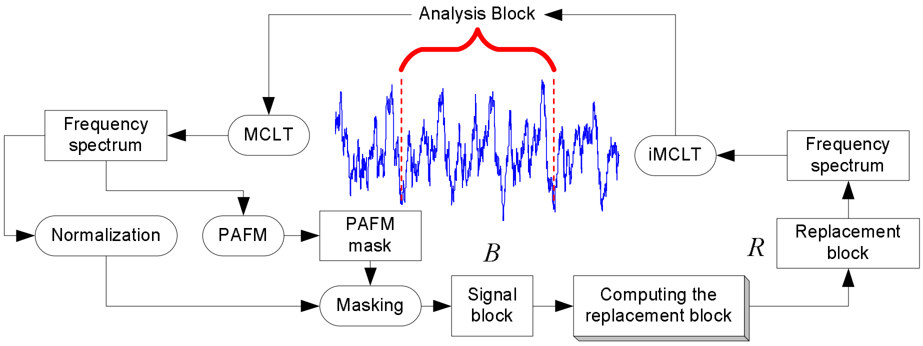

Fig. 1. Block diagram of the signal processing primitives performed as pre- and post-processing to the replacement attack.

3.1 Audio Processing for the Replacement Attack

Since most psycho-acoustic models operate in the frequency spectrum [9], we launch the replacement attack in the logarithmic (dB) frequency domain. The set of signal blocks $\mathcal{B}$ is created from the coefficients of a modulated complex lapped transform (MCLT) [9]. 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 and an $\eta = 0.25$ overlap. Each block of coefficients is normalized and psycho-acoustically masked using an off-the-shelf masking model [9]. Similarity is explored exclusively in the audible part of the frequency spectrum. Because of psycho-acoustic masking, the actual similarity function in Eqn.1 is not commutative. A replacement block is always masked with the psycho-acoustic mask of the replaced block. Figure 1 illustrates the signal processing primitives used to prepare blocks of audio content for substitution.

Watermark length is assumed to be greater than one second. In addition, we assume that watermark chips may be replicated along the time axis at most for one second[4] [8]. Thus, we restrict that for a given block its potential substitution blocks are *not* searched within one second.

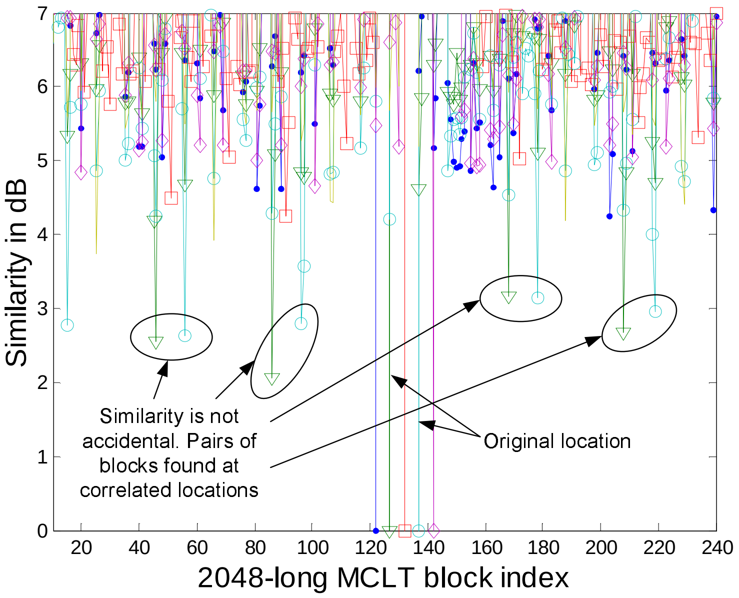

Fig. 2. Music self-similarity: a similarity diagram for five different 2048-long MCLT blocks within a techno clip with 240 MCLT blocks. Zero-similarity denotes equality. The abscissa x denotes the index of a particular MCLT block. The ordinate denotes the similarity $\phi(x, B_i)$ of the corresponding block x with respect to the selected five blocks with indices $B_i | i = \{122, 127, 132, 137, 142\}$.

3.2 Analysis of the Similarity Function

We performed several experiments in order to evaluate the effectiveness of the replacement attack. The first set of experiments aims at quantifying similarity between blocks of several audio clips marked with spread-spectrum watermarks

[4] Higher level of redundancy may enable effective watermark estimation.

at $\delta = 1$dB. In all examples, block similarity is computed over the 2–7kHz sub-band as watermark codecs commonly hide data in a sub-band that is not strongly distorted by compression and medium quality low- and high-pass filtering [8]. Figure 2 shows the values of the similarity function $\phi(B_i, B_j)$ for five 2048-long MCLT blocks at positions $i = \{122, 127, 132, 137, 142\}$ against a database of 240 blocks $j = \{1 \ldots 240\}$ within one audio clip (techno music). We observe that throughout the database four different pairs of blocks (circled in the subfigure) are found as similar below 4dB to the pair of blocks with indices 127 and 137. All similar pairs of blocks preserve the same index distance as the target pair. This points to the fact that in many cases content similarity is not a result of coïncidence, but a consequence of repetitive musical content.

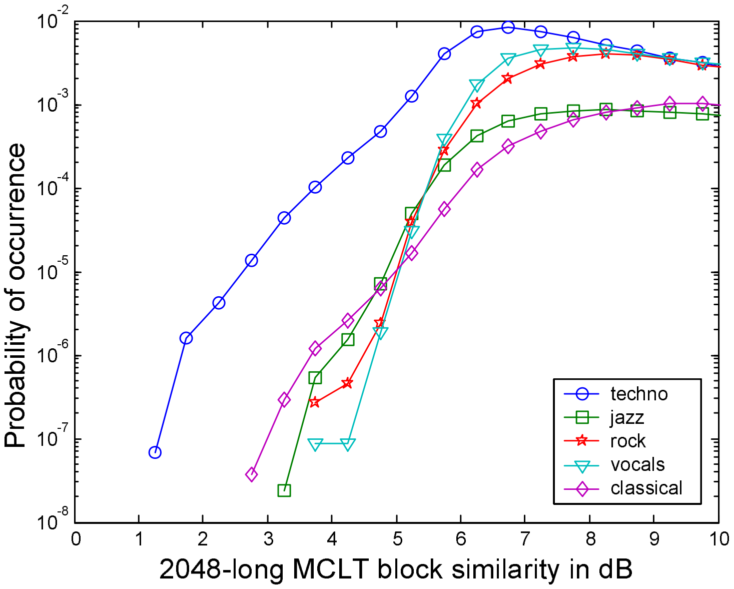

Fig. 3. Music self-similarity: probability density function of the similarity function $\phi(B_i, B_j)$ within an audio clip – five different types of music are considered: rock, classical, jazz, vocals and techno. A certain value x on the abscissa represents a histogram bin from $x - 0.25$ to $x + 0.25$ dB.

Figure 3 illustrates the probability that for a given 2048-long MCLT block B_i, there exists another block B_j within the same audio clip that is within $\phi(B_i, B_j) \in [x - 0.25, x + 0.25\}$dB, where x is a real number. This experiment was conducted for five different types of audio content: techno, jazz, rock, vocals, and classical music. For this benchmark set of distinctly different musical pieces, we conclude that the average $\phi(B_i, B_j)$ for two randomly selected blocks within an audio clip is in the range of 6–8dB. The probability of finding a similar block should rise proportionally to the size of the substitution database, especially if it consists of clips of the same genre/performer. Finally, note that electronically generated music (in our benchmark a techno song) is significantly more likely to contain perceptually correlated blocks than music that is performed by humans.

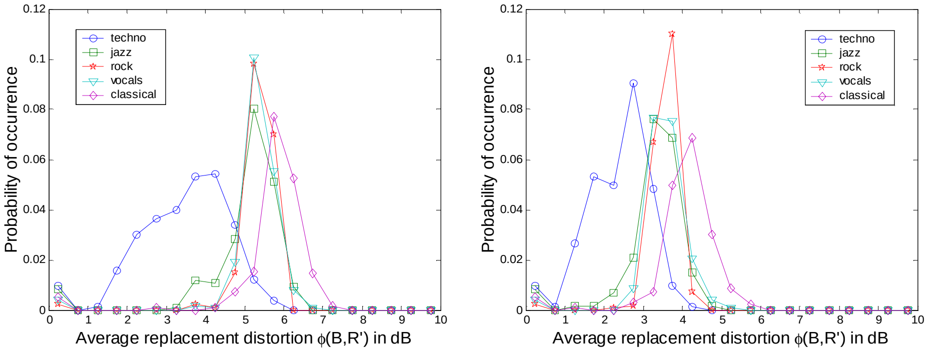

Fig. 4. Probability density function of the similarity function $\phi(B, R')$ for two different cases: $K = 1$ (left) and $K = 10$ (right).

Table 1. Improvement in signal distortion due to the replacement attack as parameter K is increased. Results are reported on the dB scale. Average block length is $\bar{m} \approx 400$.

$\phi(B,R')$	Techno	Jazz	Rock	Vocals	Classical	Average $\phi_K(B,R') - \phi_{K=1}(B,R')$
K=1	3.5714	5.2059	5.3774	5.3816	5.8963	N/A
K=10	2.3690	3.2528	3.5321	3.5193	4.0536	1.741
K=20	2.2666	3.0576	3.3792	3.3664	3.7968	1.914
K=30	2.2059	2.9255	3.3061	3.2762	3.5613	2.032
K=50	2.1284	2.6595	3.1702	3.1209	3.0635	2.253
K=100	1.9512	2.1331	2.8631	2.7439	1.8719	2.775

The second set of experiments explores the distortion that the replacement attack introduces. We consider three cases. In the first case, in the left subfigure of Figure 4, we present the probability that the replacement block R' is at distance $\phi(B, R')$ if R' equals the most similar block found in the substitution database (e.g., $K = 1$). The right subfigure presents the same metric for the case when $K = 10$ and R' is computed as described in Subsection 2.1. Finally, Table 1 quantifies the improvement in the average distortion $\phi(B, R')$ as K increases from 1 to 100. We conclude that the replacement attack in our experimental setup induces between 1.5–3dB distortion noise with respect to the marked copy – a change in fidelity that most users are willing to sacrifice for free content.

4 Effect of the Attack on Watermark Detection

In order to evaluate the effect of a replacement attack on spread-spectrum watermarks, we conducted two experiments. For both experiments, we used spread-spectrum watermarks that spread over 240 consecutive 2048-long MCLT blocks (approximately 11sec long), where only the audible frequency magnitudes in the 2–7kHz subband were marked. We did not use chip replication as its effect on watermark detection is orthogonal with respect to the replacement attack.

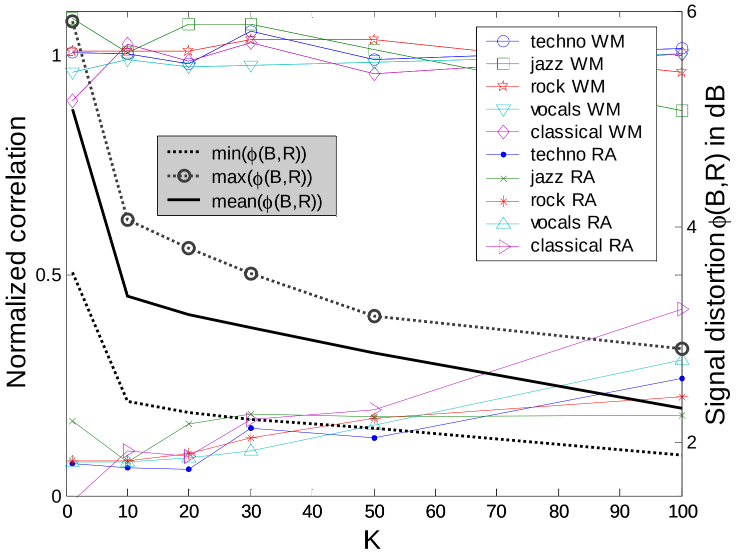

Fig. 5. Response of a spread-spectrum watermark detector to the replacement attack. The abscissa quantifies the change in parameter K from 1 to 100 for fixed watermark amplitude of $\delta = 1$dB. The left ordinate shows the increase of the normalized correlation as K increases. The results are obtained for five full songs in different genres. The right ordinate shows the corresponding minimal, maximal, and average distortion with respect to the set of benchmark clips due to the replacement attack.

Figure 5 shows how normalized correlation of a spread-spectrum watermark detector is affected by the increase of the parameter K. We performed the following experiment. We marked the first 240 2048-long MCLT blocks of five different songs (ranging from 3 to 5 minutes in duration) with a direct sequence spread spectrum watermark. The watermark amplitude was set to $\delta = 1$dB. During the attack, we replaced each target block B with its computed replacement block R following the recipe presented in Subsection 2.1. For the purpose of demonstrating the change of the correlation due to increase in K, we did not apply steps (*iii*-*iv*). When these steps are applied the minimal distortion per block is limited to αdB.

In Figure 5, we show two results. First, we show the average normalized correlation value (left ordinate) across 10 different tests for watermark detection within marked content $E[C(\mathbf{z}, \mathbf{w})] = 1$ (curves marked WM) and within marked content attacked with our attack for several values of $K = \{1, 10, 20, 30, 50, 100\}$ (curves marked RA). Second, we show on the right ordinate the signal distortion caused by the replacement attack: the minimal, average, and maximal distortion across all five audio clips. We can conclude from the diagram that for small values of K, its increase results in greatly improved distortion metrics, while for large values of K, the computed replacement vectors are too similar with respect to the target blocks which results in lower effect on the normalized correlation.

The power of the replacement attack is most notably observed by comparing the effect of adding a white Gaussian noise (AWGN) pattern $\mathbf{n} = \mathcal{N}(0, \sigma_n)$ of certain standard deviation $\sigma_n \in \{2 \dots 3\}$dB to a replacement attack of equivalent distortion. Whereas the dramatic effect of replacement can be observed in Figure 5, AWGN affects the correlation detector only negligibly. In the latter case, the expected correlation value remains the same $E[C(\tilde{\mathbf{x}} + \mathbf{n}, \mathbf{w})] = E[C(\tilde{\mathbf{x}}, \mathbf{w})]$, with increased variance $Var[C(\tilde{\mathbf{x}} + \mathbf{n}, \mathbf{w})] = Var[C(\tilde{\mathbf{x}}, \mathbf{w})] + m\sigma_n^2$. Finally, additive noise of 2–3dB in the 2–7kHz subband is a relatively tolerable modification.

Another important issue is the fact that the distortion introduced by the replacement attack is linearly proportional to the watermark amplitude. Clearly, with the increase of watermark amplitude δ, the search process of the replacement attack becomes harder for two reasons: (*i*) block contents become more randomized and (*ii*) the substituted blocks are more correlated with the original blocks. On the other hand, we have empirically concluded that watermark amplitude affects the reduction of the normalized correlation minimally. Although stronger watermarks may sound like a solution to the replacement attack, high watermark amplitudes cannot be accepted because of two reasons: first, the requirement for high-fidelity marked content and second, strong watermarks can be efficiently estimated using an optimal watermark estimator [4], i.e. estimate $\mathbf{v} = \text{sign}(\mathbf{x} + \mathbf{w})$ makes an error per bit $\varepsilon = \Pr[v_i \neq w_i] = \frac{1}{2}\text{erfc}(\frac{\sigma_x}{\delta\sqrt{2}})$ exponentially proportional to δ.

5 Conclusion

For any watermarking technology and any type of content, one powerful attack is to re-record the original content, e.g., perform again the music or capture the image of the same original visual scene. In this paper, we emulate this attack using a computing system: the replacement attack aims at replacing small pieces of the marked content with perceptually similar but unmarked[5] substitution blocks created using a library of multimedia content. The hope is that the substitutions have little correlation with the original embedded mark. Inspired by predictive coding of speech and video [15, 10], we present an algorithm for computing the replacement blocks using a least-squares linear combination of K signal blocks most similar to the target block.

Although the attack is generic and can be applied to all marking strategies, we demonstrate how it can be launched for audio content and a traditional watermarking modulation technology: direct sequence spread-spectrum. Our preliminary results demonstrate that the attack has similar effect on other marking mechanisms such as quantization index modulation.

From the presented experimental results, we conclude that an implementation of the replacement attack that considers a relatively small substitution database can create replacement blocks that are only within 1.5–3dB distance with respect to the target signal blocks. Such an attack removes approximately 80–90% of the correlation between the watermark and the marked/attacked content. Similar adversarial effects can be obtained against QIM-based watermarking schemes.

We identify two possible prevention strategies against a replacement attack. For example, a data hiding primitive may identify rare parts of the content at watermark embedding time and mark only these blocks. However this reduces significantly the practical capacity of the scheme and increases dramatically the complexity of the embedding process. In the case of spread-spectrum watermarks, longer watermarks and increased detector sensitivity may enable watermark detection at lower thresholds (e.g., $\tau < m\delta^2/10$). Unfortunately, such a solution comes at the expense of having significantly longer watermarks which results in a significantly lowered robustness with respect to de-synchronization attacks.

References

1. Andy Patrizio, "DVD piracy: It can be done," November 1st, 1999, `http://www.wired.com/news/technology/0,1282,32249,00.html`.
2. "Secure Digital Music Initiative," http://www.sdmi.org.
3. "The DVD Copy Control Association," http://www.dvdcca.org.
4. Darko Kirovski, Henrique Malvar, and Yacov Yacobi, "A dual watermarking and fingerprinting system," Tech. Rep. MSR-TR-2001-57, Microsoft Research, June 2001.
5. Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn, "Attacks on copyright marking systems," In [11], pp. 218–238.
6. Ingemar J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon, "A secure, robust watermark for multimedia," In [12], pp. 183–206.
7. Brian Chen and Gregory W. Wornell, "Quantization 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, May 2001.

[5] Or marked with a different watermark.

8. Darko Kirovski and Henrique Malvar, “Robust covert communication over a public audio channel using spread spectrum,” In [14], pp. 354–368.
9. Henrique Malvar, “A modulated complex lapped transform and its application to audio processing,” in *International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, USA, March 1999, pp. 1421–1424.
10. Mehmet Kıvanç Mıhçak, *Personal Communication.*
11. David Aucsmith, Ed., *Information Hiding: Second International Workshop*, vol. 1525 of *Lecture Notes in Computer Science*, Portland, Oregon, USA, 1998. Springer-Verlag, Berlin, Germany.
12. Ross J. Anderson, Ed., *Information hiding: first international workshop*, vol. 1174 of *Lecture Notes in Computer Science*, Isaac Newton Institute, Cambridge, UK, May 1996. Springer-Verlag, Berlin, Germany.
13. Ping Wah Wong and Edward J. Delp, Eds., *Security and Watermarking of Multimedia Contents II*, vol. 3971, San Jose, California, U.S.A., 24–26 Jan. 2000. The Society for Imaging Science and Technology (IS&T) and the International Society for Optical Engineering (SPIE), SPIE.
14. Ira S Moskowitz, Ed., *Information hiding: fourth international workshop (IH’2001)*, vol. 2137 of *Lecture Notes in Computer Science*, Pittsburgh, Pennsylvania, U.S.A., 2001. Springer-Verlag, Berlin, Germany.
15. Jerry D. Gibson, Toby Berger, David Lindbergh, and Richard L., III Baker, *Digital Compression for Multimedia : Principles and Standards*, Morgan Kaufmann Publishers, San Francisco, CA, USA, January 1998.