THE BLIND PATTERN MATCHING ATTACK ON WATERMARK SYSTEMS

Fabien A. P. Petitcolas
Microsoft Research
7 J. J. Thomson Avenue
Cambridge CB3 0FB, UK
{fabienpe@microsoft.com}

Darko Kirovski
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{darkok@microsoft.com}

ABSTRACT

Billions of dollars allegedly lost to piracy of multimedia content have recently triggered the industry to rethink the way how music and movies are distributed on the Internet. As encryption is vulnerable to digital or analog re-recording, currently almost all copyright protection mechanisms rely to certain extent on watermarking, i.e. hiding of imperceptive secrets into a host signal. In this paper, we propose a new breed of attacks on generic watermarking systems, which recognizes that multimedia content is often highly repetitive, identifies subsets of signal blocks that are similar, and finally permutes these blocks. Assuming the permuted blocks have been marked with distinct secrets, it can be shown that any watermark detector is facing a task of exponential complexity to reverse the permutations as a preprocessing step for watermark detection. In this paper, we describe the logistics of the attack and a recipe for its implementation against an audio watermarking technology.

1. INTRODUCTION

As long as the movie and music industry do not find a business model that actually benefits the consumer in downloading multimedia content from content publisher's servers, movie and music piracy is likely to reign the Internet traffic. For example, only in the month of February 2001, Napster has orchestrated download of almost 3 billion sound clips from its distributed file sharing system. Simultaneously, recently several industry-wide initiatives have had little success in preventing this trend [1], [2].

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 rely to certain extent on WMs[1], imperceptive marks hidden in host signals. Although several attack mechanisms have been largely successful in setting up robustness benchmarks for watermarking technologies (e.g. Stirmark [3]), in this paper, we propose *blind pattern matching attacks* (BPM), a new breed of attacks against generic WM systems. The strategy of a BPM attack is simple:

- partition the content into overlapping low-granularity signal blocks,
- identify subsets of perceptually similar blocks, and
- pseudo-randomly permute their locations in the signal.

As the number of perceptually unique blocks can be insufficient for a successful launch of the attack, the adversary can alternatively seek for replacement blocks in an external large multimedia library. The hope is that large percentage of the original signal will be replaced or perturbed such that WM detection is nearly impossible. Assuming that the adversary has not used an external library and that the permuted blocks have been marked with distinct secrets, it can be shown that any WM detector is facing a task of exponential complexity to reverse the permutations as a preprocessing step for correct WM detection.

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 much better source of repetitive content within a single recording.[2] 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 BPM attack and present a recipe how it can be launched on audio content.

1.1 Prior Art

As soon as people have tried to develop watermarking technologies, others have attempted to break them. Early ones, such as random geometric distortions [3] relied on the fact that most watermarking algorithms are based on some form of correlation, which itself requires good alignment properties. Breaking this alignment usually prevents reliable detection. In fact there are two main types of attacks: those who attempt to remove the WM and those who just prevent the detector from detecting them. Random geometric distortions fall in the second category.

Attacks in the first category usually try to estimate the original non-watermarked cover-signal and they usually consider the WM as noise with given statistic. For instance, Langelaar et al. showed that 3×3 median filtering gives a good approximation of original pictures in the case they have been watermarked using spread-spectrum [4]. So far, estimate-and-remove attacks have introduced fairly strong blurring effects but recent work based on maximum a posteriori WM estimation and remodulation has given promising results [5]. Unfortunately, so far most attacks have targeted still image watermarking schemes and very few have tried to deal with audio watermarking. Attacks on some audio watermarking schemes include echo removal or signal restoration [6]. These attacks are not very general as they can only be applied to certain watermarking algorithms. An obvious improvement for audio attacks is the equivalent of random geometric distortions. This is in fact a mixture of time and frequency scaling. Although robustness to this type of attack is not fully solved in the case of images, there are some audio data hiding algorithms that can cope with them already [7].

[1] WM - watermark.

[2] For example, within a common scene both background and objects experience geometric transformations significantly more frequently than changes in appearance.

2. THE BPM ATTACK

2.1 The Generic Approach

The BPM attack is not limited to a type of content or to a particular watermarking algorithm. For example, systems that modulate secrets using spread-spectrum (SS) [8] and/or quantization index modulation [9] are all prone to the BPM attack. In order to launch the attack successfully, the adversary does not need to know the details of the WM codec. In addition, the adversary needs to reduce the granularity of integral blocks of data such that no block contains enough information from which a WM can be identified individually. Note that WM detection involves processing large amount of data (for example, reliable and robust detection of audio WMs requires at least several seconds of audio [7]). Thus, blocks considered for BPM must be at least one order of magnitude smaller than WM length. For both audio and video, this requirement is not difficult to satisfy as typically blocks of 256-1024 transform coefficients for audio or bitmaps of up to 64x64 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 $x \in \mathcal{R}^N$ can be modeled as a vector, where each element $x_i \in x$ is a zero-mean independent identically distributed random variable (r.v.) with standard deviation σ_x, i.e. $x_i \sim \mathcal{N}(0, \sigma_x)$.[3] A *watermark* is defined as an arbitrary pseudo-randomly generated vector $w \in \mathcal{R}^N$, where each element $w_i \in w$ is a r.v. with standard deviation $\delta \ll \sigma_x$. For example, if direct sequence SS is used for WM modulation then $w \in \{\pm\delta\}^N$. The WM signal w is mutually independent with respect to x. The *marked signal* y is created as $y = x + w$.

The BPM attack algorithm receives as input the marked signal y and outputs its modification z. The algorithm has several steps:

[A.] ***Signal partitioning.*** In this step, content y is partitioned into a set of overlapping blocks B, where each block B_i represents a sequence of m samples of y starting at $y_{(B_i)}$.[4] 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 BPM attack. We want to select the overlap such that: (i) consecutive blocks do not have near-equivalent perceptual characteristics and (ii) for two consecutive blocks B_i and B_{i+1} starting at $y_{(B_i)}$ and $y_{(B_{i+1})}$ respectively, the block starting at y_a, $a = [(B_i) + (B_{i+1})]/2$ is not perceptually similar to B_i or B_{i+1}.

[B.] ***The similarity function.*** This is the core function of the BPM attack. It takes as an input a pair of blocks B_i and B_j and returns a real number $\phi(B_i, B_j) \geq 0$ that quantifies their similarity. Block equality is represented as $\phi(B_i, B_j) = 0$. The adversary can experiment with a number of different functions. In this section, we restrict similarity to the quadratic euclidian distance between blocks:

$$\phi(B_p, B_q)^2 = \sum_{i=0}^{m-1} \left[y_{i+(B_p)} - y_{i+(B_q)} \right]^2 . \quad (1)$$

[3]The BPM attack is not restricted to a particular signal model. However, we use the Gaussian assumption to analyze certain properties of the attack.

[4](B_x) denotes the index of the first sample in block B_x.

[C.] ***Pattern matching.*** In this step, perceptual similarities between individual signal blocks are identified. The result of this phase is a similarity bit-matrix S, with elements $S_{p,q}, p = 1..n, q = 1..n$ such that:

$$S_{p,q} = \begin{cases} 1 & , \quad m\alpha^2 \leqslant \phi(B_p, B_q)^2 \leqslant m\beta^2 \\ 0 & , \quad otherwise \end{cases} \quad (2)$$

where α^2 and β^2 are parameters that denote minimal and maximal average similarity respectively, for a pair of blocks to be considered as substitutable. The lower limit is required because substituting a block with another of exceptional similarity does not affect WM detection. The upper limit is required for attack high-fidelity.

We compute the S matrix in the following way. Initially, we compute the block-level auto-covariance of the input signal y. Then, for all pairs of blocks that correspond to a high value of the covariance matrix, we compute the accurate similarity function. The decision threshold for computing the similarity function is determined empirically. Thus, we are able to accurately estimate the perceptual similarity while retaining the algorithm complexity of an fft-based cross-correlation at $O(mnlog(n))$.

Note that the computational complexity of obtaining S can be significantly improved if we consider only block swapping. In that case, we can consider deploying a first-degree predictor that assumes that similarity is a continuous function and predicts that similar blocks are most often found after already identified similar blocks.

[D.] ***Block substitution.*** In the final step, we create the resulting attacked signal z by relocating blocks according to the realized similarities. The substitution procedure is presented using the following pseudo-code:

```
Copy z = y.
Mark all blocks in B as unvisited.
◇ Find first unvisited block B_i s.t. ∃j ≠ i|S_{i,j} = 1.
Let G ⊂ B s.t. (∀B_j ∈ G)S_{i,j} = 1 or i = j.
Let L be a random permutation of elements in G.
Reorder blocks of z with indices G according to L.
Mark blocks with indices in G as visited.
GO TO ◇
```

The effectiveness of the BPM depends on several factors. First, block size is a variable with an important trade-off. It is difficult to find large similar blocks, so the search clearly benefits from smaller blocks. On the other hand, it is difficult to estimate perceptual factors in small blocks. In addition, smaller blocks tend to preserve higher correlation between the original and the substitute that reduces the effect of BPM on the reliability of the WM detector. On the example of audio, we elaborate on this trade-off more in the next section. Second, the content itself may have little redundancy regardless of block size and relative looseness of the upper bound on similarity β. In that case, the adversary needs to search a larger database of content in hope that similar sounds or pixel-maps can be found. Finally, what is a proper minimal value of the lower bound α which allows that substituted content actually affects the WM detector? Lets assume that vector $x + w$ is deemed similar to and is replaced by vector $y + v$, where x and y are original signals marked with two distinct WMs w and v. All vectors are assumed to be of the size of a single block m. In addition, we assume that the WMs are SS sequences, which means that WM w is detected in a signal z by matched filtering: $C(z, w) = z \cdot w = \sum_{i=1}^{m} z_i w_i$. If z has been marked with w, $E[C(z, w)] = m$, else $E[C(z, w)] = 0$, with variance

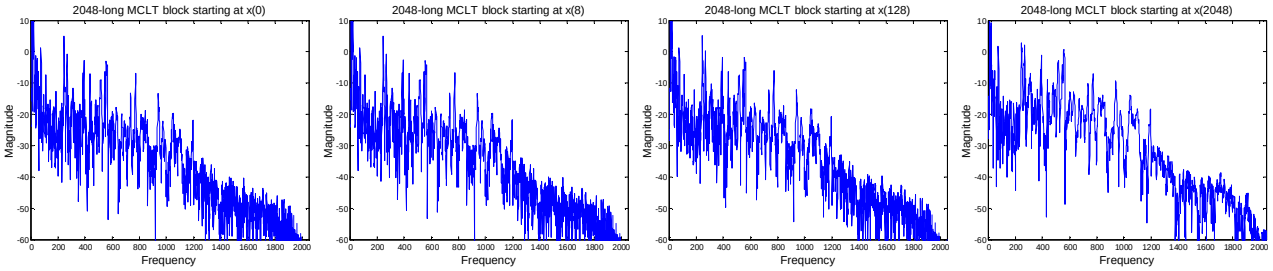

Figure 1: MCLT blocks of a signal x created starting from coefficient x_0, x_8, x_{128}, and x_{2048}. MCLT block length is 2048 frequency coefficients. As block contents change little if the shift is smaller than 512 samples (at sampling rate of 44kHz), we adopt $\eta = 0.25$.

$Var[C(z,w)] = \sigma_z\sqrt{m}$. WM is detected if $C(z,w)$ is greater than certain detection threshold δ_T. In order to have symmetric probability of a false alarm and misdetection, δ_T is commonly set to $m/2$.

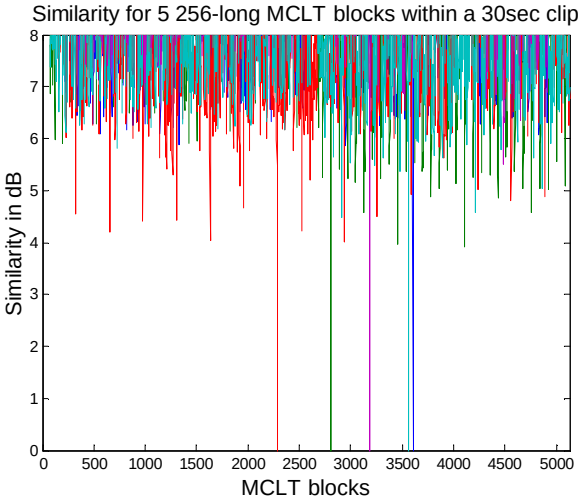

Figure 2: Similarity computation for a 30 second audio clip and 5 random 256-long MCLT blocks. Even in this small test clip by Pink Floyd, for 2 out of 5 blocks at least one substitution block within 4dB relative noise was found.

From the requirement for two blocks to be eligible for substitution in Eqn.2:

$$m\alpha^2 \leqslant \sum_{i=1}^{m} [(y_i + v_i) - (x_i + w_i)]^2 \leqslant m\beta^2 \qquad (3)$$

we can compute the expected resulting correlation $E[C(y+v,w)]$ under the assumption that vectors v and w are independent with respect to x and y:[5]

$$\begin{aligned} E[C(y+v,w)] &= E\left[\sum_{i=1}^{m}(y_i+v_i)w_i\right] \leqslant \\ &\leqslant E\left[\frac{1}{2}\sum_{i=1}^{m}(y_i-x_i)^2\right] + m(\delta^2 - \frac{\alpha^2}{2}) \end{aligned} \qquad (4)$$

where δ is WM amplitude $|v| = |w| = \delta$. Assuming that there exists true repetition of the original content ($y = x$), then setting $\alpha \geqslant \delta\sqrt{2}$ would set the expected minimum correlation to zero after substitution.

[5] $E[C(y,v)] = E[C(x,v)] = E[C(y,w)] = E[C(x,w)] = 0$.

2.2 The BPM Against Marked Audio

In this section, we demonstrate how the generic principles behind the BPM attack can be applied against an audio watermarking technology. WM length is assumed to be greater than one second. Since most psycho-acoustic models are operating in the frequency spectrum, we launch the BPM attack in the logarithmic (dB) frequency domain. The set of signal blocks B is created from the coefficients of a modulated complex lapped transform [11]. Considered MCLT analysis blocks can potentially range in size from 256 to 1024 coefficients with an $\eta = 0.25$ overlap. Figure 1 depicts how MCLT block content changes as its 2048-long analysis window shifts for 8, 128, and 2048 samples (content sampled at 44.1kHz). Each block of coefficients is psycho-acoustically masked using [11]. Similarity is explored exclusively in the audible part of the frequency spectrum. Because of psycho-acoustic masking, the similarity function is not commutative. The second operand of the function (substitution) is always masked with the psycho-acoustic mask of the first operand in addition to its own masking.

	Noise margin β [dB]						
Music	3	3.5	4	4.5	5	5.5	6
Techno	6	10	23	46	80	83	98
Jazz	0	4	13	38	77	98	98
Rock	0	0	3	44	87	100	100
Voice	0	1	7	38	83	96	98
Classical	0	1	12	50	81	98	99

Table 1: Similarity results for five typical pop, jazz, and classical 30 second soundtracks. Columns 2-8 present the percentage of 256-long MCLT blocks that are substitutable with other blocks from the same 30 second clip if the substitute is within $\beta = \{3..6\}$dB. Only the 2-7kHz subband was considered for block substitution.

We performed two types of experiments in order to evaluate the effectiveness of the BPM attack. The first set of experiments aimed at quantifying the amount of similarity between blocks of several audio clips. As an example, Figure 2 shows the similarity matrix for five randomly selected 256-long MCLT blocks within a 30 second Pink Floyd song. For the illustrated data, 2 out of 5 MCLT blocks could be substituted with other MCLT blocks that are within $\beta = 4dB$. In the example, block similarity was computed over the entire frequency spectrum. However, WMs are usually embedded within certain relatively narrow subband which may only improve the matching results [7]. Similarly, Table 1 shows that on a sample of 5 typical pop, rock, jazz, and classical music pieces, almost regularly half the MCLT blocks were substitutable with other blocks from the same 30 second clip

within a noise margin of $\beta = 4.5dB$.
The second set of experiments aimed to establish how substitutions affect a traditional SS watermarking technology. Figure 3 shows how correlation of a SS detector drops with the increase of search freedom β. Note that the results presented consider auto-correlations within a 30 second rock music clip. For longer clips and especially substitution libraries, results should be substantially better. Also, note that adding a white noise pattern of considered β dB affects the correlation detector negligibly. Finally, additive noise of 4-5dB in the 2-7kHz is a relatively tolerable modification.

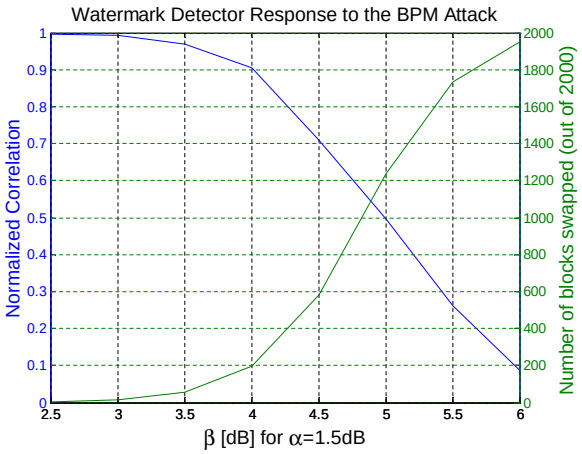

Figure 3: Response of a SS WM detector to the BPM attack. X-axis depicts the increase in β for fixed $\alpha = 1.5dB$ and WM amplitude of $1dB$. The left Y-axis shows the decrease of the normalized correlation as β increases. The right Y-axis shows the number of 256-blocks substituted for a 2000-block WM within a 5200-block audio clip.

For the same experimental set-up, Figure 4 shows the dependency between WM amplitude and correlation decrease due to the BPM attack. Clearly, with the increase of watermark amplitude, the search process of the BPM attack becomes harder for two reasons: (i) block contents become more randomized and (ii) the substituted blocks are more correlated with the original blocks because β is fixed (in this case at $5dB$). Although, stronger WMs may sound like a solution to the BPM attack, extreme values cannot be accepted because of the requirement for high-fidelity marked content and because, in this case, WM estimation becomes an effective anti-WM tool.

3. CONCLUSION

For any watermarking technology and any type of content, an ultimately powerful attack is to re-record the original content value; i.e. play again the music or capture the same original image. In this paper, we simulate this attack: given a library of multimedia content, the BPM attack aims at replacing small pieces of the marked content with perceptually similar but unmarked[6] substitutions from the library. The hope is that the substitutions have little correlation with the embedded mark. Although the attack is generic and can be applied to all marking strategies, we demonstrate how the BPM attack can be launched for audio content and a traditional SS watermarking technology. From the presented experimental results, we conclude that block substitution that creates approximately 4dB noise with respect to the marked content, is sufficient to bring a SS correlation detector to half the expected value without an attack.
At this point, the only prevention against the BPM attack would be to identify rare parts of the content at WM embedding time and mark only these blocks.

[6]Or marked with a different WM.

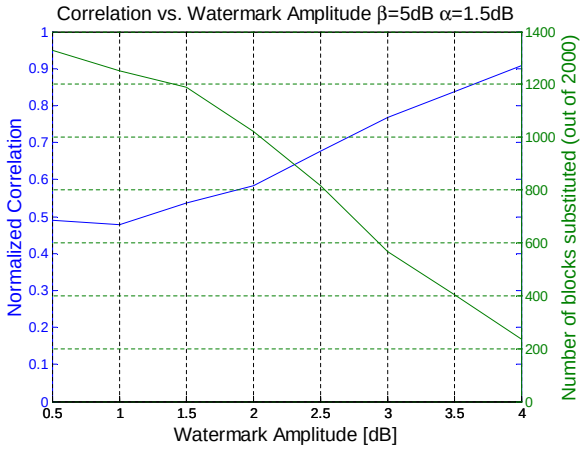

Figure 4: Effect of WM amplitude on the effectiveness of the BPM attack. X-axis depicts the amplitude δ of a SS WM applied to a rock song. The left Y-axis shows the decrease of the normalized correlation as δ increases. The right Y-axis shows the number of 256-blocks substituted for a 2000-block WM within a 5200-block audio clip.

4. REFERENCES

[1] A. Patrizio: DVD Piracy: It Can Be Done, Wired News, November 1, 1999.

[2] Secure Digital Music Initiative. `http://www.sdmi.org`.

[3] Petitcolas, F. A. P., et al.: Attacks on copyright marking systems, Information Hiding Workshop, pp.218-238, Portland, OR, USA, 1998.

[4] Langelaar, G. C., et al.: Removing Spatial Spread Spectrum Watermarks by Non-linear Filtering, European Signal Processing Conference, Rhodes, Greece, pp.2281–2284, 1998.

[5] Voloshynovskiy, S., et al.: Generalized watermarking attack based on watermark estimation and perceptual remodulation, SPIE: Security and Watermarking of Multimedia Content II, San Jose, CA, USA, 2000.

[6] Petitcolas, F. A. P., and Anderson R. J.: Evaluation of copyright marking systems, IEEE Multimedia, Vol.1, pp.574–579, Florence, Italy, 1999.

[7] Kirovski D., Malvar H.: Robust Covert Communication over a Public Audio Channel Using Spread Spectrum, Information Hiding Workshop, Pittsburgh, PA, USA, 2001.

[8] Cox, I.J., et al.: A Secure, Robust Watermark for Multimedia, Information Hiding Workshop, Cambridge, UK, 1996.

[9] Chen, B., Wornell, G.W.: Digital watermarking and Information embedding using dither modulation, Workshop on Multimedia Signal Processing, pp.273-278, Redondo Beach, CA, USA, 1998.

[10] Malvar H.: A modulated complex lapped transform and its application to audio processing, International Conference on Acoustics, Speech, and Signal Processing, pp.1421-1424, Phoenix, AZ, USA, 1999.

[11] Malvar, H.S.: Auditory masking in audio compression, Audio Anecdotes, Kluwer, New York, 2001.