ANALYSIS OF ATTACKS ON SDMI AUDIO WATERMARKS

Min Wu [*] *Scott A. Craver* [*] *Edward W. Felten* [**] *Bede Liu* [*]

[*] Dept. of Electrical Engineering [**] Dept. of Computer Science
Princeton University, Princeton, NJ 08544, U.S.A.

ABSTRACT

This paper explains and analyzes the successful attacks submitted by the authors on four audio watermark proposals during a 3-week SDMI public challenge. Our analysis points out some weaknesses in the watermark techniques currently under SDMI consideration and suggests directions for further improvement. The paper also discusses the framework and strategies for analyzing the robustness and security of watermarking systems as well as the difficulty, uniqueness, and unrealistic expectations of the attack setup.

Keywords: audio watermark, data hiding, SDMI, attack analysis

1. INTRODUCTION

Secure Digital Music Initiative (SDMI) is an international consortium that is developing open technology specifications aiming at protecting the playing, storing, and distributing of digital music [1]. Imperceptible digital watermarking has been proposed to be key elements in the SDMI systems. Digital watermarks are special signals embedded in digital audio and are extractable by detection mechanisms. Upon detection, the watermarks may direct certain actions to be taken, for example, to permit or to deny recording. An SDMI system may incorporate a combination of robust and fragile watermarks. The robust watermarks, indicating specific access policies, should survive common signal processing and attacks. The fragile watermarks may be used to indicate whether the audio has experienced certain processing such as lossy compression [2]. In early September 2000, SDMI announced a three-week public challenge for its Phase-II screening, inviting the public to evaluate the attack resistance for four watermark techniques (A, B, C, F) and two other schemes (D, E). The challenge emphasized on testing the effectiveness of robust watermarks, which is crucial in ensuring the proper functioning of the entire system.

A team of researchers from Princeton University, Rice University, and Xerox research laboratories participated this challenge and made progress on evaluating both watermark and non-watermark proposals [3]. This paper summarizes our successful attacks on the robust part of four audio watermark techniques (A, B, C, F) [1]. Also presented are analysis and implementation issues.

Prior Work on Watermark Attacks An attack on the watermarking system is successful if the original goal of embedding watermarks cannot be achieved. For robust watermarking, this means the detector is unable to detect the existence of watermark or there is ambiguity in making a definite decision. An effective attack does not have to remove the watermark. One simple example is to cause mis-synchronization via jitter [4].

Finding effective attacks and analyzing them play an important role in identifying the weaknesses and limitations of watermarking schemes, as well as in suggesting directions for further improvement. A number of attacks and some countermeasures have been reported in the literature. Most of the previous attacks target at specific types of watermarking schemes, in which analysts have full knowledge of the watermarking algorithms and are able to perform experiments with many non-watermarked, watermarked, and attacked samples, and to observe the results in real time.

SDMI Attack Setup In this challenge, the watermark embedding and detection algorithms are not known to the public. Limited information is available only through the oracle submission. After each submission, detection is performed by SDMI staff and the result is sent back with a response time of about 4-12 hours. For each of the four challenges, SDMI provided three audio samples, as illustrated in Fig. 1. They are:

- samp1?.wav (original audio with no watermark)
- samp2?.wav (samp1?.wav watermarked by Technology-?)
- samp3?.wav (a different audio watermarked by Technology-?)

where the substitution symbol "?" stands for one of the four challenges: "a", "b", "c", or "f". All audio samples are 2-minute long, sampled at 44.1 kHz with 16-bit precision. The audio contents are mostly popular music. Sample-1 for all four technologies are identical, while sample-3 are all different.

A participant of this challenge generates an attacked au-

This research is supported in part by a New Jersey State R&D Excellence Award, NSF grant MIP-9408462, and Intel Technology for Education 2000 Grant. The authors can be contacted via {minwu, sacraver}@ee.princeton.edu, felten@cs.princeton.edu, liu@ee.princeton.edu .

[1] The success was confirmed by SDMI during the 3-week challenge.

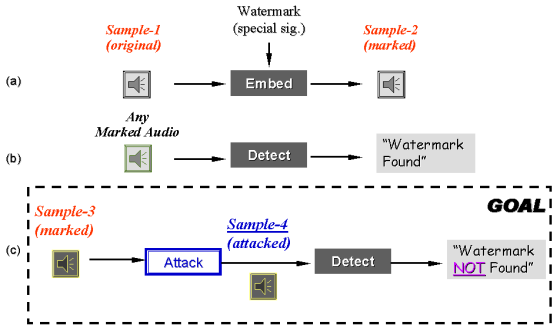

Fig. 1. Illustration of SDMI attack problem.

dio file *sample-4* from *sample-3*, then uploads it to SDMI's oracle for testing. According to SDMI's emails, a "possibly successful" attack must render the detector unable to find the watermark, while retaining the auditory quality comparable to the original one (*sample-3*). The detection response is binary, i.e., either "possibly successful" or "unsuccessful". In the unsuccessful case, there is no indication whether the detector can still find watermark or the detector can no longer find watermark but the auditory quality is considered unsatisfactory. For convenience, we shall denote the four pieces of audio as S_1, S_2, S_3, and S_4.

Comments on Attack Setup The SDMI public challenge presents an emulated rivalry environment, providing attackers with a limited amount of information and restricted access to watermark detectors in a very short time frame. The task is more difficult than the one in real world in the following aspects. First, in real world, a watermark detector encapsulated in a compliant device will be available to an attacker for unlimited uses, and the detector's response time will be instantaneous rather than hours. Second, a user of the real system will be able to distinguish whether or not a detector is able to find watermarks, regardless of the audio quality. These two aspects would enable an attacker polling a detector with different input and obtaining the corresponding output, which in turn provides a large amount of useful information for attacks. Furthermore, the SDMI business model allows a user to pass a piece of unwatermarked music through an SDMI admission device to make it SDMI-compliant thus has watermarks embedded in. This implies that a non-trivial number of original-watermarked audio pairs rather than a single pair are likely to be available to an attacker in real world. As can be seen in the next section, these pairs provide valuable information regarding how watermarks are embedded and the information can be exploited in attacks.

2. PROPOSED ATTACKS AND ANALYSIS

In this section, we first explain a general framework for tackling the attack problem. We then take two different successful attacks on Watermark **C** as examples to demonstrate our attack strategies, to describe the specific implementation, and to present analysis in detail. For completeness, the attacks for the other three watermark techniques **A**, **B**, and **F** are also briefly explained.

2.1. General Approaches to Attacks

An attacker may take one of three general approaches to tackle the problem: **(Type-1)** exploiting the design weakness via blind attack, **(Type-2)** exploring the embedding mechanism from $\{S_1, S_2\}$, the known original-watermarked pairs, or from the watermarked signal $\{S_3\}$ alone, **(Type-3)** a combination of the two.

Type-1 attacks are said to be *blind* in the sense that they do not rely on any understanding of embedding mechanism or the special properties held by watermarked signals. This approach includes commonly used robustness tests, such as compression, jittering, warping, pitch change, resampling at different rate, D/A-A/D conversion, and noise addition [5]. The counter-attack strategy for such blind attacks is to find as many weaknesses as possible and to correct them. A good design, therefore, should at least have covered most of the typical robustness tests and their combinations. One of our attacks for Watermark-C and our attack for Watermark-F are blind attacks.

Type-2 attacks are designed using the knowledge about the embedding mechanism. Such knowledge, even if not available at the start, can be obtained by studying the input-output response of the embedding system. For example, if we find the difference between S_1 and S_2 is a small signal around certain frequency, we may design an attack to distort S_3 over the corresponding frequency range. Quite a few of our attacks belong to this category. This type of attack is analogous to the *plaintext* and *ciphertext attacks* in cryptanalysis [2] [6]. The difference is that signal processing techniques are used here, including the time-domain and frequency-domain differences, the frequency response, the auto- and cross-correlation, and the cepstrum analysis. We also note that the original and watermarked signals are less likely to be available simultaneously to the public in some data hiding applications, e.g., watermarked-based authentication and DVD video watermark for copy control. Hence, Type-2 attacks may not be a major concern in those cases. But in SDMI applications where an unwatermarked music may be "admitted" into SDMI domain by embedding a watermark, any successful watermarking design has to take Type-2 attacks into consideration. One possible countermeasure is to intentionally wipe off the otherwise distinct "signature" of a particular embedding observable from the original-watermarked pairs. This process may reduce the

[2] *Plaintext attack* refers to deducing the encryption key or decrypting new ciphertexts encrypted with the same key, based on the ciphertext of several messages and their corresponding plaintext; *ciphertext attack* is based on ciphertext only.

robustness against blind attacks if the obscuring distorts the embedded watermarks, showing a tradeoff among robustness against various attacks.

Because it is not always possible to find clear clues about embedding from a limited number of original-watermarked pairs, especially when the obscuring is applied, attacks can be designed by combining the above two.

2.2. Attacks on Watermark-C

We used two attacks on Watermark-C. *Attack-C1* explores the weakness of Watermark-C under pitch change. *Attack-C2* is based on observing the difference between the original and watermarked signal $\{S_1, S_2\}$. Both attacks were confirmed as successful by SDMI oracle.

Observations from Samples of Watermark-C By taking the difference of samp1c.wav and samp2c.wav, bursts of narrow-band signal are observed, as shown in Fig. 2. These bursts appear to be around 1350 Hz.

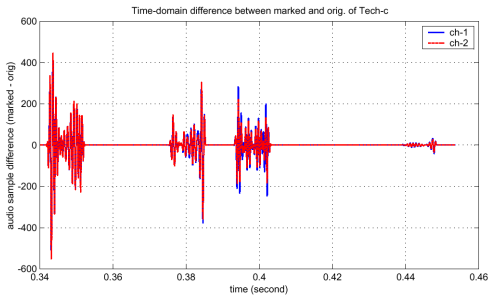

Fig. 2. Challenge-C: waveform of the difference between sample-1c and sample-2c.

Attack-C1 Attack-C1 accelerates audio samples by a small amount, which in turn changes the pitch. Blind attacks of 3% and 10% pitch increase have been applied to all four watermark proposals, and the SDMI detectors indicated that they are effective to Watermark C.

One implementation is to upsample the audio by M times followed by lowpass filtering and downsampling by N times, giving an overall resampling rate of M/N. The original sampling frequency of F_s is changed to $M/N \cdot F_s$. The resampled audio is then played or stored with the same sampling rate F_s as before. The entire process changes the pitch by a fraction of $(N-M)/M$. A precise spectrum interpretation can be obtained based on multi-rate signal processing theory. Attack-C1 can also be implemented using commercial audio editing softwares. For example, the *Effects → Pitch* menu of *GoldWave v4.19* [7] were used as an alternative way to perform pitch shift attacks.

The ability to detect pitch change varies from individual to individual and depends on whether a reference is available. While most people can discriminate pitch difference as low as 0.6% [8], it is nevertheless difficult for a person to identify small pitch changes if he/she has never heard the original before. The standard pitch itself also changed significantly in music history [9]. Our attack with 3% pitch increase (about a quarter tone) has passed SDMI's quality testing performed by "golden ears" after the challenge.

As described previously, we observed that the embedding mechanism adds a narrow band signal to the audio at around 1350Hz. Pitch change can be an effective attack because it stretches or squeezes the spectrum, causing misalignment, which in turn reduces the detector response from the popular matched-filter-type detection. One way to enhance the robustness against Attack-C1 is to estimate and undo the stretching, which is likely to be computationally expensive. Another way is to embed and/or detect watermark in a domain that is resilient to stretching/squeezing.

Attack-C2 Our second attack belongs to Type-2, attempting to jam the frequency band around 1350Hz where it was observed that a narrow-band signal had been added by the embedding mechanism. This narrow-band watermark signal has some randomness, making jamming difficult. Our successful attack is to apply notch filtering to the audio signal at selected frequencies. The filtering introduces significant changes in magnitude and phase around the notch [10], effectively damaging the embedded watermark. Specifically, we used the *Effects → Filters → Bandpass/stop* menu of the audio editor *GoldWave* to perform notch filtering, with a stop band of 1250-1450Hz and steepness of 5 (10th order).

Attack-C2 has passed SDMI's 2nd round quality testing performed by "golden ears". For signals with sufficiently rich spectrum, the magnitude and phase changes caused by notch filtering may not be detectable by a person because of frequency masking and other human auditory phenomena. In the next section, we will see that the embedding process of Watermark-B has a step of notch filtering, suggesting that Watermark-B is a potential attack on Watermark-C. It also suggests that the distortion on audio signal imposed by our Attack-C2 is comparable with that by the embedding process of Watermark-B.

2.3. Attacks on Watermark A, B & F

Watermark A Our attack on Watermark-A, referred as *copier attack*, is a Type-2 attack. By analyzing the short-time FFT of the samples, we observed regular patterns of phase difference, some of which is shown in Fig. 3. The observation leads to a time varying model describing the phase difference between sample-1a and sample-2a. Based on the model, our attack "copies" the phase change between sample-1a and sample-2a to sample-3a, aiming at recovering the phase modification done by embedding process. We also introduced some randomness in middle frequency bands during phase manipulation. A variation of this attack incorporating magnitude manipulation was also submitted.

Both were confirmed by SDMI oracle as successful.

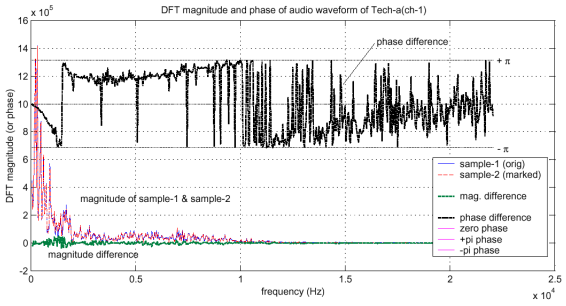

Fig. 3. Technology-A: FFT magnitude of original and watermarked signals, and phase difference between the two signals for a 1000-sample segment.

Watermark B Our attack on Watermark-B is also a Type-2 attack. A spectrum notch is observed around 2800Hz for some parts of the audio and around 3500Hz for some other parts (Fig. 4) [3]. The phase difference between original and watermarked audio signals also exhibit unique butterfly shape, indicating that notch filtering is involved in embedding. Our attack fills in those notches with random but bounded coefficient values. We also submitted a variation of this attack involving different parameters for notch description. Both were confirmed by SDMI oracle as successful.

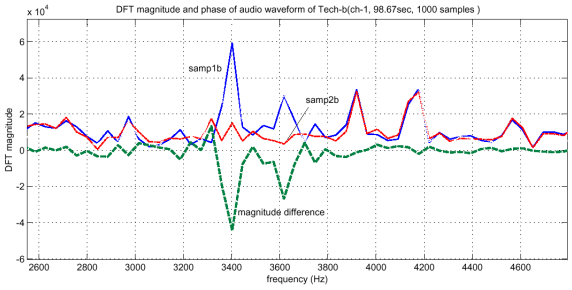

Fig. 4. Technology-B: FFT magnitudes of sample-1b and sample-2b and their difference for 1000 samples at 98.67 sec.

Watermark F Our attack on Watermark-F explores the weakness of this watermarking approach under time varying warping in time domain, thus is a Type-1 attack. In particular, we warped the time axis by inserting a periodically varying delay. The delay function comes from our study on Watermark-A, therefore the perceptual quality of our attacked audio is expected to be better than or comparable to that of the audio watermarked by Technology-A. We also submitted variations of this attack involving different warping parameters and different delay function. They were confirmed by SDMI oracle as successful.

[3] Similar techniques were found in US Patent $4,876,617$ "Signal Identification" (1989) after the challenge.

3. CONCLUSION

In this paper, we presented a general framework for analyzing the robustness and security of audio watermark systems. The framework was demonstrated by our successful attacks in the SDMI public challenge. We pointed out that (1) weaknesses in the watermarking design are very likely to be explored by an adversary as effective attacks, prompting the need of thorough testing by watermark designers; (2) a large amount of information regarding the embedding mechanism, derived from pairs of original and watermarked signals, can be used to build powerful attacks, prompting the need of obscuring distinct traces between original and watermarked signals. The second point, though not having received much attention in the literature, is important for SDMI applications.

Due to various limitations of the challenge including the very short time frame, we adopted practical strategies to increase our chance in finding successful attack(s) and in understanding all four watermark challenges. We focused on finding attacks that render mis-detection by a watermark detector without significantly degrading perceptual quality. These are crucial start points from which many optimizations, improvement, and fine-tuning can be made.

Acknowledgement The authors would like to express gratitude to Prof. P. Cook of Princeton Univ. for the discussion on perceptual aspects of digital music, and to thank their colleagues for the support in the SDMI challenge project: P. McGregor (Princeton Univ.), A. Stubblefield, B. Swartzlander, Prof. D.S. Wallach (Rice Univ.), and Dr. D. Dean (Xerox Research).

4. REFERENCES

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

[2] Secure Digital Music Initiative (SDMI): "SDMI Portable Device Specification", Part 1, ver 1.0, 1999.

[3] S.A. Craver, P. McGregor, M. Wu, B. Liu, A. Stubblefield, B. Swartzlander, D.S. Wallach, D. Dean, E.W. Felten: *Technical Report of SDMI Challenge Project*, 2000.

[4] F. Petitcolas, R. Anderson, M. Kuhn, "Attacks on Copyright Marking Systems", *2nd Workshop on Info. Hiding*, 1998

[5] Test result of International Evaluation Project for Digital Watermark Technology for Music: http://www.nri.co.jp/english/news/ 2000/001006.html , 2000.

[6] B. Schneier: *Applied Cryptography: protocol, algorithms, and source code in C*, 2nd Ed., John Wiley & Sons, 1996.

[7] GoldWave software: http://www.goldwave.com .

[8] Doug Coulter: *Digital Audio Processing*, R&D Books, 2000.

[9] Association of Blind Piano Tuners: "History of Pitch", http://www. uk-piano.org/ history/pitch.html, 2000.

[10] S.J. Orfanidis: *Introduction to Signal Processing*, Prentice Hall, 1996