Spread-Spectrum Watermarking of Audio Signals

Darko Kirovski and Henrique S. Malvar, *Fellow, IEEE*

***Abstract*—Watermarking has become a technology of choice for a broad range of multimedia copyright protection applications. Watermarks have also been used to embed format-independent metadata in audio/video signals in a way that is robust to common editing. In this paper, we present several novel mechanisms for effective encoding and detection of direct-sequence spread-spectrum watermarks in audio signals. The developed techniques aim at *i*) improving detection convergence and robustness, *ii*) improving watermark imperceptiveness, *iii*) preventing desynchronization attacks, *iv*) alleviating estimation/removal attacks, and finally, *v*) establishing covert communication over a public audio channel. We explore the security implications of the developed mechanisms and review watermark robustness on a benchmark suite that includes a combination of audio processing primitives including: time- and frequency-scaling with wow-and-flutter, additive and multiplicative noise, resampling, requantization, noise reduction, and filtering.**

***Index Terms*—Audio signals, covert communication, desynchronization, estimation attacks, spread-spectrum, watermarking.**

I. Introduction

With the growth of the Internet, unauthorized copying and distribution of digital media has never been easier. As a result, the music industry claims a multibillion dollar annual revenue loss due to piracy [1], which is likely to increase due to peer-to-peer file sharing Web communities. One source of hope for copyrighted content distribution on the Internet lies in technological advances that would provide ways of enforcing copyright in client-server scenarios. Traditional data protection methods such as scrambling or encryption cannot be used since the content must be played back in the original form, at which point, it can always be rerecorded and then freely distributed. A promising solution to this problem is marking the media signal with a secret, robust, and imperceptible watermark (WM). The media player at the client side can detect this mark and consequently enforce a corresponding e-commerce policy.

Recent introduction of a content screening system that uses asymmetric direct sequence spread-spectrum (SS) WMs has significantly increased the value of WMs because a single compromised detector (client player) in that system does not affect the security of the content [2]. In order to compromise the security of such a system without any traces, an adversary needs to break in the excess of 100 000 players for a two-hour high-definition video.

Manuscript received February 4, 2002; revised December 10, 2002. The associate editor coordinating the review of this paper and approving it for publication was Dr. Ahmed Tewfik.

The authors are with the Microsoft Research, Redmond, WA 98052 USA (e-mail: darkok@microsoft.com; malvar@microsoft.com).

Digital Object Identifier 10.1109/TSP.2003.809384

A. Watermarking Technologies

Audio watermarking schemes rely on the imperfections of the human auditory system (HAS) [3]. Numerous data hiding techniques explore the fact that the HAS is insensitive to small amplitude changes, either in the time [4] or frequency [5]–[7] domains, as well as insertion of low-amplitude time-domain echoes [8]. Information modulation is usually carried out using: SS [9] or quantization index modulation (QIM) [10]. The main advantage of both SS and QIM is that WM detection does not require the original recording and that it is difficult to extract the hidden data using optimal statistical analysis under certain conditions [11].

However, it is important to review the disadvantages that both technologies exhibit. First, the marked signal and the WM have to be perfectly synchronized at WM detection. Next, to achieve a sufficiently small error probability, WM length may need to be quite large, increasing detection complexity and delay. Finally, the most significant deficiency of both schemes is that by breaking a single player (debugging, reverse engineering, or the sensitivity attack [12]), one can extract the secret information (the SS sequence or the hidden quantizers in QIM) and recreate the original (in the case of SS) or create a new copy that induces the QIM detector to identify the attacked content as unmarked. While an effective mechanism for enabling asymmetric SS watermarking has been developed [2], an equivalent system for QIM does not exist to date.

B. Techniques for SS Watermarking of Audio

In this paper, we restrict our attention to direct-sequence SS WMs and develop a set of technologies to improve the effectiveness of their embedding and detecting in audio. WM robustness is enabled using *i*) block repetition coding for prevention against de-synchronization attacks [13] and *ii*) psycho-acoustic frequency masking (PAFM). We show that PAFM creates an imbalance in the number of positive and negative WM chips in the part of the SS sequence that is used for WM correlation detection and that corresponds to the audible part of the frequency spectrum. To compensate for this anomaly, we propose a *iii*) modified covariance test. In addition, to improve reliability of WM detection, we propose two techniques for reducing the variance of the correlation test: *iv*) cepstrum filtering and *v*) *chess* WMs. Since we embed SS WMs in the frequency domain, the energy of a WM is distributed throughout the entire synthesis block, making SS WMs audible in blocks that contain quiet periods. We solve this problem using *vi*) a procedure that identifies blocks where SS WM may be audible to decide whether to use a particular block in the WM embedding/detection process. Finally, we propose *vii*) a technique that enables reliable covert communication over a public audio channel.

In order to investigate the security of SS WMs, we explore the robustness of such a technology with respect to watermark estimation attacks [2]. To launch that attack, an adversary is assumed to know all the details of the WM codec, except the hidden secret. We present a modification to the traditional SS WM detector that *viii*) undoes the attack and, hence, forces the adversary to add an amount of noise proportional in amplitude to the recorded signal in order to successfully remove an SS WM.

We have incorporated these techniques *i)-viii*) into a system capable of reliably detecting a WM in an audio clip that has been modified using a composition of attacks that degrade the original audio characteristics beyond the limit of acceptable quality. Such attacks include fluctuating scaling in the time and frequency domain, compression, addition and multiplication of noise, resampling, requantization, normalization, filtering, and random cutting and pasting of signal samples.

In Section II, we review the basic aspects of SS watermarking, and in Section III, we describe the specifics for audio WM. We consider the overal security aspects in Section IV and present final remarks in Section V.

II. Basics of Spread-Spectrum Watermarking

The media signal to be watermarked $x \in \mathbb{R}^N$ can be modeled as a random vector, where the elements x_i are independent identically distributed (i.i.d.) Gaussian random variables, with standard deviation σ_x, i.e., $x_i \sim \mathcal{N}(0, \sigma_x)$. [1] Because x actually represents a collection of blocks of samples from an appropriate invertible transformation on the original audio signal [5], [7], [9], such modeling is arguable and is further discussed in Section V. A *watermark* is defined as a direct SS sequence w, which is a vector pseudo-randomly generated in $w \in \{\pm 1\}^N$. Each element w_i is usually called a "chip." WM chips are generated such that they are mutually independent with respect to the original recording x. The marked signal y is created by $y = x + \delta w$, where δ is the WM amplitude. The signal variance σ_x^2 directly impacts the security of the scheme: the higher the variance, the more securely information can be hidden in the signal. Similarly, higher δ yields more reliable detection, less security, and potential WM audibility.

Let $p \cdot q$ denote the normalized inner product of vectors p and q, i.e., $p \cdot q \equiv N^{-1} \sum_i p_i q_i$ with $p^2 \equiv p \cdot p$. For example, for w as defined above, we have $w^2 = 1$. A WM w is detected by correlating (or matched filtering) a given signal vector z with w:

$$C(z, w) = z \cdot w = E[z \cdot w] + \mathcal{N}\left(0, \frac{\sigma_x}{\sqrt{N}}\right). \quad (1)$$

Under no malicious attacks or other signal modifications, if the signal z has been marked, then $E[z \cdot w] = \delta$, else $E[z \cdot w] = 0$. The detector decides that a WM is present if $C(z, w) > \tau$, where τ is a detection threshold that controls the tradeoff between the probabilities of false positive and false negative decisions. We recall from modulation and detection theory that under the condition that x and w are i.i.d. signals, such a detector is optimal [14]. The probability P_{FA} of a false positive detection (false alarm) is

$$P_{\mathrm{FA}} = \Pr[C(z, w) \geq \tau | (z = x)] = \frac{1}{2}\mathrm{erfc}\left(\frac{\tau\sqrt{N}}{\sigma_x\sqrt{2}}\right) \quad (2)$$

and the probability P_{MD} of a false negative detection (misdetection) is

$$P_{\mathrm{MD}} = \Pr[C(z, w) \leq \tau | (z = x + w)] = \frac{1}{2}\mathrm{erfc}\left(\frac{(E[z \cdot w] - \tau)\sqrt{N}}{\sigma_x\sqrt{2}}\right). \quad (3)$$

Straightforward application of the principles above provides neither reliability nor robustness. In the following subsections, we outline the deficiencies of the basic SS WM paradigm and provide solutions for improved WM robustness, detection reliability, and resilience to certain powerful attacks.

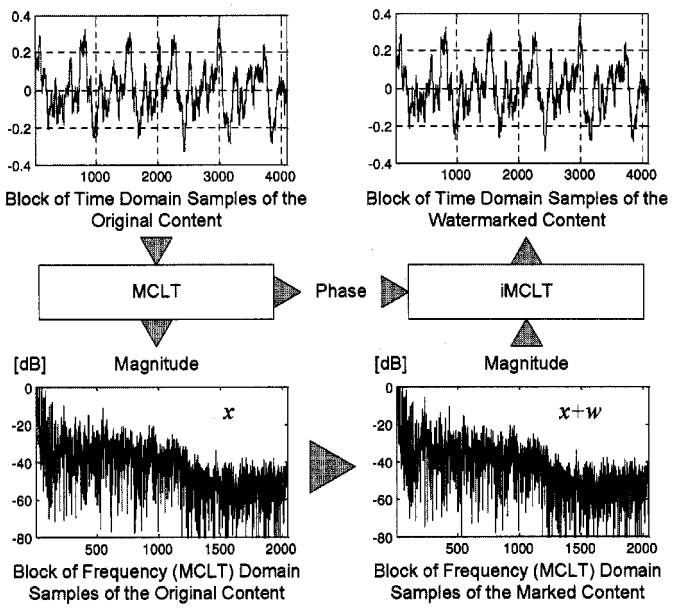

Fig. 1. Process of WM embedding: conversion of a block of time-domain samples into the MCLT domain, SS WM addition, and conversion back to the time-domain.

III. Hiding Spread-Spectrum Sequences in Audio Signals

In our watermarking system, the vector x is composed of magnitudes of several frames of a modulated complex lapped transform (MCLT) [15] in a decibel (dB) scale. The MCLT is a $2\times$-oversampled filterbank that provides perfect reconstruction. The MCLT is similar to a DFT filterbank, but it has properties that makes it attractive for audio processing, especially when integrating with compression systems, because signals can easily be reconstructed from just the real part of the MCLT [15]. After addition of the WM, we generate the time-domain marked audio signal by combining the vector $y = x + \delta w$ with the original phase of x and passing these modified frames to the inverse MCLT. Fig. 1 illustrates this process on an example time-domain frame. Typically, WM amplitude δ is set to a fixed value in the range 0.5–2.5 dB. For example, for $\delta = 1.5$ dB,

[1] $\mathcal{N}(a, b)$ denotes a Gaussian with mean a and variance b^2.

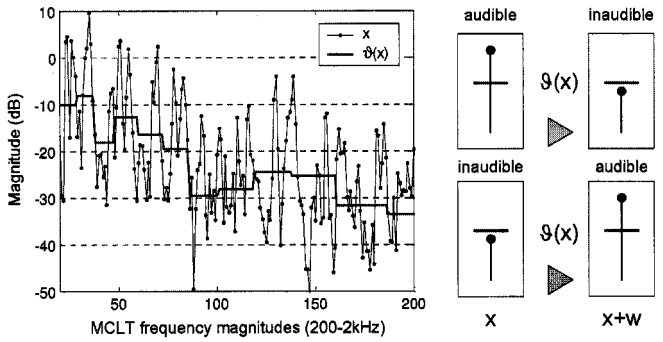

Fig. 2. PAFM: (Left) Example MCLT frequency block with an identified masking function and (right) an example of how WM addition increases the number of positive chips that correspond to the audible part of the MCLT block.

trained ears cannot statistically pass a distinction test between watermarked and original content for a benchmark suite consisting of pop, rock, jazz, classical, instrument solo, and vocal musical pieces. For the typical 44.1 kHz sampling, we use a length-2048 MCLT. Only the coefficients within 200 Hz–2 kHz are marked, and only the audible magnitudes in the same sub-band are considered during detection. Sub-band selection aims at minimizing carrier noise effects as well as sensitivity to downsampling and compression.

A. Psycho-Acoustic Frequency Masking: Consequences and Remedies

The WM detector should correlate only the audible frequency magnitudes with the WM [7] because the inaudible portions of the frequency spectrum are significantly more susceptible to attack noise. That reduces the effective watermark length because the inaudible portion often dominates the frequency spectrum of an audio signal [6].

In order to quantify the audibility of a particular frequency component, we use a simple PAFM model [16]. For each MCLT magnitude coefficient, the likelihood that it is audible averages 0.6 in the crucial 200 Hz–2 kHz subband in our audio benchmark suite. Fig. 2 illustrates the frequency spectrum of an MCLT block as well as the PAFM boundary. PAFM filtering introduces the problem of SS sequence imbalance: a problem also illustrated in Fig. 2. When embedding a positive chip ($w_i = +1$), an inaudible frequency magnitude x_i becomes audible if $x_i > \vartheta(x_i) - \delta$, where $\vartheta(\cdot)$ returns the level of audibility for the argument magnitude for a given MCLT block. Similarly, when embedding a negative chip ($w_i = -1$), an audible magnitude becomes inaudible if $x_i < \vartheta(x_i) + \delta$. We define R_A, R_+, and R_- as the ratios of frequency magnitudes that fall within the corresponding ranges

$$\begin{aligned} R_A &= \frac{|X_A|}{N} \leftarrow (\forall x_i \in X_A)\, x_i \in \Big[\vartheta(x_i) + \delta, +\infty\Big\} \\ R_+ &= \frac{|X_+|}{N} \leftarrow (\forall x_i \in X_+)\, x_i \in \Big[\vartheta(x_i), \vartheta(x_i) + \delta\Big\} \\ R_- &= \frac{|X_-|}{N} \leftarrow (\forall x_i \in X_-)\, x_i \in \Big[\vartheta(x_i) - \delta, \vartheta(x_i)\Big\}. \end{aligned} \quad (4)$$

The expectation for the relative difference ξ in the number of positive and negative chips in the correlated audible part of the SS sequence equals

$$\xi = E\left[\frac{\sum_{i=1}^{N} \tilde{w}_i}{\sum_{i=1}^{N} w_i^2}\right] = \frac{R_+ + R_-}{2R_A + R_+ + R_-} \quad (5)$$

where $\tilde{w}_i = w_i$ if corresponding $x_i + w_i$ is audible and $\tilde{w}_i = 0$ if $x_i + w_i$ is inaudible.

Asymmetric distribution of positive and negative chips in the masked SS sequence can drastically influence the convergence of the correlation test in (1). The convergence is affected because the expected value of the correlation test $E[y \cdot w]$ has an additional component proportional to ξ. For our benchmark suite, ξ averaged 0.057 at $\delta = 1$ dB, with peak values reaching $\xi \sim 0.3$ for recordings with low harmonic content. Thus, whenever PAFM is used, the normalized correlation test (1) must be replaced with a covariance test that compensates for using a nonzero-mean SS sequence. Assuming μ_1, σ_1 and μ_0, σ_0 are the mean and variance of the audible portion of x selected by positive and negative SS chips, respectively, and signal y is watermarked, the correlation test in (1) can be rewritten as

$$\begin{aligned} C(y,w) &= y \cdot w \\ &= \frac{1}{N}\left[\sum_{i=1}^{N(1+\xi)/2} y_i|_{w_i=+1} - \sum_{i=1}^{N(1-\xi)/2} y_i|_{w_i=-1}\right] \\ &= \delta + \frac{1}{2}\mathcal{N}(\mu_r, \sigma_r) \end{aligned} \quad (6)$$

where the noise component $\mathcal{N}(\mu_r, \sigma_r)$ of the detection test has a mean $\mu_r = \mu_1 - \mu_0 + \xi(\mu_1 + \mu_0)$ and variance $\sigma_r^2 = N(\sigma_1^2(1+\xi) + \sigma_0^2(1-\xi))$. The mean value μ_x of the part of the original signal x that corresponds to the audible part of y can be expressed as $2\mu_x = \mu_1 + \mu_0 + \xi(\mu_1 - \mu_0)$, whereas the mean value μ_y of the audible part of y equals $\mu_y = \mu_x + \varphi$, where $\varphi = 2\xi\delta$ if signal y is watermarked and $\varphi = 0$ in the alternate case. Thus, by using a traditional covariance test

$$V(y,w) = C(y,w) - E(y)E(w) = C(y,w) - \mu_y \xi\delta \quad (7)$$

the detector would induce a mean absolute error of $|\mu_r - \mu_y\xi\delta|$ to the covariance test because of the mutual dependency of x and w. Consider the following test:

$$\begin{aligned} C(y,w) = y \cdot w = &\frac{1}{N+\xi} \sum_{i=1}^{N(1+\xi)/2} y_i|_{w_i=+1} \\ &- \frac{1}{N-\xi} \sum_{j=1}^{N(1-\xi)/2} y_j|_{w_j=-1} \end{aligned} \quad (8)$$

which results in a noise component $\mathcal{N}(\mu_r, \sigma_r)$ for this test equal to $\mu_r = \mu_1 + \mu_0$ and $\sigma_r^2 = N(\sigma_1^2/(1+\xi) + \sigma_0^2/(1-\xi))$.

Computation of $\mu_r = \mu_1 + \mu_0$ from y can be made relatively accurate as follows. First, μ_1 and μ_0 are computed as means of the audible part of the signal y selected by positive and negative chips respectively. Then, if $\mu_1 - \mu_0 > 2\delta - \varepsilon$, we conclude that the signal has been watermarked and compensate the test in (8) for $\mu_1 + \mu_0 - 2\delta$; in the alternate case, we compensate for $\mu_1 + \mu_0$. Parameter ε is a constant equal to $\tau\sigma_r$, which ensures low likelihood of a false alarm or misdetection through selection of τ (2), (3).

An error of 2δ in the covariance test occurs if the original signal is bipartitioned with the SS chips such that $\mu_1 - \mu_0 > 2\delta - \varepsilon$. This case can be detected at WM encoding time. Then, the encoder could signal an audio signal block as *hard-to-mark*, or it could extend the length of the WM. Such cases are exceptionally rare for relatively long SS sequences and typical music content rich in sound events. Note that the exact computation of μ_1 and μ_0 would also resolve the error problem incurred in the original covariance test in (6) through exact computation of μ_r. Thus, the two tests in (6) and (8) are comparable and involve computation of similar complexity. On super-pipelined architectures, we expect the test in (8) to have better performance via loop unfolding, as it does not use branch testing.

B. Preventing the Desynchronization Attack

The correlation metrics from (1)–(3) are reliable only if the majority of detection chips are aligned with those used in marking. Thus, an adversary can attempt to desynchronize the correlation by fluctuating time- or frequency-axis scaling within the loose bounds of acceptable sound quality. To prevent such attacks, we use a multitest methodology that relies on block repetition coding of chips of the WM pattern.

It is important to define the degrees of freedom for time- and frequency-scaling that preserves the relative fidelity of the attacked recording with respect to the original. The HAS is much more tolerable to constant scaling rather than wow-and-flutter (variations in scaling over time). Hence, we adopt the following tolerance levels, which are appropriate in practice: $\gamma_T \leqslant 0.1$ for constant time-scaling and $\gamma_F \leqslant 0.05$ for constant frequency-scaling and scaling variance $\gamma_V \leqslant 0.01$ along both time and frequency.

1) Block Repetition Coding: In the first step, we provide resilience against fluctuations in playtime and pitch bending (wow-and-flutter) of up to a fixed parameter γ_V, which delimits the maximum fluctuation magnitude independently along any of these two dimensions. As common standard values for wow-and-flutter for modern turntables are significantly below 0.01, we adopt this value as our robustness limit.

We represent an SS sequence as a matrix of chips $W = \{w_{ij}\}$, $i = 1..N_F$, and $j = 1..N_T$, where N_F is the number of chips per MCLT block, and N_T is the number of blocks of N_F chips per WM. Within a single MCLT block, each chip w_{ij} is spread over a sub-band of F_i consecutive MCLT coefficients. Chips embedded in a single MCLT block are then replicated along the time axis within consecutive T_j MCLT blocks. An example of how redundancies are generated is illustrated in Fig. 3 (with fixed parameters $F_i = 3, T_j = 3$ for all i and j). Widths

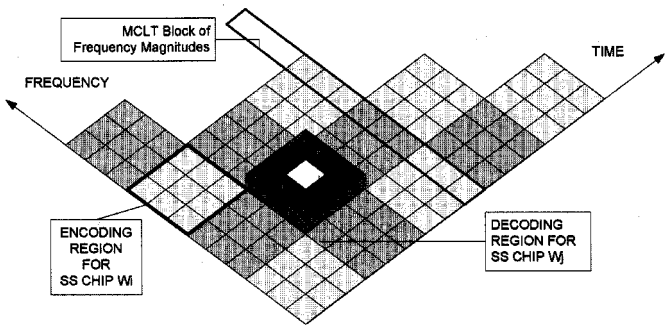

Fig. 3. Example of block repetition coding along the time and frequency domain of an audio clip. Each block is encoded with the same bit, whereas the detector integrates only the center locations of each region.

of the encoding regions F_i, $i = 1..N_F$ are computed using a geometric progression

$$\begin{aligned} F_i &= F_i'' + \eta_F + F' \\ F_i' &\geqslant \left(F_i' + \sum_{j=1}^{i-1} F_j \right) \cdot \gamma_V \\ F_i'' &\geqslant \left(F_i'' + \eta_F + F_i' + \sum_{j=1}^{i-1} F_j \right) \cdot \gamma_V \end{aligned} \tag{9}$$

where η_F is the width of the decoding region (central to the encoding region) along the frequency. Similarly, the length of the WM N_{To} in groups of constant $T_j = T_o$, $j = 1..N_{To}$ MCLT blocks watermarked with the same SS chip block is delimited by $N_{To}T_o\gamma_V < T_o - \eta_T$, where η_T is the width of the decoding region along the time-axis. Lower bound on the replication in the time domain T_o is set to 100 ms for robustness against cropping or insertion.

If a WM length of $N_{To}T_o$ MCLT blocks does not produce satisfactory correlation convergence, additional MCLT blocks ($N_T > N_{To}$) are integrated into the WM. Time-axis replication T_j, $j > N_{To}$ for each group of these blocks is recursively computed using the geometric progression (10). Within a region of F_iT_j samples watermarked with the same chip w_{ij}, only the center $\eta_F\eta_T$ samples are integrated in (1). It is straightforward to prove that such generation of encoding and decoding regions guarantees that regardless of induced wow-and-flutter limited to γ_V, the correlation test is performed in perfect synchronization. Typical redundancy parameters are i) constant replication along time axis 5–10 MCLT blocks and ii) geometrically progressed replication along the frequency axis such that typically 50–120 chips are embedded within the target sub-band 200–2 kHz.

2) Multiple Correlation Tests: The adversary can combine wow-and-flutter with a stronger constant scaling in time and frequency. Constant scaling of up to $\gamma_T < 0.1$ along the time axis and $\gamma_F < 0.05$ along the frequency axis can be performed on an audio clip with good fidelity with respect to the original recording. Resilience to static time- and pitch-scaling is obtained by performing multiple correlation tests as follows:

1) `pointer = 0; progress =` $L(1+\gamma_T)$; (L denotes WM length in MCLT blocks.

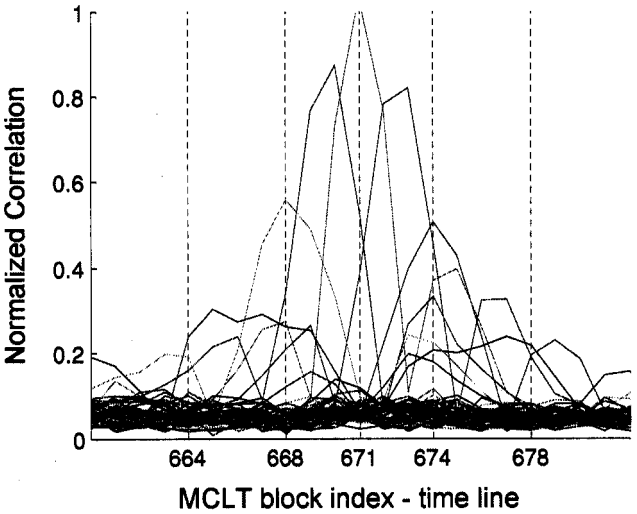

Fig. 4. Example of how a WM is detected during the search process. The correlation test that corresponds to one particular time- and frequency-scaling has synchronized the WM with the MCLT block indexed 671.

2) load buffer with MCLT co-efficients from progress consecutive MCLT blocks starting from the MCLT block indexed with pointer.
3) **for** time.scaling = $-\gamma_T$ to $+\gamma_T$ step $\gamma_V/2$ and **for** frequency.scaling = $-\gamma_F$ to $+\gamma_F$ step $\gamma_V/2$, correlate buffer with WM scaled according to time.scaling and frequency.scaling.
4) **if** (WM found in buffer with time.scaling = T) **then** progress = $L(1+T)$ **else** progress = λ.
5) pointer += progress; **goto** 2).

The search algorithm initially loads a buffer of MCLT coefficients from $L(1+\gamma_T)$ consecutive MCLT blocks. Then, the loaded contents are correlated with different scalings of the searched WM; the scalings are such that they create a grid over $\{-\gamma_T..\gamma_T, -\gamma_F..\gamma_F\}$ with $\gamma_V/2$ minimal distance between points (tests). Due to block redundancy coding, each test $\{T, F\}$ can detect a WM if the actual scaling of the clip is within the $\{T-\gamma_T..T+\gamma_T, F-\gamma_F..F+\gamma_F\}$ region. The test $\{T_m, F_m\}$ yielding the greatest correlation $C(y, w(T_m, F_m))$ is compared with the detection threshold τ to determine WM presence. If WM is found, the entire buffer is reloaded with new MCLT coefficients. Otherwise, the content of the buffer is shifted for λ MCLT blocks, and a new set of tests is performed.

In a typical implementation, for $\gamma_V = 0.02$, in order to cover $\gamma_T = 0.1$ and $\gamma_F = 0.05$, the WM detector computes 105 different correlation tests. The search step along the time axis denoted as λ typically equals between one and four MCLT blocks. An example is shown in Fig. 4. Note that the main incentive for providing such a mechanism to enable synchronization is the fact that, within the length of the WM, the adversary really cannot move away from the selected constant time and frequency scaling more than $\gamma_V/2$; such a change would induce intolerable sound quality. If the attacker is within the assumed attack bounds, the described mechanism enables the detector to conclude whether there is a WM or not in the audio clip based on the SS statistics from (1) and regardless of the presence of the attack.

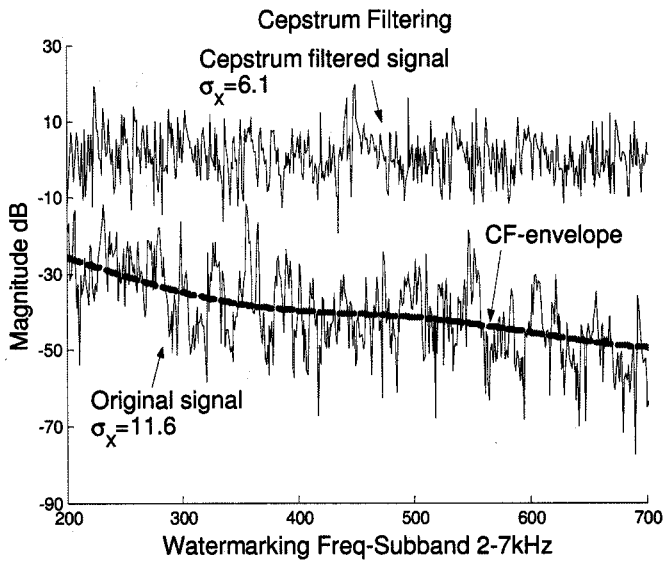

Fig. 5. Demonstration of an original MCLT block and its cepstrum filtering. The dashed line represents the CF-envelope subtracted from the original MCLT block.

C. Cepstrum Filtering

The variance σ_x^2 of the original signal directly affects the carrier noise in (1). Audio clips with large energy fluctuations or with strong harmonics are especially bound to produce large σ_x. Thus, we propose here a nonlinear processing step to reduce the carrier noise. One approach is to subtract a moving average from the frequency spectrum right before correlation: a sort of whitening step. Unfortunately, as bits of the SS sequence are spread over frequency ranges, this technique induces partial removal of the WM chips. We have developed a cepstrum filtering (CF) technique that produces significantly better results than just spectral whitening. With CF, we reduce σ_y in (1) through the following steps:

1) z = DCT$\{y\}$—compute the cepstrum of the dB magnitude MCLT vector y under test via the discrete cosine transform.
2) $z_i = 0$, $i = 1..K$—filter out the first K (typically $5 < K < 20$) cepstrum coefficients.
3) y = IDCT$\{z\}$—reconstruct the frequency spectrum via an inverse DCT. The filtered frequency spectrum replaces y in the correlation detector **(1)**.

The rationale behind CF is that large variations in y can only come from large variations in x since $|w|$ is limited to a small value $\delta \ll \sigma_x$. Thus, by filtering out large variations in y, we can reduce the carrier noise significantly, without affecting much the expected value $E[y \cdot w]$. That is particularly efficient if the WM sequence w has a nonwhite spectrum containing more noise at higher frequencies, as discussed in the next subsection. Fig. 5

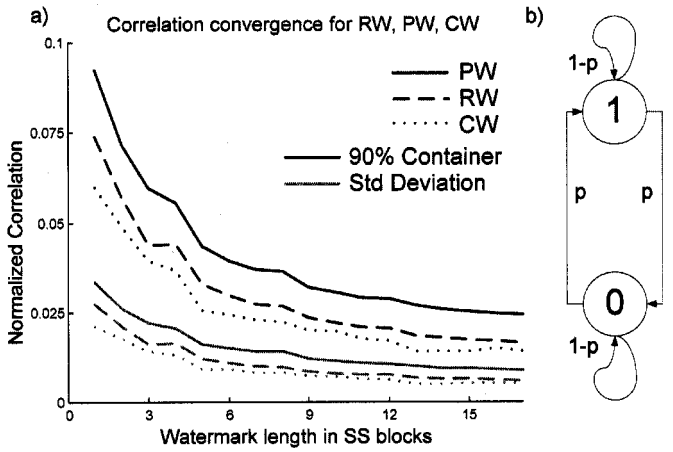

Fig. 6. (a) Convergence of the normalized correlation $C(y, w)$ with WM length for a nonwatermarked signal. Top three plots: 90% percentile limits of $C(y, w)$ (90% of the correlation values are under each curve), for a traditional purely random SS sequence, a perfect WM (PW), and a chess WM (CW). Bottom three plots: Corresponding standard deviations of $C(y, w)$ in the same order. (b) Simple state machine that produces a chess WM ($p > 0.5$).

illustrates the impact of CF on the signal variance, which is typically reduced by a factor of almost four. Thus, in order to attain the performance of CF detector, a non-CF detector must integrate almost four times more magnitude points.

D. Chess Watermarks

Because of the relatively short MCLT frames (30 ms), we assume that the audio signal has a slowly varying magnitude spectrum. Thus, for short WMs, a possible sequence in time of several consecutive positive WM chips can pose false alarms if correlated with large positive x values. In practice, that problem occurs frequently for quiet clips with strong harmonics (e.g., piano or sax solo). To alleviate the problem, it is important to attenuate the DC component of the WM chips along the time direction.

We define a *perfect WM* (PW) as a sequence of alternating positive and negative chips, along both the time and frequency axis. Correlation with PW results in highly improved correlation convergence for a nonwatermarked signal, as illustrated in Fig. 6. To leverage the convergence efficacy of PW with the security of pseudo-random SS sequences, we introduce a *chess-WM* (CW). We define a CW as a stochastic approximation to a PW by using the simple first-order state machine depicted in Fig. 6. Whereas the probability p of switching from the "**0**" state to the "**1**" state for traditional SS sequences is desired to be one-half, we built CWs to enforce frequent toggling of bits along the time axis or, equivalently, to emphasize high frequencies in the WM sequence. We typically select $p = 0.75$. For a sufficiently large N, the randomness reduction in the sequence domain does not pose a security threat, while resulting in correlation convergence similar to PW (typically $N > 200$).

E. Improving the Inaudibility of Spread-Spectrum Watermarks in Audio

SS WMs can be audible when embedded in the MCLT domain, even at low magnitudes (e.g., $\delta < 1$ dB). This can happen in blocks where certain parts (up to 10 ms) are quiet, whereas the remainder of the block is rich in audio energy. Since the SS sequence spreads over the entire MCLT block, it can cause audible noise in the quiet portion of the MCLT block (see Fig. 7).

To alleviate that problem, we detect MCLT blocks with dynamic content where an SS WM may be audible if added. The blocks are identified according to an energy criterium, for example, as descried below. WMs are not embedded nor detected in such blocks. Fortunately, such blocks do not occur often in audio content; in our benchmark set, we identified up to $\zeta < 5\%$ of MCLT blocks per WM as potential hazard for audibility. By not marking these blocks, the corresponding correlation is bound to a lower expected value $E[y \cdot w] = 1 - \zeta$, which causes only a minor effect on detector's decision. The detection of hazardous blocks is performed on each length-K MCLT block using the following algorithm.

1) Compute the interval energy level $E(i) = \sum_{j=1+K(i-1)/(2P)}^{iK/P} y_j$, $i = 1..P$ for each of the P interleaved subintervals of the tested signal y in the time-domain (commonly $K/P \geqslant 32$). Block subintervals are illustrated in **Fig. 7.**
2) **if** $(\min_{j=1}^{P}(E(j)/\sum_{i=1}^{P} E(i)) \leqslant \chi_0)$ **then** WM is audible in the block. Parameter χ_0 is empirically determined.

F. Covert Communication Over Audio Channels

SS provides only means of embedding (hiding) pseudo-random bit sequences into a given signal carrier (audio clip). One trivial way to embed an arbitrary message into a SS sequence is to use a pool of WMs such that each WM represents a symbol from an alphabet used to create the covert message. Depending on the symbol to be sent, the encoder selects one of the WMs from the pool and marks the next consecutive part of audio with this WM. The detector tries all WMs from the pool, and if any of the correlation tests yields a positive test, it concludes that the word that corresponds to the detected WM has been sent. Since a typical WM length in our implementation ranges from 11 to 22 s, to achieve a covert channel capacity of just 1 b/s, the detector is expected to perform between 210 and 221 different WM tests. Besides being computationally expensive, this technique also raises the likelihood of a false alarm or misdetection by several orders of magnitude.

Therefore, it is clear that a covert channel cannot rely solely on WM multiplicity, and thus, some form of WM modulation must be considered. A basic concept for the design of a modulation scheme is the observation that if we multiply all WM chips by -1, the normalized correlation changes sign but not magnitude. Therefore, the correlation test can detect the WM by the magnitude of the correlation and the sign carries one bit of information.

The covert communication channel that we have designed uses two additional ideas. First, to add S message bits, the SS sequence is partitioned along the time-axis into S equal-length subsets w_k, $k = 1..S$, where each w_k consists of all WM chips w_{ij} such that $(k-1)S \leqslant j < kS$. Thus, there are N_T/S

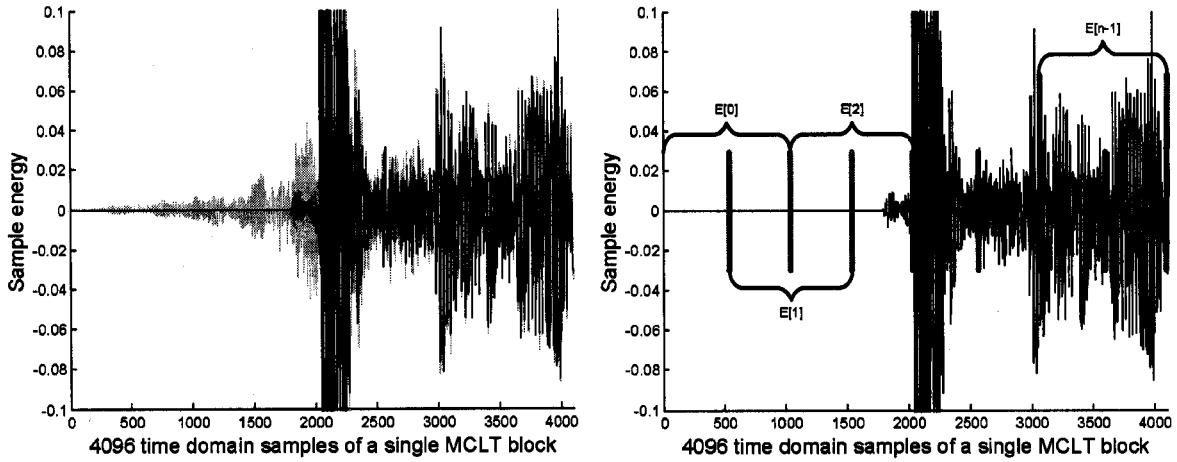

Fig. 7. Example of audibility of a SS WM when embedded in the frequency domain. The black plot denotes a single MCLT block of time domain sample of the original recording, whereas the grey line denotes the corresponding marked recording with audible noise prior to the signal peak.

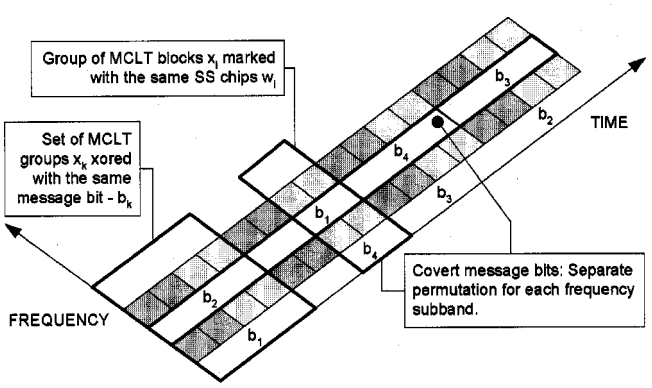

Fig. 8. Embedding a permuted covert communication channel over the temporal and spectral domain.

chip blocks of N_F chips per each w_k. Each bit b_k of a message $B \in \{\pm 1\}^S$ is used to multiply the chips of the corresponding w_k while creating the marked content $y_k = x_k + \delta b_k w_k$, where y_k and x_k are content blocks that correspond to w_k. A typical example is shown in Fig. 8.

At detection time, the squared value of each partial covariance test $C(y_k, w_k)$—computed using (1)—is accumulated to create the final test value as follows:

$$\begin{aligned} C(y,w) =& \frac{1}{S}\sum_{k=1}^{S}\left[C\left(y_k, w_k\right)\right]^2 \\ =& \frac{1}{S}\sum_{k=1}^{S}\left(E\left[y_k \cdot w_k\right] + \mathcal{N}\left(0, \sigma_x\sqrt{\frac{S}{N}}\right)\right)^2 \\ =& E[y\cdot w]^2 + 2E[y\cdot w]\mathcal{N}\left(0, \frac{\sigma_x}{\sqrt{N}}\right) \\ &+ \frac{1}{S}\sum_{k=1}^{S}\mathcal{N}^2\left(0, \sigma_x\sqrt{\frac{S}{N}}\right) \end{aligned} \tag{10}$$

Therefore, $C(y,w)$ in this case has three components: *i*) a mean and *ii*) a zero-mean Gaussian random variable (both of them equal to zero if the content is not marked) and *iii*) a sum of squares of Gaussian random variables. Thus, the likelihood of a false alarm P_{FA} (2) can be computed using the upper tail of the chi-squared pdf with S degrees of freedom:

$$\begin{aligned} P_{\mathrm{FA}} =& \Pr[C(y,w) > \tau] \\ =& \frac{1}{\Gamma\left(\frac{S}{2}\right)\sqrt{2^S}}\int_{\tau}^{\infty}\frac{z^{(S-2)/2}}{e^{z^2/2}}dz = \frac{4\tau N}{S\sigma_x^2} \end{aligned} \tag{11}$$

where $\Gamma(\cdot)$ is the Gamma function. The lower bound on the likelihood of a WM misdetection is computed according to (3) as the third component in (10) can be neglected for marked signals because it is always positive. Bits of the covert message are recovered at detection time as the sign of partial correlations $b_k = \mathrm{sign}(C(y_k, w_k))$. The likelihood of a bit misdetection P_{MDB} once a WM is detected equals

$$\begin{aligned} P_{MDB} =& \Pr\left[C\left(y_k, w_k\right) < \tau\right] \\ =& \Pr\left[\mathcal{N}\left(0, \sigma_x\sqrt{\frac{S}{N}}\right) > \delta - \tau\right] \\ <& \frac{1}{2}\mathrm{erfc}\left(\frac{(\delta-\tau)\sqrt{N}}{\sigma_x\sqrt{2S}}\right). \end{aligned} \tag{12}$$

Finally, in order to improve the robustness of each bit of the encoded covert message, we perform a secret permutation $\pi(i)$ of the message bits for each MCLT subband F_i. Thus, a permuted bit $b_{\pi(i,k)}$ is combined with chip blocks along a certain subband w_{ik}, $k = 1..S$ (each block has N_T/S chips) and then embedded in the original content as $y_{ik} = x_{ik} + \delta w_{ik} b_{\pi(i,k)}$. This procedure aims at *i*) spreading each bit of the encoded covert message throughout the entire WM for security reasons (an attacker cannot focus only on a short part of the clip hoping to remove the message bit) and *ii*) increasing the robustness of the detection algorithm because of spreading localized variances of noise over the entire length of a WM. The process of permuting bits of the message is illustrated in Fig. 8.

G. Summarizing Discussion

We have deployed the techniques described in the previous subsections to create an audio watermarking system with strong robustness with respect to common audio editing procedures. A block diagram that illustrates how the developed technologies

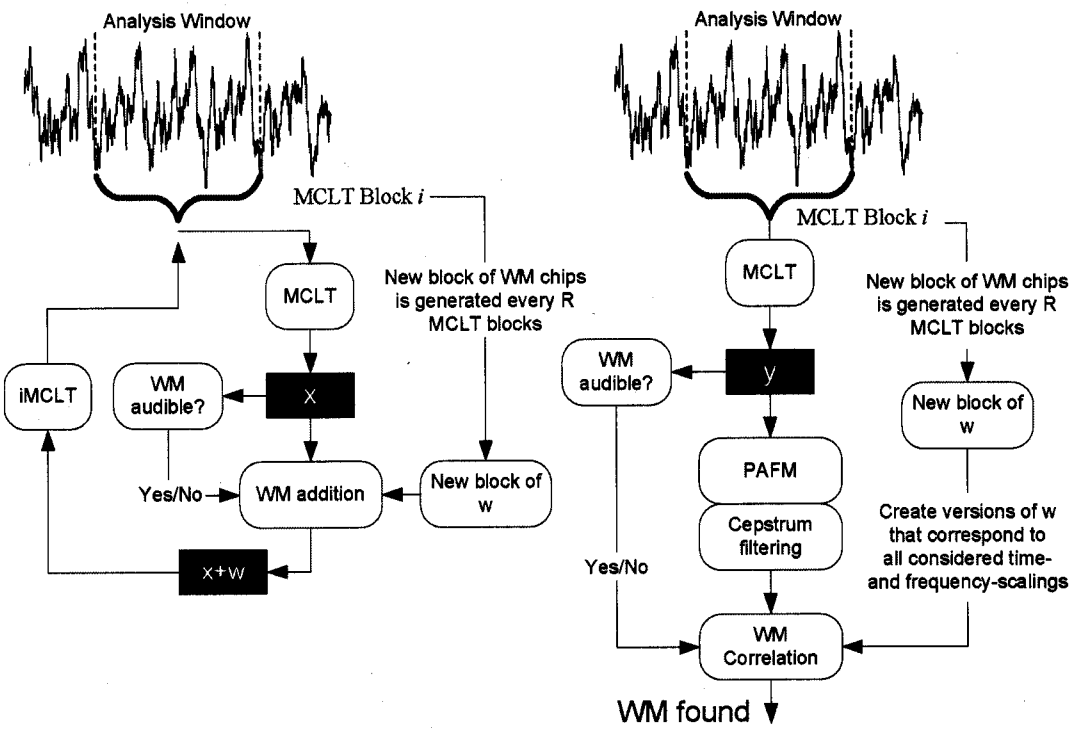

Fig. 9. Block diagram of the WM (left) embedding and (right) detection procedures.

are linked into a cohesive system for audio marking is presented in Fig. 9.

A reference implementation of our data hiding technology on an x86 platform requires 32 Kbytes of memory for code and 100 Kbytes for the data buffer. The data buffer stores averaged MCLT blocks of 12.1 s of audio (for a WM length of 11 s). WMs are searched with $\gamma_V = 0.02$, which requires ~ 40 tests per search point. Real-time WM detection under these circumstances requires about 15 MIPS, which is a small requirement for today's DSP processors. WM encoding is an order of magnitude faster, with smaller memory footprints. The achieved covert channel bit rate varies in the range of 0.5–1 b/s for $S = 4$ and a pool of 16 different WMs.

We have tested our proposed watermarking technology using a composition of common sound editing tools and malicious attacks, including all tests defined by the Secure Digital Music Initiative (SDMI) industry committee [17]. Such tests included double D/A-A/D conversion, noise addition at the -36 dB level, bandpass filtering, MP3 encoding at 64 and 32 kbps, time-scale changing of up to $\pm 4\%$, wow and flutter at 0.5%, and echo insertion of up to 100 ms. We used a data set of 80 15-s audio clips, which included jazz, classical, voice, pop, instrument solos (accordion, piano, guitar, sax, etc.) and rock. In that data set, there were no errors and from measured noise levels in the correlation metric, we estimated the error probability to be well below 10^{-6}. Error probabilities decrease exponentially fast with the increase of WM length; therefore, it is relatively easy to design a system for error probabilities below 10^{-9}, for example. Analysis of the security of embedded WMs is presented in the next section.

Fig. 10 shows the performance improvements, with the modifications described above, on our benchmark set (concatenated into a single sound clip on these diagrams): (a) and (b) versus (c) and (d) demonstrates strong gain in $C(y, w)$ variance due to cepstrum filtering, and (e) and (f) versus (g) and (h) showcases slightly reduced detection reliability due to the permuted covert communication (PCC) channel. Peaks in the correlation test clearly indicate detection and location of each WM. Note that the peak values for both detectors are virtually the same; however, the negative detection for the PCC decoder yields slightly higher variance (in our experiments, we recorded differences up to 5%).

Finally, in order to quantify the robustness of the watermarking technology with respect to a publicly available benchmark, we show the watermark detection results against the attacks in Stirmark Audio [18]. For that experiment, we have selected an audio clip rich in music events (a rhythmic latin jazz clip with trombone, piano, and alto-sax solos), watermarked it, and then detected watermarks in the original, the marked copy, and all 46 clips created by the Stirmark Audio suite of attacks. The detection results are presented in Table I. For watermarked clips, we report the minimal correlation achieved for each of the ten watermarks embedded in the audio clip. For the original clip, we report the maximal correlation value throughout the search for any of the ten watermarks. The corresponding correlation value is marked as $\overline{C(y, w)}$ in Table I. The detection threshold is set to $\tau = 0.25$, which results in an estimated probability of a false positive smaller than 10^{-6} for a variety of audio clips. From Table I, we observe that all but one attack had only minimal effect on the correlation value. The only attack that reduced significantly the correlation value (**copysample**) had a strong impact on the fidelity of the recording so that the attacked clip almost did not resemble the original. The parameters of the Stirmark Audio attack were the

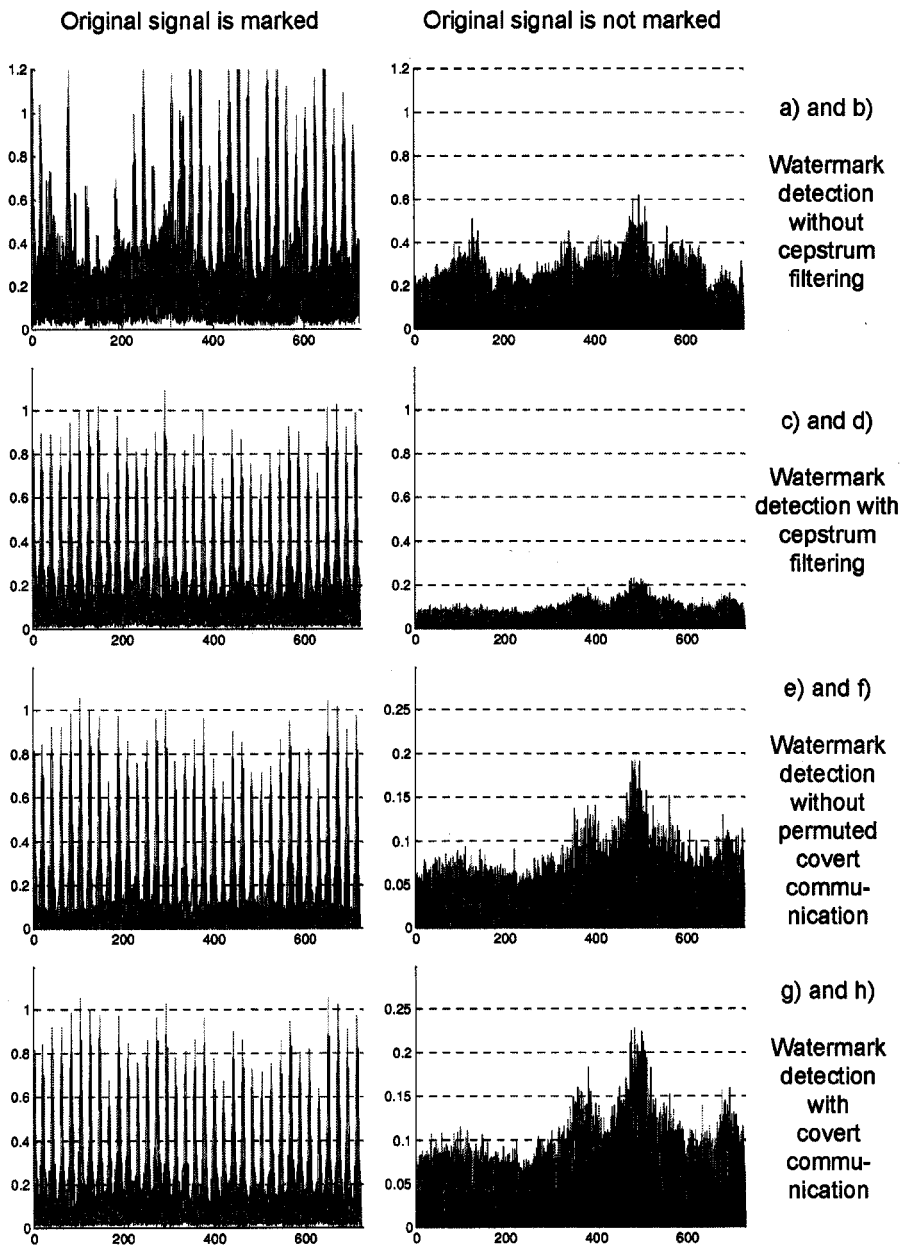

Fig. 10. Detection comparison for four different detection systems (a), (b) without and (c), (d) with cepstrum filtering and (e), (f) without and (g), (h) with a permuted covert communication channel. For each diagram, the *x-axis* depicts the timeline in MCLT blocks, whereas the *y-axis* quantifies the normalized correlation.

same as the ones included in the version of the tool available on the Web [18].

IV. Security Analysis

We now evaluate the security of our watermarking mechanisms with respect to the watermark estimation attack. As discussed in the previous section, we introduced block repetition codes and multiple correlation tests to enforce synchronization for attacks with limited variable scaling. Therefore, in improving robustness against signal deformation attacks, we introduced a certain amount of redundancy in the watermarking pattern. That improves the chances that an attacker can estimate the WM chips from the marked signal [19]. Thus, we need to quantify the efficiency of such attacks and devise new mechanisms to protect against them.

In order to simplify the formal description of block repetition codes in our audio WM codec, we now modify slightly our notation. The marked signal y is created by adding the WM with certain magnitude δ to the original

$$y = x + \delta w, w \in \{\{-1\}^m, \{1\}^m\}^n . \tag{13}$$

Vectors y and x have $N = m \times n$ samples, whereas w has n chips, each of them replicated successively m times. The WM detector correlates the averages of the central m_o elements of each region marked with the same chip, where commonly, $m_o = m/k$, and $k \in \{2, 5\}$. Such a detector can tolerate fluctuation in content scaling up to $(k-1)m/2k$ signal coefficients.

TABLE I
WATERMARK DETECTION RESULTS ON AUDIO CLIPS ATTACKED WITH THE STIRMARK AUDIO BENCHMARK. PARAMETERS OF THE ATTACKS ARE INHERITED FROM THE VERSION OF THE TOOL AVAILABLE ONLINE

Attack	$\overline{C(y,w)}$	Attack	$\overline{C(y,w)}$	Attack	$\overline{C(y,w)}$	Attack	$\overline{C(y,w)}$
Original	0.0619	Marked	1.004	addbrumm_100	1.0049	addbrumm_10100	0.9934
addbrumm_1100	0.9986	addbrumm_2100	0.9918	addbrumm_3100	0.9923	addbrumm_4100	0.9939
addbrumm_5100	0.993	addbrumm_6100	0.9973	addbrumm_7100	0.9935	addbrumm_8100	0.9942
addbrumm_9100	0.9918	addnoise_100	1.004	addnoise_300	1.0097	addnoise_500	1.0076
addnoise_700	1.0009	addnoise_900	1.0069	addsinus	1.0062	amplify	1.0075
compressor	1.0086	**copysample**	**0.0761**	cutsamples	0.9797	dynnoise	1.0097
echo	0.9128	exchange	0.9767	extrastereo_30	1.0061	extrastereo_50	1.0061
extrastereo_70	1.0061	hlpass	1.0103	invert	1.0053	fft_real_reverse	1.006
fft_stat1	1.0032	fft_test	1.0019	flippsample	1.0084	invert	1.0061
lsbzero	1.006	normalize	1.0059	rc_highpass	1.0083	rc_lowpass	1.0076
resampling	1.0092	smooth	1.0147	smooth2	1.007	stat1	1.009
stat2	1.0075	zerocross	0.9815	zerolength	0.9823	zeroremove	1.0013

The involved block repetition code improves the detection, but it also improves the efficacy of the estimation attack. If all details of the embedder are known (except w), the adversary can compute the WM estimate, amplify it with a factor $\alpha > 1$, and then subtract the amplified attack vector from the marked content [2].

Theorem 1: Given a set of m samples of x marked with the same chip w_i such that

$$y_{(i-1)m+j} = x_{(i-1)m+j} + \delta w_i, 1 \le j \le m \tag{14}$$

the optimal estimate v_i of the hidden WM chip w_i is given by

$$v_i = \text{sign}\left(\sum_{j=1}^{m}\left(x_{(i-1)m+j} + \delta w_i\right)\right). \tag{15}$$

See [2, Lemma 1] for proof. Note that $v \in \{\pm 1\}^N$.

Theorem 2: The optimal WM estimation, as presented in Theorem 1, yields the following probability of estimation error per WM chip:

$$\varepsilon = \Pr\left[v_i \neq w_i\right] = \frac{1}{2}\text{erfc}\left(\frac{\delta\sqrt{m}}{\sigma_x\sqrt{2}}\right). \tag{16}$$

See [2, Coroll. 1] for proof.

The estimation attack is performed by subtracting an amplified WM estimate αv from the marked content y:

$$z = y - \alpha v. \tag{17}$$

The maximum value of the amplification factor α depends solely on the desired level of audibility for the attack. In practice, α can be much greater than δ because the content marking entity is subject to much more stringent content fidelity constraints than an attacker.

Corollary 1: The variance $Var[z] = \sigma_z^2$ of the attacked signal depends on α as presented:

$$\sigma_z^2 = \frac{1}{n}E\left[\sum_{i=1}^{n}(y_i - \alpha v_i)^2\right] = \sigma_x^2 + \delta^2 + \alpha^2 - 2\alpha\xi$$
$$\xi = \sigma_x e^{-(\delta^2/(2\sigma_x^2))}\sqrt{\frac{2}{\pi}} + \delta\,\text{erf}\left(\frac{\delta}{\sigma_x\sqrt{2}}\right). \tag{18}$$

Proof (sketch): By replacing $(y_i - \alpha v_i)$ in (18) with $(x_i + \delta w_i - \alpha\,\text{sign}(x_i + \delta w_i))$, we obtain

$$\sigma_z^2 = \sigma_x^2 + \delta^2 + \alpha^2 - \frac{2\alpha}{n}E\left[\sum_{i=1}^{n}|x_i + \delta w_i|\right] \tag{19}$$

which proves (18) to be correct. ■

Corollary 2: After the attack, the expected correlation value computed by the WM detector equals

$$E[z \cdot w] = \delta - \alpha(1 - 2\varepsilon) = \delta - \alpha\,\text{erf}\left(\frac{\delta\sqrt{m}}{\sigma_x\sqrt{2}}\right) \tag{20}$$

with $Var[z \cdot w] \approx (\sigma_z^2)/(nm_o)$.

Fig. 11 demonstrates how $E[z \cdot w]$ and σ_z change as α increases under fixed $\delta = 1.5$, with $\sigma_x/\sqrt{m}$ varying from 2.5 to 6.5.

From (20), we compute that in order to draw the expected correlation value to $E[z \cdot w] = \theta$, the attacker has to induce $\alpha(\theta)$ equal to

$$\alpha(\theta) = \frac{\delta - \theta}{\text{erf}\left(\frac{\delta\sqrt{m}}{\sigma_x\sqrt{2}}\right)}. \tag{21}$$

If $v \neq w$ or $\alpha \neq \delta$, the estimation attack adds noise to the marked signal. Part of this noise is an accurate estimate of the WM, and it actually reverses the effect of the watermarking process. The remainder of the attack vector is applied in addition to the existing marked data.

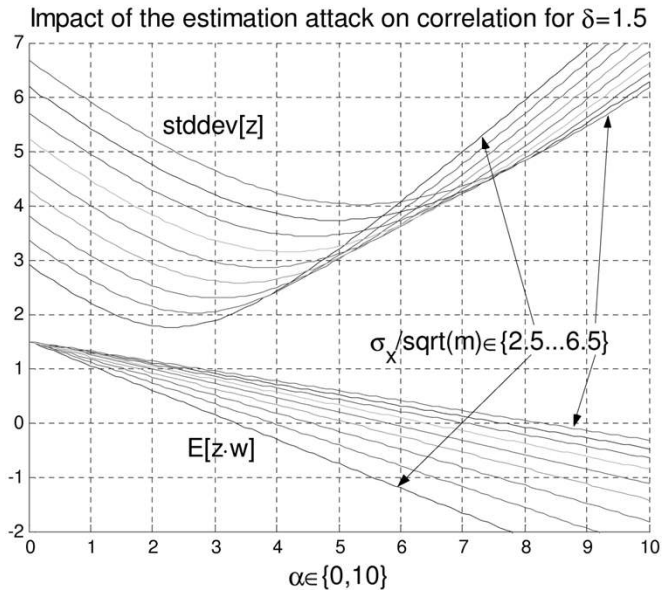

Fig. 11. Diagram of the dependency of $E[z \cdot w]$ and $Var[z]^{1/2}$ as α increases from 0 to 10 for fixed $\delta = 1.5$ and variable $(\sigma_x/\sqrt{m}) \in \{2.5 \ldots 6.5\}$.

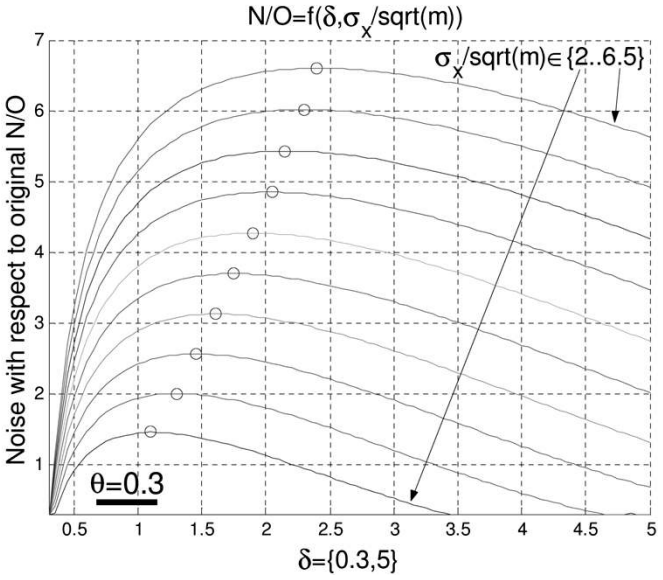

Fig. 12. Diagram of the dependency of N/O (26) with respect to δ for given $(\sigma_x/\sqrt{m}) \in \{2 \ldots 6.5\}$ and $\theta = 0.3$.

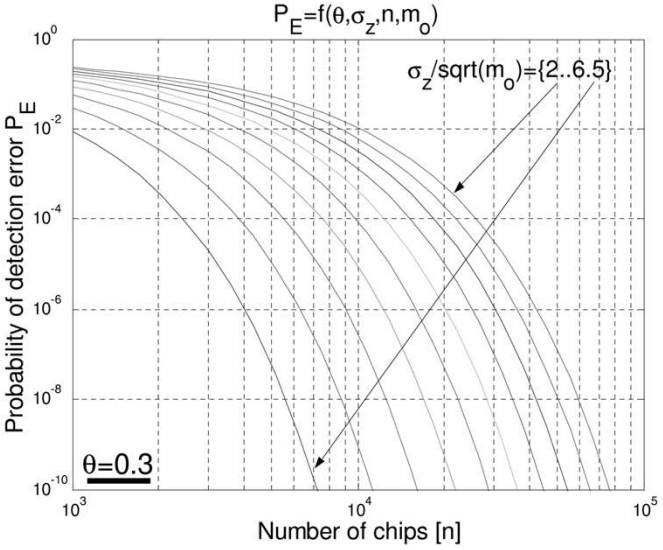

Fig. 13. Dependency diagram for P_E (24) with respect to n for given $(\sigma_z/\sqrt{m_o}) \in \{2 \ldots 6.5\}$ and $\theta = 0.3$.

Corollary 3: The estimation attack on a marked content described in (17) induces the following additive noise with respect to the original signal

$$N/O \equiv E\left[|z_i - x_i|\right] = \begin{cases} \alpha + \delta(2\varepsilon - 1), & \alpha \geq \delta \\ \delta + \alpha(2\varepsilon - 1), & \alpha < \delta \end{cases} \tag{22}$$

whereas the added noise with respect to the marked copy equals

$$\frac{N}{M} \equiv E\left[|z_i - y_i|\right] = \alpha < N/O. \tag{23}$$

A realistic attack/detect watermarking model would assume the following criteria.

Criterion 1: The amplitude of the attack α is limited by the induced noise as $N/O \leq \beta$.

Criterion 2: An attack with fixed $\alpha(\theta)$ draws the expected value of the correlation to a value $E[z \cdot w] = \theta$. For a fixed WM length n and detection decision threshold $\tau = \theta/2$ that achieves symmetric probability of false alarm P_{FA} and misdetection P_{MD}, the detection error probability $P_E = P_{\text{FA}} = P_{\text{MD}}$ is upper bounded by at most γ

$$P_E = \frac{1}{2}\text{erfc}\left(\frac{\theta\sqrt{nm_o}}{2\sigma_z\sqrt{2}}\right) \leq \gamma. \tag{24}$$

It is important to stress that the efficiency of SS watermarking and detection depends by and large on the parameters that are content dependent.

Problem 1: For a given σ_x, what is the optimal value of δ such that under the optimal estimation attack described in (17) and quantified using α, maximal N/O is induced while Criterion 2 is satisfied?

The posed problem can be solved in two steps.

1) From Criterion 2, we can compute the minimal expected value for the normalized correlation $E[z \cdot w] \geq \theta$ after the attack:

$$\theta = \frac{2\sigma_z\sqrt{2}\,\text{erf}^{-1}(1 - 2\gamma)}{\sqrt{nm_o}}. \tag{25}$$

2) From (16), (21), and (22), we can compute the dependency of the induced N/O on the WM magnitude δ:

$$N/O = f(\delta) = \frac{\delta - \theta}{\text{erf}\left(\frac{\delta\sqrt{m}}{\sigma_x\sqrt{2}}\right)} - \delta\,\text{erf}\left(\frac{\delta\sqrt{m}}{\sigma_x\sqrt{2}}\right) \tag{26}$$

from which we can numerically find the desired δ that maximizes the induced N/O.

Fig. 12 depicts the dependency of N/O with respect to δ for $(\sigma_x/\sqrt{m}) \in \{2 \ldots 6.5\}$. Optimal values $\delta(\sigma_x)$, which result in maximal N/O, are depicted using the $\{\circ\}$ symbol. Fig. 13 illustrates the probability of a detection error P_E (for $\tau = \theta/2$) with respect to a given WM length of n chips and for $(\sigma_z/\sqrt{m_o}) \in \{2 \ldots 6.5\}$ and $\theta = 0.3$.

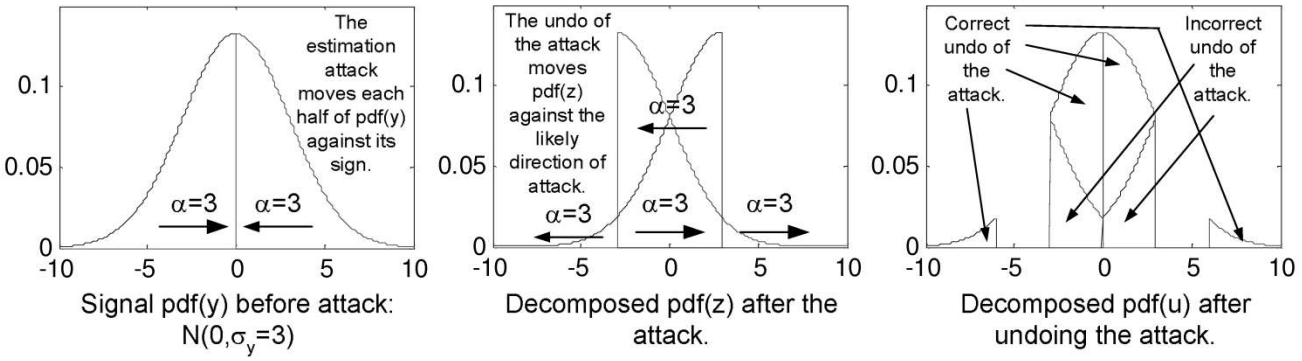

Fig. 14. Illustration of the *undo* of the estimation attack.

A. Undoing the Estimation Attack

In this subsection, we demonstrate a remedy for the estimation attack described in (17) (see Fig. 14). The main idea is to optimally reverse the attack, i.e., estimate the signal coefficient y_i from the attacked signal z_i. We also demonstrate that a slight modification to the attack in (17) succeeds in removing the WM (or disabling the detector to identify the WM) by adding additional noise to the attacked signal.

Definition 1: The *undo* operator $\mathcal{U}(z_i, \alpha)$, where $z_i, \alpha \in \mathbb{R}$, is defined as follows:

$$u_i = \mathcal{U}(z_i, \alpha) = \begin{cases} z_i + \alpha \operatorname{sign}(z_i), & |z_i| > 2\alpha \\ z_i - \alpha \operatorname{sign}(z_i), & |z_i| \leqslant 2\alpha. \end{cases} \tag{27}$$

Theorem 3: Given a signal coefficient z_i created using the estimation attack as $z_i = y_i - \alpha \operatorname{sign}(y_i)$, where y_i is a weighted sum $y_i = x_i + \delta w_i$ of a Gaussian zero-mean i.i.d. variable x_i and a SS sequence chip w_i and $\alpha \geq \delta$, optimal estimation u_i of the signal y_i such that $E[|u_i - y_i|]$ is minimal is given using the *undo* operator $u_i = \mathcal{U}(z_i, \alpha)$.

Proof (sketch): When doing the estimation attack, the adversary shifts the positive and negative pdf of the marked signal $\tilde{y}$ for α against the sign of y. The *undo* operation described in (27) retrieves all values of the original signal $y : \forall(y_i > 2\alpha)$, $u_i = y_i$. Now, let us define a subset $Y(a, b) \subset y$ s.t. $y_i \in Y(a, b)$ iff $a < y_i < b$. Since for a zero-mean Gaussian distribution of x_i and $\alpha > \delta$, $|Y(0, \alpha)| > |Y(-2\alpha, -\alpha)|$, then u_i is the optimal estimation of y_i based on a given z_i s.t. $-\alpha < z_i < 0$. Similarly, since $|Y(-\alpha, 0)| > |Y(\alpha, 2\alpha)|$, u_i is an optimal estimation of y_i based on z_i s.t. $0 < z_i < \alpha$. ■

Corollary 4: The expected value for the correlation of the recovered u and w is given by

$$E[u \cdot w] = \delta - \alpha[\operatorname{erfc}(a) - \operatorname{erfc}(b) - \operatorname{erfc}(c) + \operatorname{erfc}(d)] \tag{28}$$

where $a = (\alpha - \delta)/\sqrt{2}\sigma_x$, $b = (\alpha + \delta)/\sqrt{2}\sigma_x$, $c = (2\alpha - \delta)/\sqrt{2}\sigma_x$, and $d = (2\alpha + \delta)/\sqrt{2}\sigma_x$.

Proof (sketch): The *undo* of the estimation attack cannot recover the magnitudes of $Y(-2\alpha, -\alpha) \cup Y(\alpha, 2\alpha)$ that got mixed with $Y(-\alpha, 0) \cup Y(0, \alpha)$ during the attack. We compensate the final correlation $E[u \cdot w]$ as follows: $E[u \cdot w] = \delta - C_1 + C_2$ with

$$C_1 = \int_{\alpha}^{2\alpha} [(x - \delta)f(x - \delta) + (x + \delta)f(x + \delta)]\, dx$$
$$C_2 = \int_{\alpha}^{2\alpha} \Big[(x - \delta - 2\alpha)f(x - \delta) + (x + \delta - 2\alpha)f(x + \delta)\Big]\, dx \tag{29}$$

where $f(x + c)$ is a function of the Gaussian distribution centered at c with variance σ_x^2, which results in (28). ■

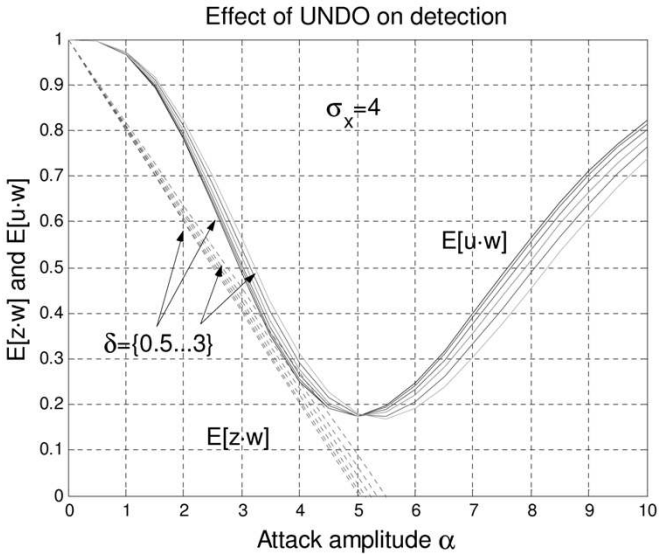

Fig. 15. Effect of the undo test on the correlation test. As α increases, the figure shows how $E[z \cdot w]$ and $E[u \cdot w]$ change for fixed $\sigma_x = 4$ and variable $\delta \in [0.5 \ldots 3]$.

Fig. 15 illustrates the effect of the *undo* operation on WM detection. Whereas the correlation value of a traditional SS WM detector $E[z \cdot w]$ inevitably converges to zero, the correlation after the *undo* operation yields $\mu = \min(E[u \cdot w]) > 0$. Thus, according to (24), for a sufficiently long SS sequence (nm_o elements of u are integrated), a detection threshold at $\tau = \min(E[u \cdot w])/2$ would yield desired detection results, regardless of the strength of the estimation attack. In the region of interest, i.e., α s.t. $(y - \alpha v) \cdot w \approx \mu$, the correlation variance

satisfies $Var[u \cdot w] \approx \sigma_z^2/n$, where σ_z is computed in Corollary 1.

The detector cannot possibly know the attack amplification value α while performing the detection. However, note that for any α, $E[\mathcal{U}(x,\alpha),w] = 0$ [with $\mathcal{U}(x,\alpha) = \{\mathcal{U}(x_i,\alpha), i = 1..n\}$], where x is a signal which does not have w embedded. Thus, the detector can perform T tests $E[\mathcal{U}(x,\alpha_t),w]$ for realistic values of $\alpha = \{\alpha_t,\ t = 1..T\}$ that can potentially break the system (e.g., $\alpha = \{3,4,5\}$).

The power of the *undo* operation is based on the inequivalent distribution of magnitudes marked with positive and negative chips. Therefore, the attacker must impose additional noise n to the attacked signal z such that the latter distributions are equalized. While the "smart noise" $-\alpha v$ draws $E[z \cdot w]$ to zero, the additional noise n enables that no *undo* operation is able to retrieve even a small part of the original distributions of the signal marked with positive and negative chips.

A modified *undo* operation

$$u_i = \mathcal{U}'(z_i,\alpha) = \begin{cases} z_i + \alpha\,\mathrm{sign}(z_i), & |z_i| \geqslant 2\alpha+\beta \\ z_i - \alpha\,\mathrm{sign}(z_i), & |z_i| < 2\alpha+\beta \end{cases}$$

with $\beta \sim \sigma_n$ may strengthen the detection procedure; however, its effectiveness is very limited. Because of the *undo* operation, the estimation attack needs to be modified as follows:

$$z = y - \alpha v + n \tag{30}$$

where n is a noise pattern aiming to equalize the distributions of magnitudes marked with positive and negative chips. For example, white noise of amplitude $\sigma_n \sim \sigma_x/2$ commonly creates a difficult task for the designer of an *undo* operation.

V. Final Remarks

We now consider three key aspects of SS-based audio watermarking.

A. Justifying the Gaussian Assumption

The linear marking (13) and detection (Corollary 4) process is performed on the audible, averaged, and cepstrum-filtered coefficients of a 2048-long MCLT analysis window [15] in the logarithmic (dB) domain. We have observed that on a great variety of audio clips, even the *individual* filtered coefficients can be accurately modeled using a Gaussian PDF. In addition, the detector averages the central m_o coefficients in each repetition block, which significantly improves the modeling accuracy due to the central limit theorem. Thus, the final working vector y extracted from the audio clip can be highly accurately macro-modeled as a Gaussian vector. Local correlations and nonstationarity are effectively cancelled using sufficiently large windows (e.g., 1024 window size at a sampling frequency of 44.1 kHz), cepstrum filtering, and running-average windowing along the time axis.

B. How Does Redundancy Impact Detector and Estimator Performance on Real-Life Data?

The reliability of detection as well as the performance of the WM estimator depend on the variance of the original working vector x. Fig. 16 illustrates the standard deviation

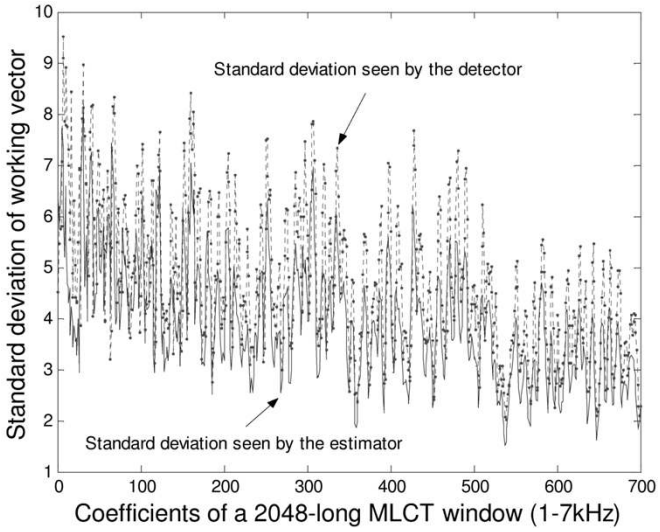

Fig. 16. Standard deviation of a typical music signal computed per transform coefficient for a 2048-long MCLT block for two different redundancy metrics 3×5 (seen by the detector) and 5×9 (seen by the estimator), where the two parameters represent corresponding redundancy along the frequency and time axis, respectively.

$\sigma_E = \sigma_x/\sqrt{m}$ that the estimator sees, assuming it knows perfectly the location of the WM and the standard deviation $\sigma_D = \sigma_x/\sqrt{m_o}$ that the detector sees while computing the correlation test. Block repetition assumed in this case is $m = 5 \times 9$ coefficients along the frequency and time axis, respectively. The corresponding region for detection is $m_o = 3 \times 5$ coefficients.

According to Fig. 16, we locate the WM to the 200 Hz–2 kHz region for three reasons. First, HAS is much more sensitive to noise in this sub-band (a noise of only 4 dB can rarely be tolerated). Second, the variance of the carrier signal is higher in this region, providing a more robust host for data hiding with respect to the estimation attack. Third, although the ratio $\sqrt{m/m_o} = \sqrt{3}$, in the proposed subband, the actual σ_E/σ_D retrieved experimentally from over 100 audio clips is only 1.18.

C. What is the Impact of the Results Obtained so far on Audio Watermarking?

We have presented a generic recipe for using SS to hide secrets in multimedia content. For a typical music content, if the SS WM is located in the 200 Hz–2 kHz sub-band, in order to draw the correlation of the new *undo* correlation test to a value that forces detection failure, the adversary needs to add total noise in the excess of 6 dB, which may be intolerable to many users. SS WM length that would enable false alarm accuracy of $P_{\mathrm{FA}} \approx 10^{-6}$ would require approximately an 80-s music frame. A WM of such length is difficult to synchronize at the detector. Although block repetition codes enable wow-and-flutter tolerance required for most low-end turntables (e.g., 0.15% playtime fluctuation), it is arguable whether a common HAS would discard such content as of no value.

On the other hand, techniques presented in this paper may provide better results for data hiding in video signals, as we estimate that per frame, significantly more chips can be embedded, resulting in shorter watermarks, i.e., higher robustness to frame dropping and limited geometric distortions.

ACKNOWLEDGMENT

The authors would like to thank Dr. R. Venkatesan, Dr. M. K. Mihçak, and M. Kesal for many useful suggestions on the optimal attacks against block repetition codes.

REFERENCES

[1] Recording Industry Association of America [Online]. Available: http://www.riaa.org.

[2] D. Kirovski, H. S. Malvar, and Y. Yacobi. (2001) A dual watermaking and fingerprinting system. Microsoft Research. [Online]. Available: http://research.microsoft.com.

[3] S. Katzenbeisser and F. A. P. Petitcolas, *Hiding Techniques for Steganography and Digital Watermarking*, S. Katzenbeisser and F. A. P. Petitcolas, Eds. Boston, MA: Artech House, 2000.

[4] P. Bassia and I. Pitas, "Robust audio watermarking in the time domain," in *Proc. EUSIPCO*, vol. 1, Rodos, Greece, Sept. 1998, pp. 25–28.

[5] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Proc. Inform. Hiding Workshop*, Cambridge, U.K., June 1996, pp. 147–158.

[6] C. Neubauer and J. Herre, "Digital watermarking and its influence on audio quality," in *Proc. 105th AES Conv.*, San Francisco, CA, Sept. 1998.

[7] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Process.*, vol. 66, no. 3, pp. 337–355, 1998.

[8] D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Proc. Inform. Hiding Workshop*, Cambridge, U.K., June 1996, pp. 293–315.

[9] W. Szepanski, "A signal theoretic method for creating forgery-proof documents for automatic verification," in *Proc. Carnahan Conf. Crime Countermeasures*, Lexington, KY, May 1979, pp. 101–109.

[10] B. Chen and G. W. Wornell, "Digital watermarking and information embedding using dither modulation," in *Proc. Workshop Multimedia Signal Process.*, Redondo Beach, CA, Dec. 1998, pp. 273–278.

[11] J. K. Su and B. Girod, "Power-spectrum condition for energy-efficient watermarking," in *Proc. Int. Conf. Image Process.*, Kobe, Japan, Oct. 1999, pp. 301–305.

[12] J. P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proc. Inform. Hiding Workshop*, Portland, OR, Apr. 1998, pp. 258–272.

[13] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE J. Select Areas Commun.*, vol. 16, pp. 474–481, May 1998.

[14] H. L. van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley, 1968.

[15] H. S. Malvar, "A modulated complex lapped transform and its application to audio processing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, May 1999, pp. 1421–1424.

[16] K. Brandenburg, "Coding of high quality digital audio," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer, 1998.

[17] Secure digital music initiative [Online]. Available: http://www.sdmi.org

[18] M. Steinebach, A. Lang, J. Dittmann, and F. A. P. Petitcolas, "Stirmark benchmark: Audio watermarking attacks based on lossy compression," in *Proc. SPIE Security Watermarking Multimedia*, vol. 4675, San Jose, CA, Jan. 2002, pp. 79–90.

[19] M. K. Mihçak, R. Venkatesan, and M. Kesal, "Cryptanalysis of discrete-sequence spread-spectrum watermarks," in *Proc. Inform. Hiding Workshop*, Noordwijkerhout, Netherlands, Oct. 2002.

Darko Kirovski received the Ph.D. degree in computer science from the University of California, Los Angeles, in January 2001.

Since April 2000, he has been a researcher at Microsoft Research, Redmond, WA. His research interests include secure systems, software delivery systems, multimedia processing and applications, intellectual property protection, and embedded system design and debugging. He has coauthored more than 50 journal and conference papers.

Dr. Kirovski received a 1999–2001 Microsoft Graduate Research Fellowship, the 1999–2000 ACM/IEEE Design Automation Conference Graduate Scholarship, and the 2002 ACM Outstanding Ph.D. Dissertation in Electronic Design Automation Award.

Henrique S. Malvar (M'79–SM'91–F'97) received the B.S. degree in 1977 from Universidade de Brasília, Brasília, Brazil, the M.S. degree in 1979 from Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, and the Ph.D. degree in 1986 from the Massachusetts Institute of Technology (MIT), Cambridge, all in electrical engineering.

From 1979 to 1993, he was with the faculty of the Universidade de Brasília. From 1986 to 1987, he was a Visiting Assistant Professor of electrical engineering at MIT and a senior researcher at PictureTel Corporation, Andover, MA. In 1993, he rejoined PictureTel, where he stayed until 1997 as Vice President of Research and Advanced Development. Since 1997, he has been a Senior Researcher at Microsoft Research, Redmond, WA, where he heads the Communication, Collaboration, and Signal Processing Research Group. His research interests include multimedia signal compression and enhancement, fast algorithms, multirate filterbanks, and wavelet transforms. He has several publications in these areas, including the book *Signal Processing with Lapped Transforms* (Boston, MA: Artech House, 1992). He is an Associate Editor for the journal *Applied and Computational Harmonic Analysis*.

Dr. Malvar is an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and a member of the Signal Processing Theory and Methods Technical Committee of the IEEE Signal Processing Society. He received the Young Scientist Award from the Marconi International Fellowship and Herman Goldman Foundation in 1981. He also received the Senior Paper Award in Image Processing in 1992 and the Technical Achievement Award in 2002, both from the IEEE Signal Processing Society.