Proc. SPIE, vol. 3971: Security and
Watermarking of Multimedia Contents II
San Jose, CA, USA, Jan. 2000.

Fundamental Performance Limits of Power-Spectrum Condition-Compliant Watermarks

Jonathan K. Su
Telecommunications Laboratory, Univ. Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
Bernd Girod
Information Systems Laboratory, Stanford Univ.
350 Serra Mall, Packard Elec. Eng. Bldg., Rm. 258
Stanford, CA 94305-9510 USA

ABSTRACT

Our previous work introduced a stochastic framework for studying watermarking. When the receiver used direct correlation detection and the attack employed linear shift-invariant (LSI) filtering, the optimal attack consisted of Wiener filtering and adding noise. Resisting the *Wiener attack* led to a *power-spectrum condition* (PSC): the watermark and original power spectra should be directly proportional. PSC-compliant watermarks are most difficult to estimate from the watermarked document in the minimum mean-squared error sense. This paper investigates the fundamental limits of PSC-compliant watermarks; it has two main parts. First, expressions are derived for the detection-error probabilities after the Wiener attack with PSC-compliant Gaussian noise. An explicit relationship between attacked-document distortion and capacity results. The second part studies optimal LSI-filtering attacks and receivers. Because direct correlation detection is only optimal for *white Gaussian noise* (WGN), the Wiener attack and PSC may actually be suboptimal. It is shown that the PSC holds for a suboptimal, effective white-noise attack. The optimum attack is derived, but there does not appear to be an analytical solution for the optimum watermark power spectrum. Numerical experiments suggest that, as a rule of thumb, white watermarks perform well at lower distortions, while PSC-compliant watermarks perform better at higher distortions.

Keywords: Digital watermarking, capacity, watermark attacks, information theory

1. INTRODUCTION

Digital watermarking is the imperceptible, secure, and robust communication of information by embedding it directly into digital documents for later retrieval. This information should be embedded in such a way that the watermarked document is perceptually the same as the original document (this is what is meant by "imperceptible"). In most watermarking applications, the information conveyed by the watermark should only be accessible by authorized parties (secure). Finally, it is usually desired that the information be retrievable even after processing of the watermarked document (robust). Any processing that may hinder communication is termed an *attack*.

This paper continues our investigation of a theoretical model for watermarking and attacks. The ultimate goal is to determine performance limits that apply to many watermarking schemes. Our approach is based on a stochastic framework and models attacks by linear shift-invariant (LSI) filtering and additive noise. This model, while ideal, allows us to apply powerful tools from probability theory, signal processing, and information theory.

We view watermarking as a communications problem and consider the perspectives of both the owner (authorized parties) and the attacker. The owner wishes to embed and retrieve as much information as possible, subject to limits such as the size and distortion of the watermarked document. The attacker wishes to prevent the communication of this information but must keep the attacked-document distortion to an acceptable level.

Sec. 2 reviews the model and some of our prior results, namely the Wiener attack and power-spectrum condition (PSC). Sec. 3 extends these results to find performance limits for watermarks that fulfill the PSC and are attacked by the Wiener attack. Sec. 4 re-examines the underlying problem, and Sec. 5 shows that the Wiener attack on PSC-compliant watermarks is equivalent to a suboptimal attack. Sec. 6 investigates the optimum attack, and conclusions appear in Sec. 7.

Further author information: Send correspondence to J. K. Su. Email: su@LNT.de. At the times of submission and acceptance of this paper, B. Girod was with the Telecommunications Laboratory, Univ. Erlangen-Nuremberg; he is now with the Information Systems Laboratory, Stanford Univ.

2. WATERMARKING AND ATTACKS

We work with M-dimensional (M-D) signals and Gaussian discrete-time/space random processes. All random processes are assumed ergodic, zero-mean, and wide-sense stationary. The M-D indexing of a signal $x[n_1, n_2, \ldots n_M]$ is abbreviated by $x[\vec{n}]$, where $\vec{n} = (n_1, \ldots, n_M)$. Likewise, M-D indexing of a frequency-domain signal $X(\omega_1, \ldots, \omega_M)$ is denoted by $X(\vec{\omega})$, with $\vec{\omega} = (\omega_1, \ldots, \omega_M)$. The M-D Dirac delta function $\delta[\vec{n}]$ is unity when $\vec{n} = (0, 0, \ldots, 0)$, and zero otherwise. Non-random quantities are denoted by a normal font (e.g., $x[\vec{n}]$, c), and random quantities by a bold font (e.g., $\mathbf{x}[\vec{n}]$, $\mathbf{c}$). We denote the set of indices forming the signal support by $\mathcal{N}$, and let N be the *support size*, i.e., the cardinality of $\mathcal{N}$.

2.1. Signal Models and Distortion

The original signal is modeled by $\mathbf{x}[\vec{n}]$; it has variance σ_x^2, autocorrelation function $R_{xx}[\vec{k}]$, and power spectrum $\Phi_{xx}(\vec{\omega})$. Similarly, the watermark is $\mathbf{w}[\vec{n}]$, which is independent of $\mathbf{x}[\vec{n}]$ and has variance σ_w^2, autocorrelation function $R_{ww}[\vec{k}]$, and power spectrum $\Phi_{ww}(\vec{\omega})$. The random formulation of the watermark is appropriate when the watermark is not known, e.g., to the attacker. Of course, during watermark embedding and retrieval by authorized parties, the watermark is known. Here, the watermark is the realization $w[\vec{n}]$ of $\mathbf{w}[\vec{n}]$. We assume that, by ergodicity, the sample averages of $w[\vec{n}]$ equal the ensemble averages of $\mathbf{w}[\vec{n}]$.

In watermarking, it is necessary to consider the distortion of various signals. For mathematical tractability, we use the *mean-squared error* (MSE) distortion measure, so the distortion D_{rx} between a signal $\mathbf{r}[\vec{n}]$ and the original $\mathbf{x}[\vec{n}]$ is $D_{rx} = \mathrm{E}\left[(\mathbf{r}[\vec{n}] - \mathbf{x}[\vec{n}])^2\right]$. We often express distortion in decibels (dB).

2.2. Watermark Embedding, Wiener Attack, and Watermark Retrieval

Watermark embedding is modeled by simple addition; the watermarked signal is $\mathbf{y}[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}]$. The distortion of the watermarked signal is $D_{yx} = \sigma_w^2$.

We assume that $\mathbf{y}[\vec{n}]$ is subjected to the *Wiener attack*,[1] which performs MMSE estimation of the watermark, subtracts the weighted estimate, and adds noise. The attacked signal $\hat{\mathbf{y}}[\vec{n}]$ is

$$\hat{\mathbf{y}}[\vec{n}] = \mathbf{y}[\vec{n}] - \gamma h[\vec{n}] * \mathbf{y}[\vec{n}] + \mathbf{v}[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}] - \gamma \hat{\mathbf{w}}[\vec{n}] + \mathbf{v}[\vec{n}], \tag{1}$$

where $h[\vec{n}]$ is the impulse response of the Wiener filter for estimating $\mathbf{w}[\vec{n}]$ from $\mathbf{y}[\vec{n}]$, $\gamma \geq 0$ is a scalar gain factor, $\hat{\mathbf{w}}[\vec{n}] = h[\vec{n}] * \mathbf{y}[\vec{n}]$ is the estimated watermark, and $\mathbf{v}[\vec{n}]$ is independent, *colored Gaussian noise*.

Given the realization $\hat{y}[\vec{n}]$, the receiver must decide whether or not the watermark realization $w[\vec{n}]$ is present. We assume a binary-hypothesis decision where H_w corresponds to $w[\vec{n}]$ present, and H_o corresponds to $w[\vec{n}]$ absent and a different watermark $w_o[\vec{n}]$ present. The latter is assumed to be a realization of the random process $\mathbf{w}_o[\vec{n}]$, which has the same properties as $\mathbf{w}[\vec{n}]$ but is independent of $\mathbf{w}[\vec{n}]$.

In many watermarking schemes,[2,3] the receiver performs *correlation detection*. It computes the time/space-averaged correlation $c = (1/N) \sum_{\vec{n} \in \mathcal{N}} (\hat{y}[\vec{n}] - a' x[\vec{n}])\, w[\vec{n}]$, where the term $(-a' x[\vec{n}])$ reflects the possibility that the receiver may have knowledge of the original $x[\vec{n}]$ and can eliminate interference from it. The detector decides in favor of H_w if $c > T$; otherwise, it decides in favor of H_o. To study the average performance of the detector, $x[\vec{n}]$ and $v[\vec{n}]$ are treated as random processes so that c may be treated as a random variable (RV) $\mathbf{c}$. The conditional moments $\mathrm{E}[\mathbf{c}|H_i]$ and $\mathrm{var}(\mathbf{v}|H_i)$, $i = w, o$, are sufficient to determine the threshold T that minimizes the average probability of error P_E.

We assume synchronization between the embedding and retrieval steps. Many attacks on watermarking schemes are based upon disrupting synchronization, but they do not remove the watermark. A more sophisticated receiver should be able to reacquire synchronization.[4,5] For this reason, we maintain this assumption.

2.3. Review of Wiener Attack and Power-Spectrum Condition (PSC)

Here we review some properties of the Wiener attack and the owner's means of countering it. For the correlation detector, the Wiener attack has the following property[1]: For a given value $\mathrm{E}[\mathbf{c}|H_w] = c_t$, the attacked signal $\hat{\mathbf{y}}[\vec{n}]$ produced by the Wiener attack has the smallest distortion $D_{\hat{y}x}$ among all LSI-filtering/additive-noise attacks. Two special cases of the Wiener attack are the following. When $\gamma = 1$ and $\sigma_v^2 = 0$, we have a "removal attack," which minimizes $D_{\hat{y}x}$. When $\gamma = \gamma_0 = \sigma_w^2/\sigma_{\hat{w}}^2$ and $\sigma_v^2 = 0$, we have an "anticorrelation attack," which makes $\mathrm{E}[\mathbf{c}|H_w] = 0$ and causes the detector to fail completely.

To resist the Wiener attack, the owner should make estimation of $\mathbf{w}[\vec{n}]$ as difficult as possible. It can be shown[6] that, to maximize the MSE $E = \mathrm{E}\big[\,(\mathbf{w}[\vec{n}] - \hat{\mathbf{w}}[\vec{n}])^2\,\big]$ of the estimated watermark, the watermark power spectrum must satisfy a necessary and sufficient condition called the *power-spectrum condition* (PSC):

$$\Phi_{ww}(\vec{\omega}) = \frac{\sigma_w^2}{\sigma_x^2}\,\Phi_{xx}(\vec{\omega}). \tag{2}$$

In this sense, "the watermark should look like the original."

When the PSC is fulfilled, $E = \alpha_{\mathrm{PSC}}\sigma_w^2$, where $\alpha_{\mathrm{PSC}} = \sigma_x^2/\left(\sigma_w^2 + \sigma_x^2\right)$. The Wiener filter impulse response becomes $h[\vec{n}] = (1-\alpha_{\mathrm{PSC}})\,\delta[\vec{n}]$. The cross-correlation and distortion become, respectively,

$$\mathrm{E}\,[\mathbf{c}|H_w] = (1 - \gamma\,(1-\alpha_{\mathrm{PSC}}))\,\sigma_w^2, \quad \text{and} \quad D_{\hat{y}x} = [1 - \gamma(2-\gamma)\,(1-\alpha_{\mathrm{PSC}})]\,\sigma_w^2 + \sigma_v^2. \tag{3}$$

Note that these expressions are independent of the dimensionality M.

PSC-compliant watermarks are most robust against the Wiener attack in the sense that they maximize $D_{\hat{y}x}$ for a given value of $\mathrm{E}\,[\mathbf{c}|H_w] = c_t$. Against PSC-compliant watermarks, the removal attack has almost no effect ($\mathrm{E}\,[\mathbf{c}|H_w] = \alpha_{\mathrm{PSC}}\sigma_w^2 \approx \sigma_w^2$). The anticorrelation attack has $\gamma_0 = 1/(1-\alpha_{\mathrm{PSC}})$, which results in $D_{\hat{y}x} \geq \sigma_x^2$; the distortion of the attacked signal is at least as large as the power of the original signal $\mathbf{x}[\vec{n}]$.

In contrast, watermarks whose power spectra are not closely matched to $\Phi_{xx}(\vec{\omega})$ are highly vulnerable to the Wiener attack.[1] For some watermarks, the anticorrelation attack can even yield $D_{\hat{y}x} \leq D_{yx}$; not only is the watermark not detected in $\hat{y}[\vec{n}]$, but $\hat{y}[\vec{n}]$ also has better quality than the unattacked, watermarked document $y[\vec{n}]$.

3. DETECTION OF PSC-COMPLIANT WATERMARKS

We now consider detection of PSC-compliant watermarks after the Wiener attack. We will study the performance when the Wiener attack parameters are selected to make detection as difficult as possible.

3.1. Optimal Detection of PSC-Compliant Watermarks

Let us assume that the additive noise $\mathbf{v}[\vec{n}]$ is also PSC-compliant, namely, $\Phi_{vv}(\vec{\omega}) = \left(\sigma_v^2/\sigma_x^2\right)\,\Phi_{xx}(\vec{\omega})$. Why might the attacker choose PSC-compliant noise? Given the received signal $\hat{\mathbf{y}}[\vec{n}]$, the receiver could perform Wiener denoising to suppress $\mathbf{v}[\vec{n}]$ before detection. The attacker turns the PSC against the owner and makes $\mathbf{v}[\vec{n}]$ as difficult as possible to remove. Since Wiener denoising to suppress PSC-compliant noise is almost completely ineffective, we do not consider it further.

3.1.1. Pre-Whitened Correlation Detection

During reception, the two potential noise sources are $\mathbf{x}[\vec{n}]$ and $\mathbf{v}[\vec{n}]$. Together they act like a single noise source whose power spectrum is directly proportional to $\Phi_{xx}(\vec{\omega})$. Then we have the conventional scenario of signal detection in the presence of colored Gaussian noise. One structure for optimal detection[7] applies a *whitening filter* to the received signal $\hat{y}[\vec{n}]$ and the watermark $w[\vec{n}]$ prior to correlation detection. Fig. 1 presents a block diagram.

Let $h'[\vec{n}]$ and $H'(\vec{\omega})$ denote the impulse response and transfer function, respectively, of the whitening filter. Then $H'(\vec{\omega}) = \left[\sigma_x^2/\Phi_{xx}(\vec{\omega})\right]^{1/2}$. (To avoid division by zero, define "$\sigma_x^2/0$" $= 0$ in this equation.) The output of a signal processed by this filter is indicated by a prime. Hence, $\mathbf{x}'[\vec{n}] = h'[\vec{n}] * \mathbf{x}[\vec{n}]$, and $\Phi_{x'x'}(\vec{\omega}) = |H'(\vec{\omega})|^2\Phi_{xx}(\vec{\omega}) = \sigma_x^2$, so that $\mathbf{x}'[\vec{n}]$ is white noise with variance σ_x^2. Similarly, $\mathbf{v}'[\vec{n}] = h'[\vec{n}] * \mathbf{v}[\vec{n}]$ is white noise with variance σ_v^2. The whitened watermark realization $w'[\vec{n}] = h'[\vec{n}] * w[\vec{n}]$ is non-random. However, we assume that ergodicity allows us to set its time/space averages equal to the ensemble averages of $\mathbf{w}'[\vec{n}] = h'[\vec{n}] * \mathbf{w}[\vec{n}]$. The whitened correlation statistic is

$$c' = \frac{1}{N}\sum_{\vec{n}\in\mathcal{N}} \left(h'[\vec{n}] * (\hat{y}[\vec{n}] - a'x[\vec{n}])\right)\left(h'[\vec{n}] * w[\vec{n}]\right) = \frac{1}{N}\sum_{\vec{n}\in\mathcal{N}} \left(\hat{y}'[\vec{n}] - a'x'[\vec{n}]\right) w'[\vec{n}]. \tag{4}$$

The receiver shown in Fig. 1 contains a scalar gain factor $a' \geq 0$, and subtracts $a'x[\vec{n}]$ from the received signal. The value of a' determines the reduction, if any, in interference from the original $x[\vec{n}]$ and depends upon the

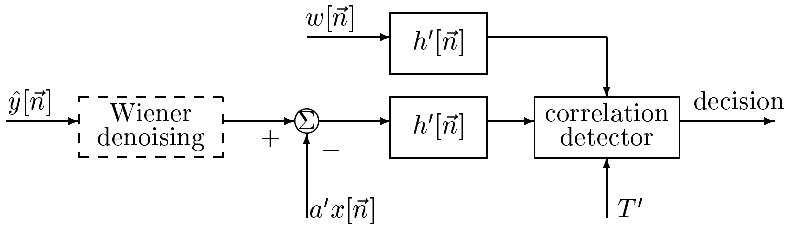

Figure 1. Block diagram of optimal detector for PSC-compliant watermarks. The impulse response of the whitening filter is $h'[\vec{n}]$. The gain factor a' depends on the detection scenario: $a' = 0$ for blind reception, $a' = 1 - \gamma\,(1 - \alpha_{\mathrm{PSC}})$ for reception-with-original.

detection scenario. On the one hand, in the case of *reception-with-original*, the receiver has access to $x[\vec{n}]$ and can set $a' = (1 - \gamma\,(1 - \alpha_{\mathrm{PSC}}))$ to eliminate the interference from $x[\vec{n}]$ that remains in $\hat{y}[\vec{n}]$. This case represents the best possible interference rejection. On the other hand, in conventional *blind detection*, the receiver has no access to the original, so we set $a' = 0$ for this case.

3.1.2. Detector Statistics

To study the expected performance of the detector, we return to the random-process interpretation of $\mathbf{x}[\vec{n}]$ and $\mathbf{v}[\vec{n}]$. Doing so allows us to treat c' as a Gaussian RV $\mathbf{c}'$, and we need the conditional means and variances of $\mathbf{c}'$ under the hypotheses H_w and H_o. The analysis is straightforward; we find

$$\mathrm{E}\left[\mathbf{c}'|H_o\right] = 0, \tag{5}$$

$$\mathrm{E}\left[\mathbf{c}'|H_w\right] = \left(1 - \gamma\,(1 - \alpha_{\mathrm{PSC}})\right)\sigma_w^2 \tag{6}$$

$$\mathrm{var}\left(\mathbf{c}'|H_i\right) = \mathrm{var}\left(\mathbf{c}'\right) = \frac{\sigma_w^2}{N}\left(\left[\left(1 - \gamma\,(1 - \alpha_{\mathrm{PSC}})\right) - a'\right]^2 \sigma_x^2 + \sigma_v^2\right), \qquad i = o, w \tag{7}$$

$$= \frac{\sigma_w^2}{N}\left(\left[(1 - a)\left(1 - \gamma\,(1 - \alpha_{\mathrm{PSC}})\right)\right]^2 \sigma_x^2 + \sigma_v^2\right), \qquad i = o, w, \tag{8}$$

where $a = a' / \left(1 - \gamma\,(1 - \alpha_{\mathrm{PSC}})\right)$, so that a' can be replaced by $a \in \{0, 1\}$, with $a = 1$ corresponding to reception-with-original and $a = 0$ to blind reception.

Besides a, there are only five parameters (σ_x^2, γ, σ_v^2, σ_w^2, N) that affect the mean and variance of $\mathbf{c}'$. The variance σ_x^2 is fixed because the original document cannot be changed. The attacker has control of two parameters: γ, which can vary between 0 and $\gamma_0 = 1/(1 - \alpha_{\mathrm{PSC}})$; and σ_v^2, which can be varied as well but must remain non-negative. The owner has two parameters: σ_w^2, which is effectively fixed because it should be as large as possible while keeping the watermark imperceptible or unobtrusive; and N, the support size, which may or may not be restricted. For example, in image watermarking, the image has a fixed number of pixels, but for audio or video watermarking, there may be so many samples available that the owner may be able to make N as large as desired. Hence, the owner has either one free parameter (N) or none at all.

One important aspect is the distortion $D_{\hat{y}x}$ of the attacked signal (Eq. (3)). The attacked-document distortion is an aspect of watermarking that differs from traditional communications problems. Although (3) has been previously reported,[6,1] it can now be combined with $\mathrm{E}\left[\mathbf{c}'|\,H_i\right]$ and $\mathrm{var}\left(\mathbf{c}'|H_i\right)$ to describe the relationship between distortion and detector performance. For example, we can examine the relationship between $D_{\hat{y}x}$ and the probability of detection error P_E.

3.2. Worst-Case Wiener Attack on PSC-Compliant Watermarks

For a given embedding ratio σ_w^2/σ_x^2 and support size N, we would like to know the worst possible detection performance that the Wiener attack can cause.

3.2.1. Attacker's Problem and Solution

The attacker wishes to find the values γ and σ_v^2 that hinder detection as much as possible while keeping $D_{\hat{y}x}$ less than or equal to some target distortion D_t, $D_{\hat{y}x,\min} \leq D_t \leq \sigma_x^2$. The lower limit $D_{\hat{y}x,\min} = \alpha_{\mathrm{PSC}}\sigma_w^2 = (1 - \alpha_{\mathrm{PSC}})\,\sigma_x^2$ corresponds to the removal attack and gives the smallest distortion that the Wiener attack can achieve.

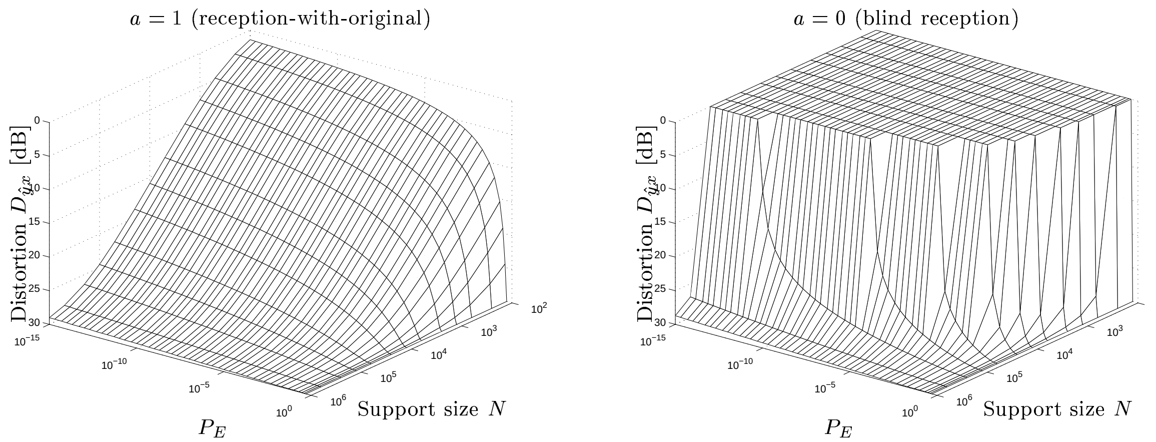

Figure 2. Support size N, probability of error P_E, and attacked-document distortion $D_{\hat{y}x}$. The embedding ratio is $\sigma_w^2/\sigma_x^2 = -30$ dB.

The solution appears in App. A and is the same for either case $j \in \{\text{orig}, \text{blind}\}$:

$$\gamma = \frac{D_t}{D_{\hat{y}x,\min}}, \quad \text{and} \quad \sigma_v^2 = -\frac{\sigma_x^2 + \sigma_w^2}{\sigma_x^4} D_t^2 + \frac{2\sigma_w^2 + \sigma_x^2}{\sigma_x^2} D_t - \sigma_w^2. \tag{9}$$

An important property is that the solution (9) is *unique*. Hence, we can truly talk about *the* worst-case detector performance.

We can use these results to find the average detector performance. For convenience, define $\mu = \mathrm{E}\left[\mathbf{c}'|H_w\right]$, and $\zeta_j = \sqrt{N \mathrm{var}\left(\mathbf{c}'\right)/\sigma_w^2}$, with γ and σ_v^2 given by (9). Then $\mu = \left(1 - D_t/\sigma_x^2\right)\sigma_w^2$, and

$$\zeta_{\text{orig}} = \sqrt{-\frac{\sigma_x^2 + \sigma_w^2}{\sigma_x^4} D_t^2 + \frac{2\sigma_w^2 + \sigma_x^2}{\sigma_x^2} D_t - \sigma_w^2}, \quad \text{and} \quad \zeta_{\text{blind}} = \sqrt{-\frac{\sigma_w^2}{\sigma_x^4} D_t^2 + \frac{2\sigma_w^2 - \sigma_x^2}{\sigma_x^2} D_t + \sigma_x^2 - \sigma_w^2}. \tag{10}$$

For a given decision threshold T', the detection error probabilities α (miss) and β (false-alarm) are

$$\alpha = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\mathrm{E}\left[\mathbf{c}'|H_w\right] - T'}{\sqrt{2}\,\mathrm{std}\left(\mathbf{c}'\right)}\right)\right] = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{N}{2\sigma_w^2}}\frac{\mu - T'}{\zeta_j}\right)\right] \tag{11}$$

$$\beta = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{T'}{\sqrt{2}\,\mathrm{std}\left(\mathbf{c}'\right)}\right)\right] \quad = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{N}{2\sigma_w^2}}\frac{T'}{\zeta_j}\right)\right]. \tag{12}$$

where $\mathrm{erf}(\cdot)$ is the error function. Assuming equal a priori probabilities for hypotheses H_w and H_o, we find the minimum probability of error to be $P_E = (\alpha + \beta)/2 = (1/2)\left[1 - \mathrm{erf}\left(\sqrt{N/2\sigma_w^2}(\mu/\zeta_j)\right)\right]$. With some basic algebraic manipulations and the inverse error function, we can solve these expressions to find the minimum support size N required to achieve a desired P_E. Or, we can use the Neyman-Pearson strategy[7,8]; for example, fix $\beta = \beta_0$ and solve (12) for T', substitute the result into (11), and solve for N in terms of α and β_0.

3.2.2. Performance of PSC-Compliant Watermarks

Fig. 2 shows the relationship between $D_{\hat{y}x}$, P_E, and N with the embedding ratio fixed at $\sigma_w^2/\sigma_x^2 = -30$ dB. This low embedding ratio reflects the fact that the watermark should be imperceptible. Not surprisingly, decreasing N or increasing $D_{\hat{y}x}$ results in an increase in P_E. The important result is that this relationship can now be computed. Hence, if a watermarking application requires a maximum value of P_E at a certain distortion $D_{\hat{y}x}$, the minimum support size N can be determined.

Fig. 3 overlays the results for blind reception and reception-with-original in a single graph. The effect of interference from the original becomes clear when we compare the performance curves for $a = 0$ and $a = 1$

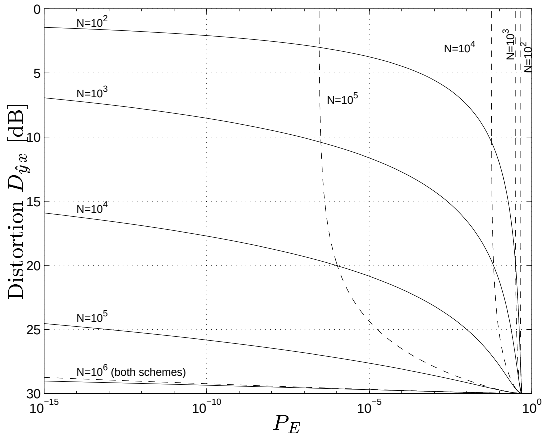

Figure 3. Probability of error P_E versus attacked-document distortion $D_{\hat{y}x}$. Solid line: $a = 1$; dashed line: $a = 0$. The embedding ratio is $\sigma_w^2/\sigma_x^2 = -30$ dB.

with N fixed. For blind reception, the interference from the original dominates P_E, so that the Wiener attack has very little effect over a large distortion range. We can also see how much N must be increased in order for blind reception to achieve the same performance point $(P_E, D_{\hat{y}x})$ as reception-with-original. For example, at $(P_E, D_{\hat{y}x}) = (10^{-6}, 20 \text{ dB})$, $N_{\text{orig}} = 10^4$, while $N_{\text{blind}} = 10^5$.

4. PROBLEM REVISITED

Let us revisit the basic watermarking problem in Sec. 1, in which the attacker tries to impair communication under a constraint on $D_{\hat{y}x}$, and the owner attempts to retrieve the watermark information from $\hat{\mathbf{y}}[\vec{n}]$ under a constraint on D_{yx}. Sec. 2.3 introduced a scenario in which the communication system is specified; in particular, the receiver used correlation detection. Then the attack (Wiener attack) and counterattack (PSC) were alternately optimized.

However, correlation detection is optimal only in the case of *white* Gaussian noise, but $\mathbf{x}[\vec{n}]$ and $\mathbf{v}[\vec{n}]$ may be colored. Hence, by restricting the form of the detector, some generality has been lost. If the detector is optimal—and therefore not restricted to correlation detector without pre-processing—the Wiener attack may be suboptimal; consequently, the PSC may be a suboptimal condition for resisting the yet-to-be-derived optimum attack.

We now investigate the case when the receiver is optimum. Consider a general LSI-filtering attack of the form

$$\hat{\mathbf{y}}[\vec{n}] = g[\vec{n}] * \mathbf{y}[\vec{n}] + \mathbf{v}[\vec{n}] = g[\vec{n}] * (\mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}]) + \mathbf{v}[\vec{n}], \tag{13}$$

where $g[\vec{n}]$ is the impulse response of the attack filter and $\mathbf{v}[\vec{n}]$ is again colored Gaussian noise. For the memoryless case, it has been shown[9] that the optimum attack has the form (13).

The distortion associated with this attack is

$$D_{\hat{y}x} = \frac{1}{(2\pi)^M} \int_\Omega \left[|G(\vec{\omega}) - 1|^2 \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega}) + \Phi_{vv}(\vec{\omega}) \right] d\vec{\omega}. \tag{14}$$

Since the receiver is assumed to be optimum, we can work directly with the channel capacity, which is

$$C = \frac{1}{(2\pi)^M} \int_\Omega \frac{1}{2} \log_2 \left(1 + \frac{|G(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega})}{(1-a)^2 |G(\vec{\omega})|^2 \Phi_{xx}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})} \right) d\vec{\omega}, \tag{15}$$

where the factor $a \in \{0, 1\}$ again reflects the cases of blind reception and reception-with-original. We may interpret (15) in the following way. Define the *effective inverse-filtered noise* by $\mathbf{n}'[\vec{n}] = (1-a)\mathbf{x}[\vec{n}] + g^{-1}[\vec{n}] * \mathbf{v}[\vec{n}]$, and define $\Phi_{n'n'}(\vec{\omega}) = (1-a)^2 \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^{-2} \Phi_{vv}(\vec{\omega})$, if $G(\vec{\omega}) \neq 0$, and $\Phi_{n'n'}(\vec{\omega}) = \infty$, if $G(\vec{\omega}) = 0$. $\Phi_{n'n'}(\vec{\omega})$ is well-defined for all frequencies, even if $G(\vec{\omega}) = 0$ at some frequencies. Then (15) can be written as

$$C = \frac{1}{(2\pi)^M} \int_\Omega \frac{1}{2} \log_2 \left(1 + \frac{\Phi_{ww}(\vec{\omega})}{\Phi_{n'n'}(\vec{\omega})} \right) d\vec{\omega}. \tag{16}$$

In the memoryless Gaussian case, the theoretical capacity of an optimal blind receiver is equal to that of a receiver with the original.[10,9] However, the enormous increase in complexity may make such systems impractical.[9] Also, this capacity result has not been extended to colored processes, and most current blind watermarking schemes are equivalent to $a = 0$. Hence, we continue to use $a \in \{0, 1\}$. Doing so allows us to examine the performance gap between current blind schemes and reception-with-original, where the latter may prove to be the theoretical capacity of an optimal blind receiver.

Strictly speaking, if $a \neq 1$, then Eq. (15) gives the maximum *achievable rate* of a suboptimal blind watermarking system, not the capacity, which is the supremum of achievable rates for all possible blind watermarking systems. Nevertheless, as it is common to use the term "capacity" when discussing the best performance of a given (suboptimal) communications system, we will also say "capacity" when $a \neq 1$.

5. EFFECTIVE WHITE-NOISE ATTACK AND THE PSC

We now show that the Wiener attack corresponds to an attack that behaves like white noise, and the PSC is the best way to resist it.

5.1. Effective White-Noise Attack

Let us suppose the attacker chooses $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to make the capacity expression (16) resemble the *white* Gaussian noise (WGN) channel. This strategy is intuitively appealing since, in traditional communications problems, communication is most difficult through the WGN channel.[11]

The attacker should set $\Phi_{n'n'}(\vec{\omega}) = \left(\sigma_{n'}^2/\sigma_w^2\right)\Phi_{ww}(\vec{\omega})$. Then $C = \frac{1}{2}\log_2\left(1 + \sigma_w^2/\sigma_{n'}^2\right)$. Since σ_w^2 is fixed, the attacker can ensure that C equals some target capacity C_t by selecting $\sigma_{n'}^2$ appropriately. It follows that

$$\Phi_{vv}(\vec{\omega}) = |G(\vec{\omega})|^2\left[\frac{\sigma_{n'}^2}{\sigma_w^2}\Phi_{ww}(\vec{\omega}) - (1-a)^2\Phi_{xx}(\vec{\omega})\right], \tag{17}$$

so that once $G(\vec{\omega})$ is known, $\Phi_{vv}(\vec{\omega})$ is also specified. Then the attacker must find $G(\vec{\omega})$ to minimize $D_{\hat{y}x}$.

The solution is given in App. B; we find

$$G(\vec{\omega}) = \frac{\Phi_{xx}(\vec{\omega})}{a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})}, \tag{18}$$

where $K = 1+\sigma_{n'}^2/\sigma_w^2$. At frequencies where both $\Phi_{xx}(\vec{\omega})$ and $\Phi_{ww}(\vec{\omega})$ are zero, the value of $G(\vec{\omega})$ is unimportant; let $G(\vec{\omega}) = 0$ at such frequencies. At frequencies where $\Phi_{xx}(\vec{\omega})$ is small, strong attenuation occurs, while at frequencies where $\Phi_{xx}(\vec{\omega})$ is large, the attenuation is weaker to avoid excessive distortion. $G(\vec{\omega})$ closely resembles a Wiener filter, but in general it does not have the form $1 - \gamma H(\vec{\omega})$, which would make it equivalent to the Wiener attack.

The attacked-document distortion becomes

$$D_{\hat{y}x} = \sigma_x^2 - \frac{1}{(2\pi)^M}\int_\Omega G(\vec{\omega})\Phi_{xx}(\vec{\omega})\,d\vec{\omega} = \sigma_x^2 - \frac{1}{(2\pi)^M}\int_\Omega \frac{\Phi_{xx}^2(\vec{\omega})}{a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})}\,d\vec{\omega}. \tag{19}$$

If the attacker simply minimizes $D_{\hat{y}x}$, C is maximized. If the attacker makes $C = 0$, then $\sigma_{n'}^2 \to \infty$, and $D_{\hat{y}x} = \sigma_x^2$, so that the attacked document is not likely to be useful. Thus, even if the owner selects $\Phi_{ww}(\vec{\omega})$ poorly, some information can still be transmitted as long as $D_{\hat{y}x} < \sigma_x^2$.

5.2. Counterattack: Power-Spectrum Condition

As a counterattack, the owner should choose $\Phi_{ww}(\vec{\omega})$ to maximize $D_{\hat{y}x}$ under the constraints $D_{yx} = \sigma_w^2$ and $C = C_t$. The latter constraint is satisfied by fixing $\sigma_{n'}^2$ appropriately. The details of the solution are described in App. C; the solution is exactly the PSC (2). Hence, the PSC is the best counterattack against the effective white-noise attack.

When the PSC is fulfilled, Eqs. (17) and (18) give

$$\Phi_{vv}(\vec{\omega}) = \frac{\left(\sigma_{n'}^2 - (1-a)^2\sigma_x^2\right)\sigma_x^2}{\left(a(2-a)\sigma_x^2 + \sigma_w^2 + \sigma_{n'}^2\right)^2}\,\Phi_{xx}(\vec{\omega}) = \frac{\sigma_v^2}{\sigma_x^2}\,\Phi_{xx}(\vec{\omega}), \quad \text{and} \quad G(\vec{\omega}) = \frac{\sigma_x^2}{a(2-a)\sigma_x^2 + K\sigma_w^2} = G_{\mathrm{PSC}}, \tag{20}$$

where G_{PSC} is constant for all $\vec{\omega}$. The corresponding attacked-document distortion is $D_{\hat{y}x} = (1 - G_{\mathrm{PSC}})\,\sigma_x^2$. To ensure $\Phi_{vv}(\vec{\omega}) \geq 0$, $\forall\vec{\omega}$, $\sigma_{n'}^2 \geq (1-a)^2\sigma_x^2$. When equality is fulfilled, the minimum possible attacked-document distortion results; namely, $D_{\hat{y}x,\min} = \alpha_{\mathrm{PSC}}\sigma_w^2$.

5.3. Capacity and Distortion

Since C is a function of $\sigma_w^2/\sigma_{n'}^2$, we can use (20) to derive an expression for C directly in terms of $D_{\hat{y}x}$. We find

$$C = \frac{1}{2}\log_2\left(1 + \frac{\left(\sigma_x^2 - D_{\hat{y}x}\right)\sigma_w^2}{\sigma_x^4 - \left(\sigma_x^2 - D_{\hat{y}x}\right)\left(a(2-a)\sigma_x^2 + \sigma_w^2\right)}\right). \tag{21}$$

Eq. (21) describes the best possible watermark performance in the presence of the effective white-noise attack. For any $D_{\hat{y}x,\min} \leq D_{\hat{y}x} \leq \sigma_x^2$, C is expressed entirely in terms of a (blind reception or reception-with-original), the respective powers σ_x^2 and σ_w^2 of the original and watermark, and the attacked-document distortion $D_{\hat{y}x}$. For $D_{\hat{y}x} > \sigma_x^2$, $C = 0$ as well, but once $C = 0$, the attacker does not need to increase $D_{\hat{y}x}$ further.

5.4. Relation to the Wiener Attack and Memoryless Case

From Sec. 2.3, the Wiener filter becomes constant when the PSC is fulfilled: $H(\vec{\omega}) = 1 - \alpha_{\mathrm{PSC}}$. In Sec. 3.2.1, we assumed the watermark $\mathbf{w}[\vec{n}]$ and noise $\mathbf{v}[\vec{n}]$ were PSC-compliant and determined the worst-case values of γ and σ_v^2 in (9) when $D_{\hat{y}x} = D_t$. In (20), $G(\vec{\omega})$ is constant and $\Phi_{vv}(\vec{\omega})$ is PSC-compliant.

This suggests that γ and σ_v^2 in (9) are related to $G(\vec{\omega})$ and $\sigma_{n'}^2$ in (20). Some algebra shows that $G(\vec{\omega}) = G_{\mathrm{PSC}} = 1 - D_t/\sigma_x^2 = 1 - \gamma\left(1 - \alpha_{\mathrm{PSC}}\right) = 1 - \gamma H(\vec{\omega})$, and $\sigma_{n'}^2 = \sigma_x^4/\left(\sigma_x^2 - D_t\right) - \left(a(2-a)\sigma_x^2 + \sigma_w^2\right)$. Hence, *the worst-case Wiener attack against a PSC-compliant watermark is equivalent to the effective white-noise attack.*

Since all power spectra have the same shape, they are effectively white relative to one another. Gaussianity implies that the random processes $\mathbf{x}[\vec{n}]$, $\mathbf{w}[\vec{n}]$, and $\mathbf{v}[\vec{n}]$ can be converted into memoryless random processes with the same respective variances σ_x^2, σ_w^2, and σ_v^2. The memoryless case has recently been addressed[10,9]; it can be shown that, for the memoryless case, the Wiener attack is the optimum attack among all possible attacks, and Eq. (21) describes the fundamental relationship between C and $D_{\hat{y}x}$. (The capacity-distortion expression derived in[9] measures distortion relative to the watermarked signal, i.e., $D_{\hat{y}y}$.) The results of Sec. 3.2 therefore also describe the theoretical performance limits for the memoryless case.

6. OPTIMUM ATTACK AND COUNTERATTACK

In a conventional additive Gaussian noise channel, $\Phi_{n'n'}(\vec{\omega})$ remains fixed and $\Phi_{ww}(\vec{\omega})$ is selected to maximize the mutual information $I(\mathbf{w}[\vec{n}]; \hat{\mathbf{y}}[\vec{n}])$. The solution for $\Phi_{ww}(\vec{\omega})$ is a water-filling rule,[11] which gives $\mathbf{w}[\vec{n}]$ a power advantage over the noise $\mathbf{n}'[\vec{n}]$. It can also be shown that the mutual information is minimized when $\Phi_{n'n'}(\vec{\omega})$ is white, which motivated the attack of Sec. 5.1. However, the hostile nature of potential attacks on watermarks means that $\Phi_{n'n'}(\vec{\omega})$ can adapt to $\Phi_{ww}(\vec{\omega})$. The attacker, rather than the owner, has a power advantage. The attacker has "the last word" on the behavior of the channel. Consequently, the attack of Sec. 5.1 is suboptimal because, by restricting the form of $\Phi_{n'n'}(\vec{\omega})$, the attack does not fully exploit the possible power advantage.

The details of the optimum attack and counterattack are the subject of a forthcoming paper.[12] We review the applicable results here.

6.1. Optimum Attack

The attacker's problem is to find $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to minimize $D_{\hat{y}x}$ for some target capacity $C = C_t$. There are no constraints on $\Phi_{n'n'}(\vec{\omega})$ (except that it be a valid power spectrum), so this attack fully exploits the available power advantage. The problem can be solved by the calculus of variations.[12] The optimum attack filter and noise power spectrum are given by

$$G(\vec{\omega}) = A_j(\vec{\omega})\,\frac{\Phi_{xx}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}, \quad \text{and} \quad \Phi_{vv}(\vec{\omega}) = \left(1 - A_j(\vec{\omega})\right)A_j(\vec{\omega})\,\frac{\Phi_{xx}^2(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}, \quad j \in \{\mathrm{blind}, \mathrm{orig}\}, \tag{22}$$

where $0 \leq A_j(\vec{\omega}) \leq 1$, $\forall\vec{\omega}$.

The weighting term $A_j(\vec{\omega})$ is parametrized by a Lagrange multiplier λ. Define $\mathrm{cl}[x] = x$, if $0 \leq x \leq 1$; $\mathrm{cl}[x] = 0$, if $x < 0$; and $\mathrm{cl}[x] = 1$, if $x > 1$. For $a = 0$ (blind reception),

$$A_{\mathrm{blind}}(\vec{\omega}) = \mathrm{cl}\left[1 + \frac{\Phi_{xx}(\vec{\omega})}{\Phi_{ww}(\vec{\omega})} - \frac{\lambda}{2\ln 2}\,\frac{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}{\Phi_{xx}^2(\vec{\omega})}\right]. \tag{23}$$

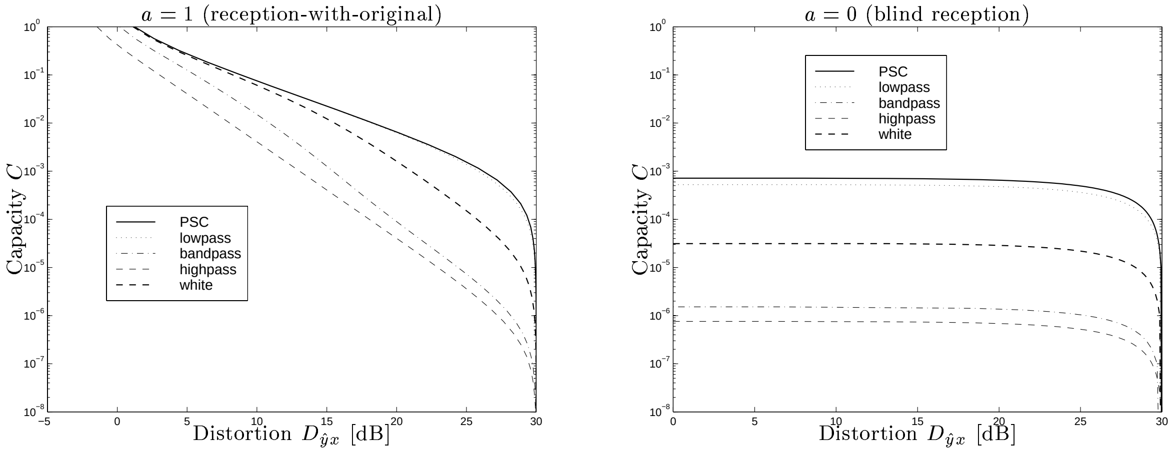

Figure 4. Performance of various watermark power spectra subject to the suboptimal, effective white-noise attack. The embedding ratio is $\sigma_w^2/\sigma_x^2 = -30$ dB.

The Lagrange multiplier can range from $\lambda_{\min}$ to $\lambda_{\max}$, where $\lambda_{\min} = 2 \ln 2 \min_{\vec{\omega}} \Phi_{xx}^3(\vec{\omega})/\Phi_{ww}(\vec{\omega})\left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)$ to $\lambda_{\max} = 2 \ln 2 \max_{\vec{\omega}} \Phi_{xx}^2(\vec{\omega})/\Phi_{ww}(\vec{\omega})$. For $a = 1$ (reception-with-original), $\lambda_{\min} = 0$, and $\lambda_{\max}$ is the same as for blind reception. The weighting term is

$$A_{\text{orig}}(\vec{\omega}) = \text{cl}\left[1 + \frac{\Phi_{ww}(\vec{\omega})}{2\Phi_{xx}(\vec{\omega})} - \frac{\sqrt{\Phi_{xx}^2(\vec{\omega})\Phi_{ww}^2(\vec{\omega}) + (2\lambda/\ln 2)\Phi_{xx}(\vec{\omega})\Phi_{ww}(\vec{\omega})\left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)}}{2\Phi_{xx}^2(\vec{\omega})}\right]. \tag{24}$$

From (14) and (15), the attacked-document distortion and capacity are, respectively,

$$D_{\hat{y}x} = \sigma_x^2 - \frac{1}{(2\pi)^M}\int_\Omega A_j(\vec{\omega})\frac{\Phi_{xx}^2(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}\, d\vec{\omega}, \tag{25}$$

$$C = \frac{1}{(2\pi)^M}\int_\Omega \frac{1}{2}\log_2\left(1 + \frac{A_j(\vec{\omega})\Phi_{ww}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega}) - A_j(\vec{\omega})\left(a(2-a)\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)}\right) d\vec{\omega}. \tag{26}$$

6.2. Optimized Counterattack

The problem of finding the optimum counterattack is quite difficult. We seek $\Phi_{ww}(\vec{\omega})$ to maximize $D_{\hat{y}x}$ for a given capacity $C = C_t$. However, the relationship between $\Phi_{ww}(\vec{\omega})$, C, and $D_{\hat{y}x}$ is very complicated due to the dependence on $A_j(\vec{\omega})$. An analytical solution may be impossible to find.

Under the assumption that all power spectra are piecewise constant, numerical methods can be applied[12] to optimize $\Phi_{ww}(\vec{\omega})$. We have implemented iterative algorithms based on marginal analysis and simulated annealing. At each iteration, they vary $\Phi_{ww}(\vec{\omega})$ slightly, search for λ to find the optimum attack such that $|C - C_t|$ is within some tolerance, and then compute the resulting distortion $D_{\hat{y}x}$. The process repeats until variations in $\Phi_{ww}(\vec{\omega})$ no longer produce increases in $D_{\hat{y}x}$.

6.3. Experimental Results

We modeled the original and example watermark power spectra as 1-D pth-order autoregressive (AR(p)) Gaussian processes, which have the form $\mathbf{x}[n] = \sum_{k=1}^{p} a_k \mathbf{x}[n-k] + \mathbf{u}[n]$, where $\mathbf{u}[n]$ is WGN. Such signal models are often used to analyze naturally correlated data such as audio, images, and video. The original was AR(1) with $a_1 = 0.95$ and $\sigma_x^2 = 10^3$. All watermarks used $\sigma_w^2 = 1$; the different models were: PSC ($a_1 = 0.95$), lowpass ($a_1 = 0.90$), bandpass ($a_1 = 0$, $a_2 = -0.9025$), highpass ($a_1 = -0.95$), and white.

Fig. 4 shows the theoretical performance (capacity versus distortion) of these watermarks after the effective white-noise attack of Sec. 5.1. The graphs show that, in agreement with Sec. 5.2, the PSC-compliant watermark

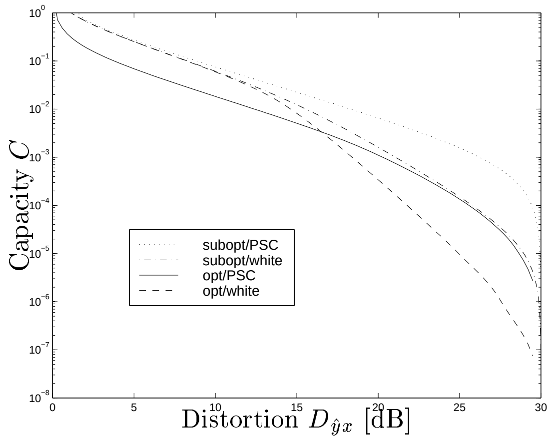

Figure 5. Performance comparison of PSC-compliant and white watermarks subject either to the suboptimal, effective white-noise attack or to the optimal attack. The embedding ratio is $\sigma_w^2/\sigma_x^2 = -30$ dB, and $a = 1$ (reception-with-original).

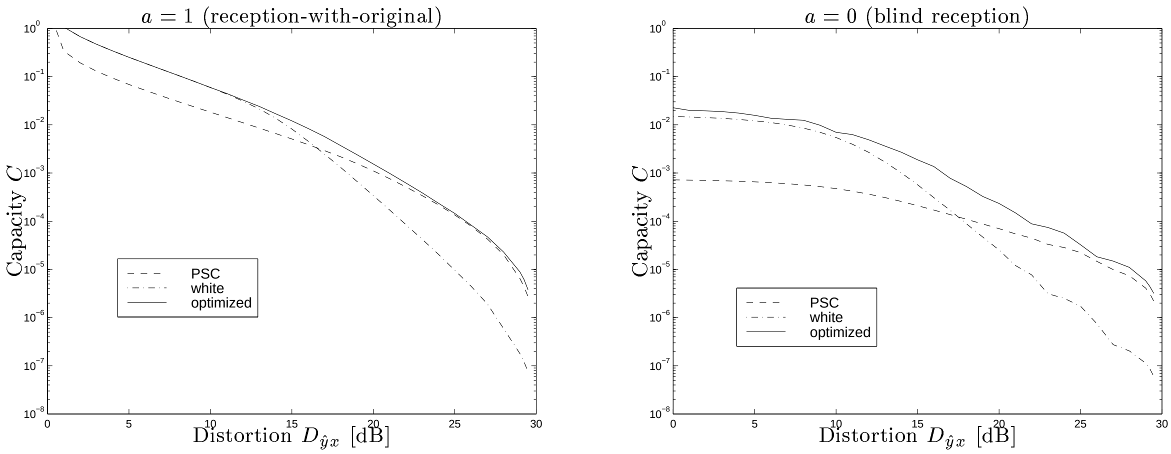

Figure 6. Performance comparison of white, PSC-compliant watermarks, and optimized watermarks subject to the optimal attack. The embedding ratio is $\sigma_w^2/\sigma_x^2 = -30$ dB.

has the best performance over all $D_{\hat{y}x}$. For blind reception, the interference from the original dominates, which explains why the curves are nearly constant over a large distortion range.

Next, Fig. 5 compares the performance of white and PSC-compliant watermarks subject to either the effective white-noise attack of Sec. 5.1 or the optimum attack of Sec. 6.1. For brevity, only the case $a = 1$ is shown; the qualitative behavior for blind reception is the same. The attack clearly improves when it fully exploits the potential power advantage. At low distortions, the white watermark outperforms the PSC-compliant one, and at high distortions, the PSC-compliant watermark performs better. In particular, the PSC is clearly not the optimum counterattack against the optimum attack.

Finally, Fig. 6 displays the performance curves for the optimum attack and the optimized watermark power spectra produced by the algorithms described in Sec. 6.2. Curves are drawn for both reception-with-original and blind reception. Consider reception-with-original: at low distortions, a white watermark performs almost as well as the optimized ones, while at high distortions, a PSC-compliant one has nearly optimal performance. In the case of blind reception, a similar qualitative statement can be made, but there remains a performance gap between the white or PSC-compliant watermark and the optimized ones. Neither power-spectral shape is the best because of the additional interference from $\Phi_{xx}(\vec{\omega})$.

7. CONCLUSIONS

The Wiener attack and PSC (2) were originally derived[6] in the context of MMSE watermark estimation and subsequent resistance against estimation. They were later extended[1] to watermark correlation detection, which is used in many current watermarking schemes. We have determined the performance limits for the Wiener attack against PSC-compliant watermarks. Sec. 5.4 showed that these are also the performance limits for the optimum attack in the memoryless case.

However, correlation detection is optimal only in the case of *white* Gaussian noise. If the correlation detector is replaced by an optimal receiver, the Wiener attack is suboptimal, and hence the PSC is a suboptimal counterattack. We showed that the Wiener attack on PSC-compliant watermarks coincides with a suboptimal, effective white-noise attack, against which the PSC is the best counterattack.

Finally, we presented the optimum attack against the optimum receiver. A closed-form expression for the optimum watermark power spectrum has not been discovered, but numerical methods have been applied to optimize it to resist the attack. Experimental results demonstrated that PSC-compliant watermarks are indeed suboptimal counterattacks. Based upon the experiments, we observed that white watermarks perform well when the attacked-document distortion is small, and PSC-compliant watermarks perform well when the distortion is large. An in-depth investigation of the optimum attack and optimized counterattack is currently in progress.[12]

ACKNOWLEDGMENT

The authors would like to thank Robert Schober for his assistance, which led to the results of Sec. 5.

APPENDIX A. WORST-CASE DETECTION OF PSC-COMPLIANT WATERMARKS

Detection becomes less reliable as the ratio $\mathrm{E}\left[\mathbf{c}'|H_w\right]/\mathrm{std}\left(\mathbf{c}'\right)$ decreases. It is convenient to use the square of this ratio, $\mathrm{E}^2\left[\mathbf{c}'|H_w\right]/\mathrm{var}\left(\mathbf{c}'\right) = N\sigma_w^2 R_j'$, $j \in \{\mathrm{orig}, \mathrm{blind}\}$, where $R_{\mathrm{orig}}' = \left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)^2/\sigma_v^2$, and $R_{\mathrm{blind}}' = \left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)^2/\left(\left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)^2\sigma_x^2+\sigma_v^2\right)$. Discard the factors N and σ_w^2 and concentrate on minimizing R_j'. Now the problem is to select γ and σ_v^2 to minimize R_j' such that $D_{\hat{y}x} \le D_t$, $0 \le \gamma \le \gamma_0$, and $\sigma_v^2 \ge 0$.

If a solution exists, it must occur when $D_{\hat{y}x} = D_t$. If there were a solution $\left(\gamma, \sigma_v^2\right)$ that did not satisfy equality, then there exists $\sigma_{v'}^2 > \sigma_v^2$ such that $R_j'\left(\gamma, \sigma_{v'}^2\right) < R_j'(\gamma, \sigma_v^2)$ and $D_{\hat{y}x}\left(\gamma, \sigma_{v'}^2\right) \le D_t$, a contradiction.

Next we seek the range of γ values that correspond to $\sigma_v^2 \ge 0$. We know $\gamma \ge 1$ because $\gamma = 1$ corresponds to the removal attack. Denote the upper limit on γ by γ_v. Set $\sigma_v^2 = 0$ in (3), which produces $\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2\gamma^2 - 2\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2\gamma - D_t + \sigma_w^2 = 0$. This equation has roots $\gamma = 1 \pm \sqrt{\left(D_t - D_{\hat{y}x,\min}\right)\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2}/\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2$. We choose the plus-case, so

$$\gamma_v = 1 + \frac{\sqrt{\left(D_t - D_{\hat{y}x,\min}\right)\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2}}{\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2} = 1 + \frac{\sigma_x^2+\sigma_w^2}{\sigma_w^2}\sqrt{\frac{D_t - D_{\hat{y}x,\min}}{\sigma_x^2+\sigma_w^2}}.$$

It is readily verified that $1 \le \gamma_v \le \gamma_0$ for $D_{\hat{y}x,\min} \le D_t \le \sigma_x^2$. Therefore, γ must lie in the interval $[1, \gamma_v]$. For any γ, the noise variance σ_v^2 that fulfills $D_{\hat{y}x} = D_t$ is given by (3), which we write as

$$\sigma_v^2 = -\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2\gamma^2 + 2\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_w^2\gamma + D_t - \sigma_w^2. \tag{27}$$

Let $a' = \varepsilon\left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)$, for some $0 \le \varepsilon \le 1$. Then the ratio R_j' can be written $R_j' = \left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)^2/\left[(1-\varepsilon)^2\left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)^2\sigma_x^2+\sigma_v^2\right]$. Substituting (27) for σ_v^2 yields $R_j' = \left(1-\gamma\left(1-\alpha_{\mathrm{PSC}}\right)\right)^2/\left(A\gamma^2+2B\gamma+C\right)$, where $A = \left(1-\alpha_{\mathrm{PSC}}\right)\left[(1-\varepsilon)^2\left(1-\alpha_{\mathrm{PSC}}\right)\sigma_x^2 - \sigma_w^2\right]$, $B = \left(1-\alpha_{\mathrm{PSC}}\right)\left[\sigma_w^2 - (1-\varepsilon)^2\sigma_x^2\right]$, and $C = (1-\varepsilon)^2\sigma_x^2 + D_t - \sigma_w^2$. Setting $\partial R_j'/\partial\gamma = 0$ and solving for γ yields $\gamma = -\left[B + \left(1-\alpha_{\mathrm{PSC}}\right)C\right]/\left[A + \left(1-\alpha_{\mathrm{PSC}}\right)B\right]$, and substituting the expressions for A, B, and C gives γ as in (9).

When we substitute γ into (27), we find σ_v^2 as in (9). It is easily confirmed that when $D_t = D_{\hat{y}x,\min}$, then $\gamma = 1$ and $\sigma_v^2 = 0$, and when $D_t = \sigma_x^2$, then $\gamma = \gamma_0$ and $\sigma_v^2 = 0$.

APPENDIX B. EFFECTIVE WHITE-NOISE ATTACK

The distortion expression (14) becomes

$$D_{\hat{y}x} = \frac{1}{(2\pi)^M} \int_\Omega \left[\left(|G(\vec{\omega}) - 1|^2 - (1-a)^2 |G(\vec{\omega})|^2 \right) \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2 K \Phi_{ww}(\vec{\omega}) \right] d\vec{\omega},$$

with $K = 1 + \sigma_{n'}^2/\sigma_w^2$. Write $G(\vec{\omega})$ in magnitude-phase form, $G(\vec{\omega}) = |G(\vec{\omega})| e^{j\theta(\vec{\omega})}$, and substitute this form into the above expression to obtain

$$D_{\hat{y}x} = \frac{1}{(2\pi)^M} \int_\Omega \left[\left(|G(\vec{\omega})|^2 - 2|G(\vec{\omega})| \cos\theta(\vec{\omega}) + 1 - (1-a)^2 |G(\vec{\omega})|^2 \right) \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2 K \Phi_{ww}(\vec{\omega}) \right] d\vec{\omega},$$

Let $D'_{\hat{y}x}$ denote the integrand of this expression.

To minimize $D_{\hat{y}x}$, compute the partial derivatives of $D'_{\hat{y}x}$ with respect to $\theta(\vec{\omega})$ and $|G(\vec{\omega})|$, and set them equal to zero. Then $\partial D'_{\hat{y}x}/\partial\theta(\vec{\omega}) = 2|G(\vec{\omega})| \sin\theta(\vec{\omega}) \Phi_{xx}(\vec{\omega}) = 0$. Thus, $\theta(\vec{\omega}) = k\pi$, and $\cos\theta(\vec{\omega}) = \pm 1$. Next, $\partial D'_{\hat{y}x}/\partial|G(\vec{\omega})| = \left[\left(2|G(\vec{\omega})| - 2\cos\theta(\vec{\omega}) - 2(1-a)^2|G(\vec{\omega})|\right) \Phi_{xx}(\vec{\omega}) + 2|G(\vec{\omega})| K \Phi_{ww}(\vec{\omega}) \right] = 0$. Solving for $|G(\vec{\omega})|$ gives $|G(\vec{\omega})| = \Phi_{xx}(\vec{\omega}) \cos\theta(\vec{\omega}) / \left(a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})\right)$. Since $|G(\vec{\omega})| \geq 0$, $\cos\theta(\vec{\omega}) = 1$; hence, $G(\vec{\omega}) = |G(\vec{\omega})|$, and choose $\theta(\vec{\omega}) = 0$, $\forall \vec{\omega}$.

APPENDIX C. POWER-SPECTRUM CONDITION

Apply the calculus of variations with the Lagrangian $J = \left((G(\vec{\omega}) - 1)^2 - (1-a)^2 G^2(\vec{\omega}) \right) \Phi_{xx}(\vec{\omega}) + G^2(\vec{\omega}) K \Phi_{ww}(\vec{\omega}) + \lambda \Phi_{ww}(\vec{\omega})$, with $G(\vec{\omega})$ given by (18). Then $dJ/d\Phi_{ww}(\vec{\omega}) = K\Phi_{xx}^2(\vec{\omega}) / \left(a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})\right)^2 + \lambda = 0$. Solving this equation for $\Phi_{ww}(\vec{\omega})$ yields $\Phi_{ww}(\vec{\omega}) = K^{-1/2} \left(\pm(-\lambda)^{-1/2} - a(2-a)K^{-1/2} \right) \Phi_{xx}(\vec{\omega})$. Hence, $\Phi_{ww}(\vec{\omega})$ is directly proportional to $\Phi_{xx}(\vec{\omega})$, which is the PSC. It is straightforward to show that there always exists λ, $-K/a^2(2-a)^2 \leq \lambda < 0$, to ensure that $\Phi_{ww}(\vec{\omega}) \geq 0$, $\forall \vec{\omega}$.

REFERENCES

1. J. K. Su and B. Girod, "Power spectrum condition for energy-efficient watermarking," in *Proc. IEEE Intl. Conf. Image Proc.*, (Kobe, Japan), Oct. 1999.
2. I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio, and video," in *Proc. IEEE Intl. Conf. Image Proc.*, vol. 3, (Lausanne, Switzerland), Sep. 1996.
3. F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Processing* **66**, pp. 283–301, 1998.
4. F. Hartung, J. K. Su, and B. Girod, "Spread spectrum watermarking: Malicious attacks and counterattacks," in *Proc. SPIE Security & Watermarking Multimedia Contents*,[13] pp. 147–158.
5. J. K. Su, F. Hartung, and B. Girod, "A channel model for a watermark attack," in *Proc. SPIE Security & Watermarking Multimedia Contents*,[13] pp. 159–170.
6. J. K. Su and B. Girod, "On the imperceptibility and robustness of digital fingerprints," in *Proc. IEEE Intl. Conf. Multimedia Comp. Sys.*, vol. 2, pp. 530–535, (Florence, Italy), Jun. 1999.
7. J. C. Hancock and P. A. Wintz, *Signal Detection Theory*, McGraw-Hill Book Co., New York, 1966.
8. M. Schwartz and L. Shaw, *Signal Processing: Discrete Spectral Analysis, Detection, and Estimation*, McGraw-Hill, New York, 1975.
9. P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding." Preprint, Sep. 1999.
10. B. Chen and G. W. Wornell, "Provably robust digital watermarking," in *Proc. SPIE Multimedia Systems and Applications II*, vol. 3845, 1999.
11. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.
12. J. K. Su, J. J. Eggers, and B. Girod, "Analysis of digital watermarks subjected to optimum linear filtering and additive noise." Preprint, Jan. 2000.
13. *Proc. SPIE Security & Watermarking Multimedia Contents*, vol. 3657, (San Jose, CA, USA), Jan. 1999.