Power-Spectrum Condition for Energy-Efficient Watermarking

Jonathan K. Su, *Member, IEEE,* and Bernd Girod, *Fellow, IEEE*

***Abstract*—This paper presents a model for watermarking and some attacks on watermarks. Given the watermarked signal, the so-called *Wiener attack* performs minimum mean-squared error (MMSE) estimation of the watermark and subtracts the weighted MMSE estimate from the watermarked signal. Under the assumption of a fixed correlation detector, the attack is shown to minimize the expected correlation statistic for the same attack distortion among linear, shift-invariant filtering attacks. It also leads to the idea of *energy-efficient watermarking*—watermarking that resists MMSE estimation as much as possible—and provides a meaningful way to evaluate robustness. The paper shows that energy-efficient watermarks must satisfy a *power-spectrum condition* (PSC), which states that the watermark's power spectrum should be directly proportional to the original signal's. PSC-compliant watermarks are proven to be most robust. Experiments with signal models and natural images demonstrate that watermarks that do not closely fulfill the PSC are vulnerable to the Wiener attack, while PSC-compliant watermarks are highly resistant to it. These theoretical and experimental results justify prior heuristic arguments that, for maximum robustness, a watermark should be closely matched to the spectral content of the original signal. The results also discourage the use of watermarks that do not approximately satisfy the PSC.**

***Index Terms*—Digital watermarking, robustness, spread spectrum, watermark attacks.**

I. Introduction

DIGITAL data, such as digital audio, images, and video, can be stored, copied, and distributed quickly, easily, and without any loss of fidelity. Although generally beneficial, these properties create problems in controlling access to or distribution of valuable digital data. Owners and authorized users of such data would like to protect them against unauthorized usage such as duplication and re-distribution.

Digital watermarking has been proposed as part of a system to protect digital data against unauthorized use [1], [2]. A digital watermarking system embeds information directly into digital data to produce watermarked data. As a result, even if copy-protection or encryption mechanisms fail, the information resides in the watermarked data. This information may then be used to determine whether or not the data was acquired through legitimate means.

In general, a digital watermark should have several different properties. The most important are imperceptibility, security, and robustness. *Imperceptibility* means that the watermarked data should be perceptually equivalent to the original, unwatermarked data. In some applications, the watermark may be perceptible as long as it is not annoying or obtrusive; however, many applications require that the watermark be imperceptible. *Security* means that unauthorized parties should not be able to detect or manipulate the watermark. Cryptographic methods are typically employed to make watermarks secure. Finally, *robustness* means that, given the watermarked data, one should not be able to make the watermark undetectable without also destroying the value or usefulness of the data.

Another characteristic of a watermarking scheme is whether or not the original data is available during detection. In some schemes [3], the watermark detector has access to the original data. Hence, interference from the original can presumably be eliminated. *Blind* schemes do not have the luxury of using the original during watermark detection [4], [5]. They typically apply some pre-processing to the received data to suppress interference from the original [4], [6].

A. Attacks on Watermarks

In this paper, we are primarily concerned with robustness. Before discussing robustness further, we need to introduce the idea of an *attack* on a watermark. An attack is any processing of the watermarked data that might damage the watermark. Attacks can be coincidental, such as JPEG compression of a legally obtained image, or hostile, such as an attempt by a multimedia pirate to destroy a watermark before re-selling watermarked data.

Examples of attacks include compression, linear filtering, geometric transformations, and D/A–A/D conversion. Some extensive lists appear in [3] and [7], but it is impossible to name all of the potential attacks. Instead, [8] provides a set of conceptual attack categories. In the present paper, we take a theoretical approach and only consider attacks that attempt to remove a watermark or to confuse the watermark detector by linear, shift-invariant (LSI) filtering.

B. Watermark Robustness

How should a watermark be structured to maximize its robustness? Cox *et al.* [3] suggest that an image watermark should be restricted to the "perceptually significant" (e.g., large-amplitude) spectral components. Large-amplitude components offer better masking potential and cannot be removed without also degrading the image. Likewise, Swanson *et al.* [2] propose the use

Manuscript received July 28, 1999; revised March 11, 2002. The associate editor coordinating the review of this paper and approving it for publication was Dr. Minerva Yeung.

J. K. Su was with the Telecommunication Laboratory, University Erlangen-Nuremberg, Erlangen, Germany. He is now with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02420-9185 USA. (e-mail: su@ll.mit.edu).

B. Girod was with the Telecommunication Laboratory, University Erlangen-Nuremberg, Erlangen, Germany. He is now with the Information Systems Laboratory, Stanford University, Stanford, CA 94320 USA.

Digital Object Identifier 10.1109/TMM.2002.806535

of frequency-domain perceptual masking models. They reason that a watermark that is well-matched to the frequency content of the original signal can be hidden effectively.

On the other hand, Hsu and Wu [9] and Piva *et al.* [5] suggest placing the watermark in the middle frequencies. Hsu and Wu explain that, with regard to imperceptibility, the human visual system is less sensitive to high spatial frequencies, but with regard to robustness, processing like compression only preserves low spatial frequencies. As a compromise, the watermark should lie in the middle frequencies. Piva *et al.* also choose middle-frequency embedding as a trade-off between imperceptibility and robustness.

Xia *et al.* [10] propose embedding an image watermark in the middle and high frequencies. They reason that the human visual system is less sensitive to noise at edges and textures, which correspond to higher-frequency content. Hence, the watermark will be less perceptible, and they claim that it remains robust against attacks such as compression and additive noise.

Zhu *et al.* [11] employ a wavelet-based scheme and argue for placing a watermark in the high frequencies to keep the watermark imperceptible. They report that this watermark remains detectable after wavelet-based compression.

Clearly, uncertainty about the proper structure of a watermark remains. Part of the difficulty in answering the question is that robustness is easy to postulate but hard to measure. Currently, it is still difficult to quantify the detectability of an attacked watermark and the quality of the attacked data. Some initial attempts have been made in [7] and [12]. They are based on selecting a distortion measure, performing a battery of attacks on different watermarks, and measuring quantities such as the probability of error after each attack. They propose a methodology for evaluating robustness experimentally, but they are specific to a given watermarking method and the set of attacks. Moreover, they lack a strong theoretical foundation and development.

This paper also attempts to answer—at least in part—the question posed at the beginning of this section. We take a theoretical approach to watermarking, which we initially presented in [13] and [14] and investigate further here. Section II introduces a general watermarking model. In Section III, we present the Wiener attack, which includes two interesting special cases: removal and anticorrelation attacks. This framework leads to the idea of energy-efficient watermarking, and it enables us to link watermark detectability to signal quality. The latter property produces a meaningful robustness criterion. Section IV explains how to resist the Wiener attack, which produces the *power-spectrum condition* (PSC), the main result of the paper. Experimental results with theoretic signal models, synthetic random signals, and natural images appear in Section V and further illustrate the importance of the PSC.

II. Watermarking Model

As a general watermarking model, we treat the watermark and original data as signals, both deterministic and random. Random signals are modeled as ergodic, zero-mean, wide-sense stationary (WSS) discrete-time (or discrete-space) random processes (DTRPs). Boldface indicates random quantities (e.g., $\mathbf{x}[n]$), and normal typeface is used for deterministic values (e.g., γ) or realizations of random quantities (e.g., $x[n]$). For ease of notation, the analysis focuses on one-dimensional signals, but the results extend directly to M-dimensional (M-D) signals as well; the M-D results are noted. To index an M-D signal $x[n_1, n_2, \ldots, n_M]$, we often use the notation $x[\vec{n}]$, where $\vec{n} = (n_1, n_2, \ldots, n_M)$. Similar notation is used for the frequency variable $\vec{\omega} = (\omega_1, \ldots, \omega_M)$.

The original signal (also called "host data" or "cover data") is represented by the process $\mathbf{x}[n]$, which has variance σ_x^2, autocorrelation function $R_{xx}[k]$, and power spectrum $\Phi_{xx}(\omega)$. Similarly, $\mathbf{w}[n]$ denotes the watermark, which has variance σ_w^2 and power spectrum $\Phi_{ww}(\omega)$. We assume that $\mathbf{x}[n]$ and $\mathbf{w}[n]$ are independent. The support of a realization is denoted by $\mathcal{N}$, and N is the number of samples in $\mathcal{N}$.

The models for embedding, distortion, and detection are first given in the context of deterministic signals $x[n]$, $w[n]$, etc. However, the analysis treats the signals as realizations of the corresponding random processes $\mathbf{x}[n]$, $\mathbf{w}[n]$, etc., and it characterizes embedding, distortion, and detection by examining the expected behavior over the ensembles of $\mathbf{x}[n]$ and $\mathbf{w}[n]$. Finally, we assume that watermark security is achieved by making $w[n]$ the output of a cryptographically secure pseudo-random number generator with a secret key known only to authorized parties.

A. Watermark Embedding

The watermarked signal $y[n]$ is simply $y[n] = x[n] + w[n]$, where $x[n]$ and $w[n]$ are realizations of the respective random processes $\mathbf{x}[n]$ and $\mathbf{w}[n]$. In the context of random processes:

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{w}[n]. \tag{1}$$

Since $\mathbf{x}[n]$ and $\mathbf{w}[n]$ are independent

$$\Phi_{yy}(\omega) = \Phi_{xx}(\omega) + \Phi_{ww}(\omega), \quad \text{and} \quad \Phi_{wy}(\omega) = \Phi_{ww}(\omega) \tag{2}$$

where $\Phi_{wy}(\omega)$ is the cross-power spectrum of $\mathbf{w}[n]$ and $\mathbf{y}[n]$.

We remark that many current watermarking methods are based on spread-spectrum communications [15], [16]. The seminal work on digital image fingerprinting by Cox *et al.* in [3] popularized the use of direct-sequence spread-spectrum for watermarking. The model (1) encompasses spread-spectrum watermarking, discussed in more detail in [4] and [17], for example.

B. Distortion Measure

To quantify signal quality, we measure the distortion between a signal $\hat{x}[n]$ and the original signal $x[n]$ via the *sample mean-squared error* (sample MSE):

$$D(\hat{x}, x) = \frac{1}{N} \sum_{n \in \mathcal{N}} (\hat{x}[n] - x[n])^2 .$$

In the context of random processes $\hat{\mathbf{x}}[n]$ and $\mathbf{x}[n]$, the sample MSE is replaced by an expectation, and the *distortion* is the (ensemble) MSE

$$D(\hat{\mathbf{x}}, \mathbf{x}) = \mathrm{E}\left[(\hat{\mathbf{x}}[n] - \mathbf{x}[n])^2\right]. \tag{3}$$

Note that $D(\hat{x}, x)$ is a sample average, while the distortion $D(\hat{\mathbf{x}}, \mathbf{x})$ is an ensemble average. We also express signal quality

as fidelity via the *original-to-noise ratio* (ONR), given by $\mathrm{ONR}(\hat{\mathbf{x}}, \mathbf{x}) = 10\log_{10}\sigma_x^2/D(\hat{\mathbf{x}}, \mathbf{x})$ dB.

For the watermarked signal $\mathbf{y}[n]$, the *embedding distortion* is $D(\mathbf{y}, \mathbf{x}) = \sigma_w^2$. The watermark signal $\mathbf{w}[n]$ should be imperceptible, so we define the *watermark-to-original ratio* (WOR) by $\mathrm{WOR} = 10\log_{10}(\sigma_w^2/\sigma_x^2)$ dB $= -\mathrm{ONR}(\mathbf{y}, \mathbf{x})$. As a rule of thumb for image watermarking, WORs below -20 dB are required to keep the watermark imperceptible. For an attacked signal $\hat{\mathbf{y}}[n]$, the *attack distortion* is $D(\hat{\mathbf{y}}, \mathbf{x})$.

C. Watermark Detection

Given a received signal $\hat{y}[n]$, the watermark detector makes a (possibly incorrect) decision about the presence or absence of $w[n]$. We assume that the detector is synchronized with the embedded watermark. A popular detection method is *correlation detection*, in which the detector computes the *sample correlation statistic*

$$s = \frac{1}{N}\sum_{n\in\mathcal{N}} \hat{y}[n]w[n] \tag{4}$$

and then compares s to a threshold T to decide whether $w[n]$ is present in $\hat{y}[n]$ ($s > T$) or not ($s \leq T$). A larger value of s corresponds to increasing confidence that $w[n]$ is indeed present in $\hat{y}[n]$, and typically T lies between 0 and σ_w^2. An important assumption in this paper is that *the detector is fixed*. The next section motivates this assumption.

In the random-signal context, during detection, the watermark signal is a particular realization $w[n]$ of $\mathbf{w}[n]$ and is completely known to the detector. Hence, when treating the correlation statistic as a random variable, we must condition on $w[n]$. Then r, the *expected value of the correlation statistic*, is

$$r = \mathrm{E}\left[\mathrm{E}\left[\mathbf{s}|\mathbf{w}[n]\right]\right] = \mathrm{E}\left[\mathbf{s}\right]. \tag{5}$$

Since usually $0 < T < \sigma_w^2$, we often normalize r by σ_w^2 to describe the relative amount of watermark power that reaches the receiver.

D. Overview of Attack and Defense

We briefly summarize the approach used to study the attack and defense. We characterize the attack and defense by examining the distortion $D(\hat{\mathbf{y}}, \mathbf{x})$ and r, the expected value of the correlation statistic. The attack is motivated as follows. Ideally, the attacker wishes to recover the original signal $x[n]$ from $y[n]$. Failing that, the attacker would like to produce an attacked signal $\hat{y}[n]$ such that $\hat{y}[n]$ has acceptable fidelity and that the watermark detector will (incorrectly) decide that $w[n]$ was not embedded in $\hat{y}[n]$. Rather than working with particular realizations, we consider random processes, so the attacker's problem is to minimize $D(\hat{\mathbf{y}}, \mathbf{x})$ such that $r = r_0$, where r_0 is the desired expected value of the correlation statistic. Note that we do *not* consider the variance $\mathrm{var}(\mathbf{s})$ because we are primarily interested in the case when $r = 0$, where detection becomes unreliable.

The defense is similarly motivated. Given the attack chosen by the attacker, the watermark signal $w[n]$ is a realization of $\mathbf{w}[n]$, which is characterized by its power spectrum $\Phi_{ww}(\omega)$. The watermarker chooses $\Phi_{ww}(\omega)$ to maximize $D(\hat{\mathbf{y}}, \mathbf{x})$ such that $r = r_0$ and $D(\mathbf{y}, \mathbf{x}) \leq D_{\mathrm{embed}}$, where the upper bound D_{embed} imposes the imperceptibility requirement. As shown below, the solution for $\Phi_{ww}(\omega)$ is equivalent to making estimation of $w[n]$ from $y[n]$ as difficult as possible, in a well-defined sense.

A key assumption used throughout this paper is that *the detector is fixed and does not compensate for the attack*. Some recent information-theoretic papers have adopted a game-theoretic approach and consider the ideal situation in which the receiver knows the attack and can compensate for it [18], [19]. There are several reasons, primarily pragmatic, for the assumption of a fixed detector. First, we assume the use of a correlation detector, which is popular in many watermarking schemes in the current literature. Correlation detection is optimal for detecting a known signal (i.e., the watermark signal) in additive white Gaussian noise (AWGN). It is suboptimal if the signal is not degraded solely by AWGN. For example, the noise may be colored (prewhitening is required prior to correlating) or non-Gaussian (e.g., a sign detector is locally optimal for additive, white Laplacian noise), or the signal may be filtered (inverse filtering is necessary). The key point is that any detector that has been designed for a specific set of assumed attacks will suffer when the actual attack differs significantly from the design assumptions. It is thus reasonable to examine the behavior of a fixed detector when its assumptions are violated. For example, Voloshynovskiy *et al.* [20] have proposed an effective attack in which outliers are introduced to confuse the correlation detector.

Second, when watermarking is viewed as a game [18], [19], the watermarker and attacker are opponents who alternately improve their respective methods. In theory, the game continues until one player wins or a stable equilibrium is reached. In practice, however, once the watermarking system has been specified and deployed, the watermarker can no longer modify it. The attacker, on the other hand, is free to develop additional, ever more insidious attacks. The watermarker can only hope that the deployed watermarking system can withstand them.

Third, for implementation reasons, it may not be feasible or cost-effective to build thousands of sophisticated watermark detectors that perform attack estimation and compensation; the simple correlation detector may be an economic, rather than engineering, choice.

III. Wiener Attack

From Section II-D, the attacker's goal is to minimize $D(\hat{\mathbf{y}}, \mathbf{x})$ such that $r = r_0$. To impose some structure on the problem, we assume that the attack consists of LSI filtering and additive noise. Let $g[n]$ and $G(\omega)$ denote the filter's impulse response and transfer function, respectively, and $\mathbf{v}[n]$ denote the noise, which has power spectrum $\Phi_{vv}(\omega)$ and is independent of $\mathbf{x}[n]$ and $\mathbf{w}[n]$. Then the attacked signal $\hat{y}[n]$ is

$$\hat{\mathbf{y}}[n] = g[n] * \mathbf{y}[n] + \mathbf{v}[n] = g[n] * (\mathbf{x}[n] + \mathbf{w}[n]) + \mathbf{v}[n]. \tag{6}$$

We formally state the attacker's problem as: Given $\Phi_{xx}(\omega)$, $\Phi_{ww}(\omega)$, and r_0, select $G(\omega)$, $\Phi_{vv}(\omega)$ to minimize $D(\hat{\mathbf{y}}, \mathbf{x})$ such that $r = r_0$. The solution is given by the following theorem, which is proved in the Appendix.

Theorem 1 (Wiener Attack): Let $\Phi_{xx}(\omega)$, $\Phi_{ww}(\omega)$, and r_0 be given. Under the constraint $r = r_0$, $D(\hat{\mathbf{y}}, \mathbf{x})$ is minimized if and only if

$$G(\omega) = 1 - \gamma H(\omega), \qquad M\text{-D: } G(\vec{\omega}) = 1 - \gamma H(\vec{\omega}),$$
$$\Phi_{vv}(\omega) = 0, \qquad \Phi_{vv}(\vec{\omega}) = 0 \tag{7}$$

where γ is a real, scalar *gain factor*, and

$$H(\omega) = \frac{\Phi_{ww}(\omega)}{\Phi_{xx}(\omega) + \Phi_{ww}(\omega)},$$
$$M\text{-D: } H(\vec{\omega}) = \frac{\Phi_{ww}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}. \tag{8}$$

With $G(\omega)$, $H(\omega)$, and $\Phi_{vv}(\omega)$ so defined, for any γ,

$$r = \sigma_w^2 - \gamma\sigma_{\hat{w}}^2 \tag{9}$$
$$D(\hat{\mathbf{y}}, \mathbf{x}) = \sigma_w^2 - \gamma(2-\gamma)\sigma_{\hat{w}}^2, \tag{10}$$
$$E = \mathrm{E}\left[(\mathbf{w}[n] - \hat{\mathbf{w}}[n])^2\right] = \sigma_w^2 - \sigma_{\hat{w}}^2 \tag{11}$$

where

$$\sigma_{\hat{w}}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_{ww}^2(\omega)}{\Phi_{ww}(\omega) + \Phi_{xx}(\omega)}\, d\omega. \tag{12}$$

Hence, to achieve $r = r_0$:

$$\gamma = \left(\sigma_w^2 - r_0\right)/\sigma_{\hat{w}}^2. \tag{13}$$

The corresponding attack distortion is

$$D(\hat{\mathbf{y}}, \mathbf{x}) = \sigma_w^2 - 2\left(\sigma_w^2 - r_0\right) + \frac{\left(\sigma_w^2 - r_0\right)^2}{\sigma_{\hat{w}}^2}. \tag{14}$$

Let $h[n]$ denote the impulse response corresponding to $H(\omega)$, so $g[n] = \delta[n] - \gamma h[n]$. Also let $\hat{\mathbf{w}}[n] = h[n] * \mathbf{y}[n]$. Observe that $H(\omega)$ is the transfer function of the Wiener filter for estimating $\mathbf{w}[n]$ from $\hat{\mathbf{y}}[n]$, so $\hat{\mathbf{w}}[n]$ is the Wiener or *linear minimum mean-squared error* (LMMSE) estimate of $\mathbf{w}[n]$ given $\hat{\mathbf{y}}[n]$. E in (11) is the MSE of the estimate. If $\mathbf{x}[n]$ and $\mathbf{w}[n]$ are further assumed to be jointly Gaussian, then the Wiener filter produces the MMSE estimate among all estimators, including nonlinear estimators.

Equation (6) becomes

$$\hat{\mathbf{y}}[n] = (\delta[n] - \gamma h[n]) * \mathbf{y}[n] + \mathbf{v}[n] = \mathbf{y}[n] - \gamma\hat{\mathbf{w}}[n] \tag{15}$$

since (7) indicates that $\mathbf{v}[n] = 0, \forall\, n$. From (15), the attack can be viewed as first computing the Wiener estimate $\hat{\mathbf{w}}[n]$ of the watermark signal $\mathbf{w}[n]$ from $\mathbf{y}[n]$ and then modifying $\mathbf{y}[n]$ by subtracting a weighted version of $\hat{\mathbf{w}}[n]$ and adding noise $\mathbf{v}[n]$. We call this attack the *Wiener attack*; a block diagram appears in Fig. 1.

A. Discussion of the Attack

The theorem indicates that the attack should *not* introduce any additive noise. Intuitively, the attacker can only affect r through $\hat{\mathbf{w}}[n]$ since $\hat{\mathbf{w}}[n]$ is the MMSE estimate of $\mathbf{w}[n]$. Any other changes to $\mathbf{y}[n]$ are uncorrelated with $\mathbf{w}[n]$ and can thus only increase $D(\hat{\mathbf{y}}, \mathbf{x})$ without reducing r. Technically, an examination of the expressions for r and $D(\hat{\mathbf{y}}, \mathbf{x})$ in Appendix B reveals that setting $\sigma_v^2 > 0$ increases $D(\hat{\mathbf{y}}, \mathbf{x})$ but does not affect r. The noise does not improve the attack, so the attacker should set $\mathbf{v}[n] = 0, \forall\, n$. This somewhat surprising result occurs because the fixed correlation detector does not compensate for the attack; if the receiver compensated for the attack, noise would be necessary [18], [19].

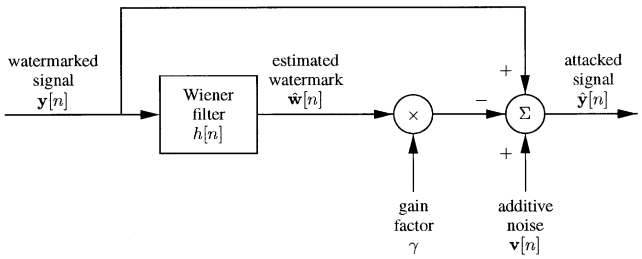


Fig. 1. Block diagram of Wiener attack.

For given power spectra $\Phi_{xx}(\omega)$ and $\Phi_{ww}(\omega)$, we can easily compute the relationship between r and $D(\hat{\mathbf{y}}, \mathbf{x})$. We only need to compute $\sigma_{\hat{w}}^2$ in (12) (e.g., by numerical integration), and then we can use (14) to find $D(\hat{\mathbf{y}}, \mathbf{x})$ for any r_0. We can also compute E via (11).

From (9) and (10), both r and $D(\hat{\mathbf{y}}, \mathbf{x})$ can be parameterized by the gain factor γ. It is now possible to relate watermark detectability, in terms of r, to the attack distortion $D(\hat{\mathbf{y}}, \mathbf{x})$. The attacker varies γ to trade off r and $D(\hat{\mathbf{y}}, \mathbf{x})$. Two values of γ result in interesting special cases of the Wiener attack.

1) Removal Attack: With $\gamma = 1$, the Wiener attack is a *removal attack*. For the attacker, this form has the appealing property that it removes as much of the watermark energy as possible while minimizing the attack distortion. This case is equivalent to Wiener denoising. The result is intuitively clear from (15), or it may be derived by taking (10) and setting $dD(\hat{\mathbf{y}}, \mathbf{x})/d\gamma = 0$. Also, $r = E$ when $\gamma = 1$.

2) Anticorrelation Attack: The attacker can instead select γ so that $r_0 = 0$, at the expense of increasing $D(\hat{\mathbf{y}}, \mathbf{x})$. We denote this special value of γ by γ_0,

$$\gamma_0 = \sigma_w^2/\sigma_{\hat{w}}^2. \tag{16}$$

This choice of γ drives r to zero with the minimum corresponding distortion $D(\hat{\mathbf{y}}, \mathbf{x})$. Since usually $0 < T < \sigma_w^2$, the probability that the detector mistakenly decides that $w[n]$ is not present in $\hat{y}[n]$ is at least 0.5. We call this attack an *anticorrelation attack*; the name emphasizes that the attack forces r to zero, as opposed to disabling detection by some other mechanism (e.g., desynchronization). We do not use the term "decorrelation attack," which could imply transforming $\mathbf{w}[n]$ or $\mathbf{y}[n]$ into uncorrelated components like the Karhunen–Loève transform.

This form of the Wiener attack is similar to an attack proposed by Langelaar *et al.* [21], who used nonlinear filtering to estimate a portion of a white-noise watermark and drive the expected correlation statistic to zero. However, the Wiener attack

is easier to analyze because of its linearity, and it permits colored watermarks.

B. Energy-Efficient Watermarking and a Robustness Criterion

We can interpret the normalized MSE E/σ_w^2 as the fraction of watermark energy that resists MMSE estimation. Since energy that can be estimated can also be removed, it is wasted. A watermark that maximizes E/σ_w^2 wastes the minimum fraction of its energy and is said to be *energy-efficient*. Since $0 \leq E/\sigma_w^2 \leq 1$, we can also compare E/σ_w^2 for different watermarks. A larger ratio means greater resistance to MMSE estimation.

In addition, we now have a well-defined way of evaluating the robustness of a watermark. Given different watermarks $\mathbf{w}_1[n]$, $\mathbf{w}_2[n]$, etc., which are characterized by their respective power spectra $\Phi_{w_1w_1}(\omega)$, $\Phi_{w_2w_2}(\omega)$, etc., the watermark $\mathbf{w}_j[n]$ that produces the largest value of $D(\hat{\mathbf{y}}_j, \mathbf{x})$ for a given value of $r = r_0$ is most robust. Similarly, if all watermarks yield the same attack distortion D_0, then the watermark with the greatest value of r_j is most robust. We thus have a meaningful way to compare the robustness of watermarks.

IV. Resisting the Wiener Attack: The Power-Spectrum Condition

Now let us consider the watermarker's perspective. From Section III, the watermarker wishes to maximize $D(\hat{\mathbf{y}}, \mathbf{x})$ under the constraints $r = r_0$ and $D(\mathbf{y}, \mathbf{x}) = \sigma_w^2 \leq D_{\text{embed}}$. So that the greatest amount of watermark energy might reach the receiver, the watermarker should choose $\sigma_w^2 = D_{\text{embed}}$. The watermarker cannot alter the original signal's power spectrum $\Phi_{xx}(\omega)$, but the watermarker has the freedom to specify the watermark's power spectrum $\Phi_{ww}(\omega)$.

From (14), $D(\hat{\mathbf{y}}, \mathbf{x})$ is maximized when $\sigma_{\hat{w}}^2$ is minimized, and from (11), $\sigma_{\hat{w}}^2$ is minimized when E is maximized. Hence, regardless of r_0, the watermarker should choose $\Phi_{ww}(\omega)$ to maximize E—and hence create an energy-efficient watermark—under the variance constraint.

The solution of this problem leads to the theorem below; the proof appears in [13].

Theorem 2 (Power-Spectrum Condition): For the watermarking model (1), E is maximized if and only if

$$\Phi_{ww}(\omega) = \frac{\sigma_w^2}{\sigma_x^2}\,\Phi_{xx}(\omega), \qquad M\text{-D: } \Phi_{ww}(\vec{\omega}) = \frac{\sigma_w^2}{\sigma_x^2}\,\Phi_{xx}(\vec{\omega}) \tag{17}$$

and, for any dimensionality M, the maximum MSE is

$$E_{\text{PSC}} = \frac{\sigma_x^2\sigma_w^2}{\sigma_w^2 + \sigma_x^2} = \alpha_{\text{PSC}}\sigma_w^2 \tag{18}$$

where $\alpha_{\text{PSC}} = \sigma_x^2/(\sigma_w^2 + \sigma_x^2)$.

We remarked in [13] that one could use a frequency-weighted MSE E_{wt} instead of E. Such a weighting might be desirable for applications like audio watermarking, which use frequency-domain perceptual masking models [22]. Unfortunately, the solution for $\Phi_{ww}(\omega)$ does not have a convenient form like (17) and does not lend itself to tractable analysis.

TABLE I
Example Values for PSC-Compliant Independent Watermarks. The Attacks Assume No Additive Noise ($\sigma_v^2 = 0$)

WOR	α_{PSC}	Removal		Anticorrelation	
		r	$\text{ONR}(\hat{\mathbf{y}}, \mathbf{x})$	$\gamma_{0,\text{PSC}}$	$\text{ONR}(\hat{\mathbf{y}}, \mathbf{x})$
−20 dB	0.9901	0.9901	20.0432 dB	101.00	0 dB
−25 dB	0.9968	0.9968	25.0137 dB	317.23	0 dB
−30 dB	0.9990	0.9990	30.0043 dB	1001.02	0 dB
−35 dB	0.9997	0.9997	35.0014 dB	3163.28	0 dB
−40 dB	0.9999	0.9999	40.0004 dB	10001.00	0 dB

A. Consequences of the Power-Spectrum Condition

We refer to (17) as the *power-spectrum condition*[1] (PSC). It states that the watermark's power spectrum should be directly proportional the original signal's power spectrum. In this sense, *the watermark should look like the original*. We say that a watermark that satisfies (17) is *spectrally matched* to the original or *PSC-compliant*. In this section, we study what happens when (17) is satisfied.

The main result is that *a spectrally-matched watermark signal is most robust*, in the sense that the attacker must introduce the greatest amount of distortion to make $r = r_0$. Important conditions are the assumptions of a fixed correlation detector and the form of the attack (LSI filtering and additive noise).

The Wiener filter transfer function (8) reduces to

$$H_{\text{PSC}}(\omega) = H_{\text{PSC}}(\vec{\omega}) = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_x^2} = 1 - \alpha_{\text{PSC}} \tag{19}$$

and the corresponding maximum MSE is given in (18). Note that the normalized MSE E/σ_w^2 for a PSC-compliant watermark is simply α_{PSC}.

From (12), $\sigma_{\hat{w}}^2 = \sigma_w^4/(\sigma_x^2 + \sigma_w^2) = (1 - \alpha_{\text{PSC}})\sigma_w^2$. Then (9) and (10) give

$$r = (1 - \gamma\,(1 - \alpha_{\text{PSC}}))\,\sigma_w^2, \tag{20}$$

$$D(\hat{\mathbf{y}}, \mathbf{x}) = (1 - \gamma(2 - \gamma)\,(1 - \alpha_{\text{PSC}}))\,\sigma_w^2. \tag{21}$$

These expressions hold regardless of the dimensionality M.

Since $\mathbf{w}[n]$ should be imperceptible, we assume $\sigma_x^2 \gg \sigma_w^2$, so $\alpha_{\text{PSC}} \approx 1$. Table I lists example values of α_{PSC} for WORs from −20 to −40 dB. We see that these watermarks have normalized MSEs close to unity. Hence, a PSC-compliant watermark can hardly be estimated by a Wiener filter.

B. Special Cases of the Wiener Attack

If the attacker sets $\gamma = 1$ for a removal attack, then the expected correlation statistic and attack distortion become $r = D(\hat{\mathbf{y}}, \mathbf{x}) = \alpha_{\text{PSC}}\sigma_w^2 \approx \sigma_w^2$. As a result, the variance of the watermark is hardly reduced by the attack, and the distortion of the attacked signal $\hat{\mathbf{y}}[n]$ is a negligible improvement over the watermarked signal $\mathbf{y}[n]$. Indeed, $\text{ONR}(\hat{\mathbf{y}}, \mathbf{x}) = (\sigma_x^2 + \sigma_w^2)/\sigma_w^2 \approx \sigma_x^2/\sigma_w^2 = \text{ONR}(\mathbf{y}, \mathbf{x})$. From the third column of Table I, it is clear that r is barely affected by this attack. The fourth column

[1]It seems likely that this result has been found previously, since it amounts to the worst case for MMSE estimation of a signal (the watermark) subject to signal-independent, additive, colored noise.

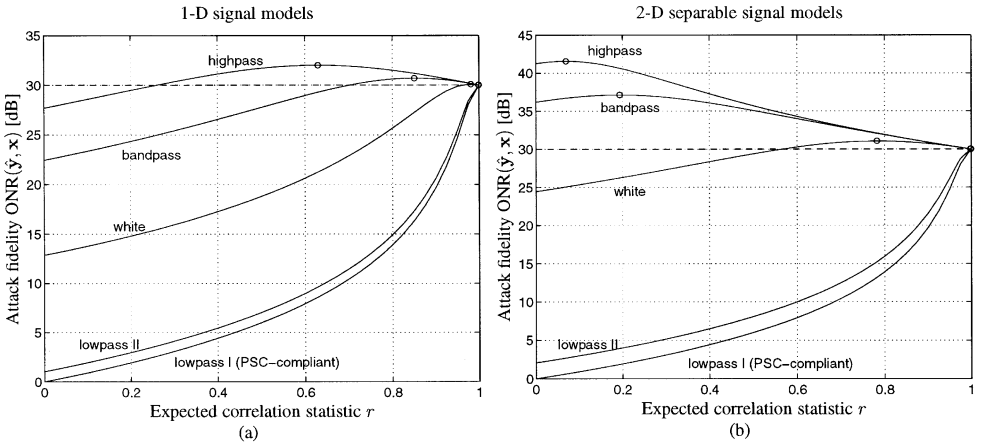


Fig. 2. Theoretical performance of watermarks using AR signal models. Original signal has a lowpass I signal model and $\sigma_w^2 = 1$. $\mathrm{WOR} = -30$ dB, which is indicated by the dashed line ($\mathrm{ONR}(\mathbf{y}, \mathbf{x}) = -\mathrm{WOR}$). Circles indicate the results of the removal attack.

shows that $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ improves by less than 0.05 dB over $\mathrm{ONR}(\mathbf{y}, \mathbf{x})$, which demonstrates the effectiveness of a spectrally-matched watermark.

Suppose instead that the attacker performs the anticorrelation attack. From (16), γ_0 becomes $\gamma_{0,\,\mathrm{PSC}} = (\sigma_w^2 + \sigma_x^2)\sigma_w^2 = 1/(1 - \alpha_{\mathrm{PSC}})$, and (10) gives $D(\hat{\mathbf{y}}, \mathbf{x}) = \sigma_x^2 + \sigma_v^2$. As a result, the attack distortion will be at least as large as the variance of the original signal, and $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x}) \leq 0$ dB. Such an attacked signal will certainly be useless. In the fifth column of Table I, the required gain factor $\gamma_{0,\,\mathrm{PSC}}$ is given for the example WORs. Note how much the estimate $\hat{\mathbf{w}}[n]$ must be amplified, which introduces large amounts of distortion; the sixth column of the table shows the corresponding $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$, which is always 0 dB.

V. Experimental Results

A. Theoretical Performance for Signal Models

For the purpose of analysis, correlated signals—such as audio or images—are often approximated by *autoregressive* (AR) random processes [23]. We denote the one–dimensional (1-D), pth-order AR process $\mathbf{x}[n]$ by AR(p), which has the model $\mathbf{x}[n] = \sum_{k=1}^{p} a_k \mathbf{x}[n-k] + \mathbf{u}[n]$, where $\mathbf{u}[n]$ is 1-D WSS white noise. "Lowpass I" and "Lowpass II" denote AR(1) models with $a_1 = 0.95$ and 0.90, respectively. "Bandpass" is an AR(2) model with $a_1 = 0$, $a_2 = -0.81$, while "Highpass" refers to an AR(1) model with $a_1 = -0.95$.

For two-dimensional (2-D) DTRPs, we employ a separable AR(p_1, p_2) model, $\mathbf{x}[n_1, n_2] = \mathbf{x}_1[n_1]\mathbf{x}_2[n_2]$, where $\mathbf{x}_1[n_1]$ and $\mathbf{x}_2[n_2]$ are 1-D AR(p_1) and AR(p_2) processes, respectively, and $p_1 = p_2 = p$. Hence, for 2-D signals, "Lowpass I" refers to the product of two 1-D AR(1) "Lowpass I" models (horizontal and vertical), and likewise for other designations such as "Bandpass." Of course, more flexible 2-D power spectra models could be employed, but these are sufficient to illustrate the main ideas in this paper.

In Sections V-A and V-B, no signals are actually generated, no watermarks are actually embedded, and no watermark detection is actually performed. Instead, we examine the theoretical relationship between r_0 and $D(\hat{\mathbf{y}}, \mathbf{x})$ using (12) and (14). Likewise, E is computed from (11).

Fig. 2 plots r against $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ for various watermarks when the WOR is -30 dB, $\sigma_w^2 = 1$, and the original has a 1-D lowpass I model [Fig. 2(a)] or a 2-D separable lowpass I model [Fig. 2(b)]. Because the 2-D signal models have greater signal separation (e.g., between a highpass and lowpass signal) than the 1-D models, the results are more dramatic for 2-D. For reference, the dashed line indicates $\mathrm{ONR}(\mathbf{y}, \mathbf{x})$, the fidelity of the unattacked, watermarked signal $\mathbf{y}[n]$. Circles show the expected correlation statistic/distortion point for the removal attack. The points where the curves intersect the vertical axis give $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ for the anticorrelation attack.

Fig. 2 shows that both bandpass and highpass watermarks are not very robust. For the 1-D case, the removal attack is fairly effective, and the anticorrelation attack can disable detection while maintaining reasonably good signal fidelity. For the 2-D case, removal is very effective, and the anticorrelation attack yields an attacked signal with much better fidelity than the unattacked, watermarked signal.

The removal attack is less successful against white watermarks. The attacked signal may not have acceptable fidelity after the anticorrelation attack for the 1-D case, but it is likely to be usable in the 2-D case.

The curves for the lowpass II watermarks show that a watermark that approximately satisfies the PSC is highly robust. The removal and anticorrelation attacks are ineffective.

PSC-compliant watermarks are clearly superior to the other watermarks. For any value of r, the fidelity $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ is far below that of the other, non-PSC-compliant watermarks. As explained in Section IV-A, (20) and (21) do not depend upon the dimensionality M, so the curves for the 1-D and 2-D PSC-compliant watermarks are identical. As $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ drops, r slowly decreases, so that a large loss in $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ is required to affect r significantly. When $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x}) = 20$ dB, r remains greater than 0.9, and even when $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x}) = 10$ dB, $r \approx 0.7$. The removal attack has almost no effect, and the anticorrelation attack completely destroys the attacked signal.

B. Theoretical Performance for Natural Images

Additional experiments were conducted on 8-bit grayscale natural images. The power spectrum $\Phi_{xx}(\omega_1, \omega_2)$ of the

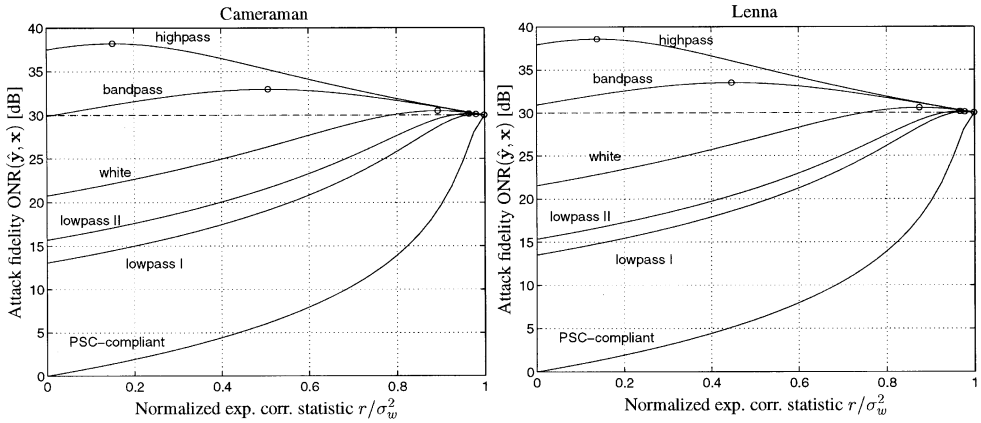


Fig. 3. Theoretical performance of watermarks for natural images. The WOR is −30 dB. Circles indicate the results of the removal attack. For Cameraman, $\mathrm{PSNR} = \mathrm{ONR} + 12.24$ dB; for Lenna, $\mathrm{PSNR} = \mathrm{ONR} + 13.76$ dB.

Fig. 4. Original Cameraman image.

$N_1 \times N_2$ original image $x[n_1, n_2]$ was estimated using the periodogram [23], $\mathrm{Per}_{xx}[k_1, k_2] = |X[k_1, k_2]|^2/(N_1 N_2)$, where $X[k_1, k_2]$ is the 2-D FFT of $x[n_1, n_2]$.

We remark that taking the full-size transform of an image may not be the best implementation for actual watermarking schemes. Also, the periodogram produces an unbiased, but not a consistent, estimate of a signal's power spectrum [23]. Nonetheless, these methods are sufficient for illustrating the relationship between theory and practice.

Fig. 4 shows theoretical performance curves for the 256 × 256 Cameraman and Lenna images. The WOR was set to −30 dB. The image-processing community often uses the *peak signal-to-noise ratio* (PSNR) as a fidelity metric for images and video. PSNR and ONR are related by $\mathrm{PSNR} - \mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x}) = 10\log_{10} 255^2 - 10\log_{10}\sigma_x^2$. The qualitative behavior is similar to the signal-model-based curves of Fig. 2. The anticorrelation attack can defeat bandpass and highpass watermarks; it may also defeat white watermarks, since $\mathrm{PSNR} \approx 33$ dB, which may be acceptable fidelity. The lowpass watermarks should leave severely distorted attacked images, and the PSC-compliant watermark should produce a worthless image; ideally it should equal the mean of $y[n_1, n_2]$.

C. Experimental Performance for Natural Images

Unlike the theoretical investigations of the preceding sections, actual watermarks were generated, embedded, and detected in the following experiments. Here we present example images for Cameraman. The original image was used during detection.

TABLE II
COMPARISON OF PREDICTED (P) AND EXPERIMENTALLY-OBTAINED (E) QUANTITIES FOR THE CAMERAMAN EXAMPLES IN FIGS. 5 AND 6. THE WOR IS −30 dB, AND THERE IS NO ADDITIVE NOISE ($\sigma_v^2 = 0$). NOTE: $\mathrm{PSNR} = \mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x}) + 12.24$ dB

Watermark		E/σ_w^2	Removal		Anticorrelation	
			r/σ_w^2	$\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$	r/σ_w^2	$\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$
Highpass	(P)	0.1514	0.1514	38.20 dB	0	37.49 dB
	(E)	0.1423	0.1517	38.47 dB	0.0004	37.81 dB
Bandpass	(P)	0.5067	0.5067	32.95 dB	0	29.88 dB
	(E)	0.5457	0.5371	32.63 dB	0.0616	29.74 dB
White	(P)	0.8935	0.8935	30.49 dB	0	20.76 dB
	(E)	0.8479	0.8474	30.72 dB	0.0004	22.54 dB
Lowpass I	(P)	0.9642	0.9642	30.16 dB	0	15.69 dB
	(E)	0.9670	0.9655	30.15 dB	0.0355	15.65 dB
PSC-comp.	(P)	0.9990	0.9990	30.00 dB	0	0.00 dB
	(E)	0.9993	0.9990	30.00 dB	0.0000	0.00 dB

To generate PSC-compliant watermarks, we set

$$w[n_1, n_2] = \sqrt{\sigma_w^2/\sigma_x^2}\,\mathrm{IFFT}\left\{\sqrt{\mathrm{Per}_{xx}[k_1, k_2]}\,U[k_1, k_2]\right\}$$

where $U[k_1, k_2]$ is the 2-D FFT of the output $u[n_1, n_2]$ of a unit-variance white Gaussian random number generator. With this construction, $\mathrm{Per}_{ww}[k_1, k_2] \approx (\sigma_w^2/\sigma_x^2)\mathrm{Per}_{xx}[k_1, k_2]$; equality would hold if $\mathrm{Per}_{uu}[k_1, k_2]$ were equal to unity for all $[k_1, k_2]$. Another way to generate a PSC-compliant watermark is to set

$$w[n_1, n_2] = \sqrt{\sigma_w^2/\sigma_x^2}\,\mathrm{IFFT}\,\{|X[k_1, k_2]|\exp j\theta[k_1, k_2]\}$$

where the phase angle $\theta[k_1, k_2]$ is chosen randomly (uniformly distributed over $[0, 2\pi)$) for each $[k_1, k_2]$ but subject to appropriate symmetry constraints to ensure that $w[n_1, n_2]$ remains real. Other watermarks were generated using the 2-D separable AR model.

Fig. 4 shows the original, unwatermarked image. In these experiments on Cameraman, $\mathrm{WOR} = -30$ dB, and $\mathrm{PSNR} = \mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x}) + 12.24$ dB.

Bandpass White Lowpass I PSC-compliant

Fig. 5. Examples of attacked images after removal attack.

Bandpass White Lowpass I PSC-compliant

Fig. 6. Examples of attacked images after anticorrelation attack.

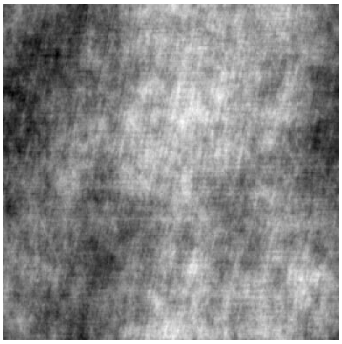


Fig. 7. Example of a PSC compliant watermark for Cameraman.

Table II summarizes the numerical results for Cameraman. The experimental results closely match the values predicted from theory (see Fig. 3). The highpass, bandpass, and white watermarks are all vulnerable to the anticorrelation attack, while the lowpass I and PSC-compliant watermarks cannot be defeated without also destroying the attacked image.

Fig. 5 shows example images resulting from the removal attack for different watermarks. (The watermarked images appear identical to the original, unwatermarked image, so they are not shown.) The removal attack is effective against the bandpas watermark, but less so against the white one. For the lowpass I watermark, the attack leaves $\mathrm{ONR}(\hat{\mathbf{y}}, \mathbf{x})$ virtually unchanged, and the attack has almost no effect against the PSC-compliant watermark.

More important, Fig. 6 shows images with the results of the anticorrelation attack. For both the bandpass and white watermarks, the attacked images still possess good visual quality and would likely be useful to the attacker. The attacked lowpass I watermarked image has been noticeably degraded and may no longer be useful. Finally, for the PSC-compliant watermark, the attacked image after the anticorrelation attack is almost perfectly flat, and it is wortless to the attacker.

An example of PSC-compliant watermark $w[n_1, n_2]$ is shown in Fig. 7. The watermark is obviously correlated, but in the spatial domain it does not resemble the original image.

VI. Summary and Conclusions

The simple models for watermarking and the Wiener attack yield insight into the structure of a watermark for improved robustness. An important assumption is the use of a fixed correlation detector that does not compensate for the effects of attack. Also, the variance of the detector statistic is not considered because we are mainly interested in the case where the expected value of the correlation statistic becomes zero. These considerations lead to the idea of energy-efficient watermarking and provide a way to link the detectability of an attacked watermark to the distortion of the attacked signal. It then becomes possible to evaluate robustness in a meaningful way. The key result is the power-spectrum condition (PSC), which states that a watermark is energy-efficient if and only if its power spectrum is directly proportional to that of the original signal. In terms of power spectra, *the watermark should look like the original*.

The PSC is intuitively satisfying. To make a watermark hard to attack, given the watermarked signal, it should be difficult to separate the watermark and original signals. The PSC holds for any signals that meet the assumptions of the model. It may therefore be applicable to digital audio, images, and video, for example.

The experimental results show that watermarks that fail to satisfy the PSC—i.e., bandpass, highpass, and white watermarks—are vulnerable to the Wiener attack. However, PSC-compliant watermarks cannot be defeated by the Wiener attack. Watermarks that are close to being PSC-compliant also resist attack well, so in practice, simple parametric models for the watermark power spectrum may be sufficient.

These results support the heuristic arguments [2], [3], [24], [25] in favor of matching a watermark to the spectral content of the signal being watermarked. They also refute earlier arguments that advocate bandpass, highpass, or white watermarks.

We remark that compression algorithms like JPEG and MPEG tend to preserve only the large-magnitude frequency components of the original signal. Similarly, perceptually-based watermarking schemes often use frequency transforms or subband decompositions and embed watermark energy roughly in proportion to the magnitude of the transform or subband coefficients of the original signal. As a consequence, compressed-domain watermarking schemes and watermarking methods that use perceptual models both have the appealing side effect that they approximately satisfy the PSC.

APPENDIX
WIENER ATTACK

A. Attack Derivation

From the attack model (6)

$$\begin{aligned} D(\hat{\mathbf{y}},\mathbf{x}) &= \mathrm{E}\left[(\hat{\mathbf{y}}[n]-\mathbf{x}[n])^2\right] \\ &= \mathrm{E}\left[((g[n]-\delta[n])*\mathbf{x}[n]+g[n]*\mathbf{w}[n]+\mathbf{v}[n])^2\right] \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big[|G(\omega)-1|^2\,\Phi_{xx}(\omega) \\ &\qquad +|G(\omega)|^2\,\Phi_{ww}(\omega)+\Phi_{vv}(\omega)\Big]\,d\omega \end{aligned} \tag{22}$$

and

$$\begin{aligned} r &= \mathrm{E}\,[\hat{\mathbf{y}}[n]\mathbf{w}[n]] = \mathrm{E}\,\{[g[n]*(\mathbf{x}[n]+\mathbf{w}[n])+\mathbf{v}[n]]\,\mathbf{w}[n]\} \\ &= \mathrm{E}\left\{\left[\sum_k g[k]\mathbf{x}[n-k]+\sum_k g[k]\mathbf{w}[n-k]+\mathbf{v}[n]\right]\mathbf{w}[n]\right\} \\ &= \sum_k g[k]R_{ww}[k] = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(\omega)\Phi_{ww}(\omega)\,d\omega. \end{aligned} \tag{23}$$

Write $G(\omega)$ in magnitude-phase form,

$$G(\omega) = |G(\omega)|\cos\phi(\omega) + j|G(\omega)|\sin\phi(\omega).$$

Then (22) and (23) give, respectively,

$$\begin{aligned} D(\hat{\mathbf{y}},\mathbf{x}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}&\left[\left(|G(\omega)|^2-2|G(\omega)|\cos\phi(\omega)+1\right)\right. \\ &\left.\times\,\Phi_{xx}(\omega)+|G(\omega)|^2\Phi_{ww}(\omega)+\Phi_{vv}(\omega)\right]\,d\omega, \end{aligned} \tag{24}$$

and

$$r = \frac{1}{2\pi}\int_{-\pi}^{\pi}|G(\omega)|\cos\phi(\omega)\,\Phi_{ww}(\omega)\,d\omega. \tag{25}$$

Then construct the Lagrangian J from the integrands of (24) and (25):

$$\begin{aligned} J = &\left[\left(|G(\omega)|^2-2|G(\omega)|\cos\phi(\omega)+1\right)\Phi_{xx}(\omega)\right. \\ &\left.+|G(\omega)|^2\Phi_{ww}(\omega)+\Phi_{vv}(\omega)\right] \\ &-\lambda|G(\omega)|\cos\phi(\omega)\,\Phi_{ww}(\omega). \end{aligned}$$

An extremum occurs when $\partial J/\partial\phi(\omega) = \partial J/\partial|G(\omega)| = \partial J/\partial\Phi_{vv}(\omega) = 0$. First, $\partial J/\partial\Phi_{vv}(\omega) = 0$ if and only if $\Phi_{vv}(\omega)=0, \forall\,\omega$. Second, $\partial J/\partial\phi(\omega) = [2\Phi_{xx}(\omega)+\lambda\Phi_{ww}(\omega)]\,|G(\omega)|\sin\phi(\omega) = 0$. Hence, $\phi(\omega) = \pm k\pi$, so $\sin\phi(\omega) = 0$ and $\cos\phi(\omega) = \pm 1$, so $G(\omega)$ is real. Third, $\partial J/\partial|G(\omega)| = 0$ yields $|G(\omega)| = \frac{\Phi_{xx}(\omega)+(\lambda/2)\Phi_{ww}(\omega)}{\Phi_{xx}(\omega)+\Phi_{ww}(\omega)}\cos\phi(\omega)$. Then $G(\omega) = |G(\omega)|\cos\phi(\omega) = \frac{\Phi_{xx}(\omega)+(\lambda/2)\Phi_{ww}(\omega)}{\Phi_{xx}(\omega)+\Phi_{ww}(\omega)}\,(\pm 1)^2$.

Finally, define $H(\omega)$ as in (8), and let $\gamma = 1-\lambda/2$; (7) follows. The solution can be shown to be a global minimum using proof-by-contradiction; the proof is omitted to conserve space.

B. Expressions for r, $D(\hat{\mathbf{y}},\,\mathbf{x})$, and E

Let $h[n]$ denote the impulse response that corresponds to $H(\omega)$, and let $\hat{\mathbf{w}}[n] = h[n]*(\mathbf{x}[n]+\mathbf{w}[n])$. Then

$$\sigma_{\hat{w}}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}|H(\omega)|^2\,(\Phi_{xx}(\omega)+\Phi_{ww}(\omega))\,d\omega.$$

Substitution of (8) for $H(\omega)$ gives (12). Also

$$\begin{aligned} R_{\hat{w}w}[0] &= \mathrm{E}\,[\hat{\mathbf{w}}[n]\mathbf{w}[n]] \\ &= \mathrm{E}\left[\sum_k h[k](\mathbf{x}[n-k]+\mathbf{w}[n-k])\mathbf{w}[n]\right] \\ &= \sum_k h[k]R_{ww}[k] = \frac{1}{2\pi}\int_{-\pi}^{\pi}H(\omega)\Phi_{ww}(\omega)\,d\omega = \sigma_{\hat{w}}^2. \end{aligned}$$

Next

$$r = \mathrm{E}\,[(\mathbf{x}[n]+\mathbf{w}[n]-\gamma\hat{\mathbf{w}}[n]+\mathbf{v}[n])\,\mathbf{w}[n]] = \sigma_w^2-\gamma R_{\hat{w}w}[0]$$

which gives (9). Similarly

$$\begin{aligned} D(\hat{\mathbf{y}},\mathbf{x}) &= \mathrm{E}\left[(\mathbf{x}[n]+\mathbf{w}[n]-\gamma\hat{\mathbf{w}}[n]+\mathbf{v}[n]-\mathbf{x}[n])^2\right] \\ &= \sigma_w^2-2\gamma R_{\hat{w}w}[0]+\gamma^2\sigma_{\hat{w}}^2+\sigma_v^2 \end{aligned}$$

which leads to (10) since $\sigma_v^2 = 0$. Then let $E = \mathrm{E}[(\mathbf{w}[n]-\hat{\mathbf{w}}[n])^2]$, so $E = \sigma_w^2-2R_{\hat{w}w}[0]+\sigma_{\hat{w}}^2$, and (11) follows.

Lastly, setting $r = r_0$ in (9) yields (13), and substituting (13) into (10) gives (14).

REFERENCES

[1] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proc. IEEE Int. Conf. Image Processing*, 1994.

[2] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking techniques," *Proc. IEEE*, vol. 86, pp. 1064–1087, Jun. 1998.

[3] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," NEC Res. Inst., Princeton, NJ, Tech. Rep. 95-10, Oct. 1995.

[4] F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Process.*, vol. 66, pp. 283–301, 1998.

[5] A. Piva, M. Barni, F. Bartolini, and V. Cappellini, "DCT-based watermark recovering without resorting to the uncorrupted original image," in *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, Oct. 1997, pp. 520–523.

[6] T. Kalker, G. Depovere, J. Haitsma, and M. J. Maes, "A video watermarking system for broadcast monitoring," in *Proc. SPIE, Security & Watermarking Multimedia Contents*, vol. 3657, San Jose, CA, Jan. 1999, pp. 103–112.

[7] M. Kutter and F. A. P. Petitcolas, "A fair benchmark for image watermarking systems," in *Proc. SPIE, Security & Watermarking Multimedia Contents*, vol. 3657, San Jose, CA, Jan. 1999, pp. 226–239.

[8] F. Hartung, J. K. Su, and B. Girod, "Spread spectrum watermarking: Malicious attacks and counterattacks," in *Proc. SPIE, Security & Watermarking Multimedia Contents*, vol. 3657, San Jose, CA, Jan. 1999, pp. 147–158.

[9] C.-T. Hsu and J.-L. Wu, "Hidden signatures in images," in *Proc. IEEE Int. Conf. Image Procesing*, vol. 3, Lausanne, Switzerland, Sep. 1996, pp. 223–226.

[10] X.-G. Xia, C. G. Boncelet, and G. R. Arce, "A multiresolution watermark for digital images," in *Proc. IEEE Int. Conf. Image Processing.*, Santa Barbara, CA, Oct. 1997, pp. 548–551.

[11] W. Zhu, Z. Xiong, and Y.-Q. Zhang, "Multiresolution watermarking for images and video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 545–550, Jun. 1999.

[12] J. Fridrich and M. Goljan, "Comparing robustness of watermarking techniques," in *Proc. SPIE, Security & Watermarking Multimedia Contents*, vol. 3657, San Jose, CA, Jan. 1999, pp. 214–225.

[13] J. K. Su and B. Girod, "On the imperceptibility and robustness of digital fingerprints," in *Proc. IEEE Int. Conf. Multimedia Computer Systems*, vol. 2, Florence, Italy, June 1999, pp. 530–535.

[14] J. K. Su and B. Girod, "Power spectrum condition for energy-efficient watermarking," in *Proc. IEEE Int. Conf. Image Processing*, Kobe, Japan, Oct. 1999.

[15] R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, "Theory of spread spectrum communications—A tutorial," *IEEE Trans. Commun.*, vol. COM-30, pp. 855–884, May 1982.

[16] P. G. Flikkema, "Spread-spectrum techniques for wireless communications," *IEEE Signal Processing Mag.*, vol. 14, pp. 26–36, May 1997.

[17] J. R. Smith and B. O. Comiskey, "Modulation and information hiding in images," in *Proc. First Info. Hiding Workshop*, vol. 1174, May 1996.

[18] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding,", preprint, Sep. 1999.

[19] J. K. Su, J. J. Eggers, and B. Girod, "Analysis of digital watermarks subjected to optimum linear filtering and additive noise," *Signal Process.*, vol. 81, pp. 1141–1175, 2001. Special issue on information theoretic issues in digital watermarking.

[20] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, "Generalized watermark attack based on stochastic watermark estimation and perceptual remodulation," in *Proc. SPIE, Security & Watermarking Multimedia Contents II*, vol. 3971, San Jose, CA, Jan. 2000, pp. 358–370.

[21] G. C. Langelaar, R. L. Lagendijk, and J. Biemond, "Removing spatial spread spectrum watermarks by nonlinear filtering," in *Proc. EUSIPCO'98*, vol. 4, 1998, pp. 2281–2284.

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

[23] M. H. Hayes, *Statistical Digital Signal Processing and Modeling.* New York: Wiley, 1996.

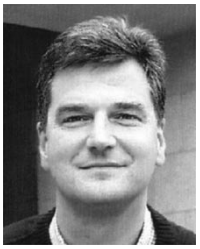
[24] D. Tzovaras, N. Karagiannis, and M. G. Strintzis, "Robust image watermarking in the subband or discrete cosine transform domain," in *Proc. EUSIPCO'98*, vol. 4, 1998, pp. 2285–2288.

[25] C. I. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 525–539, May 1998.

Jonathan K. Su (S'91–M'98) received the B.S.E.E. degree (magna cum laude) from Rice University, Houston, TX, in 1991, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, in 1993 and 1997, respectively.

In 2000, he joined the Technical Staff at the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington. From 1998 to 2000, he served as a Postdoctoral Fellow with the Telecommunications Institute I at the University of Erlangen-Nuremberg, Erlangen, Germany, where he conducted research in the field of digital watermarking. His research interests include digital image and video compression, detection and estimation, digital watermarking, and hyperspectral image processing. He is co-author (with Edward W. Kamen) of *Introduction to Optimal Estimation* (Berlin, Germany: Springer-Verlag, 1999).

Dr. Su served as co-chair of the 1999 V3D2 Watermarking Workshop held at the Institute and co-presented (with Ton Kalker) the tutorial on digital watermarking at the 2000 IEEE International Conference on Image Processing, Vancouver, BC, Canada. He is a member of Phi Beta Kappa, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.



Bernd Girod (F'98) received the M.S. degree in electrical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1980 and the Doctoral degree (with highest honors) from the University of Hannover, Germany, in 1987.

He is Professor of electrical engineering with the Information Systems Laboratory, Stanford University, California. He also holds a courtesy appointment with the Stanford Department of Computer Science. His research interests include networked multimedia systems, video signal compression, and 3-D image analysis and synthesis. Until 1987, he was Member of the Research Staff, Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, University of Hannover, working on moving image coding, human visual perception, and information theory. In 1988, he joined the Massachusetts Institute of Technology, Cambridge, first as a Visiting Scientist with the Research Laboratory of Electronics, then as an Assistant Professor of Media Technology at the Media Laboratory. From 1990 to 1993, he was Professor of computer graphics and Technical Director of the Academy of Media Arts, Cologne, Germany, jointly appointed with the Computer Science Section of Cologne University. He was a Visiting Adjunct Professor with the Digital Signal Processing Group at Georgia Tech in 1993. From 1993 until 1999, he was Chaired Professor of electrical engineering/telecommunications at the University of Erlangen-Nuremberg, Germany, and the Head of the Telecommunications Institute I, codirecting the Telecommunications Laboratory. He has served as the Chairman of the Electrical Engineering Department from 1995 to 1997, and as Director of the Center of Excellence "3-D Image Analysis and Synthesis" from 1995–1999. He has been a Visiting Professor with the Information Systems Laboratory, Stanford University, during the 1997–1998 academic year. As an entrepreneur, he has worked successfully with several start-up ventures as founder, investor, director, or advisor. Most notably, he has been a founder and Chief Scientist of Vivo Software, Inc., Waltham, MA (1993–1998); after Vivo's aquisition, since 1998, Chief Scientist of RealNetworks, Inc. (Nasdaq: RNWK); and, since 1996, an outside Director of 8×8, Inc. (Nasdaq: EGHT). He has authored or co-authored one major textbook and over 200 book chapters, journal articles, and conference papers in his field, and he holds about 20 international patents. He is a member of the editorial board of *EURASIP Signal Processing*.

Prof. Girod has served as on the editorial boards or as associate editor for several journals in his field, and is currently Area Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, as well as member of the editorial boards of *IEEE Signal Processing Magazine* and the *ACM Mobile Computing and Communication Review*. He has chaired the 1990 SPIE conference on "Sensing and Reconstruction of Three-Dimensional Objects and Scenes" in Santa Clara, CA, and the German Multimedia Conferences in Münich in 1993 and 1994, and has served as Tutorial Chair of ICASSP'97 in Münich and as General Chair of the 1998 IEEE Image and Multidimensional Signal Processing Workshop in Alpbach, Austria. He has been the Tutorial Chair of ICIP-2000 in Vancouver, BC, Canada, and the General Chair of the Visual Communication and Image Processing Conference (VCIP) in San Jose, CA, in 2001. He has been a member of the IEEE Image and Multidimensional Signal Processing Committee from 1989 to 1997 and was elected Fellow of the IEEE in 1998 "for his contributions to the theory and practice of video communications." He has been named "Distinguished Lecturer" for the year 2002 by the IEEE Signal Processing Society.