34th Asilomar Conference
Signals, Systems, Computers
Oct. 29-Nov. 1, 2000
Asilomar, CA, USA

Optimum Attack on Digital Watermarks and its Defense

Jonathan K. Su
MIT Lincoln Laboratory
Lexington, MA, USA
su@ll.mit.edu/su@LNT.de

Joachim J. Eggers
Telecomm. Laboratory
Univ. Erlangen-Nuremberg
Erlangen, Germany
eggers@LNT.de

Bernd Girod
Information Systems Lab
Stanford University
Stanford, CA, USA
girod@ee.stanford.edu

Abstract

We study the theoretical robustness of digital watermarks by viewing watermarking as communication over a hostile channel. Signals are modeled as stationary Gaussian random processes, and distortion is measured by frequency-weighted mean-squared error (MSE). The attack consists of linear shift-invariant filtering and additive Gaussian noise; it is optimized by selecting the filter and noise to minimize attacked-signal distortion under a capacity constraint. Then the defense is optimized by maximizing attacked-signal distortion under constraints on capacity and watermarked-signal distortion. We obtain performance limits and give rules-of-thumb for attack and defense. Experiments also show the influence of memory, suboptimality of additive-noise and effective white-noise attacks, and the effect of frequency-weighted distortion.

1. Introduction

Digital watermarking is the imperceptible, robust, secure communication of information in which information is embedded in and retrieved from other digital media (e.g., audio, images). This paper focuses on imperceptibility and robustness. *Imperceptibility* means the watermarked signal should be perceptually indistinguishable from the original, unwatermarked signal, and *robustness* means the information conveyed by the watermark should be reliably decodable even after processing of the watermarked signal.

An *attack* is any processing of the watermarked signal, and the processed signal is called the *attacked signal*. The attacked signal must be of sufficient perceptual quality to remain useful or valuable. The concept of robustness is intuitively clear but difficult to quantify. A watermark is said to be *robust* "if an attack cannot prevent communication of the embedded information without also rendering the attacked signal useless."

This paper takes a theoretical approach to watermarking. In the spirit of [6], we treat watermarking as a game between the owner and attacker. We also employ Kerckhoff's principle and assume that each player has complete knowledge of the other player's methods. In this way, we are able to derive performance limits and guidelines for both powerful attacks and robust watermarks.

2. Watermarking and Attack Model

We view watermarking as communication over a hostile channel and treat signals as M-dimensional (M-D), zero-mean, stationary Gaussian random processes. We will employ integrals over the M-D baseband $\Omega = [-\pi, \pi)^M$. The original signal is $\mathbf{x}[\vec{n}]$ and has power spectrum $\Phi_{xx}(\vec{\omega})$, and the watermark is $\mathbf{w}[\vec{n}]$ and has power spectrum $\Phi_{ww}(\vec{\omega})$. Let $\mathcal{X}$ and $\mathcal{W}$ denote the respective frequency supports $\Phi_{xx}(\vec{\omega})$ and $\Phi_{ww}(\vec{\omega})$. $\mathbf{x}[\vec{n}]$ and $\mathbf{w}[\vec{n}]$ are assumed independent, and the watermarked signal $\mathbf{y}[\vec{n}]$ is simply

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

$\mathbf{y}[\vec{n}]$ is sent over the channel, where it is attacked. We model the attack by linear shift-invariant (LSI) filtering and additive colored Gaussian noise (ACGN). The filter has impulse response $g[\vec{n}]$ and transfer function $G(\vec{\omega})$; the noise is $\mathbf{v}[\vec{n}]$ with power spectrum $\Phi_{vv}(\vec{\omega})$ and is independent of $\mathbf{x}[\vec{n}]$ and $\mathbf{w}[\vec{n}]$. The attacked signal is $\hat{\mathbf{y}}[\vec{n}]$,

$$\hat{\mathbf{y}}[\vec{n}] = g[\vec{n}] * \mathbf{y}[\vec{n}] + \mathbf{v}[\vec{n}], \tag{2}$$

where $*$ denotes M-D convolution.

We always assume synchronization between the encoder and decoder in our approach, since loss of synchronization does not remove the watermark signal but only makes it more difficult to locate. A more sophisticated receiver should be able to regain synchronization [4, 3].

2.1. Measuring Distortion

In watermarking, the perceptual quality or distortion of the watermarked signal and the attacked signal is important;

both signals should have acceptably low levels of distortion to remain useful to the owner or attacker. We measure distortion by the frequency-weighted *mean-squared error* (MSE) measured relative to the original signal $\mathbf{x}[\vec{n}]$.

The *embedding distortion* D_{yx} is then

$$D_{yx} = \mathrm{E}\left[\left[f[\vec{n}] * (\mathbf{y}[\vec{n}] - \mathbf{x}[\vec{n}])\right]^2\right] \quad (3)$$

$$= \frac{1}{(2\pi)^M} \int_\Omega |F(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega})\, d\vec{\omega}, \quad (4)$$

where $f[\vec{n}]$ is the impulse response of a LSI frequency-weighting filter, and $|F(\vec{\omega})|$ is the magnitude response of the filter. Ordinary MSE distortion corresponds to $f[\vec{n}] = \delta[\vec{n}]$ and $|F(\vec{\omega})| \equiv 1$.

For the *attack distortion* $D_{\hat{y}x}$, we use (2) and find

$$D_{\hat{y}x} = \frac{1}{(2\pi)^M} \int_\Omega |F(\vec{\omega})|^2 \Big[|G(\vec{\omega}) - 1|^2 \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})\Big]\, d\vec{\omega}. \quad (5)$$

Finally, define the *perceptual power* of the original by $P_x = (2\pi)^{-M} \int_\Omega |F(\vec{\omega})|^2 \Phi_{xx}(\vec{\omega})\, d\vec{\omega}$. For MSE distortion, $P_x = \sigma_x^2$.

2.2. Measuring Capacity

When $\mathbf{x}[\vec{n}]$ is available to the receiver, it does not interfere with communication of information via the watermark $\mathbf{w}[\vec{n}]$; we call this case *reception-with-original*. When the original signal $\mathbf{x}[\vec{n}]$ is not available at the receiver, we have *blind watermarking*. In conventional blind watermarking, $\mathbf{x}[\vec{n}]$ is treated as an additional source of interference. However, $\mathbf{x}[\vec{n}]$ is completely known during watermark embedding, which thus corresponds to channel coding with side information at the encoder. Costa [1] showed that, for the Gaussian case, it is possible to construct a blind coding scheme such that there is no interference from $\mathbf{x}[\vec{n}]$.

Applying Kerckhoff's principle, we assume the receiver has complete knowledge of $g[\vec{n}]$ and $G(\vec{\omega})$ and compensates for the filter's effects. We therefore write the *effective received signal* $\mathbf{z}[\vec{n}]$ as

$$\mathbf{z}[\vec{n}] = \hat{\mathbf{y}}[\vec{n}] - ag[\vec{n}] * \mathbf{x}[\vec{n}]. \quad (6)$$

The factor a, $0 \le a \le 1$, is the *original-interference suppression factor*. The case $a = 0$ corresponds to conventional blind watermarking, while $a = 1$ corresponds to reception-with-original or optimal blind watermarking. Intermediate values of a reflect the fact that a practical blind watermarking scheme may still suffer from some interference from $\mathbf{x}[\vec{n}]$. Note that $\mathbf{x}[\vec{n}]$ may not actually be available to the receiver, but the watermarking system performs as if it operated on $\mathbf{z}[\vec{n}]$.

As a result, the capacity may be written as [5]

$$C = \frac{1}{(2\pi)^M} \int_\Omega \frac{1}{2} \log_2 \left(1 + \frac{|G|^2 \Phi_{ww}}{(1-a)^2 |G|^2 \Phi_{xx} + \Phi_{vv}}\right) d\vec{\omega}, \quad (7)$$

where dependence of G, Φ_{xx}, Φ_{ww}, and Φ_{vv} on $\vec{\omega}$ is omitted to conserve space.

2.3. Well-Defined Robustness Criterion for Attack and Defense

Referring back to the intuitive definition of robustness in Sec. 1, we see that C measures a watermark's ability to communicate and $D_{\hat{y}x}$ measures the usefulness of the attacked signal. We can now evaluate robustness in a well-defined way: Given two (or more) watermarks with the same original-signal power spectrum $\Phi_{xx}(\vec{\omega})$ and the same values of D_{yx} and $D_{\hat{y}x}$, the watermark with higher (highest) capacity is more (most) robust.

Using this criterion, we define the attacker's and owner's problems precisely. **Attacker:** Given $\Phi_{ww}(\vec{\omega})$ and a target capacity $C_t \ge 0$, choose $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to minimize $D_{\hat{y}x}$ such that $C = C_t$. **Owner:** Given $G(\vec{\omega})$, $\Phi_{vv}(\vec{\omega})$, C_t, and embedding distortion D_{embed}, choose $\Phi_{ww}(\vec{\omega})$ to maximize $D_{\hat{y}x}$ such that $C = C_t$ and $D_{yx} \le D_{\text{embed}}$.

3. (Suboptimal) Effective White-Noise Attack

Before proceeding with the optimum attack, we consider a suboptimal but intuitively pleasing attack. Consider the ACGN channel with input $\mathbf{w}[\vec{n}]$, noise $\mathbf{n}_i[\vec{n}]$, and output $\mathbf{z}'[\vec{n}] = \mathbf{w}[\vec{n}] + \mathbf{n}_i[\vec{n}]$. Traditionally, $\Phi_{n_i n_i}(\vec{\omega})$ is fixed, and the signal (i.e., watermark) power spectrum $\Phi_{ww}(\vec{\omega})$ is chosen to maximize the communication rate. The solution for $\Phi_{ww}(\vec{\omega})$ is a water-filling rule that gives the signal a power advantage over the noise [2]; it can also be shown that communication is most difficult when the noise is white and Gaussian [7].

An *effective white-noise attack* based on this idea was studied in [9]. The attack selects $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ so that $\Phi_{n_i n_i}(\vec{\omega}) \propto \Phi_{ww}(\vec{\omega})$. The resulting channel is equivalent to the AWGN channel.

It can also be proven [9] that the optimum defense against this attack results when

$$\Phi_{ww}(\vec{\omega}) = \frac{\sigma_w^2}{\sigma_x^2} \Phi_{xx}(\vec{\omega}). \quad (8)$$

We call (8) the *power-spectrum condition* (PSC). A watermark that satisfies the PSC is said to be *PSC-compliant*. The capacity in this case is [8]

$$C = \frac{1}{2} \log_2 \left(1 + \frac{(P_x - D_{\hat{y}x})\, D_{\text{embed}}}{P_x^2 - (P_x - D_{\hat{y}x})\,(a(2-a)P_x + D_{\text{embed}})}\right). \quad (9)$$

4. Optimum Attack

Contrary to a conventional channel, however, in watermarking $\Phi_{ww}(\vec{\omega})$ is fixed, and then the attack chooses the filter and noise. Hence, the attacker, rather than the owner, has a potential power advantage. It was shown in [9] that the effective white-noise attack is indeed suboptimal. The optimum attack can be derived [8, 9] using the calculus of variations with Lagrangian cost function $J = (\text{integrand of } D_{\hat{y}x}) + \lambda(\text{integrand of } C)$.

The solutions for $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ are

$$G(\vec{\omega}) = A(\vec{\omega})\,\frac{\Phi_{xx}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}, \tag{10}$$

$$\Phi_{vv}(\vec{\omega}) = (1 - A(\vec{\omega}))\,A(\vec{\omega})\,\frac{\Phi_{xx}^2(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}, \tag{11}$$

where $0 \le A(\vec{\omega}) \le 1, \forall\vec{\omega}$.

For $\Phi_{xx}(\vec{\omega}) = 0$ or $\Phi_{ww}(\vec{\omega}) = 0$, $A(\vec{\omega}) = 1$, so the attack leaves such frequency components unchanged. The following equations assume that $\Phi_{xx}(\vec{\omega}) > 0$ and $\Phi_{ww}(\vec{\omega}) > 0$. Define $\text{cl}[x]$ to be the function that clips x to the interval $[0, 1]$. For $a = 0$ (conventional blind reception),

$$A(\vec{\omega}) = \text{cl}\left[1 + \frac{\Phi_{xx}(\vec{\omega})}{\Phi_{ww}(\vec{\omega})} - \frac{\lambda}{2\ln 2}\,\frac{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}{\Phi_{xx}^2(\vec{\omega})|F(\vec{\omega})|^2}\right]. \tag{12}$$

For $0 < a \le 1$, $A(\vec{\omega})$ is given by

$$A(\vec{\omega}) = \text{cl}\left[\left(1 + \frac{\Phi_{ww}(\vec{\omega})\Phi_{xx}(\vec{\omega}) - \sqrt{P(\vec{\omega})}}{2a(2-a)\Phi_{xx}^2(\vec{\omega})}\right) Q(\vec{\omega})\right], \tag{13}$$

where $P(\vec{\omega}) = \Phi_{xx}^2(\vec{\omega})\Phi_{ww}^2(\vec{\omega}) + (2\lambda/\ln 2)a(2-a)\Phi_{xx}(\vec{\omega})\,\Phi_{ww}(\vec{\omega})\,(a(2-a)\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega}))\,|F(\vec{\omega})|^{-2}$, and $Q(\vec{\omega}) = (\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega}))\,/\,(a(2-a)\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega}))$. If $a = 1$ (reception-with-original/optimal blind reception), then $Q(\vec{\omega}) = 1$.

The attack distortion $D_{\hat{y}x}$ and capacity C after this attack can now be written, respectively, as

$$D_{\hat{y}x} = P_x - \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2\,\frac{A(\vec{\omega})\Phi_{xx}^2(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}\,d\vec{\omega}, \tag{14}$$

$$C = \frac{1}{(2\pi)^M}\int_\Omega \frac{1}{2}\log_2\left(1 + \frac{A(\vec{\omega})\Phi_{ww}(\vec{\omega})}{B(\vec{\omega})}\right) d\vec{\omega}, \tag{15}$$

where $B(\vec{\omega}) = \Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})A(\vec{\omega})\big(a(2-a)\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\big)$.

4.1. Capacity and Distortion Relationship

The expressions for $A(\vec{\omega})$ are complicated, but $A(\vec{\omega})$ is completely parameterized by the Lagrange multiplier $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. The limits are always finite and given in [8, 9]. For any $\vec{\omega} \in \mathcal{W}$, $A(\vec{\omega})$ decreases monotonically from unity to zero as λ increases.

Consequently, $D_{\hat{y}x}$ is a strictly increasing function of λ, and C is a strictly decreasing function of λ. We denote these dependencies by $D_{\hat{y}x}(\lambda)$ and $C(\lambda)$, respectively. Thus, λ controls the trade-off between capacity and attack distortion. Sweeping λ from $\lambda_{\min}$ to $\lambda_{\max}$ reveals the complete performance range for a given watermark power spectrum $\Phi_{ww}(\vec{\omega})$. Define the *distortion-capacity function* by $\{(D_{\hat{y}x}(\lambda), C(\lambda)) : \lambda_{\min} \le \lambda \le \lambda_{\max}\}$.

4.2. Attack Behavior

We now characterize the behavior of the optimum attack. To satisfy the imperceptibility requirement, we may assume $\Phi_{xx}(\vec{\omega}) \gg \Phi_{ww}(\vec{\omega})$, $\forall\vec{\omega}$. Then

$$G(\vec{\omega}) \approx A(\vec{\omega}), \tag{16}$$

$$\Phi_{vv}(\vec{\omega}) \approx (1 - A(\vec{\omega}))A(\vec{\omega})\Phi_{xx}(\vec{\omega}). \tag{17}$$

As λ increases, $G(\vec{\omega})$ decreases from nearly unity to zero, and $\Phi_{vv}(\vec{\omega})$ first increases from zero to $\frac{1}{4}\Phi_{xx}(\vec{\omega})$ and then decreases back to zero. For small λ, $G(\vec{\omega}) \approx 1$ and $\Phi_{vv}(\vec{\omega}) > 0$: the attack primarily adds noise. For large λ, $G(\vec{\omega}) \to 0$ and $\Phi_{vv}(\vec{\omega}) \to 0$: the attack discards entire frequency components. Thus, the attack suggests the following **rule-of-thumb:** *At low distortions, add noise; at high distortions, throw away frequency components.*

If $\lambda = \lambda_{\min}$, then $A(\vec{\omega}) \equiv 1$. The attack becomes equivalent to MAP/MMSE estimation of $\mathbf{x}[\vec{n}]$ from $\mathbf{y}[\vec{n}]$ and minimizes $D_{\hat{y}x}$. However, $\Phi_{vv}(\vec{\omega}) \equiv 0$, so C is maximized. If $\lambda = \lambda_{\max}$, then $A(\vec{\omega}) = 0$, $\vec{\omega} \in \mathcal{W}$, so $C = 0$. If, in addition, $\mathcal{X} \subseteq \mathcal{W}$, then $D_{\hat{y}x} = P_x$, meaning the attack distortion is as large as the perceptual power of $\mathbf{x}[\vec{n}]$.

5. Optimized Defense

Finding the best defense against the optimum attack is difficult because of the complicated expressions for $A(\vec{\omega})$, $G(\vec{\omega})$, and $\Phi_{vv}(\vec{\omega})$. It is unlikely that a closed-form expression for $\Phi_{ww}(\vec{\omega})$ can be calculated, so we have used numerical methods. The watermark power spectrum $\Phi_{ww}(\vec{\omega})$ is divided into piecewise-constant frequency bands, and iterative methods based on greedy marginal analysis ("GMA") and simulated annealing (two methods called "GA/normal" and "GA/scaled") are applied [8].

During each iteration, $\Phi_{ww}(\vec{\omega})$ is perturbed slightly, and the attack is re-optimized. Attack re-optimization can be performed efficiently because $C(\lambda)$ is a decreasing function of λ. We applied a bisection search to find λ^* such that $|C(\lambda^*) - C_t|/C_t < \varepsilon$. Once λ^* has been found, $D_{\hat{y}x}(\lambda^*)$ can be computed. When the perturbations no longer produce increases in $D_{\hat{y}x}$, the algorithms stop.

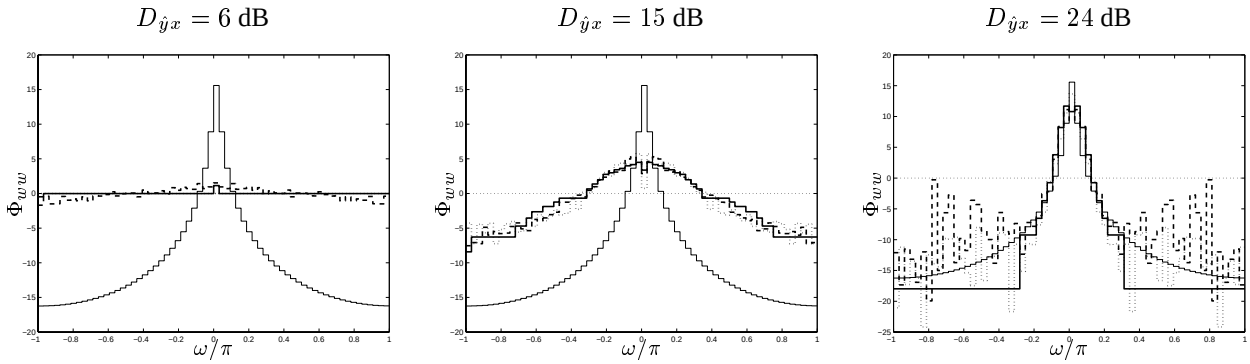

Figure 1. Example optimized watermark power spectra for selected attack distortions.

6. Experimental Results

Experiments were conducted using 1-D autoregressive (AR) processes for $\mathbf{x}[n]$ and $\mathbf{w}[n]$. AR processes are often used to model highly correlated digital signals such as audio, images, and video. We model the original signal as $\mathbf{x}[n] = a_1\mathbf{x}[n-1] + \mathbf{u}[n]$, where $a_1 = 0.95$, and $\mathbf{u}[n]$ is WGN. In all experiments, $\sigma_w^2 = 1$. Most experiments used MSE distortion and set $\sigma_x^2 = 10^3$. Then so $D_{yx} = \sigma_w^2 = 1$, and $10\log_{10}\sigma_w^2/\sigma_x^2 = -30$ dB. Due to space constraints, only results for the case $a = 1$ are shown. Similar qualitative behavior was observed for other values of a.

Note: We do *not* normalize distortion by P_x. In decibels, the *original signal-to-distortion ratio* (ODR) is ODR $= 10\log_{10} P_x/D_{\hat{y}x} = 10\log_{10} P_x - 10\log_{10} D_{\hat{y}x}$.

6.1. Rule-of-Thumb for Robustness

We first compare white ($a_1 = 0$), PSC-compliant ($a_1 = 0.95$), and optimized watermarks. Fig. 1 shows example optimized watermark power spectra as thick curves; thin curves show white (dotted) and PSC-compliant (solid) power spectra. Fig. 2 shows the distortion-capacity curves.

The results suggest a simple **rule-of-thumb:** *At low attack distortions, white watermarks have near-optimal robustness; at high distortions, PSC-compliant watermarks have near-optimal robustness*. This rule agrees with the description (Sec. 4.2) of the optimum attack: a white watermark resists additive noise well (since the noise power must be spread evenly over all frequency components), and a PSC-compliant watermark resists frequency-selective filtering well (since the attack cannot discard the frequency components where the original signal has large power).

6.2. Effect of Memory

The next experiment considers the effect of memory. The original signal $\mathbf{x}[\vec{n}]$ is modeled as being memoryless

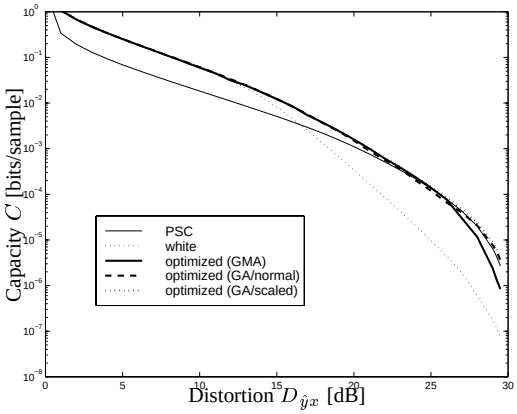

Figure 2. Robustness of different watermarks subject to optimum attack.

($a_1 = 0$) or highly correlated ($a_1 = 0.95$). Fig. 3 shows the resulting curves. They are nearly the same at low distortions because a white (memoryless) watermark is nearly optimal in this case. At high distortions, the capacity of a correlated original is much lower than that of the memoryless original. Against a correlated original, the attack can discard frequency components where $\Phi_{xx}(\vec{\omega})$ is small; this is impossible if $\mathbf{x}[\vec{n}]$ is white. Clearly, if the original signal is correlated, modeling it as being memoryless can lead to inaccurate performance predictions.

6.3. Suboptimal Attacks

The optimum attack is also compared with two suboptimal attacks. First, the frequently used *additive-noise attack*: $\hat{\mathbf{y}}[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}] + \mathbf{v}[\vec{n}]$, so $D_{\hat{y}x} = \sigma_w^2 + \sigma_v^2$. A white watermark resists the noise best, since the noise cannot gain a power advantage at any frequency. Then $C = \frac{1}{2}\log_2\left(1 + \sigma_w^2/(D_{\hat{y}x} - \sigma_w^2)\right)$. Second, the effective white-noise attack of Sec. 3. A PSC-compliant watermark provides the best defense against this attack, and the capac-

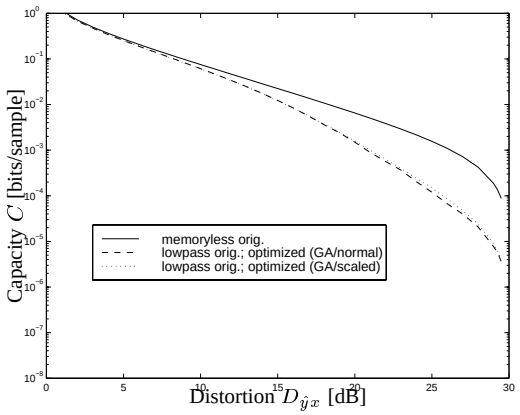

Figure 3. Robustness of memoryless and correlated originals.

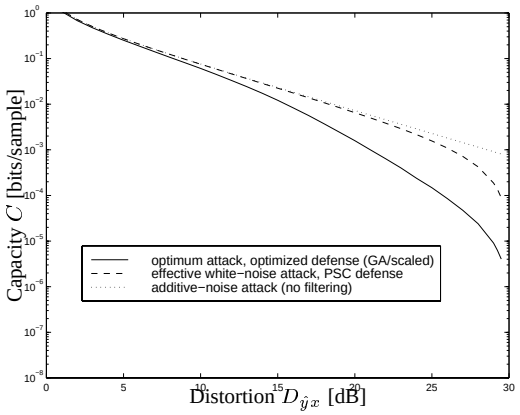

Figure 4. Robustness after different attacks.

ity is given by (9).

Fig. 4 compares the three attacks and clearly shows the suboptimality of the additive-noise and effective white-noise attacks. At low distortions, the additive-noise attack is a good approximation, but at high distortions, the optimum attack reduces capacity to 10–100 times less than that predicted by the suboptimal attacks.

6.4. Frequency-Weighted Distortion

Finally, experiments were also conducted for frequency-weighted MSE. We observed the following: At low distortions, $\Phi_{ww}(\omega)$ is such that

$$|F(\omega)|^2 \Phi_{ww}(\omega) \approx \text{constant}, \qquad (18)$$

which corresponds to a "perceptually white" power spectrum. At high distortions,

$$|F(\omega)|^2 \Phi_{ww}(\omega) \approx |F(\omega)|^2 (D_{\text{embed}}/P_x) \Phi_{xx}(\omega), \qquad (19)$$

where the right-hand side represents a frequency-weighted PSC-compliant power spectrum. Hence, the rule-of-thumb extends in a straightforward manner: *At low distortions, a "perceptually white" watermark performs nearly optimally; at high distortions, a "perceptually PSC-compliant" watermark performs nearly optimally.*

7. Conclusions

We have viewed watermarking as communications over a hostile channel and measured robustness by the attacked-signal distortion at a given capacity. We next treated watermarking as a game: for a given capacity, the attacker and owner try, respectively, to minimize or maximize the attack distortion. The results provide theoretical robustness limits.

Unlike a conventional channel, the attack adapts to the watermark, not vice versa. The optimum attack may be described as: *At low distortions, add noise; at high distortions, throw away frequency components.* Numerically optimized watermark power spectra agree with this characterization and lead to a rule-of-thumb: *At low distortions, white watermarks perform nearly optimally; at high distortions, PSC-compliant watermarks perform nearly optimally.* Additional experiments demonstrated the importance of memory, the suboptimality of some other attacks, and the extension of the rule-of-thumb to frequency-weighted distortion.

References

[1] M. H. M. Costa. Writing on dirty paper. *IEEE Trans. Inform. Theory*, IT-29:439–441, May 1983.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[3] G. Csurka, F. Deguillaume, J. J. K. Ó Ruandaidh, and T. Pun. A Bayesian approach to affine transformation resistant image and video watermarking. In *Prelim. Proc. 3rd Intl. Information Hiding Workshop*, Dresden, Germany, Sep.–Oct. 1999.

[4] F. Hartung, J. K. Su, and B. Girod. Spread spectrum watermarking: Malicious attacks and counterattacks. In *Proc. SPIE Security & Watermarking Multimedia Contents*, volume 3657, pages 147–158, San Jose, CA, USA, Jan. 1999.

[5] J. Huber. *Trelliscodierung*. Springer-Verlag, Berlin, Germany, 1992. In German.

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

[7] C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37:10–21, 1949.

[8] J. K. Su, J. J. Eggers, and B. Girod. Analysis of digital watermarks subjected to optimum linear filtering and additive noise. *Signal Processing*, to appear Spring 2001.

[9] J. K. Su and B. Girod. Fundamental performance limits of power-spectrum condition-compliant watermarks. In *Proc. SPIE Security & Watermarking Multimedia Contents II*, volume 3971, pages 314–325, San Jose, CA, USA, Jan. 2000.