Detection-Theoretic Analysis of Desynchronization Attacks in Watermarking

Pierre Moulin[1], Henrique Malvar

March, 2002

Technical Report
MSR-TR-2002-24

This paper studies the effects of desynchronization attacks such as delay and warping on the performance of watermark detection systems. First, we quantify the connection between attack channel estimation accuracy and detection performance. Second, we show how to design watermarks that minimize probability of error of the detector. Evaluation of the optimal likelihood ratio test is often computationally expensive, so as a practical alternative, we propose a family of quadratic detectors and construct the detector and family of watermarks that maximize the *deflection criterion*. For delay attacks, the deflection criterion is shown to increase quadratically with the duration of the host signal. For warping attacks, the deflection criterion increases proportionally to the duration of the signal and proportionally to the *coherence time* of the warping function. Finally, as an alternative to noncoherent watermark detection, we suggest the use of tracking techniques to estimate the warping function, followed by coherent detection.

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
`http://www.research.microsoft.com`

[1]University of Illinois, Beckman Inst., Coord. Sci. Lab, 405 N. Mathews Ave., Urbana, IL 61801; work performed in August and September 2001, while P. Moulin was a Visiting Researcher at Microsoft Research.

1 Introduction

Consider the problem of detecting a known watermark w originally embedded in a host signal s. The watermarked signal $x = s + w$ is subjected to attacks. The corrupted signal y is made available to the watermark detector, together with the reference watermark w.

Assume there is a list of possible attacks, each parameterized by some parameter $\theta \in \Theta$. For instance, consider

- addition of independent and identically distributed (i.i.d.) noise with probability density function p_θ; e.g., a Gaussian density function with mean zero and variance θ.
- compression using a particular algorithm with quality factor θ;
- delay of the watermarked signal by θ units of time;
- warping of the watermarked signal using a warping function (time-varying delay) $\theta(t)$;
- time-varying gain $\theta(t)$.

In these problems, θ is a scalar, a vector, or even a function.

While basic results of detection theory have been applied to watermarking [1], current watermarking literature does not provides satisfactory answers to complex but realistic problems such as those listed above. One approach is to use a heuristic detector (e.g., a simple correlator combined with an estimator of the unknown θ) and study its performance under a list of attacks. A more principled approach is to construct a detector that satisfies optimality properties under the same list of attacks. An attractive consequence of such an approach is that one could construct optimal watermarks. This is the approach undertaken in this paper. We focus on desynchronization attacks, but the theory is general enough to be applicable to a larger list of attacks.

2 Watermark Detection as a Composite Hypothesis Testing Problem

Statistical hypothesis testing provides a general approach to detection problems involving unknown parameters. In the absence of watermark, the received signal y is assumed to follow a particular probability distribution $p_0(y)$. In the presence of the watermark, y follows a distribution $p_\theta(y)$ which depends on the choice of the nuisance parameter $\theta \in \Theta$ by the attacker. Under these assumptions, the watermark detection problem may be formulated as a composite hypothesis test [2]:

$$\begin{cases} H_0 & : \ y \sim p_0 \\ H_1 & : \ y \sim p_\theta, \quad \theta \in \Theta. \end{cases} \tag{1}$$

Three classical techniques have been used in the detection literature to solve such problems:

1. *Bayesian approach*: a prior probability measure P is assumed over the attack channel parameter space Θ. Then integrating out θ yields a known distribution

$$p_1(y) = \int_\Theta p_\theta(y) dP(\theta). \tag{2}$$

where $dP(\theta) = \pi(\theta)\, d\theta$ if a density π exists. The Bayesian detection rule is a likelihood ratio test (LRT):

$$\frac{p_1(y)}{p_0(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3}$$

where η is the threshold of the test.

2. *Neyman-Pearson approach*: one seeks the detection test $\delta = \delta(y)$ that minimizes the probability of miss subject to a constraint on the maximum allowable probability of false alarm. The NP rule is a randomized LRT.

3. *Minimax approach*: one seeks the detection test $\delta = \delta(y)$ that minimizes $\max_\theta R(\theta, \delta)$, where $R(\theta, \delta)$ is the risk of δ conditioned on θ. The minimax rule is a randomized LRT.

4. *Generalized Likelihood Ratio Test (GLRT)*: one first estimates θ as $\hat{\theta}(y)$, and then applies the LRT

$$\frac{p_{\hat{\theta}}(y)}{p_0(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{4}$$

The first three tests have clear optimality properties but may be computationally intractable due to the need to integrate out θ in (2). The minimax approach is arguably realistic in the presence of an adversary. The GLRT has asymptotic optimality properties [3], may be computationally simple, and has been known to work reasonably well in some applications – and this even though the GLRT has generally no optimality properties for finite samples. The mean-value theorem provides a useful conceptual link between (3) and (4): if p_θ varies smoothly with θ, then given y, there exists $\tilde{\theta} \in \Theta$ such that $p_1(y) = p_{\tilde{\theta}}(y)$ [2]. In other words, the Bayes test (3) may be written in the form (4) for a particular estimator $\hat{\theta}(Y) = \tilde{\theta}(Y)$.

3 Warping Attacks

This paper focuses on a fairly challenging composite hypothesis testing problem in which the attack takes the form of a time warping of the watermarked signal. Such desynchronization attacks can disable empirically designed detectors [5]. We formulate the warping model either in a discretized or in a continuous time domain and use one or the other depending on which one is more convenient. For mathematical convenience, we assume that all signals are periodic with period equal to T in the continuous case and N in the discrete case. The host signal $s(t)$ is a periodic white Gaussian noise (WGN) process with covariance $R_s(t, t') = E[s(t)s(t')] = \sigma_S^2 \eth(t - t')$ for $t, t' \in \mathbb{R}$, where $\eth(t) = \frac{1}{T}\sum_{k\in\mathbb{Z}} \delta(t - kT)$ is an infinite train of Dirac impulses (also known as the *shah* function).

The white noise assumption for $s(t)$ may seem restrictive, but the white noise model is often applicable in a transform domain (e.g., wavelet coefficients, lapped orthogonal transform, etc.)

Two models are considered for data collection. In each case, the watermark $w(t), t \in [0, T]$ is a periodic and continuous function ($w(0) = w(T)$). The warping function is real-valued and is denoted by $\theta(t), t \in [0, T]$ for the continuous-time model, and by $\theta(n), n \in \{0, \cdots, N-1\}$ for the discrete-time model. [1]

Discrete-Time Data:

$$\begin{cases} H_0 & : \ y(n) = s(n) & , \ n \in \{0, 1, \cdots, N-1\} \\ H_1 & : \ y(n) = w(n - \theta(n)) + s(n - \theta(n)) & , \ n \in \{0, 1, \cdots, N-1\}. \end{cases} \tag{5}$$

Continuous-Time Data:

$$\begin{cases} H_0 & : \ y(t) = s(t) & , \ 0 \le t \le T \\ H_1 & : \ y(t) = w(t - \theta(t)) + s(t - \theta(t)) & , \ 0 \le t \le T. \end{cases} \tag{6}$$

The attack channel parameter is a slowly-varying sequence $\theta(n)$ (discrete case) or function $\theta(t)$ (continuous case). For instance, we may assume that the rate of variation of θ is no greater than some specified ϵ:

$$\Theta = \{\theta \ : \ |\theta(n) - \theta(n-1)| \le \epsilon\} \quad \text{(discrete case)} \tag{7}$$

$$\Theta = \{\theta \ : \ |\theta'(t)| \le \epsilon\} \quad \text{(continuous case)}. \tag{8}$$

For instance, in audio watermarking, we would typically have $\epsilon = 0.04$ [5]. If $\epsilon = 0$, the warping attack reduces to a fixed (but unknown) delay.

We would further like to assume that the statistics of $s(t)$ are indistinguishable from those of $s(t - \theta(t))$. Otherwise the host signal itself would serve as a synchronization signal, thereby helping the detector. Strictly speaking, this assumption is incompatible with our above assumption on the statistics of $s(t)$. Hence, to make the analysis tractable, we have decided to study the hypothesis test

$$\begin{cases} H_0 & : \ y(t) = s(t) & , \ 0 \le t \le T \\ H_1 & : \ y(t) = w(t - \theta(t)) + s(t) & , \ 0 \le t \le T, \end{cases} \tag{9}$$

which serves as an approximation to the original detection problem.

4 Relation to Communication Problems

The communication literature contains a rich variety of signal detection problems closely related to the watermarking problem. Detection of a known signal (without any unknown parameter θ) is a coherent detection problem. When signals undergo delays or time-varying delays (same as time warping [6]), the detection problem is said to be noncoherent [7, 2]. If partial information about the delay or time-varying delay is available, the detection problem is said to be partially coherent.

[1] To qualify as a warping function, θ should have the property that $t - \theta(t)$ is strictly increasing, i.e., $\theta'(t) > -1$ for all t. This condition is not imposed in our analysis, so we sometimes end up considering a broader class of attacks.

Much of the signal detection theory developed in the communication literature has been applied to narrowband signals. Noncoherent detection of wideband signals (such as in spread-spectrum applications) is much more elaborate, and techniques such as transmission of a known training sequence are often used to facilitate detection. Such techniques are not applicable to watermarking, so we shall develop solutions based on first principles rather than specific techniques from spread-spectrum communications.

5 Analysis of Delay Attacks

We first ask how well the warping function θ can be estimated, and what are the effects of estimation errors on detection performance. High estimation accuracy does not necessarily translate into high detection performance, because the sensitivity to estimation errors may be high.

Consider first the case of a simple delay $\theta \in [0, T]$. The likelihood functional for θ is [2]

$$l(\theta, y) = -\frac{1}{2\sigma_S^2} \int_0^T (y(t) - w(t-\theta))^2 \, dt.$$

5.1 Coherent Detector

If the delay θ is known, we have a coherent detection problem [2].

Define the *normalized, deterministic autocorrelation function* of the watermark as

$$R_w(t) = \frac{1}{T} \int_0^T w(t')w(t'+t) \, dt', \tag{10}$$

which has a maximum at $t = 0$. Observe that

$$R_w''(0) = \frac{1}{T} \int_0^T w(t')w''(t') \, dt' = \frac{1}{T} \int_0^T |w'(t')|^2 \, dt' \tag{11}$$

where the second equality is obtained using integration by parts.

Under our model assumptions, if θ is known, the LRT becomes a simple correlation test [2]:

$$c_\theta = \int_0^T y(t)w(t-\theta) \, dt \quad \begin{matrix} H_1 \\ \gtrless \\ H_0 \end{matrix} \quad \eta \tag{12}$$

where the *correlation statistic* c_θ has mean 0 and $TR_w(0)$ under H_0 and H_1, respectively, and has variance $\sigma_S^2 T R_w(0)$ under both H_0 and H_1. For Bayesian detection under equal priors on H_0 and H_1, the threshold of the LRT is $\eta = \frac{T}{2} R_w(0)$, and the probability of error is

$$P_e = Q\left(\frac{1}{2}\sqrt{SNR}\right). \tag{13}$$

Here we have defined $Q(u) = \int_u^\infty \phi(v) \, dv$, and $\phi(u) = (2\pi)^{-1/2} \exp\{-\frac{u^2}{2}\}$. Also

$$SNR \triangleq \frac{TR_w(0)}{\sigma_S^2}, \tag{14}$$

where the numerator represents total watermark energy. Note that detector performance (13) depends on the energy of the watermark and not on its spectral contents.

5.2 Estimation Accuracy

The Fisher information for estimation of θ is given by [2]

$$\begin{aligned} J(\theta) &= E_{Y|\theta}\left[\frac{\partial l(\theta, Y)}{\partial \theta}\right]^2 \\ &= \frac{1}{\sigma_S^2}\int_0^T |w'(t-\theta)|^2\, dt \\ &= \frac{1}{\sigma_S^2}\int_0^T |w'(t)|^2\, dt. \end{aligned} \tag{15}$$

Also (15) yields

$$J(\theta) = \frac{TR_w''(0)}{\sigma_S^2}. \tag{16}$$

The Fisher information is independent of θ in this case. It is inversely proportional to the noise variance σ_s^2 and proportional to the total energy $\|w'\|^2$ in the watermark derivative (typically increases proportionally to T). Hence high estimation accuracy might be achieved if T is large, or if $w(t)$ possesses significant high-frequency content. [2]

5.3 Mismatched Detector

If the correlation detector uses a *mismatched* value $\theta + \delta$ instead of the true θ, then the correlation statistic $c_{\theta+\delta}$ has means 0 and $TR_w(\delta)$ under H_0 and H_1, respectively, and variance $\sigma_S^2 TR_w(0)$ under both H_0 and H_1. If $\eta = \frac{T}{2}R_w(0)$ again (as would be the GLRT choice), then

$$P_e(\delta) = \frac{1}{2}\left[Q\left(\frac{\sqrt{TR_w(0)}}{2\sigma_S}\right) + Q\left(\frac{TR_w(\delta) - TR_w(0)/2}{\sigma_S\sqrt{TR_w(0)}}\right)\right]$$

which is necessarily greater than $P_e(0)$ in the matched case. If SNR is large and $R_w(\delta) << R_w(0)$, then $P_e(\delta) \approx \frac{1}{2}$, and the mismatched detector is effectively disabled.

The performance under small errors ($\delta \to 0$) is tractable and particularly insightful. We derive

$$\frac{dP_e(\delta)}{d\delta} = -\frac{T}{2}R_w'(\delta)\phi\left(\frac{TR_w(\delta) - TR_w(0)/2}{\sigma_S\sqrt{TR_w(0)}}\right)$$

$$\left.\frac{dP_e(\delta)}{d\delta}\right|_{\delta=0} = 0 \tag{17}$$

$$\left.\frac{d^2P_e(\delta)}{d\delta^2}\right|_{\delta=0} = -\frac{T}{2}R_w''(0)\frac{1}{\sigma_S\sqrt{TR_w(0)}}\phi\left(\frac{\sqrt{TR_w(0)}}{2\sigma_S}\right) > 0. \tag{18}$$

Equation (18) shows that detector sensitivity to mismatches is greatest when $R_w(t)$ has a narrow peak at $t = 0$. But this is precisely the condition required for accurate estimation of θ, see (16).

[2] Recall that the variance of any unbiased estimator of θ is lower bounded by $1/J(\theta)$, and that this bound is achievable by the maximum-likelihood estimator under standard asymptotic conditions (here $T \to \infty$).

Commonly-used pseudo-random white noise watermarks are in this category. Conversely, detectors for lowpass watermarks [5] have lower sensitivity to mismatch errors, but are more likely to make large estimation errors.

To complete the analysis, we must relate δ to our statistical model. If δ is the error of the maximum-likelihood estimator (MLE) of θ, then δ is random with mean zero and variance approaching $J^{-1}(\theta)$ (Fisher information) as $SNR \to \infty$ [2]. Using (16), (18) and (11), we obtain an increase in probability of error of

$$\begin{aligned} \Delta P_e &= \frac{1}{2} E\left[\delta^2 \left.\frac{d^2 P_e(\delta)}{d\delta^2}\right|_{\delta=0}\right] = \frac{1}{2} J^{-1}(\theta) \left.\frac{d^2 P_e(\delta)}{d\delta^2}\right|_{\delta=0} \\ &= \frac{1}{2} SNR^{-1/2} \phi\left(\frac{1}{2}\sqrt{SNR})\right) \sim \frac{1}{4} Q\left(\frac{1}{2}\sqrt{SNR}\right) \end{aligned} \tag{19}$$

where the last expression follows from the asymptotic relation $Q(u) \sim u^{-1}\phi(u)$ as $u \to \infty$. Observe that ΔP_e is independent of the spectral characteristics of w. The performance of the mismatched detector is somewhat worse than that of the matched detector, see (13). In particular, this conclusion applies to the GLRT.

5.4 Bayesian Detector

Assume that the unknown delay θ is random with a p.d.f. $\pi(\theta), \theta \in [0, T)$. Then from (2) and (3), we obtain the LRT

$$L(y) = \frac{1}{T}\int_0^T L(\theta, y)\pi(\theta)\, d\theta \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{20}$$

where

$$\begin{aligned} L(\theta, y) &= \exp\left\{\frac{1}{\sigma_S^2}\left[\int_0^T y(t)w(t-\theta)\, dt - \frac{R_w(0)}{2}\right]\right\} \\ &= \exp\left\{\frac{1}{\sigma_S^2}\left[c_\theta - \frac{R_w(0)}{2}\right]\right\}. \end{aligned}$$

The integral (20) is intractable, but asymptotic ($SNR \to \infty$) approximations can be derived based on Laplace's integral expansion technique [13, 14]. The general idea is that only a narrow range of values of θ contribute to the mixture integral (20), which may be approximated as [13, 14]

$$\hat{p}_1(y) = p_{\hat{\theta}}(y)\frac{\pi(\hat{\theta})}{\sqrt{J(\hat{\theta})}}$$

where $\hat{\theta}(y) = argmax_{\theta\in\Theta} p_\theta(y)$ is the maximum-likelihood estimate of θ. The LRT (3) takes the form

$$\frac{p_{\hat{\theta}}(y)}{p_0(y)} \underset{H_0}{\overset{H_1}{\lessgtr}} \eta\frac{\sqrt{J(\hat{\theta})}}{\pi(\hat{\theta})} = \eta\frac{T^{3/2}\sqrt{R_w''(0)}}{\sigma_S}.$$

For our delay problem, applying this technique to (20) assuming that $\pi(\theta)$ is the uniform distribution over $[0, T]$ yields the approximation

$$\hat{L}(y) = L(\hat{\theta}, y) \frac{\sigma_S}{T^{3/2}\|w'\|}.$$

The LRT takes a simplified form: it has the same structure as the GLRT, but uses a threshold $\eta' = \frac{TR_w(0)}{2} + \ln \frac{T^{3/2}\|w'\|}{\sigma_S}$:

$$c = \int_0^T y(t) w(t - \hat{\theta})\, dt \quad \begin{matrix} H_1 \\ \gtrless \\ H_0 \end{matrix} \quad \eta'. \tag{21}$$

We can then derive the asymptotic probability of error,

$$P_e \sim aQ\left(\frac{1}{2}\sqrt{SNR}\right) \quad \text{as } SNR \to \infty$$

where a is some constant larger than 1; compare with (13).

5.5 Quadratic Noncoherent Detector

The optimal Bayesian test (3) is expensive to implement due to the need to compute the correlation statistic c_θ for all values of θ. A suboptimal but often good approach consists of using a quadratic detection test [8]. The benefits of this approach in the context of detection of narrowband signals with drifting phase have been demonstrated in papers by Foschini *et al.* [9] and Veeravalli and Poor [10]. We first define the quadratic detection statistic and derive the deflection criterion which serves as a performance index for the detection test. The deflection criterion is then used to derive properties of optimal watermarks.

5.5.1 Decision Statistic

Assume θ is random over the interval $[0, T]$, with a distribution $\pi(\theta)$. We further assume that θ is independent of $s(t), t \in [0, T]$. The test statistic c_θ in (12) cannot be used because θ is unknown, but consider its mean-square average:

$$\begin{aligned} z &= \int_0^T c_\theta^2\, \pi(\theta)\, d\theta \qquad (22) \\ &= \int_0^T \int_0^T \int_0^T y(t) y(t') w(t - \theta) w(t' - \theta)\, dt\, dt'\, \pi(\theta)\, d\theta \end{aligned}$$

which can be written in the form

$$z = \int_0^T \int_0^T y(t) y(t') R_w(t, t')\, dt\, dt' \tag{23}$$

where

$$R_w(t, t') = \int_0^T w(t - \theta) w(t' - \theta)\, \pi(\theta)\, d\theta \tag{24}$$

is a *weighted watermark autocorrelation sequence*. Computation of (23) is attractive because integration over θ is done offline via (24).

Note the following properties of R_w:

Symmetry: $R_w(t,t') = R_w(t',t)$;

Maximum value: $|R_w(t,t')| \leq R_w(t,t)$;

Uniformly distributed θ: If $\pi(\theta)$ is the uniform distribution over the interval $[0,T]$, then $R_w(t,t') = R_w(t-t')$ depends only on the difference between the time arguments.

Random watermarks: If $w(t)$ is a realization of a periodic, wide-sense stationary random process with correlation sequence $r_w(t)$, then $E_W[R_w(t,t')] = r_w(t-t')$. Moreover, if T is large and the support of $\pi(\theta)$ is sufficiently broad [3], then $R_w(t,t') \approx E_W[R_w(t,t')]$.

Also note that if $\pi(\theta)$ is a distribution concentrated near some time t_0, then $R_w(t,t')$ represents a local correlation function.

Instead of (23), we may want to consider the more general quadratic test statistic

$$z = \int_0^T \int_0^T y(t)y(t')K(t,t')\,dtdt' \tag{25}$$

where $K(t,t')$ is an arbitrary symmetric Hilbert-Schmidt kernel. Let us derive a test based on (25).

5.5.2 Deflection Criterion

Computing the first two moments of Z in (25) under H_0 and H_1, we obtain

$$\begin{aligned}
E[Z|H_0] &= \int_0^T \int_0^T E[y(t)y(t')|H_0]K(t,t')\,dt\,dt' \\
&= \int_0^T \int_0^T R_s(t,t')K(t,t')\,dt\,dt' \\
&= \sigma_S^2 \int_0^T K(t,t)\,dt, \qquad (26) \\
E[Z|H_1] &= \int_0^T \int_0^T E[y(t)y(t')|H_1]K(t,t')\,dt\,dt' \\
&\overset{(a)}{=} \int_0^T \int_0^T [E_\theta[w(t-\theta)w(t'-\theta)] + R_s(t,t')]K(t,t')\,dt\,dt' \\
&= \int_0^T \int_0^T [R_w(t,t') + R_s(t,t')]K(t,t')\,dt\,dt' \\
&= \int_0^T \int_0^T R_w(t,t')K(t,t')\,dt\,dt' + \sigma_S^2 \int_0^T K(t,t)\,dt, \qquad (27)
\end{aligned}$$

[3] Additional technical conditions apply.

where equality (a) is due to the independence of θ and $s(t), t \in [0,T]$. After some algebraic manipulations, one can derive [10]

$$Var[Z|H_0] = Var[Z|H_1] = 2\sigma_S^4 \int_0^T \int_0^T K^2(t,t')\,dt\,dt'. \tag{28}$$

The quadratic test takes the form

$$z \underset{H_0}{\overset{H_1}{\gtrless}} \eta = \frac{1}{2}\int_0^T \int_0^T R_w(t,t')K(t,t')\,dt\,dt'. + \sigma_S^2 \int_0^T K(t,t)\,dt. \tag{29}$$

The deflection criterion (also termed deflection coefficient or *generalized signal-to-noise ratio*) for quadratic detection is defined as [11]

$$d^2 = \frac{(E[Z|H_1] - E[Z|H_0])^2}{Var[Z|H_0]} = \frac{(\int_0^T \int_0^T R_w(t,t')K(t,t')\,dt\,dt')^2}{2\sigma_S^4 \int_0^T \int_0^T K^2(t,t')\,dt\,dt'}. \tag{30}$$

This criterion would determine the probability of error of the test (29) if the distributions of Z under H_0 and H_1 were Gaussian. Of course they are not Gaussian in this problem, and the deflection coefficient only serves as a tractable measure of separability of the two distributions.

By application of the Cauchy-Schwarz inequality, the choice of $K(t,t')$ that maximizes d^2 turns out to be $\alpha R_w(t,t')$ where α is an arbitrary nonzero constant. Hence the optimal quadratic decision statistic is (22), and leads to the deflection criterion

$$d^2 = \frac{\int_0^T \int_0^T R_w^2(t,t')\,dt\,dt'}{2\sigma_S^4}. \tag{31}$$

5.5.3 Optimal Watermark Design

The use of d^2 in (31) as a performance criterion for quadratic detection also suggests its use as a criterion for watermark design. The dependency of d^2 on w is via the correlation function $R_w(t,t')$. Assume either that $\pi(\theta)$ is uniform over $[0,T]$, or that the stochastic model of Sec. 5.5.1 for the watermark can be used. Hence $R_w(t,t') = R_w(t-t')$, and (31) becomes

$$d^2 = \frac{T\int_0^T R_w^2(t)\,dt}{2\sigma_S^4}. \tag{32}$$

Assume the fixed energy constraint

$$R_w(0) = \frac{1}{T}\|w\|^2 \le \sigma_w^2. \tag{33}$$

Maximizing (32) over R_w subject to the constraint (33), we obtain

$$R_w(t,t') = \sigma_w^2, \quad \forall t,t'.$$

The maximum is achieved by the constant watermark $w(t) \equiv \sigma_w$. For this watermark,

$$d_{opt}^2 = \frac{T^2\sigma_w^4}{2\sigma_S^4} = \frac{1}{2}SNR^2. \tag{34}$$

Of course, a constant watermark does not convey any information about the value of θ.

5.5.4 Discussion

Remark #1. For a sinusoidal watermark $w(t) = \sqrt{2}\sigma_w \cos(2\pi kt/T + \phi)$, where $k \in \mathbb{N}_0$ and ϕ is an arbitrary phase factor, we have $R_w(t) = \sigma_w^2 \cos(2\pi kt/T)$ and $d^2 = \frac{1}{2}d_{opt}^2$, independently of the values of k and ϕ.

Remark #2. For narrowband watermarks, $d^2 \approx \frac{1}{2}d_{opt}^2$.

Remark #3. In order to make $d^2 \propto \int_0^T R_w^2(t)\,dt$ large, efficient watermarks should have a *long correlation time*. Correlation time may be defined as the smallest T_c such that $|R_w(t)| \leq \beta R_w(0)$ for all $t \in [T_c, T/2]$, where $\beta < 1$ is some fixed constant.

Remark #4. Among all watermarks with correlation time T_c and $\beta = 0$, the optimal choice is

$$R_w(t) = \begin{cases} \sigma_w^2 & : \ 0 \leq t < T_c \\ 0 & : \ T_c \leq t < T, \end{cases} \tag{35}$$

leading to

$$d^2 = T\frac{T_c\sigma_w^4}{2\sigma_S^4} = \frac{T_c}{T}d_{opt}^2. \tag{36}$$

While there is a substantial reduction of performance if $T_c << T$, the deflection criterion is still increasing (linearly instead of quadratically) with T.

Remark #5. The use of a sinusoidal watermark is unrealistic in watermarking applications because a clever attacker would identify its presence and filter it out (instead of implementing a delay operation). Nevertheless the above analysis demonstrates the advantages of watermarks with *long correlation time*, because such watermarks spread out $R_w(t,t')$ over the entire square $[0,T]^2$, leading to a large value of the deflection coefficient (22). Conversely, for watermarks with a *short correlation time*, $R_w(t,t')$ is concentrated near the vicinity of the main diagonal of the square $[0,T]^2$, leading to a smaller value of the deflection coefficient.

Remark #6. If $w(t)$ is a realization from a periodic, wide-sense stationary random process with correlation $r_w(t) = E[w(t')w(t'+t)]$, then d^2 is a random variable which converges almost surely to

$$d^2 = \frac{T\int_0^T r_w^2(t)\,dt}{2\sigma_S^4}$$

as $T \to \infty$, by the strong law of large numbers. Hence the previous remarks about the benefits of long correlation times apply to the stochastic case as well.

Remark #7. If σ_S^2 is known only approximately (say is estimated from the data), the detection test (29) can be used with the approximate σ_S^2, with little performance loss in the case of large SNR. Indeed, comparison of (26) and (27) shows that $\frac{E[Z|H_1]}{E[Z|H_0]} \sim SNR$ when SNR is large.

Remark #8. The various test statistics considered so far may be written in the form $z = \max_\theta c_\theta$ (GLRT statistic), $z = \int_0^T \exp\{c_\theta/\sigma_s^2\}\pi(\theta)\,d\theta$ (sufficient statistic used by Bayes test), or $z = \int_0^T c_\theta^2\pi(\theta)\,d\theta$ (quadratic statistic). All these statistics may be thought of as measures of peakness of the function c_θ.

6 Analysis of Warping Attacks

This section extends the results of Sec. 5 to more general warping attacks of the form (6). Warping destroys long-term correlations, so it will not come as a surprise that warping attacks are *much more effective* than delay attacks.

6.1 Quadratic Noncoherent Detector

When $y(t)$ is given by the warping model (9), the test statistic (25) can still be used for quadratic detection. The mean and variance of Z are still given by (26), (27) and (28), with the correlation function now given by

$$R_w(t,t') = \int_0^T \int_0^T w(t-\theta_t)w(t'-\theta_{t'})\,\pi(\theta_t,\theta'_t)\,d\theta_t d\theta_{t'} \tag{37}$$

where $\pi(\theta_t,\theta'_t)$ now denotes the joint p.d.f. of θ_t and $\theta_{t'}$. Examples of computation of $R_w(t,t')$ can be found in the optical-communications literature, when $w(t)$ is a sinusoid and θ_t is a Brownian motion (model for phase noise). Then the warped sinusoid $w(t-\theta_t)$ has a Lorentzian spectrum [9, 10]. In our problem, $w(t)$ is neither a sinusoid nor even a narrowband signal.

We model θ_t as a periodic stationary stochastic process. The kernel K in (25) that maximizes the deflection coefficient is still R_w. To gain some insight into this problem, make two fairly mild assumptions:

A1. $\int_0^T w(t-\theta_t)\,\pi(\theta_t)\,d\theta = 0$ for all t,

A2. θ_t and $\theta_{t'}$ are independent for $T_c \le |t-t'| < T$.

The parameter T_c is large if the warping functions vary slowly. According to Assumption A2, the correlation time of θ_t is at most T_c. Then for all such t,t' such that $T_c \le |t-t'| < T$, we have

$$\begin{aligned} R_w(t,t') &= \int_0^T w(t-\theta_t)w(t'-\theta_{t'})\,\pi(\theta_t)\pi(\theta'_t)\,d\theta_t d\theta_{t'} \\ &= \left(\int_0^T w(t-\theta_t)\,\pi(\theta_t)\,d\theta\right)\left(\int_0^T w(t'-\theta_{t'})\pi(\theta'_t)d\theta_{t'}\right) \\ &= 0. \end{aligned}$$

Hence the warped watermark also has correlation time limited to T_c. Under the watermark energy constraint (33), it is easily seen that the optimal correlation function and deflection coefficient are given in (35) and (36).

Piecewise-Constant "Warping" Functions. Consider the following piecewise-constant model for the warping function. Let $t_k = \frac{k}{K}T$ for $0 \le k < K$, and assume that θ_t is equal to θ_{t_k} for all $t \in [t_k, t_k+1)$. Here $\theta_{t_k}, 0 \le k < K$ are independent random variables with a uniform distribution over $[0,T]$. (Hence θ_t is nonmonotonic and is an interval permutation function rather than a warping function: this is a broader class of attacks than the one we originally considered.)

Each interval has length $T_c = T/K$. Then $R_w(t,t') = 0$ if t and t' do not belong to the same interval, and so

$$\begin{aligned} d^2 &= \frac{\int_0^T \int_0^T R_w^2(t,t')\,dt\,dt'}{2\sigma_S^4} \\ &= \frac{\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \int_{t_k}^{t_{k+1}} R_w^2(t,t')\,dt\,dt'}{2\sigma_S^4} \\ &\leq \frac{\sum_{k=0}^{K-1} (t_{k+1}-t_k)^2 R_w^2(0)}{2\sigma_S^4} \\ &= \frac{TT_c R_w^2(0)}{2\sigma_S^4}, \end{aligned} \tag{38}$$

where the right side is the upper bound on d^2 for warping functions with correlation time T_c, and watermarks with energy constraint (33). The bound can be *nearly attained* by letting $w(t)$ be a narrowband signal over each interval $[t_k, t_k + 1)$, with possibly a different center frequency in each interval.

Randomized basis functions. The watermark should be randomized for security purposes, and be difficult to estimate. We propose a watermark construction based on randomized basis functions which nearly achieves the upper bound (36) on the deflection coefficient. The security of this design remains to be investigated. Assume $K = T/T_c$ is an integer, and consider a distribution $p(u), 0 \leq u \leq T_c$ which is concentrated near $u = T_c$. Specifically, assume that $E[U] = \alpha T_c$ and $Var[U] = \beta T_c^2$, where $\alpha \approx 1$ and $\beta << 1$. Let $t_{-1} = 0$ and generate i.i.d. random variables $u_0, u_1, u_2, \cdots$. Generate the increasing random sequence $t_k = t_{k-1} + u_k$ for $k = 0, 1, 2, \cdots$, and stop as soon as $T_k \geq T$. Let K denote the final value of k. Note that the random variables K and $\sum_{k=0}^{K-1}(t_{k+1} - t_k)^2$ converge almost surely to $\frac{T}{\alpha T_c}$ and $\alpha T T_c(1 + \frac{\beta}{\alpha^2})$, respectively, as $\frac{T}{T_c} \to \infty$.

Assume again a piecewise-constant model for θ_t. For simplicity, assume that θ_t is constant over each interval $[t_k, t_{k+1}]$ (a similar analysis applies if θ_t is constant over a different set of intervals, of average length T_c.) Then from (38), we obtain

$$d^2 \leq \frac{\sum_{k=0}^{K-1}(t_{k+1} - t_k)^2 R_w^2(0)}{2\sigma_S^4} \approx \frac{TT_c(\alpha + \beta/\alpha) R_w^2(0)}{2\sigma_S^4} \tag{39}$$

where the approximation is accurate if $T >> T_c$. This example illustrates the fact that little optimality is lost using randomized basis functions.

6.2 Estimation of Warping Functions

Estimating a warping function is considerably more involved than estimating a simple delay. The fact that the warping function varies slowly suggests that accurate estimation is still possible. Sec. 6.2.2 outlines a possible approach in which the warping function is modeled as a Markov random process.

6.2.1 Fisher Information

In order to derive bounds on the estimation accuracy of the warping function $\theta(t)$, we use a parametric model:

$$\theta(t) = \sum_{k=0}^{K-1} \theta_k \varphi_k(t), \quad 0 \le t \le T. \tag{40}$$

Now the Fisher information matrix for estimation of the K-vector $[\theta_0, \cdots, \theta_{K-1}]$ is given by [2]

$$\begin{aligned} J_{kl}(\theta) &= -E_{Y|\theta}\left[\frac{\partial^2 l(\theta, Y)}{\partial\theta_k \partial\theta_l}\right] \\ &= \frac{1}{\sigma_S^2}\int_0^T |w'(t-\theta(t))|^2 \varphi_k(t)\varphi_l(t)\, dt, \quad 0 \le k, l < K. \end{aligned} \tag{41}$$

The expression (15) is a special case of (41). If the basis functions $\{\varphi_k\}$ are nonoverlapping, or at least are orthogonal with respect to $|w'(t-\theta(t))|^2$, then $J(\theta)$ is diagonal. If $w'(t)$ is a wide-sense stationary process with mean zero, variance $\sigma_{w'}^2$, then $J_{kl}(\theta)$ is a random variable whose expectation is $\sigma_{w'}^2 \int_0^T \varphi_k(t)\varphi_l(t)\, dt$. Moreover, if the support of the basis functions $\{\varphi_k\}$ is "long enough" (this notion will not be formalized here), the distribution of $J_{kl}(\theta)$ is concentrated in the vicinity of its average. As in Sec. 5.2, the dependency of $J(\theta)$ on θ is mild or even inexistent.

If the vector $[\theta_0, \cdots, \theta_{K-1}]$ is random with p.d.f. $\pi(\theta)$, define

$$J_{kl}^{\pi}(\theta) = -\frac{\partial^2 \ln \pi(\theta)}{\partial\theta_k \partial\theta_l}, \quad 0 \le k, l < K. \tag{42}$$

If $\pi(\theta)$ is a Gaussian p.d.f. with mean zero and covariance matrix R, then $J^{\pi}(\theta) = R^{-1}$.

For any unbiased estimator $\hat{\theta}(Y)$ of θ, we have [15]

$$Cov(\theta(\hat{Y})) \ge (J^{\pi}(\theta) + J(\theta))^{-1}, \tag{43}$$

where the inequality indicates that the difference between the left and rides sides is a nonnegative definite matrix.

Example. To illustrate (43), assume that $K = 2^M$ is a power of 2. Consider a system of (scaled) Haar basis functions:

$$\theta(t) = \theta_{00} + \sum_{m=1}^{M} \sum_{l=0}^{2^{M-m}} \theta_{lm}\varphi_{lm}(t), \quad 0 \le t \le T.$$

Here $T_c = 2^{-M}T$, $\varphi_0(t) \equiv 1$, and $\{\varphi_{lm}\}$ are scaled Haar wavelets at resolutions $2^m T_c$, for $0 \le m \le M$. Specifically,

$$\varphi_{lm}(t) = \begin{cases} 1 & : \ 2^m(l-1)T_c \le t < 2^m(l-\frac{1}{2})T_c \\ -1 & : \ 2^m(l-\frac{1}{2})T_c \le t < 2^m l T_c \\ 0 & : \ 0 \le t < 2^m(l-1)T_c \text{ or } 2^m l T_c \le t \le T \end{cases}$$

for $0 \le l < 2^{M-m}$ and $1 \le m \le M$. Assume that $\{\theta_{lm}\}$ are independent Gaussian random variables with mean 0 and variance ∞ if $k = 0$ and $C2^{mD}T_c$ if $k \ge 1$. This models a warping function with

fractal characteristics and smoothness determined by the exponent D. The mean value θ_{00} of this warping function (average delay) is completely unknown a priori. Then

$$\begin{aligned} J(\theta) &= \sigma_{w'}^2 T_c I \\ J^{\pi}(\theta) &= (CT_c)^{-1} diag\{0, 2^{-D}, 2^{-D}, \cdots, 2^{-MD}\}. \end{aligned}$$

The resulting Fisher information matrix $J(\theta) + J^{\pi}(\theta)$ is diagonal. The Cramer-Rao lower bound on the variance of any unbiased estimator of θ_{00} is equal to $(\sigma_{w'}^2 T_c)^{-1}$. The Cramer-rao lower bound on the variance of any unbiased estimator of θ_{lm} is equal to $(\sigma_{w'}^2 T_c + (C2^{mD} T_c)^{-1})^{-1}$ if $m \geq 1$. So θ_{lm} may be easy to estimate even though θ_{00} is hard to estimate.

6.2.2 Tracking

Assume that

$$\theta(n) = \theta(n-1) + u(n), \quad n \in \{0, 1, \cdots, N-1\} \tag{44}$$

where $u(n)$ is a white noise process with mean zero and variance σ_u^2. (so the process (44) has independent increments.) Also assume that $s(n)$ satisfies the first-order autoregressive model

$$s(n) = \rho s(n-1) + v(n), \quad n \in \{0, 1, \cdots, N-1\} \tag{45}$$

where $|\rho| < 1$, and $v(n)$ is a white noise process with mean zero and variance σ_v^2.

The warping sequence θ is related to the observations y via the nonlinear model

$$y(n) = w(n - \theta(n)) + s(n), \quad n \in \{0, 1, \cdots, N-1\} \tag{46}$$

which is a discrete-time equivalent of (9). We view θ and s as the state of a dynamic system whose observations are given by (46). The model (44), (45), (46) suggests the use of an extended Kalman filter (EKF), or a recursive Bayesian filter.

Extended Kalman filter. The basic idea is to linearize the observation model (46) in $\theta(n)$ around $\theta(n-1)$:

$$\begin{aligned} \tilde{y}(n) &\stackrel{\triangle}{=} y(n) - [w(n - \theta(n-1)) + \theta(n-1) w'(n - \theta(n-1))] & (47) \\ &\approx -\theta(n) w'(n - \theta(n-1)) + s(n). & (48) \end{aligned}$$

For simplicity, we first assume that the signal process $s(n)$ is white, in which case $\rho = 0$ and $\sigma_v^2 = \sigma_S^2$. (If s is not white, we may consider whitening the observations as a preprocessing step. The effects of this whitening on the warping have to be determined.)

The predicted value of $\theta(n)$ based on observations up to time $n-1$ is the same as the estimated value of $\theta(n-1)$ based on the same observations and is denoted by $\hat{\theta}(n) = \hat{\theta}(n|n-1) = \hat{\theta}(n-1|n-1)$. This prediction is updated as follows:

$$\hat{\theta}(n) = \hat{\theta}(n-1) + K(\tilde{y}(n) - h\hat{\theta}(n-1)) \tag{49}$$

where

$$K = \frac{\Sigma_{n-1} h}{h^2 \Sigma_{n-1} + \sigma_S^2} \tag{50}$$

is the Kalman gain, $h = -w'(n - \theta(n-1))$ is the state transition parameter, Σ_n is the variance of $\hat{\theta}(n)$, and the term that multiplies K is the innovation in $y(n)$. The variance Σ_n satisfies the Riccati-type equation

$$\Sigma_n = \frac{\Sigma_{n-1}}{h^2\Sigma_{n-1}/\sigma_S^2 + 1} + \sigma_u^2. \tag{51}$$

Recursive Bayesian Filter. The linearization (48) is accurate for lowpass watermarks, but its validity is highly questionable for pseudo-noise sequences. Moreover, the performance of Kalman and EKF filters can be poor when noise statistics are strongly non-Gaussian. This suggests the use of more advanced recursive estimation methods such as *particle filtering*, which have recently gained popularity in the signal processsing literature [16].

References

[1] J. R. Hernández and F. Pérez-González, "Statistical Analysis of Watermarking Schemes for Copyright Protection of Images," *Proc. IEEE*, Vol. 87, No.7, pp. 1142—1166, 1999.

[2] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd Ed., Springer-Verlag, 1994.

[3] O. Zeitouni, "When is the Generalized Likelihood Ratio Test Optimal?", *IEEE Trans. Info. Theory*, Vol. 38, No. 5, pp. 1597—1602, 1992.

[4] S. Verdu, *Multiuser Detection*, Cambridge U. Press, Cambridge, U.K., 1998.

[5] D. Kirovski and H. Malvar, "Robust Covert Communication over a Public Audio Channel using Spread Spectrum," *Proc. Information Hiding Workshop*, Pittsburgh, PA, 2001.

[6] M. A. Blanco and F. S. Hill, Jr., "On Time Warping and the Random Delay Channel," *IEEE Trans. on Information Theory*, Vol. 25, No. 2, pp. 155—166, 1979.

[7] R. A. Roberts, "On the Detection of a Signal Known Except for Phase," *IEEE Trans. on Information Theory*, Jan. 1965.

[8] C. R. Baker, "Optimum Quadratic Detection of a Random Vector in Gaussian Noise," *IEEE Trans. on Comm.*, Vol. 14, No. 6, pp. 802—805, 1966.

[9] G. J. Foschini, L. J. Greenstein and G. Vanucci, "Noncoherent Detection of Lightwave Signals Corrupted by Phase Noise," *IEEE Trans. on Comm.*, Vol. 36, No. 3, pp. 306—314, 1988.

[10] V. Veeravalli and H. V. Poor, "Quadratic Detection of Signals with Drifting Phase," *J. Acoustic Soc. of America*, Vol. 89, No. 2, pp. 811—819, Feb. 1991.

[11] R. J. Barton and H. V. Poor, "On Generalized Signal-to-Noise Ratios in Signal Detection," *Mathematics of Control, Signals and Systems*, Vol. 5, No. 1, pp. 81—91, 1992.

[12] P. Moulin and A. Ivanovic, "The Watermark Selection Game," *Proc. CISS Conf. on Info. Sci. and Systems*, Baltimore, MD, March 2001.

[13] B. S. Clarke and A. R. Barron, " Information–Theoretic Asymptotics of Bayes Methods," *IEEE Trans. on Information Theory*, Vol. 36, No. 3, pp. 453—471, May 1990.

[14] T. A. Severini, "Likelihood Functions for Inference in the Presence of Nuisance Parameter," *Biometrika*, Vol. 85, No. 3, pp. 507—522, 1998.

[15] H. L Van Trees, *Detection, Estimation and Modulation Theory*, Wiley, New York, 1968.

[16] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Trans. on Signal Processing*, Vol. 50, No. 2, pp. 174—188, Feb. 2002.