Detection-Theoretic Analysis of Desynchronization Attacks in Watermarking *

Pierre Moulin, Alexia Briassouli
University of Illinois
Beckman Inst., Coord. Sci. Lab & ECE Dept.
405 N. Mathews Ave.
Urbana, IL 61801
USA
{moulin,briassouli}@ipf.uiuc.edu

Henrique Malvar
Microsoft Research
One Microsoft Way
Redmond, WA 98052
USA
malvar@microsoft.com

Abstract. This paper studies the effects of desynchronization attacks such as delay and warping on the performance of blind watermark detection systems. Detection performance is measured using a probability-of-error criterion. 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 correlation time of the warping function. Piecewise-sinusoidal watermarks are shown to be nearly optimal in that case. Some examples are presented.

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 $p_0$; e.g., a Gaussian density function with mean zero and variance $\theta$.
- compression using a particular algorithm with quality factor $\theta$;
- delay of the watermarked signal by $\theta$ 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, $\theta$ is a scalar, a vector, or even a function.

While basic results of detection theory have been applied to watermarking [1, 2], current watermarking literature does not provide 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 $\theta$) and study its performance under a list of attacks. A more principled approach, which we follow here, 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 can construct optimal watermarks. 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 $p_0(y)$ which depends on the

*Work performed in August and September 2001, while P. Moulin was a Visiting Researcher at Microsoft Research. This work was partially supported by NSF grants CCR 00-81368 and CDA 96-24396.
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 [3]:

$$
\begin{align*}
H_0 & : y \sim p_0 \\
H_1 & : y \sim p_0, \quad \theta \in \Theta.
\end{align*}
$$

(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 set $\Theta$. Then integrating out $\theta$ yields a known distribution

$$p_1(y) = \int_\Theta p_0(y) dP(\theta).$$

(2)

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

$$
\frac{p_1(y)}{p_0(y)} \overset{H_1}{\gtrsim} \eta, \quad H_0
$$

(3)

where $\eta$ 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 $\delta$ conditioned on $\theta$. The minimax rule is a randomized LRT.

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

$$
\frac{p_1(y)}{p_0(y)} \overset{H_1}{\gtrsim} \eta.
$$

(4)

The first three tests have clear optimality properties but may be computationally intractable due to the need to integrate out $\theta$ in (2). The minimax approach is arguably realistic in the presence of an adversary. The GLRT has asymptotic optimality properties [4], 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.

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 $R_s(t, t') = E[s(t)s(t')] = \sigma_s^2 \delta(t-t')$ for $t, t' \in \mathbb{R}$, where $\delta(t) = \sum_{k \in \mathbb{Z}} \delta(t-kT)$ 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 $\theta(n), n \in \{0, \ldots, N-1\}$ for the discrete-time model.  

\textbf{Discrete-Time Data}:

$$
\begin{align*}
H_0 & : y(n) = s(n), \\
& \quad n \in \{0, 1, \ldots, N-1\} \\
H_1 & : y(n) = w(n - \theta(n)) + s(n - \theta(n)), \\
& \quad n \in \{0, 1, \ldots, N-1\}.
\end{align*}
$$

(5)

\textbf{Continuous-Time Data}:

$$
\begin{align*}
H_0 & : y(t) = s(t), \\
& \quad 0 \leq t \leq T \\
H_1 & : y(t) = w(t - \theta(t)) + s(t - \theta(t)), \\
& \quad 0 \leq t \leq T.
\end{align*}
$$

(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 $\theta$ is no greater than some specified $\epsilon$:

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

$$
\Theta = \{\theta : |\theta(t)| \leq \epsilon\} \quad \text{(continuous case)}
$$

(7)

(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

\footnote{To qualify as a warping function, $\theta$ 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.}
make the analysis tractable, we have decided to study the hypothesis test

\[
\begin{align*}
H_0 : & \quad y(t) = s(t) \quad , \quad 0 \leq t \leq T \\
H_1 : & \quad y(t) = w(t - \theta(t)) + s(t) \quad , \quad 0 \leq t \leq T ,
\end{align*}
\]

(9)

which serves as an approximation to the original detection problem.

4 Related 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 \( \theta \)) 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, 3]. If partial information about the delay or time-varying delay is available, the detection problem is said to be partially coherent.

Much of the signal detection theory developed in the communication literature has been applied to narrow-band 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

Consider first the case of a simple delay \( \theta \in [0, T] \). The likelihood functional for \( \theta \) is [3]

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

5.1 Coherent Detector

If the delay \( \theta \) is known, we have a coherent detection problem [3].

Define the normalized, deterministic autocorrelation function of the watermark as

\[
R_w(t) = \frac{1}{T} \int_0^T w(t')w(t' + t) dt',
\]

(10)

which has a maximum at \( t = 0 \).

Under our model assumptions, if \( \theta \) is known, the LRT becomes a simple correlation test [3]:

\[
c_\theta = \int_0^T y(t)w(t - \theta) dt \quad \gtrless \quad \eta \quad (11)
\]

where the correlation statistic \( c_\theta \) has mean 0 and \( TR_w(0) \) under \( H_0 \) and \( H_1 \), respectively, and has variance \( \sigma^2 \) 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

\[
P_e = Q \left( \sqrt{SNR} \right). \quad (12)
\]

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

\[
SNR = \frac{TR_w(0)}{\sigma^2}, \quad (13)
\]

where the numerator represents total watermark energy. Note that detector performance (12) depends on the energy of the watermark and not on its spectral contents.

5.2 Quadratic Noncoherent Detector

When \( \theta \) is random, the optimal Bayesian test (3) is expensive to implement due to the need to compute the correlation statistic \( c_\theta \) for all values of \( \theta \). 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.

5.2.1 Decision Statistic

Assume \( \theta \) is random over the interval \( [0, T] \), with a distribution \( \pi(\theta) \). We further assume that \( \theta \) is independent of \( s(t), t \in [0, T] \). The test statistic \( c_\theta \) in (11) cannot be used because \( \theta \) is unknown, but consider its mean-square average:

\[
z = \int_0^T c_\theta^2 \pi(\theta) d\theta \quad (14)
\]

\[
= \int_0^T \int_0^T \int_0^T \int_0^T y(t)w(t - \theta)w(t' - \theta) \ dt \ dt' \ \pi(\theta) \ d\theta
\]

which can be written in the form

\[
z = \int_0^T \int_0^T y(t)R_w(t, t') dt \ dt'
\]

(15)
where
\[
R_w(t, t') = \int_0^T w(t-\theta)w(t'-\theta) \pi(\theta) \, d\theta
\]  
(16)

is a weighted watermark autocorrelation sequence. Computation of (15) is attractive because integration over $\theta$ is done offline via (16).

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 $\theta$:** 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 \(^2\), 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 (15), we may want to consider the more general quadratic test statistic
\[
z = \int_0^T \int_0^T y(t)y(t')K(t, t') \, dt \, dt'
\]  
(17)

where $K(t, t')$ is an arbitrary symmetric Hilbert-Schmidt kernel. Let us derive a test based on (17).

### 5.2.2 Deflection Criterion

Computing the first two moments of $Z$ in (17) under $H_0$ and $H_1$, we obtain
\[
E[Z|H_0] = \int_0^T \int_0^T E[y(t)y(t')]H_0 K(t, t') \, dt \, dt'
\]
\[
= \int_0^T \int_0^T R_s(t, t')K(t, t') \, dt \, dt'
\]
\[
= \sigma_S^2 \int_0^T K(t, t) \, dt,
\]  
(18)

\[
E[Z|H_1] = \int_0^T \int_0^T E[y(t)y(t')]H_1 K(t, t') \, dt \, dt'
\]
\[
= \int_0^T \int_0^T [E_\theta[w(t-\theta)w(t'-\theta)]
\]
\[
+ R_s(t, t')]K(t, t') \, dt \, dt'.
\]

where equality (a) is due to the independence of $\theta$ and $s(t), t \in [0, T]$. After some algebraic manipulations, one can derive \(^{[10]}\)
\[
Var[Z|H_0] = 2\sigma_S^4 \int_0^T \int_0^T K^2(t, t') \, dt \, dt'.
\]  
(20)

The quadratic test takes the form
\[
z \geq \eta = \frac{1}{2} \int_0^T \int_0^T R_w(t, t')K(t, t') \, dt \, dt'
\]
\[
+ \sigma_S^2 \int_0^T K(t, t) \, dt.
\]  
(21)

The deflection criterion (also termed deflection coefficient or generalized signal-to-noise ratio) for quadratic detection is defined as \(^{[11]}\)
\[
d^2 = \frac{(E[Z|H_1] - E[Z|H_0])^2}{Var[Z|H_0]}
\]
\[
= \frac{(\int_0^T \int_0^T R_w(t, t')K(t, t') \, dt \, dt')^2}{2\sigma_S^4 \int_0^T \int_0^T K^2(t, t') \, dt \, dt'}. \tag{22}
\]

This criterion would determine the probability of error of the test (21) 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 $\alpha$ is an arbitrary nonzero constant. Hence the optimal quadratic decision statistic is (14), and leads to the deflection criterion
\[
d^2 = \frac{\int_0^T \int_0^T R_w(t, t') \, dt \, dt'}{2\sigma_S^4}. \tag{23}
\]

**5.2.3 Optimal Watermark Design**

The use of $d^2$ in (23) 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

\(^{2}\) Additional technical conditions apply.
model of Sec. 5.2.1 for the watermark can be used. Hence \( R_w(t, t') = R_w(t - t') \), and (23) becomes
\[
d^2 = \frac{T \int_0^T R_w^2(t) dt}{2\sigma_w^2}.
\] (24)
Assume the fixed energy constraint
\[
R_w(0) = \frac{1}{T} \| w \|^2 \leq \sigma_w^2.
\] (25)
Maximizing (24) over \( R_w \) subject to the constraint (25), we obtain
\[
R_w(t, t') = \sigma_w^2, \quad \forall t, t'.
\]
The maximum is achieved by the constant watermark \( w(t) = \sigma_w \). For this watermark,
\[
d^2_{opt} = \frac{T^2 \sigma_w^4}{2\sigma_w^2} = \frac{1}{2} SNR^2.
\] (26)

5.2.4 Discussion

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

Remark #2. For narrowband watermarks, \( d^2 \approx \frac{1}{T} d^2_{opt} \).

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
\[
R_w(t) = \begin{cases} \sigma_w^2 & : 0 \leq t < T_c \\ 0 & : T_c \leq t < T, \end{cases}
\] (27)
leading to
\[
d^2 = T \frac{T_c \sigma_w^4}{2\sigma_w^2} = \frac{T_c}{T} d^2_{opt}.
\] (28)
While there is a substantial reduction of performance if \( T_c \ll T \), 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 (14). 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
\[
d^2 = \frac{T \int_0^T r_w^2(t) dt}{2\sigma_w^2}
\]
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 \( \sigma_w^2 \) is known only approximately (say is estimated from the data), the detection test (21) can be used with the approximate \( \sigma_w^2 \), with little performance loss in the case of large SNR. Indeed, comparison of (18) and (19) shows that \( \frac{E[Z^2/R]}{E[Z]} \sim SNR \) when \( SNR \) is large.

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

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 (17) can still be used for quadratic detection. The mean and variance of \( Z \) are still given by (18), (19) and (20), with the correlation function now given by
\[
R_w(t, t') = \int_0^T \int_0^T w(t - \theta) w(t' - \theta') \pi(\theta, \theta') d\theta d\theta'.
\] (29)
where \( \pi(\theta_1, \theta_1') \) now denotes the joint probability density function of \( \theta_1 \) and \( \theta_1' \). Examples of computation of \( R_w(t, t') \) can be found in the optical-communications literature, when \( w(t) \) is a sinusoid and \( \theta_1 \) 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 \(\theta_t\) as a periodic stationary stochastic process. The kernel \(K^*\) in (17) 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. \(\theta_t\) and \(\theta_{t'}\) are independent for \(T_c \leq |t - t'| < T\).

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

\[
R_w(t, t') = \int_0^T w(t - \theta_t)w(t' - \theta_{t'}) \pi(\theta_t)\pi(\theta_{t'}) \, d\theta_t \, d\theta_{t'}
\]

\[
= \left( \int_0^T w(t - \theta_t) \pi(\theta_t) \, d\theta_t \right) \times \left( \int_0^T w(t' - \theta_{t'}) \pi(\theta_{t'}) \, d\theta_{t'} \right)
\]

\[
= 0.
\]

Hence the warped watermark also has correlation time limited to \(T_c\). Under the watermark energy constraint (25), it is easily seen that the optimal correlation function and deflection coefficient are given in (27) and (28).

**Piecewise-Constant “Warping” Functions.** Consider the following piecewise-constant model for the warping function. Let \(t_k = kT\) for \(0 \leq k < K\), and assume that \(\theta_t = \theta_t(k)\) for all \(t \in [t_k, t_{k+1})\). Here \(\theta_t(k), 0 \leq k < K\) are independent random variables with a uniform distribution over \([0, T]\). (Hence \(\theta_t\) 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

\[
d^2 = \frac{\int_0^T \int_0^T R_w^2(t, t') \, dt \, dt'}{2\sigma_S^4}
\]

\[
= \frac{\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \int_{t_k}^{t_{k+1}} R_w^2(t, t') \, dt \, dt'}{2\sigma_S^4}
\]

\[
\leq \frac{\sum_{k=0}^{K-1} (t_{k+1} - t_k)^2 R_w^2(0)}{2\sigma_S^4}
\]

\[
= \frac{TT_cR_w^2(0)}{2\sigma_S^4},
\]

where the right side is the upper bound on \(d^2\) for warping functions with correlation time \(T_c\), and watermarks with energy constraint (25). 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.

**7 Numerical Results**

We performed several experiments to evaluate the performance of our design. The primary goal was numerical evaluation of the probability of error (which as mentioned earlier is not guaranteed to correlate with the value of the deflection criterion) under different choices of the watermark and different choices of the desynchronization attack. A related goal was to investigate the effect of different designs of the watermark and of the desynchronization attack on detector performance. In all cases, the data \(y(n), 1 \leq n \leq N\) was a discrete-time signal of length \(N = 200\), and Monte-Carlo runs (averaging over all random variables) were used to determine empirical probabilities of error.

**Watermarks.** We considered piecewise-sinusoidal watermarks such as that plotted in Fig. 1a. The intervals have uniform length \(N/N_c\), where the correlation time \(N_c\) (a submultiple of \(N\)) is user-selected. The phase of the first sinusoid is zero, and the phase of the following ones are chosen so as to preserve continuity of the underlying analog waveform.

**Attacks.** We considered three types of attacks. See Fig. 1b,c for examples of the last two attacks applied to the piecewise-sinusoid of Fig. 1a.

1. **Pure Delays.** The delay \(\theta\) was a uniformly distributed random variable in \([0, 1, \ldots, N]\).

2. **Piecewise-Constant Delays.** We considered \(N_c \in \{20, 40, 100\}\). The delays in different intervals (of length \(N_c\)) were independent random variables uniformly distributed over \([0, 1, \ldots, N]\).

3. **AR(1) Warping Function.** Here the attacker chooses \(\theta(n) = \rho\theta(n - 1) + e(n)\), where \(e(n)\) is a white Gaussian noise process with mean 0 and variance \(\sigma_e^2\). The expected mean-square first-order difference of \(\theta(n)\), \(\frac{1}{1+\rho} \sigma_e^2\), was constrained to be equal to \(\epsilon^2 = 0.04^2\) (see (7) and Fig. 2). The correlation time \(N_c\) was defined as \(\frac{\ln 0.5}{\ln \rho}\), so that \(\rho^{N_c} = 0.5\).

As expected, the performance of piecewise-sinusoidal watermarks was virtually independent of the individual values of the sinusoid frequencies.

Typical performance results are shown in Fig. 3 and 4. The probability of error of the detector is plotted as a function of \(\sigma_w^2/R_w(0)\) for piecewise-constant delay attacks and for AR(1) warping attacks, respectively.
Consider Fig. 3 first. Detector performance is worse when the number of intervals used by the attacker is large, as expected: the attack is more complicated. This observation is also consistent with the deflection criterion analysis in (30). Of course, detection performance is also worse when $\sigma_2^2/R_w(0)$ is large.

Next, consider Fig. 4. In this case the performance of the piecewise-sinusoidal watermarks is studied for an AR(1) warping attack, which could potentially be more effective than a piecewise-constant delay attack (because piecewise-sinusoidal watermarks have no optimality property in this case) or less effective (because temporal coherence of the warping function at all times is guaranteed with high probability). To make meaningful comparisons with Fig. 3, we plotted curves corresponding to similar correlation times. The results are strikingly similar to those of Fig. 3.

Finally, we verified that detector performance for some “natural” but suboptimal watermark designs is worse than it is for piecewise-sinusoidal watermarks. For instance, choosing $\sigma_2^2/R_w(0) = 22$ dB and a piecewise-constant delay attack with $N_c = 40$, we obtained a probability of error of 0.256 for white-Gaussian-noise watermarks instead of 0.061 for the piecewise-sinusoidal watermarks of Fig. 3.

References


Figure 1: Piecewise-sinusoidal watermark \([5 \text{ pieces, } N_c = 40]\) (a) before attack; (b) after a piecewise-constant delay attack; (c) after an AR(1) warping attack.

Figure 2: AR(1) warping attack \(\theta(n) (\rho = 0.9828, N_c = 40)\).

Figure 3: Probability of error for piecewise-sinusoidal watermarks under piecewise-constant delay attacks \((N_c = 200, 100, 40, 20)\).

Figure 4: Probability of error for piecewise-sinusoidal watermarks under AR(1) warping attacks \((N_c = \infty, 100, 40, 20)\).