Audio watermarking under desynchronization and additive noise attacks

Abdellatif Zaidi, Remy Boyer, Pierre Duhamel

▶ **To cite this version:**

Abdellatif Zaidi, Remy Boyer, Pierre Duhamel. Audio watermarking under desynchronization and additive noise attacks. IEEE Transactions on Signal Processing, 2006, 54 (2). hal-00575667

HAL Id: hal-00575667
https://hal.science/hal-00575667

Submitted on 10 Mar 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Audio Watermarking Under Desynchronization and Additive Noise Attacks

Abdellatif Zaidi, *Student Member, IEEE*, Rémy Boyer, *Member, IEEE*, and Pierre Duhamel, *Fellow, IEEE*

***Abstract*—Digital watermarking is often modeled as the transmission of a message over a noisy channel denoted as "watermark channel." Distortions introduced by the watermark channel result mainly from attacks and may include interference from the original signal. One of the main differences with classical transmission situations stems from the fact that *perceived* distortions have to be taken into account. However, measuring the perceived impact an attack has on a watermarked signal is currently an unsolved problem. Possible means of circumventing this problem would be 1) to define the distortion in a so-called "perceived domain" and define an "ad hoc" equivalence between objective and perceived distortion or 2) to define an "equivalent distortion" by removing from the attack noise the part that is correlated to the host signal. This paper concentrates on the second approach and first shows that the resulting "equivalent" attack is a particular case of a thoroughly studied channel: filtering plus additive noise. However, the approach in this paper emphasizes the fact that the additive noise in the model has to be decorrelated with the signal. Then, the formalism is applied to (desynchronization plus noise) attacks on audio signals. In this context, this paper provides the corresponding capacities, as well as optimal "attack" and "defense" strategies in a game theory context.**

***Index Terms*—Communication with side information, desynchronization, dirty-paper coding, spread-spectrum (SS), watermark channel, watermarking game theory.**

I. Introduction

DIGITAL watermarking can be viewed as a communication problem. An information m to be sent to the receiver is encoded into a signal $\mathbf{w}$ called the watermark, which is then embedded into the media signal $\mathbf{x}$, referred to as *the cover signal*, to form the watermarked data $\mathbf{s}$.[1] This watermarked data is sent to the receiver through a channel, denoted as *the watermark channel*, where it might be further processed or even replaced by some other data. This process is also denoted as *the attack*. In the context of robust watermarking, the goal of an attacker is to impair or even remove the embedded watermark information without impairing the cover signal. Conversely, the aim of the defender[2] is to design the transmitter in such a way that the watermark is still there, as long as the attack results in received signals of sufficient quality. This so-called robust watermarking was first proposed for multimedia copyright protection [1] and then for many other possible applications. Rather than considering a given application, this paper is concerned with the estimation of the channel parameters and the tuning of watermarking systems so that the attacker and defender strategies are optimized.

Robust digital watermarking differs from traditional communication in that the watermark should in general have the three following contradicting requirements.

1) *Imperceptibility*: After embedding, the watermarked document $\mathbf{s}$ should remain perceptually equivalent to the original signal $\mathbf{x}$. Usually, this is translated by the fact that the embedding distortion D_E should be upper bounded (disregarding for a while that the perceived distortion is difficult to estimate).
2) *Robustness*: The watermark must be robust toward common degradations. Depending on applications, these degradations result from benign processing and transmission; in other cases, they result from deliberate attacks.
3) *Capacity*: The embedder should be able to transmit the maximum amount of information through the watermark channel. The amount of information a watermark carries is called the *payload*. A rate of payload that is reliably detectable and recoverable at the receiver side is called an *achievable rate*. The data hiding capacity is the supremum of all achievable rates.

Over the last years, several watermarking schemes have been developed [2]–[4] for a large variety of data types. These schemes can be broadly divided in two main classes: 1) host-interference nonrejecting methods and 2) host-interference rejecting methods. Host interference nonrejecting methods do not allow the encoder to exploit knowledge of the host signal $\mathbf{x}$. The simplest methods consist of adding a pseudo-noise sequence to the host signal and are often referred to as spread-spectrum (SS) (they can be either blind [5] or nonblind [6]). When the knowledge of the host signal at the encoder is adequately exploited in system design, the resulting information embedding system can be host-interference free. Examples include quantization index modulation (QIM) [7], [8], dither modulation (DM) [9], [10], and the famous scalar Costa scheme (SCS) [11]. In QIM-based watermarking schemes, for example, decoding is achieved without any knowledge of the original signal. We may view the design of QIM systems as the simultaneous design of an ensemble of source codes (quantizers) and channel codes (signal constellations). For sufficiently large codebooks, blind recovery is possible

Manuscript received July 27, 2004; revised February 14, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zixiang Xiong.

The authors are with the CNRS/LSS, 91192 Gif-sur-Yvette, France (e-mail: abdellatif.zaidi@lss.supelec.fr; remy.boyer@lss.supelec.fr; pierre.duhamel@lss.supelec.fr).

Digital Object Identifier 10.1109/TSP.2005.861106

[1]In the rest of this paper, we will use interchangeably the set of terms *host* and *cover* for the original signal $\mathbf{x}$ and also the set of terms *composite*, *watermarked*, and *public* for the signal $\mathbf{s}$.

[2]In this paper, the words "receiver" and "defender" are equivalently used to refer to the watermark detector. The word "defender" is especially used in a context of Game Theory, by opposition to that of an "attacker."

1053-587X/$20.00 © 2006 IEEE

since the noise introduced by a quantizer is approximately white and uncorrelated with the cover signal. Performances of these watermarking schemes have been studied in terms of capacity and robustness against different types of attacks. When an attacker disrupts a watermark communication, it usually results in two more or less correlated effects: 1) decreasing detection reliability and 2) host signal quality degradation. The relative strengths of these effects naturally depend on the attack but also on the watermarking scheme itself.

A complete characterization of the watermark channel is not available and seems to be very difficult in a general setting. Initially, the analysis of the watermark channel was limited to the additive white Gaussian noise (AWGN) channel where the attack effect is assumed to be additive Gaussian noise-like. Recently, however, theoretical analysis of more sophisticated watermark channels have been published. In [12], Eggers *et al.* proposed a channel model for digital watermarking facing attacks by amplitude scaling and additive white noise (SAWN). In [13], the authors focused on the reindexing channel, which they showed to behave like a linear filter in average. The first intent of this paper is to give means of appropriately characterizing channel attacks in terms of distortion, with an attempt to attenuate the discrepancy between the objective distortion and the perceived one. Our main motivation is to use the proposed approach to derive new insights into the desynchronization plus noise attack, which will be modeled as an additive white Gaussian noise and jitter (AWGN&J) channel. A desynchronization attack can yield a very high probability of error by simply resampling the received signal at other time instants. Throughout this paper, the term "jitter" denotes an attack that introduces random shifts in the nominal sampling instants.

The watermark channel of interest is first presented in Section II. This channel is characterized as follows: 1) the attack is shown to be equivalent to an amplitude scaling plus additive noise, uncorrelated to the signal, and 2) the distortion is measured with respect to the watermarked signal and not to the original signal as considered in [12] and [14] (Section III). Given this channel characterization, the impact of the AWGN&J channel is then evaluated in terms of capacity and error probability, in the case of SS and Costa-based watermarking schemes (Section IV). The trend is put on imperfectly synchronized blind SS-based watermarking. In Section V, the watermarking game [15] is formulated and solved using the introduced *objective* and *perceived* distortions when hiding SS sequences in audio signals. Solving the game sheds light on defender and attacker optimal strategies and provides answers to questions such as the following.

- For a given (perceived) distortion budget, and from the attacker point of view, what part should be allocated to desynchronization, and what part should be allocated to additive noise?
- From the defender point of view, what is the worst distortion?
- Knowing that, is it possible to find countermeasures, so that this distortion is reduced to a tolerable amount?

Notation: In the following, we use boldface font to indicate vectors, e.g., $\mathbf{x}$. $x[n]$ refers to the nth sample of time-dependent signal $x(t)$ sampled at rate $f = (1/T)$ (i.e., $x[n] = x(nT)$) and x_n to the nth element of a vector $\mathbf{x}$. All vectors are row vectors. Random variables are written in *sans serif* font, e.g., x for a scalar random variable and $\mathbf{x}$ for a vector random variable. If $\mathbf{x}$ is an element of a vector space $\mathcal{K}$, $\bar{\mathbf{x}}$ is its conjugate transpose and $\|\mathbf{x}\|$ is its normalized Euclidean norm. For two length-N vectors $\mathbf{u}$ and $\mathbf{v}$, $< \mathbf{u}, \mathbf{v} >$ denotes their normalized inner product, i.e., $\langle \mathbf{u}, \mathbf{v} \rangle \triangleq (1/N) \sum_{i=1}^{N} u_i v_i$. We write $\mathsf{x} \sim p_X(\mathbf{x})$ to indicate that a random variable x is distributed as $p_X(\mathbf{x})$. In this case, $E(\mathsf{x})$ denotes its expectation. The Gaussian distribution with mean μ and square deviation σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. Finally, if $(p, q) \in \mathbb{Z}^2$, $[p : q]$ denotes all integers between p and q.

II. Watermark Channel and Its Model

A watermark channel refers to all operations a watermarked signal may be subject to. These include intentional and nonintentional manipulations. Initially, the watermarking system was designed independently of the channel characterization, while more recent works tune the system to be robust under the worst possible attack in a given category. Thus, one difficulty is to define tractable channel models that accurately fit the possible impairments (either intentional or nonintentional).

The classical communication channels [Binary Symmetric Channel (BSC), AWGN, Rayleigh,...] are not likely to accurately model a watermark channel in real-world scenarios. A better understanding of the watermark channel can be achieved by considering attacks not through their nature but through their impact on the watermarked signal: attacks on the cover signal can in general be modeled easily by filtering plus additive noise. In a general setting, a straightforward model may involve a signal-dependent noise. That is, the noise may be highly correlated with the cover signal. This paper studies a special case of this filtering plus noise channel and provides some tools for increasing its usefulness (through the noise decorrelation process). The proposed approach is then used to focus on desynchronization attacks. Research to assess the impact of desynchronization attacks in digital watermarking has been carried out in two different directions:

1) some watermarking methods attempt to overcome desynchronization attacks by embedding the watermark in an "invariant domain" as in [16] and [17];
2) other schemes are based on an estimation of the attack parameters followed by a compensation as in [18].

Recently, the AWGN&J channel was introduced by Baggen [19] in the context of data storage applications to study the effect of timing jitter in the capacity of magnetic recording media. Insights from this model are used in this paper to investigate the effect of desynchronization attacks on several watermarking schemes.

A. A Distortion Model for a Watermark Attack

Let m be the message to be transmitted. m is usually first encoded into a watermark $\mathbf{w}$ and then embedded into the cover signal $\mathbf{x}$. The resulting composite (watermarked) signal is $\mathbf{s} = \mathbf{x} + \mathbf{w}$. Consider then a general attack $\mathcal{A}$ over the watermark

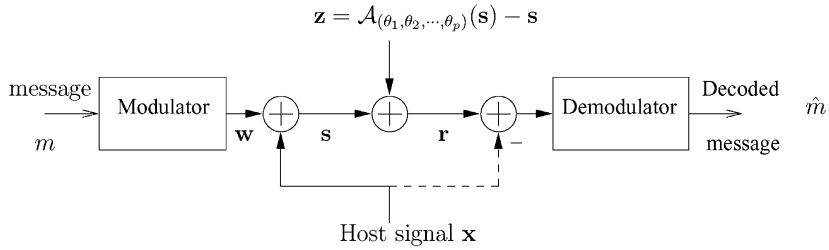


Fig. 1. Abstract communication model for blind (solid line) and nonblind (dashed) watermarking.

channel. In an attempt to fool the receiver, the attacker may use a set of admissible attack parameters $\{\theta_1, \theta_2, \ldots, \theta_p\}$ from some finite domain Θ. The attacker processes the signal $\mathbf{s}$ in such a way that the received signal $\mathbf{r}$ is given by $\mathbf{r} = \mathcal{A}_{(\theta_1,\theta_2,\ldots,\theta_p)}(\mathbf{s})$. Equivalently, the received (attacked) signal $\mathbf{r}$ can be written as the sum of the watermarked signal $\mathbf{s}$ and an interfering signal $\mathbf{z} = \mathbf{r} - \mathbf{s}$. Of course, $\mathbf{z}$ is $(\theta_1, \theta_2, \ldots, \theta_p)$-dependent, and it fully characterizes the attack $\mathcal{A}$, i.e.,

$$\mathbf{z} = \mathcal{A}_{(\theta_1,\theta_2,\ldots,\theta_p)}(\mathbf{s}) - \mathbf{s}. \tag{1}$$

Thus, the watermarking system can be modeled as depicted in Fig. 1. The distortion resulting from the channel attack is generally measured by

$$D_a \triangleq \|\mathsf{r} - \mathsf{s}\| = \|\mathsf{z}\|. \tag{2}$$

After the channel attack, the watermarked signal must remain of sufficient quality. Thus, the channel attack $\mathcal{A}(\theta_1, \theta_2, \ldots, \theta_p)$ has to be upper bounded by a maximum distortion $D_{a\max}$. Clearly, "sufficient quality" should correspond to a perceived distortion but is often measured by (2). This results in

$$\|\mathsf{z}\| \leq D_{a\max}. \tag{3}$$

B. Outline of Our Approach

There are two problems with the classical channel description using the difference signal as given in Section II-A. First, denoting this difference signal z as "noise" is not always accurate: z may contain parts of the composite signal s. In such a situation, z should not be treated as independent noise. Also, "useful" components of z, i.e., those that are highly correlated to the desired signal s must not be counted as noise and should be considered as "useful." Second, the distortion measure D_a does not perceptually characterize the attack $\mathcal{A}$ effect on the watermarked signal $\mathbf{s}$. To cope with these problems, one can note that the attacker effect, that is the additional signal z, can be decomposed into two parts: one that is correlated to the desired signal s and one that is not. The first part is somehow useful and should be "included" in the desired signal s. The second being decorrelated with s can be reasonably considered as noise and will be denoted as "attacker noise" hereafter. The overall approach is equivalent to removing from the signal z the part that is correlated to the watermarked signal and characterizing the attack $\mathcal{A}$ by the remaining part only, i.e., the attacker noise. One straightforward advantage is that the attacker-induced perceived distortion is, likewise, readily measured by the energy (or power) of the "noise" part. This decorrelation-based approach was used previously to model quantization noise (when the high resolution assumption is not valid). More formally, our proposal is to use a "scale plus additive noise" channel model and impose the noise to be uncorrelated with the host signal, as follows:

$$\mathbf{r} = k_z\mathbf{s} + \mathbf{n_z} \text{ under the constraint that } E(\mathbf{s}\bar{\mathbf{n}}_{\boldsymbol{z}}) = 0. \tag{4}$$

Coefficient k_z is easily obtained by imposing $E(\mathbf{n_z}\bar{\mathbf{s}}) = k_z E(\mathbf{s}\bar{\mathbf{s}}) + E(\mathbf{n_z}\bar{\mathbf{s}}) = 0$, which gives

$$k_z = \frac{E(\mathbf{r}\bar{\mathbf{s}})}{E(\mathbf{s}\bar{\mathbf{s}})}. \tag{5}$$

The residual noise $\mathbf{n}_z$ is then given by

$$\mathbf{n}_z = \mathbf{r} - \frac{E(\mathbf{r}\bar{\mathbf{s}})}{E(\mathbf{s}\bar{\mathbf{s}})}\mathbf{s}. \tag{6}$$

Note that, disregarding the value-metric scaling coefficient k_z, the resulting model (4) is additive—just like that given by (1), $\mathbf{r} = \mathbf{s} + \mathbf{z}$. The main difference, however, consists of the fact that unlike signal $\mathbf{z}, \mathbf{n}_z$ is uncorrelated with $\mathbf{s}$. In addition, by opposition to some recent watermarking-related works where specific attacks are addressed as in [12], [15], and [20], we proceed differently here: Given a general attack $\mathcal{A}$ that processes the signal s in such a way that the received signal is $\mathbf{r} = \mathcal{A}_{(\theta_1,\theta_2,\ldots,\theta_p)}(\mathbf{s})$, we begin writing this received signal as $\mathbf{r} = \mathbf{s}+\mathbf{z}$. Next, we derive coefficient k_z and signal $\mathbf{n}_z$ according to (5) and (6) such that the constraint in (4) is satisfied. As a result, the decorrelation process results in a model (4), apparently common at first glance (i.e., of the form $\mathbf{r} = A\mathbf{s}+\mathbf{v}$). However, important differences are 1) parameters k_z and $\mathbf{n}_z$ are not "explicit" in the channel attack, and 2) they depend on the transmitted signal s itself. Another fundamental difference comes from the fact that if ever the signal v involves a part that is correlated to s, the communication model will remove it and include it with composite signal s. The above model will be shown to be particularly useful with desynchronization attacks. Note also that the subscript z in k_z and $\mathbf{n}_z$ is used to point out the model parameters' dependency on the attack $\mathbf{z}$, as clearly shown by (5) and (6). For convenience, we will simply use k and $\mathbf{n}$ to characterize the channel attack every time no ambiguity is possible.

C. Objective and Perceived Distortion Measure

A very simple computation allows the computation of the "error signal" variance in terms of the "noise signal" variance,

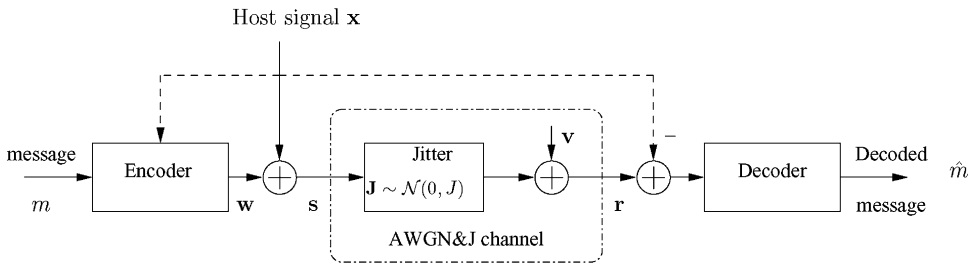


Fig. 2. Additive white Gaussian noise and jitter channel.

which in most circumstances is much smaller. Due to decorrelation between s and n, the objective distortion defined by (2) becomes

$$D_a \triangleq \|\mathbf{r}-\mathbf{s}\| = |k-1|^2\sigma_s^2+\sigma_n^2. \tag{7}$$

Thus, the communication model (4) shows a scale factor k (a luminance change for images, a sound level change for audio signals) and an additive noise n. Both the scaling and the noise inhibit reliable detection of the watermark at the receiver side. However, only the noise n should be considered in evaluating host signal quality loss. Consequently, we assume in this paper that, for perceived distortion measure, the scale factor does not contribute to the distortion. Hence, rather than assuming that the MSE (the norm of the error signal) is a good model for the perceived distortion, we shall devote this role to σ_n^2. Obviously, more accurate models exist, involving human perception models, but the model (4) seems to be a good tradeoff between accuracy and tractability. Simulation results based on real audio signals in the presence of desynchronization attacks show its accuracy.

III. AWGN&J Channel Classical Model

In this section, after a short presentation of the AWGN&J channel, AWGN&J desynchronization effects are investigated differently: 1) using common intersymbol interference (ISI) assumptions commonly known the communication theory in Section III-A and 2) using the model (4) in Section III-B. Both approaches are finally compared. In other words, we will compare the distortions resulting from the two writings of the jittered signal.[3] This comparison will confirm the accuracy of the model (4) and emphasize its particular usefulness for desynchronization attacks characterization.

The AWGN&J channel is an AWGN channel in which the signal s is, in addition to the independent identically distributed (i.i.d.) Gaussian noise v, randomly sampled, as shown in Fig. 2. More precisely, the receiver has to decide on the presence of the watermark based on $s_J[n]+v[n]=s[nT+\tau]+v[n]$ rather than $s[n]+v[n]=s[nT]+v[n]$. The delay τ can be larger than one sampling period T. However, in most cases, the receiver can compensate for any time shifts multiple of T with relatively easy resynchronization procedures. A very easy method will be described in Subsection V-A. In the following, we assume that $\tau=\delta T$ is a fraction of the sampling period T, i.e., $\delta\in[0,1]$.

[3]Note that a plain comparison consists in comparing the distortions resulting from writing the received signal as 1) $\mathbf{r}=\mathbf{s}_J+\mathbf{v}$ and 2) $\mathbf{r}=k\mathbf{s}+\mathbf{n}+\mathbf{v}$, which amounts to comparing $\mathbf{s}_J$ to $\mathbf{r}=k\mathbf{s}+\mathbf{n}$.

The deviation δ is a realization of the process J at time nT and J is assumed to be Gaussian, $\mathsf{J}\sim\mathcal{N}(0,J)$. Depending on the desynchronization (constant shift or random sampling), δ can either be random or constant. Both cases are addressed hereafter. The resulting watermarking communication over an AWGN&J channel is similar to that described in Section II except that, this time, the watermarked signal $\mathbf{s}$ is replaced by $\mathbf{s_J}$ such that $\mathbf{r}=\mathbf{s_J}+\mathbf{v}$. The jittered signal $\mathbf{s_J}$ will be denoted by $\mathbf{s_f}$ in case of a constant scaling and by $\mathbf{s_r}$ in case of random sampling.

Some studies of desynchronization attacks using the AWGN&J channel model [21]–[23] or not [14] already exist. However, in these works, the desynchronization noise is expressed using the ISI term and is assumed to be uncorrelated with the watermarked signal. This assumption, while valid in a traditional communication context, cannot hold in the context of watermarking due to the correlation of signals. Instead, this ISI term must first be processed to remove from it the part that is correlated to the watermarked signal s. Only after that, the remaining part can be assumed to be noise-like. This is a straightforward application of the model (4) above. Using this model will shed light on AWGN&J desynchronization and will prove the inaccuracy of the classical ISI approach. The latter is stated in the following subsection.

A. ISI Approach to AWGN&J Channel Desynchronization

Under appropriate band-limited assumptions, the time-continuous signal $s(t)$ can be reconstructed without error from the sequence $\{s[n]\}_{n\in\mathbb{Z}}$ according to Shannon–Nyquist interpolation, as follows:

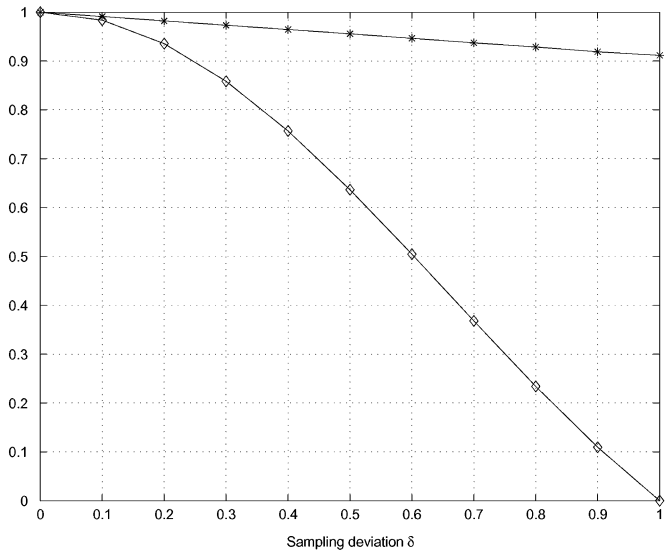
$$s(t)=\sum_{n\in\mathbb{Z}} s[n]\mathrm{sinc}\left(\frac{t}{T}-n\right). \tag{8}$$

This expression will be used to derive expressions for desynchronization noise and induced distortions in presence of a jitter. Whenever required, indexes f and r will refer, respectively, to fixed and random jitters. Equation (8) can be put in the form

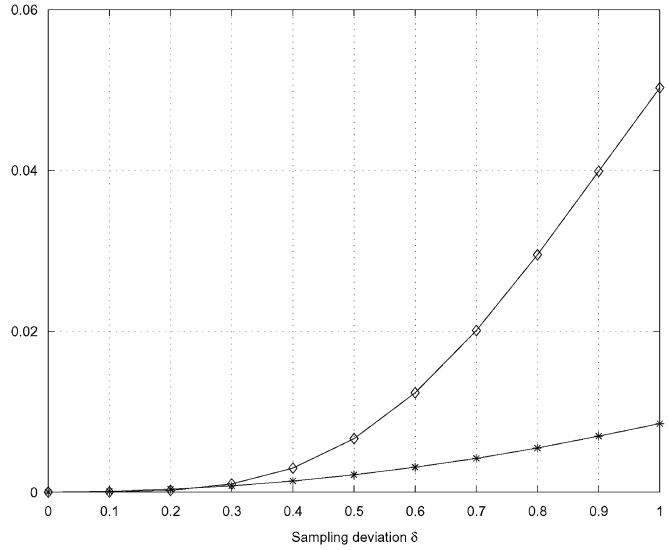
$$s_J[n]=\mathrm{sinc}(\delta)s[n]+\sum_{k\in\mathbb{Z}\setminus\{n\}} s[k]\mathrm{sinc}(n-k+\delta). \tag{9}$$

This equation shows that introducing a constant time shift is equivalent to filtering the watermarked signal or, alternatively, to first attenuating the watermarked signal $s(t)$ and then adding a signal-dependent noise $z_f(t)$ given by

$$z_f(t)=\sum_{k\in\mathbb{Z}\setminus\{n\}} s[k]\mathrm{sinc}(n-k+\delta). \tag{10}$$



(a)



(b)

Fig. 3. Effect of a constant time scaling $\Delta = \delta T$ is investigated differently i) as an additional noise of power $\sigma^2_{z_f}$ resulting from the ISI term (diamond) as considered in [14] and ii) using the proposed model (asterisk). Corresponding scale factors and desynchronization noises are compared. (a) Diagram of dependency of the scale factor k_f on the deviation δ with respect to $\mathrm{sinc}(\delta)$. (b) Equivalent white noise power $\sigma^2_{n_f}$ with respect to $\sigma^2_{z_f}$ stemming from the plain model. Results are obtained with document-to-watermark ratio (DWR) = 20 dB.

The signal $z_f(t)$ can be seen in the context of digital communication as the ISI term. Moreover, in case of a constant shifting, the scaling does not change the overall energy of the signal $s(t)$. Thus, under the uncorrelation hypothesis assumed in [14] and using (9), the distortion due to adding the signal $z_f(t)$ writes

$$\sigma^2_{z_f} = (1 - \mathrm{sinc}(\delta)^2)\sigma^2_s. \tag{11}$$

In the case of random resampling, the variable δ is random. The corresponding distortion can be expressed as in (11) with an additional expectation over all possible values of δ. A much simpler alternative expression of $s_r[n]$ can be obtained by using the

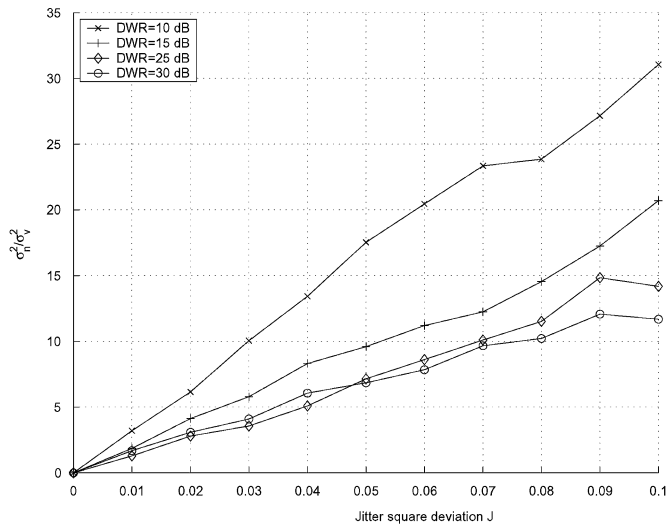


Fig. 4. Diagram of dependency of the desynchronization noise $\sigma^2_{n_r}$ on the jitter square deviation J. Dependency on the composite signal $\mathbf{s} = \mathbf{x} + \mathbf{w}$ is illustrated through that on the document-to-watermark ratio (DWR): the jitter becomes stronger with strong watermarks (lower DWRs).

Taylor–Young series expansion around τ. At first order, $s_r(t) = s(t) + \tau(d/dt)s(t)$. The effect of the jitter can then be viewed as the introduction of an additional signal $z_r(t)$ given by

$$z_r(t) = \tau \frac{d}{dt} s(t). \tag{12}$$

Clearly, the signal $z_r(t)$ depends on the watermarked signal $s(t)$. The corresponding distortion is given by

$$\sigma^2_{z_r} = JE\left(\left[\frac{d}{dt}x(t) + \frac{d}{dt}w(t)\right]^2\right). \tag{13}$$

Note that ISI signals $\mathbf{z}_f$ and $\mathbf{z}_r$ rise directly from interpolation in case of constant shift and random resampling, respectively. Hence, *a priori*, these terms are not necessarily decorrelated from $\mathbf{s}$. An additional decorrelation process (as described in the model (4)) is needed to extract the corresponding *noise* parts.[4] However, in this subsection, we forget for a while the correlation with $\mathbf{s}$ and derive insights into the AWGN&J channel using conventional ISI assumptions.[5] In this case, based on (11) and (13), one can already give some specificities of watermarking channels including jitter.

1) The influence of the jitter depends on the watermarked signal power $\sigma^2_s = \sigma^2_x + \sigma^2_w$. Hence, the well-known embedder strategy consisting of increasing the watermark power σ^2_w to improve detector performance in case of AWGN attacks is no longer the optimum strategy, since at the same time, it enforces the impact of the desynchronization attack by increasing the attack distortion (see Fig. 4).

[4]These noise parts given by model (4) will be denoted by $\mathbf{n}_f$ and $\mathbf{n}_r$, respectively. Fig. 3(b) shows that $\sigma^2_{n_f} \ll \sigma^2_{z_f}$. Therefore, the signal $\mathbf{z}_f$ contains parts of $\mathbf{s}$ that have to be removed from it in order to get the noise part (i.e., $\mathbf{n}_f$). Note also that simulation results with real audio signals support the fact that $\mathbf{z}_f$ and $\mathbf{s}$ are highly correlated.

[5]That is, the ISI term is uncorrelated with the signal $s(t)$ being interpolated.

2) Since the jitter noise is somehow proportional to the original signal, embedding the watermark into a transform domain where the original data is less powerful may alleviate the effect of the jitter.

In the following section, the AWGN&J channel is characterized using the model (4). The goal is, as stated before, to compare the resulting distortions to those being derived using the ISI approach and given by (11) and (13).

B. AWGN&J Channel in Light of Model (4)

Expressing differently the jittered signal $\mathbf{s_J}$, 1) using the model (4) and 2) using (9), we get $k\mathbf{s}+\mathbf{n}=\mathbf{s_J}$. Constant and random time shifts are treated separately.

1) Constant Time Shift: As mentioned before, the scaling does not change the overall energy of the signal $\mathbf{s}$ in case of a constant time shift. This can be shown to result in $k\in[-1,1]$ and $\sigma_n^2=(1-k^2)\sigma_s^2$. In addition, using the model (4), it follows that

$$k_f=\mathrm{sinc}(\delta)+\frac{\langle \mathbf{z}_f,\mathbf{s}\rangle}{||\mathbf{s}||} \tag{14a}$$

$$\mathbf{n}_f=\mathbf{z}_f-\frac{\langle \mathbf{z}_f,\mathbf{s}\rangle}{||\mathbf{s}||}\mathbf{s}. \tag{14b}$$

Fig. 3(a) depicts the dependency of the equivalent scale factor k_f on the sampling deviation δ. It can be seen that k_f decreases with δ but has a much smaller dependency on δ than the factor $\mathrm{sinc}(\delta)$ corresponding to the ISI approach. Note that curves in Fig. 3(a) correspond to a document-to-watermark ratio $(\mathrm{DWR}=10\log_{10}(\sigma_x^2/\sigma_w^2))$ of 20 dB and a watermark-to-noise ratio $(\mathrm{WNR}=10\log_{10}(\sigma_w^2/\sigma_v^2))$ of 0 dB, which are typical values in watermarking systems. Smaller values of k_f can be obtained with stronger watermarked signals. It is worth noting that the model parameter k_f given by (14a) is larger than the scale factor $\mathrm{sinc}(\delta)$ of expression (9) obtained with the ISI approach. In order to further outline the accuracy of the model (4), we compare the power $\sigma_{n_f}^2$ of the noise $\mathbf{n}_f$ to that of the ISI term $\mathbf{z}_f$. The result is depicted in Fig. 3(b). We see that $\sigma_{n_f}^2$ naturally increases with the shift δ. However, unlike the scale factor, $\sigma_{n_f}^2$ is smaller than $\sigma_{z_f}^2$. As stated before, writing the jittered signal $\mathbf{s_f}$ as the sum of two signals, one that is proportional (highly correlated) to it and another that is decorrelated from it, permits the extraction of the noise part $\mathbf{n}_f$. Since $\sigma_{n_f}^2$ is smaller than $\sigma_{z_f}^2$ and $\sigma_{n_f}^2$ is the power of the exact noise term in $\mathbf{r}_f$, it follows that $\mathbf{z}_f$ should not be totally accounted to noise. The difference $\sigma_{z_f}^2-\sigma_{n_f}^2$ corresponds to the power of the part of $\mathbf{z}_f$ that is falsely attributed to noise in the ISI approach.

2) Random Time Shift: Consider now the random jitter case. Again, we have

$$\mathbf{s}+\mathbf{z}_r=k_r\mathbf{s}+\mathbf{n}_r \tag{15}$$

with $\mathbf{n}_r$ uncorrelated with $\mathbf{s}$. Parameters k_r and $\mathbf{n}_r$ can be derived in a way similar to the constant shift case. Intuitively, however, unlike a constant shift, the random variable τ in $z_r(t)$ ensures enough randomness this time so that the objective error may be reasonably considered as uncorrelated with the watermarked signal $s(t)$ (this is checked below by simulation; see Fig. 4). $z_r(t)$ can hence be assimilated to a signal-dependent noise that is approximately decorrelated from $\mathbf{s}$. Therefore, it follows that

$$k_r\approx 1 \tag{16a}$$

$$\mathbf{n}_r\approx\mathbf{z}_r. \tag{16b}$$

Desynchronization experiments, including real audio signals sampled at $f_e=44.1$ kHz show that k_r is most of the time very close to unity and that for a jitter square deviation $J\in[0,1]$, we have $k_r\geq 0.97$. In addition, these tests show that the embedded watermark is inaudible as long as $J\leq 0.04$. Of course, this threshold depends on the signal used and should not be taken for granted, but it already gives an idea about the jitter square deviation range of interest. For this range, simulations show that $k_r\geq 0.99$. The uncorrelation assumption is, unlike the constant time shift, approximately valid for practically all relevant jitter attacks. The jitter acts then as an additive noise of power $\sigma_{n_r}^2=\sigma_{z_r}^2$. However, this noise is dependent on the composite signal $\mathbf{s}=\mathbf{x}+\mathbf{w}$. Fig. 4 illustrates this dependency: here, the cover signal power σ_x^2 is maintained fixed. That (σ_w^2) of the watermark $\mathbf{w}$ varies according to $\mathrm{DWR}=10\log_{10}(\sigma_x^2/\sigma_w^2)\in\{10,15,25,30\}$ dB. Also, the additive Gaussian noise power $\sigma_v^2=\sigma_x^2 10^{-3}$ is fixed. We see that 1) the effect of the jitter (strength of desynchronization noise $\mathbf{n}_r$) increases with the jitter square deviation J (the dependency is approximately linear), and 2) as the power σ_w^2 increases (DWR decreases), the jitter becomes stronger. This illustrates the remark above: increasing the watermark power for more reliable detection in an AWGN&J channel enforces at the same time the effect of the jitter.

In light of the stated above comparison, we conclude the following.

1) The decorrelation hypothesis between $\mathbf{z}_f$ and $\mathbf{s}$ is in general not consistent. The ISI term $\mathbf{z}_f$ is highly correlated to $\mathbf{s}$.
2) Removing from the ISI term the signal-like term results in a more accurate characterization of the attack where the *real* scale factor k_f is larger than $\mathrm{sinc}(\delta)$ and the *real* equivalent additive noise $\mathbf{n}_f$ is much weaker than the ISI term $\mathbf{z}_f$.
3) The objective distortion induced by the scaling attack is

$$D_{af}=|k_f-1|^2\sigma_s^2+\sigma_{n_f}^2. \tag{17}$$

The perceived distortion is given by $\sigma_{n_f}^2$, which is much smaller than that rising directly from the plain model.
4) The *random* jitter has additive signal-dependent-like behavior.

The AWGN&J channel has been characterized in terms of jitter-induced distortions. Since the capacity of any watermarking scheme depends mainly on these distortions,[6] one important thing is to evaluate the performances of this scheme over an AWGN&J channel. The distortions expressed above will help estimate the real performances loss. To that end, two watermarking schemes taken, respectively, from the interference-rejecting and nonrejecting watermarking methods are considered.

[6]It also depends on the embedding distortion $D_E=\sigma_w^2$. In addition, for blind SS embedding, the host signal itself accounts for self-noise and must be included in the channel distortion as shown by (24a).

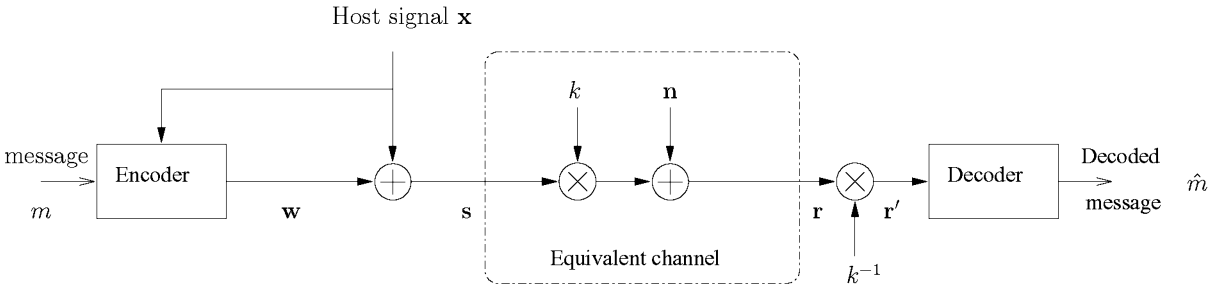


Fig. 5. Blind watermarking as writing on dirty paper over channel (4).

For the former, a brief overview of communication with state information at the encoder is given.

IV. Optimal and Suboptimal Watermarking Schemes Facing AWGN&J Attacks

We assume watermarking of an i.i.d. Gaussian original signal $\mathbf{x} \sim \mathcal{N}(0, \sigma_x^2)$ over a watermark channel characterized by its attack $\mathcal{A}$ such that $\mathbf{r} = k\mathbf{s} + \mathbf{n}$. Such a channel may represent the traditional AWGN channel, the SAWGN channel investigated in [12], or the AWGN&J depicted above or any other watermarking channel (attack). Only the pair $(k, \mathbf{n})$ would vary accordingly. The receiver compensates for the scaling by dividing $\mathbf{r}$ by k to produce the preprocessed signal

$$\mathbf{r}' = \mathbf{x} + \mathbf{w} + \frac{\mathbf{n}}{k}. \tag{18}$$

Thus, the watermark receiver sees an additive white noise (AWN) channel with the effective noise $\mathbf{n}' = \mathbf{n}/k$, with variance σ_n^2/k^2. The watermark capacity for communicating over this effective channel depends only on the cover signal $\mathbf{x}$ and the ratio of the embedding distortion $D_E = \|\mathbf{s} - \mathbf{x}\| = \sigma_w^2$ by the effective channel noise σ_n^2/k^2. The noise power σ_n^2 is related to D_a by (7) which enables the computation of the ratio $\frac{k^2 D_E}{\sigma_n^2}$ as

$$\frac{k^2 D_E}{\sigma_n^2} = \frac{k^2 D_E}{D_a - (k-1)^2(\sigma_x^2 + D_E)}. \tag{19}$$

A. Ideal Costa Scheme

Rather than considering watermarking as communication over a very noisy channel where the host signal $\mathbf{x}$ acts as self-interference (as in SS), it has recently been realized [24], [25] that blind watermarking can be viewed as *communication with side information at the encoder*. The relevant work is the initial Costa "Writing on Dirty Papers" [26]. Fig. 5 depicts a block diagram of blind watermark communication over the channel (4) where the encoder exploits the side-information about the host signal. The scheme originally conceived by Costa is called the Ideal Costa Scheme (ICS) and emerges as a universally good encoding strategy for coding with side information available at the encoder. Based on a huge random codebook, Costa showed that optimal transmitter encodes its message "in the direction" of the interfering signal $\mathbf{x}$ such that the latter does not affect the capacity of the channel, achieving thus the standard Gaussian channel capacity. In our case, the effective watermark-to-noise power ratio $(\sigma_w^2)/(\sigma_{n'}^2)$ is given by (19). Hence, the communication rate under an attack of the form (4) writes

$$\mathcal{R}_{\mathrm{ICS}}^{\mathcal{A}} = \frac{1}{2}\log_2\left(1 + \frac{k^2 D_E}{D_a - (k-1)^2\sigma_s^2}\right). \tag{20}$$

For given D_E, D_a, and σ_x^2, capacity is defined as the supremum of all achievable rates. Alternatively, Moulin *et al.* showed in [27] that the hiding capacity may be formulated as a min–max problem between the information hider and the attacker. The information hider wants a guaranteed rate of reliable transmission under any attack that satisfies a upper-bound constraint on D_a. Conversely, the attacker wants to minimize this rate for any information hiding strategy that satisfies an upper-bound constraint on the embedding distortion D_E. Later, in [28], Moulin *et al.*, with a differently defined distortion measure, have shown that the optimum attack over all possible attacks is a specific scale plus additive white Gaussian noise (SAWGN). For the channel (4) investigated here, capacity is then obtained by minimizing (20) over all possible attacks $k \in [1 - \sqrt{D_a/(\sigma_x^2 + D_E)}, 1 + \sqrt{D_a/(\sigma_x^2 + D_E)}]$. The constraint on admissible scale factor set corresponds to the expression inside the function $\log(\cdot)$ in (20) strictly larger than unity. Otherwise, capacity would be negative and the watermarking system design meaningless. Details of the resolution are skipped here since a very similar game, where the objective function is the detection probability, will be thoroughly studied in Section V. The resolution gives $k_{\mathrm{opt}} = 1 - D_a/(\sigma_x^2 + D_E)$ and

$$\mathcal{C}_{\mathrm{ICS}}^{\mathcal{A}} = \frac{1}{2}\log_2\left(1 + \frac{D_E(\sigma_s^2 - D_a)}{\sigma_s^2 D_a}\right) < \frac{1}{2}\log_2\left(1 + \frac{D_E}{D_a}\right). \tag{21}$$

Note that in general $D_a \ll \sigma_x^2 + D_E$ such that $k_{\mathrm{opt}} \in [1 - \sqrt{D_a/(\sigma_x^2 + D_E)}, 1 + \sqrt{D_a/(\sigma_x^2 + D_E)}]$ is well satisfied. In addition, the term on the right-hand side of (21) is the achievable capacity if there were no attack (which is that of an AWGN channel with SNR D_E/D_a).

B. Traditional Spread-Spectrum

A simplified diagram of basic SS-based watermarking over the channel (4) is shown in Fig. 6. Blind and nonblind reception refer to the fact of having access or not to the cover signal $\mathbf{x}$ at the receiver side. If yes, the decoder subtracts the cover signal $\mathbf{x}$ from the received signal $\mathbf{r}'$ prior to decoding. If not, the decoder performance suffers greatly from host-signal interference. Blind

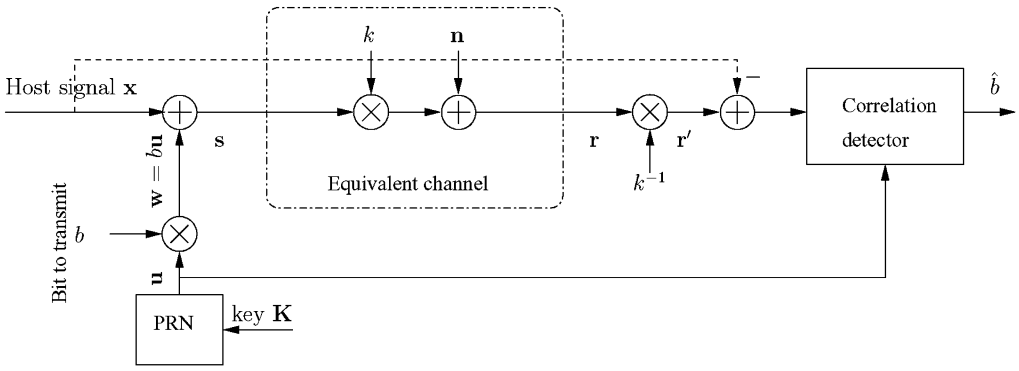


Fig. 6. Blind (solid line) and nonblind (dashed line) SS-based watermarking over channel (4).

and nonblind communication rates under an attack of the form (4) write

$$\mathcal{R}^{\mathcal{A}}_{\text{Blind SS}} = \frac{1}{2}\log_2\left(1+\frac{k^2 D_E}{k^2\sigma_x^2 + D_a - (k-1)^2\sigma_s^2}\right) \tag{22a}$$

$$\mathcal{R}^{\mathcal{A}}_{\text{Non-blind SS}} = \frac{1}{2}\log_2\left(1+\frac{k^2 D_E}{D_a - (k-1)^2\sigma_s^2}\right). \tag{22b}$$

Again, capacity is obtained through a min–max problem resolution. The set of admissible scale factors for the nonblind case is the same as before. For blind SS, it is given by (23), shown at the bottom of the page. The optimization results in the same saddle-point $k_{\text{opt}} = 1 - \frac{D_a}{\sigma_x^2 + D_E}$ as before, which satisfies (23) and for which transmission rates write

$$\mathcal{C}^{\mathcal{A}}_{\text{Blind SS}} = \frac{1}{2}\log_2\left(1+\frac{D_E(\sigma_s^2 - D_a)}{\sigma_s^2 D_a + (\sigma_s^2 - D_a)\,\sigma_x^2}\right) \tag{24a}$$

$$\mathcal{C}^{\mathcal{A}}_{\text{Non-blind SS}} = \frac{1}{2}\log_2\left(1+\frac{D_E\left(\sigma_s^2 - D_a\right)}{\sigma_s^2 D_a}\right). \tag{24b}$$

Capacity loss of both ICS and SS is depicted in Fig. 7. As shown by (21) and (24a), the attack (4) results in significant capacity loss especially for very low WNRs D_E/D_a. As for the AWGN channel, ICS outperforms SS for almost all values of D_E/D_a. Note, however, that ICS-capacity reduction is larger than that for SS: ICS is less robust than SS facing attacks of the form (4). This fact will be supported by simulations over an AWGN&J channel [see Fig. 8(a) and (b)]. Also, in case of very strong attacks, ICS and SS capacities fall to the same values and ICS presents no gain over SS. These attacks are, however, sufficiently strong to practically impair any communication and are, consequently, not relevant in real applications. For reasonable WNRs ($10\log_{10}(D_E/D_a) > -16$ dB), ICS remains more efficient.

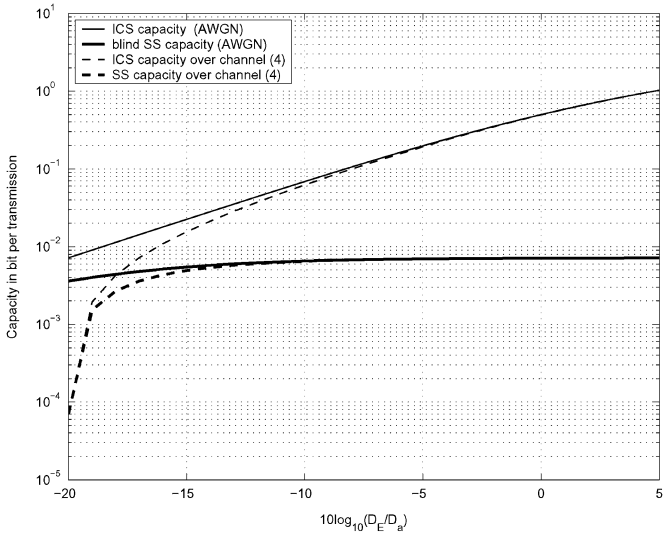


Fig. 7. Capacity loss of both ICS and blind SS scheme under the influence of an attack of the form (4). The result is depicted for DWR $= 10\log_{10}(\sigma_x^2/D_E) =$ 20 dB. For strong attacks, ICS and SS capacities fall to same values. ICS becomes more sensitive than SS.

Now, focus on the special case of an AWGN&J channel. This channel has been shown to be a special case of attacks of the form (4), with parameters k and $\mathbf{n}$ given by (14a), (14b), (16a), and (16b). Hence, ICS and blind SS capacities over an AWGN&J channel are readily given by (21) and (24a), respectively, and are shown in Fig. 7. However, since these capacities are obtained through a min–max resolution, they correspond to the achievable rate under the optimum attack (k_{opt}). More insights can be obtained using achievable rates (20) and (22a) instead of capacities as stated below.

C. Application to AWGN&J Channels

Here, unlike capacity that is derived analytically using k_{opt}, we want to see how transmission rates $\mathcal{R}_{\text{ICS}}^{\text{AWGN\&J}}$ and

$$k \in \left[\frac{\left(\sigma_x^2 + D_E\right) - \sqrt{\left(\sigma_x^2 + D_E\right)^2 - \left(\sigma_x^2 + D_E - D_a\right)D_E}}{D_E}, \frac{\left(\sigma_x^2 + D_E\right) + \sqrt{\left(\sigma_x^2 + D_E\right)^2 - \left(\sigma_x^2 + D_E - D_a\right)D_E}}{D_E}\right]. \tag{23}$$

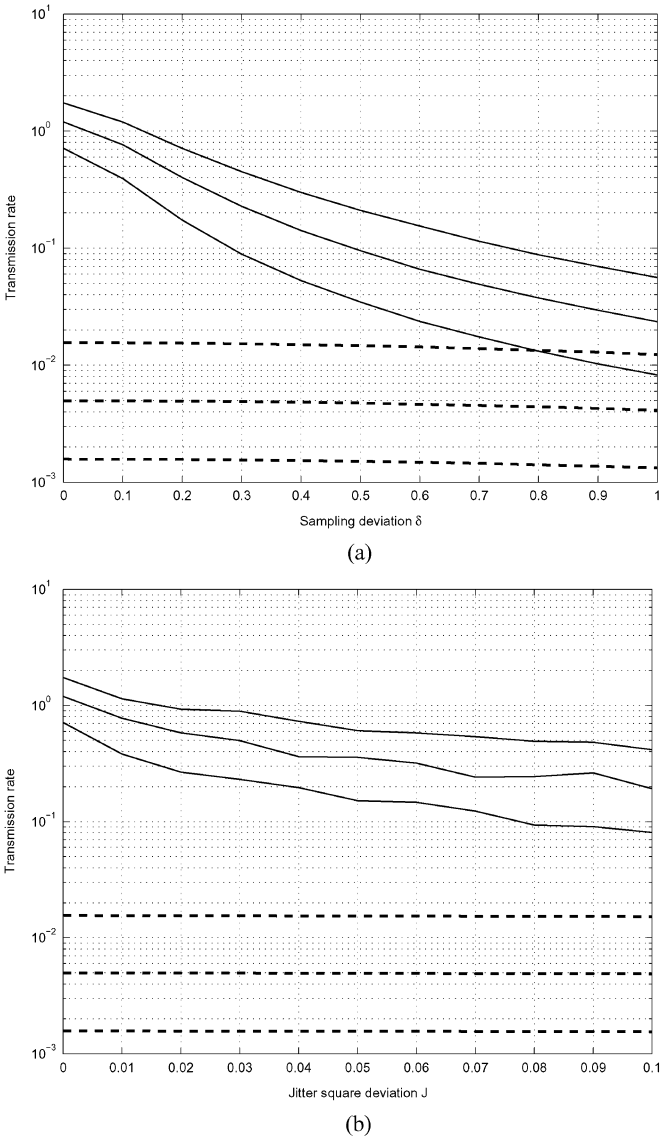


Fig. 8. Transmission rate loss of both ICS (solid line) and blind SS (dashed line) over an AWGN&J channel. Gaussian noise $\mathbf{v}$ is such that $\sigma_v^2 = 10^{-3}\sigma_x^2$. (a) Composite signal $\mathbf{s}$ is scaled with $\Delta = \delta T$. (b) $\mathbf{s}$ is randomly resampled using the jitter $\mathsf{J} \sim \mathcal{N}(0, J)$. With both schemes and under both attacks, transmission rate degrade with DWR. From bottom to top: DWR $= 25$, 20, and 15 dB.

$\mathcal{R}_{\text{blind-SS}}^{\text{AWGN\&J}}$ degrade in presence of a jitter J. That is, we are interested in the current jitter being used and not the optimal one as in capacity analysis. Simulations are required to compute parameter k and noise $\mathbf{n}$ in (20) and (22a). We proceed as follows: Given some jitter (shift δ), the composite signal $\mathbf{s}$ is interpolated resulting in $\mathbf{s}_{\boldsymbol{J}}$. Next, the equivalent attack (scaling k and noise $\mathbf{n}$) is derived, and white Gaussian noise $\mathbf{v}$ is added. The received signal is $\mathbf{r} = k\mathbf{s} + \mathbf{v}_{\text{eq}}$, where $\mathbf{v}_{\text{eq}} = \mathbf{v} + \mathbf{n}$ is the overall channel noise.

1) Rate Loss Under Constant Time Shift Attacks: Fig. 8(a) depicts transmission rates given by (20) and (22a) using expressions given by (14a) and (14b) for the scale factor k and the noise $\mathbf{n}$. For both ICS and SS, these are shown for three values of DWR $= 10\log_{10}(\sigma_x^2/D_a)$. We observe that the ICS transmission rate drastically decreases if the sampling deviation δ increases. This illustrates the loss in ICS capacity already shown in Fig. 7 and particularly apparent for low WNR $= 10\log_{10}(D_E/D_a)$: As δ is close to unity, the jitter-induced distortion D_a is large and WNR is low. It is worth noting that IC- rate degradation reveals a more general setting: Almost all quantization-based embedding schemes are highly sensitive to scaling. When scaled, the received signal is rounded to a bad quantization cell center. Blind SS, however, is almost insensitive to scaling but performs far below ICS. This is particularly useful for the design of watermarking systems in situations the transmitted signal may be scaled in the channel: ICS should be preferred to SS for applications where a great amount of information is to be transmitted. However, SS may be used for applications where the transmission rate is not the main issue and where robustness against scaling is highly appreciated. The latter applications are referred to as "one-bit watermarking" problems in digital watermarking. Another important remark rises from comparing the transmission rates corresponding to the same shift δ but different values of DWR. It can be seen that the higher the DWR, the larger the rate loss. This is not contradictory with (11) because the embedding distortion D_E is reduced as well so that the transmission rate broadly decreases for large DWR.[7]

2) Rate Loss Under Random Jitter Attacks: The effect of a random jitter $\mathsf{J} \sim \mathcal{N}(0, J)$ combined with an AWGN $\mathbf{v}$ attack on a watermarked signal $\mathbf{s} = \mathbf{x} + \mathbf{w}$ is depicted in Fig. 8(b). For the same reason as above, we concentrate on attacks with jitter square deviation $J < 0.04$. Again, we use (20) and (22a), where k_r and $\mathbf{n}_r$ are replaced by (16a) and (16b), respectively. As for the constant time shift case, we observe that ICS-rate reduction is larger than that of SS, which is almost insensitive to the jitter. Also, though large DWRs result in small distortions as previously shown by (13), the decrease in the embedding distortion D_E causes the transmission rate to degrade.

Now compare the ICS-rate loss to that in case of a constant shift attack. It is worth noting that the rate loss is larger when facing constant shifting. This is not completely surprising: Remember that the random jitter attack has been shown to behave as additive noise. With ICS, whose practical implementations are forms of quantization, scaling is more harmful than adding noise. This fact will be supported by the game theory resolution in Section V. The remainder of the paper is devoted to providing insights into both the *optimum attack* and the *optimum defense*. By "optimum," it is meant the "the best strategy" in a game theory context. The Watermarking Game does not have universal solutions, and both attacker and defender should adapt to each other. Here, the game is first briefly reviewed and then solved in the case of an AWGN&J attack and blind SS. We also provide a simple means of circumventing constant time shift attacks.

V. Game Theory Approach to AWGN&J Channels

In a robust watermarking transmission context, the embedder must design his embedding scheme so that the watermark survives the worst possible attack. Conversely, the attacker has to perform the optimal attack that best impairs the watermark, for

[7]Note that small increasing of the DWR is obtained through decreasing the embedding distortion $D_E = \sigma_w^2$. Cover signal and Gaussian noise powers σ_x^2 and σ_v^2 are maintained fixed as above.

a given distortion budget. The resulting optimization problem (game theory problem) is often formulated as a max–min (or min–max) problem. The criterion to be optimized is the detection (or, equivalently, error) probability in case of one-bit watermarking and the watermarking capacity in case of data hiding. Since capacity has already been optimized in Section IV and since for many watermarking applications, the most significant criterion is reliable detection, we concentrate on the one-bit watermarking. We consider the criterion of detection probability. The game watermarking has been thoroughly studied in the case of an AWGN channel [20], [29], [30]. In [15], Moulin *et al.* discussed the case of attacks by filtering and additive noise. In [12], Eggers *et al.* considered attacks by amplitude scaling and additive noise. But, by opposition to the following, only objective distortions were used and were evaluated with respect to the original host signal, not to the watermarked signal.

In an AWGN&J channel, the attacker can desynchronize the signal and add noise as well. In this section, we answer the following three questions.

1) If ever the attacker has a perceived distortion budget, with two ways of using it—either by introducing jitter or by using additive noise or any combination of both—what is the best strategy?
2) Conversely, what is the best tuning for the defender knowing the best potential attack?
3) Is there means for the defender to find countermeasures to the attacker strategy (put some limits to the efficiency of its optimal strategy)?

We begin by answering 3). The main difficulty in synchronizing a randomly scaled received signal stems, as stated above, from the fact that random time scaling the watermarked signal broadly behaves like adding noise (disregarding that this noise is signal-dependent as given by (12)). In case of a constant time shift, however, the receiver should be able to reverse the effect of scaling. The main solutions that have been proposed can be divided into two categories: 1) embedding of a pilot sequence as classically used in traditional communication and 2) using a correlation-based alignment algorithm [31]. While pilot sequences present an additional source of weakness if ever intercepted by an attacker, the algorithm in [31] has good matching properties but requires the availability of (a copy of) the original signal at the receiver side.[8] This algorithm consists of computing the maximum normalized correlation between the pirated (attacked) signal and the original. Here, we propose a cross-correlation-based matching process that we denote by "multiple correlation test." This procedure is similar to that in [31] but does not require knowledge of the original content at the receiver. Blind resynchronization is made possible by using the watermark instead of the original signal for correlation computation. Having access to the watermark $\mathbf{w}$ at the receiver side is commonly assumed in a one-bit watermarking context. The aim is to mark user-specific contents with the same small watermark.

[8]In [31], Schonberg *et al.* proposed this algorithm in the context of fingerprinting, which is indeed an application where availability of the original signal is usually assumed. Here, we focus on detecting the same watermark embedded in several different contents.

A. Preventing Constant Shift

Suppose the attacker performs, in addition to the AWGN $\mathbf{v}$, a time shift $\Delta = \delta T$ with $\delta \in [0,1]$. The restriction $\delta \in [0,1]$ is due to the fact that, with a cross-correlation-based resynchronization procedure, the defender can compensate for any T-multiple time shift: The receiver searches for the maximum cross correlation between the received (attacked) signal $\mathbf{r}$ and the watermark $\mathbf{w}$ and realigns $\mathbf{r}$ before proceeding to detection so that he or she gets rid of any T-multiple scaling. We concentrate then on the case $\delta \in [0,1]$. As stated above, the received watermarked (and attacked) signal is given by $\tilde{r}(t) = r(t+\Delta) = \tilde{x}(t) + \tilde{w}(t) + v(t)$, where $\tilde{x}(t), \tilde{w}(t)$ and $\tilde{r}(t)$ are, respectively, desynchronized signals $x(t), w(t)$ and $r(t)$. The analysis below shows that desynchronization is much more harmful than white noise. Therefore, it is very important for the defender to maintain this contribution to a reasonable level. One possible way is to interpolate the received signal, so that the receiver performs several shifts of the watermark $w(t)$ along the time axis, and proceed to correlation tests with $\tilde{r}$ for each of these shifted versions. Depending on the number of correlation tests the decoder can perform, the receiver can maintain the maximum time shift to a desired bounded value. A large number of tests at the receiver side, however, increases the computational complexity. Therefore, there will be a tradeoff between computational complexity and optimality. Suppose the receiver is able to perform $M(\geq 2)$ tests. Let $\tilde{\mathbf{w}}^{(k)}$ denote the watermark signal shifted by $kT/M, k \in [1 : M-1]$, that is $\tilde{w}^{(k)}[n] = \tilde{w}[nT + kT/M]$ and $c^{(k)} = (\langle \tilde{\mathbf{r}}, \tilde{\mathbf{w}}^{(k)} \rangle)/(\sqrt{||\tilde{\mathbf{r}}||||\tilde{\mathbf{w}}^{(k)}||})$, the correlation coefficient between the received signal and $\tilde{w}^{(k)}(t)$, with $c^{(0)} \triangleq (\langle \tilde{\mathbf{r}}, \mathbf{w} \rangle)/(\sqrt{||\tilde{\mathbf{r}}||||\mathbf{w}||})$. In order to bound Δ within an interval of length T/M, the receiver determines $k_0 \in [1 : M-1]$ according to

$$k_0 = \underset{k \in [1:M-1]}{\operatorname{argmax}} \, c^{(k)}. \tag{25}$$

k_0 represents the location index for which $\tilde{r}(t)$ optimally matches the signal $r(t)$ when scaled back by $k_0 T/M$. Next, the receiver proceeds to detection using the aligned signal $\tilde{r}(t - k_0 T/M)$. Likewise, the residual desynchronization is smaller than T/M, and the attacker should not waste energy in further desynchronizing the watermarked signal $s(t)$. The cost the receiver has to pay in order to maintain the maximum time shift to a (small) bounded value is the computation of M correlations. From the analysis outlined in the first part of the paper, this bounding (a parameter of the transmitter) is absolutely required; otherwise, desynchronization-induced noise would increase to very large values, resulting in very poor detection performances.

B. Game Theoretical Formulation

We consider the embedding of one bit of information $b \in \{0,1\}$ into an original data $\mathbf{x}$ of length N, assumed to be Gaussian, $\mathbf{x} \sim \mathcal{N}(0, \sigma_x^2)$. The watermark signal is given by $\mathbf{w} = b\mathbf{u}$—where chips w_i are mutually independent with respect to $\mathbf{x}$. The sequence $\mathbf{u}$ is produced by a pseudorandom number generator (PRN) using a secret key $\mathbf{K} \in K$. Its elements are equal to $+\sigma_u$ or $-\sigma_u$ (see Fig. 6). Also, according to

Kerkhoff's principle, we suppose the attacker knows the used watermarking scheme. The embedder, however, not having access to the attacker scale factor k, does not normalize the received signal. In this context, the attacker may either add white Gaussian noise, desynchronize the watermarked signal, or perform both operations as long as the overall attack distortion D_a is upper-bounded by a certain tolerance level $D_{a\max}$. On the other side, the embedder chooses the appropriate length N of original data, the number M of correlations to be performed, and watermark power σ_w^2 subject to a certain maximum embedding distortion $D_{E\max}$.

1) Detection Probability: Detection is based on the sign of $r = \frac{\langle \mathbf{r},\mathbf{u}\rangle}{\|\mathbf{u}\|}$ where the received signal is $\mathbf{r} = k\mathbf{s} + \mathbf{n} + \mathbf{v}$ and $r = kb + kx + n + v$, with

$$\begin{cases} x = \frac{\langle \mathbf{x},\mathbf{u}\rangle}{\|\mathbf{u}\|} \sim \mathcal{N}\left(0, \frac{\sigma_x^2}{N\sigma_w^2}\right) \\ v = \frac{\langle \mathbf{v},\mathbf{u}\rangle}{\|\mathbf{u}\|} \sim \mathcal{N}\left(0, \frac{\sigma_v^2}{N\sigma_w^2}\right) \\ n = \frac{\langle \mathbf{n},\mathbf{u}\rangle}{\|\mathbf{u}\|} \sim \mathcal{N}\left(0, \frac{\sigma_n^2}{N\sigma_w^2}\right). \end{cases}$$

Therefore, the probability density function (PDF) of r is given by $r \sim \mathcal{N}(kb, (k^2\sigma_x^2 + \sigma_n^2 + \sigma_v^2)/(N\sigma_w^2))$. Natural performances measure for the one-bit watermarking problem are probability of false positive (false alarm) P_{FA}, false negative (miss detection) P_{MD} and probability of detection P_{d}. The watermark detection problem can be formulated as a hypothesis test:

$$\begin{cases} H_0 : b = 0 :\Rightarrow \text{no watermark}, \\ H_1 : b = 1 \Rightarrow \text{watermark found}. \end{cases}$$

The detector decides that a watermark is present if $r > \nu$, where ν is some detection threshold that controls the trade-off between false positive and false negative decisions. These probabilities are given by

$$\begin{aligned} P_{\mathrm{FA}} &= P(r > \nu \,|\, H_0) \\ &= \frac{1}{2}\mathrm{erfc}\left(\nu\sqrt{\frac{N\sigma_w^2}{2\left(k^2\sigma_x^2 + \sigma_v^2 + \sigma_n^2\right)}}\right) \qquad (26a) \\ P_{\mathrm{MD}} &= P(r < \nu \,|\, H_1) \\ &= \frac{1}{2}\mathrm{erfc}\left((k-\nu)\sqrt{\frac{N\sigma_w^2}{2\left(k^2\sigma_x^2 + \sigma_v^2 + \sigma_n^2\right)}}\right) \qquad (26b) \\ P_{\mathrm{d}} &= P(r > \nu \,|\, H_1) \\ &= \frac{1}{2}\mathrm{erfc}\left((\nu-k)\sqrt{\frac{N\sigma_w^2}{2\left(k^2\sigma_x^2 + \sigma_v^2 + \sigma_n^2\right)}}\right). \qquad (26c) \end{aligned}$$

The parameter k is unknown. Noncoherent detection theory provides several techniques to solve detection problems with unknown parameters. Below the Neyman–Pearson approach is first reviewed and then applied to derive consistent choice of parameter ν. More details about such choice can be found in [32].

2) Neyman–Pearson Criterion for Threshold Selection: Subject to a constraint on the maximum acceptable probability of false positive (false alarm), the test consists of minimizing the probability of false negative (miss-detection).

For example, a maximum allowable probability of false alarm $P_{\mathrm{FA}_{\max}} = 10^{-6}$ leads to a threshold $\nu \approx 3.3\sqrt{2\sigma_r^2}$. In [6], it is stated that to improve the characteristics of robustness against attacks, the new threshold should be evaluated directly on the watermarked and possibly attacked signal $\mathbf{r}$. This results in the following choice:

$$\begin{cases} \nu \approx 3.3\sqrt{2\frac{k_f^2\sigma_x^2 + \sigma_v^2 + \sigma_{n_f}^2}{N\sigma_u^2}} & \text{for a constant shift} \\ \tilde{\nu} \approx 3.3\sqrt{2\frac{\sigma_x^2 + \sigma_v^2 + JE\left[\left(\frac{d}{dt}s(t)\right)^2\right]}{N\sigma_u^2}} & \text{for a random shift.} \end{cases}$$

3) Max–Min Criterion: Over an AWGN&J channel, the detection probability, denoted by $P_d^{\mathrm{AWGN\&J}}$, is given by (26c). Also, we assume as stated in [14], that meaningful embedding and attack distortions should satisfy

$$0 \le D_{E\max} \le D_{a\max} \le \sigma_x^2. \qquad (27)$$

Taking into account the defender ability to perform the *multiple correlation* test described above, the scaling must satisfy $\delta \le \frac{1}{M}$. Due to correlation computing cost, we suppose $M \le M_{\max}$, M is a parameter of the defender. For the same reason (correlation based detection cost), very large values of the signal length N are not allowed ($N \le N_{\max}$). As for the bound on M, that on N should ensure good compromise between detection performance and computing complexity. The embedder wants to maximize $P_d^{\mathrm{AWGN\&J}}$, and the attacker wants to minimize it under constraints pair $(D_{E\max}, D_{a\max})$. The problem is then naturally formulated as a game between the embedder and the attacker and can be written as

$$\max_{D_E \le D_{E\max}} \min_{D_a \le D_{a\max}} P_d^{\mathrm{AWGN\&J}}. \qquad (28)$$

This optimization problem is solved in the following section for both constant and random scalings.

C. Solving the Watermarking Game

The attack distortion has been shown above to be given by $D_a = |k-1|^2\sigma_s^2 + \sigma_n^2 + \sigma_v^2$. Let us first determine the part of the distortion budget that the attacker should allocate to noise and that to allocate to jitter, so that the detection performances are maximally reduced. By considering the proposed model $k\mathbf{s} + \mathbf{n} + \mathbf{v}$ in which $\mathbf{n}$ and $\mathbf{s}$ are uncorrelated as required by model (4), there is *a priori* no difference in nature between $\mathbf{n}$ and $\mathbf{v}$ (disregarding $\mathbf{n}$ dependency on the signal $\mathbf{s}$). Hence, we divide the global attack distortion D_a into two parts: D_k due to scaling and D_v due to the additive noise such that

$$\begin{aligned} D_v &= \sigma_v^2 + \sigma_n^2 = \alpha D_a \qquad (29a) \\ D_k &= (k-1)^2(\sigma_w^2 + \sigma_x^2) = (1-\alpha)D_a. \qquad (29b) \end{aligned}$$

$\alpha \in [0,1]$ characterizes the tradeoff between the two components of the global distortion. We will refer to the case $\alpha = 1$ as the *all-noise* case since the overall attack is equivalent to that of adding the noise quantity $\mathbf{n} + \mathbf{v}$. Similarly, we will refer to the case $\alpha = 0$ as the *all-desynchronization* case since it corresponds to a channel attack by time-axis scaling only. Any other attack with $\alpha \in]0,1[$ will be termed as *mixed* since both adding noise and scaling are required.

1) Case of a Constant Time Shift Attack: Prior to revealing optimum attacker and defender strategies, we suppose the proper resynchronization procedure investigated above was used and reformulate the optimization problem. The *multiple*

correlation test does not change the criterion to be optimized. However, the ranges of the optimization variables M, N and α are modified (as it will be shown in (33)). Intuitively, this follows from the fact that countermeasurements performed by the defender naturally reduce the set of admissible parameters for the attacker. In our case, a lower bound on δ can be shown to result in a lower bound on the attack scale factor k: Let $h(\cdot)$ be the function relating k_f to δ. For an explicit expression of $k_f = h(\delta)$, we need to combine (14a) and (10). Namely, we need to invert (10) to get δ and replace it in (14a). Unfortunately, no explicit formula for parameter δ can be derived from (10). However, parameter k dependence on δ is depicted in Fig. 3(a). Using this curve will be shown to be sufficient to bypass the difficulty raised above.[9] Of course, this results in an approximate solution, but it is already enough to answer questions raised while formulating the game. Also, the curve depicted in Fig. 3(a) corresponds to specific values of WNR $=$ 0 dB and DWR $=$ 20 dB, but this would not change the concluding remarks related to relative noise and desynchronization effects and stated at the end of this section. In other words, not knowing the watermark power σ_w^2 does not matter since we only use the monotonously decreasing property of $h(\cdot)$ in solving the game. This property implies a lower bound on the set of admissible values for the scale factor k. The constraint $\Delta \leq T/M$ gives δ in $[0, 1/M]$. There exists then a lower bound on k_f, say $k_{\min} \in [0,1]$ such that $k_{\min} = h(1/M)$ and $k_f \geq k_{\min} \forall \delta \in [0, 1/M]$. Using (29b), we obtain

$$k_f = 1 - \sqrt{\frac{(1-\alpha)D_a}{\sigma_w^2 + \sigma_x^2}}. \tag{30}$$

Similarly, lower bound constraint $k_{\min}$ on k_f implies a similar constraint on α: There exists $\alpha_{\min} \in [0,1]$ such that $\alpha \in [\alpha_{\min}, 1] \forall \delta \in [0, \frac{1}{M}]$, which when combined with (30), gives

$$\alpha_{\min} = 1 - \frac{(1-k_{\min})^2(\sigma_x^2 + \sigma_w^2)}{D_a}. \tag{31}$$

Furthermore, inequality $\alpha_{\min} \geq 0$ gives $D_k \leq D_{k\max} = (1 - \alpha_{\min})D_a$. The latter upper bound on D_k can be understood this way: A part of the overall distortion D_a must be allocated to noise. It is worth noting that scenarios corresponding to $\alpha = \alpha_{\min}$ and $\alpha = 1$ are worthy of some discussion.

1) $\alpha = \alpha_{\min}$: With respect to cases $\alpha = 0$ (all desynchronization) and $\alpha = 1$ (all noise), this case corresponds to a mixed situation where the attacker should both add noise and desynchronize the signal. The global objective distortion D_a results then from both i) an attack by amplitude scaling causing an objective distortion $D_k = (k_{\min} - 1)^2 \sigma_s^2$ and ii) an attack by additive noise of power $D_v = D_a - (k_{\min} - 1)^2 \sigma_s^2$. With regard to these distortions, one can note the following:
   a) increasing the watermarked signal power σ_s^2 enforces the distortion D_k due to the scale factor with respect to that of the equivalent noise $\mathbf{v}_{\text{eq}} = \mathbf{v} + \mathbf{n}_z$;
   b) increasing the admissible set of correlations M causes the distortion D_k to decrease; conversely, the additive noise distortion D_v increases.

   Imposing a lower bound on α gives $D_k \leq (1 - \alpha_{\min})D_a$ and prevents the receiver from the all-desynchronization attack. However, this is achieved by the cost of a certain signal processing complexity at the receiver side implicitly shown here through the defender parameter M.
2) $\alpha = 1$: This is the case of an attenuating additive noise $\mathbf{v}$. The attack is of type AWGN and traditional watermarking game solutions apply. Most prominent examples of these can be found in [15] and [20].

We now rewrite the detection probability (26c) with k_f and $\sigma_v^2 + \sigma_{n_z}^2$ expressed by (30) and (29a), respectively. The resulting formula can be expressed as a function of both the setting of defender parameters $\{N, M, \sigma_w^2\}$ and that of the attacker $\{\alpha, D_a\}$, as

$$P_d^f\left(\left[N, M, \sigma_w^2\right], [\alpha, D_a]\right) = \frac{1}{2}\text{erfc}\left((\nu - k_f)\sqrt{\frac{N}{2}\frac{\sigma_w^2}{\left(1 - \sqrt{\frac{(1-\alpha)D_a}{\sigma_x^2+\sigma_w^2}}\right)^2 \sigma_x^2 + \alpha D_a}}\right) \tag{32}$$

where $\nu_f = 3.3\sqrt{\frac{2}{N}\frac{k_f^2\sigma_x^2 + \alpha D_a}{\sigma_w^2}}$. Note that P_d^f depends on M through the admissible set values of parameter α. Consequently, the max–min problem (28) specializes as

$$\max_{[N,M,\sigma_w^2]} \min_{[\alpha, D_a]} P_d^f\left(\left[N, M, \sigma_w^2\right], [\alpha, D_a]\right) \tag{33}$$

where

$$\begin{cases} N \leq N_{\max} \\ M \in [0, M_{\max}] \\ 0 < \sigma_w^2 \leq D_{E\max} \\ \alpha \geq 1 - \frac{(1-k_{\min})^2(\sigma_x^2+\sigma_w^2)}{D_a} \\ D_a \leq D_{a\max}. \end{cases}$$

We now turn to the attacker and defender optimum strategies.

a) Optimum Attack: For a given set of defender parameters $\{N, M, \sigma_w^2\}$, the detection probability $P_d^f([N, M, \sigma_w^2], [\alpha, D_a])$ writes as a function of the attacker parameters pair (D_a, α). A two-dimensional (2-D) plot of this function is shown in Fig. 9(a). We see that the detection probability decreases with D_a. The optimal attack corresponds, as intuitively expected, to a maximized global distortion $D_a = D_{a\max}$. Minimizing then P_d^f over $\alpha \in [\alpha_{\min}, 1]$ for given values of N, M, and σ_w^2 provides the optimal scaling attack. Fig. 9(b) clearly shows that the detection probability is maximally reduced for $\alpha = \alpha_{\min}$: This corresponds to a *mixed* attack and refers to the fact that desynchronization is much more efficient than noise in impairing the detection probability for a given distortion budget. Note that without the multiple correlation procedure, the optimal solution would be $\alpha = 0$, that is the so denoted all-desynchronization.

In summary: when given the possibilities of adding white noise, desynchronizing the watermarked signal by constant scaling or performing both operations, desynchronization turns out to be optimal. However, to cope with appropriate defender countermeasurements (the multiple correlation test described above), the attacker is constrained to a maximum allowable attack distortion budget D_a. Thus, its best strategy is first to

[9] It will be shown that for the final solution, we need just bounds on the value-metric scaling parameter k. These can already be obtained from Fig. 3(a).

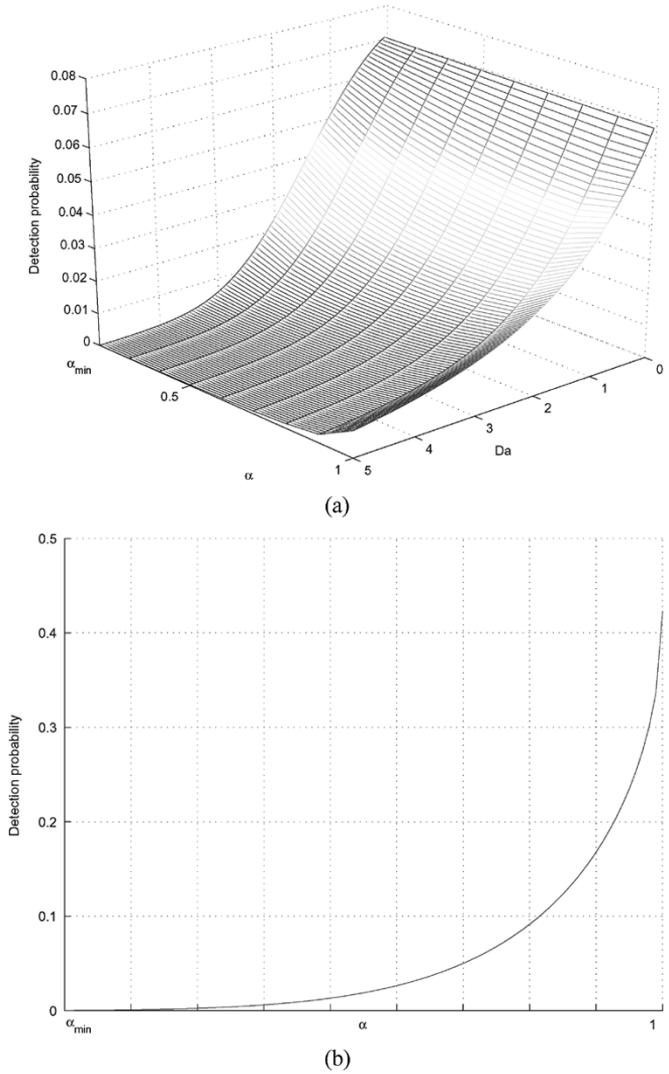


Fig. 9. Optimum attack: Detection probability has to be minimized over the set of attacker parameters $\{\alpha, D_a\}$. The 2-D plot (top) shows that maximally reduced detection is obtained with large attack distortion $D_a = D_{a\max}$. The corresponding detection probability $P_{\rm d}(D_a = D_{a\max})$ plotted over $\alpha \in [\alpha_{\min}, 1]$ (bottom) shows that the optimal attack is mixed, $\alpha_{\rm opt} = \alpha_{\min}$.

desynchronize the signal and then fulfill the remainder of the distortion budget by adding the appropriate noise amount.

As a result, the received signal corresponding to the worst attack can be expressed as

$$\mathbf{r} = \left(1 - \sqrt{\frac{(1-\alpha_{\min})D_a}{\sigma_w^2 + \sigma_x^2}}\right)\mathbf{s} + \mathbf{v}_{\rm eq}, \tag{34}$$

where $\mathbf{v}_{\rm eq}$ is such that $\sigma_{v_{\rm eq}}^2 = \alpha_{\min} D_a$.

Recall that in real-world scenarios, the attacker chooses the parameters Δ and σ_v^2 (not k_f and α). The optimal attack turns to correspond to the combination of the following single attacks:

- a time shift $\Delta = (T/M)$;
- an additive noise of power $\sigma_v^2 = D_a - \sigma_{s_J}^2$, where $s_J(t) = s(t + (T/M))$.

b) Optimized Defense: We now turn to characterize the optimized defense that best prevents the defender from the worst attack ($\alpha = \alpha_{\min}$). After replacing attacker parameters by corresponding optimum values derived above, the detection probability (32), depending only on N, M, and σ_w^2, writes

$$P_d^f\left([N, M, \sigma_w^2]\right) = \frac{1}{2}\mathrm{erfc}\left((\nu_f - k_{\min})\sqrt{\frac{N}{2}\frac{\sigma_w^2}{k_{\min}^2\sigma_x^2 + \alpha D_a}}\right). \tag{35}$$

The aim of the defender is to maximize this worst detection probability (35). With qualitative considerations, we can already determine the optimum value of M: The detection probability depends on M through $k_{\min}$. Larger values of $k_{\min}$, corresponding to a tight range of δ, are better for the defender. Then, disregarding computational complexity, M should be maximized. An optimum defender choice would then intuitively correspond to $M = M_{\max}$. The resulting P_d^f depicted in Fig. 10(a) shows that the watermark embedding power should be maximized, namely $\sigma_{w\rm opt}^2 = D_{E\max}$. Also, the parameter N should have the largest possible value, i.e., $N = N_{\max}$. Hence, the optimum defense corresponds to the set of defender parameters chosen to be maximal ($N = N_{\max}, M = M_{\max}$ and $\sigma_{w\rm opt}^2 = D_{E\max}$). This is not surprising and is rather consolidating. One important issue, however, is to compare the robustness of the optimized defense against the mixed attack (shown to be optimal) to that facing the all-noise attack. Fig. 10(b) depicts the detection probability (35) for different values of the watermark power σ_w^2. We observe the following.

- For the same watermark power σ_w^2, we have $P_d^f(\alpha = \alpha_{\min}) < P_d^f(\alpha = 1)$, (the mixed attack is stronger than the all-noise attack). In other words, to achieve the same detection probability, the embedding distortion of a watermark facing the mixed attack must be larger than that of a watermark facing the all-noise attack.
- The slope of the detection probability curve in case of the all-noise attack is larger than that of the mixed attack: A part of the watermark power σ_w^2 enforces the attack impact in the latter case. This fact has already been outlined in Subsection III-A with a nonoptimized defense. Unfortunately, it remains valid with an optimized defense as well.

2) Case of Random Jitter: This attack has been shown to be equivalent to an additive noise $\mathbf{v}_{\rm eq} = \mathbf{n}_r + \mathbf{v}$. Again, suppose $D_v = \sigma_v^2 = \alpha D_a$ and $D_{J_r} = \sigma_{n_r}^2 = (1-\alpha)D_a$. The resulting detection probability is

$$P_d^r = \frac{1}{2}\mathrm{erfc}\left((\nu_r - 1)\sqrt{\frac{N}{2}\frac{\sigma_w^2}{\sigma_x^2 + D_a}}\right) \tag{36}$$

with $\nu_r = 3.3\sqrt{(2/N)(\sigma_x^2 + D_a)/\sigma_w^2}$. The threshold ν_r depends only on the global attack distortion D_a. This would suggest that, from a strict theoretical game-solving point of view, the situation is equivalent to that of the Gaussian watermarking game [20] (under the hypothesis of a Gaussian jitter noise $\mathbf{n}_r$). One can see, however, that from the defender point of view fighting against a random jitter attack is more difficult than that of facing a Gaussian noise. At least, the perceived quality degradation will be greater with the jitter. This means that jittering the watermarked signal $\mathbf{s}$ would remain optimal from the attacker's

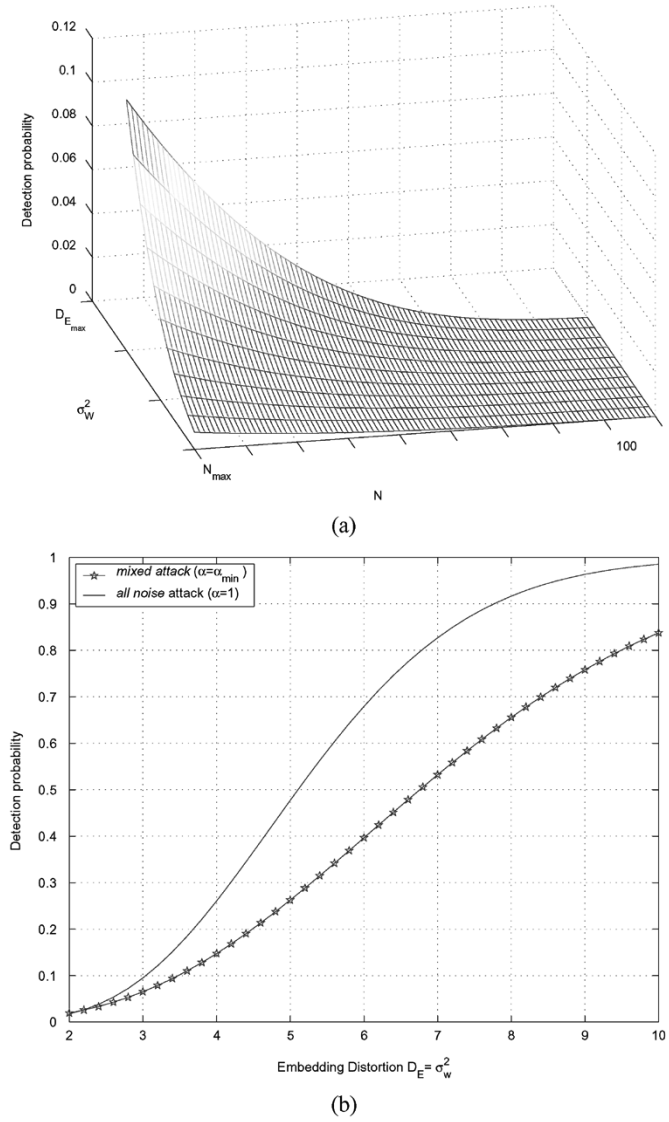


Fig. 10. Optimum defense: Detection probability has to be maximized over the set of defender parameters $\{N, D_E\}$. The 2-D plot (top) shows that reliable detection is obtained with large embedding distortion $D_E = D_{E\max}$ and $N = N_{\max}$. Bottom: The detection probability resulting from solving the game (mixed attack) is compared to that of the all-noise attack. For the same embedding distortion, $Pd(\alpha = \alpha_{\min})$ is smaller than $Pd(\alpha = 1)$.

point of view. This claim is enforced by the fact that, unlike the Gaussian noise $\mathbf{v}$, the jitter noise $\mathbf{n}_r$ depends on the watermarked signal (as suggested by (13)) and is, consequently, significantly increased whenever the defender wants to combat it by increasing the watermark power σ_w^2. In addition, the host signal contributes itself to enforce the jitter effect through σ_x^2 in (13). Then, attributing the hole distortion to the jitter noise $(\sigma_{n_r}^2 = D_a)$ and using (13), the optimal jitter square deviation J must satisfy $J = D_a/(E[((d/dt)s(t))^2])$. The optimum defense, again, corresponds to $\sigma_w^2 = D_{E\max}$.

3) Discussion: Results following from the analysis are summarized below.

1) Facing AWGN attacks, increasing the watermark power is always positive from the embedder's point of view.
2) Under constant scaling attacks, the following two contradicting effects related to deliberately increasing the watermark power appear:
   a) *a positive effect*: increasing the watermark power results in a more reliable detection;
   b) *a negative effect*: increasing the watermark power enforces the desynchronization attack.
3) From the optimized defense analysis, one can see that even in the worst case, that is "the mixed attack," increasing the watermark power remains optimal. Expressed differently, the so-called positive effect always overcomes the negative effect under constant time shift attacks.
4) The multicorrelation test alleviates the impact of the (all-desynchronization) attack (optimum when no countermeasure is taken).
5) Even if the random jitter behavior is noise-like, its dependency on both the host signal and the watermark makes it optimal from an attacker's point of view.

VI. Conclusion

In this paper, we first investigated the general watermarking channel $\mathcal{A}$. Our main motivation was to evaluate the perceived impact an attacker has on a watermarked signal. Our approach consists in removing from the equivalent additive signal $\mathbf{z} = \mathbf{r} - \mathbf{s}$, very often assumed to be uncorrelated with the watermarked signal $\mathbf{s}$, the part that is signal-like. The equivalent attack turns to be a particular case of well-studied channel attack: attacks by filtering and additive noise. This additive noise referred to as the desynchronization noise has been shown to more accurately characterize the attack impact on the original watermarked signal quality loss. Our approach has then been applied to the desynchronization attacks modeled by attacks by jitter plus noise, the AWGN&J channel. Performance loss of the most common watermarking schemes in the presence of such attacks have then been derived. Finally, we investigated optimal attacker and defender strategies in a watermarking game theory context. Results outline a somewhat intuitive result: Desynchronization attacks are much more harmful than additive noise. This was the motivation for providing means to the defender to limit this contribution. Finally, the best strategies for the defender and attacker were described.

References

[1] I. Cox, M. Miller, and A. McKellips, "Electronic watermarking: The first 50 years," in *Proc. Int. Workshop Multimedia Signal Processing*, 2001, pp. 225–230.
[2] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," in *Proc. IEEE Int. Conf. Communications*, vol. 2, BC, Canada, Jun. 1998, pp. 823–827.
[3] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE*, vol. 87, pp. 1079–1107, 1999.
[4] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—A survey," *Proc. IEEE*, vol. 87, pp. 1062–1078, Jul. 1999.
[5] W. Szepanski, "A signal theoretic method for creating forgery-proof documents for automatic verification," in *Proc. Carnahan Conf. Crime Counter-Measures*, Lexington, KY, May 1979, pp. 101–109.
[6] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[7] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.
[8] ——, "Preprocessed and postprocessed quantization index modulation methods for digital watermarking," in *Proc. SPIE Security Watermarking of Multimedia Contents II*, vol. 3971, San Jose, CA, Jan. 2000, pp. 48–59.
[9] ——, "Digital watermarking and information embedding using dither modulation," in *Proc. Workshop Multimedia Signal Processing*, Dec. 1998, pp. 273–278.
[10] M. Ramkumar, "Data hiding in multimedia: Theory and applications,", New Jersey Institute of Technology, Kearny, NJ, Nov. 1998.
[11] J. J. Eggers, J. K. Su, and B. Girod, "A blind watermarking scheme based on structured codebooks," in *Proc. Int. IEE Colloquim Secure Image Authentication*, London, U.K., Apr. 2000, pp. 1–6.
[12] J. J. Eggers, R. Bäuml, and B. Girod, "Digital watermarking facing attacks by amplitude scaling and additive white noise," in *Proc. Int. ITG Conf. Source Channel Coding*, Berlin, Germany, Jan. 2002, pp. 28–30.
[13] J. K. Su, F. Hartung, and B. Girod, "A channel model for a watermark attack," in *Proc. SPIE Security Watermarking of Multimedia Contents*, San Jose, CA, Jan. 1999, pp. 159–170.
[14] R. Bäuml, J. J. Eggers, and J. Huber, "A channel model for watermarks subject to desynchronization attacks," in *Proc. Int. ITG Conf. Source Channel Coding*, Berlin, Germany, Jan. 2002, pp. 28–30.
[15] P. Moulin and A. Ivanovic, "The zero-rate spread spectrum watermarking game," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1098–1117, Apr. 2003.
[16] J. J. K. ORuanaidh and T. Pun, "Rotation, scale and translation invariant digital image watermarking," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, 1997, pp. 536–539.
[17] M. Kutter, "Watermarking resisting to translation, rotation and scaling," in *Proc. SPIE Multimedia Systems Applications*, vol. 3528, Nov. 1998, pp. 423–431.
[18] D. Kirovski and H. S. Malvar, "Spread spectrum watermarking of audio signals," *IEEE Trans. Signal Process. (*Special Issue on Data Hiding*)*, vol. 51, no. 4, pp. 1020–1033, Apr. 2003.
[19] C. M. M. J. Baggen, "An information theoretic approach to timing jitter," Ph.D. dissertation, Univ. of California, San Diego, CA, 1993.
[20] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1639–1667, Jun. 2002.
[21] V. Licks, F. Ourique, R. Jordan, and F. Pérez-González, "The effect of the random jitter attack on the bit error rate, performance of spatial domain image watermarking," in *Proc. IEEE Int. Conf. Image Processing (ICIP) 2003*, vol. 2, Barcelona, Spain, 2002, pp. 28–30.
[22] V. Licks, F. Ourique, R. Jordan, and G. Heileman, "Performance of dirty-paper codes for additive white Gaussian noise," presented at the IEEE Workshop of Statistical Signal Processing (WSSP03), MO, 2003.
[23] S. Pateux, G. L. Guelvouit, and J. Delhumeau, "Capacity of data-hiding system subject to desynchronization," presented at the SPIE, San Jose, CA, Jan. 2004.
[24] B. Chen and G. Wornell, "Achievable performance of digital watermarking systems," in *Proc. Int. Conf. Multimedia Computing Systems*, vol. 87, Florence, Italy, Jun. 1999, pp. 13–18.
[25] I. Cox, M. Miller, and A. McKellips, "Watermarking as communication with side information," in *Proc. Int. Conf. Multimedia Computing Systems*, Jul. 1999, pp. 1127–1141.
[26] M. H. M. Costa, "Writing on dirty papers," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.
[27] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563–593, Mar. 2003.
[28] P. Moulin, M. K. Mihcak, and G.-I. A. Lin, "An information-theoretic model for image watermarking and data hiding," presented at the IEEE Int. Conf. Image Processing, Vancouver, BC, Canada, Sep. 2000.
[29] A. S. Cohen and A. Lapidoth, "The capacity of the vector Gaussian watermarking game," in *IEEE Proc. Int. Symp. Information Theory*, Washington, DC, Jun. 2001, p. 5.
[30] K. Shimizu and E. Aiyoshi, "Necessary conditions for min-max problems and algorithms by a relaxation procedure," *IEEE Trans. Autom. Control*, vol. AC-25, no. 1, pp. 62–66, Jan. 1980.
[31] D. Schonberg and D. Kirovski, "Fingerprinting and forensic analysis of multimedia," presented at the Multimedia Security Workshop, Magdeburg, Germany, Sep. 2004.
[32] A. Piva, M. Barni, F. bartolini, and V. Capellini, "Threshold selection for correlation-based watermark deletion," in *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, Island of Rhodes, Greece, Sep. 1998, pp. 337–355.

Abdellatif Zaidi (S'04) was born in Tunisia in 1978. He received the Eng. degree in electrical engineering from the Ecole Nationale Supérieure de Techniques Avancées (ENSTA), Paris, France, in 2002 and the M.Sc. and Ph.D. degrees in digital communications systems with distinction from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2002 and 2005, respectively.

From 2002 to 2005, he was with the Signals and Systems Laboratory (LSS) at the National Center of Scientific Research (CNRS), Gif-sur-Yvette, France, working toward the Ph.D. degree. His research interests cover a broad area of topics in digital communications, signal processing and information theory, including information embedding, source/channel coding, detection and estimation, coding for side-informed channels, and multiuser information theory.

Rémy Boyer (M'04) received the B.Sc. degree (engineer diploma, with distinction) from the Ecole Supérieure d'Informatique, d'Electronique et d'Automatique (ESIEA), Paris, France, Paris, France, in 1999 and the M.Sc. and Ph.D. degrees in signal processing from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1999 and 2002, respectively.

From 2002 to 2003, he was a Visiting Researcher at the University of Sherbrooke, QC, Canada. He is currently an Associate Professor at the University of Paris XI, Orsay, France, and Permanent Researcher at the Signals and System Laboratory (LSS) of the National Center of Scientific Research (CNRS), Gif-sur-Yvette, France. His research activities include sinusoidal audio coding, source separation, watermarking, and decomposition of multiway arrays.

Pierre Duhamel (M'87–SM'87–F'98) was born in France in 1953. He received the Eng. degree in electrical engineering from the National Institute for Applied Sciences (INSA) Rennes, France, in 1975 and the Dr.Eng. degree and the Doctoratés Sciences degree, both from Orsay University, Orsay, France, in 1978 and in 1986, respectively.

From 1975 to 1980, he was with Thomson-CSF, Paris, France, where his research interests were in circuit theory and signal processing, including digital filtering and analog fault diagnosis. In 1980, he joined the National Research Center in Telecommunications (CNET), Issy les Moulineaux, France, where his research activities were first concerned with the design of recursive charge-coupled device filters. Later, he worked on fast algorithms for computing Fourier transforms and convolutions and applied similar techniques to adaptive filtering, spectral analysis, and wavelet transforms. From 1993 to September 2000, he was a Professor at Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, where his research activities focused on signal processing for communications. From 1997 to 2000, he was head of the Signal and Image Processing Department of ENST. He is currently with CNRS/LSS (Laboratoire de Signaux et Systemes/National Center of Scientific Research), Gif-sur-Yvette, France, where he is developing studies in signal processing for communications (including equalization, iterative decoding, and multicarrier systems) and signal/image processing for multimedia applications, including source coding, joint source/channel coding, watermarking, and audio processing.

Dr. Duhamel was Chairman of the IEEE DSP committee from 1996 to 1998 and a member of the SP for IEEE Com committee until 2001. He was Co-General Chair of the 2001 International Workshop on Multimedia Signal Processing, Cannes, France, and he will be Co-technical Chair of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2006, Toulouse, France. He was the IEEE Distiguished Lecturer for 1999. A paper he coauthored on subspace-based methods for blind equalization received the "Best Paper Award" from the IEEE TRANSACTIONS ON SIGNAL PROCESSING in 1998. He was also awarded the "Grand Prix France Telecom" by the French Science Academy in 2000. He was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1989 to 1991, an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, and a Guest Editor for the Special Issue on Wavelets of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.