submitted to
Electronic Imaging 2002
San Jose, California, USA
Jan. 19-25, 2002

A Channel Model for Watermarks Subject to Desynchronization Attacks

Robert Bäuml, Joachim J. Eggers, Roman Tzschoppe and Johannes Huber
Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstr. 7/NT, 91058 Erlangen, Germany
{baeuml,eggers,roman,huber}@LNT.de

ABSTRACT

One of the most important practical problems of blind Digital Watermarking is the resistance against desynchronization attacks, one of which is the Stirmark random bending attack in the case of image watermarking.

Recently, new blind digital watermarking schemes have been proposed which do not suffer from host-signal interference. One of these quantization based watermarking scheme is the Scalar Costa Scheme (SCS). We present an attack channel for SCS which tries to model typical artefacts of local desynchronization. Within the given channel model, the maximum achievable watermark rate for imperfectly synchronized watermark detection is computed. We show that imperfect synchronization leads to inter-sample-interference by other signal samples, independent from the considered watermark technology. We observe that the characteristics of the host signal play a major role in the performance of imperfectly synchronized watermark detection.

Applying these results, we propose a resynchronization method based on a securely embedded pilot signal. The watermark receiver exploits the embedded pilot watermark signal to estimate the transformation of the sampling grid. This estimate is used to invert the desynchronization attack before applying standard SCS watermark detection. Experimental results for the achieved bit error rate of SCS watermark detection confirm the usefullness of the proposed resynchronization algorithm.

1. INTRODUCTION

Digital watermarking is the art of communicating information, "the watermark message", by embedding it into digital multimedia documents, called "host documents" or "host signals", to produce "marked signals". The embedded *watermark* should be reliably decodable even after further processing of the marked data, which is also denoted as *attack* against the embedded watermark. Such processing can be simple D/A-A/D conversion of the document, but also a malicious attempt to impair watermark reception. Digital watermarking has gained a lot of attention in the recent years for its potential in several areas like proof of ownership and copyright enforcement. For instance, the embedded watermark can provide information about the copyright holder of a document or indicate the copy-state of the digital content.

The research community has come up with a vast variety of watermarking algorithms for different types of multimedia data, e.g., natural images, audio, video. Depending on the data type, watermark embedding is implemented in the spatial or time domain, or in transform domains like the DFT/DCT-spectrum or a wavelet domain. The constraints of a certain application for digital watermarking must be taken into account during the design of a watermarking scheme. One important aspect is the availability of the original document at the watermark receiver. In many applications, the original document cannot be used during watermark reception, which is denoted as *blind watermark reception* or more generally *blind watermarking*. Blind watermarking is considered throughout this paper.

We consider digital watermarking as a communication problem, where the watermark communication channel is characterized by possible attacks against the embedded watermark. A complete characterization of the watermark channel is currently not available, though theoretical analyses of specific attack scenarios have been published within the last two years.[6, 10, 11, 14] One specifically interesting attack is the addition of white Gaussian noise (AWGN), since the analysis of extended attack scenarios can often be based on the analysis of the AWGN attack. Further, the AWGN attack can be applied easily so that each watermarking scheme should show good robustness at least against this type of attack. The design of watermarking schemes facing AWGN attacks and the resulting watermark capacity is reviewed in Section 2. In particular, the *Scalar Costa Scheme*

(SCS) watermarking[7,8] is described, which is currently the most powerful practical blind watermarking scheme in terms of watermark capacity in the case of AWGN attacks.

The goal of this paper is to extend the characterization of the watermark attack channel with respect to *desynchronization attacks*. During the analysis of AWGN attacks, it is assumed that the watermark receiver can look for the watermark information exactly at the same position where it has been embedded, which is denoted as *perfectly synchronized reception*. However, in real-world scenarios, this assumption does not hold necessarily. It is even possible that an attacker intentionally modifies the watermarked document in order to desynchronize the watermark receiver. Note that, for simplicity, the term "synchronization" is used here in a quite general way, although, in a strict sense, synchronization is only relevant for time depending data. A more detailed description of possible desynchronization attacks and the state-of-the-art in solving the synchronization problem for watermark receivers is given in Section 3. Next, we present a model for imperfectly synchronized watermark reception in Section 4. Based on this model, we analyze in Section 5 the watermark capacity of SCS watermarking depending on the synchronization accuracy. In Section 6 a practical approach to the resynchronization problem is presented based on a penalized MLSE. Simulation results for the synchronization accuracy and the BER in the uncoded case are included. Section 7 concludes the most important results and gives an outlook on future research on desynchronization attacks.

2. BLIND WATERMARKING FACING AWGN ATTACKS

We consider digital watermarking a communications problem which can be described as follows: The encoder derives from the *watermark message* m and the host signal $\mathbf{x}$ an appropriate watermark signal $\mathbf{w}$ which is added to the host signal to produce the watermarked signal $\mathbf{s}$. $\mathbf{w}$ must be chosen such that the distortion between $\mathbf{x}$ and $\mathbf{s}$ is negligible. Next, the watermarked signal $\mathbf{s}$ might be processed, which gives a signal $\mathbf{r}$. Such processing potentially impairs watermark communication and thus is denoted as an *attack* against the embedded digital watermark. In general, attacks against digital watermarks are only constrained with respect to the distortion between $\mathbf{x}$ and $\mathbf{r}$. Finally, the receiver must be able to decode the watermark message from the received (attacked) signal $\mathbf{r}$. Both, encoding and decoding, depend on a key sequence $\mathbf{k}$, which ensures that only authorized parties can embed, decode, and modify the embedded watermark message m. Fig. 1 depicts the described blind watermark communication scenario, where an attack by an additive white Gaussian noise (AWGN) signal $v \sim \mathcal{N}(0, \sigma_v^2,)$ is assumed. Further, the analysis is constrained to independent identically distributed (IID) Gaussian original signals $x \sim \mathcal{N}(0, \sigma_x^2)$. In this paper, $\mathbf{x}$,$\mathbf{w}$,$\mathbf{s}$,$\mathbf{r}$ $\mathbf{v}$ and $\mathbf{k}$ are vectors, and $x[n]$,$w[n]$,$s[n]$,$r[n]$,$v[n]$ and $k[n]$ refer to their respective nth elements.

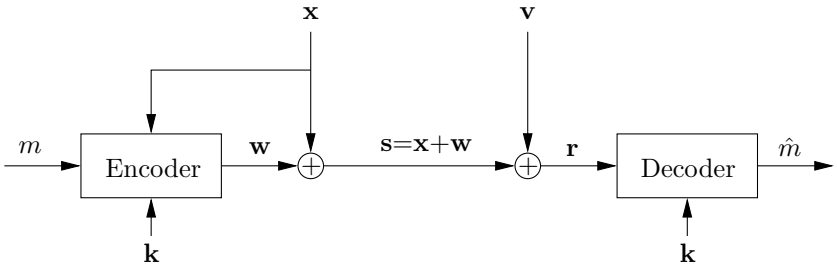

Figure 1: Blind watermarking facing an AWGN attack.

Up to now, the most popular watermark embedding technique is based on the addition of a watermark signal $\mathbf{w}$ which is chosen independently from the host signal $\mathbf{x}$. Here, we assume a Gaussian watermark signal with $w \sim \mathcal{N}(0, \sigma_w^2,)$. This watermarking technique is also denoted as *spread-spectrum* (SS) watermarking, a term derived from spread-spectrum communication, although used in a slightly different way. For blind SS watermark reception, the unknown host signal $\mathbf{x}$ is considered as unavoidable interference. The watermark capacity of SS watermarking for Gaussian host signals and AWGN attacks is $C = 0.5 \log_2(1 + \sigma_w^2/(\sigma_x^2 + \sigma_v^2))$ bit/sample, which can be easily derived from the capacity of an AWGN channel.[4] Unfortunately, in realistic watermarking scenarios we have $\sigma_x^2 \gg \sigma_w^2$ to ensure imperceptibility of the watermark signal. Thus, the capacity of SS watermarking is limited by huge host-signal interference.

In 1999, it has been realized that blind watermarking can be considered communication with side information at the encoder,[2,5] which is obvious from the block diagram in Fig. 1. Costa[3] showed theoretically that for a Gaussian host signal of power σ_x^2, a watermark signal of power σ_w^2, and AWGN of power σ_v^2 the maximum rate of reliable communication (capacity) is $C = 0.5 \log_2(1 + \sigma_w^2/\sigma_v^2)$ bit/sample, independent of σ_x^2. The result is surprising since it shows that the host signal $\mathbf{x}$ need not be considered as interference at the decoder although the decoder does not know $\mathbf{x}$.

Costa's ideal scheme involves a *random* codebook which must be available at the encoder and the decoder. Unfortunately, for good performance the codebook must be so large that neither storing it nor searching it is practical. Thus, for practical application, the random codebook is replaced by a structured codebook, in particular a product codebook of dithered uniform scalar quantizers. The such simplified scheme is denoted as *Scalar Costa Scheme* (SCS).[7,8] This paper focusses on SCS watermarking, so that a brief review of the basic principle is given in the following.

For SCS watermarking, the watermark message m is encoded into a sequence of watermark letters $\mathbf{d}$, where $d[n] \in \mathcal{D} = \{0, 1\}$ in case of binary SCS. Each of the watermark letters is embedded into the corresponding host elements $x[n]$. The embedding rule for the nth element is given by

$$\begin{aligned} a[n] &= \Delta\left(\frac{d[n]}{2} + k[n]\right) \\ s[n] &= x[n] + \alpha\left(\mathcal{Q}_\Delta\left\{x[n] - a[n]\right\}\right) - \alpha\left(x[n] - a[n]\right), \end{aligned} \tag{1}$$

where $\mathcal{Q}_\Delta\{\cdot\}$ denotes scalar uniform quantization with step size Δ. The key $\mathbf{k}$ is a pseudo-random sequence with $k[n] \in (0, 1]$. This embedding scheme depends on two parameters: the quantizer step size Δ and the scale factor α. Both parameters can be jointly optimized to achieve a good trade-off between embedding distortion and detection reliability for a given noise variance of an AWGN attack.[7]

Watermark decoding from the received signal $\mathbf{r}$ is based on the pre-processed received signal $\mathbf{y}$. The extraction rule for the nth element is

$$y[n] = \mathcal{Q}_\Delta\left\{r[n] - k[n]\Delta\right\} - (r[n] - k[n]\Delta), \tag{2}$$

where $|y[n]| \leq \Delta/2$. $y[n]$ should be close to zero if $d[n] = 0$ was sent, and close to $\pm\Delta/2$ for $d[n] = 1$.

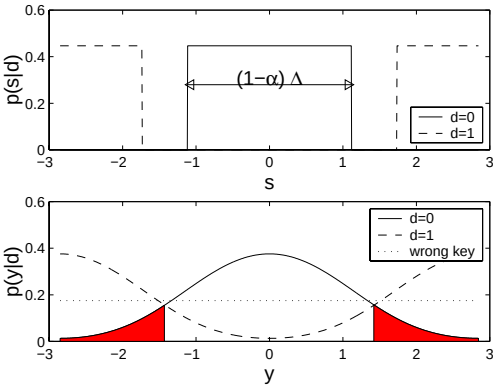

Figure 2. One period of the PDFs of the transmitted and the received signal for binary SCS (σ_w^2=1, WNR = 2 dB, Δ = 5.7, α = 0.61). The filled areas represent the probability of detection errors assuming $d = 0$ was sent. The dotted line in the lower plot depicts the PDF when detecting with a wrong key $\mathbf{k}$.

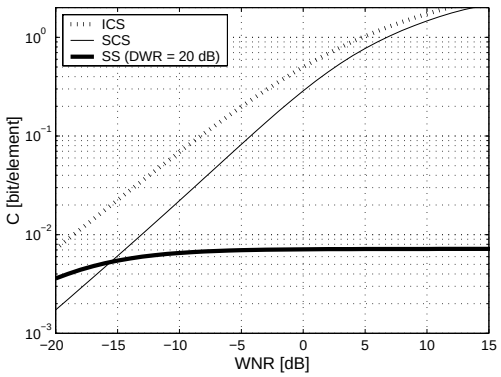

Figure 3. Watermark capacities in case of AWGN attacks for the Ideal Costa Scheme (ICS), the Scalar Costa Scheme (SCS), and spread-spectrum (SS) watermarking.

The basic properties of binary SCS watermarking can be demonstrated by the probability density function (PDF) of the transmitted signal $\mathbf{s}$ and the PDF of the extracted signal $\mathbf{y}$. Note that conditioning on the key

sequence $\mathbf{k}$ is assumed in the following. The upper plot of Fig. 2 depicts one period of the PDF of the transmitted samples s conditioned on the sent watermark letter d. For $d = 0$, the transmitted value s is concentrated around integer multiples of Δ. Contrary, for $d = 1$, s is concentrated around $\Delta/2$ plus integer multiples of Δ. The lower plot shows the PDF of the extracted samples y after AWGN attack conditioned on the sent watermark letter d. $p_y\left(y[n]|d[n]\right)$ is computed numerically.[7] We observe that the PDFs of y in case of $d = 0$ and in case of $d = 1$ can be still distinguished. Note that the distribution of $p_y\left(y[n]|d[n]\right)$ will be uniform for any possible $\mathbf{r}$ when the PDF is determined over all possible keys. This is indicated by the dotted line in the lower plot of Fig. 2.

Fig. 3 shows the watermark capacities for the mentioned watermarking schemes, namely the ideal Costa scheme (ICS), SCS, and SS and for AWGN attacks with varying *Watermark-to-Noise power Ratio* (WNR) of $\mathrm{WNR} = 10\log_{10}\sigma_w^2/\sigma_v^2$ [dB]. Only SS watermarking suffers from host-signal interference, which limits the achievable capacity. The shown capacity of SS watermarking is for the realistic *Document-to-Watermark power Ratio* (DWR) of $\mathrm{DWR} = 10\log_{10}\sigma_x^2/\sigma_w^2 = 20$ dB. SCS watermarking does not achieve the capacity of an ideal Costa scheme, but comes close to that for a large range of practically relevant WNRs. In particular, SCS watermarking achieves significantly larger watermark capacities than blind SS watermarking for $\mathrm{WNR} > -15$ dB.

3. DESYNCHRONIZATION ATTACKS

The analysis of desynchronization attacks against digital watermarks and the development of efficient counter-attacks is still one of the most demanding problems in the field of digital watermarking. In this section, we illustrate the problem of desynchronization attacks in the case of watermarked image data, and give a brief overview of the state-of-the-art in this research area.

Desynchronization attacks have been a problem for a considerable time, especially in the field of image watermarking. Early desynchronization attacks consisted of rather simple global affine transformations. Robustness against such global desynchronization attacks can be achieved by watermark embedding into transform invariant domains. For instance, watermark embedding in the log-polar domain enables robustness against rotation, translation and scaling of the watermarked image.[9] Further, global affine transformations can be estimated relative easily due to the small number of free attack parameters. The estimation of these parameters is usually based on a known embedded synchronization pattern, where the estimation accuracy increases with the image size.

One of the most popular software tools for attacks on image watermarks is the StirMark package,[13] which offers a wide range of different attacks to render watermark extraction hard to impossible. One of the most effective attacks within StirMark is the random bending attack which exloits the inability of the human eye to detect small local geometric distortions. For this attack, a smooth transformation of the sampling grid is applied which desynchronizes a simple watermark detector. Thus, pre-processing prior to standard watermark detection is required to enable watermark detection.

Counter-attacks against the StirMark random bending attack have been investigated mainly for non-blind watermarking,[12] where the knowledge of the original image can be exploited to achieve synchronization of the watermark detector. A promising approach for blind watermarking is based on a model for the local transformations, e.g., local affine transformations, where a synchronization pattern is used to estimate the model parameter. As for global transform models, the synchronization accuracy increases with the number of pixels available for the parameter estimation. In practice, it is highly unlikely that the original sampling grid can be reconstructed perfectly. Therefore, we investigate in this paper the influence of inaccurately synchronized watermark detection. We assume that resynchronization has been performed on the received data so that only a jitter in the sampling grid remains as effective distortion. All effects are viewed in the coordinate domain, where warping effects can be handeled easiest.

It has to be noted that desynchronization attacks are also applicable to other media, e.g. audio data, though the specific attack model may need to be adapted to the given media type. For instance, the amount of subjectively acceptable local modifications of the sampling grid may differ significantly between image data and audio data.

Synchronization is also a major issue in communications, especially in wireless communication, and has been solved satisfactorily for current applications. Unfortunately, the methods developed in these fields cannot be

easily transferred to the synchronization problem in digital watermarking. Typical synchronization problems have been solved for proper models of specific transmission channels. Such models are still missing for desynchronization attacks against digital watermarks. One major problem is that the attacker has many degrees of freedom to implement desynchronization attacks and at the same time has malicious intent.

4. A CHANNEL MODEL FOR DESYNCHRONIZATION ATTACKS

In this section, a channel model for imperfectly synchronized watermark detection is developed. We assume that coarse resynchronization as been applied, e.g., based on the estimation of parameters of local transforms using an embedded synchronization pattern. The artefacts of imperfect resynchronization are similar across all resynchronization methods in the sense that the estimated sampling grid generally contains a certain deviation from the original sampling grid. For simplicity, one-dimensional signals are considered subsequently. The extension to multi-dimensional signals, e.g., image or video data, is straight forward. The developed model gives insights into the principle limits of watermark detection after desynchronization attacks.

Let $s[n] = x[n] + w[n]$ denote the discrete watermarked signal. This signal corresponds to the critically sampled continuous signal $s(t)$, which is bandlimited to $f_G = 2/T$, where T denotes the width of one sampling interval. Then, $\hat{s}[n] = s(nT + T_\Delta)$ denotes the resampled signal, where an offset of T_Δ in the sampling grid has been introduced. Assuming ideal interpolation, $\hat{s}[n]$ can be computed from $s[n]$ with

$$\hat{s}[n] = \sum_{\nu=-\infty}^{\infty} s((n+\nu)T) \cdot \mathrm{sinc}\,(\nu T + T_\Delta)\,, \tag{3}$$

where $\mathrm{sinc}\,(x) = sin(\pi x)/\pi x$. Further signal distortions due to attack operations are described by an additive noise source $v[n]$ with power σ_v^2, so that the received attacked signal is given by

$$r[n] = \hat{s}[n] + v[n]. \tag{4}$$

The described channel model is depicted in Fig. 4.

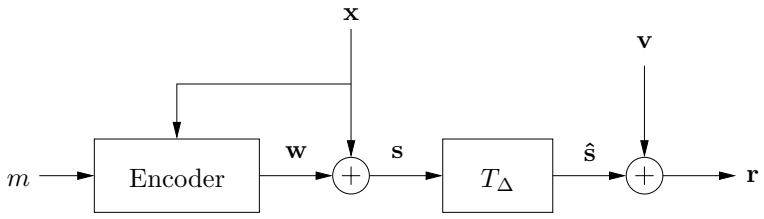

Figure 4: Channel model for a desynchronization attack

Note that considering only a constant sampling offset by T_Δ is not very restrictive. The model can be extended without difficulties to a sampling offset $T_\Delta[n]$ so that $\hat{s}[n] = s(nT + T_\Delta[n])$. However, in this paper we focus on a constant offset T_Δ which gives already important insights concerning the required resynchronization accuracy for watermark detection.

Next, the nth received signal sample $r[n]$ is decomposed into a component derived from the nth watermarked sample $s[n]$ and additional contributions from samples $s[n+\nu]$, with $\nu \neq 0$, which gives

$$\begin{aligned} r[n] &= \sum_{\nu=-\infty}^{+\infty} s((n+\nu)T) \cdot \mathrm{sinc}\,(\nu T + T_\Delta) + v[n] \\ &= \underbrace{(w[n] + x[n]) \cdot \mathrm{sinc}\,(T_\Delta)}_{\hat{s}_e[n]} + \end{aligned}$$

$$+ \quad \underbrace{\sum_{\substack{\nu=-\infty \\ \nu\neq 0}}^{+\infty} s((n+\nu)T)\cdot \mathrm{sinc}\,(\nu T + T_\Delta)}_{z[n]}$$

$$+ \quad v[n]. \tag{5}$$

$\hat{s}_e[n]$ corresponds to the original watermarked signal sample, and $z[n]$ describes *Inter-Symbol-Interference* (ISI). In common communication scenarios without side-information at the encoder, ISI by $s[n+\nu]$, for $\nu \neq 0$, can in principle be inverted to avoid degradation of detection performance. However, when exploiting side-information at the encoder, as in Costa's scheme or its practical version SCS, little is known about possible exploitations of ISI. Thus, we assume that ISI is unavoidable interference for SCS watermark detection. Further, we assume in the following that the watermarked signal is white and Gaussian distributed with a power of $\sigma_s^2 = \sigma_x^2 + \sigma_w^2$.[7]

As we can derive from $\hat{s}_e[n]$, the signal bearing component in our model, containing $w[n]$, is attenuated by $\mathrm{sinc}\,(T_\Delta)$. Thus, we can determine the power $\sigma_{\hat{w}}^2$ of the attenuated watermark $\hat{w}[n]$ after the warping operation by

$$W(T_\Delta) = \sigma_{\hat{w}}^2 = \sigma_w^2 \cdot \mathrm{sinc}\,(T_\Delta)^2 . \tag{6}$$

In turn, the resulting noise power $N(T_\Delta)$ contains now the ISI term from $z[n]$ and the AWGN $v[n]$:

$$N(T_\Delta) = \sigma_s^2 \cdot (1 - \mathrm{sinc}\,(T_\Delta)^2) + \sigma_v^2 = \sigma_z^2 + \sigma_v^2. \tag{7}$$

In the following, we investigate our channel model for $\sigma_v^2 = 0$ and $\sigma_v^2 = \sigma_w^2$. Fig. 5 shows the resulting effective watermark-to-noise power ratio $10\log_{10}(W/N)$ for DWR = 15 dB, 20 dB, and 25 dB. We observe that the power of the AWGN $v[n]$ does not play a dominant role if the relative sampling offset T_Δ/T is larger than about 0.1 to 0.3, depending on the DWR. Further, a significant influence of the DWR appears. This result is a consequence of the assumption that the entire signal $z[n]$ is unavoidable noise.

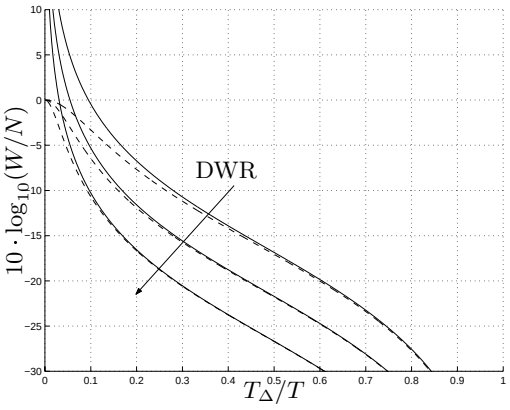

Figure 5. W/N of $\hat{s}[n]$ depending on the sampling deviation T_Δ. The depicted results are for $\sigma_v^2 = 0$ (–) and $\sigma_v^2 = \sigma_w^2$ (- -) and for DWR = 15 dB, 20 dB, 25 dB.

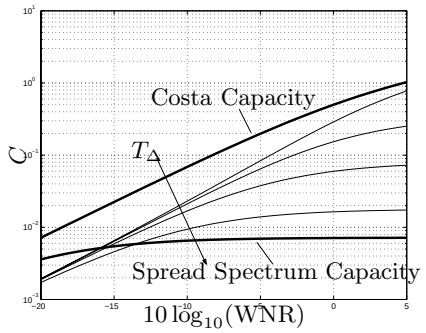

Figure 6. Capacity of SCS watermarking under the influence of a sampling grid deviation T_Δ and AWGN with different WNR. The depicted results are for DWR = 20 dB and T_Δ/T=0, 0.05, 0.1, 0.2, as indicated by the arrow. The upper broad curve represents the channel capacity for an ideal Costa watermarking algorithm with perfect synchronization; the lower broad curve is the capacity for SS watermarking with perfect synchronization.

5. WATERMARK CAPACITY FOR IMPERFECTLY SYNCHRONIZED RECEPTION

In Section 2, SCS watermarking has been introduced as a powerful blind watermarking technology. Significant gains over state-of-the-art SS watermarking are predicted due to the host-signal independence of blind SCS

watermarking. However, the described channel model for imperfectly synchronized watermark detection shows that the strength of ISI interference is strongly dependent on the host signal, in particular on the DWR. In SCS, the side-information about the host signal $\mathbf{x}$ at the encoder is exploited in a quite simple way. That is, the watermark sample $w[n]$ is chosen such that interference from $x[n]$ during blind watermark detection vanishes or is negligible at least. The influence of samples $x[n+\nu]$, for $\nu \neq 0$, which contribute strongest to the total ISI, is not considered during SCS watermark embedding. Thus, the performance of SCS in case of desynchronization attacks is no longer host signal independent. As soon as there is a desynchronization attack and this attack cannot be reversed perfectly, SCS suffers from host signal interference similar to SS watermarking. Here, the capacity of SCS watermarking after AWGN and desynchronization attacks is derived using the model described in Section 4. This allows us to investigate the remaining advantage of SCS over SS watermarking.

We assume that the ISI $z[n]$ has a Gaussian distribution, which is reasonable for a white and Gaussian host signal $\mathbf{x}$. Then, the capacity of SCS watermarking after AWGN and desynchronization attacks can be obtained from the the capacity of SCS watermarking facing a simple AWGN attacks using the effective watermark-to-noise ratio $10\log_{10}(W/N)$ as derived in Section 4.

Fig. 6 shows the capacity of SCS watermarking after AWGN attacks with WNR=$-20,\dots 5$ dB and grid shifts $T_\Delta/T = 0, 0.05, 0.1$, and 0.2 at DWR = 20 dB. We observe that ISI is less important for strong AWGN attacks since in these cases the AWGN dominates ISI for larger grid shifts. The SCS watermark capacity is still reasonably high, even under moderate synchronization errors up to $T_\Delta = T/10$. The comparison to SS watermarking with perfect synchronization shows that only for very strong AWGN attacks (WNRs below -15 dB) SCS watermarking with imperfect synchronization ($T_\Delta/T > 0.1$) performs worse than SS watermarking.

6. PRACTICAL SYNCHRONIZATION

In order to apply the results from the previous investigations, we need to determine a base algorithm. The scenario in mind is a desynchronisation attack, where a white Gaussian signal is watermarked and attacked by a warping operation applied to the sample grid of the watermarked data. The counterattack tries to estimate the warping operation by the use of a well-known pilot sequence embedded in the original domain. A Viterbi algorithm working on a tree-structure is applied to estimate the warping operation in the *Maximum Likelihood Sequence Estimation* (MLSE) sense. In a further effort, a penalized MLSE estimation, exploiting a-priori knowledge about the warping operation, is adopted.

6.1. Desynchronization by grid-warping

One kind of desynchronization attacks can be modelled by a smooth warping function $w(t)$, which maps the argument t of the continuous signal $s(t)$ to t' producing $s(w(t)) = s(t')$. Interpreting $s(w(t))$ as the attacked signal, we can set

$$r(t) = s(w(t)). \tag{8}$$

For illustration, one period of $s(t) = sin(2\pi t)$, warped by w(t) $= (1-\xi)t^2 + \xi t$, $t = 0\dots 1, \xi = \sqrt{2}$, is depicted

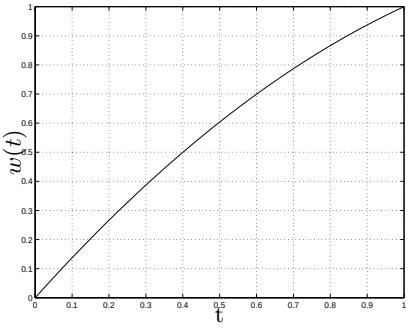

Figure 7. An exemplary warping function $w(t)$.

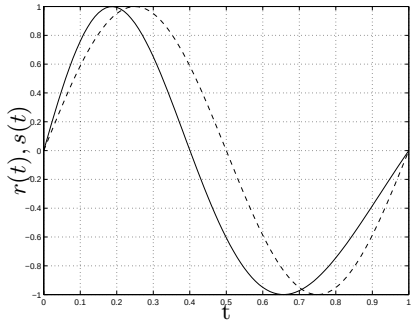

Figure 8. Illustrative warping of one period of a sine function with the original function s(t) (- -) and the warped function r(t) (–).

in Fig. 8. The respective warping function deployed is depicted in Fig. 7.

With the assumption of $s[n]$ being the sampled version of the bandlimited signal $s(t)$, as presented in Section 4, we can generate $r[n]$ as the sampled version of $r(t)$. Due to the nonlinear warping operation, which yields $r(t)$ from $s(t)$, a bandwith expansion can not be avoided in general. Thus, the sampling process from $r(t)$ to $r[n]$ can introduce irreversible distortion due to aliasing, and even with perfect knowledge of $w(t)$, $s[n]$ cannot be reconstructed perfectly from $r[n]$ in the general case. Especially our assumption about a white host signal $x[n]$ worsens the situation in this model compared to a realistic scenario, as natural host signals usually concentrate their energy in the low frequencies leading to less distortion due to lowpass filtering prior to sampling $r(t)$.

6.2. Pilot sequence embedding

The pilot sequence $\mathbf{p}$ contains well known symbols embedded via SCS watermarking as presented in Section 2. For convenience, we choose $\mathbf{p} = \mathbf{0}$. Δ is to be determined for embedding with respect to the noise power σ_z^2 of the interference introduced by ISI. From the previous analysis we can derive that the noise introduced by ISI is Gaussian at a constant T_Δ . Now consider T_Δ a random variable with equal probability in the interval $[-T_{\Delta,max}; T_{\Delta,max})$ under the assumption that the correct position of the sample to be synchronized on is located at $T_\Delta = 0$

$$p_{T_\Delta}(\tau) = \begin{cases} \frac{1}{2T_{\Delta,\max}} & , \quad -T_{\Delta,\max} \leq \tau \leq T_{\Delta,\max} \\ 0 & , \quad |\tau| > T_{\Delta,\max} \end{cases} \tag{9}$$

$T_{\Delta,\max}$ is here the maximum distance form the correct sampling position, any $T_\Delta > T_{\Delta,\max}$ belongs to a different (incorrect) sampling position. (9) is in general valid as long as the mean warping distance $w(t)-t$ of the samples is significantly larger than the distance between the interpolated samples and the warping function $w(t)$ itself is reasonably smooth. With the noise variance σ_z^2 from (7) the noise PDF for a fix T_Δ can be written as

$$\begin{aligned} p_z(Z|T_\Delta) &= \frac{1}{\sigma_z\sqrt{2\pi}} \mathrm{e}^{-\frac{Z^2}{2\sigma_z^2}} = \\ &= \frac{\mathrm{e}^{-\frac{Z^2}{2\sigma_s^2 \cdot \sqrt{1-\mathrm{sinc}(T_\Delta)^2}}}}{\sigma_s\sqrt{2\pi(1-\mathrm{sinc}\,(T_\Delta)^2)}} \end{aligned} \tag{10}$$

Interpreting T_Δ as a random variable, about which the decoder has no knowledge, with realization τ, the resulting PDF $p_z(Z)$ seen by the application is

$$\begin{aligned} p_z(Z) &= \int_{-\infty}^{+\infty} p_{T_\Delta}(\tau) \cdot p_z(Z|\tau) d\tau \\ &= \frac{1}{2T_{\Delta,\max}} \int_{-T_{\Delta,\max}}^{T_{\Delta,\max}} \frac{\mathrm{e}^{-\frac{Z^2}{2\sigma_s^2 \cdot \sqrt{1-\mathrm{sinc}(\tau)^2}}}}{\sigma_s\sqrt{2\pi(1-\mathrm{sinc}\,(\tau)^2)}} d\tau. \end{aligned} \tag{11}$$

For a decent implementation, $T_{\Delta,\max}$ has to be chosen such, that the ISI does not significantly disturb the pilot sequence estimation process. From Fig. 5 we can roughly estimate the influence of $T_{\Delta,\max}$ on the pilot sequence estimation. It is obvious, that, depending on the DWR used for watermark embedding, a relatively small T_Δ is required to achieve a moderate ISI.

A numerical evaluation of (12) with $T_{\Delta,\max} = T/32$ is depicted in Fig. 9. Obviously, the resulting noise PDF does not resemble a Gaussian distribution any more. In the existing literature[7] Δ has been optimized for maximum transinformation in the case of AWGN attacks. Though this attack case is no more valid, it can be shown that Δ as optimized for AWGN attacks is only very weakly dependent on the underlying attack noise PDF, so in our case Δ is chosen only according to σ_z^2 and σ_w^2 without the presumption of an AWGN attack. The resulting attack noise power $\bar{\sigma}_z^2$ can be calculated with (7)

$$\bar{\sigma}_z^2 = \int_{-\infty}^{+\infty} p_{T_\Delta}(\tau) \cdot \sigma_s^2 \cdot (1-\mathrm{sinc}\,(\tau)^2) d\tau = \frac{\sigma_s^2}{2T_{\Delta,\max}} \int_{-T_{\Delta,\max}}^{T_{\Delta,\max}} (1-\mathrm{sinc}\,(\tau)^2) d\tau. \tag{12}$$

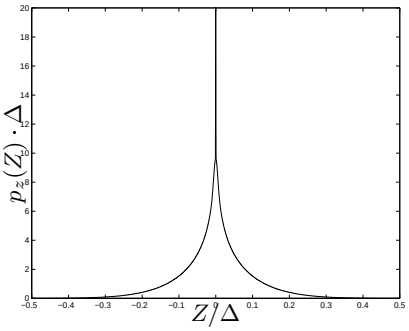

Figure 9. Interference noise PDF $p_z(Z)$ for $T_{\Delta,\max} = T/32$ and DWR = 20 dB. For the case of Dither Modulation, the processed receive signal $y[n]$ yields the same PDF.

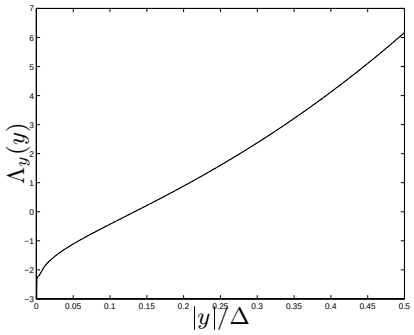

Figure 10. Log-likelihood ratio $\Lambda_y(y)$ for $T_{\Delta,\max} = T/32$

A Taylor approximation for small T_Δ yields

$$\bar{\sigma}_v^2 \approx \sigma_s^2 \frac{2\pi^2}{9} T_{\Delta,\max}^3. \tag{13}$$

For a given DWR and $T_{\Delta,\max}$, α and Δ[7] can be written as

$$\Delta = \sigma_s \sqrt{\frac{12}{\mathrm{DWR}+1} + \frac{2.71 \cdot 8\pi^2}{3} T_{\Delta,\max}^3} \tag{14}$$

$$\alpha = \sqrt{\frac{1}{1 + \frac{2\pi^2}{9}(\mathrm{DWR}+1) T_{\Delta,\max}^3}}. \tag{15}$$

It is obvious that in the case of pure desynchronization attacks, where $T_{\Delta,\max}$ is chosen such that the remaining ISI power $\bar{\sigma}_v^2$ is relatively low compared to the watermark pilot signal, α approaches 1. This means that the pilot signal is embedded by Dither Modulation[1] in practice.

6.3. Synchronization based on an MLSE estimator

In the following investigation, a well known pilot signal $p[n]$ is embedded into the host data $x[n]$. Without loss of generality, the pilot sequence is set $p[n] = 0 \forall n$. A key $k[n]$ as presented in Section 2 is required for embedding and extraction to ensure security and providing means to distinguish subsequent pilot symbols. Binary SCS watermarking is utilized for the embedding process, which reduces to Dither Modulation for small $T_{\Delta,\max}$. The received signal $r[n]$ is upsampled by $\rho = T/2T_{\Delta,\max}$, using a $\mathrm{sinc}\,(\cdot)$-interpolation filter, to produce $r_\rho[n_\rho]$. Following an MLSE approach, a Viterbi algorithm starting from a pair $(r_\rho[0], p[0])$ tries to estimate the next sample that contains the following pilot symbol with maximal probability. In general, we can not assume to have a-priori knowledge about the starting point $r_\rho[0]$ where the first pilot symbol $p[0]$ is embedded. Though, after a certain number of steps, the algorithm will synchronize onto the pilot sequence. Thus, in a practical application, one can assume to know the end point of the pilot sequence und can run the algorithm backwards again. For this reason we assume to know the starting point $(r_\rho[0], p[0])$.

The log-likelihood ratio $\Lambda_y(y)$ is derived from $p(y|0)_k$ and $p(y|0)_{\bar{k}}$, where $p(y|0)_k$ denotes the PDF of $y[n]$ with support in $(-\Delta/2; \Delta/2]$ under the assumption that the correct key k is used. $p(y|0)_{\bar{k}}$ denotes the PDF of $y[n]$ over all incorrect keys $\bar{k}$, which leads to a constant distribution in $(-\Delta/2; \Delta/2]$, as depicted in Fig. 2. In the case, that the sampling value $r_\rho[n_\rho]$ is the nearest possible sampling value to the correct embedding position, the PDF $p(y|0)_k$ is valid. The expected PDF over all other sampling positions is $p(y|0)_{\bar{k}}$. Under these assumptions, the log-likelihood ratio for the metric calculation is

$$\Lambda_y(y) = -\ln \frac{p(y|0)_k}{p(y|0)_{\bar{k}}} \tag{16}$$

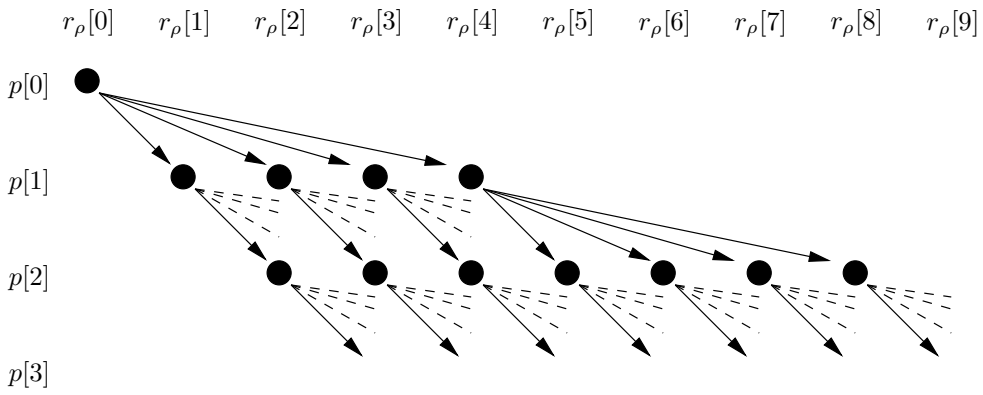

Figure 11: Viterbi tree illustration

The resulting log-likelihood function for $T_{\Delta,\max} = T/32$ and DWR = 20 dB, as used for the metric calculation in the following, is depicted in Fig. 10.

Fig. 11 illustrates the principle of the Viterbi algorithm working on a tree structure. From the starting sample $r_\rho[0]$ with the first pilot symbol $p[0]$ embedded, the next pilot symbol $p[1]$ is considered to be located in one of the samples $r_\rho[\nu], \nu = 1 \ldots N$, with $N > \rho$. N determines the maximum possible step size from one sample to another. It is crucial for the algorithm to have a sufficiently large step size N to be able to track the pilot signal in areas enlarged by the warping. For the case of no desynchronization attack, $p[n]$ is always embedded in $r_\rho[\rho \cdot n] = r[n]$, which is obvious from the fact that the original samples $r[n]$ are located at positions $n_\rho = \rho \cdot n$ in $r_\rho[n_\rho]$ after upsampling by a factor of ρ. In each step of the Viterbi algorithm, one pilot symbol $p[n]$ is considered. Of all transitions to a pair $(r_\rho[n_\rho], p[n])$ from $(r_r[n_\rho - \nu], p[n-1])$, only the transition with the lowest accumulated metric $\lambda(n_\rho, n)$, called the survivor, is considered, other transitions are ignored. The accumulated metric at $(r_\rho[n_{\rho,0}], p[n_0])$ is calculated as

$$\lambda(n_{\rho,0}, n_0) = \sum_{\nu=0}^{n} \Lambda(y_\rho[n_\rho(\nu), \nu]) \tag{17}$$

with $n_\rho(\nu)$ representing the index n_ρ of the pair $(r_\rho[n_\rho], p[n])$ under the prerequisite that the path ending in $(r_\rho[n_{\rho,0}], p[n_0])$ is selected. The received preprocessed signal from (2) is here calculated as

$$y_\rho[n_\rho, n] = \mathcal{Q}_\Delta \left\{ r_\rho[n_\rho] - k[n]\Delta \right\} - \left(r_\rho[n_\rho] - k[n]\Delta \right) \tag{18}$$

Bear in mind that the sum over the log-likelihood ratios requires subsequent values to be independent to achieve optimality. This is not the case with a smooth warping function, where subsequent step values $\Delta_\rho(\nu) = n_\rho(\nu) - n_\rho(\nu - 1)$ are very closely correlated. Without any further assumptions about the warping function, it is not possible to find the correct path through the tree reliably.

Simulations have shown, that due to the high number of possible paths, often an incorrect path is selected when no further restrictions are placed upon the path. But for a successful desynchronization attack, a distortion constraint has to be fulfilled, which usually leads to very smooth warping functions. Utilizing this a-priori knowledge about the warping function, a penalty is introduced depending on the variation of $\Delta_\rho(\nu)$. An optimal penalty factor depends on the characteristics of the warping function, which in general is defined by the attacker. The warping function can have any properties and is only constrained by subjective quality measures, so only very simple heuristics are applied to find a penalty factor. We found, that a penalty factor γ for the metric calculation of $\gamma(\nu) = (\Delta_\rho(\nu) - \Delta_\rho(\nu - 1))^2 + 1/10$ performed very well in our simulations. The corresponding metric $\lambda_p(n_{\rho,0}, n_0)$ is calculated as

$$\lambda_p(n_{\rho,0}, n_0) = \sum_{\nu=0}^{n} \gamma(\nu) \cdot \Lambda(y_\rho[n_\rho(\nu), \nu]) \tag{19}$$

6.4. Simulation results

For the simulation, a DWR of 20 dB is assumed, which presents a realistic value for watermark embedding. A white Gaussian host signal $\mathbf{x}$ of length $n_{\max} = 200$ samples is generated. A pilot signal $\mathbf{p}$ is generated and embedded into $\mathbf{x}$ to produce $\mathbf{s}$. To avoid effects with non-ideal interpolation filters, all further calculations are performed after upsampling of $\mathbf{s}$ by a factor of 2. A warping function $w(t)$ as depicted in Fig. 7 with $\xi = 1.05$ is utilized such that the sample with index n is moved to the new position $\tilde{n} = n_{\max} \cdot w(n/n_{\max})$, $\tilde{n} \in \mathcal{R}$. For the oversampling factor ρ during the resynchronization $\rho = 16$ is assumed, which in turn leads practically to Dither Modulation during the embedding process according to (15).

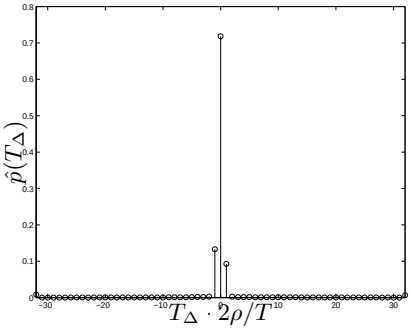

Figure 12. Histogram for the measured sampling deviation from the best match n_ρ.

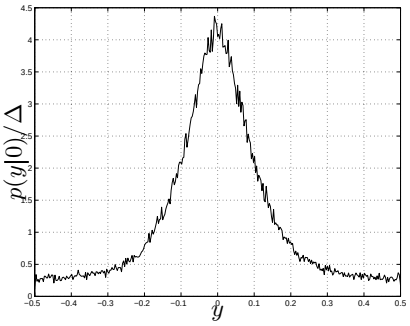

Figure 13. Measured PDF of y after resynchronization of r.

A histogram of the deviation of the estimated $\hat{n}_\rho$ from the best match position n_ρ is depicted in Fig. 12. The histogram was averaged over 1000 independent simulations with different host data $\mathbf{x}$ and keys $\mathbf{k}$. Obviously, the synchronization works well under the given conditions. Deviations from the best match are restricted to one sample distance in most cases, which is equivalent to $|T_\Delta| \leq 3T/64$, taking the overall oversampling rate into account.

For the BER measurement, the pilot $\mathbf{p}$ and a message $\mathbf{m}$ are interleaved embedded such that $\mathbf{p}$ is embedded at odd positions of the sample index of $\mathbf{x}$ and $\mathbf{m}$ is embedded at even positions. Without loss of generality, we set $\mathbf{p} = \mathbf{0}$ and $\mathbf{m} = \mathbf{0}$ again. A lowpass hostsignal with cut-off frequency $\Omega_c = \pi/2$ is utilized to provide better means to interpolate the warping function. This is in general justified for realistic signals which typically have lowpass characteristics.

The estimated warping function is linearly interpolated to estimate the positions of the message symbols. After resynchronization, the PDF of the preprocessed receive signal $\mathbf{y}$ with respect to $\mathbf{m}$ is measured. The resulting PDF is depicted in Fig. 13. Applying hard-decision with a decision threshold of $y_{\text{threshold}} = \Delta/4$ for the estimation of the sent message $\mathbf{m}$ yields $\hat{\mathbf{m}}$. From the PDF $p(y|0)$ we can can estimate BER ≈ 0.15. Applying a decent channel code, e.g. a turbo code rate 1/3, negligible error rates can be expected.

7. CONCLUSIONS AND FUTURE RESEARCH

Robust watermark detection after desynchronization attacks is still an important problem in the field of digital watermarking. In this paper, a channel model for imperfectly synchronized watermark detection has been investigated. The focus of our analysis is on blind scalar Costa scheme (SCS) watermarking, which is for perfectly synchronized detection independent from the host signal statistics and thus outperforms the popular spread-spectrum (SS) watermarking by far. We observed that SCS suffers from inter-symbol-interference (ISI) in the case of imperfectly synchronized watermark detection. We investigated the SCS watermark capacities after AWGN attacks and imperfect resynchronizion. One important result is that, for realistics DWRs, a synchronization error up to 10 % of the sampling interval is acceptable. For such accurate resynchronization, SCS watermarking performs for weak to medium-strong attacks still significantly better than SS watermarking. Nevertheless, our analysis highlights the fact that very exact resynchronization plays a major role for this watermarking method to keep up a reasonable watermark capacity.

A practical resynchronization scheme by embedding a pilot sequence into the data has been presented. Utilizing a penalized MLSE approach, a good estimate of the warping function can be derived. In our case, a deviation of no more than one sample in the oversampled domain was observed in most cases.

In a further effort, the pilot signal is embedded together with an information bearing watermark. After resynchronization, a remaining BER of 15% is observed in the uncoded case. Together with reasonable channel coding, this presents a good base for low error rate watermarks.

Further work has to be carried out on improved robustness of the resynchronization. All practical resynchronization simulations have been performed without additional distortion by quantization or imperfect interpolation filters. Malicious attacks beyond desynchronization have not been considered as well in this scenario. For real applications, these circumstances have to be taken into account, providing a wide range for future research.

REFERENCES

1. B. Chen and G. W. Wornell. Digital watermarking and information embedding using dither modulation. In *Proc. of IEEE Workshop on Multimedia Signal Processing (MMSP-98)*, pages 273–278, Redondo Beach, CA, USA, Dec. 1998.
2. B. Chen and G. W. Wornell. Provably robust digital watermarking. In *Proceedings of SPIE: Multimedia Systems and Applications II (part of Photonics East '99)*, volume 3845, pages 43–54, Boston, MA, USA, September 1999.
3. M. H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
5. I. J. Cox, M. L. Miller, and A. L. McKellips. Watermarking as communications with side information. *Proceedings of the IEEE, Special Issue on Identification and Protection of Multimedia Information*, 87(7):1127–1141, July 1999.
6. J. J. Eggers and B. Girod. Quantization effects on digital watermarks. *Signal Processing*, 81(2):239–263, February 2001.
7. J. J. Eggers, J. K. Su, and B. Girod. A blind watermarking scheme based on structured codebooks. In *Secure Images and Image Authentication, Proc. IEE Colloquium*, pages 4/1–4/6, London, UK, April 2000.
8. J. J. Eggers, J. K. Su, and B. Girod. Performance of a practical blind watermarking scheme. In *Proc. of SPIE Vol. 4314: Security and Watermarking of Multimedia Contents III*, San Jose, Ca, USA, January 2001.
9. M. Kutter. Watermarking resisting to translation, rotation and scaling. In *Proc. of SPIE: Multimedia systems and application*, volume 3528, pages 423–431, Boston, USA, November 1998.
10. P. Moulin and M. K. Mihçak. The data-hiding capacity of image sources. preprint, June 2001.
11. P. Moulin and J. A. O'Sullivan. Information-theoretic analysis of information hiding. Preprint, September 1999.
12. I. B. Ozer, M. Ramkumar, and A. N. Akansu. A new method for detection of watermarks in geometrically distorted images. In *Proceedings of the IEEE Intl. Conference on Speech and Signal Processing 2000 (ICASSP 2000)*, Istanbul, Turkey, June 2000.
13. F. A. P. Petitcolas and M. G. Kuhn. StirMark image watermark benchmark software. Technical report, available at http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/, October 1998.
14. J. K. Su, J. J. Eggers, and B. Girod. Analysis of digital watermarks subjected to optimum linear filtering and additive noise. *Signal Processing, Special Issue on Information-Theoretic Issues in Digital Watermarking*, 81(6), June 2001.