A Channel Model for Watermarks Subject to Desynchronization Attacks

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

Abstract—
Digital watermarking has been proposed for the copyright protection of multimedia documents, e.g., natural images, audio, video. One important unsolved problem in the field of digital watermarking is the synchronization of the watermark receiver after D/A-A/D conversion or local desynchronization attacks, e.g., StirMark’s random bending attack. In this paper, we present a channel model for watermark reception with non-perfect synchronization. Under the given channel model the capacity of SCS watermarking and imperfectly synchronized reception is derived. One important result is that the capacity of SCS watermarking depends on the document power as long perfect synchronization is not achieved. For realistic document-to-watermark power ratios, synchronization errors up to 10% of the sampling interval can be accepted without too significant losses off the SCS watermark capacity.

I. 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 it’s 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, e.g., [5], [10], [13], [9]. 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 II. In particular, the Scalar Costa Scheme (SCS) watermarking [6], [7] 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 III. Next, we present a model for imperfectly synchronized watermark reception in Section IV. Based on this model, we analyze in Section V the watermark capacity of SCS watermarking depending on the synchronization accuracy. Section VI concludes the most important results and
gives an outlook on future research on desynchronization attacks.

II. BLIND WATERMARKING FACING AWGN ATTACKS

We consider digital watermarking a communication problem which can be described as follows: The encoder derives from the watermark message \( m \) and the host signal \( x \) an appropriate watermark signal \( w \) which is added to the host signal to produce the watermarked signal \( s \). \( w \) must be chosen such that the distortion between \( x \) and \( s \) is negligible. Next, the watermarked signal \( s \) might be processed, which gives a signal \( r \). 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 \( x \) and \( r \). Finally, the receiver must be able to decode the watermark message from the received (attacked) signal \( r \). Both, encoding and decoding, depend on a key sequence \( k \), 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. Furthermore, the analysis is constrained to independent identically distributed (IID) Gaussian original signals \( x \sim \mathcal{N}(0, \sigma_x^2) \). In this paper, \( x,w,s,r,v \) and \( k \) are vectors, and \( x[n],w[n],s[n],r[n],v[n] \) and \( k[n] \) refer to their respective \( n \)th elements.

![Fig. 1. Blind watermarking facing an AWGN attack.](image)

Up to now, the most popular watermark embedding technique is based on the addition of a watermark signal \( w \) which is chosen independently from the host signal \( x \). Here, we assume a Gaussian watermark signal with \( v \sim \mathcal{N}(0, \sigma_v^2) \). 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 \( x \) 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 [3]. 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 [4], [1], which is obvious from the block diagram in Fig. 1. Costa [2] showed theoretically that for a Gaussian host signal of power \( \sigma_x^2 \), a watermark signal of power \( \sigma_w^2 \), and AWGN of power \( \sigma_v^2 \), the maximum rate of reliable communication (capacity) is \( C = 0.5 \log_2(1 + \sigma_w^2 / \sigma_v^2) \) bit/sample, independent of \( \sigma_x^2 \). The result is surprising since it shows that the host signal \( x \) need not be considered as interference at the decoder although the decoder does not know \( x \).

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) [6], [7]. This paper focuses 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 \( d[n] \) where \( d[n] \in D = \{0, 1\} \) in case of binary SCS. Each of the watermark letters is embedded into the corresponding host elements \( x[n] \). The embedding rule for the \( n \)th element is given by

\[
\begin{align*}
a[n] &= \Delta \left( \frac{d[n]}{2} + k[n] \right) \\
s[n] &= x[n] + \alpha (Q_\Delta \{ x[n] - a[n] \}) + \alpha (a[n] - x[n])
\end{align*}
\]

(1)

where \( Q_\Delta \{ \cdot \} \) denotes scalar uniform quantization with step size \( \Delta \). The key \( k \) is a pseudo-random sequence with \( k[n] \in \{0, 1\} \). This embedding scheme depends on two parameters: the quantizer step size \( \Delta \) and the scale factor \( \alpha \). 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. Optimal values for \( \Delta \) and \( \alpha \) are given in [6].

Watermark decoding from the received signal \( r \) is based on the pre-processed received signal \( y \). The extraction rule for the \( n \)th element is

\[
y[n] = Q_\Delta \{ r[n] - k[n]\Delta \} + k[n]\Delta - r[n]
\]

(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 \( d[n] = 1 \).

The basic properties of binary SCS watermarking can be demonstrated by the probability density function (PDF) of the transmitted signal \( s \) and the PDF of the extracted signal \( y \). Note that conditioning on the key sequence \( k \) is assumed in the following. The upper plot of Fig. 2 depicts one period of
SS watermarking suffers from host-signal interference, which limits the achievable capacity. The shown capacity of SS watermarking is for the realistic Document-to-Noise power Ratio (DWR) of DWR = 10\log_{10} \sigma_{w}^{2}/\sigma_{n}^{2} = 20 \, \text{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 WNR > -15 \, \text{dB}.

### III. 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 [8]. Further, global affine transformations can be estimated relatively 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 [12], 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 exploits 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. An example transformation is presented in Fig. 4, where the transformation is obvious from the prior knowledge of the viewer about the image. The same transformation imposed on a natural image yields negligible visual artefacts, while synchronization of standard watermark detectors is destroyed. The human visual perception is a complex process which takes a lot of prior knowledge about the image into account. Thus, it is difficult to find an objective quality measure which reflects the subjective quality loss due to the random bending attack. Note that the popular mean-squared error (MSE) distortion measure is not
appropriate within the context of desynchronization attacks.

Counter-attacks against the StirMark random bending attack have been investigated mainly for non-blind watermarking, e.g., [11], 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 handled 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 im-
plement desynchronization attacks and at the same time has malicious intent.

IV. 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 \( T_\Delta \) in the sampling grid has been introduced. Assuming ideal interpolation, \( \hat{s}[n] \) can be computed from \( s[n] \) with

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

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

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

The described channel model is depicted in Fig. 5.

Note that considering only a constant sampling offset by \( T_\Delta \) 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 \( T_\Delta \) which gives already important insights concerning the required resynchronization accuracy for watermark detection.

Next, the \( n \)th received signal sample \( r[n] \) is decomposed into a component derived from the \( n \)th
As we can derive from Fig. 7, very little deviation from the exact sampling grid produces relatively large disturbance. In the best shown case, where the DWR is 15 dB, a deviation of 10% from the exact grid leads to additional noise from ISI which is as strong as the AWGN. In cases of larger DWRs, the influence of ISI increases and thus even more accurate resynchronization is required.

\[
\begin{align*}
\text{W}[T_{\Delta}] &= \sigma^2_w = \sigma^2_w \cdot \text{sinc}^2(T_{\Delta})^2. \\
N(T_{\Delta}) &= \sigma^2_v \cdot (1 - \text{sinc}^2(T_{\Delta}))^2 + \sigma^2_v.
\end{align*}
\]

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

\[
\text{AIR} = \frac{\sigma^2_v}{\sigma^2_w \cdot (1 - \text{sinc}^2(T_{\Delta}))^2}, \quad (6)
\]

and represents the influence of the noise produced by the sampling jitter compared to the AWGN \( v \). As we can derive from Fig. 7, very little deviation from the exact sampling grid produces relatively large disturbance. In the best shown case, where the DWR is 15 dB, a deviation of 10% from the exact grid leads to additional noise from ISI which is as strong as the AWGN. In cases of larger DWRs, the influence of ISI increases and thus even more accurate resynchronization is required.

In Section II, 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. Further, a significant influence of the AWGN appears. This result is a consequence of the assumption that the entire signal \( s_z[n] \) is unavoidable noise.

Another interesting value is the *Attack to Interference Power Ratio* (AIR), which is defined as

\[
\text{AIR} = \frac{\sigma^2_v}{\sigma^2_w \cdot (1 - \text{sinc}^2(T_{\Delta}))^2}, \quad (6)
\]

and represents the influence of the noise produced by the sampling jitter compared to the AWGN \( v \). As we can derive from Fig. 7, very little deviation from the exact sampling grid produces relatively large disturbance. In the best shown case, where the DWR is 15 dB, a deviation of 10% from the exact grid leads to additional noise from ISI which is as strong as the AWGN. In cases of larger DWRs, the influence of ISI increases and thus even more accurate resynchronization is required.
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 IV. This allows us to investigate the remaining advantage of SCS over SS watermarking.

We assume that the ISI $s_z[n]$ has a Gaussian distribution, which is reasonable for a white and Gaussian host signal $x$. 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 $10 \cdot \log_{10}(W/N)$ as derived in Section IV. Fig. 8 shows the resulting capacity curves for $\sigma_w^2 = 0$ and $\sigma_v^2 = \sigma_w^2$, and for three different DWRs. As already suggested by the analysis in Section IV, a relative deviation $T_{\Delta}/T$ of more than 10% results in a significant reduction of the SCS watermark capacity. ISI dominates the influence of AWGN for large $T_{\Delta}/T$.

Next, the influence of variable AWGN power $\sigma_v^2$ for differently fixed grid shifts $T_{\Delta}$ and DWR = 20 dB is investigated. Fig. 9 shows the capacity of SCS watermarking after AWGN attacks with WNR = $-20, \ldots, 5$ dB and grid shifts $T_{\Delta}/T = 0, 0.05, 0.1, \text{ and } 0.2$. 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 ($T_{\Delta}/T \approx 0.1$) performs worse than SS watermarking.

VI. Conclusions and Future Research

Robust watermark detection after desynchronization attacks is still an important open 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 case of imperfectly synchronized watermark detection. This is especially true for large document-to-watermark power ratios (DWRs), where ISI dominates other attack distortions. Thus, the property of SCS being host signal independent is no more true under desynchronization attacks. We investigated the SCS watermark capacities after AWGN attacks and imperfect resynchronization. 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.

The presented model for desynchronization attacks is very simple, but gives already significant insights into the synchronization problem of digital watermarking. In future, the model has to be extended to include random synchronization errors. Further, amplitude synchronization before SCS watermark detection has to be investigated as well.
References


