4th Intl. ITG Conference
on Source and Channel Coding
Berlin, Jan. 28-30, 2002

Digital Watermarking facing Attacks by Amplitude Scaling and Additive White Noise

Joachim J. Eggers, Robert Bäuml

Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstr. 7/NT, 91058 Erlangen, Germany
eggers@LNT.de

Bernd Girod

Information Systems Laboratory
Stanford University
Stanford, CA 94305-9510, USA
girod@ee.stanford.edu

Abstract —

Digital watermarking is a technology which potentially can be used to enforce the copyrights and integrity of digital multimedia data. In this paper, a communications perspective on digital watermarking is used to compute upper performance limits on blind digital watermarking for simple AWGN attacks and attacks by amplitude scaling and additive white noise. We show that the latter case can be translated into effective AWGN attacks, which enables a straight forward capacity analysis based on the previously obtained watermark capacities for AWGN attacks. We analyze the watermark capacity for different theoretical and practical blind watermarking schemes. This analysis shows that the practical ST-SCS watermarking achieves at least 40 % of the capacity of an ideal blind watermarking scheme.

I. Introduction

Digital media has replaced analog media in many applications within the last decade. The success of the digital representation of analog media is mainly due to properties like simple noise-free transmission over general purpose channels, compact storaging, perfect copying, and simple editing. Not only various advantages of the new digital technology have been realized, but also several drawbacks. Most problems with digital media are related to intellectual property rights and trustworthiness of the content. *Digital Watermarking* is one approach to enforce copyrights or to ensure the integrity of digital media. Here, digital watermarking is considered as the *imperceptible, robust, secure communication* of information by embedding it in and retrieving it from other digital data. The basic idea is that the embedded information – the watermark message – travels with the multimedia data wherever the watermarked data goes. This watermark message is then exploited to resolve ownership disputes, to implement playback control, to differentiate between different copies of the same content, or to verify the integrity of the digital data.

Over the last years, many different watermarking schemes for a large variety of data types have been developed. Most of the work considers still image data, but watermarking of audio and video data is popular as well. Theoretical limits of digital watermarking have been investigated since about 1999 [1, 2, 3]. Here, the amount of reliably communicable watermark information dependent on the statistics of the original data and the strength of attacks against the embedded watermarks is analyzed.

In this paper, we present a theoretical analysis of the performance limits of different watermarking schemes against amplitude scaling and additive white noise attacks. A general communication model for digital watermarking is presented in Section II. In Section III, digital watermarking facing an attack by additive white Gaussian noise (AWGN) is reviewed. We consider four different watermarking technologies. The simple AWGN attack is extended in Section IV to attacks by amplitude scaling and additive white (Gaussian) noise (SAW(G)N). We show that the SAW(G)N attack can be translated into an *effective* AW(G)N attack, so that a performance analysis can be based on the results given in Section III. We demonstrate that previous work on this subject is not complete due to an ineff£cient restriction of the watermark embedding. Further, the loss of suboptimal embedding schemes compared with an ideal embedding scheme is investigated. The new results presented in this paper are concluded in Section V.

II. A Communications Approach to Digital Watermarking

We consider digital watermarking a communications problem. Fig. 1 depicts a general perspective on digital watermarking. A *watermark message* m is embedded into the *original data* $\mathbf{x}$ of length L_x to produce the *watermarked data* $\mathbf{s}$. The embedding process is dependent on the key K and must be such, that the quality difference between $\mathbf{x}$ and $\mathbf{s}$ (*embedding distortion* D_{E}) is not too large. For embedding, a key sequence $\mathbf{k}$ of appropriate length is derived from the key K. The difference $\mathbf{w} = \mathbf{s} - \mathbf{x}$ is denoted the *watermark signal*. The watermarked data $\mathbf{s}$ might be further processed or even replaced by some other data. This process, denoted *attack*, produces the *attacked data* $\mathbf{r}$. The attack can be any processing such that the quality difference between $\mathbf{x}$ and $\mathbf{r}$ (*attack distortion* D_{A}) is acceptable. Usually, the goal of the attack is to impair or even remove the embedded watermark information. The attacked data $\mathbf{r}$

is equivalent to the *received data* $\mathbf{r}$, which is input to the watermark reception process. Watermark reception denotes both, *decoding* of a received watermark message $\hat{m}$ using key K and, watermark *detection*, meaning the hypothesis test whether $\mathbf{r}$ is watermarked or not. In this paper, we focus on the reliable decoding of the watermark message. In some applications of digital watermarking, the original data $\mathbf{x}$ might be available to the watermark receiver as indicated with the dotted arrow in Fig. 1, however, in many applications it is not available. We focus on *blind watermarking* which denotes the scenario where the watermark receiver operates without access to the original data $\mathbf{x}$. Here, $\mathbf{x}$,$\mathbf{w}$,$\mathbf{s}$,$\mathbf{r}$, and $\mathbf{k}$ are vectors, and x_n,w_n,s_n,r_n, and k_n refer to their respective nth elements.

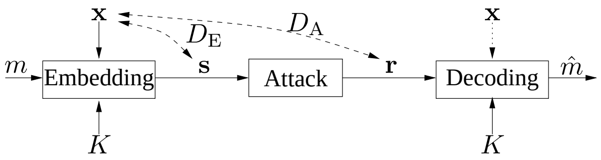

Fig. 1: General model of digital watermarking.

Digital watermarking is inherently related to stochastic description of multimedia data. There is no use for watermarking of data that is perfectly known to attackers. Any modification of the data could be inverted perfectly, leading to trivial watermark removal. Thus, essential requirements on data being robustly watermarkable are that there is enough randomness in the structure of the original data and that quality assessments can be made only in a statistical sense. Therefore, the original data $\mathbf{x}$ is considered a realization of a discrete random process x. Here, random variables are written in Sans Serif font, e.g., x for a scalar random variable and x for a vector random variable. In this paper, independent, identically distributed (IID) data elements are assumed so that it is sufficiently to characterize the element-wise probability density function (PDF) $p_{\mathsf{x}}(x)$. The definition of an appropriate data quality measure depends strongly on the data at hand. However, in many cases a (weighted) mean-squared error distortion measure allows a meaningful quality assessment. Thus, the embedding and attack distortions measures are defined here by $D_{\mathrm{E}} = \mathrm{E}\left\{(\mathsf{s}-\mathsf{x})^2\right\}$ and $D_{\mathrm{A}} = \mathrm{E}\left\{(\mathsf{r}-\mathsf{x})^2\right\}$, respectively. Note that $D_{\mathrm{E}} = \sigma_{\mathsf{w}}^2$ for a mean-free watermark signal $\mathbf{w}$.

In watermarking applications, the embedder tries to communicate as much watermark information as possible while maintaining a sufficient high data quality. Contrary, an attacker tries to impair watermark communication while impairing the data quality as little as possible. Therefore, digital watermarking scenarios can be considered a game between the watermark embedder and the attacker [2]. In [2], the watermark capacity C is defined as the supremum of all achievable watermark rates for a given pair (D_{E},D_{A}). Any processing which achieves (D_{E},D_{A}) has to be considered. A complete solution to this general watermarking game is currently not available. Thus, we consider sub-optimal watermarking schemes, e.g., spread-spectrum (SS) watermarking, and sub-optimal attack channels, e.g. AWGN attacks. In this case, the watermark capacity is the supremum of all achievable rates for the constrained watermarking scheme and/or the constrained attack. The present constraints are indicated by suffixes, e.g., $C_{\mathrm{SS}}^{\mathrm{AWGN}}(D_{\mathrm{E}}, D_{\mathrm{A}}) = C_{\mathrm{SS}}^{\mathrm{AWGN}}(\mathrm{WNR})$ denotes the capacity of spread-spectrum watermarking facing an AWGN channel, which is completely determined by the watermark-to-noise power ratio WNR for a fix document-to-watermark power ratio DWR.

III. Digital Watermarking facing AWGN Attacks

Watermarking of an IID Gaussian original signal $\mathsf{x} \sim \mathcal{N}(0, \sigma_{\mathsf{x}}^2)$ and an attack by AWGN $\mathsf{v} \sim \mathcal{N}(0, \sigma_{\mathsf{v}}^2,)$ is reviewed for four different watermarking schemes. The AWGN attack is of interest since it is so simple that it can be easily applied in any watermarking scenario. Thus, the performance of a watermarking scheme facing an AWGN attack can be considered an upper bound on the general watermark capacity. Further, the extended attack considered in Section IV can be analyzed using the results obtained for the AWGN attack.

A. Spread-Spectrum Watermarking

The term *spread-spectrum (SS) watermarking* has been established in the watermarking community for watermark embedding by the addition of a statistically independent pseudo-noise signal $\mathbf{w}$ with power σ_{w}^2 which is derived from the watermark message m and the key K. Fig. 2 depicts a block diagram for blind and non-blind spread-spectrum watermarking. Spread-spectrum watermarking is one of the first methods used for watermarking (e.g., [4, 5]) and is still the most popular one. For the given assumptions about the original signal, the attack noise, and for a Gaussian watermark signal $\mathsf{w} \sim \mathcal{N}(0, \sigma_{\mathsf{w}}^2)$ the SS watermark capacity is given by the capacity of an AWGN channel, which is [6]

$$C_{\mathrm{non-blind\,SS}}^{\mathrm{AWGN}} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_{\mathsf{w}}^2}{\sigma_{\mathsf{v}}^2}\right) \tag{1}$$

for non-blind SS watermarking, and

$$C_{\mathrm{blind\,SS}}^{\mathrm{AWGN}} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_{\mathsf{w}}^2}{\sigma_{\mathsf{x}}^2 + \sigma_{\mathsf{v}}^2}\right) \tag{2}$$

for blind SS watermarking. The suffix "blind" is suppressed in the remainder since the focus of this paper is on blind watermarking.

Note that $\sigma_{\mathsf{x}}^2 \gg \sigma_{\mathsf{w}}^2$ and $\sigma_{\mathsf{x}}^2 \gg \sigma_{\mathsf{v}}^2$ due to the quality constraints for watermark embedding and attacks on watermarks, respectively. Thus, the performance of blind SS watermarking facing an AWGN attack is mainly determined by the $\mathrm{DWR} = 10\log_{10}\sigma_{\mathsf{x}}^2/\sigma_{\mathsf{w}}^2[\ \mathrm{dB}]$. This shows that blind watermark reception suffers

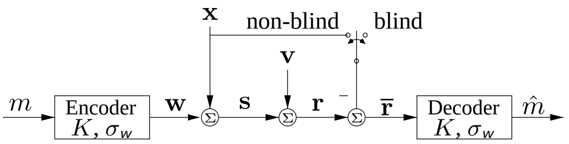

Fig. 2: Blind/non-blind watermark transmission in the presence of an AWGN attack.

significantly from original signal interference. Contrary, the performance of non-blind SS watermarking facing an AWGN attack is completely independent from the characteristics of the original signal $\mathbf{x}$. Here, the performance depends solely on the $\mathrm{WNR} = 10\log_{10}\sigma_{\mathsf{w}}^2/\sigma_{\mathsf{v}}^2[\ \mathrm{dB}]$.

B. Watermarking as Communication with Side-Information at the Encoder

In 1998, it has been realized [7, 8] that considering blind watermarking as *communication with side-information at the encoder* enables the design of improved blind watermarking schemes with reduced interference from the original signal. Fig. 3 depicts a block diagram of blind watermark communication, where the encoder exploits the side-information about the original signal.

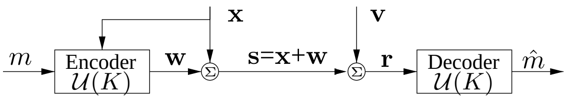

Fig. 3: Watermark communication facing an AWGN attack as communication with side-information.

Chen and Wornell introduced an important but almost forgotten paper by Costa into the watermarking community. Costa[9] showed theoretically that for the communication scenario depicted in Fig. 3 with a Gaussian original signal of power σ_{x}^2, a watermark signal of power σ_{w}^2, and AWGN of power σ_{v}^2 the capacity is

$$C_{\mathrm{ICS}}^{\mathrm{AWGN}} = \frac{1}{2}\log_2\left(1+\frac{\sigma_{\mathsf{w}}^2}{\sigma_{\mathsf{v}}^2}\right), \qquad (3)$$

independent of σ_{x}^2. The suffix “ICS” stands for *ideal Costa scheme*, and is used here to distinguish the theoretical performance limit from that of suboptimal schemes discussed below. The result (3) is surprising since it shows that the original data $\mathbf{x}$ need not be considered as interference at the decoder although the decoder does not know $\mathbf{x}$. Costa presents a theoretic scheme which involves a random codebook $\mathcal{U}^{L_x}$ which is

$$\begin{aligned}\mathcal{U}^{L_x} \quad &= \quad \{\mathbf{u}_l = \mathbf{w}_l + \alpha\mathbf{x}_l \mid l \in \{1,2,\ldots,L_{\mathcal{U}}\},\\ &\qquad \mathsf{w}\sim\mathcal{N}(0,\sigma_{\mathsf{w}}^2 I_{L_x}), \mathsf{x}\sim\mathcal{N}(0,\sigma_{\mathsf{x}}^2 I_{L_x})\},\end{aligned} \qquad (4)$$

where $\mathbf{w}$ and $\mathbf{x}$ are realizations of two L_x-dimensional independent random processes x and w with Gaussian PDF. $L_{\mathcal{U}}$ is the total number of codebook entries and I_{L_x} denotes the L_x-dimensional identity matrix. For secure watermarking, the codebook choice must be dependent on a key K. There exists at least one such codebook such that for $L_x \rightarrow \infty$ the capacity (3) is achieved. Note that the optimum choice of the parameter α depends on the WNR and is given by

$$\alpha = \alpha^* = \frac{\sigma_{\mathsf{w}}^2}{\sigma_{\mathsf{w}}^2+\sigma_{\mathsf{v}}^2} = \frac{1}{1+10^{-\mathrm{WNR}/10}}. \qquad (5)$$

C. Practical Communication Derived from Costa’s Scheme

Costa’s scheme allows significant gains over common blind SS watermarking. Unfortunately, for good performance, $\mathcal{U}$ must be so large that neither storing it nor searching it is practical. In [10], we proposed to replace Costa’s random codebook by a structured codebook, in particular a product codebook of dithered uniform scalar quantizers, and called this scheme *SCS* (Scalar Costa Scheme). Similar suboptimal approaches to implement Costa’s scheme have been developed by Chen and Wornell [7] and by Ramkumar and Akansu [11]. A detailed description and performance analysis of SCS is given in [10, 12]. Here, only an outline of SCS is given.

In SCS, the watermark message m is encoded into a sequence of watermark letters $\mathbf{d}$, where the elements d_n belong to a D-ary alphabet $\mathcal{D} = \{0,1,\ldots,D-1\}$. In many practical cases, binary SCS watermarking ($d_n \in \mathcal{D} = \{0,1\}$) is sufficient. Each of the watermark letters is embedded into the corresponding original signal elements x_n. For example, x_n could be a signal sample or a frequency coefficient of multimedia data. The embedding rule for the nth element is given by

$$\begin{aligned}\tilde{x}_n &= x_n - \Delta\left(\frac{d_n}{D}+k_n\right)\\ q_n &= \mathcal{Q}_\Delta\left\{\tilde{x}_n\right\} - \tilde{x}_n\\ s_n &= x_n + \alpha q_n,\end{aligned} \qquad (6)$$

where $\mathcal{Q}_\Delta\left\{\cdot\right\}$ denotes scalar uniform quantization with step size Δ. The key $\mathbf{k}$ is a pseudo-random sequence with $k_n \in (0,1]$. Note that the obtained watermark signal $\mathbf{w}$ is mean-free and statistically independent from the original signal $\mathbf{x}$. The SCS embedding scheme depends on two parameters: the quantizer step size Δ and the scale factor α. For given α and embedding distortion σ_{w}^2, the step size is $\Delta = \sqrt{12}\sigma_{\mathsf{w}}/\alpha$. The parameter α is optimized for each WNR to achieve a good trade-off between embedding distortion and decoding reliability. The optimal value of α in SCS must be computed numerically[10]. A good approximation is given by

$$\alpha_{\mathrm{opt}} \approx \sqrt{\frac{\sigma_{\mathsf{w}}^2}{\sigma_{\mathsf{w}}^2+2.71\sigma_{\mathsf{v}}^2}}. \qquad (7)$$

For positive WNRs, α^* is even a slightly better approximation, but for negative WNRs, α^* is too large.

At the decoder, the received data $\mathbf{r}$ is extracted to obtain the data $\mathbf{y}$. The extraction rule for the nth element is

$$y_n = \mathcal{Q}_\Delta \left\{ r_n - k_n \Delta \right\} + k_n \Delta - r_n. \qquad (8)$$

For binary SCS, $|y_n| \leq \Delta/2$, where y_n should be close to zero if $d_n = 0$ was sent, and close to $\pm\Delta/2$ for $d_n = 1$. The extracted data $\mathbf{y}$ can be used as soft-input for common channel decoding algorithms.

The capacity $C_{\mathrm{SCS}}^{\mathrm{AWGN}}(\mathrm{WNR})$ has to be computed numerically [10]. The obtained results are shown in Section III.E.

D. Spread-Transform Watermarking

Redundant embedding of watermark message bits is required in most watermarking application. A general approach to spread watermark message bits over many original signal elements is called spread-transform (ST) watermarking. ST watermarking has been proposed by Chen and Wornell [7]. In ST watermarking, the watermark is not directly embedded into the original signal $\mathbf{x}$, but into the projection $\mathbf{x}^{\mathrm{ST}}$ of $\mathbf{x}$ onto a random sequence $\mathbf{t}$. The spreading factor τ denotes the number of elements of $\mathbf{x}$ being projected on $\mathbf{t}$ to obtain one element of $\mathbf{x}^{\mathrm{ST}}$.

The basic idea behind ST watermarking is that any component of the channel noise $\mathbf{v}$ being orthogonal to the spreading vector $\mathbf{t}$ does not impair watermark decoding. Thus, an attacker, not knowing the exact spreading direction $\mathbf{t}$, has to introduce much larger distortions to impair a ST watermark as strong as a watermark embedded directly into $\mathbf{x}$. For an AWGN attack, the effective WNR_τ after ST with spreading factor τ is given by

$$\mathrm{WNR}_\tau = \mathrm{WNR}_1 + 10 \log_{10} \tau. \qquad (9)$$

Thus, doubling the spreading length τ gives an additional power advantage of 3 dB for the watermark in the ST domain.

Below, the combination of ST watermarking with SCS watermarking is denoted as ST-SCS watermarking. ST watermarking effectively performs a mapping of the WNR according to (9). This can be exploited to compute the capacities of ST-SCS watermarking via

$$C_{\mathrm{ST-SCS},\tau}^{\mathrm{AWGN}}(\mathrm{WNR}) = \frac{C_{\mathrm{ST-SCS},1}^{\mathrm{AWGN}}(\mathrm{WNR} + 10 \log_{10} \tau)}{\tau}. \qquad (10)$$

The spreading factor τ is chosen to achieve maximum capacity. We showed in [12] that $\tau = 1$ for all $\mathrm{WNR} \geq \mathrm{WNR}_{\mathrm{crit}}$. For $\mathrm{WNR} < \mathrm{WNR}_{\mathrm{crit}}$, the optimal τ is such that the effective WNR is equal to $\mathrm{WNR}_{\mathrm{crit}}$. We determined numerically that $\mathrm{WNR}_{\mathrm{crit}} \approx 0.01$ for SCS watermarking. $C_{\mathrm{ST-SCS}}^{\mathrm{AWGN}}(\mathrm{WNR})$ denotes the capacity of ST-SCS watermarking with optimum spreading factor τ.

E. Watermark Capacity Comparison for AWGN Attacks

A detailed capacity analysis of SCS and ST-SCS watermarking and a comparison to SS and ICS watermarking in case of AWGN attacks is given in our previous work [10, 12]. Here, we summarize the most important results. Fig. 4 shows the capacities $C_{\mathrm{ICS}}^{\mathrm{AWGN}}(\mathrm{WNR})$, $C_{\mathrm{SCS}}^{\mathrm{AWGN}}(\mathrm{WNR})$, $C_{\mathrm{ST-SCS}}^{\mathrm{AWGN}}(\mathrm{WNR})$, and $C_{\mathrm{SS}}^{\mathrm{AWGN}}(\mathrm{WNR})$. The original signal power has only an influence for SS watermarking. The shown capacity $C_{\mathrm{SS}}^{\mathrm{AWGN}}(\mathrm{WNR})$ is for $\mathrm{DWR} = 15\ \mathrm{dB}$. We observe, that ST-SCS and SCS watermarking do not achieve the capacity of ICS, but are not too far from ICS either. ST-SCS watermarking gives an advantage over SS watermarking only for $\mathrm{WNR} < \mathrm{WNR}_{\mathrm{crit}} \approx 0.01\ \mathrm{dB}$. Blind SS watermarking suffers significantly from original signal interference. For weak to moderately strong attacks (i.e., WNRs greater than about $-10\ \mathrm{dB}$) SCS watermarking outperforms SS watermarking by far since the capacity of SCS is not reduced by original signal interference. For very strong attacks ($\mathrm{WNR} < -15\ \mathrm{dB}$), blind SS watermarking achieves higher capacities than (ST-)SCS watermarking since here the attack distortion becomes more important than the original signal interference. However, note that ICS outperforms blind SS watermarking at all attack distortion levels. Costa's proof shows that ICS is the optimal scheme under all possible schemes for the considered communication scenario.

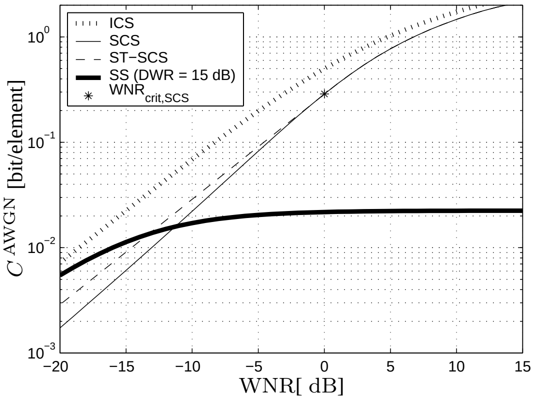

Fig. 4: Capacity of blind watermarking schemes facing an AWGN attack.

IV. Watermarking facing SAW(G)N Attacks

Watermarking of mean-free IID original signals with power $\sigma_x^2 = \mathrm{E}\left\{x^2\right\}$ is considered. The PDF of the original signal is not specified, in particular, we do not constrain the discussion to Gaussian signals. However, we investigate the watermarking game for attacks constrained to an amplitude scaling by a fix factor g_a and independent additive white noise $\mathbf{v}$ with variance σ_v^2 as depicted in Fig. 5. Subsequently, we denote this attack as amplitude scaling and additive white noise (SAWN) attack. Several of the subsequently derived results are independent from the noise PDF. When results are specific to Gaussian noise we denote the attack as amplitude scaling and additive white Gaussian noise (SAWGN) attack. Further, we do not necessarily restrict the attacker to use mean-free noise. However, we assume that the addition of a DC component can be inverted perfectly at the receiver, thus a DC-offset in the added noise sequence has no effect on the performance of the watermarking scheme and consequently is neglected here.

A. SAWN Attacks and Effective AWN Attacks

Fig. 5 depicts the investigated communication scenario. The shown scenario is more general than those investigated by Moulin et al. [2, 13] and Su et al. [14] due to the amplitude scaling by g_e at the embedder's side. The embedder chooses $\mathbf{w}'$ and g_e to transmit the watermark message m with embedding distortion D_{E}. The attacker chooses g_a and $\mathbf{v}$ constrained to the attack distortion D_{A} to disturb the watermark communication as much as possible. To solve the game between embedder and attacker, we assume that both know their opponents strategy. In our analysis, this means that the attacker knows the used watermarking scheme, the power σ_x^2 of the original signal and the introduced embedding distortion D_{E}, but does not know the exact realization of the original signal $\mathbf{x}$ and of the watermark signal $\mathbf{w}$. The receiver knows the scale factor $g = g_e g_a$, the noise variance σ_v^2 and a possibly non-zero mean of $\mathbf{v}$, but does not know the exact realizations of $\mathbf{x}$ and $\mathbf{v}$. The embedder knows all parameters known to the receiver plus the original signal $\mathbf{x}$ and the watermark message m to be sent. The assumptions about the knowledge of the attacker are in line with Kerkhoff's principle, which states that the security of a system should only depend on a secret key and not on the secrecy of the algorithm itself. A method for the estimation of the parameter required at the watermark receiver is described in [15].

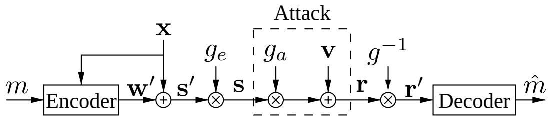

Fig. 5: Watermark communication facing an amplitude scaling and additive white noise attack. The receiver compensates for the introduced amplitude scaling before decoding $\hat{m}$.

The watermarking game in case of SAWN attacks is analyzed with respect to the watermark capacity $C^{\mathrm{SAWN}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}})$. It is obvious that $C^{\mathrm{SAWN}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}) = 0$ for $g_a = 0$ since the entire signal is deleted. Thus, it is meaningless to design a watermarking scheme if such a strong attack is allowed. If $g_a \neq 0$, the receiver can compensate for the amplitude scaling attack by dividing $\mathbf{r}$ by $g = g_e g_a$ to produce the pre-processed signal

$$\begin{aligned} \mathbf{r}' &= g^{-1}\mathbf{r} \\ &= g^{-1}(g_a \mathbf{s} + \mathbf{v}) \\ &= g^{-1}(g_a g_e(\mathbf{x} + \mathbf{w}') + \mathbf{v}) \\ &= \mathbf{x} + \mathbf{w}' + \mathbf{v}', \end{aligned} \tag{11}$$

with $\mathbf{v}' = g^{-1}\mathbf{v}$. Thus, after compensating for the amplitude scaling, the watermark receiver sees an AWN attack with the *effective* noise $\mathbf{v}'$ with variance $\sigma_{v'}^2 = \sigma_v^2/g^2$. We observe that scaling by $g < 1$ increases the effective noise power.

Fig. 5 and (11) reveal the similarity of the investigated blind watermarking scenario and the communication scenario with side-information at the encoder and AWGN channel. $\mathbf{w}'$ is the transmitted signal, $\mathbf{x}$ is the channel state known to the encoder, and $\mathbf{v}' = g^{-1}\mathbf{v}$ is the channel noise. Therefore, a blind watermarking system facing an SAWN attack can be designed similarly to a communication system with side-information at the encoder and *effective* AWN attack with noise variance $\sigma_{v'}^2$.

The watermark capacity for communication over an effective AWN channel with fix noise PDF depends only on the chosen embedding scheme, the original signal and the power ratio

$$\zeta(\sigma_{w'}^2, \sigma_{v'}^2) = \frac{\sigma_{w'}^2}{\sigma_{v'}^2}. \tag{12}$$

$\sigma_{w'}^2$ and $\sigma_{v'}^2$ of the effective AWN attack model can be related to D_{E}, D_{A}, g_e, and $g = g_e g_a$ for watermarking facing an SAWN attack. We assume $\mathrm{E}\{w'x\} = 0$ in the following derivations, which is fulfilled for all watermarking schemes discussed in Section III. Then, the embedding distortion D_{E} is given by

$$\begin{aligned} D_{\mathrm{E}} &= \mathrm{E}\left\{(s-x)^2\right\} = \mathrm{E}\left\{((1-g_e)x - g_e w)^2\right\} \\ &= (1-g_e)^2\sigma_x^2 + g_e^2\sigma_{w'}^2. \end{aligned} \tag{13}$$

Solving for $\sigma_{w'}^2$ gives

$$\sigma_{w'}^2 = \frac{D_{\mathrm{E}} - (1-g_e)^2\sigma_x^2}{g_e^2}. \tag{14}$$

For independent noise $\mathbf{v}$, the SAWN attack distortion is given by

$$\begin{aligned} D_{\mathrm{A}} &= \mathrm{E}\left\{(x-r)^2\right\} \\ &= \mathrm{E}\left\{(x - g(x+w') - v)^2\right\} \\ &= \sigma_x^2(1-g)^2 + \sigma_{w'}^2 g^2 + \sigma_{v'}^2 g^2, \end{aligned} \tag{15}$$

which gives

$$\sigma_{v'}^2 = \frac{D_{\mathrm{A}} - \sigma_x^2(1-g)^2 - \sigma_{w'}^2 g^2}{g^2}. \tag{16}$$

Finally, the power ratio in (12) can be expressed with (14) and (16) as

$$\zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}, g_e, g) = \frac{g^2(D_{\mathrm{E}} - \sigma_x^2(1-g_e)^2)}{(\sigma_x^2 - D_{\mathrm{E}})g^2 - (\sigma_x^2 - D_{\mathrm{A}})g_e^2 + 2\sigma_x^2 g g_e(g_e - g)}, \tag{17}$$

which enables the computation of the watermark capacity for an SAWN attack based on the capacity for the effective AWN attack model.

B. Solving the Watermarking Game for SAWN Attacks

The capacity $C^{\mathrm{AWN}}(\zeta)$ is for all efficient watermarking schemes monotonously increasing. Therefore, it is possible to reformulate the watermarking game constrained to SAWN attacks and specific embedding schemes using $\zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}})$ as objective function. For a certain embedding scheme and fix D_{E}, the embedder chooses g_e which maximizes $\zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}})$. The attacker chooses g_a which achieves D_{A} and minimizes $\zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}})$. Thus, the solution to the considered watermarking game is equivalent to the solution of the min-max problem

$$\begin{aligned}
&\zeta_{\mathrm{opt}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}) \\
&\quad= \min_{g_a} \max_{g_e} \zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}, g_e, g_a) \\
&\quad= \min_{g} \max_{g_e} \zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}, g_e, g). \qquad (18)
\end{aligned}$$

The solution of the min-max problem in (18) is equivalent to the saddlepoint of $\zeta(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}, g_e, g)$ over all positive (g_e,g=$g_e g_a$) if such a unique saddlepoint exists. Common analysis shows that a unique saddlepoint is given by

$$\begin{aligned}
g_{e,\mathrm{opt}} &= \frac{\sigma_x^2 - D_{\mathrm{E}}}{\sigma_x^2}, && (19)\\
g_{\mathrm{opt}} &= \frac{\sigma_x^2 - D_{\mathrm{A}}}{\sigma_x^2}, && (20)\\
g_{a,\mathrm{opt}} &= \frac{g_{\mathrm{opt}}}{g_{e,\mathrm{opt}}} = \frac{\sigma_x^2 - D_{\mathrm{E}}}{\sigma_x^2 - D_{\mathrm{A}}}. && (21)
\end{aligned}$$

We observe that $g_{e,\mathrm{opt}}$ is independent from any attack parameter and $g_{a,\mathrm{opt}}$ depends only on σ_x^2, D_{E}, and D_{A}, which are all independent from a specific watermark embedding scheme with $\mathrm{E}\{w'x\} = 0$. Thus, embedder and attacker can easily choose their optimum scale factors g_e and g_a, respectively. We assume in the remainder that always the optimum values of g_e and g_a are chosen so that the suffix "opt" can be neglected.

The optimum values of $\sigma_{w'}^2$ and $\sigma_{v'}^2$ can be derived from (14),(16), (19), and (20) to be

$$\begin{aligned}
\sigma_{w'}^2 &= \frac{\sigma_x^2 D_{\mathrm{E}}}{\sigma_x^2 - D_{\mathrm{E}}}, && (22)\\
\sigma_{v'}^2 &= \frac{{\sigma_x^2}^2 (D_{\mathrm{A}} - D_{\mathrm{E}})}{(\sigma_x^2 - D_{\mathrm{A}})(\sigma_x^2 - D_{\mathrm{E}})}. && (23)
\end{aligned}$$

Thus, the solution to (18) is

$$\zeta_{\mathrm{opt}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}) = \frac{D_{\mathrm{E}}(\sigma_x^2 - D_{\mathrm{A}})}{\sigma_x^2 (D_{\mathrm{A}} - D_{\mathrm{E}})}. \qquad (24)$$

Note that for $g_{e,\mathrm{opt}}$, the embedding distortion D_{E} never exceeds the original signal power σ_x^2, which can be concluded from (19). Further, it can be observed that $g_{e,\mathrm{opt}} \approx 1$ for practically relevant ratios $D_{\mathrm{E}}/\sigma_x^2$. Thus, the amplitude scaling by g_e at the embedder's side does not give a significant performance improvement over the schemes considered in [2, 13, 14], but in principle an improvement is achieved. Further note that we defined the watermark signal $\mathbf{w}$ to be the difference between the original and the watermarked signal which is here

$$\mathbf{w} = \mathbf{s} - \mathbf{x} = (g_e - 1)\mathbf{x} + g_e \mathbf{w}'. \qquad (25)$$

Thus, for $\mathrm{E}\{w'x\} = 0$, $\mathbf{x}$ and $\mathbf{w}$ are correlated for all $g_e \neq 1$, which contradicts earlier results obtained in [2].

We observe that the minimum attack distortion D_{A}, achieved for the weakest attack ($\sigma_{v'}^2 = 0$), is $D_{\mathrm{A}} = D_{\mathrm{E}}$. Therefore, meaningful embedding distortions and attack distortions are constrained to

$$0 \leq D_{\mathrm{E}} \leq D_{\mathrm{A}} \leq \sigma_x^2. \qquad (26)$$

Note that the upper limit on D_{E} has no meaning in practical watermarking scenarios. In general, the embedder can choose any distortion level between zero and σ_x^2 since he has the first move in the considered game.

C. Watermarking of IID Gaussian Signals

The optimization of the parameter σ_v^2 and g_a of an SAWN attack required only weak assumptions on the statistics of the original signal and the considered watermarking schemes. In particular, we exploited the knowledge of the original signal power σ_x^2, of the embedding distortion D_{E}, and constrained the embedding schemes to schemes with $\mathrm{E}\{xw'\} = 0$ and a monotonously increasing $C^{\mathrm{AWN}}(\zeta)$. Now, the specific case of Gaussian original signals and blind watermarking using an ideal Costa scheme is considered. Further, it is assumed that the attacker uses Gaussian noise. For such a scenario, the capacity for communication over an effective AWGN channel is given by

$$C_{\mathrm{ICS}}^{\mathrm{AWGN}} = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_{w'}^2}{\sigma_{v'}^2} \right). \qquad (27)$$

With (22) and (23) for $\sigma_{w'}^2$ and $\sigma_{v'}^2$, respectively, we obtain the capacity in case of SAWGN attacks, which is

$$C_{\mathrm{ICS}}^{\mathrm{SAWGN}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}) = \frac{1}{2} \log_2 \left(\frac{D_{\mathrm{A}}(\sigma_x^2 - D_{\mathrm{E}})}{\sigma_x^2 (D_{\mathrm{A}} - D_{\mathrm{E}})} \right). \qquad (28)$$

The watermarking game for Gaussian original signals has been first investigated by Moulin et al.[2], however, with a differently defined attack distortion measure. Moulin et al. derive that the optimum attack under all possible attacks is a specific SAWGN attack, the Gaussian test channel (GTC). Later, in [13], Moulin et al. consider the same attack distortion measure used here and derive that a specific SAWGN attack is again the optimum attack under all possible attacks. However, details of the proof are currently not available to the authors. Note, that the capacity in (28) is slightly larger than the one given in [13], which suggests that the solution in [13] is not as general as assumed before. The important difference between the analysis here and the one given in [13] is the introduced amplitude scaling by g_e as final step during watermark embedding. The model in [13] does not consider such an amplitude scaling an thus restricts the embedding process in an inefficient way. The scaling by g_e prevents that an attacker first improves the quality of the watermarked signal $\mathbf{s}$ before adding noise with increased power.

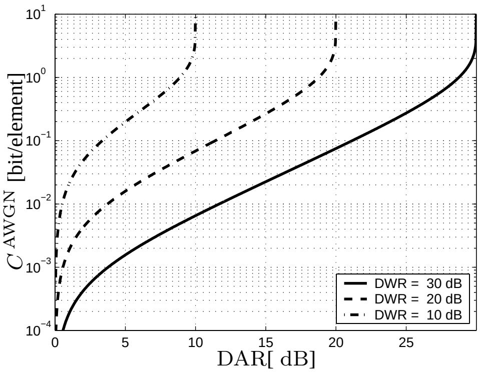

Fig. 6: Watermark capacities for Gaussian original signals and ICS watermarking.

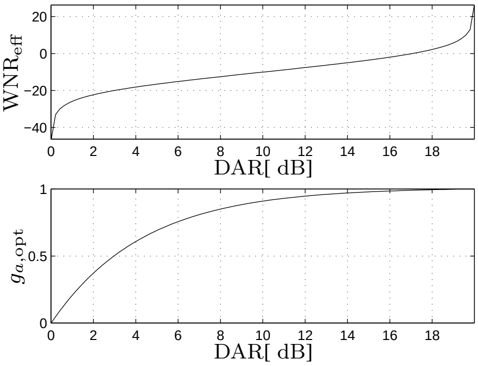

Fig. 7: Effective WNR and optimum scale factor g_{opt} for $\mathrm{DWR} = 20$ dB.

Fig. 6 depicts $C_{\mathrm{ICS}}^{\mathrm{SAWGN}}$ for three different DWRs. The capacities are plotted over the entire range of achievable document-to-attack power ratios ($\mathrm{DAR} = 10\log 10(\sigma_x^2/D_{\mathrm{A}})[\text{ dB}]$). Note that the DWR and the DAR can be considered as a document quality measure, where large values indicate good document quality. The watermark capacity goes to infinity when DAR tends to the DWR, that means no attack occured. Contrary, the watermark capacity goes to zero for DAR approaching its lower limit zero when the entire signal is erased.

The upper plot in Fig. 7 shows the relationship between the DAR after the optimized SAWGN attack and the corresponding effective WNR for communication over an AWGN channel. Note that the effective WNR goes to plus or minus infinity as the DAR achieves it upper or lower limit, respectively. The lower plot in Fig. 7 depicts the corresponding optimal scale factor g_a. For weak attacks, where the quality loss is less than 6 dB, the attacker mainly adds noise. However, when increased attack distortions are accessible, the attacker more and more scales down the watermarked signal. At the limit $\mathrm{DAR} = 0$, the entire watermarked signal is erased by choosing $g_a = 0$.

D. Suboptimum Watermarking Schemes

In this section, the performance of suboptimum watermarking schemes facing an SAWGN attack is discussed. We compare SS, SCS, and ST-SCS watermarking with ICS watermarking.

The capacity $C_{\mathrm{SS}}^{\mathrm{SAWGN}}$ can be derived from (2), which gives

$$C_{\mathrm{SS}}^{\mathrm{SAWGN}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}}) = \frac{1}{2}\log_2\left(\frac{\sigma_x^2(\sigma_x^2 - D_{\mathrm{E}})}{\sigma_x^2(\sigma_x^2 - D_{\mathrm{E}}) - D_{\mathrm{E}}(\sigma_x^2 - D_{\mathrm{A}})}\right) \quad (29)$$

The capacity $C_{\mathrm{SCS}}^{\mathrm{AWGN}}$ has been derived numerically, thus, an analytical expression of $C_{\mathrm{(ST-)SCS}}^{\mathrm{SAWGN}}$ is not available. In the following comparison, we exploit the derived mapping of the DAR after SAWGN attack onto the effective WNR for AWGN channels to compute the capacities $C_{\mathrm{(ST-)SCS}}^{\mathrm{SAWGN}}(\sigma_x^2, D_{\mathrm{E}}, D_{\mathrm{A}})$. For this, the capacity $C_{\mathrm{SCS}}^{\mathrm{AWGN}}(\mathrm{WNR})$ is computed numerically for $\mathrm{WNR} = -20\text{ dB}\ldots 20\text{ dB}$. Effective $\mathrm{WNR} > 20$ dB occur for DARs close to the DWR. Here, linear extrapolation of $C_{\mathrm{SCS}}^{\mathrm{AWGN}}(\mathrm{WNR})$ for $\mathrm{WNR} > 20$ dB with its derivative at $\mathrm{WNR} = 20$ dB is applied. For $\mathrm{WNR} < -20$ dB, a reasonable extrapolation of the numerically derived capacity curve can be obtained with help of the spread transform. Thus, in a strict sense, the shown curves for $C_{\mathrm{SCS}}^{\mathrm{SAWGN}}$ are valid for ST-SCS watermarking where the ST is active only for $\mathrm{WNR} < -20$ dB. Contrary, $C_{\mathrm{ST-SCS}}^{\mathrm{SAWGN}}$ denotes the capacity of ST-SCS watermarking where the ST is active for all $\mathrm{WNR} < \mathrm{WNR}_{\mathrm{crit}} \approx 0.01$ dB.

Fig. 8 compares the performance of the considered watermarking schemes for three different levels of the embedding strength ($\mathrm{DWR} = 10$ dB, 20 dB, 30 dB). The plots in the left column of Fig. 8 show the capacity in units of bit/element. The capacity $C_{\mathrm{SS}}^{\mathrm{SAWGN}}$ is limited over the entire range of DARs due to the original signal interference. Consequently, SS watermarking performs significantly better for decreased DWR. The minimum usable DWR is application dependent. Contrary, all other considered techniques could achieve in principle an infinite large capacity for DARs close to the DWR. The capacity of ICS and (ST-)SCS watermarking depends on the DWR, as well. However, in particular for strong attacks ($\mathrm{DAR} \to 0$), the influence of the DWR is almost negligible. The influence of the DWR increases for increased DARs which is obvious since the capacities tend to infinity when the DAR achieves the DWR. We can also observe that the capacities of (ST)-SCS watermarking follow in general the behavior of the capacity of ICS. At a certain DAR, depending on the DWR, (ST-)SCS performs worse than SS watermarking. However, it is unlikely that such low DARs are allowed in practical scenarios.

The capacity ratio $\nu_C = C^{\mathrm{SAWGN}}\ /\ C_{\mathrm{ICS}}^{\mathrm{SAWGN}}$ of the suboptimal schemes and an ideal Costa scheme is depicted in the right column of Fig. 8. First, we observe that (ST-)SCS achieves for weak attacks almost the capacity of ICS, where for stronger attacks a significant

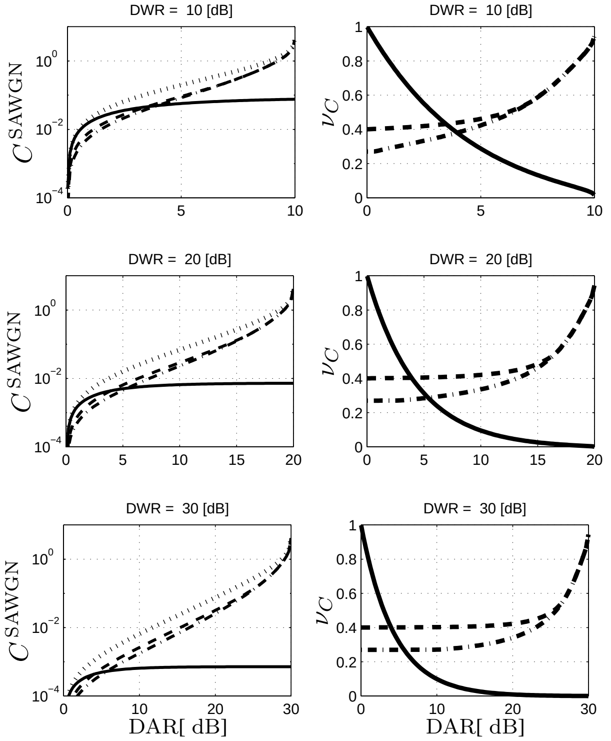

Fig. 8: Capacity comparison for watermarking facing the SAWGN attack for DWR= 10 dB, 20 dB, 30 dB in case of white Gaussian original signals (ICS: "· · ·"; SS:"–"; SCS:"-·-·"; ST-SCS:"- -").

capacity reduction is visible. However, we can also observe that the capacity reduction for ST-SCS and SCS is limited and never below 40% and 28%, respectively. This result is independent from the considered DWR. SS watermarking achieves the performance of ICS for very strong attacks (DAR ≈ 0). However, for weaker attacks, the capacity loss becomes more and more severe. For very weak attacks, the relative capacity of SS compared to an ideal scheme tends to zero.

V. Conclusions

A communications perspective of digital watermarking is presented, where the performance of digital watermarking schemes is characterized by the watermark capacity for certain embedding schemes and constrained attacks. The watermark capacity of four different blind watermarking schemes facing AWGN attacks is reviewed. The considered blind watermarking schemes are common blind spread-spectrum (SS) watermarking, the ideal Costa scheme (ICS) which exploits the side-information about the original signal at the encoder, the practical but suboptimal scalar Costa scheme (SCS) and its combination with spread-transform watermarking (ST-SCS). Next, attacks by amplitude scaling and additive white (Gaussian) noise (SAW(G)N) are discussed. We show that these attacks can be translated into effective AWGN attacks, so that the previously presented capacity results for AWGN attacks can be used for the capacity analysis of SAWGN attacks. We also show, that optimal watermark embedding in case of SAWGN attacks produces watermark signals which are correlated with the original signal. This contradicts previous results presented by Moulin et al. [2]. The watermark capacities for ICS, (ST-)SCS, and SS embedding and SAWGN attacks are compared. An important results is that the practical ST-SCS watermarking scheme achieves at least 40 % of the capacity of ICS.

VI. Acknowledgements

The authors thank Jonathan Su for helpful discussions which improved the clarity of Section IV significantly.

VII. References

[1] B. Chen and G. W. Wornell, "Provably robust digital watermarking," in *Proceedings of SPIE: Multimedia Systems and Applications II (part of Photonics East '99)*, vol. 3845, (Boston, MA, USA), pp. 43–54, September 1999.

[2] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding." Preprint, September 1999.

[3] J. K. Su and B. Girod, "Power-spectrum condition for energy-efficient watermarking," in *Proceedings of the IEEE Intl. Conference on Image Processing 1999 (ICIP '99)*, (Kobe, Japan), October 1999.

[4] F. Hartung and B. Girod, "Digital watermarking of raw and compressed video," in *Proceedings EUROPTO/SPIE European Conference on Advanced Imaging and Network Technologies*, (Berlin, Germany), October 1996.

[5] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, 1997.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

[7] B. Chen and G. W. Wornell, "Achievable performance of digital watermarking systems," in *Proceedings of the IEEE Intl. Conference on Multimedia Computing and Systems (ICMCS '99)*, vol. 1, pp. 13–18, (Florence, Italy), pp. 13–18, June 1999.

[8] 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*, vol. 87, pp. 1127–1141, July 1999.

[9] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, May 1983.

[10] 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*, (London, UK), pp. 4/1–4/6, April 2000.

[11] M. Ramkumar and A. N. Akansu, "Self-noise suppression schemes in blind image steganography," in *Proceedings of SPIE: Multimedia Systems and Applications II (part of Photonics East '99)*, vol. 3845, (Boston, MA, USA), pp. 55–65, September 1999.

[12] 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.

[13] P. Moulin, M. K. Mihcak, and G.-I. A. Lin, "An information-theoretic model for image watermarking and data hiding," in *Proceedings of the IEEE Intl. Conference on Image Processing 2000 (ICIP 2000)*, (Vancouver, Canada), September 2000.

[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*, vol. 81, June 2001.

[15] J. J. Eggers, R. Bäuml, and B. Girod, "Estimation of amplitude modifications before SCS watermark detection." submitted, January 2002.