Inverse Mapping of SCS-Watermarked Data

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

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

ABSTRACT

Scalar Costa Scheme (SCS) watermarking is a practical version of Costa's capacity achieving blind watermarking scheme. In this paper, inversion of SCS, that is the removal of the embedded signal, is investigated. For the noiseless case, where no attack is performed on the watermarked data, SCS watermarking can be inverted perfectly. For the case of an AWGN attack, an MMSE estimate for the original hostdata is presented, which gives an optimum estimate of the hostdata prior to watermark embedding.

1 Introduction

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. Usually there is no intention to remove the watermark within those applications. But one can think of scenarios as well where it is desirable to completely remove a watermark to restore the original data exactly. Examples are information hiding applications with medical images [5] or multiple watermark reception. In applications dealing with medical images, the goal is mainly to recover the original signal with a minimum amount of distortion. In multiple watermark reception, the interference of the first decoded watermark on other embedded watermarks should be minimized. For this, the already decoded watermark is exploited to remove the corresponding embedding distortion as much as possible.

In Section 2 we will present a brief review about the underlying waterwarking scheme and Section 3 covers the inversion of SCS watermarking in the noiseless and noisy case.

2 SCS watermarking

A general model for the communication of a message via watermarking can be described as follows: The encoder derives from the watermark message and the host data $\mathbf{x}$ an appropriate watermark sequence $\mathbf{w}$ which is added to the host data to produce the watermarked data $\mathbf{s}$. $\mathbf{w}$ must be chosen such that the distortion between $\mathbf{x}$ and $\mathbf{s}$ is negligible. Next, an attacker might modify the watermarked data $\mathbf{s}$ into data $\mathbf{r}$ to impair watermark communication. The attack is only constrained with respect to the distortion between $\mathbf{x}$ and $\mathbf{r}$. Finally, the decoder must be able to detect the watermark message from the received data $\mathbf{r}$. In *blind* watermarking schemes, the host data $\mathbf{x}$ is not available to the decoder. The codebook used by the watermark encoder and decoder is randomized dependent on a key $\mathbf{k}$ to achieve secrecy of watermark communication. In this paper, $\mathbf{x}$,$\mathbf{w}$,$\mathbf{s}$,$\mathbf{r}$ and $\mathbf{k}$ are vectors, and x_n,w_n,s_n,$r_n$and $k_n$refer to their respective nth elements, with random variables written in sans serif, e.g. x.

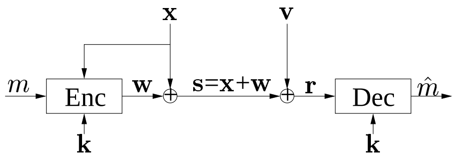

Figure 1: Watermark encoding followed by an AWGN attack.

It has been shown that blind watermarking can be considered communication with side information at the encoder [1]. Costa [2] 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 $C = 0.5\log_2(1+\sigma_{\mathsf{w}}^2/\sigma_{\mathsf{v}}^2)$ 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 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 it is replaced by a structured codebook, in particular a product codebook of dithered uniform scalar quantizers. The name *SCS* (Scalar Costa Scheme) [3] is derived from this codebook structure.

The watermark message m is encoded into a sequence of watermark letters $\mathbf{d}$, where $d_n \in \mathcal{D} = \{0, 1\}$ in the 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

$$\begin{array}{rcl} a[n] & = & \Delta\left(\dfrac{d_n}{2}+k_n\right) \\ x_{q,n} & = & \mathcal{Q}_\Delta\left\{x_n - a[n]\right\} \\ s_n & = & x_n + \alpha(x_{q,n} - (x_n - a[n])), \end{array} \tag{1}$$

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 $k_n \in (0,1]$. 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. Optimal values for Δ and α are given in [3].

Watermark detection is based on the pre-processed received data $\mathbf{y}$. The extraction rule for the nth element is

$$y_n = \mathcal{Q}_\Delta\left\{r_n - k_n\Delta\right\} + k_n\Delta - r_n, \qquad (2)$$

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 $d_n = 1$.

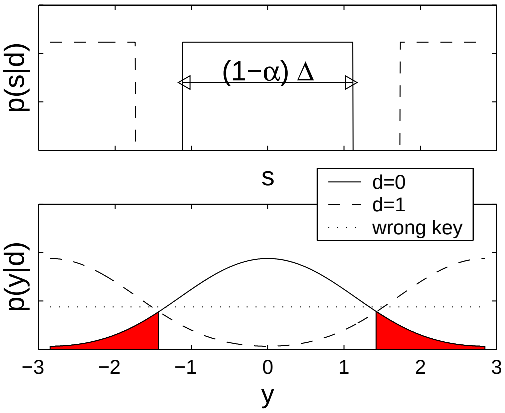

Figure 2: One period of the PDFs of the sent and the received signal for binary SCS (σ_w^2=1, WNR $=$ 3dB, $\Delta = 6$, $\alpha = 0.58$). The filled areas represent the probability of detection errors assuming $d = 0$ was sent. The dotted line in the lower plot depicts the PDF when detecting with a wrong key $\mathbf{k}$.

The upper plot of Fig. 2 depicts one period of the PDF of the sent elements s conditioned on the sent watermark letter and $k_n = 0$. The lower plot shows the PDF of the pre-processed received elements y after AWGN attack conditioned on the sent watermark letter. The derivation of $p_y\left(y_n|d_n\right)$ is given in [3].

3 Inverse SCS

Perfect recovery of the original signal from the received signal might be impossible in many practical cases, e.g., attack noise cannot be removed in general. However, in some cases it is sufficient to produce a signal that is closer to the original signal than the received signal. In this section, ways to invert SCS watermarking are discussed. In practice, the receiver sees an attacked watermarked signal. Here, a simple AWGN attack is considered again. For completeness the noiseless case is discussed first. Throughout the section, it is assumed that the transmitted sequence of watermark letters $\mathbf{d}$ and the correct key sequence $\mathbf{k}$ are perfectly known, e.g., correct decoding has been performed, which can be treated without loss of generality as $\mathbf{d} = \mathbf{0}$ and $\mathbf{k} = \mathbf{0}$. The effect of possible remaining bit errors after error correction decoding, and thus imperfect knowledge of $\mathbf{d}$, is not investigated. However, it is obvious that for low bit-error rates the influence of the incorrect inverse mapping applied to those samples with incorrectly received dither samples $\hat{d}_n$ on the overall quality improvement by inverse SCS is negligible.

3.1 Inverse SCS in the Noiseless Case

For SCS watermark embedding, the quantization error $q_n = x_{q,n} - x_n$ is scaled by α to obtain the watermark sample w_n that is embedded into x_n by simple addition as given in (1). An alternative formulation of the SCS embedding rule is

$$s_n = x_{q,n} - (1-\alpha)q_n. \qquad (3)$$

This shows that the original signal can be recovered from r_n for $r_n = s_n$ by extracting the value

$$\overline{y}_n \quad = \quad x_{q,n} - r_n = (1-\alpha)q_n, \qquad (4)$$

and inverting the watermark embedding by

$$r_n = s_n = x_{q,n} - \frac{\overline{y}_n}{1-\alpha} = x_{q,n} - \frac{1-\alpha}{1-\alpha}q_n = x_n. \qquad (5)$$

The perfect invertibility of SCS is illustrated by the input-output characteristic of SCS embedding for $\alpha = 0.6$, $d_n = 0$, and $k_n = 0$ and the corresponding inverse SCS shown in Fig. 3. The input-output characteristic of SCS embedding is a strictly increasing function so that the inverse mapping in the noiseless case exists. This inverse mapping is obtained by mirroring the input-output characteristic of SCS embedding at that for the identity mapping $x_n = s_n$.

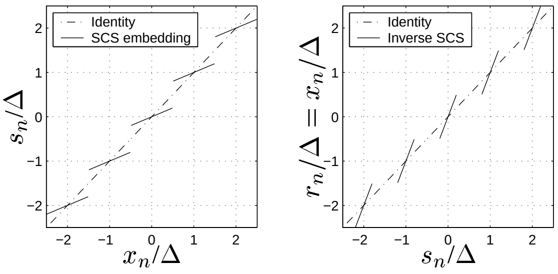

Figure 3: Input-output characteristic for SCS embedding (left) and inverse SCS (right) in the noiseless case. The example is for $\alpha = 0.6$, $d_n = 0$, and $k_n = 0$.

3.2 Inverse SCS after AWGN Attack

Inversion of SCS watermarking after transmission over an AWGN channel is considered. It is impossible to reconstruct $\mathbf{s}$ or the additive noise signal $\mathbf{v}$ from the received signal $\mathbf{r}$ even with perfect knowledge of $\mathbf{d}$ because the transmitted value s_n depends also on the original signal value x_n that is not known to the receiver. Consequently, it is impossible to recover the host signal perfectly, however, one can at least try to find an estimate $\hat{\mathbf{x}}$ so that $\mathrm{D}\left(\mathbf{x}, \hat{\mathbf{x}}\right) \leq \mathrm{D}\left(\mathbf{x}, \mathbf{r}\right)$, where again the MSE distortion measure is adopted.

3.2.1 Estimation of the Original Signal

The minimum mean-squared error (MMSE) estimate $\hat{x}_n$ of the original signal sample x_n should be derived for each received sample r_n. IID signals are assumed so that the sample index n is suppressed in the following. With help of the known key sequence sample k and known watermark letter d, the deviation $\overline{y} \in [-\frac{\Delta}{2}, \frac{\Delta}{2})$ from the next valid SCS codebook entry r_q is given by

$$\overline{y} = r - r_q, \tag{6}$$

with

$$r_q = \mathcal{Q}_\Delta \left\{ r - \Delta \left(\frac{d}{D} + k \right) \right\} + \Delta \left(\frac{d}{D} + k \right). \tag{7}$$

For AWGN attacks, the most likely corresponding quantized original signal sample is $\hat{x}_q = r_q$. Thus, the MMSE estimate $\hat{x}$ is

$$\begin{aligned} \hat{x}(r_q, \overline{y}) &= \arg \min_{x_t \in \mathbb{R}} \mathrm{E}\left\{ (x_t - x)^2 | r = r_q + \overline{y} \right\} \\ &= r_q + \hat{y}(r_q, \overline{y}), \end{aligned} \tag{8}$$

with

$$\hat{y}(r_q, \overline{y}) = \arg \min_{\hat{y}_t \in [-\frac{\Delta}{2}, \frac{\Delta}{2})} \mathrm{E}\left\{ (r_q + \hat{y}_t - x)^2 | r = r_q + \overline{y} \right\} \tag{9}$$

where r_q is no longer considered within the minimization, and $\hat{y}(r_q, \overline{y}) \in [-\frac{\Delta}{2}, \frac{\Delta}{2})$ has to be chosen such that the MSE $\mathrm{E}\left\{ (\hat{\mathsf{x}} - \mathsf{x})^2 \right\}$ is minimized. Straightforward analysis shows that $\hat{y}(r_q, \overline{y})$ has to be computed by

$$\begin{aligned} \hat{y}(r_q, \overline{y}) &= \mathrm{E}\left\{ \mathsf{x} | r = r_q + \overline{y} \right\} - r_q \\ &= \int_{-\infty}^{\infty} x \, p_{\mathsf{x}}(x | r = r_q + \overline{y}) \, \mathrm{d}x - r_q. \end{aligned} \tag{10}$$

Thus, the estimation problem is reduced to finding the conditional PDF $p_{\mathsf{x}}(x | r = r_q + \overline{y})$. It is assumed that $p_{\mathsf{x}}(x | r = r_q + \overline{y})$ is independent from r_q, which is approximately valid for AWGN attacks and an almost flat PDF $p_{\mathsf{x}}(x)$ in the range of one quantization interval, e.g., fine quantization, so that $\hat{y}(r_q, \overline{y}) = \hat{y}(\overline{y})$. Thus, the random variable $\overline{\mathsf{y}}$ with support in $[-\frac{\Delta}{2}, \frac{\Delta}{2})$ is introduced and the PDF $p_{\mathsf{x}}(x | r = r_q + \overline{y}) = p_{\mathsf{x}}(x | \overline{\mathsf{y}} = \overline{y})$ is considered in the following.

First, Bayes' rule is applied which yields

$$p_{\mathsf{x}}(x | \overline{y}) = \frac{p_{\mathsf{x}}(x) \, p_{\overline{\mathsf{y}}}(\overline{y} | x)}{p_{\overline{\mathsf{y}}}(\overline{y})}. \tag{11}$$

It can be shown that under the assumption of AWGN attacks, where $\mathbf{v}$ realizes a Gaussian noise process with variance σ_{v}^2 , a sufficiently accurate approximation for $p_{\overline{\mathsf{y}}}(\overline{y})$ is obtained by considering only $x \in [-\frac{3\Delta}{2}, \frac{3\Delta}{2})$. Assuming a reasonably flat PDF $p_{\mathsf{x}}(x)$ in the range of a few quantizer steps Δ , $p_{\mathsf{x}}(x | \overline{y})$ can be approximated as

$$p_{\mathsf{x}}(x | \overline{y}) \approx \frac{p_{\overline{\mathsf{y}}}(\overline{y} | x)}{\int_{-\frac{3\Delta}{2}}^{\frac{3\Delta}{2}} p_{\overline{\mathsf{y}}}(\overline{y} | x) \mathrm{d}x} \forall x \in [-\frac{3\Delta}{2}, \frac{3\Delta}{2}) \tag{12}$$

with an appropriate analytical approximation for $p_{\overline{\mathsf{y}}}(\overline{y} | x)$. The desired mapping $\overline{y} \rightarrow \hat{y}$ can be computed numerically with (10) for $r_q = 0$ and $x \in [-\frac{3\Delta}{2}, \frac{3\Delta}{2})$. A more detailed derivation and description of the numerical evaluation of (10) is omitted here due to space constraints, but can be found in [4].

3.2.2 Achievable Distortion Reduction

The derived mapping from $\overline{y}$ to $\hat{y}$ is illustrated for different channel noise variances. Further, the achievable distortion improvement is investigated.

Fig. 4 depicts the PDFs $p_{\overline{\mathsf{y}}}(\overline{y})$ and $p_{\hat{\mathsf{y}}}(\hat{y})$ to demonstrate the result of the mapping operation for $\mathrm{WNR} = \mathrm{WNR}_{\mathrm{design}} = 0$ dB . In this case, values near $\overline{y} = 0$ are pushed in the direction of $\Delta/2$. This could have been expected since the SCS embedding rule pushes the original samples in the direction of 0. More interesting is the mapping for values close to $\Delta/2$. In this range, it is more likely that the channel noise, not SCS embedding, has pushed the watermarked data into the direction of $\Delta/2$. Thus, $|\hat{y}| < |\overline{y}|$ for $\overline{y}$ close to $\pm\Delta/2$.

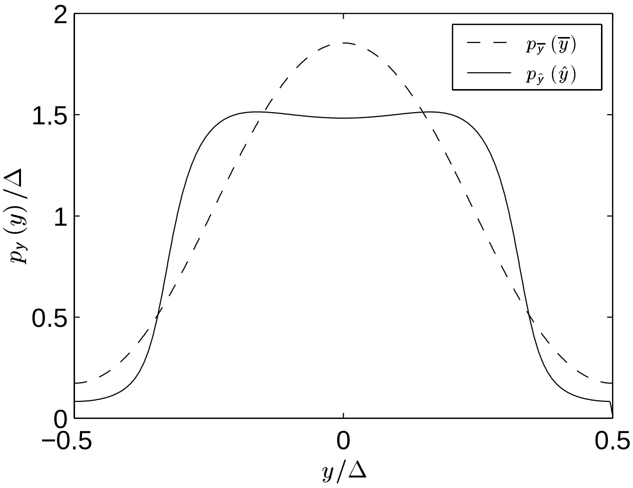

Figure 4: PDFs of received extracted data before and after inverse SCS mapping ($\mathrm{WNR} = \mathrm{WNR}_{\mathrm{design}} = 0$ dB)

Fig. 5 shows the mapping $\overline{y} \rightarrow \hat{y}$ for $\mathrm{WNR} = 0$ dB and $\mathrm{WNR} = 6$ dB where the step size Δ is such that maximum capacity is achieved for $\mathrm{WNR}_{\mathrm{design}} = 0$ dB . For the case of $\mathrm{WNR} = 6$ dB the SCS watermark has been designed very conservatively so that Δ is much larger than necessary for the actual channel noise.

As we can derive from the mapping for $\mathrm{WNR} = 6$ dB in Fig. 5, the samples are moved consequently to the interval boundaries in this case. In the noiseless case, the PDF of $\hat{y}$ is assumed to be uniform over the range $(-\Delta/2, \Delta/2]$. Further, the mapping almost never pushes samples to the interval center. Only values very close to the interval boundaries are moved a little bit in the direction of the interval center.

Fig. 6 shows the mapping rule of $\overline{y}$ and $\hat{y}$ for $\mathrm{WNR} = \mathrm{WNR}_{\mathrm{design}} = 5$ dB. The same tendency as for $\mathrm{WNR} =$

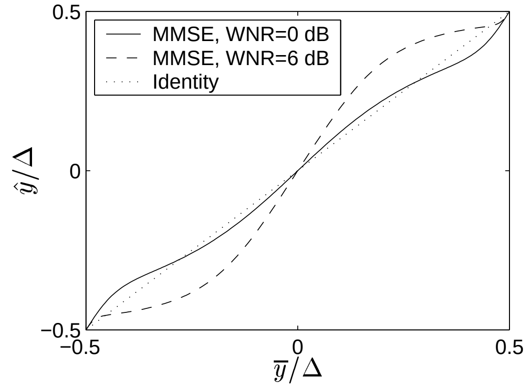

Figure 5: Inverse MMSE SCS mapping for $\mathrm{WNR} = 0$ dB and $\mathrm{WNR} = 6$ dB with $\mathrm{WNR}_{\mathrm{design}} = 0$ dB .

0 dB can be observed, however, in particular, the values close to $\pm\Delta/2$ are moved further into the center of the quantization interval. The reason for this is that for higher WNR, the optimal value of α is higher, and thus the watermarked data is concentrated more tightly around the interval center. Only large noise samples could have pushed the data close to $\pm\Delta/2$. Since the Gaussian PDF decreases exponentially, it is more likely that the received y belongs to the current quantization interval, than that it has been pushed by noise into the current quantization interval.

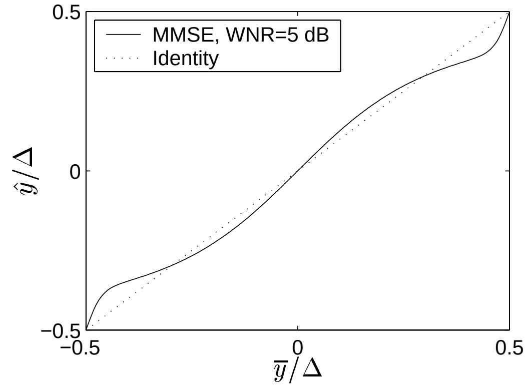

Figure 6: Inverse SCS mapping for $\mathrm{WNR} = 5$ dB .

Finally, the achieved distortion improvement is investigated. The improvement is measured in terms of the difference between the DAR (document-to-attack power ratio) before (DAR_r) and after ($\mathrm{DAR}_{\hat{x}}$) the mapping, which is given by

$$\mathrm{DAR}_{\mathrm{imp}} = \mathrm{DAR}_{\hat{x}} - \mathrm{DAR}_r = 10\log_{10}\frac{\mathrm{D}\left(x,r\right)}{\mathrm{D}\left(x,\hat{x}\right)}\ \mathrm{dB}. \quad (13)$$

Simulations show, that, depending on the actual WNR used, distortion improvements up to $\mathrm{DAR}_{\mathrm{imp}} = 0.04$ dB can be achieved in the case of AWGN attacks with $\sigma_v^2 = \sigma_{v;\max}^2$. This result is disappointing, as the gain is negligible in practical cases. Obviously, the optimal quantizer step size in SCS is such that, after AWGN attacks, the watermark embedding distortion is no longer invertible. Yet, for an over-design of the SCS quantizer step size $\Delta$for a noise power being 6 dB above the given channel noise power, the maximum distortion improvement is about 2.2 dB. Although this improvement might be of interest in practice, it is important to emphasize that such an improvement could be obtained only for very mild channel conditions.

4 Conclusion

It has to be concluded that the inversion of SCS watermarking after AWGN attacks is practically impossible or at least inefficient for attack scenarios where the actual attack noise matches the expected attack noise the watermark has been designed for. Nevertheless, the derived inverse SCS mapping might be useful in several cases. Suppose the owner of a signal stores only the SCS watermarked version and erases the original. In this case, the SCS watermark might be designed for strong robustness, that is, low WNRs. However, even without an explicit attack, the watermarked signal is slightly distorted due to quantization, which might occur when storing the data. This quantization can be approximated by low-power noise. In such a scenario, the inverse scaling derived for the noiseless case might be not appropriate, but the MMSE estimation removes a good deal of the distortion introduced by the SCS watermark, as mentioned for an overdesign of 6 dB. A typical environment for such a scenario can be found when watermarking medical images.

References

[1] B. Chen and G. Wornell. Preprocessed and postprocessed quantization index modulation methods for digital watermarking. In *Proc. of SPIE Vol. 3971: Security and Watermarking of Multimedia Contents II*, pages 48–59, San Jose, Ca, USA, January 2000.

[2] M. H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[3] 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.

[4] Joachim J. Eggers. *Information Embedding and Digital Watermarking as Communication with Side Information*. PhD thesis, Lehrstuhl für Nachrichtentechnik I, Universität Erlangen-Nürnberg, Erlangen, Germany, November 2001. preprint.

[5] B. Macq and F. Dewey. Trusted headers for medical images. In *V^3D^2 Watermarking Workshop*, Erlangen, Germany, October 1999.