

PROVABLY OR PROBABLY ROBUST DATA HIDING?

Félix Balado and Fernando Pérez-González

Signal Theory and Communications Department
University of Vigo, E-36200 Vigo, Spain
{fiz, fperez}@tsc.uvigo.es

ABSTRACT

It has been claimed that quantization-based data-hiding methods offer an advantage over spread-spectrum schemes. In this paper we review this assertion by presenting a new look on the assumptions made for these claims, and we give a new performance comparison between both approaches under random additive channel distortions. The existence of a threshold in the distortion level for the goodness of each of the methods is shown.

1. INTRODUCTION

Lately, the watermarking community has witnessed the emergence of a new data hiding philosophy based on the use of side information. The first deliberate use of this concept dates back to the identification of the watermarking problem not just with a classical communications problem, but with a special communications scenario for which, in the absence of attacks, the channel state is known beforehand by the transmitter [1]. The revisiting of an important result by Costa [2], that showed how to obtain host signal interference rejection under certain conditions, gave the first rigorous theoretical framework for addressing data hiding with side information. Two equivalent practical implementations of this result were proposed: Distortion Compensated Quantization Index Modulation (DC-QIM) [3] and Scalar Costa Scheme (SCS) [4], based on dithered quantizers. These methods present the characteristic of performing detection in a deterministic way, i.e. without considering the host signal statistics; thus, we will term them as *known host-state* methods.

On the other hand we have the methods that had previously received most attention, i.e. spread-spectrum-like methods. These schemes do always present host signal interference, even in the absence of attacks. Commonly, they just make limited or null use of side information (e.g. only

to undertake perceptual analysis), but in their best performing versions they feature maximum likelihood (ML) decoding. This implies that knowledge of the host signal statistical distribution is needed, what suggests to name them *known host-statistics* methods.

1.1. The need for a new comparison

Several claims that known host-state methods offer superior performance than known host-statistics ones have been made. Nevertheless, previous comparisons presented some flaws that demand a more detailed review of these claims. First of all, only a fair measure can be used for a fair comparison: parameters such as the probability of decoding error (P_e) or capacity are significant because they provide factual design parameters; alternatively, a comparison based on “penalty distortions” or signal-to-noise ratio (SNR) advantages is condemned to be skewed, because its translation into objective parameters measuring actual performance depends on the (disregarded) underlying host signal statistics. Taking into account that channel distortions (either intentional or unintentional) constitute a basic assumption of the data hiding game, in this paper we focus on “robustness in probability” and show that “provable robustness” (i.e. zero-error probability for certain amplitude-bounded distortions) does not necessarily lead to the former.

Secondly, the key issue of distortion measurement cannot be overlooked. The measure used in the assumptions of Costa’s result —mean squared error (MSE)— seems to be inadequate for the data hiding problem, in which only *local* measures have a perceptual meaning whatever the domain used. With a MSE measure perceptually unacceptable local distortions could be globally compensated to meet the established restriction. A rough approximation to a local measure is the pointwise distortion measure that we will use here.

In addition, the Gaussianity hypotheses of Costa’s theorem are difficult to verify in practice. Gaussian analysis would prove interesting if it would yield an effective bound to performance; unfortunately, we will see that this is not always the case. In this situation, disregarding the host signal statistical distribution in detection equals to throwing away

Work partially supported by the *Xunta de Galicia* under project PGIDT01 PX13204PM, the European project Certimark (Certification of Watermarking Technologies), IST-1999-10987, and the CYCIT project AMULET, reference TIC2001-3697-C03-01.

valuable information that can prove useful when unknown attacking distortions affect the watermarked signal.

2. PERFORMANCE UNDER RANDOM ADDITIVE ATTACKS

We want to compare the impact of a simple random additive distortion on the performance of both approaches using P_e . We will restrict the analysis to the case where only one host signal sample x is used to hide one symbol bit $b = \pm 1$ of information. Even though this case is of little practical interest due to the high P_e associated, it does show the behavior that appears when more signal samples (higher dimensions) are used for the embedding; this more general case has been analyzed in [5] and confirms this assertion. Without loss of generality we will write the watermarked signal as $y = x + w$; after undergoing the additive random distortion independent from x we will have a received signal $z = y + n$.

We will term as D_w the embedding distortion, that is, the variance of the watermark, and as $D_c = \sigma_n^2$ the channel (attacking) distortion. As for the highest level of D_c permissible we will assume that it can only be as high as D_w . For convenience we define the square root ratio $\xi \triangleq \sqrt{D_w/D_c}$, and we call *watermark-to-noise ratio* to $\text{WNR} = 20 \log_{10} \xi$. The pointwise distortion criterion in this scenario is the same as the MSE one, but, as discussed in Sect. 1.1, this issue has to be carefully considered in multidimensional cases as those studied in [5].

The problem of determining the probability distribution function (pdf) $f_n(n)$ of the worst random distortion is a hard problem that largely depends on the data hiding method and host signal used. For this reason we choose to make the comparison when n is a Gaussian random variable with variance $D_c = \sigma_g^2$, that has always been used as a reference channel in previous analyses, and when n is uniform noise between $-\eta$ and η ($D_c = \eta^2/3$), that will prove especially harmful for quantization methods.

2.1. Known host-state methods

We will analyze QIM and DC-QIM implemented by means of dithered uniform quantizers, also called Dither Modulation (DM). The centroids of the quantizers $Q_{-1}(x)$ and $Q_1(x)$ are given by the lattices $\Lambda_{-1} = 2\Delta\mathbb{Z} + d$ and $\Lambda_1 = 2\Delta\mathbb{Z} + d + \Delta$, with d a possibly key-dependent arbitrary value that we assume without loss of generality to be zero for the analysis. Notice that the exact knowledge of the quantization centroids would permit a perfect attack (i.e. $P_e = 0.5$).

In QIM the watermarked signal turns to be the quantization centroid closest to x , i.e. $y = Q_b(x)$ for the symbol b ; the watermark is just the quantization error $w = e = Q_b(x) - x$. On the other hand DC-QIM takes as the watermark the

scaled quantization error, i.e. $w = v \cdot e = v(Q_b(x) - x)$, where the constant v is alleged to work in the same way as the optimizable constant α used in [2]. Thanks to the small size of Δ , e is assumed to be a uniform random variable in the intervals $[-\Delta, \Delta)$ and $[-v\Delta, v\Delta)$ respectively, therefore yielding $D_w = \Delta^2/3$ and $D_w = v^2\Delta^2/3$. We must remark that for DC-QIM (SCS) it is irrelevant that the hypotheses of Costa's result are only cast for Gaussian host signals, because uniformity is assumed to hold inside the quantization cells regardless of the host signal statistics. This implies that this implementation does not fully use Costa's conditions.

Moreover, we can use the uniform quantization error to generate the watermark following any arbitrary law, $w = T(e)$, not just $T(e) = v \cdot e$ as in DC-QIM. Notice that, as it happens with e , $T(e)$ is also orthogonal to x in the sense specified in [4]. This permits to improve DC-QIM for some channel distortions (as we will show in Sect. 3) with a method we propose next. Consider a DC-QIM variant that we name Gaussian DC-QIM (GDC-QIM) in which $y = x + T(e) = Q_b(x) + G(e)$, where $G(\cdot)$ is the mapping of a uniform variable into a zero-mean Gaussian variable with variance σ_v^2 . For this purpose, the quantized error e is mapped to $[0, 1)$ and fed into the inverse of the complementary cumulative Gaussian distribution function¹, as routinely made to generate an arbitrary pdf from an uniform one. As $w = e + G(e)$, the embedding distortion in this case is

$$D_w = \int_{-\Delta}^{\Delta} \left(e + \sigma_v Q^{-1} \left(\frac{e + \Delta}{2\Delta} \right) \right)^2 \frac{de}{2\Delta} = \frac{\Delta^2}{2} H \left(\frac{\sigma_v}{\Delta} \right) \quad (1)$$

where $H(\tau) \triangleq \int_{-1}^1 (e + \tau Q^{-1}((e+1)/2))^2 de$.

For all these methods decoding is simply performed by quantizing z (no statistical knowledge about x needed), that amounts to a minimum Euclidean distance decoder, i.e. $\hat{b} = \arg \min_{-1,1} \|z - Q_b(z)\|^2$. With this, it is straightforward to determine P_e since by symmetry and assuming that $b = -1$ is sent

$$P_e = P \{ \|z - Q_1(z)\|^2 < \|z - Q_{-1}(z)\|^2 \} \quad (2)$$

or, equivalently, we can write

$$P_e = \int_{\mathcal{R}_1} f_r(r + Q_{-1}(x)) dr, \quad (3)$$

with $r = w + n$ and \mathcal{R}_1 the decision region associated to $\hat{b} = 1$. Therefore $f_r(r)$ is the convolution of $f_n(n)$ with the pdf of the watermark. For QIM this simply means that $f_r(r) = f_n(r)$. On the other hand, for DC-QIM and GDC-QIM $f_r(r)$ is the convolution of $f_n(n)$ with a uniform and a normal distribution respectively.

¹ $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{\tau^2}{2}} d\tau$

Uniform noise. For QIM, as $f_r(r) = 1/2\eta$ for $|r| < \eta$, considering the decision regions it is possible to write that

$$P_e = \begin{cases} 0, & \xi \geq 2 \\ 1 - \frac{\xi}{2}, & 2/3 \leq \xi < 2 \end{cases} \quad (4)$$

where ξ takes the value Δ/η . For DC-QIM we have that, for $\eta \geq (1-v)\Delta$, the resulting pdf is

$$f_r(r) = \begin{cases} \frac{1}{2\eta}, & |r| \leq \eta - (1-v)\Delta \\ \frac{\eta + (1-v)\Delta - |r|}{4(1-v)\Delta\eta}, & \eta - (1-v)\Delta < |r| \leq \eta + (1-v)\Delta \end{cases} \quad (5)$$

Considering only the case $v \geq 1/2$, we have now that

$$P_e = \begin{cases} 0, & \xi \geq \frac{v}{v-1/2} \\ \frac{(v-(v-1/2)\xi)^2}{4v(1-v)\xi}, & \frac{v}{3/2-v} \leq \xi < \frac{v}{v-1/2} \\ \frac{5/2-2v}{3/2-v} - \frac{3/2-v}{v}\xi, & \xi < \frac{v}{3/2-v} \end{cases} \quad (6)$$

with ξ taking now the value $v \cdot \Delta/\eta$. The case $v < 1/2$ is less interesting because we will have in any case $P_e \neq 0$. Last, we consider GDC-QIM where

$$f_r(r) = \frac{1}{2\eta} \left(\mathcal{Q} \left(\frac{r-\eta}{\sigma_v} \right) - \mathcal{Q} \left(\frac{r+\eta}{\sigma_v} \right) \right). \quad (7)$$

Now $\xi = \sqrt{3D_w}/\eta$ and P_e is computed using (3) by means of numerical integration; this probability can be written as a function of σ_v/Δ and numerically optimized for this parameter.

Gaussian noise. In this case, it is possible to analytically determine the value of P_e for QIM, which results in

$$P_e = 2 \sum_{k=0}^{\infty} \left\{ \mathcal{Q} \left(\frac{(4k+1)}{2} \sqrt{3}\xi \right) - \mathcal{Q} \left(\frac{(4k+3)}{2} \sqrt{3}\xi \right) \right\}, \quad (8)$$

where $\xi = \Delta/(\sqrt{3}\sigma_g)$. For DC-QIM $f_r(r)$ is the same pdf as (7), but replacing η and σ_v by $(1-v)\Delta$ and σ_g respectively. In this case P_e is found through numerical integration, and $\xi = v\Delta/(\sqrt{3}\sigma_g)$. As for GDC-QIM, $f_r(r)$ is just a Gaussian with variance $\sigma_g^2 + \sigma_v^2$. A formula similar to (8) applies for P_e and the parameter ξ takes the value $\sqrt{D_w}/\sigma_g$. As in the uniform case, P_e can be numerically computed and optimized for the parameter σ_v/Δ .

2.2. Known host-statistics methods

As a typical spread-spectrum amplitude modulation method we choose to compute the watermark given b as $w = s \alpha \cdot b$, where s is a pseudorandom variable for which $E\{s^2\} = 1$ and $\alpha > 0$ a quantity chosen so the perceptual constraint is met. For this method, $D_w = \alpha^2$. We will just consider the case where x is modeled by a Laplacian pdf with variance σ_x^2 , which is adequate when the domain chosen is the

discrete cosine transform (DCT). Thus, the pdf of x has the form $f_x(x) = \beta/2 \exp(-\beta|x|)$, with $\beta = \sqrt{2}/\sigma_x$. For convenience we define $\lambda \triangleq \sqrt{\sigma_x^2/D_w}$ and call *document-to-watermark ratio* to $DWR = 20 \log_{10} \lambda$. The optimal ML decoder for this scheme decides $\hat{b} = +1$ if $|z - s \alpha| > |z + s \alpha|$ or, equivalently, if $zs > 0$. Assuming that s takes the values ± 1 with probability $1/2^2$, by symmetry the bit error probability will be

$$P_e = P\{z > 0 \mid b = 1\} = \int_{-\infty}^0 f_r(r - \alpha) dr \quad (9)$$

where $f_r(r) = f_n(r) * f_x(r)$. It is interesting to note that in the case when there is no distortion present in the channel, i.e. $n = 0$, we have that $f_r(r) = f_x(r)$. In this case $P_e = \exp(-\sqrt{2}/\lambda)/2$, with $\lambda = \sigma_x/\alpha$, and the method is *not provably robust*, i.e. $P_e > 0$.

Uniform noise. In this case we have that

$$f_r(r) = \begin{cases} (1 - e^{-\beta\eta} \cosh(\beta r))/2\eta, & |r| < \eta \\ e^{-\beta|r|} \sinh(\beta\eta)/2\eta, & |r| \geq \eta \end{cases} \quad (10)$$

It is straightforward now the calculation of P_e using (9)

$$P_e = \begin{cases} \lambda\xi \frac{e^{-\sqrt{2}/\lambda}}{2\sqrt{6}} \sinh\left(\frac{\sqrt{6}}{\lambda\xi}\right), & \xi > \sqrt{3} \\ \frac{1}{2} - \frac{\xi}{2\sqrt{3}} + \lambda\xi \frac{e^{-\sqrt{6}/(\lambda\xi)}}{2\sqrt{6}} \sinh(\sqrt{2}/\lambda), & \xi \leq \sqrt{3} \end{cases} \quad (11)$$

with $\xi = \sqrt{3}\alpha/\eta$.

Gaussian noise. The desired pdf is now

$$f_r(r) = \frac{\beta}{2} e^{\frac{\beta^2 \sigma_g^2}{2}} \left\{ e^{-\beta r} \mathcal{Q} \left(\frac{-r + \beta \sigma_g^2}{\sigma_g} \right) + e^{\beta r} \mathcal{Q} \left(\frac{r + \beta \sigma_g^2}{\sigma_g} \right) \right\} \quad (12)$$

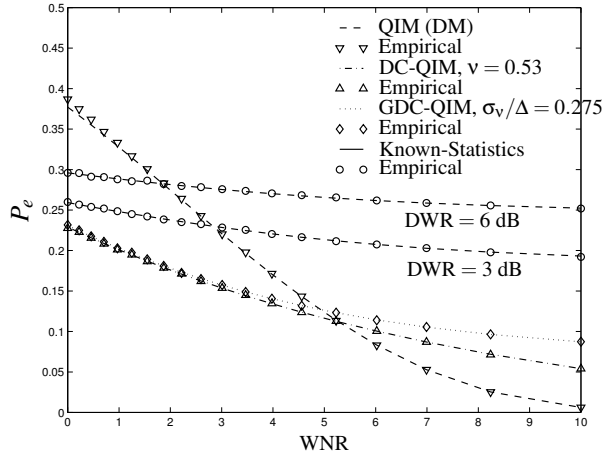
For this case $\xi = \alpha/\sigma_g$, and numerical integration is required for computing P_e inserting (12) into (9).

3. COMPARISONS AND CONCLUSIONS

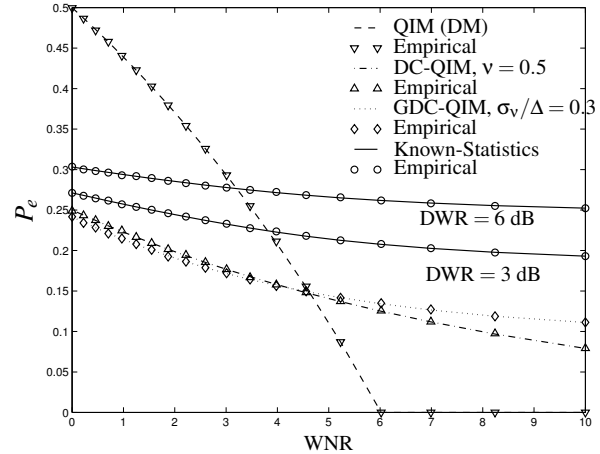
In Fig. 1 we present P_e for the analyzed methods under the proposed distortions, using a decreasing distortion level. In both cases we see the typical performance of QIM, that degrades rapidly when increasing the distortion. Note in Fig. 1(b) that, as uniform noise is amplitude bounded, we can get $P_e = 0$ for a WNR > 6 dB which provides a certain degree of *provable robustness*. Nevertheless when WNR = 0 dB, $P_e = 0.5$ and so the channel becomes useless.

The plots of DC-QIM and GDC-QIM have been minimized on v and σ_v/Δ for the highest distortion (WNR = 0 dB). It is still possible for DC-QIM to have $P_e = 0$ if $v > 1/2$ with uniform noise, but at too high WNR values for the plot

²This choice minimizes the bit error probability while meeting the constraint $E\{s^2\} = 1$. The proof of this fact follows from the convexity of the integral of the tail of $f_r(r)$.



(a) Gaussian noise



(b) Uniform noise

Figure 1: Performance Comparison

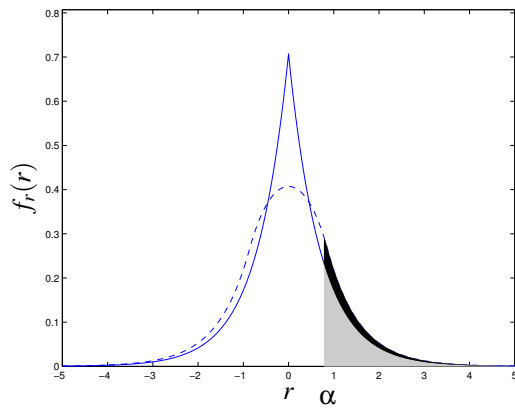


Figure 2: Tail integral variation when uniform noise corrupts a Laplacian pdf (DWR = 6 dB, WNR = 4 dB)

shown. Notice that, even when GDC-QIM cannot defeat DC-QIM under Gaussian distortion, it does improve DC-QIM in the uniform case under 4 dB. This is a simple proof that DC-QIM is not optimal in the sense of measuring P_e . Note also that the proposed GDC-QIM is just an illustration of the fact that DC-QIM can be improved; the search for better alternatives is open. For instance, even better results could be achieved by taking the transformation $G(\cdot)$ to be a truncated Gaussian pdf instead of a true Gaussian. Therefore, notice that the possibility of more successful schemes employing side information in data hiding is still open.

It is very interesting to see that known host-statistics methods are DWR-dependent, while known host-state ones are almost insensitive to this parameter. Lower values of DWR are exploited by the statistical detection to achieve lower P_e values in all the WNR range. Also note that, even when $P_e \neq 0$ for any value of WNR, the degradation with growing distortions is much more graceful than the other

methods: the reason is that the tail integral of the overall pdf that determines P_e is only slightly modified (see Fig. 2). This means that, if values of WNR under 0 dB were allowed as done in [4] and others, the known-statistics method for DWR = 6 dB in our example would defeat DC-QIM for both Gaussian and uniform noise under WNR ≈ -2 dB. Last, further improvements are expectable for Generalized Gaussian models with shape parameters lower than one.

4. REFERENCES

- [1] Ingemar J. Cox, Matthew L. Miller, and Andrew L. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1127–1141, July 1999.
- [2] Max H.M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.
- [3] Brian Chen and Gregory W. Wornell, "Preprocessed and postprocessed quantization index modulation methods for digital watermarking," in *Proc. of SPIE*, San José, USA, January 2000, Security and Watermarking of Multimedia Contents, pp. 48–59.
- [4] Joachim J. Eggers and Bernd Girod, "Quantization watermarking," in *Proc. of SPIE*, San José, USA, January 2000, Security and Watermarking of Multimedia Contents II.
- [5] Fernando Pérez-González, Félix Balado, and Juan R. Hernández, "Performance analysis of existing and new methods for data hiding with known host information in additive channels," *IEEE Trans. on Signal Processing*, 2001, Submitted.