Worst Case Additive Attack against Quantization-Based Watermarking Techniques

J.E. Vila-Forcén, S. Voloshynovskiy, O. Koval, T. Pun
University of Geneva, Department of Computer Science. 24 rue Général-Dufour, CH 1211, Geneva, Switzerland
F. Pérez-González
University of Vigo, Signal Theory and Communications Department. E-36200 Vigo, Spain
Contact email: svolos@cui.unige.ch

***Abstract*— In the scope of quantization-based watermarking techniques and additive attacks, there exists a common belief that the worst case attack (WCA) is given by additive white Gaussian noise (AWGN). Nevertheless, it has not been proved that the AWGN is indeed the WCA within the class of additive attacks against quantization-based watermarking. In this paper, the analysis of the WCA is theoretically developed with probability of error as a cost function. The adopted approach includes the possibility of masking the attack by a target probability density function (*PDF*) in order to trick smart decoding. The developed attack upper bounds the probability of error for quantization-based embedding schemes within the class of additive attacks.**

I. Introduction

THE design of the worst case attack is an active line of research in communications. The knowledge of the WCA points out the weakness or the strength of a given technique and allows to create a fair benchmark for different communications methods.

The study of the WCA within the scope of digital data-hiding technologies has even higher importance in many practical applications. Currently, the benchmarking of most data-hiding techniques is performed against the additive white Gaussian noise attack [1], [2]. However, the real attacker is an aggressive character whose primary goal is to destroy or to prevent the reliable decoding of hidden information. This goal is achieved by applying the least favorable conditions against particular data-hiding systems.

Considering different data-hiding benchmarking methods, the two most widely used are presented. The first one is based on estimation of the information-theoretic limits, i.e., the capacity. According to the second method, the practical efficiency of the embedding strategies is analyzed with respect to the bit error rate.

The common belief about superiority of quantization-based methods is established based on their testing in the AWGN scenario. Motivated by the fact, first shown in [3] that the AWGN is not the WCA against the quantization-based techniques, one can conclude that the WCA against these methods is still an open problem.

Thus, we formulate the goal of this paper as the investigation of the worst additive attack strategy for a fixed quantization-based watermarking communications scenario which is depicted in Fig. 1. In this set-up, the message b should be communicated via the host channel X and the attacking channel $p(y|y')$. The message b is encoded into a watermark W taking into account the host state X. We assume here that both the encoder and the decoder are fixed while the attacker can apply any additive attack changing the transition probability $p(y|y')$.

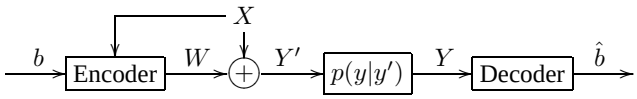


Fig. 1. Communications scenario.

To underline the importance of this problem, we would like to point out that its solution will allow us to justify a valid benchmarking set-up instead of applying commonly accepted AWGN as the WCA.

This paper is organized as follows. The fundamentals of the quantization-based techniques under analysis are described in Section II. The problem formulation is presented in Section III. The justification of the proposed set-up and constraints is explained in Section IV. Finally, in Section V the probability of error is analyzed for the quantization-based techniques according to the proposed WCA and compared with the AWGN and uniform noise attacks.

II. Quantization-based techniques

Within the class of quantization-based techniques, we will concentrate our analysis on both dither modulation (DM) [1] and distortion compensated DM (DC-DM) [2]. Both are approximations of Costa [4] scheme with a structured codebook where binning strategy is exploited.

Dither modulation: The watermarked data is obtained as:

$$y' = Q_{b_i}(x), \tag{1}$$

where different quantizers $Q_{b_i}(\cdot)$ are used to embed a symbol b_i into the host x as it is shown in Fig. 2 for the binary case. The quantizers will be constructed for both DM and DC-DM using subtractive dithering [5] that corresponds in the binary case to:

$$Q_{b_i}(x) = Q(x + d_{b_i}) - d_{b_i}, \tag{2}$$

where $d_0 = -\Delta/2$ and $d_1 = \Delta/2$ assuming that the quantization bin width is 2Δ. As a result of the quantization

embedding, the watermark distortion can be expressed as:

$$D_W = E\left[|Y' - X|^2\right] = \frac{\Delta^2}{3}, \tag{3}$$

where $E[\cdot]$ denotes the expected value. The output *PDF* $f_{Y'}(\cdot)$ is a quantized version of the host *PDF* $f_X(\cdot)$.

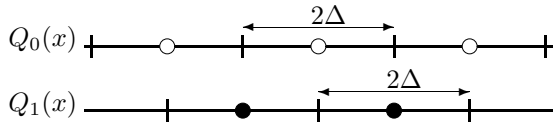


Fig. 2. DM embedding quantizers, binary case.

Distortion-compensation DM: The DC-DM technique produces the watermarked data as:

$$y' = x + \alpha(Q_{b_i}(x) - x), \tag{4}$$

where α is the distortion compensation parameter $0.5 \leq \alpha \leq 1$. If $\alpha = 1$, the DC-DM simplifies to the DM. For $\alpha < 0.5$, the embedding scheme is not errorless even in a noise free scenario: thus, to avoid degradation of the performance, we constrain $\alpha \geq 0.5$. For different values of α, the watermarked sample will have intermediate values between the original x and its quantized version $Q_{b_i}(\cdot)$. The watermark distortion for the DC-DM embedding is:

$$D_W = \alpha^2 \frac{\Delta^2}{3}, \tag{5}$$

where it is evident that if $\alpha < 1$ the value of the distortion of the DM (3) is larger than the one of DC-DM (5). This provides an opportunity to increase Δ and subsequently to improve on the performance for the same D_W.

The output *PDF* $f_{Y'}(\cdot)$ of the DC-DM technique is a train of uniform pulses instead of a train of δ functions due to the distortion compensation, as it is presented in Fig. 3, where $K = (1-\alpha)\Delta$.

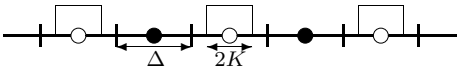


Fig. 3. DC-DM output *PDF* $f_{Y'}(\cdot)$.

In both the DM and DC-DM, the probability of error can be calculated as the integral of the equivalent noise *PDF* over the support of the *wrong* bins $\mathcal{S}$ with respect to the communicated message bit [3]:

$$P_e = \int_{\mathcal{S}} f_Y(x)dx. \tag{6}$$

The error region $\mathcal{S}$ is illustrated in Fig. 4 assuming that one of the symbols is sent for the binary signaling case and hard decision is used at the decoder. The errorless region will be denoted as the complementary set $\bar{\mathcal{S}}$. In case of the AWGN, equation (6) is equal to the integral over the hatched area on Fig. 5.

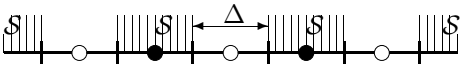


Fig. 4. DM and DC-DM error region $\mathcal{S}$ for the binary case.

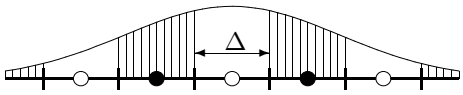


Fig. 5. Received *PDF* $f_Y(\cdot)$ in the case of AWGN attack for one bin. The error region is hatched.

III. Problem formulation

McKellips and Verdu considered the problem of worst case additive attack design for the PAM, having chosen the probability of error P_e as a cost function [6]. In their approach, the attack has bounded energy. Besides the energy, they also considered a possibility of masking the noise *PDF* by a target distribution $f_T(\cdot)$ with variance σ_T^2 in order to trick a *smart decoding* strategy when the decoder can estimate the attack *PDF* to choose an appropriate decoding strategy. The Kullback-Leibler distance (KLD) [7] between the target distribution and the optimized WCA *PDF* $f_Z(\cdot)$ is proposed as an additional optimization constraint.

Adopting their approach to our quantization-based set-up, we impose constraints on the attack power, target *PDF* and tolerance in the KLD sense. The objective of optimization is to find the *PDF* of the attack which satisfies the above constraints while producing the largest possible value of the cost function, i.e., the probability of error. This *PDF* of the attack will constitute the least favorable distribution for the decoder for the given class of attacking strategies.

The optimization problem can be formulated as follows:

$$\max_{f_Z(x)} P_e = \max_{f_Z(x)} \int_{\mathcal{S}} f_Y(x)dx \tag{7}$$

subject to the constraints:

$$\int_{-\infty}^{\infty} f_Z(x)dx = 1, \tag{8}$$

$$\int_{-\infty}^{\infty} x^2 f_Z(x)dx \leq \sigma_Z^2, \tag{9}$$

$$\int_{-\infty}^{\infty} f_Z(x) \log \frac{f_Z(x)}{f_T(x)}dx \leq \beta. \tag{10}$$

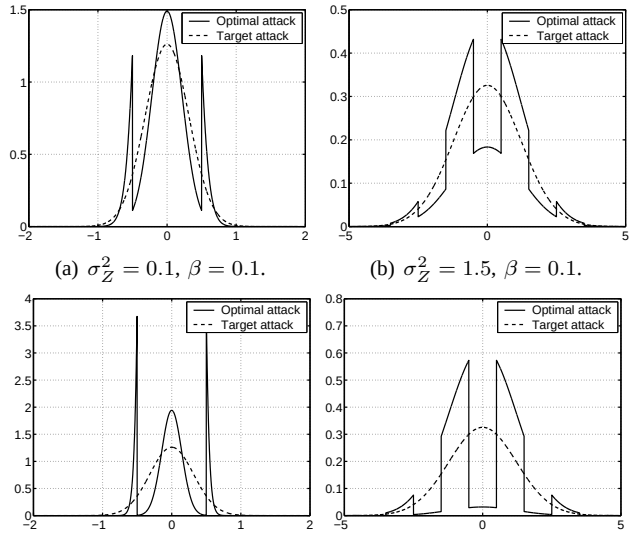
$f_Y(\cdot)$ is the *PDF* of the received signal $Y = X + W + Z$. X represents the host data, W stands for the quantization-based watermark [1], [2] and Z denotes the attack. σ_Z^2 in (9) constrains the attack variance assuming $\mu_Z = \int_{-\infty}^{\infty} x f_Z(x)dx = 0$, and β in (10) determines the tolerance in the KLD sense. The constraint (8) follows from *PDF* definition, which also imposes $f_Z(x) \geq 0, \forall x \in \mathcal{R}$.

The optimization is performed using the Lagrangian multiplier method:

$$\begin{aligned} L(f_Z(x), \lambda_1, \lambda_2, \lambda_3) &= \int_{\mathcal{S}} f_Y(x)dx \\ &+\lambda_1 \left(\int_{-\infty}^{\infty} f_Z(x)dx - 1 \right) + \lambda_2 \left(\int_{-\infty}^{\infty} x^2 f_Z(x)dx - \sigma_Z^2 \right) \\ &+\lambda_3 \left(\int_{-\infty}^{\infty} f_Z(x) \log \frac{f_Z(x)}{f_T(x)}dx - \beta \right). \end{aligned} \tag{11}$$

IV. Problem solution

In general, no close analytical solution of (7) exists and numerical computation is required. For different values of the tolerance β the resulting WCA *PDF* varies from the target *PDF* to a set of δ-functions. Fig. 6 represents a Gaussian target *PDF* with variance $\sigma_T^2 = \sigma_Z^2$ and the WCA *PDF* against the DM [1] where the bin width is equal to 1. It is assumed that the embedded symbol is located at the origin. Fig. 6(a) and 6(b) show the WCA when the tolerance factor β with the target attack is small for different attack powers. The WCA is masked by the target *PDF* (Gaussian). Fig. 6(c) and 6(d) represent the same analysis for larger tolerance factors, where the *PDF* shape of the result varies to increase the probability of error for the same attack energy.



(a) $\sigma_Z^2 = 0.1$, $\beta = 0.1$. (b) $\sigma_Z^2 = 1.5$, $\beta = 0.1$.

(c) $\sigma_Z^2 = 0.1$, $\beta = 0.5$. (d) $\sigma_Z^2 = 1.5$, $\beta = 0.5$.

Fig. 6. Attack examples for different tolerance and power attack values. WCA *PDF* and Gaussian target *PDF*.

The solution of the above problem (11) when $\beta \to \infty$ suggests three-δ functions. Two of them are located in the two error regions $\mathcal{S}$ that are the closest to the origin, and the third one is located at the origin to contribute to the *PDF* but not to the the attack power. However, a hypothetical *smart decoder* can apply more sophisticated decoding strategies than the originally proposed minimum distance decoder [1], [2]. This motivates the attacker to mask the applied attack to avoid the possibility to be detected while degrading communications performance as much as possible.

Such an attack provides a probability of error equal $P_e = 1$ which is easily reversed to $P_e = 0$ inverting the bits value. In fact, from the mutual information point of view there is no uncertainty. Therefore we will restrict the probability of error to $0 \leq P_e \leq 0.5$.

The 3-δ attack *PDF* is presented in Fig. 7. For both DM and DC-DM, it is neccesary to optimize the parameters T and V in order to maximize the probability of error as $\delta \to \infty$. The amplitude of the δ-functions V and $1 - 2V$ guarantees that the integral of the *PDF* is always 1 for $0 \leq V \leq 0.5$. In practice, we will always constrain $V < 0.25$ in order to have a maximum probability of error equals to 0.5.

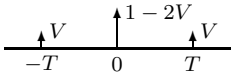


Fig. 7. 3-δ attack scheme, $0 \leq V \leq 0.5$.

The energy of the proposed attack is given by ($\mu_Z = 0$):

$$D_Z = \sigma_Z^2 = \int_{-\infty}^{\infty} x^2 f_Z(x) dx = 2VT^2. \tag{12}$$

Dither modulation: The convolution between the output *PDF* $f_{Y'}(\cdot)$ of the DM technique and the proposed attack (Fig. 7) is a set of delta functions. The optimal point T where to fix the position of the deltas T is $T = \frac{\Delta}{2} + \epsilon$ where $\epsilon > 0, \epsilon \to 0$. In this case, the probability of error which is equal to the integral over the error region $\mathcal{S}$ is $P_e = 2V$. Using equations (3) and (12), it is possible to write:

$$\text{WNR} = 10 \log_{10} \frac{D_W}{D_Z} = 10 \log_{10} \frac{4}{3P_e}, \tag{13}$$

where WNR denotes the watermark-to-noise ratio. The above equation (13) produces:

$$P_e = \frac{4}{3} 10^{-\frac{\text{WNR}}{10}}. \tag{14}$$

Distortion-compensation DM: The derivation of the WCA when $\delta \to \infty$ for the DC-DM is similar to the DM but now the received signal *PDF* $f_Y(\cdot)$ is the convolution of a set of delta functions with a pulse, which is the result of the DC-DM embedding scheme. The convolution is presented in Fig. 8 for one of the periodical bins on Fig. 4 where:

$$V_1 = \frac{1 - 2V}{2K}, \qquad V_2 = \frac{V}{2K}, \qquad T_2 = T + K, \tag{15}$$

and the nearest error bin of the error region $\mathcal{S}$ represented in Fig. 4 initiates at M.

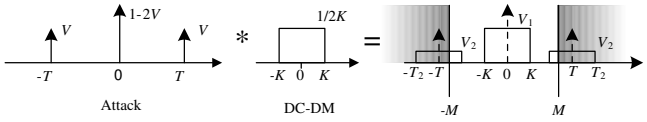


Fig. 8. Convolution of 3-δ attack with DC-DM output *PDF*.

The probability of error of the 3-δ attack for DC-DM is:

$$P_e = \frac{V}{K} (T + K - M). \tag{16}$$

The maximization of equation (16), using (5), (12), $K = (1-\alpha)\Delta$ and $M = \Delta/2$ from DC-DM embedding (4), Fig. 3, is achieved by:

$$T = \begin{cases} (2\alpha - 1)\Delta, & \alpha < \frac{5}{6}, \\ (\frac{3}{2} - \alpha)\Delta, & \alpha \geq \frac{5}{6}. \end{cases} \tag{17}$$

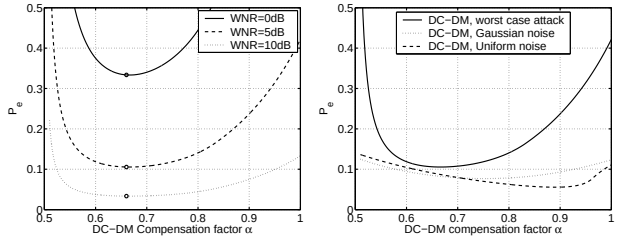
The result from equation (17), using (5) and (12) simplifies (16) to:

$$P_e = \begin{cases} \frac{\sigma_Z^2}{4(1-\alpha)(2\alpha-1)\Delta^2} & \alpha < \frac{5}{6}, \\ \frac{\sigma_Z^2}{\left((\frac{3}{2}-\alpha)\Delta\right)^2} & \alpha \geq \frac{5}{6}, \end{cases} \tag{18}$$

and the WNR $= 10 \log_{10} \frac{\alpha^2 \Delta^2}{3\sigma_Z^2}$.

V. Results of computer modeling

In the previous section, the analysis of the probability of error has been performed. Fig. 9 presents the results of the above asymptotic case ($\beta \to \infty$), for which the WCA *PDF* is composed of a set of three δ-functions. In Fig. 9(a), the probability of error is shown as a function of the DC-DM distortion compensation factor α for different WNR. To prove the efficiency of the proposed attack, we compare the probability of error for the developed attack vs. Gaussian and uniform attacks (Fig. 9(b)). The DM is shown as a special case of the DC-DM if $\alpha = 1$. Therefore, it becomes evident from Fig. 9 that the proposed attack produces larger probability of error than the AWGN and uniform noise attacks.

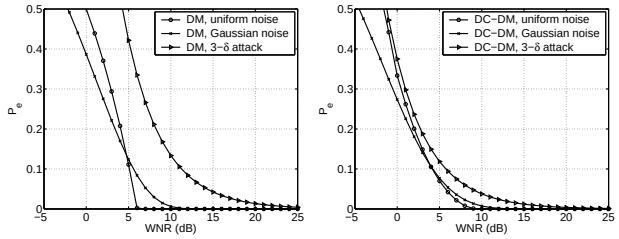


(a) WCA for 3 WNR values 0dB, 5dB and 10dB

(b) Worst case, Gaussian and uniform noise attacks for WNR= $5dB$

Fig. 9. Performance comparison for $\beta \to \infty$.

The choice of a Gaussian is due to common belief that it was the WCA. In our simulation AWGN is the target attack $f_T(\cdot)$. In [3], Perez-Gonzalez *et al* demonstrate that uniform noise attack is worse than the Gaussian for some WNR and DC-DM distortion compensation factor. The results on Fig. 10 show that our developed 3-δ attack is always worse than the Gaussian and uniform noise attacks, for all WNR and DC-DM compensation factors.



(a) DM benchmarking.

(b) DC-DM benchmarking, $\alpha = 0.75$.

Fig. 10. WCA benchmarking.

A very important conclusion from the previous results is the existence of a compensation parameter that can be chosen in advance at the encoder to guarantee a given system performance disregarding the attack *PDF* for a given WNR. The optimal DC-DM compensation parameter α can be chosen equal $\alpha = 2/3$ in order to guarantee that the probability of error is bounded, and its maximum value does not exceed the P_e of the WCA. In previous schemes [2], [4] the optimal compensation parameter α was obtained when both the *PDF* and the power of the noise were known at the encoder. Nevertheless, a mismatch between their prediction and the real noise can considerably degrade the system performance.

Here, we demonstrate that $\alpha = 2/3$ is the optimal DC-DM compensation parameter for all WNR since it is the minimum from (14) and (18) as it is presented in Fig. 9(a). Therefore, a blind encoder can use this value without any knowledge of the additive attack *PDF*. The maximum probability of error is determined only by the WNR.

VI. Conclusions

In this paper, the worst case energy constrained additive attack against quantization-based watermarking techniques was developed. Depending on the tolerance factor, it is possible to hide the statistical properties of the attack preventing the *smart decoding*. The closed form analytical solution for the problem in the asymptotic case ($\beta \to \infty$) was obtained. The superiority of the designed attack over the classically used AWGN and uniform noise attacks was shown. Finally, we want to emphasize that an optimal compensation factor α for the DC-DM method exists, which independently on the WNR gives the optimal system performance undergoing the developed attack.

VII. Acknowledgment

This paper was partially supported by SNF Professorship grant No PP002-68653/1, Interactive Multimodal Information Management (IM2) project and by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT. The authors are thankful to the members of Stochastic Image Processing group at University of Geneva for many helpful and interesting discussions. The information in this document reflects only the author's views, is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

References

[1] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory.*, vol. 47, pp. 1423–1443, May 2001.

[2] J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, "Scalar costa scheme for information embedding," *IEEE Transactions on Signal Processing*, vol. vol. 51, no. no. 4, pp. 1003–1019, April 2003.

[3] F. Perez-González, F. Balado, and J. R. Hernández, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Trans. on Signal Processing, Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery*, vol. 51, no. 4, April 2003.

[4] M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[5] B. Chen and G. Wornell, "Dither modulation: a new approach to digital watermarking and information embedding," in *IS&T/SPIE's 11th Annual Symposium, Electronic Imaging 1999: Security and Watermarking of Multimedia Content I SPIE.*, vol. 3657, San Jose, California, USA, January 1999, pp. 342–353.

[6] A. Mckellips and S. Verdu, "Worst case additive noise for binary-input channels and zero-threshold detection under constraints of power and divergence," *IEEE Transactions on Information Theory*, vol. 43, no. 4, pp. 1256–1264, July 1997.

[7] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley and Sons, New York, 1991.