Performance Analysis of ST-DM Watermarking in Presence of Nonadditive Attacks

Franco Bartolini, *Member, IEEE*, Mauro Barni, *Member, IEEE*, and Alessandro Piva

***Abstract*—The performance of spread-transform dither modulation (ST-DM) watermarking in the presence of two important classes of non additive attacks, such as the gain attack plus noise addition, and the quantization attack are evaluated. The analysis is developed under the assumption that the host features are independent and identically distributed Gaussian random variables, and that a minimum distance criterion is used to decode the hidden information. The theoretical bit-error probabilities are derived in closed form, thus permitting to evaluate the impact of the considered attacks on the watermark at a theoretical level. The analysis is validated by means of extensive Monte Carlo simulations. In addition to the validation of the theoretical analysis, Monte Carlo simulations permitted to abandon the hypothesis of normally distributed host features, in favor of more realistic models adopting a Laplacian or a generalized Gaussian probability density function. The general result of our analysis is that the excellent performance of ST-DM are confirmed in all cases with the only noticeable exception of the gain attack.**

***Index Terms*—Gain attack, nonadditive watermarking channel, quantization attack, ST-DM watermarking.**

I. Introduction

SPURRED by the theoretical analysis of the watermarking problem, quantization index modulation (QIM) watermarking [1] has rapidly become popular as one of the best performing blind watermarking strategies among those developed since watermarking was brought to the attention of signal processing researchers almost ten years ago. The main reason behind the good performance of QIM is that it represents a simple, yet powerful, way to apply the informed embedding principle first established by Cox *et al.* [2] and further developed in the works by Chen and Wornell [1], [3], Moulin [4], and Cohen and Lapidoth [5], echoing previous information theoretic results by Costa [6] and Gelf and Pinsker [7], in which the so-called channel with known state, or side information, at the encoder is studied.[1]

Among the wide class of QIM watermarking techniques, a prominent position is occupied by the spread-transform dither modulation (ST-DM) algorithm [1], which couples the effectiveness of QIM schemes and conventional spread-spectrum systems. According to the ST-DM scheme, each bit is embedded within the host signal by quantizing the correlation between the host feature sequence and a reference spreading sequence. Quantization is performed according to one out of two quantizers depending on the sign of the to-be-hidden bit (we assume that, before embedding, the information sequence is mapped into an antipodal sequence). At a practical level, the performance of the ST-DM algorithm, together with those of a large class of QIM methods, have been carefully analyzed by Gonzalez *et al.* [11] by assuming that the watermark is impaired by an additive attacker. Though the analysis carried out by Gonzalez *et al.* confirms the outstanding properties of ST-DM (and some other algorithms strongly related to it, such as distortion-compensated ST-DM, quantization projection, and distortion-compensated quantization projection), a deeper analysis is required to assess the performance of ST-DM in practical scenarios where a much wider variety of attacks have to be considered.

In this work, we analyze the performance of ST-DM in the presence of more realistic attacks, namely the gain attack (multiplication by an unknown scale factor) plus noise addition, and the quantization attack. These are very important attacks, since they model some of the most common manipulations the watermarked data usually undergo. More specifically, the gain attack properly models linear filtering and, to a lesser extent, histogram equalization or loudness changes, and the quantization attack gives a good indication of the performance of ST-DM in the presence of lossy compression such as JPEG coding for the case of still images.

The theoretical analysis is validated, and somewhat extended, through Monte Carlo simulations, allowing to evaluate the performance of ST-DM when the host features do not follow a Gaussian distribution. The figure of merit used to measure the effectiveness of ST-DM is the bit-error rate, since we deem that this better reflects the way watermarking is used in practical applications. To make the comparison with other algorithms possible, the analysis is carried out by clearly defining the operating conditions in terms of payload, data-to-watermark ratio (DWR), and watermark-to-noise ratio (WNR).

This paper is organized as follows. In Section II, the theoretical framework used throughout the paper is described, with a brief introduction to ST-DM watermarking and the definition of the figures of merit used for the analysis. In Section III, the bit-error probability in presence of a gain attack plus noise addi-

Manuscript received October 31, 2003; revised January 17, 2004. This work was supported by the Italian Ministry for Education, University and Research (MIUR) under Grant 2003094459 (WAVECOP project). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Reginald L. Lagendijk.

F. Bartolini (deceased) was with the Department of Electronics and Telecommunications, University of Florence, 50139-Firenze, Italy.

M. Barni is with the Department of Information Engineering, University of Siena, 53100-Siena, Italy (e-mail: barni@dii.unisi.it).

A. Piva is with the National Inter-university Consortium for Telecommunications (CNIT), University of Florence, 50139-Firenze, Italy (e-mail: piva@lci.det.unifi.it).

Digital Object Identifier 10.1109/TSP.2004.833868

[1] It has to be noted that QIM does not represent the only way to turn Costa's ideas into practice (see, for example, [8]–[10])

tion is derived theoretically, under the assumption of Gaussian noise. The analysis is validated and extended to the case of non-Gaussian host features through Monte Carlo simulations. Section IV is devoted to the theoretical analysis of the bit-error rate when a quantization attack is present. Finally, in Section V, some conclusions are drawn.

II. Definition of the Theoretical Framework

The ST-DM algorithm belongs to the wider class of QIM watermarking algorithms. According to the QIM approach, watermarking is achieved through the quantization of the host feature vector on the basis of a set of predefined quantizers, where the particular quantizer used by the embedder depends on the to-be-hidden message $\mathbf{b}$. Stated in another way, the to-be-hidden message modulates the quantizer index, hence justifying the QIM appellative. The simplest way to design a QIM watermarking system consists in associating each bit of $\mathbf{b}$, say b_i, to a single host feature f_i and let b_i determine which quantizer, chosen between two uniform scalar quantizers, is used to quantize f_i. To be specific, the two codebooks $\mathcal{U}_0$ and $\mathcal{U}_1$ associated respectively to $b = 0$ and $b = 1$ are defined as

$$\mathcal{U}_0 = \{k\Delta + d, k \in \mathbb{Z}\} \tag{1}$$

$$\mathcal{U}_1 = \left\{k\Delta + \frac{\Delta}{2} + d, k \in \mathbb{Z}\right\} \tag{2}$$

where d is an arbitrary parameter, possibly depending on a secret key to improve security. In the following, we will assume $d = \Delta/4$, since in this way a lower distortion is obtained (see below). Watermark embedding is achieved by applying either the quantizer $\mathcal{Q}_0$ associated to $\mathcal{U}_0$

$$\mathcal{Q}_0(f) = \arg \min_{u_{0,i} \in \mathcal{U}_0} |u_{0,i} - f| \tag{3}$$

where f is the feature hosting b and $u_{0,i}$ are the elements of $\mathcal{U}_0$, or the quantizer associated to $b = 1$

$$\mathcal{Q}_1(f) = \arg \min_{u_{1,i} \in \mathcal{U}_1} |u_{1,i} - f|. \tag{4}$$

By letting f_w indicate the marked feature, we have

$$f_w = \begin{cases} \mathcal{Q}_0(f), & b = 0 \\ \mathcal{Q}_1(f), & b = 1. \end{cases} \tag{5}$$

Of course, embedding each bit into a single feature yields a very fragile watermark; hence, it is customary to let each bit be hosted by r features $\mathbf{f} = (f_1, f_2 \ldots f_r)$, e.g., by repeatedly inserting b into $f_1 \ldots f_r$ by means of (5) (DM-with-bit-repetition). The ST-DM algorithm permits to better exploit the availability of r host features to host a single bit b. According to the ST-DM approach, the correlation between the host feature vector $\mathbf{f}$ and a reference spreading signal $\mathbf{s}$ is quantized instead of the features themselves. In a more precise way, let us assume that $\mathbf{s}$ is a unit-norm binary pseudo-random sequence taking values $\pm 1/\sqrt{r}$ (this choice guarantees that watermark distortion is spread uniformly over all the features[2]). The embedder calculates the correlation between $\mathbf{f}$ and $\mathbf{s}$, as follows:

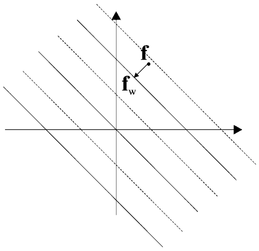


Fig. 1. Geometrical representation of ST-DM watermarking. Points on solid lines form the $\mathcal{U}_0$ codebook, whereas dashed lines correspond to $\mathcal{U}_1$.

$$\rho_f = \mathbf{f} \cdot \mathbf{s} = \sum_{i=1}^{r} f_i s_i \tag{6}$$

and then it subtracts the projection of $\mathbf{f}$ on $\mathbf{s}$ from $\mathbf{f}$ and adds a new vector component along the direction of $\mathbf{s}$ resulting in the desired quantized autocorrelation, say ρ_w

$$\mathbf{f}_w = \mathbf{f} - \rho_f \mathbf{s} + \rho_w \mathbf{s} \tag{7}$$

where ρ_w is calculated by applying (5) to ρ_f. With regard to the quantization step, let us remember that the components of $\mathbf{s}$ take only values $\pm 1/\sqrt{r}$. If ρ_f is quantized with step Δ, then the maximum distortion along each feature component is $\Delta 2/\sqrt{r}$. A geometric interpretation of ST-DM watermarking for $r = 2$ is given in Fig. 1. Solid lines represent quantized values corresponding to $b = 0$, whereas dashed lines are relative to $b = 1$. For any host feature vector $\mathbf{f}$, embedding $b = 0$ is obtained by projecting $\mathbf{f}$ over the closest solid line.

Further improvements to the basic ST-DM algorithms may be obtained either by introducing distortion-compensated (DC-STDM) [1] or by considering the enlarged class of distortion-compensation quantization projection (DC-QP) schemes [11]. However, the improvement with respect to plain ST-DM is marginal [11] and at the expenses of requiring some additional knowledge about the attack power, in that the variance of the additive noise attack brought to the watermarked signal must be known. For this reason, we did not to include DC-STDM and DC-QP in our analysis.

As to decoding, a minimum distance decoder is adopted:

$$b^* = \arg \min_{b \in \{0,1\}} \min_{u_{b,i} \in \mathcal{U}_b} |u_{b,i} - \rho'| \tag{8}$$

where by ρ', we indicate the correlation between the watermarked and possibly attacked features and the spreading vector $\mathbf{s}$. Note that that this is the optimum decoding strategy only if adjacent entries of the codebooks can be assumed to be equiprobable, which is not the case for large values of r.

[2] Note that in this way, we are neglecting perceptual considerations that could require distortion not to be uniformly distributed among different perceptual components.

In order to compare systems based on different embedding/recovery rules and operating in different host domains, a general theoretical framework and a set of common, objective, parameters must be defined. Such parameters will have to measure detection reliability, watermark obtrusiveness and attack strength. In all the cases we are interested in average measures, where the average is taken over the host feature sequence, i.e., we fix the spreading sequence $\mathbf{s}$ and the to-be-hidden message $\mathbf{b}$ and we average across the feature sequence $\mathbf{f}$.

With regard to the host features, unless we specify otherwise, we will model them as a sequence of independent and identically distributed random variables following a zero-mean Gaussian probability density function (pdf). The reliability of the watermark is measured by means of the actual bit-error rate, and its obtrusiveness through the document-to-watermark ratio, expressing the ratio between the power of the host features and that of the watermark. To be specific, by letting

$$\mathbf{w} = \mathbf{f}_w - \mathbf{f} \tag{9}$$

be the embedded watermark, and by exploiting the stationarity of $\mathbf{f}$, we have

$$\text{DWR} = \frac{\sum_i E[f_i^2]}{\sum_i E[w_i^2]} = \frac{E[f^2]}{E[w^2]} \tag{10}$$

where $E[f^2]$ and $E[w^2]$ indicate, respectively, the power of the host features and that of the watermark. Note that dependence on i may be neglected due to the stationarity of $\mathbf{f}$ and to the independence of $E\left[w_i^2\right]$ on $\mathbf{b}$ and $\mathbf{s}$ (see Section II-A).

The strength of the attacks may be measured by the WNR, giving the ratio between the power of $\mathbf{w}$ and that of the noise introduced by attacks. More specifically, by indicating with $\mathbf{f}'_w$ the attacked host signal, we let

$$\boldsymbol{\nu} = \mathbf{f}'_w - \mathbf{f}_w \tag{11}$$

and

$$\text{WNR} = \frac{\sum_i E\left[w_i^2\right]}{\sum_i E\left[\nu_i^2\right]}. \tag{12}$$

We already noted that due to the symmetry of the problem and the stationarity of $\mathbf{f}$, $E\left[w_i^2\right]$ does not depend on i; hence, if $E\left[\nu_i^2\right]$ does not depend on i as well, we can use the simplified expression

$$\text{WNR} = \frac{E[w^2]}{E[\nu^2]}. \tag{13}$$

As we will show later, this is the case with the gain attack, but not with the quantization attack, for which, due to the dependence of quantization noise on the spreading sequence $\mathbf{s}$, the more general expression (12) must be used.

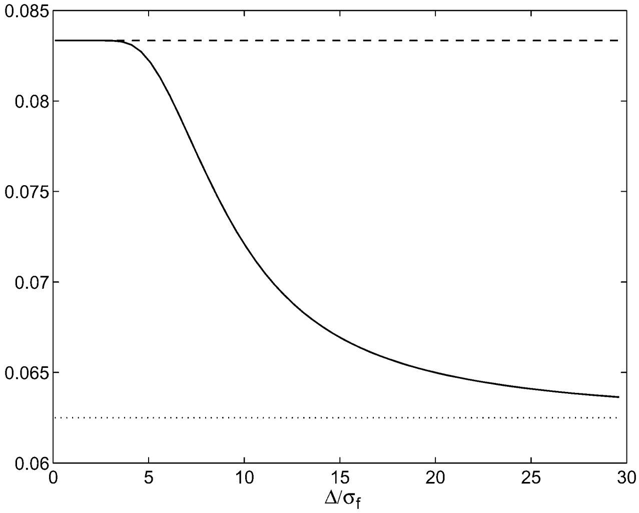


Fig. 2. Sketch of α, (16), versus Δ/σ_f. Dashed line: the value of α under the assumption of a uniform quantization error $(\alpha = 1/12)$. Dotted line: the limit for $\Delta/\sigma_f \rightarrow \infty (\alpha = 1/16)$.

A. DWR Computation for ST-DM

The computation of the actual DWR in the ST-DM case requires some care, since the common assumption that the quantization noise along each component of the host feature sequence is uniformly distributed does not hold. To be specific, let us start by noting that, due to (7), it follows that

$$E[w^2] = \frac{E\left[(\rho_w - \rho_f)^2\right]}{r}. \tag{14}$$

Hence, the DWR depends on the quantization noise affecting the correlation ρ. In order to compute $E\left[(\rho_w - \rho_f)^2\right]$, we can observe that due to the normality of the host features, and to the fact that the spreading vector has unitary norm, ρ_f follows a Gaussian pdf with zero mean and variance $\sigma_\rho^2 = \sigma_f^2$. In addition, a lower quantization error can be obtained by letting $d = \Delta/4$, so that two symmetric codebooks are associated to $b = 0$ and $b = 1$. Under this assumption, and by assuming that $b = 0$ and $b = 1$ are equiprobable, we can write

$$\begin{aligned} E\left[(\rho_w - \rho_f)^2\right] &= E\left[(\rho_w - \rho_f)^2 | b = 0\right] \\ &= \sum_{i=-\infty}^{\infty} \int_{i\Delta - \Delta/4}^{i\Delta + 3\Delta/4} \left(x - \frac{\Delta}{4} - i\Delta\right)^2 f_{\rho_f}(x)dx \end{aligned} \tag{15}$$

with

$$f_{\rho_f}(x) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{x^2}{2\sigma_f^2}\right) \tag{16}$$

where we have exploited the symmetry of the problem with respect to the cases $b = 0$ and $b = 1$. By simple algebra, it can be shown that the above quantity assumes the form

$$E\left[(\rho_w - \rho_f)^2\right] = \alpha\Delta^2 \tag{17}$$

with α depending on the ratio Δ/σ_f. In particular, it can easily be shown that

$$\alpha(\Delta/\sigma_f) = \sum_{i=-\infty}^{\infty} \int_{(i-1/4)\Delta/\sigma_f}^{(i+3/4)\Delta/\sigma_f} \left(\frac{u}{\Delta/\sigma_f} - \frac{1}{4} - i\right)^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du. \tag{18}$$

Interestingly, the values of α stemming from (15) can be compared to the result obtained by assuming that the quantization error is uniformly distributed $(\alpha = 1/12)$. Such a comparison is shown in Fig. 2, where α is plotted against Δ/σ_f. As it can be seen, the quantization error is always smaller than that obtained by assuming a uniform distribution (dashed line), and tends to 1/16 for large values of Δ/σ_f. As already noted in [11], this gives the ST-DM an extra advantage with respect to DM with bit repetition, since a larger quantization step can be used for a given DWR.

In Sections III and IV, ST-DM performances are evaluated as a function of the true DWR as resulting from (14) and (15), i.e.,

$$\mathrm{DWR} = \frac{r\sigma_f^2}{\alpha\Delta^2}. \tag{19}$$

III. Gain Attack

The performances of ST-DM have been extensively studied in [11] by assuming that the attack is limited to the addition of Gaussian noise. In this section we extend the analysis by assuming that, prior to noise addition, the marked host features are scaled by an unknown scale factor g. While corresponding to very common manipulations such as brightness or loudness changes, this attack is not easily taken into account by models describing the distortion introduced by the attacker in terms of the mean square error (mse) between the marked and the attacked signals. In fact, whereas the introduction of a scale factor may correspond to very high mse values, the perceived degradation of the host signal is very small (sometimes the scaled signal looks nicer than the original one).

A. Theoretical Analysis

Our goal is to find a closed-form expression for the bit-error rate as a function of DWR, WNR, and g, where the WNR only takes into account the distortion introduced by the additive part of the attack [i.e., we do not use exactly (11) and (12)]. To do so, let us observe that after scaling and noise addition, the marked features can be written as

$$\mathbf{f}'_w = g\mathbf{f}_w + \mathbf{n} \tag{20}$$

where by $\mathbf{n}$, we indicated the Gaussian noise (with zero mean and variance σ_n^2) added by the attacker. It is, thus, immediate to verify that in this case

$$\mathrm{WNR} = \frac{\alpha\Delta^2}{r\sigma_n^2}. \tag{21}$$

Furthermore, it is easy to see that the correlation between the attacked features and the reference direction $\mathbf{s}$ is given by

$$\rho' = g\rho_w + n_\rho = g\mathcal{Q}_{0/1}(\rho_f) + n_\rho \tag{22}$$

with $n_\rho = \mathbf{n} \cdot \mathbf{s}$ normally distributed with variance $\sigma_{n_\rho}^2 = \sigma_n^2$. We can now compute the error probability conditioned to the embedding of a bit 0 as

$$P_{e|0} = \sum_{i=-\infty}^{\infty} p(u_{0,i}) P_{e|u_{0,i}} \tag{23}$$

with

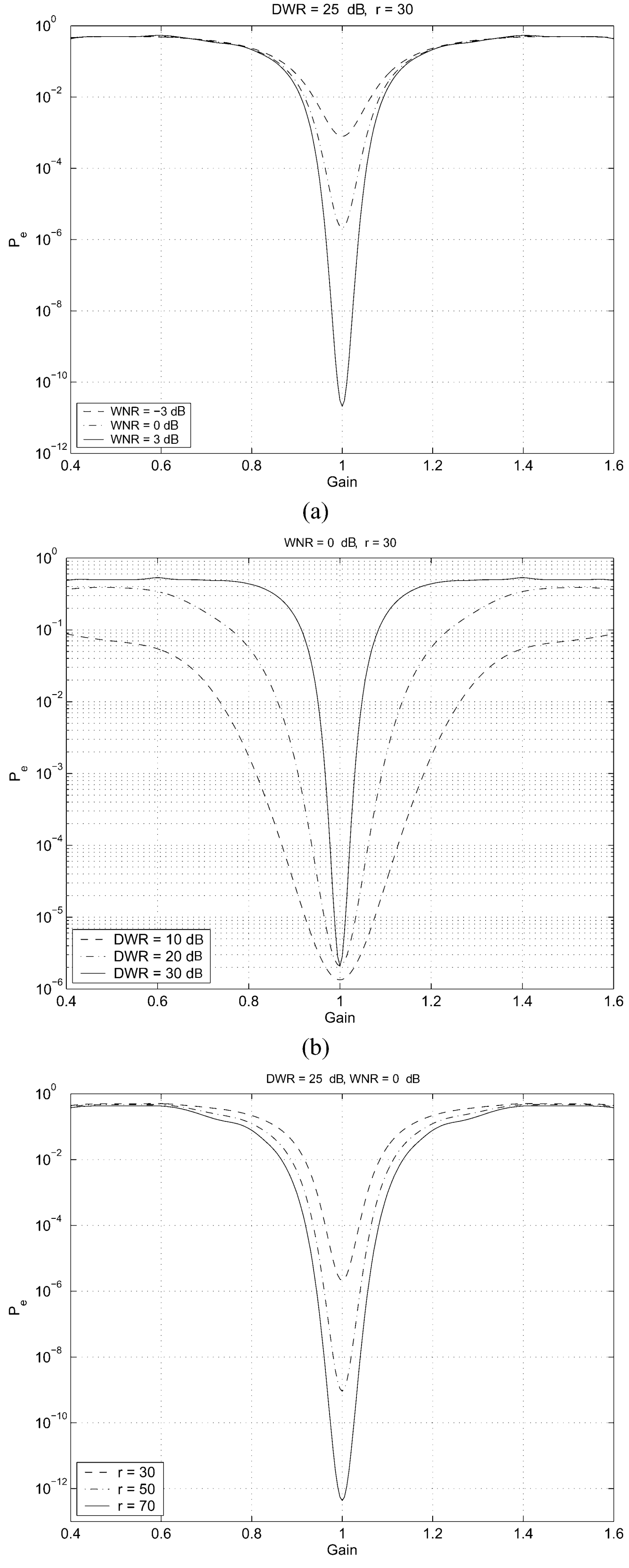
$$\begin{aligned} P_{e|u_{0,i}} &= P\left(\rho' \in \bigcup_j \left[\left(j + \frac{1}{2}\right)\Delta, (j+1)\Delta\right] |u_{0,i}\right) \\ &= \sum_{j=-\infty}^{\infty} \int_{(j+1/2)\Delta}^{(j+1)\Delta} \frac{1}{\sqrt{2\pi\sigma_{n_\rho}^2}} \\ &\quad \times e^{\left(-(\rho' - g(i+1/4)\Delta)^2/2\sigma_{n_\rho}^2\right)} d\rho' \\ &= \sum_{j=-\infty}^{\infty} \int_{(j+1/2-g(i+1/4))(\Delta/\sigma_{n_\rho})}^{(j+1-g(i+1/4))(\Delta/\sigma_{n_\rho})} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \end{aligned} \tag{24}$$

and

$$\begin{aligned} p(u_{0,i}) &= P\left(\rho_f \in \left[\left(i - \frac{1}{4}\right)\Delta, \left(i + \frac{3}{4}\right)\Delta\right]\right) \\ &= \int_{(i-1/4)(\Delta/\sigma_f)}^{(i+3/4)(\Delta/\sigma_f)} \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)} dx. \end{aligned} \tag{25}$$

For the symmetry of the problem (remember that we let $d = \Delta/4$), we have that $P_{e|1} = P_{e|0}$, thus yielding $P_e = P_{e|1} = P_{e|0}$.

The error probability given by the above equations is plotted in Fig. 3 for several values of DWR, WNR, r and g. Specifically, Fig. 3(a) shows the bit-error probability for DWR $= 25$ dB, $r = 30$ and for WNR $= -3$ dB, 0 dB, and $+3$ dB. As can be seen, the performance of ST-DM decreases rapidly as soon as the value of g departs from 1, up to a point that for $g < 0.9$ and $g > 1.1$, the error probability is unacceptably high. At the same time, we can see that the influence of WNR on this behavior is negligible. This is not the case when DWR is varied since, as shown in Fig. 3(b), for lower values of DWR, the range of admissible g is wider. This is a very interesting result, since as opposed to the AWGN case, where the performance are almost insensitive to DWR, robustness against the gain attack may be augmented by increasing the watermark strength. Finally, in Fig. 3(c), the error probability for different values of r is shown. Even in this case, the range of admissible g's increases for higher values of r; however, the improvement is less evident than in the DWR case. This improvement is due to the fact that having fixed the DWR, increasing r implies an increase of the quantization step Δ [see (19)], producing an effect similar to that just described for the variable DWR case: The difference in this case is that by



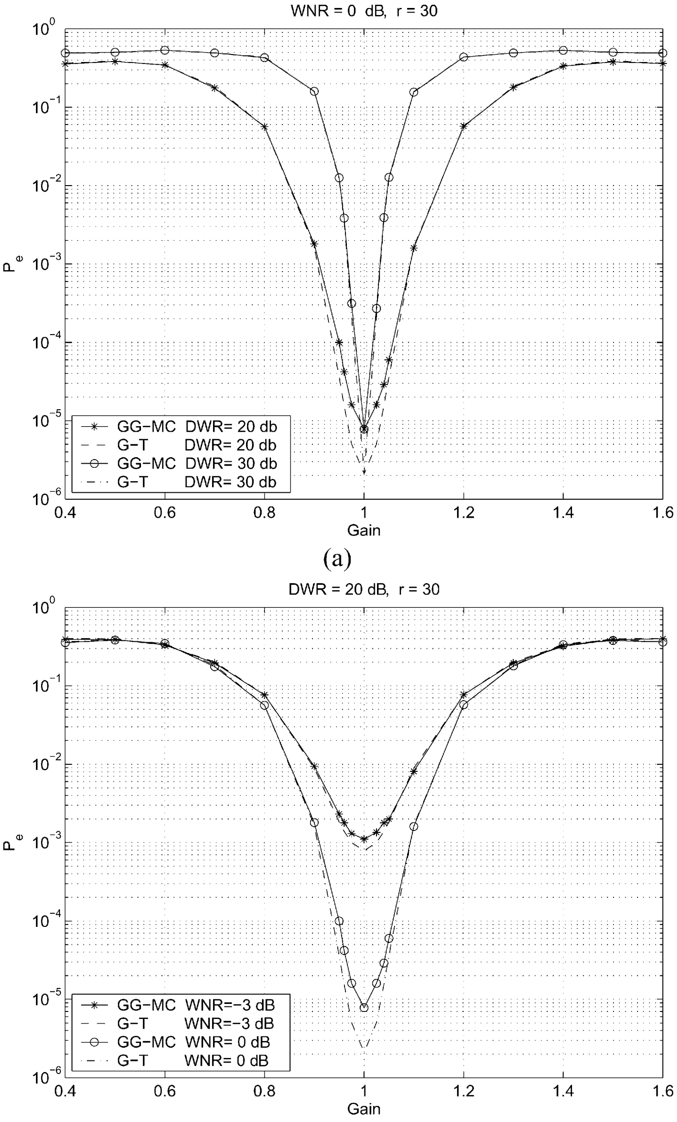
(a)

(b)

(c)

Fig. 3. Bit-error probability in the presence of gain attack plus Gaussian noise addition, for different values of: (a) WNR, (b) DWR, and (c) r.

increasing r, the watermarking payload is decreased, whereas this does not occur if DWR is decreased.



(a)

(b)

Fig. 4. Bit-error probability in the presence of gain attack plus gaussian noise addition, for generalized Gaussian features ($c = 1$, i.e., Laplacian) (GG-MC). The bit-error probability obtained in the case of Gaussian features is reported as well (G-T).

B. Monte Carlo Simulations

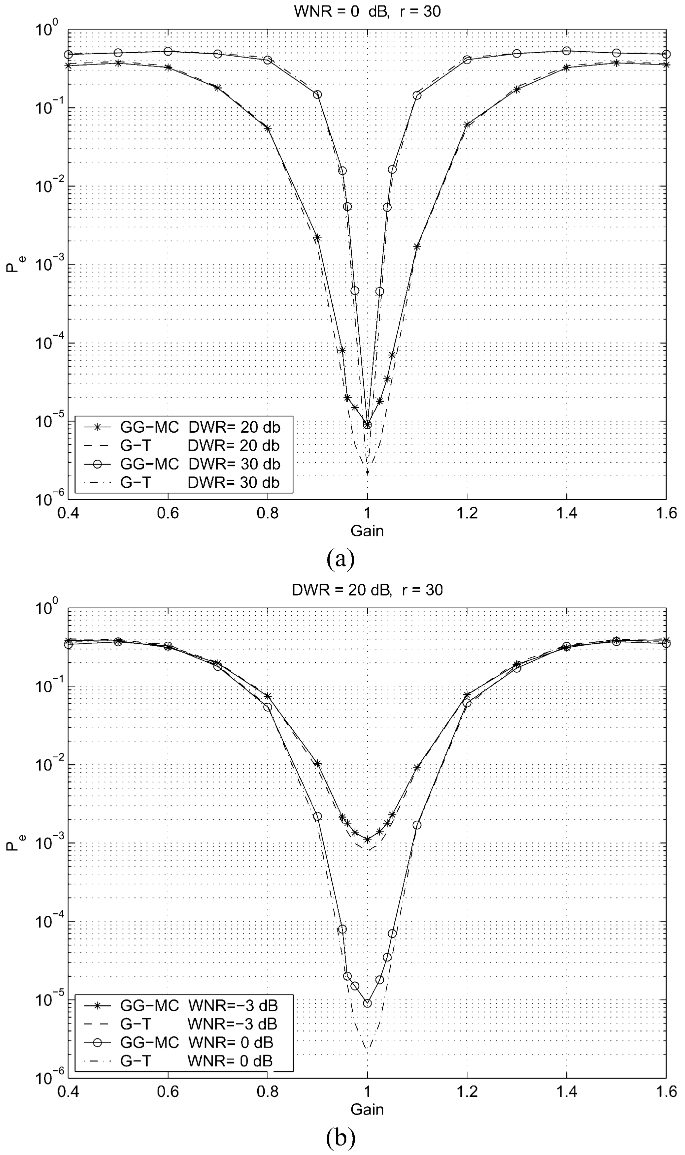
For the analysis carried out so far, we have assumed that the host features follow a Gaussian pdf, which is only rarely the case. For example, by considering the case of still images, pixel values in the spatial domain are better modeled by a uniform pdf, whereas DCT coefficients are conveniently described by means of a generalized Gaussian pdf, given by

$$p(x) = A\exp(-|\beta x|^c) \tag{26}$$

where the parameters A and β can be expressed as a function of the shape parameter c and the standard deviation of x

$$\beta = \frac{1}{\sigma_x}\sqrt{\frac{\Gamma\left(\frac{3}{c}\right)}{\Gamma\left(\frac{1}{c}\right)}} \tag{27}$$

$$A = \frac{\beta c}{2\Gamma\left(\frac{1}{c}\right)} \tag{28}$$



(a)

(b)

Fig. 5. Bit-error probability in the presence of gain attack plus Gaussian noise addition, for generalized Gaussian features ($c = 0.5$) (GG-MC). The bit-error probability obtained in the case of Gaussian features is reported as well (G-T).

and $\Gamma(x)$ is the standard gamma function. The parameter c controls the shape of $p(x)$. For example, it can be seen that the Gaussian and the Laplacian distributions are special cases of the generalized Gaussian pdf, given by $c = 2$ and $c = 1$, respectively. We, then, used Monte Carlo simulations to extend theoretical analysis to the case of host features following a generalized Gaussian pdf. In Figs. 4 and 5, the results we obtained for $c = 1$ (i.e., for Laplacian features) and $c = 0.5$ with Monte Carlo simulations (GG-MC) are given and compared with those theoretically obtained with Gaussian host features (G-T). As can be seen, at least in the case of generalized Gaussian features that we are considering here, the pdf of the host features has a minor impact on the robustness of the watermark, even if a certain deterioration of the performance may be appreciated, especially for $g \simeq 1$. A possible explanation for this relative insensitivity of the performance on the distribution of the host features is that the sole term that depends on this characteristic is the $p(u_{0,i})$ probability, as given by (25), but if r is large enough, the central limit theorem can be invoked to sustain that the distribution of ρ_f remains Gaussian, regardless of the distribution of the host features themselves.

IV. Quantization Attack

Let us now consider another attack that commonly affects multimedia documents, i.e., feature quantization. This is the kind of attack that occurs when the document is compressed. In particular we will consider the case in which the compression and the watermarking domains coincide.

A. *Theoretical Analysis*

The model we will use for analyzing this attack is the following:

$$f'_{w,i} = \mathcal{Q}_{\Delta_a}(f_{w,i}) = f_{w,i} + q_i \tag{29}$$

where Δ_a represents the quantization process defining the attack and q_i is the corresponding (feature dependent) quantization noise. In this case, the correlation between the attacked features and the reference direction $\mathbf{s}$ is given by

$$\rho' = \sum_{i=1}^{r} f_{w,i} s_i + \sum_{i=1}^{r} q_i s_i = \rho_w + q_a. \tag{30}$$

For the symmetry of the problem, we can condition our analysis to the embedding of a 0 bit. We have

$$\begin{aligned} P_{e|0} &= \sum_{k=-\infty}^{\infty} p(u_{0,k}) P_{e|u_{0,k}} \\ &= \sum_{k=-\infty}^{\infty} p(u_{0,k}) \\ &\quad \times P\left(\rho' \in \bigcup_j \left[\left(j+\frac{1}{2}\right)\Delta, (j+1)\Delta\right] |u_{0,k}\right) \\ &= \sum_{k=-\infty}^{\infty} p(u_{0,k}) \sum_{j=-\infty}^{\infty} \int_{(j+1/2)\Delta}^{(j+1)\Delta} f_{\rho'|u_{0,k}}(\rho')d\rho' \\ &= \sum_{k=-\infty}^{\infty} p(u_{0,k}) \sum_{j=-\infty}^{\infty} \int_{(j+1/2)\Delta}^{(j+1)\Delta} f_{q_a|u_{0,k}}(\rho' - u_{0,k})d\rho' \\ &= \sum_{k=-\infty}^{\infty} p(u_{0,k}) \sum_{j=-\infty}^{\infty} \int_{(j-k+1/4)\Delta}^{(j-k+3/4)\Delta} f_{q_a|u_{0,k}}(\rho')d\rho' \end{aligned} \tag{31}$$

where we have exploited the fact that $u_{0,k} = k\Delta + \Delta/4$, and we have indicated with $f_{x|y}(x)$ the probability density function of x conditioned to y. In order to evaluate the error probability, we need the pdf of the random variable q_a conditioned to *transmission* of the codebook entry $u_{0,k}$. Let us start by observing that q_a is the weighted sum of the quantization noise values q_i affecting each watermarked feature. As such, each q_i will depend on the corresponding $f_{w,i}$. The analysis is complicated by the fact that, at least in principle, the watermarked features $f_{w,i}$ are not independent. To see this, let us decompose the vector $\mathbf{f}$ into a component $\mathbf{f}^{\perp}$ orthogonal to $\mathbf{s}$ and a component parallel to $\mathbf{s}$.

Of course, ST-DM will only affect the parallel part by replacing it with $\rho_w\mathbf{s}$, i.e.,

$$\mathbf{f}_w = \mathbf{f}^\perp + \rho_w\mathbf{s} \tag{32}$$

where the components of $\mathbf{f}^\perp$ are a linear combination of the original features f_i that we assumed to be Gaussian i.i.d. random variables. The distribution of $\mathbf{f}^\perp$ will thus be a multivariate Gaussian with zero mean and covariance matrix C to be calculated. As to $\mathbf{f}_w$ it is immediate to see that it still follows a multivariate Gaussian distribution, with

$$E[\mathbf{f}_w|u_{0,k}] = u_{0,k}\mathbf{s} \tag{33}$$

and covariance matrix C

$$E\left[(\mathbf{f}_w - \rho_w\mathbf{s})^t(\mathbf{f}_w - \rho_w\mathbf{s})\right] = E\left[(\mathbf{f}^\perp)^t\mathbf{f}^\perp\right] \tag{34}$$

where we have assumed all vectors to be row vectors. In order to calculate $E\left[(\mathbf{f}^\perp)^t\mathbf{f}^\perp\right]$, let us introduce the projection operators that project the vector $\mathbf{f}$ over the space orthogonal to $\mathbf{s}$. From linear algebra, we know that such an operator has the form $P_s^\perp = I - \mathbf{s}^t\mathbf{s}$; hence, we can write

$$\begin{aligned} E\left[(\mathbf{f}^\perp)^t\mathbf{f}^\perp\right] &= E\left[(I-\mathbf{s}^t\mathbf{s})\mathbf{f}^t\mathbf{f}(I-\mathbf{s}^t\mathbf{s})\right] \\ &= E\left[\mathbf{f}^t\mathbf{f}\right] - \mathbf{s}^t\mathbf{s}E\left[\mathbf{f}^t\mathbf{f}\right] - E[\mathbf{f}^t\mathbf{f}]\,\mathbf{s}^t\mathbf{s} \\ &\quad + \mathbf{s}^t\mathbf{s}E\left[\mathbf{f}^t\mathbf{f}\right]\mathbf{s}^t\mathbf{s} \\ &= \sigma_f^2 I - 2\sigma_f^2\mathbf{s}^t\mathbf{s} + \sigma_f^2\mathbf{s}^t\mathbf{s}\mathbf{s}^t\mathbf{s} \\ &= \sigma_f^2\left(I - \mathbf{s}^t\mathbf{s}\right) \end{aligned} \tag{35}$$

where in the last equality, we exploited the fact that $\mathbf{s}\mathbf{s}^t = \|\mathbf{s}\|^2 = 1$. By remembering that $s_i = \pm 1/\sqrt{r}$, we can conclude that

$$C_{ij} = E\left[f_i^\perp f_j^\perp\right] = \begin{cases} \left(1-\frac{1}{r}\right)\sigma_f^2, & \text{if } i = j \\ \pm\frac{1}{r}\sigma_f^2, & \text{if } i \neq j \end{cases} \tag{36}$$

which proves the dependency between $f_{w,i}$ coefficients. If r is large enough, though, the correlation between $f_{w,i}$ vanishes, thus permitting us to consider the q_i terms independent of each other. Furthermore, if r is large enough, we can exploit the Central Limit Theorem, and approximate q_a, as given by (30), by a Gaussian random variable with mean (we avoid explicitly indicating the conditioning to $u_{0,k}$ for notation simplicity)

$$\mu_{q_a} = \sum_{i=1}^{r} s_i\mu_{q_i} \tag{37}$$

and variance

$$\sigma_{q_a}^2 = \sum_{i=1}^{r} s_i^2\sigma_{q_i}^2 = \frac{1}{r}\sum_{i=1}^{r}\sigma_{q_i}^2. \tag{38}$$

The mean μ_{q_i} and the variance $\sigma_{q_i}^2$ resulting from the quantization with a step size Δ_a of a Gaussian random variable having mean $u_{0,k}s_i = (k\Delta + \Delta/4)\,s_i$ and variance (approximately) σ_f^2 remain to be estimated. It is easy to demonstrate that

$$\begin{aligned} \mu_{q_i} &= \sum_{l=-\infty}^{\infty}\int_{\Delta_a(l-1/2)}^{\Delta_a(l+1/2)}(q - l\Delta_a)\frac{1}{\sqrt{2\pi\sigma_f^2}} \\ &\quad\times e^{-((q-k\Delta-\Delta/4)^2/2\sigma_f^2)}dq \\ &= \sum_{l=-\infty}^{\infty}\int_{(\Delta_a(l-1/2)-u_{0,k}s_i)/\sigma_f}^{(\Delta_a(l+1/2)-u_{0,k}s_i)/\sigma_f}(\sigma_f t + u_{0,k}s_i - l\Delta_a) \\ &\quad\times\frac{1}{\sqrt{2\pi}}e^{-(t^2/2)}dt \\ &= u_{0,k}s_i - \Delta_a\sum_{l=-\infty}^{\infty} l \\ &\quad\times\int_{(\Delta_a(l-1/2)-u_{0,k}s_i)/\sigma_f}^{(\Delta_a(l+1/2)-u_{0,k}s_i)/\sigma_f}\frac{1}{\sqrt{2\pi}}e^{-(t^2/2)}dt. \end{aligned} \tag{39}$$

Similarly, $\sigma_{q_i}^2$ can be computed based on the mean square value (MSV) of q_i that results as

$$\begin{aligned} \mathrm{MSV}_{q_i} &= \sum_{l=-\infty}^{\infty}\int_{\Delta_a(l-1/2)}^{\Delta_a(l+1/2)}(q - l\Delta_a)^2\frac{1}{\sqrt{2\pi\sigma_f^2}} \\ &\quad\times e^{-((q-k\Delta-\Delta/4)^2/2\sigma_f^2)}dq \\ &= \sigma_f^2 + \frac{1}{r}u_{0,k}^2 + \Delta_a^2\sum_{l=-\infty}^{\infty} l^2 \\ &\quad\times\int_{(\Delta_a(l-1/2)-u_{0,k}s_i)/\sigma_f}^{(\Delta_a(l+1/2)-u_{0,k}s_i)/\sigma_f}\frac{1}{\sqrt{2\pi}}e^{-(t^2/2)}dt \\ &\quad - 2\sigma_f\Delta_a\sum_{l=-\infty}^{\infty} l\int_{(\Delta_a(l-1/2)-u_{0,k}s_i)/\sigma_f}^{(\Delta_a(l+1/2)-u_{0,k}s_i)/\sigma_f}\frac{t}{\sqrt{2\pi}} \\ &\quad\times e^{-(t^2/2)}dt \\ &\quad - 2u_{0.k}s_i\Delta_a\sum_{l=-\infty}^{\infty} l\int_{(\Delta_a(l-1/2)-u_{0,k}s_i)/\sigma_f}^{(\Delta_a(l+1/2)-u_{0,k}s_i)/\sigma_f}\frac{1}{\sqrt{2\pi}} \\ &\quad\times e^{-(t^2/2)}dt. \end{aligned} \tag{40}$$

Note that MSV_{q_i} also influences the WNR of the attack, that, according to (12), results as

$$\mathrm{WNR} = \frac{\alpha\Delta^2}{\sum\limits_{k=-\infty}^{k=-\infty} p(u_{0,k})\sum\limits_{i=1}^{r}\mathrm{MSV}_{q_i}}. \tag{41}$$

B. Monte Carlo Simulations

As a first step, we verified the validity of the assumptions leading to the theoretical bit-error probability derived above. Such assumptions, i.e., independence of q_i and use of the Central Limit Theorem to calculate the pdf of q_a, are certainly valid for large values of r, however their impact on the accuracy of the theoretical results must be verified experimentally. In Fig. 6, the theoretical bit-error probability (G-T) is compared to the results obtained by means of Monte Carlo simulation, for the case of $\mathrm{DWR} = 20$ and $r = 30$ (G-MC); as it can be seen the agreement of theoretical results with simulations is excellent. In Fig. 7, the

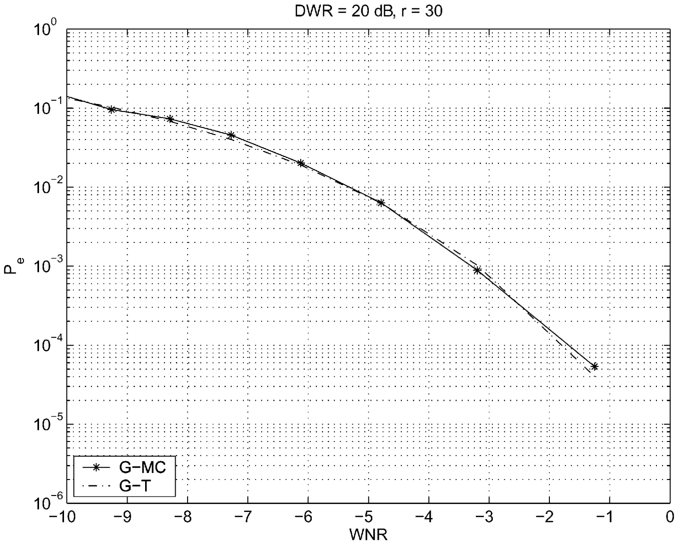


Fig. 6. Comparison between the theoretical error probability and the bit-error rate obtained through Monte Carlo simulations for various values of WNR for the quantization attack.

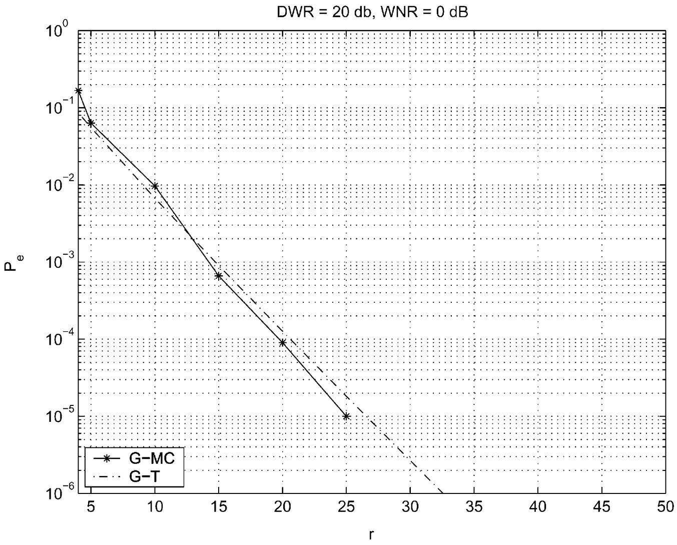
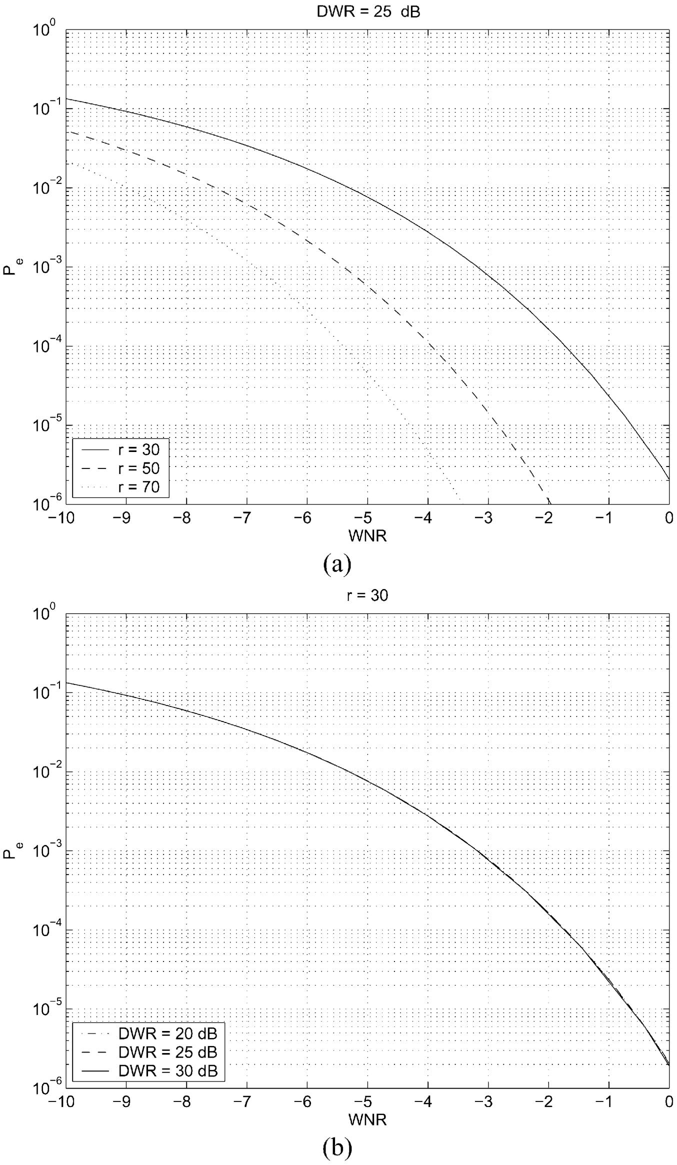


Fig. 7. Comparison between the theoretical error probability and the bit-error rate obtained through Monte Carlo simulations for various values of r for the quantization attack.

analysis is repeated for several values of r (DWR $=$ 20 and WNR $=$ 0): the agreement between theory and simulations is very good, even for rather low values of r.

Given that the theoretical analysis is confirmed by numerical simulations, we can use the theoretical probability of error to get more insight into the performance of ST-DM in the presence of quantization attack. This is the goal of Fig. 8, where the bit-error probability is depicted as a function of WNR for several values of r and DWR. Upon inspection of the results, the positive effect obtained by increasing r comes out, together with the insensitivity of P_e with respect to DWR (the three plots for DWR $=$ 20, 25, and 30 are almost superimposed). It can be useful to understand how the WNR calculated on the watermarked and attacked features is related, for example, to the JPEG compression quality factor, as this is the usual parameter considered to characterize the strength of the JPEG attack. This relation will depend on the type of features that are chosen for watermarking and the strength of the watermark: We consider here the case in which the watermark is embedded into the DCT coefficients belonging to the third, fourth, fifth, and sixth diagonals of each 8 × 8 block of an image (a total of 18 coefficients per block are watermarked). As to watermark strength, we let DWR = 25 dB. In Table I, the WNR obtained by considering a 10% quality factor is reported for some standard images (the popular ImageMagick tool has been used for JPEG compression): It is apparent that in



(a)

(b)

Fig. 8. Theoretical bit-error probability in the presence of quantization attack for Gaussian host features.

TABLE I
WNR OBTAINED BY JPEG-COMPRESSING SOME STANDARD IMAGES WITH A QUALITY FACTOR OF 10% AND DWR $=$ 25 dB

Image	Lenna	Boat	Bridge
WNR (db)	-1.1	-0.9	-6.3

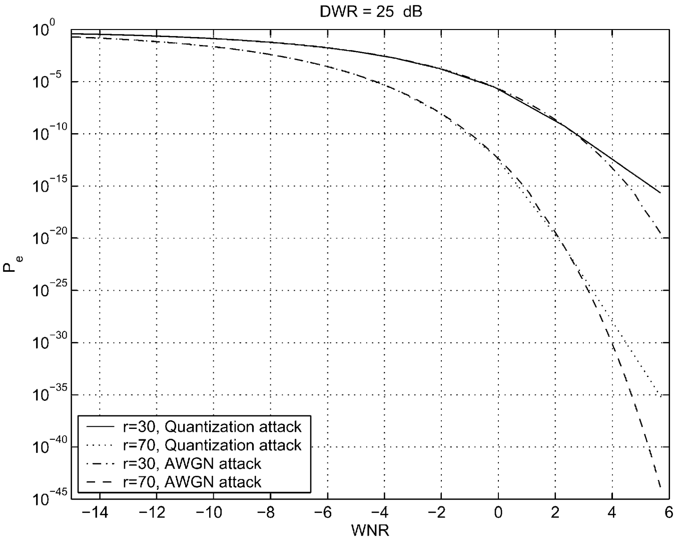


Fig. 9. Bit-error probability in the presence of additive Gaussian noise and quantization attack.

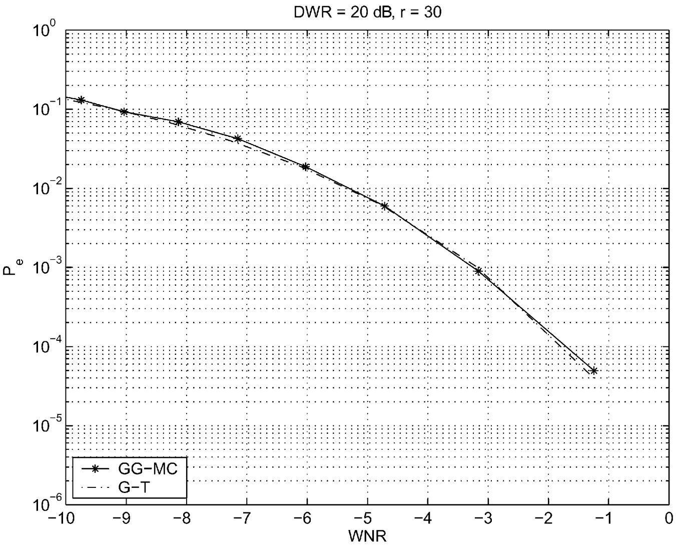


Fig. 10. Bit-error probability in the presence of quantization attack for Gaussian (G-T) and generalized Gaussian host features (GG-MC).

order to obtain WNR values lower than zero, a very low quality factor must be considered, thus validating the significance of the experiments we conducted.

In Fig. 9, the effect of the quantization attack is compared with that obtained by means of an additive Gaussian attack, as it can be seen that the two effects are almost identical. This is not really surprising if we consider (30) and that, thanks to the CLT, the effect of the quantization noise on the correlation between the feature vector and the spreading direction $\mathbf{s}$ can be considered to have an almost Gaussian distribution, thus yielding to a model of the attack that is almost identical to that of the simple Gaussian noise addition.

Finally, Monte Carlo simulations have been used to extend the analysis to the case of host features following a generalized Gaussian pdf (GG-MC). The results we obtained (Fig. 10) ensure that the shape of the pdf has a very low impact on the robustness against the quantization attack.

V. Conclusions and Future Works

In this work, we have analyzed the performance of ST-DM watermarking in the presence of non additive noise. In particular, due to the importance they assume in multimedia signal processing applications, we considered the gain attack plus additive Gaussian noise and the quantization attack. By modeling the host features as i.i.d. Gaussian random variables, we managed to derive a closed form expression for the bit-error probability as a function of DWR and WNR. Though some of the results were expected, the availability of an exact expression for the bit-error represents a very important tool which allows to ground the design of any ST-DM watermarking system on a solid scientific basis.

In general, the excellent robustness of ST-DM watermarking is confirmed by our analysis, especially with regard to the quantization attack, which demonstrated to be approximately as harmful as the conventional AWGN attack. The only noticeable exception is represented by the gain attack, since even for values of g as close to 1 as 0.9 or 1.1, the error probability becomes excessively high. Interestingly such an effect can be limited by increasing r or, even better by increasing DWR.

A further step toward an even more realistic analysis of the performance of ST-DM, in terms of more realistic attacks and use of host features that do not exactly obey to a theoretical pdf, requires that experimental tests are performed. It is our intention, then, to extend the present analysis so to encompass the watermarking of true multimedia documents, namely still images, and to evaluate the robustness of ST-DM against common manipulations such as linear filtering, histogram equalization and JPEG coding.

Acknowledgment

This paper is dedicated to Franco Bartolini, who passed away on January 1, 2004, just after the present paper was finished. Both our work and our life have greatly benefited by the scientific collaboration with Franco and, most of all, by his friendship. We will surely miss him in the days and years to come. (Mauro and Alessandro).

References

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

[2] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. IEEE*, vol. 87, pp. 1127–1141, July 1999.

[3] B. Chen and G. Wornell, "Achievable performance of digital watermarking schemes," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems, ICMCS '99*, vol. 1, Florence, Italy, June 1999, pp. 13–18.

[4] P. Moulin, "The role of information theory in watermarking and its application to image watermarking," *Signal Processing*, vol. 81, no. 6, pp. 1121–1139, 2001.

[5] A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1639–1667, June 2002.

[6] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 439–441, May 1983.

[7] S. I. Gelf and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[8] J. Chou, S. S. Pradhan, L. El Ghaoui, and K. Ramchandran, "Watermarking based on duality with distributed source coding and robust optimization principles," in *Proc. 7th IEEE Int. Conf. Image Processing, ICIP'00*, vol. 1, Vancouver, BC, Canada, Sept. 2000, pp. 585–588.

[9] J. Chou and K. Ramchandran, "Robust turbo-based data hiding for image and video sources," in *Proc. 9th IEEE Int. Conf. Image Processing, ICIP'02*, vol. 2, Rochester, NY, Sept. 2002, pp. 133–136.

[10] M. L. Miller, G. J. Doerr, and I. J. Cox, "Applying informed coding and embedding to design a robust high-capacity watermark," *IEEE Trans. Image Processing*, vol. 13, pp. 792–807, June 2004.

[11] F. Perez-Gonzalez, F. Balado, and J. R. Hernandez, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Trans. Signal Processing*, vol. 51, pp. 960–980, Apr. 2003.

Mauro Barni (S'90–M'96) graduated in electronic engineering in 1991 and received the Ph.D. degree in informatics and telecommunications in October 1995, both from from the University of Florence, Florence, Italy.

From 1991 through 1998, he was with the Department of Electronic Engineering, University of Florence. Since September 1998, he has been with the Department of Information Engineering, University of Siena, Siena, Italy, where he is an associate professor. His main interests are in the field of digital image processing and computer vision. His current research activity is focused on the application of image processing techniques to copyright protection and authentication of multimedia data (digital watermarking). He is author/co-author of more than 130 papers published in international journals and conference proceedings and holds three patents in this field. He is on the editorial board of the *Eurasip Journal of Applied Signal Processing*.

Dr. Barni serves as associate editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, the IEEE SIGNAL PROCESSING MAGAZINE (Column and Forum section). Dr. Barni is a member of the IEEE Multimedia Signal Processing Technical Committee (MMSP-TC) and the Signal Processing Conference Board.

Franco Bartolini (M'96) was born in Rome, Italy, in 1965. He graduated (cum laude) in electronic engineering in 1991 and received the Ph.D degree in informatics and telecommunications in November 1996, both from the University of Florence, Florence, Italy. He passed away on January 1, 2004.

From November 2001 until 2003, he was an Assistant Professor with the University of Florence. His research interests include digital image sequence processing, still and moving image compression, nonlinear filtering techniques, image protection and authentication (watermarking), image processing applications for the cultural heritage field, signal compression by neural networks, and secure communication protocols. He published more than 130 papers on these topics in international journals and conferences. He held three Italian and one European patent in the field of digital watermarking.

Dr. Bartolini was a member of the Program Committee of the SPIE/IST Workshop on Security, Steganography, and Watermarking of Multimedia Contents as well as the Technical Program Co-Chair for IEEE MMSP Workshop 2004. Dr. Bartolini was a member of SPIE and IAPR.

Alessandro Piva graduated (cum laude) in electronic engineering and received the Ph.D degree in informatics and telecommunications, both from the University of Florence, Florence, Italy, in February 1999.

He was with the Department of Electronics and Telecommunications, University of Florence, as a Postdoctoral Researcher. Since July 2002, he has been a Research Scientist with the National Inter-university Consortium for Telecommunications (CNIT), University of Florence. His research activity is focused on multimedia systems, digital image sequence processing, video and image digital watermarking, image processing techniques for cultural heritage applications and secure communication protocols, and low bit-rate video transmission over wireless networks. He has published more than 70 papers on these topics in international journals and conferences. He holds three Italian and one European patent in the field of digital watermarking.

Dr. Piva was Co-Guest Editor for t6he March 2004 Special Issue of the IEEE TRANSACTIONS ON IMAGE PROCESSING on "Image Processing for Cultural Heritage."