Quantifying Security Leaks in Spread Spectrum Data Hiding: A Game-Theoretic Approach. *

Pedro Comesaña and Fernando Pérez-González

Dept. Teoria de la Señal y Comunicaciones. ETSI Telecom., Universidad de Vigo, 36200 Vigo, Spain

ABSTRACT

A game-theoretic approach is introduced to quantify possible information leaks in spread-spectrum data hiding schemes. Those leaks imply that the attacker knows the set-partitions and/or the pseudorandom sequence, which in most of the existing methods are key-dependent. The bit error probability is used as payoff for the game. Since a closed-form strategy is not available in the general case, several simplifications leading to near-optimal strategies are also discussed. Finally, experimental results supporting our analysis are presented.

1. INTRODUCTION

In the last years there has been an increasing interest by the research community in finding ways of quantifying the security of data-hiding schemes, since cryptography-based approaches to security fall short of watermarking requirements. In this paper we take a step in this direction and analyze and quantify the loss of performance of spread-spectrum data-hiding systems when some information on the structure of the watermark is leaked.

Most spread-spectrum watermarking schemes usually partition the available set of host samples in order to achieve diversity for each hidden bit. This is generally done by making this partition depend on a key that is known to the embedder and the decoder, but not to the attacker. The leakage of information regarding this key is what is analyzed in this paper. This leakage may be due to improper key handling, collusion attacks or even the mere structure of the watermark; think, for instance, of periodic watermarks, which have been proposed as an efficient way of preserving synchronization.

Some researchers have dealt with game-theoretic aspects of data hiding capacities.[1, 2] However payoffs other than channel capacity are also possible in the data hiding game, as already suggested in[3] and developed in.[4] Here, we build on this idea to determine optimal playing strategies by considering that the bit error rate (BER) for the hidden information defines the payoff in the game.

Three agents generally play the data hiding game: embedder, attacker and decoder, as depicted in Figure 1. Each one has different objectives and constraints which frequently lead to colliding interests. First of all, the embedder is responsible for hiding the information in the host image in a secret way, trying to hinder the estimation of the watermark by the attacker. In most applications, the embedder faces very strict perceptual constraints so as to guarantee that there is no loss of value in the watermarked image. In fact, we depart from other game-theoretic works in imposing more severe constraints in order to guarantee the imperceptibility of the watermark, leaving the embbeder out of the game.

Second, the watermark is manipulated by the attacker. The attacker can play an active role by frustrating or at least making it difficult the decoding process. Needless to say, the attacker's strategy will also be constrained by a distortion measure with some perceptual meaning. In this paper we will consider only additive noise attacks. As we will show, when the attacker knows some information about the partitions, his/her averaging strategy can be significantly improved by taking into account this information.

*Work partially funded by *Xunta de Galicia* under grants PGIDT01 PX132204PM and PGIDT02 PXIC32205PN; CYCIT, AMULET project, reference TIC2001-3697-C03-01; and FIS, IM3 Research Network, reference FIS-G03/185.

Further author information: (Corresponding author: F.P-G.)
F.P-G.: E-mail: fperez@tsc.uvigo.es, Telephone: +34 986 812124
P.C.A.: E-mail: pcomesan@gts.tsc.uvigo.es, Telephone: +34 986 812683

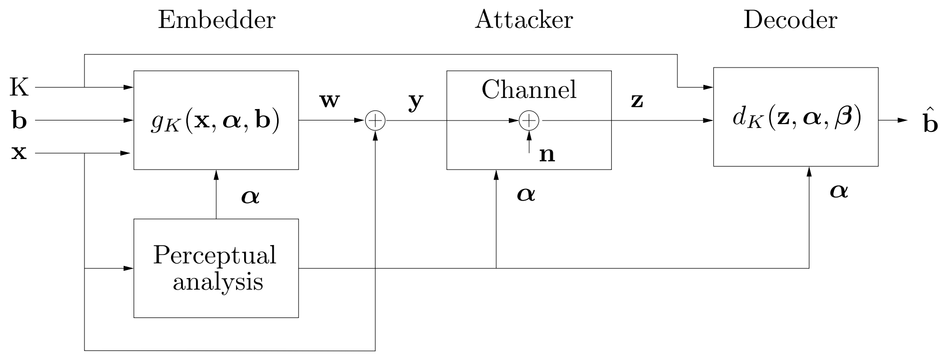

Figure 1. Data hiding model.

The decoder is in charge of retrieving the information in a reliable way. He/she could have information about the attacker's strategy (this is more likely in unintentional attacks), and in this case, he/she could use this information to improve the performance of the system. Note that the converse could also be true, but we must note once again that the decoder does have the final word. Since the decoder knows the partition, his strategy will be to select a decoding strategy that minimizes the bit error probability for each bit. Note that this implies that the objective functions for the attacker and the decoder are not the same (as opposed to most previous game-theoretic approaches) for the non-leakage case.

Finally, the results obtained with these two approaches will be compared from a empirical point of view and the gap between them will be measured. This gap is an effective performance indicator that serves as quantifier of the loss attributable to the leakage of the key used to spread the hidden data.

The paper is structured as follows: Sect. 2 is devoted to formulate Spread-Spectrum methods in matrix form; the optimal decoding filter is presented in Sect. 3, while the optimal attack is computed in Sect. 4. The optimal filter and attack for the case where the complete covariance matrix is replaced by a local version, are recomputed in Sect. 5 and Sect. 6, respectively. In Sect. 7 a further simplication is introduced and the corresponding results are discussed. Experimental results are shown in Section 8 while Section 9 contains our conclusions and future lines of research.

2. A MATRIX FORMULATION OF SPREAD-SPECTRUM

In a watermarking spread-spectrum system the watermarked signal can be described as

$$\mathbf{y} = \mathbf{x} + \mathbf{D}_{\alpha}\mathbf{D}_{\mathbf{S}}\mathbf{P}\mathbf{b} \tag{1}$$

where $\mathbf{x}$ is a zero-mean host signal arranged in a length-L vector, with covariance matrix $\mathbf{R}_{\mathbf{X}}$ whose eigenvalues are $\lambda_{x_j}, j = 1, \cdots, L$. $\mathbf{D}_{\alpha}$ is a diagonal matrix whose elements α denote the greatest amplitude modification in the host signal that renders no perceptual modification. This vector is computed following a perceptual method (for example applying[5, 6] when the watermarking process is performed in the DCT domain). $\mathbf{D}_{\mathbf{S}}$ is a diagonal matrix whose elements contain the spreading sequence which will be assumed to take values in $\{\pm 1\}$. $\mathbf{b}$ are the M symbols to be hidden ($\{\pm 1\}$) and $\mathbf{P} = (\mathbf{p_1 p_2} \cdots \mathbf{p_M})$ is the matrix representing the coefficients assigned to each bit (with size $L \times M$ and values in $\{0, 1\}$), thus $p_{j,i}$ will take the value 1 if the j-th coefficient is used to hide the i-th bit, and is zero otherwise. Let $\mathcal{S}_i$ denote the set of coefficients of the i-th column of $\mathbf{P}$ taking the 1 value, or equivalently, the set of coefficients devoted to convey the i-th bit.

This signal can be modified by the sum of additive noise (attacks) and filtering (due to intentional attacks or aimed at reducing the host signal interference). Therefore, the signal received at the decoder can be written as

$$\mathbf{z} = \mathbf{H}(\mathbf{x} + \mathbf{n}) + \mathbf{H}\mathbf{D}_{\alpha}D_{\mathbf{S}}P\mathbf{b} \tag{2}$$

where $\mathbf{n}$ is zero-mean additive noise independent of $\mathbf{x}$ with covariance matrix $\mathbf{R}_{\mathbf{n}}$ and eigenvalues denoted by $\lambda_{n_j}, j = 1, \cdots, L$. $\mathbf{H}$ corresponds to the attacking or equalizing filter.

If the decoding process is based on a cross correlation (see[7] for a discussion on its optimality), the correlation for the i-th bit, denoted by r_i can be written as

$$r_i = \mathbf{p_i}^T \mathbf{D_S} \mathbf{B} \mathbf{z}, \quad 1 \leq i \leq M \tag{3}$$

being $\mathbf{B}$ an arbitrary filter that is introduced with the purpose of optimizing the decoding process.

Since the number of coefficients devoted to transmit a bit is suppossed to be large, the Central Limit Theorem (CLT) can be applied to show that the probability density function (pdf) of the random variable r_i can be well-approximated by a Gaussian distribution with first and second order moments:

$$\mathrm{E}\{r_i\} = \mathbf{p}_i{}^T \mathbf{D_S} \mathbf{B} \mathbf{H} \mathbf{D}_\alpha \mathbf{D_S} \mathbf{P} \mathbf{b}, \ i = 1, \cdots, M \tag{4}$$
$$\mathrm{E}\{r_i^2\} = \mathbf{p}_i{}^T \mathbf{D_S} \mathbf{B} \mathbf{H} \mathrm{E}\{\mathbf{a}\mathbf{a}^T\} \mathbf{H}^T \mathbf{B}^T \mathbf{D_S} \mathbf{p}_i, \ i = 1, \cdots, M \tag{5}$$

where $\mathbf{a} \triangleq \mathbf{x} + \mathbf{n} + \mathbf{D}_\alpha \mathbf{D_S} \mathbf{P} \mathbf{b}$, and averages are taken over $\mathbf{x}$ and $\mathbf{n}$.

Unfortunately, the above statistics do not depend only on b_i, but also on $b_k, k \neq i$. This means that in order to take a global decision on the available information, one has to jointly consider all the possible components in $\mathbf{b}$. Nevertheless, if we assume that $b_i, i = 1, \cdots, M$ are i.i.d. random variables, independent of $\mathbf{x}$ and $\mathbf{n}$, and taking values in $\{\pm 1\}$ with the same probability ($\mathrm{E}\{b_i\} = 0$, $\mathrm{E}\{b_i^2\} = 1$, $\mathrm{E}\{\mathbf{b}\mathbf{b}^T\} = \mathbf{I}_{M \times M}$), it is possible to write

$$\mathrm{E}\{r_i | b_i = \mathrm{b}_i\} = \mathbf{p}_i{}^T \mathbf{D_S} \mathbf{B} \mathbf{H} \mathbf{D}_\alpha \mathbf{D_S} \mathbf{p}_i \mathrm{b}_i, \quad i = 1, \cdots, M \tag{6}$$
$$\mathrm{Var}\{r_i | b_i = \mathrm{b}_i\} = \mathbf{p}_i{}^T \mathbf{D_S} \mathbf{B} \mathbf{H} \mathbf{A} \mathbf{H}^T \mathbf{B}^T \mathbf{D_S} \mathbf{p}_i, \quad i = 1, \cdots, M \tag{7}$$

being

$$\mathbf{A} \triangleq \mathrm{E}\{\mathbf{a}\mathbf{a}^T\} = \mathbf{R_X} + \mathbf{R_n} \tag{8}$$

Here, expectations are taken over $\mathbf{x}$, $\mathbf{n}$ and the other components of $\mathbf{b}$.

Therefore, we can approximate the probability of error for the i-th bit as

$$P_e(i) = P(\hat{b}_i \neq \mathrm{b}_i) \approx Q\left(\sqrt{\frac{\mathrm{E}^2\{r_i | b_i = \mathrm{b}_i\}}{\mathrm{Var}\{r_i | b_i = \mathrm{b}_i\}}}\right) \tag{9}$$

3. OPTIMAL DECODING FILTER

In this section we will derive the optimal decoding filter when the full covariance matrices of the host signal and the noise are used. The optimal decoding filter $\mathbf{B}$ is defined as the one that minimizes $P_e(i)$. Since the $Q(\cdot)$ function is monotonically decreasing, while the square root is monotonically increasing, this is equivalent to maximizing the following SINR (Signal to Interference plus Noise Ratio),

$$\max_{\mathbf{B}} \mathrm{SINR} \triangleq \frac{\mathrm{E}^2\{r_i | b_i = \mathrm{b}_i\}}{\mathrm{E}\{r_i^2 | b_i = \mathrm{b}_i\} - \mathrm{E}^2\{r_i | b_i = \mathrm{b}_i\}} = \left(\frac{\mathrm{Var}\{r_i | b_i = \mathrm{b}_i\}}{\mathrm{E}^2\{r_i | b_i = \mathrm{b}_i\}}\right)^{-1} \tag{10}$$

which is tantamount to

$$\min_{\mathbf{B}} \frac{\mathrm{Var}\{r_i | b_i = \mathrm{b}_i\}}{\mathrm{E}^2\{r_i | b_i = \mathrm{b}_i\}} \tag{11}$$

Recalling Cauchy-Schwarz inequality

$$\frac{< \mathbf{a}_1, \mathbf{a}_1 >}{|< \mathbf{a}_1, \mathbf{a}_2 >|^2} \geq \frac{1}{< \mathbf{a}_2, \mathbf{a}_2 >} \tag{12}$$

where $< \mathbf{x}, \mathbf{y} >$ denotes the scalar product of $\mathbf{x}$ and $\mathbf{y}$, and equality holds if and only if $\mathbf{a}_1$ and $\mathbf{a}_2$ are colinear. After replacing

$$\begin{aligned} \mathbf{a}_1 &\triangleq \mathbf{A}^{1/2}\mathbf{H}^T\mathbf{B}^T\mathbf{D}_{\mathbf{S}}\mathbf{p}_i && (13) \\ \mathbf{a}_2 &\triangleq \mathbf{A}^{-1/2}\mathbf{D}_{\alpha}\mathbf{D}_{\mathbf{S}}\mathbf{p}_i && (14) \end{aligned}$$

into (12) we see that the minimum of (11) is obtained when

$$\mathbf{B} = K\mathbf{D}_{\alpha}\mathbf{A}^{-1}\mathbf{H}^{-1} \tag{15}$$

with K any real constant. Consequently,

$$\min_{\mathbf{B}} \frac{\mathrm{Var}\{r_i|b_i = \mathrm{b}_i\}}{\mathrm{E}^2\{r_i|b_i = \mathrm{b}_i\}} = \frac{1}{\mathbf{p}_i{}^T\mathbf{D}_{\mathbf{S}}\mathbf{D}_{\alpha}\mathbf{A}^{-1}\mathbf{D}_{\alpha}\mathbf{D}_{\mathbf{S}}\mathbf{p}_i} \tag{16}$$

So this will be the optimal strategy for the decoder, assuming he/she knows the attack. Note that the filtering effect due to $\mathbf{H}$ will be completely removed, as long as this filter is invertible.

4. OPTIMAL ATTACK

In this section we will analyze the optimal attack in three different scenarios of information leakage: 1) Perfect Knowledge of $\mathbf{P}$ and $\mathbf{D}_{\mathbf{S}}$ (no leakage); 2) perfect knowledge of $\mathbf{P}$, $\mathbf{D}_{\mathbf{S}}$ unknown, and 3) both $\mathbf{P}$ and $\mathbf{D}_{\mathbf{S}}$ unknown.

We will introduce also three constraints upon $\mathbf{R}_{\mathbf{n}}$, aimed at forcing it to be a positive definite matrix, bounding the noise power and limiting the per-sample noise variance.

4.1. Perfect Knowledge of P and $\mathbf{D}_{\mathbf{S}}$

The objective of the attacker will be

$$\begin{aligned} \max_{\mathbf{R}_{\mathbf{n}}} P_e &= \max_{\mathbf{R}_{\mathbf{n}}} \frac{1}{M}\sum_{i=1}^{M} P_e(i) \\ &\approx \max_{\mathbf{R}_{\mathbf{n}}} \frac{1}{M}\sum_{i=1}^{M} Q\left(\sqrt{\frac{\mathrm{E}^2(r_i|b_i = \mathrm{b}_i)}{\mathrm{Var}(r_i|b_i = \mathrm{b}_i)}}\right) && (17) \end{aligned}$$

with the following constraints

$$\begin{aligned} \frac{1}{L}\sum_{k=1}^{L}\lambda_{n_k} &\leq D_c && (18) \\ \lambda_{n_j} &\leq \frac{L}{M}\cdot D_c, \quad j = 1, \cdots, L && (19) \\ \lambda_{n_j} &\geq 0, \quad j = 1, \cdots, L && (20) \end{aligned}$$

being $D_c \triangleq \frac{1}{L}\sum_{j=1}^{L}(y_j - z_j)^2$ the squared-error due to the attack. The purpose of constraint (19) is to limit the noise power per sample.

Since the attacker knows that the decoder has access to $\mathbf{P}$ and $\mathbf{D}_{\mathbf{S}}$, and he/she assumes the decoder knows his/her attack (conservative point of view), his/her objective can be rewritten as

$$\max_{\mathbf{R}_{\mathbf{n}}} \sum_{k=1}^{M} Q\left(\sqrt{\mathbf{p}_i{}^T\mathbf{D}_{\mathbf{S}}\mathbf{D}_{\alpha}\mathbf{A}^{-1}\mathbf{D}_{\alpha}\mathbf{D}_{\mathbf{S}}\mathbf{p}_i}\right) \tag{21}$$

where the dependence with $\mathbf{R}_{\mathbf{n}}$ appears through (8). The solution of (21) is very involved and requires the use of numerical methods. Moreover, the attacker needs perfect knowledge of $\mathbf{P}$ and $\mathbf{D}_{\mathbf{S}}$, which in a practical secure system will be obviously not the case.

4.2. Perfect knowledge of $\mathbf{P}$, unknown $\mathbf{D_S}$.

The objective of the attacker is still (21), but in this case we must average over the possible spreading sequences to compute $\mathrm{E}\{r_i|b_i = \mathrm{b}_i\}$ and $\mathrm{Var}\{r_i|b_i = \mathrm{b}_i\}$. Therefore,

$$\begin{aligned} \mathrm{E}\{r_i|b_i = \mathrm{b}_i\} &= K\gamma_i b_i, \quad i = 1, \cdots, L && (22)\\ \mathrm{Var}\{r_i|b_i = \mathrm{b}_i\} &= K^2\gamma_i && (23)\end{aligned}$$

where

$$\begin{aligned} \gamma_i &\triangleq \mathrm{E}_{\mathbf{S}}\left\{{\mathbf{p}_i}^T\mathbf{D_S}\mathbf{D}_\alpha(\mathbf{R_X}+\mathbf{R_n})^{-1}\mathbf{D}_\alpha\mathbf{D_S}\mathbf{p}_i\right\}\\ &= {\mathbf{p}_i}^T\mathbf{D}_\alpha\mathrm{diag}\{(\mathbf{R_X}+\mathbf{R_n})^{-1}\}\mathbf{D}_\alpha\mathbf{p}_i && (24)\end{aligned}$$

After substituting (22-23) into (17) it is easy to see that the maximization in (17) is equivalent to minimizing (24). Nonetheless, to the best of our knowledge, a closed-form expression for $\mathbf{R_n}$ minimizing (24) is not possible, so numerical methods are required.

4.3. $\mathbf{P}$ and $\mathbf{D_S}$ unknown

In this case the attacker has no information about the set of coefficients devoted to each bit. This lack of information can be modeled as a random variable. Those sets can be built following a spatial criterion, a frequential one, or merely a equiprobable pseudorandom choice. The latter is the case we will pursue here. Since the construction of these sets is completely random from the attacker's point of view, we can focus our analysis on what would happen with a generic index i, given that all the indices will have the same average performance.

The elements of $\mathbf{P}$, $p_{i,j}$, will be now characterized with i.i.d. random variables (possibly overlapping sets are included in this model). These random variables will take the value 0 with probability $p_0 = \frac{M-1}{M}$ and 1 with $p_1 = \frac{1}{M}$. In this way each bit is hidden in average in the same number of coefficients following a uniform distribution.

From (21), and taking into account that we now deal with an "average bit", the attacker would have to minimize

$$\gamma_i = \mathrm{E}_{\mathbf{P}}\left\{{\mathbf{p}_i}^T\mathbf{D}_\alpha\mathrm{diag}\{(\mathbf{R_X}+\mathbf{R_n})^{-1}\}\mathbf{D}_\alpha\mathbf{p}_i\right\} = \frac{1}{M}\mathrm{trace}\left\{\mathbf{D}_\alpha\mathrm{diag}\{(\mathbf{R_X}+\mathbf{R_n})^{-1}\}\mathbf{D}_\alpha\right\} \quad (25)$$

Again, it is not possible to achieve a closed-form expression for computing the value of $\mathbf{R_n}$ minimizing (25), so general theoretical conclusions cannot be drawn.

5. NEAR-OPTIMAL APPROACH FOR THE DECODER

When writing the expressions for the conditional expectations in (6) and (7) no special structure is assumed for the host covariance matrix. However, it is well-known that host-image statistics typically exhibit a local behavior that can be exploited to arrive at simplified expressions for the decoder. Thus, we will assume that the host coefficients devoted to sending a particular bit are located far enough so as to justify the assumption of statistical independence between any pair of them.

Let N_1 and N_2 denote the respective width and height of the host image, and let $x[k_1, k_2]$ denote the host image coefficient at position $\mathbf{k} \triangleq (k_1, k_2)$. Then, we will consider that the only non-negligible correlations with $x[k_1, k_2]$ are those with those coefficients belonging to a square with side $2R+1$, denoted by $\mathcal{V}_{\mathbf{k}}$, that is,

$$E\{x[k_1, k_2]x[j_1, j_2]\} = 0, \text{ if } |k_1 - j_1| > R, \text{ or } |k_2 - j_2| > R \quad (26)$$

for any $k_1, j_1 \in \{1, \cdots, N_1\}$, $k_2, j_2 \in \{1, \cdots, N_2\}$.

In this setting, (3) can be replaced by

$$r_i = \mathbf{p}_i^T\mathbf{D}_{\mathbf{S}_i}\mathbf{B}_i\mathbf{z}_i \quad (27)$$

where $\mathbf{p}_i$ is the subset of elements of the i-th column of $\mathbf{P}$, corresponding to coefficients in the set $\mathcal{T}_i = \bigcup_{\mathbf{j}\in\mathcal{S}_i} \mathcal{V}_{\mathbf{j}}$, $\mathbf{D}_{\mathbf{S}_i}$ is the diagonal matrix with those values of $\mathbf{D}_{\mathbf{S}}$ corresponding to the coefficients in $\mathcal{T}_i$, B_i is a decoding filter that will be optimized by the decoder for the i-th bit, and $\mathbf{z}_i$ are the values of $\mathbf{z}$ corresponding to the received signal coefficients in $\mathcal{T}_i$ arranged in a vector. Following the assumptions at the beginning of this section, let $\mathbf{x}_i$ be a vector containing those host image coefficients belonging to $\mathcal{V}_{\mathbf{j}}$, for any $\mathbf{j} \in \mathcal{S}_i$, and arranged in such a way that all coefficients in $\mathcal{V}_{\mathbf{j}}$ for a given $\mathbf{j}$ occupy successive positions. Similarly, we define the vector $\mathbf{n}_i$ as that containing those elements of $\mathbf{n}$ belonging to $\mathcal{T}_i$ and arranged in the same order as for $\mathbf{x}_i$. Also, let $\mathbf{R}_{\mathbf{X}_i}$ and $\mathbf{R}_{\mathbf{n}_i}$ denote the respective covariance matrices of $\mathbf{x}_i$ and $\mathbf{n}_i$.

Taking all this into account, we can rewrite (6) and (7) in the following way

$$\begin{aligned} \mathrm{E}\{r_i|b_i = \mathrm{b}_i\} &= \mathbf{p}_i{}^T \mathbf{D}_{\mathbf{S}_i} \mathbf{B}_i \mathbf{H}_i \mathbf{D}_{\alpha_i} \mathbf{D}_{\mathbf{S}_i} \mathbf{p}_i \mathrm{b}_i, \; i = 1, \cdots, M & (28) \\ \mathrm{E}\{r_i^2|b_i = \mathrm{b}_i\} &= \mathbf{p}_i{}^T \mathbf{D}_{\mathbf{S}_i} \mathbf{B}_i \mathbf{H}_i \mathbf{A}_i \mathbf{H}_i^T \mathbf{B}_i^T \mathbf{D}_{\mathbf{S}_i} \mathbf{p}_i, \; i = 1, \cdots, M & (29) \end{aligned}$$

where $\mathbf{D}_{\alpha_i}$ is defined similarly to $\mathbf{D}_{\mathbf{S}_i}$, $\mathbf{H}_i$ are the elements of $\mathbf{H}$ contained in $\mathcal{T}_i$, and $\mathbf{A}_i$ is defined as

$$\mathbf{A}_i \triangleq \mathbf{R}_{\mathbf{X}_i} + \mathbf{R}_{\mathbf{n}_i} \tag{30}$$

It is interesting to note that from the way the elements of $\mathbf{x}_i$ are ordered, it is obvious that $\mathbf{R}_{\mathbf{X}_i}$ will be a block-diagonal matrix (with blocks of size $(2R+1)^2 \times (2R+1)^2$) and $\mathbf{R}_{\mathbf{n}_i}$ will have a similar structure if the noise is locally correlated. We will assume that the sets $\mathcal{S}_j$ are sparse enough, so all coefficients in $\mathcal{T}_i \backslash \mathcal{S}_i$ belong to different sets $\mathcal{S}_k$, with $k \neq i$.

Following a strategy similar to that employed in Sect. 3, it is clear that the best decoding filter for the i-th bit becomes

$$\mathbf{B}_i = K_i \mathbf{D}_{\alpha_i} \mathbf{A}_i^{-1} \mathbf{H}_i^{-1} \tag{31}$$

This is obviously a suboptimal solution, since we are not using all the available information about the covariance matrix. Nevertheless, most of the relevant information for the bits in $\mathcal{S}_j$ is being used, so the results are expected to be quite close to those obtained with the full covariance matrix.

Obviously, this approximation affords a significant reduction in the computational cost required to obtain (31) compared to using the full covariance matrix, as it was done in (15), since now we only have to invert size-$(2R+1)^2 \times (2R+1)^2$ matrices.

6. NEAR-OPTIMAL APPROACH FOR THE ATTACKER

In this section we will analyze the strategy for the attacker when identical considerations as in the previous section are made, including the decoding filter obtained in (31).

6.1. Perfect Knowledge of P, unknown $\mathbf{D}_{\mathbf{S}}$

In this case (22) and (23) are still valid, but the value of γ_i is replaced by

$$\begin{aligned} \gamma_i &= \mathrm{E}_{\mathbf{S}} \left\{ \mathbf{p}_i^T \mathbf{D}_{\mathbf{S}_i} \mathbf{D}_{\alpha_i} \left(\mathbf{R}_{\mathbf{X}_i} + \mathbf{R}_{\mathbf{n}_i} \right)^{-1} \mathbf{D}_{\alpha_i} \mathbf{D}_{\mathbf{S}_i} \mathbf{p}_i \right\} = \mathbf{p}_i^T \mathbf{D}_{\alpha_i} \mathrm{diag} \left[\left(\mathbf{R}_{\mathbf{X}_i} + \mathbf{R}_{\mathbf{n}_i} \right)^{-1} \right] \mathbf{D}_{\alpha_i} \mathbf{p}_i \\ &= \sum_{j\in\mathcal{S}_i} \alpha_j^2 g_{i,j,j}^2 \end{aligned} \tag{32}$$

where $\mathrm{diag}(\mathbf{G}_i)$ is the diagonal matrix containing the diagonal elements of $\mathbf{G}$ and $g_{i,j,k}$ denotes the j,k-th entry of the matrix $\mathbf{G}_i$ defined next

$$\mathbf{G}_i \triangleq \left(\mathbf{R}_{\mathbf{X}_i} + \mathbf{R}_{\mathbf{n}_i} \right)^{-1} \tag{33}$$

Expression (32) can be rewritten as

$$\gamma_i = (1, \cdots, 1) \mathbf{\Delta}_i \mathrm{diag} \left[\mathbf{G}_i \right] \mathbf{\Delta}_i (1, \cdots, 1)^T \tag{34}$$

where $\mathbf{\Delta}_i$ is a matrix with size $(|\mathcal{S}_i| \cdot (2R+1)^2) \times (|\mathcal{S}_i| \cdot (2R+1)^2)$ and whose j,k-th element is given by

$$\mathbf{\Delta}_{i,j,k} = \begin{cases} \alpha_j & \text{if } j = k \in \mathcal{S}_i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } 1 \leq j,k \leq |\mathcal{S}_i| \cdot (2R+1)^2 \tag{35}$$

Therefore,

$$\gamma_i = \text{trace}\left[\mathbf{\Delta}_i \mathbf{G}_i \mathbf{\Delta}_i\right] \tag{36}$$

Minimizing (36) is again equivalent to the maximization in (17). Numerical methods must be as well applied to compute the value of $\mathbf{R}_{\mathbf{n}_i}$ which maximizes the probability of error P_e.

6.2. P and $\mathbf{D_S}$ unknown

Following the characteristics of $\mathbf{P}$ described in Sect. 4.3, and averaging over all possible partitions, we can write γ_i as

$$\gamma_i = \text{E}_{\mathbf{P}}\left\{\mathbf{p}_i^T \mathbf{D}_{\alpha_i} \text{diag}\left[\mathbf{G}_i\right] \mathbf{D}_{\alpha_i} \mathbf{p}_i\right\} \tag{37}$$

It is important to take into account that all the matrices and the vector $\mathbf{p}_i$ in (37) have $|\mathcal{S}_i| \cdot (2R+1)^2$ dimensions.

A great simplification of (37) is afforded when $\mathbf{R_X}$ is diagonal (the samples of the original host signal are uncorrelated) and $\mathbf{R_n}$ is restricted to have a similar structure. In this case, the problem can be solved by means of Lagrange's multipliers, using the following functional

$$\varphi = \frac{1}{M} \sum_{k=1}^{L} \frac{\alpha_k^2}{\sigma_{x_{k,k}}^2 + \sigma_{n_{k,k}}^2} + \lambda \left(\sum_{k=1}^{L} \sigma_{n_{k,k}}^2 - L \cdot D_c \right) \tag{38}$$

where $\sigma_{x_{k,k}}^2 = \text{E}\{x^2[k]\}$ and a similar definition applies to $\sigma_{n_{k,k}}^2$. Differentiating (38) with respect to $\sigma_{n_{j,j}}^2$ and equation to zero, we can write

$$\frac{\partial \varphi}{\partial \sigma_{n_{j,j}}^2} = -\frac{\alpha_j^2}{M\left(\sigma_{x_{k,k}}^2 + \sigma_{n_{k,k}}^2\right)^2} + \lambda = 0 \tag{39}$$

whose solution is given by

$$\sigma_{n_{j,j}}^2 = \xi \alpha_j - \sigma_{x_{j,j}}^2 \tag{40}$$

for some ξ is discussed below. Taking into account the constraints (18-20), we have a waterfilling solution, with the form

$$\sigma_{n_{j,j}}^2 = \min\left[\max\left(\xi \alpha_j - \sigma_{x_{j,j}}^2, 0\right), \frac{L}{M} D_c\right] \tag{41}$$

being ξ a suitably chosen parameter so that (41) verifies (18).

6.3. Discussion

Moulin and Ivanović have recently studied[4] a problem whose formulation is similar to the one in Sect.6.2. The main differences between this and the approach in[4] are:

- In[4] the analysis focuses on the game between embedder and attacker, while we consider the attacker and decoder as the players in the game.
- The decoder in[4] implements the optimal maximum *a posteriori* decision rule. This explains why the decoder is left out of the game.

- In contrast, we have considered that the embedder uses all the allowed amplitude modification leading to an unnoticeable distortion, as given by the perceptual mask. This is the reason why the embedder is left out of the game.

Although the approach to the problems in these two papers is quite different, for theoretical insight diagonal host and noise covariance matrices must be assumed in both cases. Moreover, in Moulin and Ivanovicś work, both embedding and attacking filters have to be replaced by identity matrices.

The result given in[4] is

$$\sigma^2_{n_{j,j}} = \begin{cases} \xi\alpha_j - \sigma^2_{x_{j,j}}, & \text{if } \frac{\sigma^2_{x_{j,j}}}{\alpha_j} \leq \xi \\ 0 & \text{elsewhere} \end{cases} \quad \text{for } 1 \leq j \leq L \tag{42}$$

where ξ is the unique solution to the nonlinear equation (obtained after taking into account a missing term)

$$\sum_{j=1}^{L} \left\{ \max\left[\xi\alpha_j - \sigma^2_{x_{j,j}}, 0\right] + \alpha_j^2 \right\} = D_{xz} \tag{43}$$

where D_{xz} is the mean squared error between $\mathbf{x}$ and $\mathbf{z}$.

Even though our solution and (42) are obviously different (due to the different assumptions), it is remarkable that in both cases they have a similar structure, based on the waterfilling principle.

7. A POTENTIAL FURTHER SIMPLIFICATION

In this section we discuss a potential further simplification over the development in Sect. 6, and discuss its problems.

After a careful look at (30), one might think of replacing $\mathbf{A}_i$ by $\mathbf{R}_{\mathbf{X}_i}$, since the dominant contribution in $\mathbf{A}_i$ is clearly that due to the host covariance matrix. In this case, the decoding filter in (31) would become $\mathbf{B}_i = K_i \mathbf{D}_{\alpha_i} \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{H}_i^{-1}$. This simplification would lead to the following expressions

$$\begin{aligned} \mathrm{E}(r_i | b_i = \mathrm{b}_i) &= \mathbf{p}_i{}^T \mathbf{D}_{\mathbf{S}_i} \mathbf{D}_{\alpha_i} \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{D}_{\alpha_i} \mathbf{D}_{\mathbf{S}_i} \mathbf{p}_i \mathrm{b}_i, \quad i = 1, \cdots, M &(44)\\ \mathrm{E}(r_i^2 | b_i = \mathrm{b}_i) &= \mathbf{p}_i{}^T \mathbf{D}_{\mathbf{S}_i} \mathbf{D}_{\alpha_i} \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{A}_i \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{D}_{\alpha_i} \mathbf{D}_{\mathbf{S}_i} \mathbf{p}_i, \quad i = 1, \cdots, M &(45) \end{aligned}$$

so the objective of the attacker would be to maximize $\mathrm{E}(r_i^2 | b_i = \mathrm{b}_i)$, or equivalently,

$$\max_{\mathbf{R}_{\mathbf{n}_i}} \mathbf{p}_i{}^T \mathbf{D}_{\mathbf{S}_i} \mathbf{D}_{\alpha_i} \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{R}_{\mathbf{n}_i} \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{D}_{\alpha_i} \mathbf{D}_{\mathbf{S}_i} \mathbf{p_i} = \max_{\mathbf{R}_{\mathbf{n}_i}} \operatorname{trace}\left(\boldsymbol{\Delta}_i \mathbf{R}^{-1}_{\mathbf{X}_i} \mathbf{R}_{\mathbf{n}_i} \mathbf{R}^{-1}_{\mathbf{X}_i} \boldsymbol{\Delta}_i \right), \quad i = 1, \cdots, M \tag{46}$$

Let $\mathbf{Q} = \mathbf{R}^{-1}_{\mathbf{X}_i}$, then we can write (46) as

$$\max_{\mathbf{R}_{\mathbf{n}_i}} \sum_{j \in \mathcal{S}_i} \alpha_j^2 \sum_{k \in \mathcal{V}_j} \sum_{l \in \mathcal{V}_j} q_{j,k} q_{j,l} R_{n_{k,l}} \tag{47}$$

Unfortunately, it is difficult to introduce in this expression the constraints for the positive-definiteness of $\mathbf{R}_{\mathbf{n}}$. However, if we further restrict $\mathbf{R}_{\mathbf{n}}$ to be diagonal, we have

$$\max_{\mathbf{R}_{\mathbf{n}}} \sum_{j \in \mathcal{S}_i} \alpha_j^2 \sum_{k \in \mathcal{V}_j} q_{j,k}^2 R_{n_{k,k}} \tag{48}$$

so the solution would be to allocate all the available power to that coefficient k having the largest value of $\sum_{j \in \mathcal{S}_i} \alpha_j^2 \sum_{k \in \mathcal{V}_j} q_{j,k}^2$. The explanation for this outcome is simple: were a similar attack be used in Sects. 4 to 6, the decoding filter would reject the sample that monopolizes all the available noise power, so this would not be a profitable strategy for the attacker. In contrast, the filter $\mathbf{B}_i$ that has been used throughout this section does not depend on $\mathbf{R}_{\mathbf{n}}$, a fact that can be used by the attacker to his/her advantage.

8. EXPERIMENTAL RESULTS

In this section we quantify the performance loss (as measured by the bit error rate) that results when the attacker has full knowledge of the partitions that define the sets of coefficients devoted to hiding each bit.

First, we have chosen a toy model to exemplify this loss when $\mathbf{R_X}$ is not diagonal. We have created a 16×16 synthetic image, for which the covariance at a given pixel is non-zero only for the nearest neighboring pixels. For each pixel, both the variance and the perceptual mask are chosen at random. We have computed the noise having diagonal covariance matrix which maximizes P_e for both cases where the partition matrix $\mathbf{P}$ is known and unknown. Even in this too simple case the computational resources required by the attacker to obtain the optimal attack are very expensive. While the objective function is not too difficult to compute, a number of evaluations must be performed in order to achieve the final result. For this reason, a more realistic example has not being implemented for the time being.

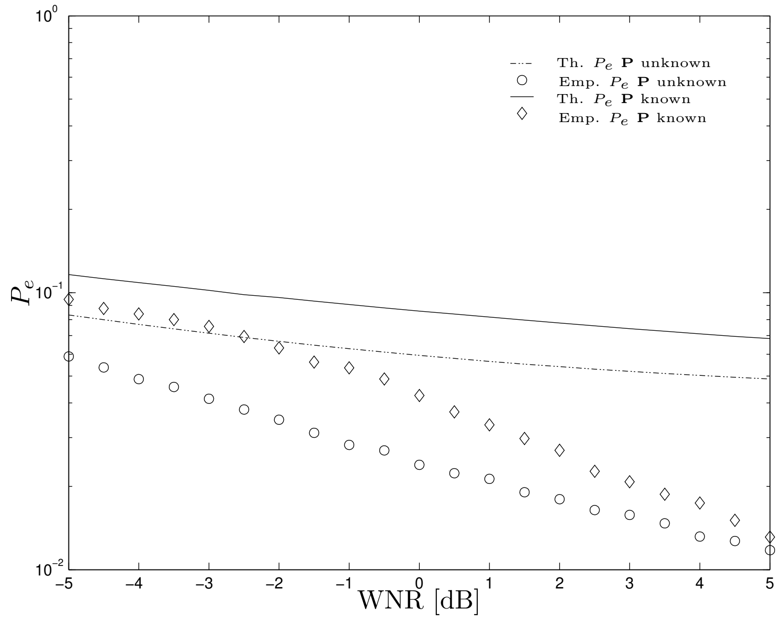

Figure 2. Theoretical and empirical P_e vs. WNR when both $\mathbf{P}$ is known and unknown. $M = 8$.

In fact, the number of samples of the *host image* is so small that the discrepancy between the theoretical and empirical results shown in Fig. 2 is partly justified. In any case, the values of P_e when $\mathbf{P}$ is known to the attacker are clearly above those obtained when $\mathbf{P}$ is not known.

It is important to remark that considering a diagonal covariance matrix implies also a simplification. To get the really optimal attack, we would have to consider the whole covariance matrix in the optimization, although this seems to be clearly impractical. Nevertheless, other simplifications than a diagonal covariance matrix could have been done. For example, one could think of optimizing a covariance matrix kernel which would be replicated all over the image. This is equivalent to constraining to a wide-sense stationary attack. Such model could be further improved by taking into account the restrictions given by the perceptual mask.

We have also considered the case of watermarking in the DCT domain in Fig. 3. In this situation it is well known that the samples can be assumed independent, so the covariance matrix $\mathbf{R_X}$ is approximately diagonal. We have watermarked the mid-frequency coefficients and in the optimization of $\mathbf{R_n}$ we have constrained it to be diagonal too. Since the power distribution for a given set is computed following (41), it is only necessary to optimize the power devoted to each $\mathcal{S}_i$. Fig. 3 shows that P_e is almost linear with the WNR and how the gap between the cases where $\mathbf{P}$ is and is not known increases with the WNR.

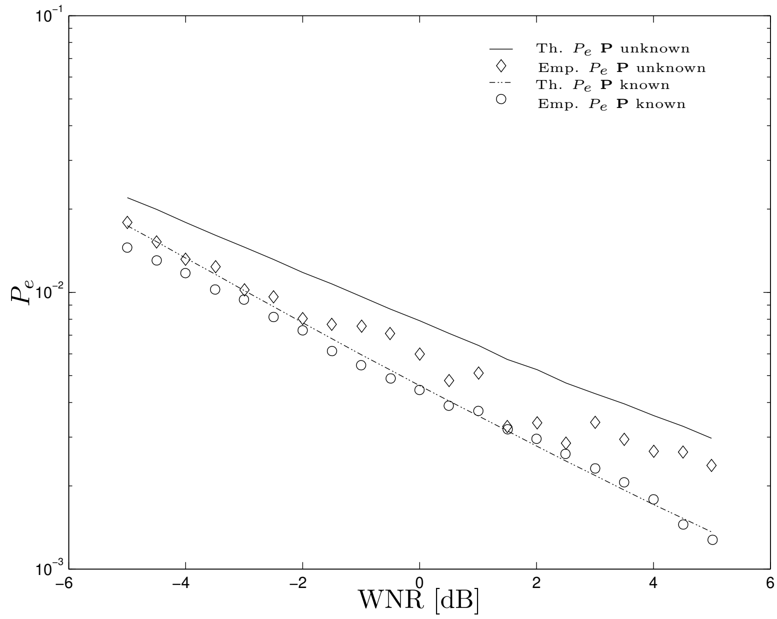

Figure 3. Theoretical and empirical P_e vs. WNR when both $\mathbf{P}$ is known and unknown, watermaring in the DCT domain. $L/M = 50$.

9. FINAL REMARKS

In this paper we have shown preliminary results discussing what should be the optimal strategies for the attacker and the decoder when the former has some extra knowledge about the sets of coefficients devoted to convey each bit. It was shown that for the non-diagonal case the computation of the optimal attacking strategy is not a trivial task; whereas the decoder will have to compute only once the inverse of $(\mathbf{R_X} + \mathbf{R_n})$ to get the optimal decoding pre-filter, the attacker will have to calculate it a number of times in order to find the optimal $\mathbf{R_n}$. Even when this matrix is constrained to have a specific structure this operation is computationally very expensive.

A different situation arises when the watermark is introduced in the DCT domain, since there $\mathbf{R_X}$ can be assumed to take diagonal form. Thus, constraining $\mathbf{R_n}$ to have the same diagonal structure leads to an easy-to-compute attacking strategy.

REFERENCES

1. A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE Transactions on Information Theory* **48**, pp. 1639–1667, June 2002.
2. J. J. Eggers and B. Girod, *Informed Watermarking*, Kluwer Academic Publishers, 2002.
3. P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. on Information Theory* **49**, pp. 563–593, March 2003.
4. P. Moulin and A. Ivanovic, "The zero-rate spread-spectrum watermarking game," *IEEE Trans. on Signal Processing* **51**, pp. 1098–1117, April 2003.
5. A. B. Watson, "Dct quantization matrices visually optimized for individual images," in *Proceedings of the SPIE*, 1993. in Human Vision, Visual Processing and Digital Display III.
6. A. J. A. Jr and H. A. Peterson, "Luminance-model-based dct quantization for color image compression," in *Proceedings of the SPIE*, B. E. Rogowitz, ed., 1992. in Human Vision, Visual Processing and Digital Display IV.

7. J. R. Hernández and F. Pérez-González, "Statistical analysis of watermarking schemes for copyright protection of images," *Proceedings of the IEEE* **87**, pp. 1142–1166, July 1999. Special Issue on Identification and Protection of Multimedia Information.