DITHER-MODULATION DATA HIDING WITH DISTORTION-COMPENSATION: EXACT PERFORMANCE ANALYSIS AND AN IMPROVED DETECTOR FOR JPEG ATTACKS

Fernando Pérez-González, Pedro Comesaña and Félix Balado

Dept. Tecnologías de las Comunicaciones. ETSI Telecom., Universidad de Vigo, 36200 Vigo, Spain
email: fperez@tsc.uvigo.es, pcomesan@gts.tsc.uvigo.es, fiz@tsc.uvigo.es

ABSTRACT

The binary Distortion Compensated Dither-Modulation (DC-DM), which can be regarded to as a baseline for quantization-based data-hiding methods, is rigorously analyzed. A novel and accurate procedure for computing the exact probability of bit error is given, as well as an approximation amenable to differentiation which allows to obtain the optimal weights in a newly proposed decoding structure, for significant improvements on performance. The results are particularized for a JPEG compression scenario which allows to show their usefulness. Experimental results validating the proposed theory are presented.

1. INTRODUCTION.

Although quantization-based methods have been presented since the beginnings of watermarking, it was not until very recently that the idea was revisited from a sound theoretical perspective in the form of a data hiding scheme known as Quantization Index Modulation (QIM) [1], which hides information by constructing a data-driven set of quantizers. This was later connected to an old paper by Costa [2] to realize that by adding back a fraction of the quantization error, performance could be significantly improved. This scheme was thus termed Distortion Compensated QIM (DC-QIM).

The original proposal of DC-QIM can be adapted for using many off-the-shelf vector quantizers. In particular, a considerable attention has been paid to the special case called Dither Modulation (DM) [1] —or formally equivalent schemes [3], [4]—, which has the advantage of its simplicity. Here we consider a (binary) multidimensional extension of Distortion Compensated DM (DC-DM) which can be regarded as a baseline for more sophisticated QIM schemes [1]. Surprisingly, even for this very simple scheme a thorough performance analysis (e.g. measured in terms of the bit error rate, BER) lacks in the literature and only very rough—and thus of little use—approximations based on the so-called "union bound" are known [1].

The main contributions of the present paper, all focused on the binary DC-DM method and, to the authors' knowledge, novel, are: 1) to provide a procedure for the *exact* computation of the BER that improves on the method published in [5] (and which was only an upper bound); 2) to give an approximation for the BER, more accurate than the union-bound and amenable to analytical optimization; 3) to extend the performance analysis to noise correlated to the host image, in particular, to JPEG "attacks"; 4) to significantly improve the hidden information decoder by using a weighted Euclidean distance; and 5) to give an analytical expression for the optimal weights, which produce considerable gains when used on typical images.

Throughout the paper we will assume that the host image coefficients are arranged in a vector $\boldsymbol{x}$, so that the watermarked image can be written as $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}$, being $\boldsymbol{w}$ the watermark. The information to be hidden is represented by a vector $\boldsymbol{b}$ with N binary antipodal components, i.e., $b_j = \pm 1, j = 1, \cdots, N$. Following most existing schemes, we will consider that the $j-$th bit is hidden in a key-dependent set of coefficients $\mathcal{S}_j$ with cardinality L_j, for all $j = 1, \cdots, N$. The total set of coefficients devoted to data hiding is denoted by $\mathcal{S} \triangleq \bigcup_{i=1}^{N} \mathcal{S}_i$. For convenience, $\boldsymbol{w}_j$ stands for the vector comprising those samples with indices belonging to $\mathcal{S}_j$. We will also assume that prior to decoding the watermarked image is sent through an additive probabilistic noise channel, so that the image at its output $\boldsymbol{z}$ can be written as $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{n} = \boldsymbol{x} + \boldsymbol{w} + \boldsymbol{n}$, where $\boldsymbol{n}$ is the noise vector. By virtue of the pseudorandom choice of the indices in $\mathcal{S}$ we may assume that the samples in $\boldsymbol{n}$ are also mutually independent, with zero mean and variances $\sigma_{n_i}^2, i \in \mathcal{S}$.

To measure the impact of the attack, we will follow the popular *watermark-to-noise* ratio (WNR), defined as $\text{WNR} \triangleq 10 \log_{10} \sum_{i \in \mathcal{S}} \text{E}\{w_i^2\} / \sum_{i \in \mathcal{S}} \sigma_{n_i}^2$.

2. BASIC CONCEPTS OF DC-DM.

Structured quantization-based methods ([3], [1]) hide information by constructing a set of vector quantizers $\boldsymbol{Q}_{\boldsymbol{b}}(\cdot)$, each representing a different codeword $\boldsymbol{b}$. So, given a host vector $\boldsymbol{x}$ and an information codeword $\boldsymbol{b}$, the embedder constructs the watermarked vector $\boldsymbol{y}$ by simply quantizing $\boldsymbol{x}$ with $\boldsymbol{Q}_{\boldsymbol{b}}(\cdot)$, i.e. $\boldsymbol{y} = \boldsymbol{Q}_{\boldsymbol{b}}(\boldsymbol{x})$.

Here we will analyze the simplest (and most studied) implementation of these methods, the binary Distortion Compensated Dither Modulation (DC-DM) [1]. In the binary DC-DM the watermark samples in the set $\mathcal{S}_j, j = 1 \cdots, N$,

This work was partially funded by the *Xunta de Galicia* under projects PGIDT01 PX132204PM and PGIDT02 PXIC32205PN, and the CYCIT project AMULET, reference TIC2001-3697-C03-01.

are given by $\boldsymbol{w}_j = \nu_j \boldsymbol{e}_j$, i.e. the L-dimensional quantization error $\boldsymbol{e}_j \triangleq \boldsymbol{Q}_{b_j}(\boldsymbol{x}_j) - \boldsymbol{x}_j$, weighted by an optimizable distortion-compensating parameter ν_j, $0 < \nu_j \leq 1$. Then, we will have $\boldsymbol{y}_j = \boldsymbol{Q}_{b_j}(\boldsymbol{x}_j) - (1-\nu_j)\boldsymbol{e}_j, \;\; j = 1, \cdots, N$

The uniform quantizers $\boldsymbol{Q}_{-1}(\cdot)$ and $\boldsymbol{Q}_{+1}(\cdot)$ are such that the corresponding centroids are the points in the lattices

$$\begin{aligned} \Lambda_{-1} &= 2(\Delta_1\mathbb{Z}, \cdots, \Delta_L\mathbb{Z})^T + \boldsymbol{d} \\ \Lambda_{+1} &= 2(\Delta_1\mathbb{Z}, \cdots, \Delta_L\mathbb{Z})^T + (\Delta_1, \cdots, \Delta_L)^T + \boldsymbol{d} \end{aligned} \quad (1)$$

with $\boldsymbol{d} \in \mathbb{R}^L$ a key-dependent dithering vector. Note that, in contrast to [1], our setup allows for different quantization steps to be used in each dimension to better account for perceptual constraints.

If the quantization step in each dimension is small enough, we can consider that the quantization error e_i in each dimension will be uniformly distributed between $[-\Delta_i, \Delta_i)$, being $2\Delta_i$ the quantization step. Thus, the embedding distortion in each dimension will be $\mathrm{E}\{w_i^2\} = \nu_j^2\Delta_i^2/3$.

Finally, decoding is implemented as

$$\begin{aligned} \hat{b}_j &= \arg\min_{-1,1} \left\{ \left(\boldsymbol{z}_j - \boldsymbol{Q}_{b_j}(\boldsymbol{z}_j)\right)^t \boldsymbol{B}_j \left(\boldsymbol{z}_j - \boldsymbol{Q}_{b_j}(\boldsymbol{z}_j)\right) \right\}, \\ & \quad j = 1, \cdots, N. \end{aligned} \quad (2)$$

where $\boldsymbol{B}_j = \mathrm{diag}(\beta_{j_1}/\Delta_{j_1}^2, \cdots, \beta_{j_L}/\Delta_{j_{L_j}}^2)$ and $\mathcal{S}_j = \{j_1, \cdots, j_{L_j}\}$. These weighting vectors $\boldsymbol{\beta}_j$ allow to improve decoding when additional information about the noise pdf is available, as we will confirm in Section 5. On the other hand, the normalization by Δ_i in the i-th dimension is reasonable if one thinks that noise variance will be roughly proportional to Δ_i^2 to reduce the perceptual impact of the attack. For simplicity, in next section we will analyze the case in which no weights other than the normalization by Δ_i are used, that is, $\boldsymbol{\beta}_i = 1$. The analysis given here can be readily extended for an arbitrary weights vector.

3. PERFORMANCE ANALYSIS AND NUMERICAL COMPUTATION.

To analyze the performance of this scheme in terms of the bit error probability (P_e), we will define

$$\begin{aligned} u_i &\triangleq z_i - Q_{b_j}(z_i) \\ &= Q_{b_j}(x_i) - (1-\nu_j)e_i + n_j - Q_{b_j}(z_i) \\ &= 2l\Delta_i - (1-\nu_j)e_i + n_i \end{aligned} \quad (3)$$

for all $i \in \mathcal{S}$ and some integer l. Since u_i is a quantization error generated by a quantizer of step size $2\Delta_i$, then u_i must belong to $[\Delta_i, \Delta_i)$, and l in (3) takes the appropriate value so that this is accomplished. Consequently, the pdf of u_i can be written as

$$f_{u_i}(u_i) = \begin{cases} \sum_{l=-\infty}^{\infty} f_{u_i'}(u_i - 2l\Delta_i), & u_i \in [-\Delta_i, \Delta_i) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $u_i' \triangleq n_i - (1-\nu_j)e_i$, is a random variable with pdf

$$f_{u_i'}(u_i') = f_{n_i}(u_i') * \frac{1}{(1-\nu_j)} f_{e_i}(u_i'/(1-\nu_j)) \quad (5)$$

Alternatively, we can write $f_{u_i}(u_i) = f_{u_i''}(u_i + \Delta_i)$, where

$$f_{u_i''}(u_i'') = f_{n_i}(u_i'' - \Delta_i) \circledast_{2\Delta_i} \frac{f_{e_i}(u_i''/(1-\nu_j))}{(1-\nu_j)} \quad (6)$$

being $\circledast_{2\Delta_i}$ the circular convolution of size $2\Delta_i$ operator. This circular convolution includes the *aliasing* effect which is evident in (4); also, the shifts of size Δ_i in $f_{u''}$ and f_n are due to the fact that the circular convolution is defined in $[0, 2\Delta_i)$, while we are interested in $[-\Delta_i, \Delta_i)$. A similar technique has been used in [3], where the role of the circular convolution is played by the sampling of the characteristic function with period $\frac{2\pi}{2\Delta}$, which is known to have an aliasing effect, since it is equivalent to the convolution in the time domain with an impulse train with period 2Δ.

To follow a strategy similar to the one described in [5] and given (2) with $\boldsymbol{\beta}_i = 1$, we will define $\boldsymbol{v}$ as the vector with components $v_i \triangleq u_i/\Delta_i$, $i = 1, \cdots, L_j$, so we can write the bit error probability for the j-th hidden bit as

$$\begin{aligned} P_e(j) &= P\{\|\boldsymbol{v}_j'\|^2 > \|\boldsymbol{v}_j' - (1, \cdots, 1)^T\|^2\} \\ &= P\left\{ \sum_{i \in \mathcal{S}_j} v_i' > L_j/2 \right\}, \end{aligned} \quad (7)$$

where $\boldsymbol{v}'$ is an auxiliary random vector with independent components such that $\boldsymbol{v}' \triangleq |\boldsymbol{v}|$ with pdf (assuming that v_i is symmetric)

$$f_{v_i'}(v_i') \triangleq \begin{cases} 2\Delta_i f_{u_i}(v_i'\Delta_i), & \text{if } 0 \leq v_i' \leq 1 \\ 0, & \text{otherwise} \end{cases}, i \in \mathcal{S}_j \quad (8)$$

If we define $r_j \triangleq \sum_{i \in \mathcal{S}_j} v_i'$, then the computation of $P_e(j)$ is equivalent to integrating the tail of the pdf of r_j from $L_j/2$, but since the v_i' are independent random variables, the pdf of r_j is just the convolution of the pdf's of v_i', $i \in \mathcal{S}_j$. An efficient way to compute it is with the DFT. To that end, let $\boldsymbol{\Phi}_{v_i'} \triangleq \mathrm{DFT}_{L_jT}\left(f_{v_i'}(t/T)\right)$ be the $L_j \cdot T$-point DFT of the sequence obtained by sampling $f_{v_i'}$ at $\frac{t}{T}$, $t \in \{0, \cdots, T-1\}$. From this, it is straightforward to write

$$\Phi_{r_j}(l) = \prod_{i \in \mathcal{S}_j} \Phi_{v_i'}(l), \quad (9)$$

for $l = 1, \cdots, L_jT - 1$.

Finally, let $f_{r_j}[r_j] = \mathrm{IDFT}_{L_jT}(\boldsymbol{\Phi}_{r_j})$, then $P_e(j)$ can be computed as

$$P_e(j) = \sum_{k=\left\lceil \frac{L_j(T-1)+1}{2} \right\rceil}^{L_jT} f_{r_j}[k], \quad \text{for all } j \in \{1, \cdots, N\}. \quad (10)$$

where the lower index in the summation can be seen as corresponding to $L_j/2$ in (7).

4. JPEG COMPRESSION.

In the previous development we have assumed that the noise $\boldsymbol{n}$ is independent of $\boldsymbol{x}$. This is clearly not the case if the

attack is a coarse quantization, like the popular JPEG compression, which is supposed to be one of the most likely unintentional attacks. In this section we develop a method for estimating P_e for a given quality factor.

Assuming the bits to transmit are equiprobable, and due to the symmetry of the JPEG compression, we will concentrate in the case when $b = -1$, without loss of generality. Given a JPEG quantization step δ_i corresponding to the i-th dimension, we are interested in computing the probability associated to each JPEG centroid, noting that this probability will depend not only on the pdf of the host image (here assumed to be Laplacian with parameter λ) but also on that of the watermark. To that end, we have to determine the limits of each quantization bin; the DC-DM centroid associated to the k-th JPEG bin with limits $a_{i_k}^{\pm} = k\delta_i \pm \delta_i/2$ (the upper or lower limit, depending of the sign, in the i-th dimension) is

$$Q_{-1}(a_{i_k}^{\pm}) = d_i + 2\Delta_i \cdot \text{round}\left(\frac{a_{i_k}^{\pm} - d_i}{2\Delta_i}\right) \tag{11}$$

so the offset between the JPEG centroid and the DC-DM centroid is $e_y(a_{i_k}^{\pm}) \triangleq a_{i_k}^{\pm} - Q_{-1}(a_{i_k}^{\pm})$. This offset corresponds to the watermarked image and it can be shown to map back into the host image as

$$e_x(a_{i_k}^{\pm}) = \frac{\min\{\max[e_y(a_{i_k}^{\pm}), -(1-\nu_j)\Delta_i], (1-\nu_j)\Delta_i\}}{(1-\nu_j)},$$

for all $i \in \mathcal{S}_j$, $j = 1, \cdots, N$. Therefore, if we define $\gamma_{j_k}^{\pm} \triangleq Q(a_{j_k}^{\pm}) + e_x(a_{j_k}^{\pm})$, we can see that it corresponds to the upper (lower) limit of the JPEG quantization bin for the k-th JPEG centroid in the i-th dimension of the host image. Now we can compute the probability of occurrence for this centroid as $P_{i_k} = P(x_i \leq \gamma_{i_k}^{+}) - P(x_i \leq \gamma_{i_k}^{-})$ with

$$P(x_i \leq \tau) = \begin{cases} \frac{1}{2}e^{\lambda_i \tau}, & \text{if } \tau \leq 0 \\ 1 - \frac{1}{2}e^{-\lambda_i \tau}, & \text{if } \tau > 0 \end{cases} \tag{12}$$

The parameter λ_j for the Laplacian distribution can be estimated using the maximum likelihood criterion.

Notice that P_{i_k} play the same role as $f_{u_i'}$ in 4. Once we know the probability of a representative set of JPEG centroids, it is necessary to "fold" them onto the interval $[-\Delta_i, \Delta_i)$ as in (4), and then follow a similar strategy to that developed in the previous section to compute $P_e(j)$ by convolving the pdf's of the JPEG centroids in each dimension. In our practical implementation we have used the DFT method introduced in Section 3.

5. OPTIMAL DECODING WEIGHTS.

In the general formulation of Sect. (2) we considered the possibility of computing an Euclidean distance weighted by a vector $\boldsymbol{\beta}_j$. In this Section we will show how these optimal decoding weights can be determined and which kind of knowledge about the noise pdf is required. Following the development in Section 3, it is easy to show that (7) now becomes

$$P_e(j) = P\left\{\sum_{i\in\mathcal{S}_j} \beta_i v_i' > \frac{1}{2}\sum_{i\in\mathcal{S}_j}\beta_i\right\} \tag{13}$$

The random variables v_k' adding up in the leftmost sum in (13) are independent and similarly distributed (it is not necessary that they are identically distributed, but they must not be too different). Then, when the cardinality L_j of each subset $\mathcal{S}_j$ is large enough and under some additional conditions discussed in [5], it is possible to resort to the Central Limit Theorem (CLT), to write that

$$P_e(j) \approx \mathcal{Q}\left(\frac{\frac{1}{2}\sum_{k\in\mathcal{S}_j}\beta_k - \sum_{k\in\mathcal{S}_j}\beta_k \text{E}\{v_k'\}}{\sqrt{\sum_{k\in\mathcal{S}_j}\beta_k^2 \text{Var}\{v_k'\}}}\right) \tag{14}$$

where $\mathcal{Q}(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\frac{\tau^2}{2}} d\tau$. Recalling that the $\mathcal{Q}(\cdot)$ function is monotonically decreasing, it follows that $P_e(j)$ is minimized when its argument, that we will call in short SNR_j, is maximized. Then, the optimal decoding weights can be found by differentiating the argument of $\mathcal{Q}(\cdot)$ in (14) with respect to $\beta_i, i \in \mathcal{S}_j$, $j = 1, \cdots, N$:

$$\begin{aligned}\frac{\partial \text{SNR}_j}{\partial \beta_i} &= \rho_j\left(\frac{1}{2} - \text{E}\{v_i'\}\right) - \frac{\beta_i \text{Var}\{v_i'\}}{\rho_j} \\ &\cdot \left(\frac{1}{2}\sum_{k\in\mathcal{S}_j}\beta_k - \sum_{k\in\mathcal{S}_j}\beta_k \text{E}\{v_k'\}\right)\end{aligned} \tag{15}$$

where ρ_j is an adequate constant. Setting (15) to zero and operating we find that the optimal decoding weights β_i^*, for all $i \in \mathcal{S}$ can be written as

$$\beta_i^* = K \cdot \frac{\left(\frac{1}{2} - \text{E}\{v_i'\}\right)}{\text{Var}\{v_i'\}} \tag{16}$$

where K is an irrelevant positive real constant, since $\boldsymbol{\beta}^*$ can be scaled without any impact on performance. Also, it is very interesting to note some of the β_i^* may be negative. This will happen when the random variable v_i' is such that $\text{E}\{v_i'\} > 1/2$, which may occur for large distortions.

Finally, as it can be inferred from (16), in order to compute the optimal decoding weights, knowledge of $\text{E}\{v_i'\}$ and $\text{Var}\{v_i'\}$ is required. In the case of JPEG compression, since the quantization table is generally available to the decoder, which will then be able to compute the joint pdf of $\boldsymbol{v}'$ using the procedure outlined in Section 4, and, consequently, to derive the optimal decoding weights for minimum BER data extraction.

6. EXPERIMENTAL RESULTS.

In order to validate the analytical results presented heretofore, we have watermarked the image *Lena* with size 256×256 in the DCT domain taking into account the same perceptual properties as in previous works and we have represented the BER vs. WNR curves for different noise distributions and decoders. First of all (Fig. 1), we have studied

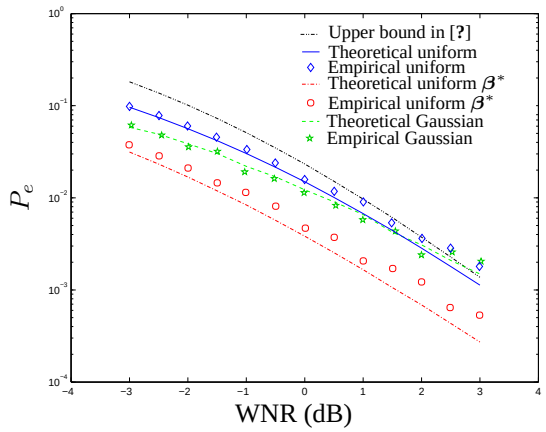

Fig. 1. BER versus WNR for DC-DM ($L = 10$ and $\nu = 0.5$) with additive noise proportional to JPEG with QF = 85.

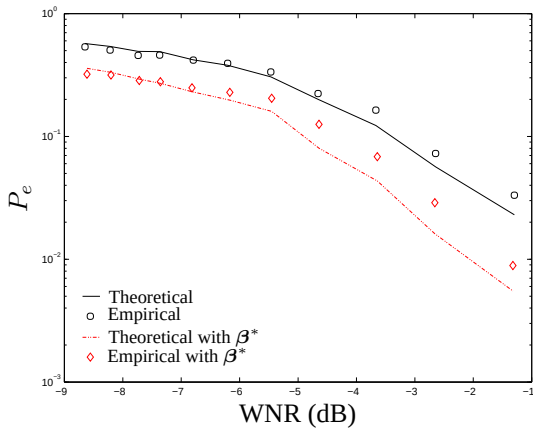

Fig. 2. BER versus WNR for DC-DM ($L = 20$ and $\nu = 0.5$) with JPEG compression with QF between 70 and 90.

additive noise attacks, with both uniform and Gaussian distributions, and with variances which depend for each DCT coefficient on the corresponding squared quantization step used in JPEG compression with a quality factor (QF) of 85. The resulting noise is scaled in order to work with different WNR operating points. The theoretical curves for the uniform case correspond to using (14), while for Gaussian noise the DFT technique explained in Sect. 3 was employed. Fig. 1 also plots the upper bound previously published in [5]. For the uniform case, Fig. 1 clearly shows the improvement on performance that results when the procedure in Sect. 5 is used. The slight difference between empirical and theoretical results is due to the non-uniformity of the image within the quantization step. This also explains why that difference is larger when the WNR increases.

Finally, in Fig. 2 we depict the BER vs. WNR when the same image is compressed with QF's ranging from 70 to 90, comparing the empirical results with the analytical ones obtained by following the procedure described in 4. The results obtained show that binary DC-DM performs poorly in front of JPEG compressions. Note that this fact is already very accurately predicted by our theory. Moreover, without using the optimal weights an improvement does not follow by incrementing the size L of the partitions devoted to a particular bit; in fact if $\mathrm{E}\{v_i'\} > \frac{1}{2}$ and $\beta_i = 1$, for all $i \in S_j$, P_e will increase with L . This is a quite remarkable result which basically implies that although the multidimensional extension of the scalar DC-DM scheme can be regarded to as a repetition code, for small WNR's (high values of $\mathrm{E}\{v_i'\}$) there might be no advantage in using an unweighted decoder, in evident contrast to what happens in spread-spectrum and quantized-projection data hiding [6]. Nevertheless if the optimal weights are used we do obtain such an improvement. The explanation is clear: in that case we are using information about the distribution of v_i', so the argument of (14) will increase with L and therefore P_e will decrease. Finally, in Fig. 2 we have also represented the empirical and theoretical results obtained when the optimal decoding weights are used, to demonstrate how the BER can be reduced by following this strategy.

7. CONCLUSIONS

In this paper we have presented a theoretical analysis for the binary DC-DM data hiding method, which can be considered as a reference for other more sophisticated quantization-based schemes. An accurate analysis was lacking in the data hiding literature and only rough upper bounds were available. The procedure here given not only allows to assess beforehand the bit error rate performance of the DC-DM method, but to improve the detector by exploiting any available knowledge about the noise joint pdf. This is particularly so for JPEG compression, a case that has been treated here in some detail.

8. 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] M. H. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory*, vol. 29, pp. 439–441, May 1983.

[3] J. J. Eggers and B. Girod, *Informed Watermarking*. Kluwer Academic Publishers, 2002.

[4] M. Ramkumar, A. Akansu, and X. Cai, "Floating signal constellations for multimedia steganography," in *IEEE ICC*, (New Orleans, USA), pp. 249–253, June 2000.

[5] F. Pérez-González and F. Balado, "Nothing but a kiss: A novel and accurate approach to assessing the performance of multidimensional distortion-compensated dither modulation," in *Proc. of the 5th International Workshop on Information Hiding*, Lecture Notes in Computer Science, (Noorwijkerhout, The Netherlands), Springer-Verlag, October 2002.

[6] F. Pérez-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*, vol. 51, pp. 960–980, April 2003. Special Issue "Signal Processing for Data Hiding in Digital Media & Secure Content Delivery".