

Iterative Estimation of Amplitude Scaling on Distortion Compensated Dither Modulation

K. M. Whelan, F. Balado, G.C.M. Silvestre, N.J. Hurley

Department of Computer Science,
University College Dublin, Ireland.

ABSTRACT

The vulnerability of quantization-based data hiding schemes to amplitude scaling requires the formulation of countermeasures to this relatively simple attack. Parameter estimation is one possible approach, where the applied scaling is estimated from the received signal at the decoder. This estimate can be used to correct the mismatch with respect to the quantization step assumed by the decoder prior to decoding. In this work we first review previous approaches utilizing parameter estimation as a means of combating the scaling attack on DC-DM. We then present a method for joint iterative decoding and maximum likelihood estimation of the scaling factor for this quantization-based method. The estimation method exploits the reliabilities provided by the near-optimal iterative decoding process in order to successively refine the estimate. The complexity of this problem is tackled using the expectation maximization algorithm. By performing estimation in cooperation with the decoding process, reliable estimation is possible at very low watermark-to-noise power ratios by using sufficiently low rate codes.

Keywords: Side-informed data hiding, DC-DM, Iterative Decoding, ML Estimation, Amplitude Scaling.

1. INTRODUCTION

Quantization-based data hiding schemes have become the most promising methods to approach the capacity limit stated by Costa.¹ Among these, two completely equivalent *scalar* methods have been proposed, namely Distortion-Compensated Dither Modulation (DC-DM) with uniform scalar quantizers by Chen and Wornell,² and the Scalar Costa Scheme (SCS) by Eggers et. al.³ We will refer to the aforementioned scalar scheme as DC-DM in the sequel, due to the precedence of this name. While relatively simple, the achievable rate of this scheme has been shown to be asymptotically (for large constellation sizes and large WNR) only 1.53 dB away from the data hiding capacity for the additive white Gaussian noise channel.^{2,3}

In practical data hiding however other attack channels must be considered in assessing overall system performance. Of particular relevance to quantization-based data hiding methods is the amplitude scaling attack, which involves scaling of the watermarked signal values by a constant factor. This is a particularly challenging attack for quantization-based embedding schemes such as DC-DM as the amplitude scaling of the watermarked signal creates a mismatch with respect to the quantization step assumed by the decoder. As a consequence, and without measures to combat this attack, the performance of this scheme decreases rapidly as the scaling factor applied departs from unity even in the absence of any additional attacking noise.

There have been a number of proposals in the data hiding literature to counter the amplitude scaling attack. We may broadly classify these proposals into two categories: a) those which aim at achieving intrinsic amplitude scaling invariance by employing embedding methods whose performance are, by construction, invariant to scaling, for example^{4,5} ; b) those which perform estimation of the amplitude scaling at the decoder in an attempt to reverse the applied scaling prior to decoding. Our focus in this work will be on the later approach.

Among previous estimation approaches, which we will discuss in more detail in Section 3, firstly there are those which employ the assistance of pilot symbols or templates to aid in the estimation. While no doubt useful

Further author information: (Send correspondence to K. Whelan)

- E-mail: kevin.whelan@ihl.ucd.ie, Telephone: +353 1 716 2454

- This work is funded under Enterprise Ireland ATRP program, research grant ATRP2002/230 and by the European Commission through the IST Programme under Contract IST-2002-507609 SIMILAR.

in practice there are a number of drawbacks to this type of approach, most notably an efficiency loss with respect to the maximum achievable rate. Partly due to this loss, other approaches have investigated blind estimation, where estimation is performed without recourse to pilot symbols or templates. Blind estimation methods also carry an efficiency loss, although, ideally, it is only caused by the intrinsic nonzero variance of any estimation process. A group of works following this strategy and related to our approach are those in,^{6,7} who propose methods for blind estimation of amplitude scalings for side-informed data hiding.

Nevertheless, the blind estimation approach that we pursue here markedly differs from the strategies followed in the aforementioned works. These methods were solely based on invariants of the watermarked signal itself and/or on its statistical properties. Our approach does use this type of information, but it also builds on the redundant structure imposed on the watermark by near-optimal channel coding. Notice also that these error-correcting codes create an overhead, similar to the pilot sequences. Nonetheless, this overhead is compulsory in any scheme trying to approach channel capacity, and it is asymptotically minimal by definition.

The key element of the approach that we will follow here is the iterative refinement of the estimations. As we will see, this refinement is afforded by the increasingly more accurate reliability information on the decoded symbols that is available to the decoder. Consequently, the decoding and estimation processes may be intertwined allowing for reliable estimation at low watermark-to-noise power ratios – a scenario where other blind estimation approaches will most likely fail. The roots of this type of strategy lie in the development of iteratively decodable channel codes (i.e. turbo codes, low-density parity-check codes), with coding gains that enable reliable communication near the Shannon limit.⁸ Different researchers in the digital communications field have realized that the so-called *turbo principle* can also be exploited for purposes other than the decoding process. Among these applications we may cite equalization,⁹ SNR estimation,¹⁰ or synchronization.¹¹ Our strategy in the present work will be the application of this philosophy in the estimation of amplitude scalings on DC-DM.

1.1. Notation and Preliminaries

Scalar random variables are denoted by capital letters, e.g., X , while their realizations are indicated with lowercase types, e.g., x . An exception to this is given by the estimate of the amplitude scaling factor γ , which is denoted by $\hat{\gamma}$ for keeping the usual notation in estimation. All vectors are row vectors, and are denoted by bold types. Vectors may be written in capital, e.g. \mathbf{X} , or lowercase letters, e.g. \mathbf{x} , according to the aforementioned conventions. The probability distribution function (pdf) of a continuous random variable X is denoted by $p_X(x)$, whereas if X is discrete its probability mass function (pmf) is designated by $P_X(X = x)$. For the sake of simplicity some notational shortcuts will be used. The subscripts of the probability functions will be dropped wherever it is clear the random variable they refer to. Also, $P_X(X = x)$ will be denoted by $P(x)$ wherever the meaning of this probability is unambiguous. Last, it will be understood that X represents any of the random variables in $\mathbf{X} = (X_1, \dots, X_N)$ if they are identically distributed.

The host data will be denoted by the N -length random vector \mathbf{X} . Except otherwise indicated, we will limit our exposition to zero-mean Gaussian independent identically distributed (i.i.d.) host data with variance σ_x^2 . This choice allows for analytic tractability in many cases and has been the usual benchmark host in most prior data hiding research. We will also assume without loss of generality that the information symbols to be embedded are statistically independent and equally likely. Two measurements that we will use are the *watermark-to-noise ratio* (WNR) and the *host-to-watermark ratio* (HWR). These parameters are defined as the ratio between the average energy of the involved signals.

2. PROBLEM FORMULATION

We will review here the basics of the DC-DM data hiding scheme and formalize the notion of an amplitude scaling attack. For the sake of simplicity we will consider only a binary scheme, which does not imply a significant loss in the achievable rate with respect to schemes using larger alphabets for moderate to low WNR.³ In any case, all the results that we will present are straightforwardly extendable to higher-dimensional constellations.

In binary DC-DM employing scalar uniform quantizers each sample of the watermarked signal carries one information symbol $b_k \in \{\pm 1\}$. This symbol is hidden by quantizing a sample of the host signal x_k to the nearest

centroid $Q_{b_k}(x_k)$ of the shifted lattice Λ_{b_k} given by

$$\Lambda_{b_k} \triangleq 2\Delta \mathbb{Z} + \Delta \frac{(b_k + 1)}{2} + d_k,$$

with \mathbb{Z} the only one-dimensional lattice (integer lattice), 2Δ the quantization step size, and \mathbf{d} a key-dependent pseudorandom dither value deterministically known to both the encoder and the decoder. If we define next the quantization error with respect to Λ_{b_k} as

$$e_k \triangleq x_k - Q_{b_k}(x_k) = x_k \bmod \Lambda_{b_k}, \quad (1)$$

then, the watermarked signal at the sample k when b_k is embedded is obtained as

$$y_k = x_k - \alpha \cdot e_k = Q_{b_k}(x_k) + (1 - \alpha) \cdot e_k,$$

i.e, the watermark $w_k \triangleq y_k - x_k$ is built using the quantization error weighted by an optimizable constant $0 \leq \alpha \leq 1$. Due to perceptual reasons we have that the inequality $\Delta^2/3 \ll \sigma_x^2$ usually holds true. Then, for a wide range of host signals, $P_{X_k}(x_k)$ can be taken to be roughly constant within one quantization bin. Therefore, we will assume that the quantization error inside a bin is viewed by the decoder as a uniformly distributed random variable in the interval $[-\Delta, \Delta)$. Accordingly, the watermark is also viewed as uniformly distributed in a single bin. Notice that the aforementioned perceptual restrictions usually impose a high HWR = σ_x^2/σ_w^2 .

In uncoded DC-DM the decoder acts by quantizing a received vector \mathbf{z} , which is a distorted version of \mathbf{y} . This sample-by-sample quantization amounts to finding the closest centroid of $\Lambda_{-1} \cup \Lambda_1$ in a Euclidean distance sense. Then, the decoding decisions are

$$\hat{b}_k = \arg \min_{b \in \{\pm 1\}} |z_k - Q_b(z_k)|, \quad (2)$$

for $k = 1, \dots, N$. Last, notice that this particular formulation of DC-DM (SCS) is completely equivalent to those ones which apply a scaling α before quantization.

We assume that, before reaching the decoder, the watermarked signal has been subjected to an amplitude scaling attack of the form

$$\mathbf{Z} = \gamma \cdot (\mathbf{Y} + \mathbf{G}), \quad (3)$$

where \mathbf{G} is an i.i.d. Gaussian noise vector with known variance σ_g^2 and independent of \mathbf{X} , and γ a positive unknown scalar that we wish to estimate. The notation with random variables is due to the fact that the decoder sees all the signals in (3) as random variables. A similar scenario has been assumed in other works devoted to the estimation of amplitude scaling,^{3, 6, 7, 12} although the later work does not assume gaussianity. Joint estimation for the noise variance could also be considered along the guidelines given here. For convenience, we choose a scenario in which scaling is applied after noise addition as in this case the WNR stays constant after the scaling. Notice however that some previous works have assumed the scaling is applied before the noise addition. In any case, the two can be made equivalent by appropriate scaling of the σ_g^2 . Therefore any reference made in the sequel to an amplitude scaling attack will refer to (3) where it assumed σ_g^2 has been scaled accordingly.

Clearly, the attack (3) requires a scaling of Λ_{b_k} by γ at the decoder in order to achieve the same decoding performance as the case where no scaling is applied. Our task then, will be to estimate γ at the decoder and apply the appropriate scaling using this estimate.

3. PREVIOUS APPROACHES TO ESTIMATION OF AMPLITUDE SCALING ON DC-DM

Before proceeding we will review in this section previous approaches to the estimation of amplitude scaling on DC-DM. We first consider those approaches which employ pilots symbols or templates, following which we examine blind estimation strategies. As we will see, many of these approaches exploit the structure induced on the host signal pdf by DC-DM embedding.

3.1. Pilot Assisted Estimation

Among these works we may cite Eggers et al.³ who propose the use of pilot symbols to aid in the estimation. These symbols form a preestablished sequence \mathbf{b}^p , agreed on by the encoder and decoder. We assume, without loss of generality, that the sequence of pilot symbols are embedded as the first L elements of \mathbf{b} . The sequence \mathbf{b} is securely embedded in the host signal using a dither \mathbf{d} . Prior to decoding, the decoder computes $\tilde{p}(\mathbf{z}^p|\mathbf{b}^p, \mathbf{d}^p)$ where the superscript p refers to the first L elements of the corresponding vector and $\tilde{p}(\cdot)$ represents an empirical histogram of the true pdf. The estimation relies on the fact that only by considering values of \mathbf{d}^p within certain contiguous intervals of its distribution, a number of empirical histograms which display local maxima at intervals of approximately $2\gamma\Delta$ can be constructed. The periodicities in these histograms can be used to estimate the scaling by Fourier analysis as

$$\hat{\gamma} = \arg \max_{\tilde{\gamma}} \left| \Phi \left(\frac{1}{2\tilde{\gamma}\Delta} \right) \right|, \quad (4)$$

with $\Phi(\cdot)$ the combined spectra of the histograms. A similar approach is implicitly followed by Shterev et al.⁷ in the absence of a dither. The use of a sufficiently long pilot sequence with both of these methods allows for reliable estimation at very low WNRs.

Moulin et al.¹² propose to embed a pilot signal or template \mathbf{s} to aid in the estimation of amplitude scalings. The shape of the pilot signal is optimized through a game on the fundamental lower bound on the variance of the maximum likelihood (ML) estimator of γ . The payoff of this game is $J_s(\gamma)$, the Fisher information of the parameter γ for a given \mathbf{s} . The encoder then chooses a signal \mathbf{s} to maximize the Fisher information while an attacker chooses γ to minimize $J_s(\gamma)$, under specified embedding and attack distortion constraints, respectively. This strategy by the encoder guarantees a good performance of

$$\hat{\gamma} = \arg \max_{\tilde{\gamma}} \log p(\mathbf{z}|\mathbf{s}; \tilde{\gamma}). \quad (5)$$

While estimation approaches based on pilot symbols or signals are no doubt useful in practice, they inherently imply an efficiency loss with respect to the maximum achievable rate. This loss is due to a decrease in the number of conveyable information symbols in the first case, and to a decrease in the available watermark power in the second one, assuming that the perceptual constraints are constant. Moreover, as argued in,¹² pilot symbols also mean a loss of adaptivity and security in data hiding, despite the fact that they might be key-dependent.

3.2. Blind Estimation

In a generalization of their pilot assisted estimation method in,⁷ Shterev et al. propose not to devote any of the embedded information symbols \mathbf{b} , to pilot symbols and apply the same estimation technique. The estimate of γ in this case is computed as

$$\hat{\gamma} = \arg \max_{\tilde{\gamma}: \tilde{\gamma} < T} \left| \Phi \left(\frac{1}{\tilde{\gamma}\Delta} \right) \right|, \quad (6)$$

with $\Phi(\cdot)$ the Fourier transform of the empirical pdf $\tilde{p}(\mathbf{z})$. The restriction on $\tilde{\gamma}$ in (6) represents the necessity to eliminate the peak around zero frequency from the maximization. The performance of this method obviously suffers in comparison with the pilot case, with reliable estimation impossible below WNR ≈ 0 dB where isolation of a dominant spectral component becomes impossible.

Using the periodicities in the received signal pdf, as done in (4) and (6) is one way to exploit the structure induced on the the host signal pdf by DC-DM embedding for the estimation of amplitude scalings. An alternative approach to take advantage of this structure is proposed by Lee et al.⁶ where the peaks in pdf of the received signal are exploited in the estimation. To this end, the pdf of \mathbf{z} is modelled as a weighted mixture of L Gaussian distributions with means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_L)$ and standard deviations $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_L)$ i.e. $p(\mathbf{z}) \approx \sum_{k=1}^L \beta_k p_k(\mathbf{z}; \mu_k, \sigma_k^2)$ where β_k is the weight on the mixture component $p_k(\cdot) \sim \mathcal{N}(\mu_k, \sigma_k^2)$. Ideally, if the means approximately correspond to the peaks of $p(\mathbf{z})$, the distance between adjacent elements of $\boldsymbol{\mu}$ should then give an estimate of the correct quantization step size for decoding. The Expectation-Maximization (EM) algorithm¹³ is

used to iteratively estimate $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$. At the i -th iteration $\hat{\boldsymbol{\mu}}^{(i-1)}, \hat{\boldsymbol{\sigma}}^{(i-1)}$ and $\hat{\boldsymbol{\beta}}^{(i-1)}$ are used to compute $\hat{\boldsymbol{\mu}}^{(i)}$ and $\hat{\boldsymbol{\beta}}^{(i)}$. The scaling factor is then estimated as

$$\hat{\gamma}^{(i)} = \arg \min_{\tilde{\gamma}} \sum_{k=1}^L \beta_k^{(i)} \left(\hat{\mu}_k^{(i)} - \tilde{\gamma} \cdot \hat{\mu}_k^{(i-1)} \right)^2. \quad (7)$$

This new estimate is then used to compute $\hat{\boldsymbol{\sigma}}^{(i)}$ and the process is iterated a number of times to obtain the final estimate of γ . While successful in the higher WNR range, this approach suffers difficulties in the lower WNR range as the peaks in the pdf of the received signal become less and less discernible. In this case $\hat{\boldsymbol{\mu}}$ does not correspond to the peaks of the pdf and reliable estimation of γ becomes impossible with this approach.

A method which is not explicitly based on exploiting the peaks or periodicities in the received signal pdf is proposed by Shterev et al.⁷ and extended in,¹⁴ where ML estimation of the scaling factor is proposed. Statistical characterization of the involved signals allows for formulation of the ML estimator. The ML estimate is computed as

$$\hat{\gamma} = \arg \max_{\tilde{\gamma}} \log p(\mathbf{z}; \tilde{\gamma}). \quad (8)$$

Numerical techniques are required to solve (8), due to the complexity of the pdfs involved. The results presented in¹⁴ indicate the success of this approach, even at low WNRs. These results are not surprising as knowledge of unscaled host signal variance is assumed and therefore all knowledge is available to build the asymptotically optimal ML estimator. It is somewhat unrealistic to assume the *exact* knowledge of σ_x^2 at the decoder in a practical scenario. Indeed, under the assumption of this knowledge we might think of building other, much simpler, estimators exploiting knowledge of σ_x^2 , for example $\hat{\gamma}^2 = \frac{1}{N\sigma_x^2} \sum_{i=1}^N z_k^2$. Also, in order to locate the maximum of the likelihood function, an extremely fine search of the scaling parameter space is performed. This strategy is computationally expensive, possibly becoming impractical depending on the extent of the search.

Finally, an alternative estimation strategy involves conditioning the embedding quantization step size on moments of the host signal which scale in proportion with the attacked watermarked signal. The decoder can then use these scaled moments in the estimation of the correct step size required for decoding. This approach was suggested in,³ and put into practice in¹⁵ for local environments. Another promising method following this strategy was recently proposed and analyzed in,¹⁶ showing the potentiality of this method in combatting the amplitude scaling attack on quantization-based data hiding methods.

4. ITERATIVE MAXIMUM LIKELIHOOD DECODING AND ESTIMATION

Notice that many of the above proposals present drawbacks at low WNRs, in the case of pilots this is due to the large decrease in efficiency with respect to the maximum achievable rate. In the case of the blind methods such as^{6,7} these WNRs pose difficulties for the estimation process, as discussed above. The problem of reliable estimation at low WNRs motivates our proposal to use iteratively decodable channel codes in conjunction with DC-DM and to perform estimation jointly with the decoding process. Therefore, as we will see in what follows, by using sufficiently low rate codes the estimation process can be provided with reliable information on the decoded symbols which is exploited in the estimation of the amplitude scaling.

We describe next how to perform ML estimation of the amplitude scaling on DC-DM jointly with the iterative decoding process. The complexity of this task will be addressed using the EM algorithm, following the same guidelines as those provided in.¹⁷

4.1. Channel Coding

In order to make the embedding efficient with respect to the achievable rate, we will hide a binary codeword $\mathbf{c} = (c_1, \dots, c_N)$ in \mathbf{x} instead of hiding N uncoded bits using DC-DM. The codeword is obtained by encoding a binary information vector $\mathbf{b} = (b_1, \dots, b_M)$, $M < N$, using a rate $r_c = M/N$ error-correcting code. Without loss of generality we will also consider that the codeword symbols are given in antipodal form, i.e., $c_k \in \{\pm 1\}$. We will center our attention on the simplest case of parallel concatenated codes with iterative decoding, i.e.,

turbo codes. In any case, the procedure that will be presented here is extensible to other iteratively decodable codes. We recall that, in this case, the parallel concatenated codewords are formed as the concatenation of three elements⁸

$$\mathbf{c} = (\mathbf{c}^s, \mathbf{c}^{p1}, \mathbf{c}^{p2}). \quad (9)$$

The subvector $\mathbf{c}^s = (c_1, \dots, c_M) = \mathbf{b}$ is just the systematic output. On the other hand, the subvectors \mathbf{c}^{p1} and \mathbf{c}^{p2} are the parity output of a certain recursive systematic convolutional (RSC) encoder, and they have the same length $(N - M)/2$. These parities are obtained when the input to the RSC is \mathbf{b} and a pseudorandomly interleaved version of \mathbf{b} , $\Pi(\mathbf{b})$, respectively. Notice that, without loss of generality and in order to simplify notation, we are assuming that each codeword symbol c_k is embedded in the host sample with the same index x_k , $k = 1, \dots, N$. In practice, a key-dependent pseudorandom permutation may be used for introducing uncertainty in the position of the coded symbols.

4.2. ML Decoding and Estimation

Let us assume first that decoder has received \mathbf{z} and knows the parameter γ . Now, thanks to the model $p_{\mathbf{Z}}(\mathbf{z}; \gamma)$ and to the code structure, the decoder may undertake near-ML decoding instead of just performing symbol-by-symbol hard decisions using (2). This is done by computing the log-likelihood ratio (LLR)

$$\lambda_k \triangleq \log \frac{P(b_k = +1 | \mathbf{z}; \gamma)}{P(b_k = -1 | \mathbf{z}; \gamma)}, \quad (10)$$

for $k = 1, \dots, M$. The LLRs are computed iteratively by the decoding process until convergence is achieved. One decoding iteration comprises two semi-iterations consisting in decoding sequentially $(\mathbf{z}^s, \mathbf{z}^{p1})$ and $(\mathbf{z}^s, \mathbf{z}^{p2})$. Notice that we are adopting for \mathbf{z} the same superscript notation used for \mathbf{c} in (9). A decoding semi-iteration computes the *a posteriori* probabilities in (10) using the BCJR algorithm, and the model of \mathbf{Z} , exploiting *a priori* information given by the previous decoding semi-iteration. Therefore, this information i.e. $\boldsymbol{\lambda}^{(i-1)} = (\lambda_1^{(i-1)}, \dots, \lambda_M^{(i-1)})$ is used for computing $\boldsymbol{\lambda}^{(i)}$, where the superscript (i) indicates the i -th decoding semi-iteration and $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$. After any decoding step the decisions on the decoded symbols are simply

$$\hat{b}_k = \text{sgn } \lambda_k,$$

for $k = 1, \dots, M$. After convergence this procedure amounts to near-ML decoding of a random-like code, accounting for the near-optimal properties of turbo codes. For a detailed description of the iterative decoding process see for example.⁸

Nevertheless, the statistical model needed for this decoding process is dependent on the unknown parameter by hypothesis. Therefore, the receiver has to estimate γ from \mathbf{z} in order to undertake decoding. A commonly favored estimation strategy is the ML approach which although not optimal for finite data, asymptotically yields the minimum variance unbiased estimator. Following this approach the receiver computes the ML estimate as

$$\hat{\gamma}_{\text{ML}} = \arg \max_{\tilde{\gamma}} \log p(\mathbf{z}; \tilde{\gamma}), \quad (11)$$

where we use the log-likelihood for convenience. We have already seen this approach used in (5) and (8) to estimate amplitude scalings, although (5) differs slightly from (11) and (8) in that an additional pilot signal is used by the estimator to ensure good performance. The essential difference between (11) and (8) lies in the use of a channel code for the estimation.

Notice that the computation of the likelihood in (11) is hindered by the dependencies introduced in \mathbf{Z} by the codeword. For this reason, and being aware of the suboptimality of this strategy, we will resort to solve (11) only for the subvector \mathbf{z}^s of \mathbf{z} corresponding to the systematic part of the codeword. As the scaling attack in (3) does not introduce dependencies among the elements of \mathbf{Y}^s , the independence hypothesis also holds true for the elements of \mathbf{Z}^s , and the simpler ML problem can be written as

$$\hat{\gamma}_{\text{ML}} = \arg \max_{\tilde{\gamma}} \log p(\mathbf{z}^s; \tilde{\gamma}) \quad (12)$$

$$= \arg \max_{\tilde{\gamma}} \sum_{k=1}^M \log p(z_k; \tilde{\gamma}). \quad (13)$$

In any case, we will see that the subvectors corresponding to the parities \mathbf{z}^{P_1} and \mathbf{z}^{P_2} can be used to improve the estimator (12) beyond what we could obtain with \mathbf{z}^s alone.

Unfortunately, a tractable and straightforward solution to this problem is not possible in many cases. A convenient way to address the problem (13) is provided by the EM algorithm formalized by Dempster et al.,¹³ which aims at finding the ML solution iteratively using two alternating steps. We have seen in Section 3.2, the use of the EM algorithm to estimate amplitude scalings. The key difference between that approach and the one we will follow is the use of channel coding. We will see next how the structure imposed on the watermark by near-optimal channel coding can be exploited by EM, using the reliabilities on the embedded information symbols provided by the decoding process.

The first step of the EM algorithm, called E-step, involves the computation of the functional

$$\mathcal{Q}(\hat{\gamma}^{(i)}, \tilde{\gamma}) \triangleq \sum_{\mathbf{b} \in \mathcal{B}} P(\mathbf{b} | \mathbf{z}^s; \hat{\gamma}^{(i)}) \cdot \log p(\mathbf{z}^s | \mathbf{b}; \tilde{\gamma}), \quad (14)$$

with $\mathcal{B} = \{\pm 1\}^M$, $\hat{\gamma}^{(i)}$ the estimate of γ at semi-iteration i , and $\tilde{\gamma}$ a free variable for optimizing $\hat{\gamma}$ at the next iteration. The embedded information vector \mathbf{b} , unknown to the receiver, plays the role of the *unobserved data* in the terminology employed in.¹³ We may see that the computation of (14) requires having a pmf of these unobserved data, conditioned to the received signal \mathbf{z}^s and an estimate of the parameter $\hat{\gamma}^{(i)}$. This pmf could be computed from \mathbf{z}^s by applying Bayes rule. But actually, the iterative decoding process does nothing else than obtaining more accurately this distribution by exploiting the codeword redundancy and the *a priori* information. This probability distribution, computed by the turbo decoding process using the channel model estimate at a given semi-iteration, is

$$P(b_k | \mathbf{z}; \hat{\gamma}^{(i)}) = \frac{1}{1 + \exp(-b_k \lambda_k)}, \quad (15)$$

for $k = 1, \dots, M$, and noting that $\lambda^{(i)}$ depends on \mathbf{z} , $\hat{\gamma}^{(i)}$, and $\lambda^{(i-1)}$. For the sake of keeping notation simple, we will denote the pmf (15) just as $P(b_k)$ in the remainder of the paper, keeping always in mind the dependencies stated above.

Now, it is possible to take the second step, called M-step, which is the maximization problem

$$\hat{\gamma}^{(i+1)} = \arg \max_{\tilde{\gamma}} \mathcal{Q}(\hat{\gamma}^{(i)}, \tilde{\gamma}). \quad (16)$$

It is guaranteed that, for the succession built in this way, $p(\mathbf{z}; \hat{\gamma}^{(i+1)}) \geq p(\mathbf{z}; \hat{\gamma}^{(i)})$.¹³ An issue is that solving the optimization problem (16) can be involved in some cases. For this reason, it is worth noting that (slower) convergence can be also achieved by finding at each M-step a $\tilde{\gamma}$ that increases the functional (14), instead of maximizing it.¹³ To sum up, the EM algorithm uses an estimate $\hat{\gamma}^{(i)}$ to obtain a better reestimate $\hat{\gamma}^{(i+1)}$ that can be used to improve iterative decoding by recomputing (15) and so on. In this way, turbo decoding is intertwined with the estimation problem. Notice that (16) is an iterative estimation of γ , using EM such as (7). It is the use of a channel code which makes (16) a better strategy.

From a practical point of view, the EM functional (14) can be considerably simplified using the hypotheses of mutual independence among the elements of \mathbf{Z}^s and \mathbf{B} , respectively, and noting that $p(z_k | \mathbf{b}; \tilde{\gamma}) = p(z_k | b_k; \tilde{\gamma})$. Then, (14) may be rewritten as

$$\mathcal{Q}(\hat{\gamma}^{(i)}, \tilde{\gamma}) = \sum_{\mathbf{b} \in \mathcal{B}} \prod_{j=1}^M P(b_j) \cdot \sum_{k=1}^M \log p(z_k | b_k; \tilde{\gamma}) = \sum_{k=1}^M \sum_{b_k \in \{\pm 1\}} P(b_k) \log p(z_k | b_k; \tilde{\gamma}) \cdot \sum_{\mathbf{b} \in \mathcal{B}} \prod_{\substack{j=1 \\ j \neq k}}^M P(b_j).$$

As the summation over all the probabilities of any marginal pmf of $P(\mathbf{b}) = \prod_{k=1}^M P(b_k)$ equals one, we have finally that

$$\mathcal{Q}(\hat{\gamma}^{(i)}, \tilde{\gamma}) = \sum_{k=1}^M \sum_{b_k \in \{\pm 1\}} P(b_k) \log p(z_k | b_k; \tilde{\gamma}). \quad (17)$$

4.3. Blind Performance Assessment

An interesting aside from the discussion in the preceding section is that the LLRs (10) can be exploited in more ways than the one previously described. A fact that we will find useful in practical joint estimation and decoding is that the actual probability of bit error at the decoder, which is defined as $P_b \triangleq \frac{1}{M} \sum_{k=1}^M \Pr(\hat{b}_k \neq b_k)$, can be blindly estimated at any decoding stage. The decoder can do this in a relatively accurate way exploiting (10) provided by the decoding process. There are several approaches to undertake this estimation, as for instance¹⁸

$$\hat{P}_b = \frac{1}{M} \sum_{k=1}^M \frac{1}{1 + \exp |\lambda_k|}. \quad (18)$$

We will see in Section 8 that this blind performance estimation may be used to improve the initialization of the iterations.

5. FOURIER SERIES APPROXIMATION TO PDF OF DC-DM UNDER GAUSSIAN NOISE

Computation of the EM functional (17) requires a statistical characterization of Z for a given γ and conditioned to an embedded symbol b , i.e., $p(z|b; \gamma)$. This distribution is also required to compute the channel reliabilities of the embedded symbols which are passed to the BCJR algorithm on the first decoding semi-iteration. As we will see next, the exact expression for this pdf is not amenable to analysis and so we will pursue an approximation. Following the derivation of this approximation we will evaluate its accuracy in a probability of decoding error sense.

5.1. Exact PDF of DC-DM under Gaussian Noise

The pdf we require i.e. $p(z|b; \gamma)$ can be obtained from the pdf of $R \triangleq Y + G$ by a simple random variable transformation so our initial task will be to compute the pdf of R conditioned to b . This distribution is given by

$$p_R(r|b) = \sum_{q \in \Lambda_b} P(Q_b(X) = q) \cdot p_T(r - q), \quad (19)$$

with $p_T(\cdot)$ defined as the convolution of $p_G(\cdot)$ with a uniform distribution on the interval $[-(1 - \alpha)\Delta, (1 - \alpha)\Delta)$. While (19) would be the optimal pdf for iterative decoding, this pdf is by construction not amenable to analysis.^{3, 14} We will therefore pursue an approximation to (19).

5.2. Weighted Fourier Series Approximation

Consider the case where $\text{HWR} \rightarrow \infty$, then (19) resembles a periodic function on the lattice $2\Delta\mathbb{Z}$. Therefore, following the approach of Forney et al.¹⁹ who propose a method to approximate a periodic function on a lattice, it is possible to obtain the Fourier series representation of $p_R(r|b)$ using the dual lattice of $2\Delta\mathbb{Z}$. According to the same authors, if the noise variance is not too low we may approximate (19) by the lower frequency terms of the Fourier series, having in this case

$$p_R(r|b) \approx \frac{1}{2\Delta} \left(1 - b \cdot \eta \cos\left(\frac{\pi r}{\Delta}\right) \right), \quad (20)$$

where $\eta \triangleq 2 \text{sinc}(1 - \alpha) \exp(-\pi^2 \sigma_g^2 / (2\Delta^2))$.

If the pdf of X were uniform and the HWR high, the truncated Fourier series (20) would be a sufficient approximation to (19) up to a normalization factor. Of course, this approximation is worse for the values close to the limits of the uniform distribution. Similarly, we may think of weighting (20) by the distribution of the host when this is not uniform. Then, weighting that expression by a zero-mean Gaussian pdf with variance σ_x^2 , and normalizing the resulting function to retain a pdf, we may approximate the pdf of R by

$$p_R(r|b) \approx \frac{\exp\left(-\frac{r^2}{2\sigma_x^2}\right)}{\sqrt{2\pi}\sigma_x} \cdot \frac{1 - b \cdot \eta \cos\left(\frac{\pi r}{\Delta}\right)}{1 - b \cdot \eta \exp\left(-\frac{\pi^2 \sigma_x^2}{2\Delta^2}\right)}. \quad (21)$$

A similar model is proposed in³ for estimation of amplitude scalings using pilot symbols, although with a definition somewhat more heuristic than the one given here.

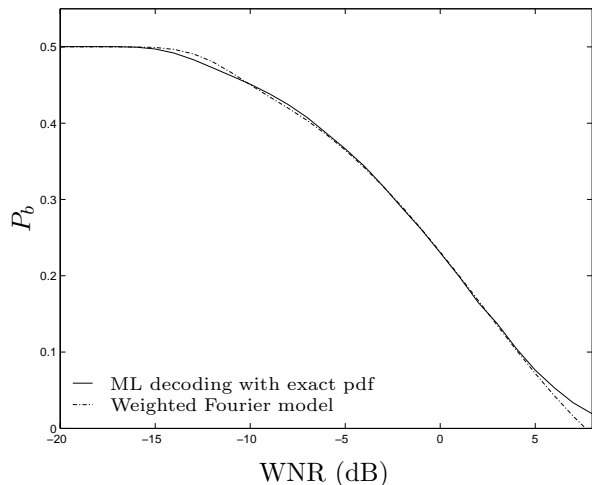


Figure 1. Theoretical P_b for uncoded DC-DM under ML decoding with exact pdf compared to that computed using weighted Fourier model, HWR = 20 dB. α optimized at each WNR.

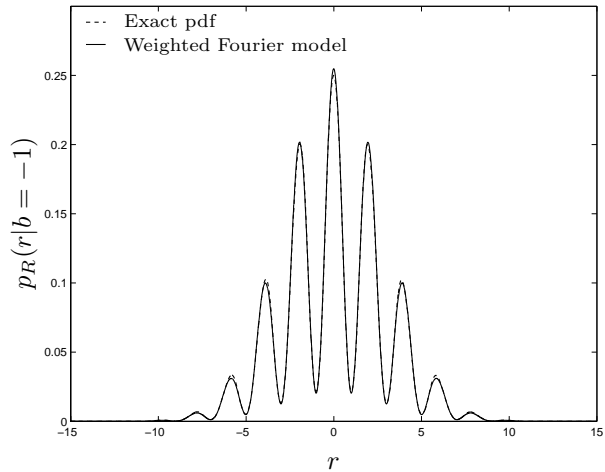


Figure 2. Exact pdf for DC-DM under Gaussian noise compared with weighted Fourier series model. WNR = 0 dB, HWR = 20 dB.

5.3. Accuracy of Fourier Series Approximation

It is important to assess the accuracy of the proposed weighted Fourier model (21) in relation to the exact pdf (19). As our main interest will be whether this model is accurate in a probability of decoding error sense, we will compute P_b for uncoded DC-DM under ML decoding and compare with that computed using the weighted Fourier model for decoding.

Figure 1 shows the theoretical P_b as a function of WNR for ML decoding with (19) and decoding using (21). Notice that the proposed approximation is accurate for decoding across a wide range of WNRs, only deviating significantly from the exact P_b above WNR \approx 3 dB. This deviation is due to only using the low frequency terms in the Fourier model. As the noise variance decreases, the peaks of the pdf (19) around the centroids of Λ_b become more pronounced. Therefore in order to accurately approximate the increasingly rapid changes in the pdf between neighboring centroids, higher frequency terms are needed in the Fourier model. In fact, using only two terms leads to the Fourier model becoming negative at high WNRs, hence ceasing to be a pdf. This accounts for the anomalous values of P_b in Figure 1, in the range WNR $>$ 3 dB. Notice that the Fourier model can easily be modified to include higher frequency terms, although this would greatly complicate the analysis that will follow. Finally, Figure 2 shows the exact pdf (19) compared with (21).

6. MAXIMIZATION OF EM FUNCTIONAL

Using the Fourier model described in the foregoing section, we now compute the EM functional (17) and perform the required maximization (16).

Assuming the same definition of R as in the previous section, the scaling of R by a factor γ will yield the pdf sought for computation of (17), that is

$$p_Z(z|b; \gamma) = \frac{p_R(z/\gamma|b)}{\gamma}. \quad (22)$$

Notice that (22) requires knowledge of σ_x^2 . A reasonable way to estimate this variance from the received signal is $\hat{\sigma}_x^2 = \hat{\sigma}_z^2/\gamma^2$, where $\hat{\sigma}_z^2 \triangleq \sum_{k=1}^M z_k^2/M$. We might think of other different estimators making use of σ_g^2 ; however, perceptual constraints strongly limit their potential performance improvements, while, as we will verify next, the proposed estimator allows for some convenient simplifications in the expressions required for the maximization problem. Taking now the logarithm of (22) and disregarding the terms that do not affect the maximization, the

EM functional (17) can be written as

$$\mathcal{Q}(\hat{\gamma}^{(i)}, \tilde{\gamma}) \approx \sum_{k=1}^M \sum_{b_k \in \{\pm 1\}} P(b_k) \cdot b_k \left\{ \exp\left(-\frac{\pi^2 \hat{\sigma}_z^2}{2\tilde{\gamma}^2 \Delta^2}\right) - \cos\left(\frac{\pi z_k}{\tilde{\gamma} \Delta}\right) \right\}, \quad (23)$$

where we have used the aforementioned estimate $\hat{\sigma}_x^2$ and the approximation $\log(1+x) \approx x$. This approximation holds for small values of $|x|$ and, hence, it is accurate in this case as $\eta < 1$ in the range of validity of the model. Last, notice that (23) does not involve σ_g^2 . The authors in¹⁴ observed an insensitivity of their ML estimation strategy to the value of σ_g^2 . This observation is supported by the independence of the above expression with respect to σ_g^2 , as this parameter does not affect the maximization of (23).

We now perform the maximization step which involves finding the value of $\tilde{\gamma}$ that maximizes (23). This value is given by the solution to the following expression

$$\frac{\partial \mathcal{Q}(\hat{\gamma}^{(i)}, \tilde{\gamma})}{\partial \tilde{\gamma}} = \sum_{k=1}^M \sum_{b_k \in \{\pm 1\}} P(b_k) \cdot b_k \left\{ \exp\left(-\frac{\pi^2 \hat{\sigma}_z^2}{2\tilde{\gamma}^2 \Delta^2}\right) \cdot \frac{\pi \hat{\sigma}_z^2}{\tilde{\gamma} \Delta} - \sin\left(\frac{\pi z_k}{\tilde{\gamma} \Delta}\right) \cdot z_k \right\} = 0. \quad (24)$$

This equation has to be solved numerically, and it presents multiple roots. Still, it can be verified that its behavior is approximately linear in the region around $\hat{\gamma}^{(i)}$, which would be our initial guess for an iterative numerical solution. So, in order to get an explicit solution, we may approximate the first derivative using the first two terms of its Taylor series about $\hat{\gamma}^{(i)}$. The computation may be further simplified by neglecting the exponential summand in (24), as it tends to zero for high HWR assuming that the scaling factor is close to one. We will see in Section 8 that the method is only effective within a certain environment around this value. Under these conditions, the M-step (16) can be solved as

$$\hat{\gamma}^{(i+1)} \approx \frac{\sum_k \sum_{b_k} P(b_k) \cdot b_k \sin\left(\frac{\pi z_k}{\hat{\gamma}^{(i)} \Delta}\right) \cdot z_k}{\sum_k \sum_{b_k} P(b_k) \cdot b_k \cos\left(\frac{\pi z_k}{\hat{\gamma}^{(i)} \Delta}\right) \cdot \frac{\pi z_k^2}{\hat{\gamma}^{(i)2} \Delta}} + \hat{\gamma}^{(i)}, \quad (25)$$

for $k = 1, \dots, M$. Even in the event that this approximation is not able to exactly reach the M-step maximum, it is usually enough for increasing the EM functional. As remarked in Section 4, this suffices for the convergence of EM.

7. CRAMER-RAO BOUND

As the estimators obtained with ML are asymptotically unbiased as $M \rightarrow \infty$, the Cramer-Rao lower bound (CRB) on the variance of the parameter $\hat{\gamma}$ may be used to evaluate its efficiency. We will compute next an approximate analytical expression for the CRB on $\text{Var}\{\hat{\gamma}\}$ using (22). The performance of the proposed estimator will be compared to this approximate bound and the exact bound in Section 8.

It is straightforward to show that the basic regularity condition for the existence of a valid CRB²⁰ is satisfied. Obtaining first $p(z; \gamma) = \sum_b p(z|b; \gamma)/2$ we may see that whereas $\frac{\partial}{\partial \gamma} \log p(z; \gamma)$ is an odd function of z , $p(z; \gamma)$ is even, which implies that $\text{E} \left\{ \frac{\partial}{\partial \gamma} \log p(z; \gamma) \right\} = 0$. Now, differentiating again $\frac{\partial}{\partial \gamma} \log p(z; \gamma)$ with respect to γ , and using the same approximation on the logarithm as in (23), we have that

$$\frac{\partial^2 \log p(z; \gamma)}{\partial \gamma^2} = \frac{1}{\gamma^2} - \frac{3z^2}{\gamma^4 \sigma_x^2} + \zeta \sin\left(\frac{\pi z}{\gamma \Delta}\right) \frac{2\pi z}{\gamma^3 \Delta} + \zeta \cos\left(\frac{\pi z}{\gamma \Delta}\right) \frac{\pi^2 z^2}{\gamma^4 \Delta^2}, \quad (26)$$

with $\zeta \triangleq \eta^2 \exp(-\pi^2 \sigma_x^2 / (2\Delta^2))$. Now, integrating (26) over $p(z; \gamma)$, and making the simplification $\zeta \approx 0$, which holds true for high HWR values, it is tedious but straightforward to show that the bound may be approximated as

$$\text{CRB}(\gamma) \triangleq -\frac{1}{M} \text{E} \left\{ \frac{\partial^2 \log p(z; \gamma)}{\partial \gamma^2} \right\}^{-1} \approx \frac{\gamma^2}{2M}. \quad (27)$$

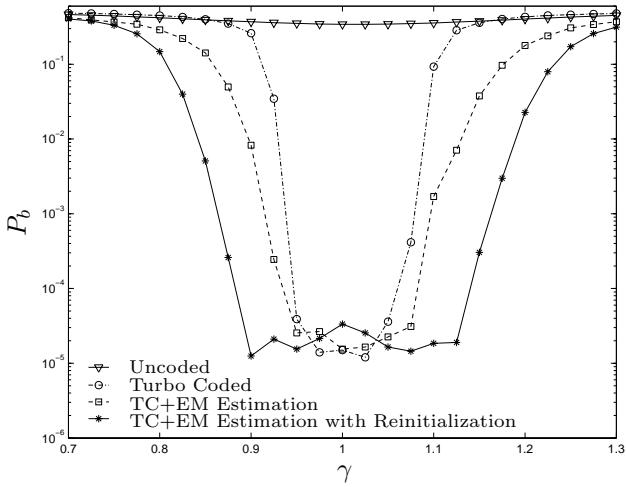


Figure 3. Performance of DC-DM under an amplitude scaling γ . $r_c = 1/15$, HWR = 20 dB, WNR = -4 dB, $\alpha = 0.3$, $\hat{\gamma}^{(0)} = 1$.

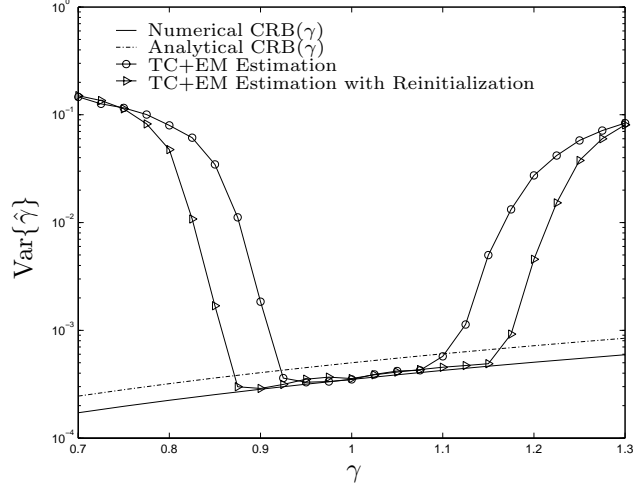


Figure 4. Variance of the estimator of the amplitude scaling γ . $r_c = 1/15$, HWR = 20 dB, WNR = -4 dB, $\alpha = 0.3$, $\hat{\gamma}^{(0)} = 1$.

8. EXPERIMENTAL RESULTS

In this section we undertake empirical tests for the proposed strategy, using the probability of decoding error as the performance measurement. The component codes of the turbo codes used in the experiments have been chosen by trial-and-error, without extensive optimizations. Unless otherwise indicated, an interleaver of size $M = 1000$ is chosen. A maximum of 20 decoding iterations are performed, and the cross-entropy criterion²¹ is used for early stopping of the process. The value of the DC-DM parameter α is chosen empirically to correspond to the approximate WNR at which the turbo cliff occurs for a particular code on the Gaussian channel.

We may see in Figure 3 the performance of the proposed scheme. For comparison purposes, the performances of the uncoded case and of the turbo-coded case with no estimation are also included in the plot. The initialization of the method is done using $\hat{\gamma}^{(0)} = 1$. In this case the algorithm achieves the same performance as for the case where the scaling is known by the decoder ($\gamma = 1$) for the range $\gamma \in (0.95, 1.08)$. An additional performance improvement is also offered over the remaining range of γ considered in the plot. While this improvement is modest compared with that of the turbo code acting alone, we see in the same figure that further improvements are possible through a judicious initialization of the EM algorithm.

This alternative initialization strategy is based on exploiting the blind performance assessment (18). Starting from $\hat{\gamma}^{(0)} = 1$, a number of decoding iterations are performed and \hat{P}_b is then examined. If the decoding decisions are reliable the algorithm is let to run until convergence; alternatively, a reinitialization is performed. Two values of γ either side of the current estimate $\hat{\gamma}^{(i)}$ are chosen, and one decoding semi-iteration is performed using both new values. Again, we employ \hat{P}_b to determine the best scaling, which is used to run the algorithm until convergence. This simple strategy, which comes at a minimal increase in complexity, enlarges the range of scalings which the algorithm can correct, as shown in Figure 3.

Last, Figure 4 shows the empirical variance of $\hat{\gamma}$ compared with the Cramér-Rao lower bound. Notice that the bound is approximately achieved in the range around unity corresponding to near-errorless decoding in Figure 3, and that the analytical bound is reasonably good.

9. CONCLUSIONS

We have presented a method for joint iterative decoding and ML estimation of amplitude scalings on DC-DM. We have shown that this method allows for reliable estimation at low WNRs, albeit in a limited range of scalings, by using low rate channel codes. The limitation of the method is clearly due to the severity with which an amplitude scaling negatively impacts the channel reliabilities required by the iterative decoder. On the positive side, we must realize that the strategy proposed is inherently afforded in many situations, as it does not require anything

else than the near-optimal iteratively decodable codes present in most advanced systems. Therefore, it is almost always integrable with other solutions like pilot estimation to achieve overall better system performance. Also, only constant amplitude scaling have been considered. Nevertheless, and considering that the Cramér-Rao bound decreases as $O(M^{-1})$, we might think of applying the estimation method presented to amplitude scalings only *locally* constant, for sufficiently large local environments.

REFERENCES

1. M. H. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory* **29**, pp. 439–441, May 1983.
2. B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory* **47**, pp. 1423–1443, May 2001.
3. J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding," *IEEE Trans. Signal Processing* **51**, pp. 1003–1019, Apr. 2003.
4. H. Malvar and D. Florêncio, "Improved spread spectrum: a new modulation technique for robust watermarking," *IEEE Trans. Signal Processing* **51**, pp. 898–905, Apr. 2003.
5. A. Abrardo and M. Barni, "Orthogonal dirty paper coding for informed data hiding," in *Procs. of SPIE: Security, Steganography, and Watermarking of Multimedia Contents VI*, **5306**, (San José, USA), Jan. 2004.
6. K. Lee, D. S. Kim, T. Kim, and K. A. Moon, "EM estimation of scale factor for quantization-based audio watermarking," in *Procs. of the 2nd International Workshop on Digital Watermarking, Lecture Notes in Computer Science* **2939**, pp. 316–327, Springer-Verlag, (Seoul, Korea), Oct. 2004.
7. I. Shterev, R. Lagendijk, and R. Heusdens, "Statistical amplitude scale estimation for quantization-based watermarking," in *Procs. of SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI*, **5306**, (San Jos, USA), Jan. 2004.
8. C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," in *Proc. IEEE Int. Conf. on Communications*, pp. 1064–1070, (Geneva, Switzerland), May 1993.
9. C. Douillard, M. Jézéquel, C. Berrou, A. Picart, P. Didier, and A. Glavieux, "Iterative correction of inter-symbol interference: turbo equalization," *European Trans. Telecommun.* **6**, pp. 507–511, sep-oct 1995.
10. T. A. Summers and S. G. Wilson, "SNR mismatch and online estimation in turbo decoding," *IEEE Trans. Communications* **46**, pp. 421–423, Apr. 1998.
11. J. R. Barry, A. Kavčić, S. W. McLaughlin, A. Nayak, and W. Zeng, "Iterative timing recovery," *IEEE Signal Processing Mag.*, Jan. 2004.
12. P. Moulin, "Embedded-signal design for channel parameter estimation," in *Procs. of the IEEE Workshop on Statistical Signal Processing*, (St Louis, USA), Sept. 2003. Parts I & II.
13. A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society, Series B* **39**(1), pp. 1–38, 1977.
14. R. Lagendijk and I. Shterev, "Estimation of attacker's scale and noise variance for qim-dc watermark embedding," in *Proc. of the IEEE International Conf. on Image Processing*, (Singapore), Oct. 2004.
15. J. Oostveen, T. Kalker, and M. Staring, "Adaptive quantization watermarking," in *Procs. of SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI*, **5306**, (San José, USA), Jan. 2004.
16. F. Pérez-González, M. Barni, A. Abrardo, and C. Mosquera, "Rational dither modulation: a novel data-hiding method robust to value-metric attacks," in *IEEE International Workshop on Multimedia Signal Processing*, (Siena, Italy), Sept. 2004.
17. F. Balado, F. Pérez-González, and P. Comesaña, "Blind iterative decoding of side-informed data hiding using the expectation-maximization algorithm," in *Procs. of SPIE: Security, Steganography, and Watermarking of Multimedia Contents VI*, **5306**, (San José, USA), Jan. 2004.
18. P. Hoeher, I. Land, and U. Sörger, "Log-likelihood values and Monte Carlo simulation - Some fundamental results," in *Proc. 2nd Int. Symp. on Turbo Codes & Rel. Topics*, pp. 43–46, (Brest, France), Sept. 2000.
19. G. D. Forney, M. D. Trott, and S.-Y. Chung, "Sphere-bound-achieving coset codes and multilevel coset codes," *IEEE Trans. Inform. Theory* **46**, pp. 820–850, May 2000.
20. S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, vol. I, Prentice Hall, 1993.
21. J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Transactions on Information Theory* **42**, pp. 429–445, March 1996.