AN INFORMATION-THEORETIC APPROACH TO THE DESIGN OF ROBUST DIGITAL WATERMARKING SYSTEMS

Brian Chen and Gregory W. Wornell

Department of Electrical Engineering and Computer Science,
and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA

ABSTRACT

A variety of emerging applications require the design of systems for embedding one signal within another signal. We describe a new class of embedding methods called quantization index modulation (QIM) and develop a realization termed coded dither modulation in which the embedded information modulates the dither signal of a dithered quantizer. We also develop a framework in which one can analyze performance trade-offs among robustness, distortion, and embedding rate, and we show that QIM systems have considerable performance advantages over previously proposed spread-spectrum and low-bit modulation systems.

1. INTRODUCTION

A variety of related applications have emerged recently that require the design of systems for embedding one signal, sometimes called an "embedded signal" or "watermark", within another signal, called a "host signal". The embedding must be done such that the embedded signal causes no serious degradation to its host. At the same time, the host always carries the embedded signal, which can only be removed by causing significant damage to the host. These applications include copyright notification and enforcement, authentication, and transmission of auxiliary information. These and other applications are described in [1], which also provides an overview of several proposed information-embedding algorithms.

Many previously proposed algorithms belong to one of two classes: (1) additive techniques such as spread-spectrum in which a small pseudo-noise signal is added to the host signal and (2) quantize-and-replace strategies that replace a quantized host signal with another quantization value. A common example belonging to the second class is low-bit(s) modulation (LBM) in which the least significant bit(s) of the host signal are replaced by the embedded signal.

This work has been supported in part by ONR under Grant No. N00014-96-1-0930, by AFOSR under Grant No. F49620-96-1-0072, and by a NDSEG Fellowship.

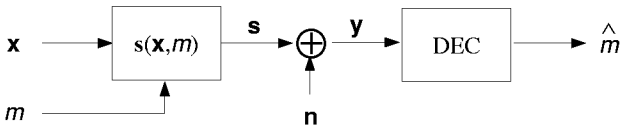

Figure 1: General information embedding problem model. An integer m is embedded in the host signal $\mathbf{x}$. A perturbation vector $\mathbf{n}$ corrupts the composite signal $\mathbf{s}$. The decoder extracts an estimate $\hat{m}$ of m from the channel output $\mathbf{y}$.

In this paper, motivated by information-theoretic perspectives, we describe a new class of information-embedding systems, quantization index modulation (QIM) systems [2], that efficiently perform the trade-offs among the robustness of the embedding, the degradation to the host signal caused by the embedding, and the amount of data embedded. We also demonstrate that coded dither modulation, a convenient implementation of a QIM system, offers significant advantages over previously proposed spread-spectrum and LBM techniques. For the information-theoretic analysis, see [4].

2. PROBLEM MODEL

Many information-embedding applications can be described by Fig. 1. We wish to embed some information m in some host signal vector $\mathbf{x} \in \Re^N$. This host signal could be a vector of pixel values or Discrete Cosine Transform (DCT) coefficients from an image, for example. We wish to embed at a rate of R bits per dimension (bits per host signal sample) so we can think of m as an integer chosen from the set $\{1, 2, \ldots, 2^{NR}\}$. An embedding function maps $\mathbf{x}$ and m to a composite signal $\mathbf{s} \in \Re^N$ subject to some distortion constraint such as, for example, the squared-error distortion constraint

$$D(\mathbf{s}, \mathbf{x}) = \frac{1}{N}\|\mathbf{s} - \mathbf{x}\|^2 \leq D_{\max}, \qquad \forall m. \tag{1}$$

The composite signal is passed through a channel, where it is subjected to various common signal processing manipula-

tions such as lossy compression, addition of random noise, and resampling, as well as deliberate attempts to remove the embedded information. We model the combined effects of these manipulations by the addition of a noise or perturbation vector $\mathbf{n} \in \Re^N$, which can be random or deterministic, signal independent or signal dependent. Thus, this channel model is completely general. However, we assume that the channel output $\mathbf{y}$ must still be a fair representation of the original signal so in this paper we either bound the energy of the perturbation vector,

$$\|\mathbf{n}\|^2 \leq N\sigma_n^2, \tag{2}$$

or bound the distortion $D_{\mathbf{y}}$ between $\mathbf{y}$ and $\mathbf{x}$. The decoder forms an estimate $\hat{m}$ of m based on $\mathbf{y}$. We quantify the robustness of the system by the maximum allowable σ_n^2 such that we can still guarantee that $\hat{m} = m$. Alternatively, we can characterize the reliability of the system by the probability of a message error $\Pr[\hat{m} \neq m]$ or by bit-error rate [3] [4]. In any case, the problem we face is to design an embedding function $\mathbf{s}(\mathbf{x}, m)$ that achieves the best possible trade-off among the three parameters rate, distortion, and robustness (or reliability).

3. QUANTIZATION INDEX MODULATION

We can view the embedding function $\mathbf{s}(\mathbf{x}, m)$ as an ensemble of functions of $\mathbf{x}$, indexed by m. We denote the functions in this ensemble as $\mathbf{s}(\mathbf{x}; m)$ to emphasize this view.

In quantization index modulation (QIM) systems [2], these functions are quantizers, which is convenient for at least two reasons. First, each individual quantizer is designed such that one can satisfy the distortion constraint. Second, the reconstruction points of each quantizer in the ensemble are "far away" in some sense from the reconstruction points of every other quantizer so that the system is robust to noise. Quantization index modulation refers to modulating an index or sequence of indices with the embedded information and quantizing the host signal with the associated quantizer or sequence of quantizers.

Figure 2 illustrates QIM information embedding for the $N = 2$ and $R = 1/2$ case. In this example, one bit is to be embedded so that $m \in \{1, 2\}$. The reconstruction points in $\Re^N$ of the two required quantizers are represented in Fig. 2 with $\times$'s and $\circ$'s. If $m = 1$, for example, $\mathbf{x}$ is quantized with the $\times$-quantizer, i.e., $\mathbf{s}$ is chosen to be the $\times$ closest to $\mathbf{x}$. If $m = 2$, $\mathbf{x}$ is quantized with the $\circ$-quantizer.

A few parameters of the ensemble conveniently characterize the performance of a QIM system. The number of quantizers in the ensemble equals the number of possible values for m, and hence, determines the information-embedding rate. The size and shape of the quantization cells determine the embedding-induced distortion. Finally, the minimum distance $d_{\min}$ between the sets of reconstruction points of different quantizers in the ensemble determines the robustness of the embedding, where the minimum distance is defined as

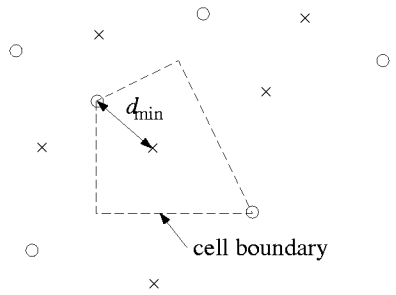

Figure 2: Quantization index modulation. The reconstruction points marked with $\times$'s and $\circ$'s belong to two different quantizers. The minimum distance $d_{\min}$ measures the robustness to noise, and the sizes of the quantization cells determine the embedding-induced distortion.

$$d_{\min} \stackrel{\Delta}{=} \min_{(i,j):i\neq j} \min_{(\mathbf{x}_i,\mathbf{x}_j)} \|\mathbf{s}(\mathbf{x}_i; i) - \mathbf{s}(\mathbf{x}_j; j)\|. \tag{3}$$

Intuitively, the minimum distance measures the size of noise vectors that can be tolerated by the system. For example, with bounded noise energy (2) a minimum distance decoder, which chooses the reconstruction point closest to the channel output, is guaranteed to not make an error as long as

$$\frac{d_{\min}^2}{4N\sigma_n^2} > 1. \tag{4}$$

Alternatively, for additive white Gaussian noise with variance σ_n^2, the error probability is $\sim Q(d_{\min}/(2\sigma_n))$ at high signal-to-noise ratio [5].

4. CODED DITHER MODULATION

Dithered quantizers [6] are quantizer ensembles where the quantization cells and reconstruction points of every quantizer in the ensemble are shifted versions of some base quantizer $\mathbf{q}(\cdot)$. The shift is given by a dither vector $\mathbf{d}$. In a dither modulation system, the dither vector is modulated by the embedded information. Specifically, we define a dither vector $\mathbf{d}(m)$ for each possible value of m so that $\mathbf{s}(\mathbf{x}; m) = \mathbf{q}(\mathbf{x} + \mathbf{d}(m)) - \mathbf{d}(m)$.

As a simple example, we consider the case where $\mathbf{q}(\cdot)$ is a uniform, scalar quantizer with step size Δ. A simple way to ensure that two dithered quantizers are the maximum possible distance from each other in some arbitrary number of dimensions L is to impose the constraint [4]

$$d_i(2) = \begin{cases} d_i(1) + \Delta/2, & d_i(1) < 0 \\ d_i(1) - \Delta/2, & d_i(1) \geq 0 \end{cases}, \quad i = 1, \ldots, L, \tag{5}$$

where $d_i(1)$ and $d_i(2)$ are the ith components of the two dither vectors.

Such dithered quantizers can be used to embed N/L bits $\{z_1, z_2, \ldots, z_{N/L}\}$ as follows: we choose a pair of dither subvectors of length L satisfying (5), associating each subvector with a 0 or 1, and concatenate the N/L subvectors associated with the bit sequence $z_1, \ldots, z_{N/L}$ to form the overall dither vector of length N. We call this type of information embedding "binary dither modulation". If the bit sequence $z_1, \ldots, z_{N/L}$ is a coded bit sequence obtained by applying error correction coding to the NR bits in m, then we refer to this embedding method as "coded binary dither modulation".

If the error correction code is a binary block code with a minimum Hamming distance of d_H and rate of k_u/k_c, then

$$L = \frac{N}{NRk_c/k_u} = \frac{1}{R}(k_u/k_c).$$

Also, because any two coded bit sequences differ by at least d_H bits, (5) implies that the reconstruction points of any given quantizer in the resulting ensemble are shifted by $\pm\Delta/2$ in each dimension relative to the points of any other quantizer over at least Ld_H dimensions. Then, the minimum distance (3) is

$$d^2_{\min} = Ld_H\left(\frac{\Delta}{2}\right)^2 = \left(d_H\frac{k_u}{k_c}\right)\frac{1}{R}\left(\frac{\Delta}{2}\right)^2. \quad (6)$$

If the quantization cells are sufficiently small such that $\mathbf{x}$ can be modeled as uniformly distributed within each cell, the expected squared-error distortion per sample (1) of a uniform, scalar quantizer is $E[D(\mathbf{s},\mathbf{x})] = \Delta^2/12$. Thus, with bounded noise energy and a minimum distance decoder, this expression can be combined with (4) and (6) to compactly express the trade-off among distortion, robustness and rate as

$$\boxed{\gamma_c\frac{3}{4}\frac{1}{NR}\frac{E[D(\mathbf{s},\mathbf{x})]}{\sigma_n^2} > 1,} \quad (7)$$

where $\gamma_c = d_H(k_u/k_c)$ is the performance gain from error correction coding.

The non-zero minimum distance of QIM systems offers quantifiable robustness to noise. In contrast, spread-spectrum systems offer relatively little robustness to noise. These systems embed information by adding a pseudo-noise vector $\mathbf{w}(m)$ to the host signal, i.e., $\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{w}(m)$. The minimum distance of these systems is actually zero, which can be seen by setting $\mathbf{x}_j = \mathbf{x}_i + \mathbf{w}(i) - \mathbf{w}(j)$ during the minimization over $(\mathbf{x}_i, \mathbf{x}_j)$ in (3). Thus, although these systems may be effective when the host signal is known at the decoder, in the often more typical case where the host signal is not known, they offer no guaranteed robustness to noise, and hence, no expression analogous to (7) exists. Although LBM systems have non-zero $d_{\min}$, analysis in [4] establishes that LBM is worse than binary dither modulation by 2.43 dB in this case.

5. TAMPER RESISTANCE

In the analysis of Sec. 4, the robustness measure was the maximum energy between $\mathbf{s}$ and $\mathbf{y}$ that could be tolerated for error-free decoding. In some scenarios, however, it may be more appropriate to consider the distortion between $\mathbf{y}$ and $\mathbf{x}$. For example, attackers with partial knowledge of the host signal, which may be in the form of a probability distribution, can actually calculate this distortion to assure that their attacks do not excessively degrade the host signal.

Furthermore, the attacker may have full knowledge of the embedding and decoding processes. For example, in a copyright ownership notification system, everyone could embed the ASCII representation of a copyright notice such as, "Property of ..." in their copyrightable works. Such a system is analogous to the system currently used to place copyright notices in (hardcopies of) books and requires no central authority to store, register, or maintain separate keys or watermarks for each user. The widespread use of such a "no-key" or "universally-accessible" system requires only standardization of the decoder so that everyone will agree on the decoded watermark, and hence, the owner of the copyright.

In this section, we examine the robustness of QIM, spread spectrum, and LBM systems to attacks from adversaries that have a distortion constraint, partial knowledge of the host signal, and full knowledge of the embedding and decoding processes including any keys. We show that of the three systems considered, only QIM systems are sufficiently robust that even a fully-informed attacker must degrade the host signal quality to remove the watermark.

The measure of robustness is $D_{\mathbf{y}}$, the minimum expected squared-error per letter distortion between $\mathbf{y}$ and $\mathbf{x}$ that an attacker would need to impose in order to cause a decoding error. We use $D_{\mathbf{s}}$ to denote the expected distortion between $\mathbf{s}$ and $\mathbf{x}$ (i.e., due to embedding). The ratio between $D_{\mathbf{y}}$ and $D_{\mathbf{s}}$ is the distortion penalty that the attacker must pay to remove the watermark, and hence, is a figure of merit measuring the trade-off between robustness and embedding-induced distortion at a given rate. Distortion penalties for QIM, spread spectrum, and LBM systems are derived below and are shown in Table 1. A distortion penalty less than 1 (0 dB) indicates that the attacker can actually improve the signal quality and remove the watermark simultaneously.

5.1. Quantization Index Modulation

We first consider the robustness of quantization index modulation. We assume that all reconstruction points $\mathbf{s}$ lie at the centroids of their respective quantization cells.

We use $\mathcal{R}$ to denote the quantization cell containing $\mathbf{x}$ and $E_{\mathbf{x}|\mathcal{R}}[\cdot]$ to denote expectectation taken over the conditional probability density function of $\mathbf{x}$ given that $\mathbf{x} \in \mathcal{R}$.

Table 1: Attacker's distortion penalties. The distortion penalty is the additional distortion that an attacker must incur to successfully remove a watermark.

Embedding System	Distortion Penalty $(D_{\mathbf{y}}/D_{\mathbf{s}})$
Quant. Index Mod.	$1+\frac{1}{4}\frac{d^2_{\min}/N}{D_{\mathbf{s}}} > 0$ dB
Binary Dith. Mod.	$1+\gamma_c\frac{3/4}{NR} > 0$ dB
Spread Spectrum	$-\infty$ dB
Low-bit(s) Modulation	≤ 0 dB

Then, since $\mathbf{s}$ is the centroid of $\mathcal{R}$,

$$E_{\mathbf{x}|\mathcal{R}}[\mathbf{s}-\mathbf{x}] = \mathbf{0}. \tag{8}$$

Also, $D_{\mathbf{s}} = \frac{1}{N}E_{\mathbf{x}|\mathcal{R}}\left[\|\mathbf{s}-\mathbf{x}\|^2\right]$.

The most general attack can always be represented as $\mathbf{y} = \mathbf{s}+\mathbf{n}$, where $\mathbf{n}$ may be a function of $\mathbf{s}$. The resulting distortion is

$$D_{\mathbf{y}} = \frac{1}{N}E_{\mathbf{x}|\mathcal{R}}\left[\|(\mathbf{s}-\mathbf{x})+\mathbf{n}\|^2\right] = D_{\mathbf{s}} + \frac{\|\mathbf{n}\|^2}{N},$$

where we have used (8) to eliminate the cross term $\mathbf{n}^{\mathrm{T}}E_{\mathbf{x}|\mathcal{R}}[\mathbf{s}-\mathbf{x}]$. For a successful attack, $\|\mathbf{n}\| \geq d_{\min}/2$ so our figure of merit for an index modulation system is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} \geq 1+\frac{1}{4}\frac{d^2_{\min}/N}{D_{\mathbf{s}}}. \tag{9}$$

In the special case of coded binary dither modulation with uniform, scalar quantization considered Sec. 4, the distortion is $D_{\mathbf{s}} = \Delta^2/12$, and (6) gives the squared minimum distance, $d^2_{\min} = \gamma_c(\Delta^2/(4R))$. Thus, the attacker's distortion penalty (9) in this case is

$$\boxed{\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} \geq 1+\gamma_c\frac{3/4}{NR},}$$

which we see grows with the strength of the coding applied.

5.2. Spread-spectrum Modulation

The embedding function of a spread-spectrum system is $\mathbf{s} = \mathbf{x}+\mathbf{w}(m)$. Because an attacker with full knowledge of the embedding and decoding processes can decode the message m, the attacker can completely remove the watermark by subtracting $\mathbf{w}(m)$ from $\mathbf{s}$ to obtain the original host signal, i.e., $\mathbf{y} = \mathbf{s}-\mathbf{w}(m) = \mathbf{x}$. Hence, the resulting distortion penalty is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} = \frac{0}{D_{\mathbf{s}}} = -\infty \text{ dB}.$$

Because the spread-spectrum embedding function combines $\mathbf{x}$ and $\mathbf{w}(m)$ in a simple linear way, anyone that can extract the watermark, can easily remove it. In contrast, the quantization that occurs in QIM systems effectively hides the exact value of $\mathbf{x}$ even when m is known.

5.3. Low-bit(s) Modulation

The embedding function of a LBM system can be written as $\mathbf{s} = \mathbf{q}(\mathbf{x})+\mathbf{d}(m)$, where $\mathbf{q}(\cdot)$ represents the coarse quantizer that determines the most significant bits and $\mathbf{d}$ represents the effect of the (modulated) least significant bits. Because the embedding never alters the most significant bits of the host signal, $\mathbf{q}(\mathbf{s}) = \mathbf{q}(\mathbf{x})$. One possible attack is to simply remodulate the least significant bits with some other message m', i.e., $\mathbf{y} = \mathbf{q}(\mathbf{s})+\mathbf{d}(m') = \mathbf{q}(\mathbf{x})+\mathbf{d}(m')$. Since both $\mathbf{s}$ and $\mathbf{y}$ are low-bit(s) modulated versions of $\mathbf{x}$, the distortions must be equal, particularly if the distortions are averaged over all possible choices of m and m'. Thus, the attacker's distortion penalty in this case is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} = 1 = 0 \text{ dB}.$$

This expression does not depend on γ_c so error correction coding does not improve LBM in this context. Finally, the argument above is for a particular attack, not necessarily the best attack, and thus, establishes only an upper bound on the distortion penalty.

6. REFERENCES

[1] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, vol. 86, pp. 1064–1087, June 1998.

[2] B. Chen and G. W. Wornell, "System, method, and product for information embedding using an ensemble of non-intersecting embedding generators." U.S. patent pending. For licensing information, contact: MIT Technology Licensing Office.

[3] B. Chen and G. W. Wornell, "Digital watermarking and information embedding using dither modulation," in *Proc. IEEE Workshop Multimedia Signal Processing (MMSP-98)*, (Redondo Beach, CA), Dec. 1998.

[4] B. Chen and G. W. Wornell, "Dither modulation and quantization index modulation: New methods for digital watermarking and information embedding." Preprint.

[5] E. A. Lee and D. G. Messerschmitt, *Digital Communication*. Kluwer Academic Publishers, 2nd ed., 1994.

[6] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Prentice-Hall, 1984.